



UNIVERSITY OF
LIVERPOOL

Image-based Semantic Segmentation of
Large-scale Terrestrial Laser
Scanning Point Clouds

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of
Doctor in Philosophy

By

Yuanzhi Cai

May 2023

PGR Declaration of Academic Honesty

NAME (Print)	Yuanzhi Cai
STUDENT NUMBER	201218781
SCHOOL/INSTITUTE	School of Engineering
TITLE OF WORK	Image-based Semantic Segmentation of Large-scale Terrestrial Laser Scanning Point Clouds

This form should be completed by the student and appended to any piece of work that is submitted for examination. Submission by the student of the form by electronic means constitutes their confirmation of the terms of the declaration.

Students should familiarise themselves with Appendix 4 of the PGR Code of Practice: PGR Policy on Plagiarism and Dishonest Use of Data, which provides the definitions of academic malpractice and the policies and procedures that apply to the investigation of alleged incidents.

Students found to have committed academic malpractice will receive penalties in accordance with the Policy, which in the most severe cases might include termination of studies.

STUDENT DECLARATION

I confirm that:

- I have read and understood the University's PGR Policy on Plagiarism and Dishonest Use of Data.
- I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study.
- I have not copied material from another source nor committed plagiarism nor fabricated, falsified or embellished data when completing the attached material.
- I have not copied material from another source, nor colluded with any other student in the preparation and production of this material.
- If an allegation of suspected academic malpractice is made, I give permission to the University to use source-matching software to ensure that the submitted material is all my own work.

SIGNATURE.....*Yuanzhi Cai*.....
.....

DATE.....*01/05/2023*.....
.....

Abstract

Image-based Semantic Segmentation of Large-scale Terrestrial Laser Scanning Point Clouds

by Yuanzhi Cai

Large-scale point cloud data acquired using terrestrial laser scanning (TLS) often need to be semantically segmented to support many applications. To this end, various three-dimensional (3D) methods and two-dimensional (i.e., image-based) methods have been developed. For large-scale point cloud data, 3D methods often require extensive computational effort. In contrast, image-based methods are favourable from the perspective of computational efficiency. However, the semantic segmentation accuracy achieved by existing image-based methods is significantly lower than that achieved by 3D methods. On this basis, the aim of this PhD thesis is to improve the accuracy of image-based semantic segmentation methods for TLS point cloud data while maintaining its relatively high efficiency.

In this thesis, the optimal combination of commonly used features was first found, and an efficient manual feature selection method was proposed. It was found that existing image-based methods are highly dependent on colour information and do not provide an effective means of representing and utilising geometric features of scenes in images. To address this problem, an image enhancement method was developed to reveal the local geometric features in images derived by the projection of point cloud coordinates. Subsequently, to better utilise neural network models that are pre-trained on three-channel (i.e., RGB) image datasets, a feature extraction method (LC-Net) and a feature selection method (OSTA) were developed to reduce the higher dimension of image-based features to three. Finally, a stacking-based semantic segmentation (SBSS) framework was developed to further improve segmentation accuracy. By integrating SBSS, the dimension-reduction method (i.e. OSTA) and locally enhanced geometric features, a mean Intersection over Union (mIoU) of 76.6% and an Overall Accuracy (OA) of 93.8% were achieved on the Semantic3D (Reduced-8) benchmark. This set the state-of-the-art (SOTA) for the semantic segmentation accuracy of image-based methods and is very close to the SOTA accuracy of 3D method (i.e., 77.8% mIoU and 94.3% OA). Meanwhile, the integrated method took less than 10% of the processing time (52.64s versus 563.6s) of the fastest SOTA 3D method.

Abstract

Declaration for Authorship

This thesis has not been submitted in support of an application for a degree at this or any other university. It is the result of my own work and does not include any work done in collaboration. Parts of this thesis have been published in the following peer-reviewed journal articles. I confirm that I have the copyright to include these articles in this thesis and that UOL has the right to make this thesis publicly available. The first and third articles were published in open access journals. The second and fourth articles were published in the IEEE, and the IEEE copyright statement can be found in the appendix.

1) **Cai, Y.**, Huang, H., Wang, K., Zhang, C., Fan, L., Guo, F., 2021b. Selecting Optimal Combination of Data Channels for Semantic Segmentation in City Information Modelling (CIM). *Remote Sens.* 13, 1367. <https://doi.org/10.3390/rs13071367> **(Chapter 3)**

2) **Cai, Y.**, Fan, L., Atkinson, P.M., Zhang, C., 2022a. Semantic Segmentation of Terrestrial Laser Scanning Point Clouds Using Locally Enhanced Image-Based Geometric Representations. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. <https://doi.org/10.1109/TGRS.2022.3161982> **(Chapter 4)**

3) **Cai, Y.**, Fan, L., Zhang, C., 2022b. Semantic Segmentation of Multispectral Images via Linear Compression of Bands: An Experiment Using RIT-18. *Remote Sens.* 14, 2673. <https://doi.org/10.3390/rs14112673> **(Chapter 5)**

4) **Cai, Y.**, Fan, L., Fang, Y., 2023a. SBSS: Stacking-Based Semantic Segmentation Framework for Very High-Resolution Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.* 61, 1–14. <https://doi.org/10.1109/TGRS.2023.3234549> **(Chapter 7)**

Declaration for Authorship

I declare that these publications are purely my own work as the first authorship. The ideas and experiments were developed by myself, discussed and approved by my supervisors. The co-authors have permitted these papers to appear in this thesis. The co-authors of these publications have signed below to confirm this.

Signed:

Yuanzhi Cai *Yuanzhi Cai*

Lei Fa *Lei Fa*

Cheng Zhang *Cheng Zhang*

Fangyu Guo *Fangyu Guo*

Peter Atkinson *PA*

Yuan Fang *Yuan Fang*

Hong Huang *Hong Huang*

Kaiyang Wang *Kaiyang Wang*

Acknowledgements

I would like to express my sincere gratitude to my supervisors Dr Lei Fan and Dr Cheng Zhang, for their guidance, support, and encouragement throughout my PhD journey. I would also like to express my sincere appreciation to the civil engineering faculty at XJTLU for providing a nurturing academic environment that has allowed me to grow and develop as a researcher.

Acknowledgements

Table of Contents

PGR Declaration of Academic Honesty	iii
Abstract	v
Declaration for Authorship	vii
Acknowledgements	ix
Table of Contents	xi
List of Abbreviations.....	xvii
List of Figures	xix
List of Tables.....	xxiii
Chapter 1: Introduction	27
1.1. Background	27
1.2. Aim and objectives.....	28
1.3. Thesis layout	30
Chapter 2: Literature Review	33
2.1. Benchmark data.....	33
2.2. Deep learning for point cloud semantic segmentation.....	33
2.2.1. Three-dimensional methods.....	34
2.2.2. Image-based methods.....	36
2.2.3. Performance comparison of 3D and image-based methods on Semantic3D.....	36
2.3. Reflections.....	37
Chapter 3: Manual feature selection	39
3.1. Introduction	40

Table of Contents

3.2.	Materials and Methodology	43
3.2.1.	Paradigms for semantic segmentation.....	43
3.2.2.	Study Materials	44
3.2.3.	Methodology	46
3.2.4.	Experiment arrangement	48
3.3.	Results	51
3.4.	Discussion	53
3.5.	Summary	59
Chapter 4:	Locally enhanced image-based geometric features.....	61
4.1.	Introduction	62
4.2.	Methodology	67
4.2.1.	Study data.....	67
4.2.2.	Segmentation accuracy metrics	68
4.2.3.	Point cloud to image projection	69
4.2.4.	Enhancement of image-based geometric features	72
4.2.5.	Semantic segmentation network structure.....	74
4.2.6.	Pretraining of network and transfer learning.....	77
4.3.	Experiment and results	78
4.3.1.	Information loss from point clouds to images.....	78
4.3.2.	Effect of local enhancement area on the segmentation results.....	80
4.3.3.	Selecting combinations of feature channels	83
4.3.4.	Final performance of HR-EHNet	84
4.4.	Discussion	87
4.5.	Summary	92

Table of Contents

Chapter 5: Automatic feature extraction.....	93
5.1. Introduction.....	94
5.2. Materials and Methods.....	97
5.2.1. Study Data.....	97
5.2.2. LC-Net (contains modifications that do not affect the results in this chapter).....	99
5.2.3. Network Structure.....	101
5.2.4. Training Setting.....	104
5.2.5. Comparisons.....	105
5.3. Results.....	107
5.4. Discussion.....	110
5.5. Summary.....	115
Chapter 6: Automatic feature selection.....	117
6.1. Introduction.....	118
6.2. Methodology.....	121
6.2.1. OSTA.....	121
6.2.2. Establishment of benchmarks.....	123
6.2.3. Evaluation metrics.....	127
6.3. Experiments and results.....	128
6.3.1. Semantic segmentation on benchmarks.....	128
6.3.2. Efficiency of OSTA.....	130
6.3.3. Ablation study.....	131
6.4. Discussion.....	135

Table of Contents

6.4.1.	Most robust combination.....	135
6.4.2.	Coastal aerosol band for cloud detection	137
6.4.3.	Training with channel combinations other than the selected one	138
6.4.4.	Limitation of OSTA	138
6.5.	Summary	139
Chapter 7:	Stacking-based semantic segmentation framework	141
7.1.	Introduction	142
7.2.	SBSS framework.....	146
7.2.1.	Overview of SBSS framework.....	146
7.2.2.	Error correction module	148
7.2.3.	Error correction scheme	150
7.3.	Experiments and results	153
7.3.1.	Datasets and implementation details.....	153
7.3.2.	Scale related segmentation error	154
7.3.3.	Segmentation network choice	156
7.3.4.	Ablation study on the test setting.....	156
7.3.5.	Complexity and the speed of the SBSS framework.....	158
7.3.6.	Quantitative results on Cityscapes, UAVid, LoveDA and Potsdam test sets	160
7.3.7.	Qualitative Analysis of the Segmentation Results	163
7.3.8.	Exploring the potential for higher accuracy	167
7.3.9.	Integrating developed methods for the semantic segmentation of TLS point clouds (additional results to the published version).....	168
7.4.	Future work	170

Table of Contents

7.5. Summary 171

Chapter 8: Conclusion 173

8.1. Key results..... 173

8.2. Future work 177

References 179

Publications 219

Table of Contents

List of Abbreviations

CIM	City Information Modelling
TLS	Terrestrial Laser Scanning
LiDAR	Light Detection And Ranging
ALS	Aerial Laser Scanning
MLS	Mobile Laser Scanning
mIoU	mean Intersection over Union
OA	Overall Accuracy
CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
SGS	Supervised Grid Search
DF	Direct Feed
PCA	Principal Component Analysis
SSN	Semantic Segmentation Network

List of Abbreviations

List of Figures

Figure 3-1: Detailed steps of the data preprocess.	46
Figure 3-2: The entropy of different channel data.	49
Figure 3-3: Mean intersection over union (mIoU) on test point clouds.....	52
Figure 3-4: Overall accuracy (OA) on test point clouds.....	52
Figure 3-5: Training accuracy for combinations of 8C (all the channels), RGB (color), IRGB (intensity and color), and IRGBD (intensity, color and depth) using networks of Inception-ResnetV2 backbone.....	56
Figure 3-6: Feature maps and segmentation results for four combinations for the building-road joint image.....	57
Figure 3-7: Feature maps and segmentation results for four combinations in the street view image.....	57
Figure 3-8: Summary of the average time of single training for nine network structures.	58
Figure 4-1: The distribution of the classes of points in Semantic3D dataset.....	69
Figure 4-2: Key stages in the projection process: (a). The raw input point cloud, (b). All points scaled to a spherical surface at a distance of 1 from the origin (i.e., the center of the scanner), (c). The panoramic image rasterized from the spherical surface.....	70
Figure 4-3: Illustrations of image enhancement effects: (a). The panoramic image projected from RGB channels, (b). The panoramic image projected from Z coordinate, (c). A local RGB image extracted from the box in (a), (d). The distribution histogram of the pixel values in (c), (e). The local Z coordinate image extracted from the box in (b), (f). The distribution histogram of the	

List of Figures

pixel values in (e), (g). The enhanced local Z coordinate image. (h) The distribution histogram of the pixel values in (g), (i). The enhanced Z coordinate image without overlapping. (j). The enhanced Z coordinate image with overlapping.....	72
Figure 4-4: Illustration of the HR-EHNet network structure: upsampling and downsampling were implemented by bilinear interpolation and strided 3×3 convolution, respectively; The colored blocks that represent multiple residual convolution operations were performed.	75
Figure 4-5: Plot of accuracy (OA and mIoU) versus angular resolution.	79
Figure 4-6: Effects of an excessive angular resolution on the projected image: (a). Many black empty pixels for an angular resolution of $1/50$ degree, (b). A continuous image without empty pixels for an angular resolution of $1/20$ degree.	80
Figure 4-7: Impacts of the local enhancement area on the enhancement results: (a). 128×128 pixels, (b). 32×32 pixels, (c). 8×8 pixels.....	81
Figure 4-8: Impacts of the local enhancement area on the enhancement results: (a). OA, (b). mIoU.....	82
Figure 4-9: (a). The pseudo color images of <i>IZeDe</i> feature channels for the point clouds in Semantic3D (reduced-8) test set, (b). The corresponding segmentation results (The legend is only for the visualization of the segmentation results in (b)).	86
Figure 4-10: Incorrect RGB information in TLS point cloud data: (a). The RGB image contains the cyclist that were not scanned by TLS, (b). The enhanced Z image for the same scene.	88

Figure 4-11: Comparisons between *IZeDe* and *IRGBZeDe* on segmentation results for three scenes: (a). The pseudo color images of *IZeDe*, (b). The segmentation results using the corresponding *IZeDe* images, (c) The RGB images, (d). The segmentation results using the feature combination of *IRGBZeDe*..... 90

Figure 5-1: Percentage of each class in RIT-18: (a) training image (%), (b) test image (%). 98

Figure 5-2: Linear combinations of two adjacent bands without any overlap for the RIT-18 dataset. 100

Figure 5-3: Apply LC-Net to the networks. 101

Figure 5-4: Basic block designs for ResNet, HRNet, and Swin. 102

Figure 5-5: RIT-18 segmentation results for Swin-tiny on the test image: (a) ground truth, (b) segmentation map using the DF approach, (c) segmentation map using the SGS approach, (d) segmentation map using LC-Net..... 109

Figure 5-6: RIT-18 segmentation results of the test image: (a) ground truth, (b) segmentation map from ResNet50+ LC-Net, (c) segmentation map from HRNet-w18+ LC-Net, (d) segmentation map from Swin-tiny+ LC-Net. 112

Figure 5-7: (a) RGB image (bands 3, 2, and 1) of the RIT-18 test image; (b) pseudo color image of bands 4–6 of the RIT-18 test image. 113

Figure 5-8: The effect of the addition of LC-Net on the segmentation accuracy of different classes against the DF approach (%). 114

Figure 6-1: Percentages of training iterations of the three stages in OSTA and the learning rate schedule. 122

Figure 6-2: Semantic segmentation results on benchmark data. 129

List of Figures

Figure 7-1: Workflow of one iteration of the SBSS framework, in which Y_i and X_{i+1} are the input maps; $Y_{i \rightarrow i+1}, s, c$ is the output map of one iteration and the segmentation map (i.e., a new Y_i) for the next iteration; the loop is ended when $i+1 = n$ (i.e., the number of scales used in SBSS); the area(s) A are determined by the error correction scheme, and are also applied to X_{i+1} and $Y_{i \rightarrow i+1}$	147
Figure 7-2: Error correction module.	149
Figure 7-3: Structures of the error correction network (left), the stem block (middle), and the residual block (right).....	149
Figure 7-4: Qualitative comparisons between MS and SBSS-MS on the Cityscapes, UAVid, LoveDA and Potsdam validation sets.	158
Figure 7-5: Segmentation accuracy (IoU %) of buildings using different resizing scales. The horizontal coordinates (of the top 4 plots) for SS refer to the resizing scale used. The horizontal coordinates (of the bottom 4 plots) for MS and SBSS represent the utilisation of all the scales that are equal to and smaller than the current scale (e.g., the coordinate value 1 represents the case where the scales 0.5, 0.75 and 1 were all used).	164
Figure 7-6: Visual comparisons between SS, MS and SBSS-MS on the UAVid validation set. The numbers in brackets represent the resizing scales used.	165
Figure 7-7: Visual comparisons between SBSS-MS using different set of resizing scales on the UAVid validation set.	166
Figure 7-8: Qualitative comparisons between SBSS-MS with different setting on the LoveDA validation set.....	168

List of Tables

Table 2-1: Summary of point cloud semantic segmentation datasets	33
Table 2-2: Quantitative comparison of existing methods on Semantic3D (Reduced-8) (%).....	36
Table 3-1: Summary of basic information of 15 labelled point clouds.....	45
Table 3-2: Combinations of channels.	50
Table 3-3: Average improvement by adding different channels.....	53
Table 3-4: The mIoU of seven networks regarding different combinations of channels (the highest mIoU for each network is marked as green).	54
Table 3-5: The OA of seven networks regarding different combinations of channels (the highest OA for each network is marked as yellow).....	54
Table 3-6: Ranking of the mIoU performance of 13 channel combinations for seven networks.	55
Table 3-7: Ranking of the OA performance of 13 channel combinations for seven networks.	55
Table 3-8: Quantitative results of different approaches on Semantic3D (reduced- 8). Accessed on 16 March 2021 (the overperformed methods are marked in grey).	59
Table 4-1: quantitative results of different channel combinations on the semantic3d training set (five-fold cross-validation).	83
Table 4-2: Impacts of retaining or replacing the first layer of the pre-trained network on the segmentation results when <i>IZeDe</i> were used as the input channels (five-fold cross-validation).....	85

List of Tables

Table 4-3: Quantitative results (%) of different approaches on Semantic3D (reduced-8).	86
Table 4-4: The times taken by each step of HR-EHNet to process the Semantic3D (reduced-8) test dataset.....	87
Table 5-1: Summary of five commonly used multispectral datasets.	98
Table 5-2: Detailed architecture specifications of ResNet50 and Swin-tiny, where the bracket indicates a residual block, and the number outside the brackets is the number of stacked blocks for the stage.....	103
Table 5-3: Detailed architecture specifications of HRNet-w18, where the bracket indicate a residual block, and the number outside the brackets is the number of stacked blocks for the stage.	104
Table 5-4: Summary of the key performance of PCA, DF, LC-NET, and CoinNet (%).	107
Table 5-5: Comparison of accuracies and computational cost using PCA, DF, LC-Net, and SGS (%). The numbers shown in the “Input Method” column for SGS indicate which three bands were used as the input images to the subsequent networks.	108
Table 5-6: Performance of LC-Net and LCAB (%).	110
Table 5-7: Final weights in LC-Net used in the three networks considered.	114
Table 6-1: Summary of benchmark data used.	125
Table 6-2: Evaluation metrics used in OSTA.	128
Table 6-3: Summary of the semantic segmentation results on benchmark data.	130
Table 6-4: Efficiency metrics and calculation of OSTA.....	131
Table 6-5: Results of replacing pruning criteria.....	132

Table 6-6: Results of replacing pruning strategy.....	132
Table 6-7: Results of removing linear warmup and/or supernet training stage.	133
Table 6-8: Results of directly fine-tuning the trained supernet with selected combinations by OSTA.....	133
Table 6-9: Results of replacing for pretrained weights in SSN.	134
Table 6-10: Summary of recurring channel combinations in the Top 10 of SGS.	135
Table 6-11: Detailed segmentation results for combinatio in Table 6-10.....	135
Table 6-12: Replication of Table 6-11, re-arranged in the vertical direction according to the combination used. The bolded one indicate the best accuracy for that class were achieved by using corresponding combination.	137
Table 6-13: Top 10 in SGS for cloud detection benchmark data.....	138
Table 7-1: The explanations of the abbreviations used in Figure 7-1.....	147
Table 7-2: Comparison of the total size of the images to be processed by MS and SBSS-MS.	151
Table 7-3: Comparison of the total size of the images to be processed by SS and SBSS-SS.....	152
Table 7-4: Summary of four datasets used.....	153
Table 7-5: Training setting for the segmentation network.....	154
Table 7-6: Input scales that achieve the highest segmentation accuracy for different classes.....	155
Table 7-7: Segmentation accuracy (mIoU) on validation sets using single scale tests (%).	156

List of Tables

Table 7-8: The quantitative results of the ablation studies on validation sets of four datasets.....	157
Table 7-9: Comparison of the efficiency of the SBSS framework with other methods on cityscapes validation set.	159
Table 7-10: Quantitative comparison results on the cityscapes test set. the input patch sizes used in SBSS-MS and SBSS-SS are 1024×512 and 512×256 respectively.....	161
Table 7-11: Quantitative comparison results on the uavid test set (%).	162
Table 7-12: Quantitative comparison results on the loveda test set (%).	162
Table 7-13: Quantitative comparison results on the potsdam test set (%).	163
Table 7-14: Quantitative comparison for SBSS-MS using different input methods on the LoveDA test set (%).	167
Table 7-15: Training setup for the integrated method.....	169
Table 7-16: Quantitative comparison of different methods on Semantic3D (Reduced-8) (%).	169

Chapter 1: Introduction

1.1. Background

Terrestrial laser scanning (TLS) can acquire up to hundreds of millions of millimetre-accurate three-dimensional (3D) data points (i.e., point cloud) within minutes. This data acquisition technology is used in many applications in civil engineering, such as building/city information modelling (BIM/CIM) and structure/slope deformation monitoring. All of which can benefit from semantic segmentation (i.e., assign a class label to each data point) of TLS point clouds. For example, accurate semantically segmented point clouds can be used to improve the speed and accuracy of registration, generate semantically enhanced BIM/CIM models, and automate deformation monitoring.

Existing methods for semantic segmentation of point clouds can be divided into two categories based on the dimension of data representation (e.g., point, voxel and pixel) used by their classifiers, namely 3D methods and two-dimensional (2D) methods (i.e., image-based methods). Because it is inherently more suitable to represent the spatial information of a point cloud in 3D, 3D methods often achieve higher segmentation accuracies than those of image-based methods. However, 3D methods are relatively inefficient, with their processing time growing exponentially with the volume of input point cloud data. Image-based methods project a point cloud as multichannel image(s) (different channels contain different features) and then segment multichannel image(s) using 2D semantic segmentation networks. The segmentation accuracies of image-based methods are relatively lower due to the projection-induced loss and distortion of

Chapter 1: Introduction

spatial information of the point cloud. However, thanks to high computational efficiency of 2D semantic segmentation networks, the processing time required by image-based methods can be several orders of magnitude lower than that of 3D methods.

In summary, semantic segmentation of point clouds using existing methods suffers from either low accuracy (image-based methods) or low efficiency (3D methods). However, achieving accurate and efficient semantic segmentation of TLS point clouds is highly desirable for civil engineering applications, especially for time-sensitive tasks such as deformation monitoring.

1.2. Aim and objectives

The aim of this thesis is to establish new approaches to improve the accuracy of image-based methods for semantic segmentation of TLS point clouds while high computational efficiency is maintained. To achieve this aim, the following objectives and questions are considered.

Objective 1. Select the optimal feature combination from commonly used image-based features for semantic segmentation of TLS point clouds.

Question 1-1: Can higher semantic segmentation accuracy be achieved using fewer features than using all available features?

Question 1-2: Can an efficient manual feature selection method be developed to select the optimal feature combination?

Objective 2. Develop novel image-based geometric features to improve segmentation accuracy of TLS point clouds.

Question 2-1: Can novel image-based geometric features be developed to improve segmentation accuracy?

Question 2-2: Can accurate semantic segmentation of TLS point clouds be achieved without using colour information?

Question 2-3: How to better utilise a model pre-trained on a large-scale RGB image dataset to improve segmentation accuracy?

Objective 3. Develop novel dimension reduction methods to transform multichannel images into 3-channel images to better utilise model(s) pre-trained on large-scale RGB image datasets.

Question 3-1: For multichannel images that are not derived from TLS point cloud data, e.g., multispectral images, is it still true that using an appropriate 3-channel combination will give a higher semantic segmentation accuracy than using all available channels?

Question 3-2: Can a novel feature extraction method be developed to avoid repetitive testing of different channel combinations while achieving semantic segmentation accuracy comparable to the optimal 3-channel combination for multichannel images?

Question 3-3: Can a novel feature selection method be developed to efficiently and automatically select the optimal 3-channel combination for the semantic segmentation of multichannel images?

Question 3-4: For the developed feature extraction method and feature selection method, which one gives a higher semantic segmentation accuracy?

Chapter 1: Introduction

Objective 4. Develop a novel framework to improve the semantic segmentation accuracy of images.

Question 4-1: Can a novel image semantic segmentation framework be developed to improve segmentation accuracy?

Question 4-2: How accurate can the image-based semantic segmentation method be achieved on TLS point clouds by integrating the methods developed in this thesis?

1.3. Thesis layout

The organisation of the subsequent chapters of this thesis is illustrated in Figure 1-1, starting with a literature review chapter, followed by five chapters corresponding to the five methods developed for the four objectives in Section 1.2, and ending with a conclusion chapter. The content of these chapters is summarised as follows.

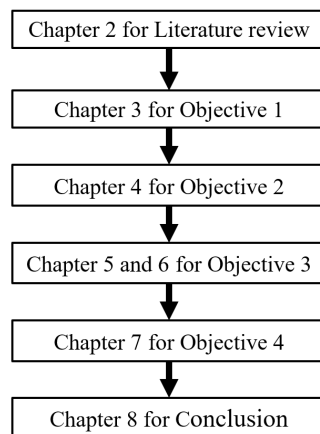


Figure 1-1: Flowchart for achieving each objective.

Chapter 2: Literature review

This chapter provides a concise overview of deep learning-based semantic segmentation methods for point clouds, discusses their pros and cons, and reflects on potential improvement for image-based methods.

Chapter 3: Manual feature selection (published)

This chapter presents a manual feature selection framework for the semantic segmentation of multichannel images. The optimal combination of commonly used features is selected for image-based semantic segmentation of TLS point clouds. The importance and patterns of feature selection are discussed.

Chapter 4: Locally enhanced image-based geometric features (published)

This chapter presents locally enhanced image-based geometric features for the semantic segmentation of TLS point clouds. The feasibility of excluding colour information and the importance of retaining the first layer's weights of the pre-trained model (i.e., the importance of dimension reduction) were investigated.

Chapter 5: Automatic feature extraction (published)

This chapter presents an automatic feature extraction method that compresses multispectral images into 3-band images for semantic segmentation. The semantic segmentation accuracies achieved using all bands and using different 3-band combinations are also presented for the multispectral dataset used.

Chapter 6: Automatic feature selection (preparing for publication)

This chapter presents an automatic feature selection method for the semantic segmentation of multichannel images. For the case of using a 3-channel image as input, the feature selection method developed in this chapter and the feature extraction

Chapter 1: Introduction

method developed in Chapter 5 are compared on a multichannel benchmark dataset derived from TLS point clouds and a multispectral benchmark dataset.

Chapter 7: Stacking-based semantic segmentation framework (published)

This chapter presents a stacking-based semantic segmentation framework that improves segmentation accuracy by learning the preferred resizing scales for different object classes. Furthermore, for the image-based semantic segmentation of TLS point clouds, the accuracy achieved by integrating methods developed in this thesis is demonstrated in the results section of this chapter.

Chapter 8: Conclusion

This chapter summarises the key results from Chapters 3 to 7 including the answers to the research questions in Chapter 1, followed by key recommendations for future work.

Chapter 2: Literature Review

2.1. Benchmark data

Several datasets have been established to evaluate the performance of deep learning algorithms for point cloud segmentation. The characteristics of these datasets are summarised in Table 2-1. These datasets are acquired by different types of sensors, including RGB-D cameras, Mobile Laser Scanners (MLS), Aerial Laser Scanners (ALS) and Terrestrial Laser Scanners (TLS).

Table 2-1: Summary of point cloud semantic segmentation datasets

	Sensors	Total points (M)	Scans	Points per scans (M)	RGB
Oakland (Munoz et al., 2009)	MLS	2	17	0.1	No
Totonto-3D (Tan et al., 2020)	MLS	78	4	19.5	Yes
Paris-Lille-3D (Roynard et al., 2018)	MLS	143	3	47.7	No
S3DIS (Armeni et al., 2016)	RGB-D	273	272	1.0	Yes
IQmulus (Vallet et al., 2015)	MLS	300	10	30.0	No
DALES (Varney et al., 2020)	ALS	505	40	12.6	No
Semantic3D (Hackel et al., 2017)	TLS	4009	30	133.3	Yes
SemanticKITTI (Behley et al., 2019)	MLS	4549	43552	0.1	No

As the only TLS dataset, Semantic3D is used in this thesis. It contains a training set and a test set, each having 15 annotated point clouds collected in urban scenes. The points are labelled as eight classes (i.e., man-made terrain, natural terrain, high vegetation, low vegetation, buildings, hard scape, scanning artefacts, cars).

2.2. Deep learning for point cloud semantic segmentation

The goal of semantic segmentation of point clouds is to classify points into subsets according to their semantics. It involves learning both global features of a point cloud and fine-grained details of each data point. As mentioned in Chapter 1, existing segmentation methods can be categorised into 3D methods and image-based methods. The evolution of 3D and image-based methods is briefly introduced in Sections 2.2.1

Chapter 2: Literature Review

and 2.2.2, respectively, and their performances in terms of accuracy and efficiency on Semantic3D are shown in Section 2.2.3.

2.2.1. Three-dimensional methods

Depending on the type of main segmentation network/module used, 3D methods can be divided into point MLP methods, point convolution methods, graph-based methods and 3D CNN methods.

2.2.1.1. Point MLP methods

The pioneering work of point MLP methods is PointNet (Charles et al., 2017). In PointNet, point-wise (local) features are learned using weight-sharing MLP, and global features are learned with symmetric pooling. A series of point MLP methods have been developed subsequently, with a focus on the development of better global feature learning techniques. Representative ones include neighbouring feature pooling (Engelmann et al., 2019; Hu et al., 2021; Jiang et al., 2018; Qi et al., 2017; Zhang et al., 2019; H. Zhao et al., 2019), attention-based aggregation (Chen et al., 2019; Wen et al., 2020; Jiancheng Yang et al., 2019; C. Zhao et al., 2019), local-global feature concatenation (Arandjelovic et al., 2018; Wen et al., 2020; Y. Zhao et al., 2019) and recurrent neural modules (Engelmann et al., 2017; Huang et al., 2018; Ye et al., 2018).

2.2.1.2. Point convolution methods

Point convolution methods aim to develop convolution operators that can handle a raw point cloud (Hua et al., 2018; Jeppesen et al., 2019; S. Wang et al., 2018). The most famous one is Kernel Point Convolution (KPConv) (Thomas et al., 2019). The convolution weights of KPConv are based on Euclidean distances between kernel points. The locations of kernel points are chosen based on an optimisation function for maximizing the coverage of a spherical space. Based on KPConv, researchers have developed a series of modified methods (Lai et al., 2022; Y. Li et al., 2022; Lin et al.,

2021; Liu et al., 2020; M. Xu et al., 2021; Yan et al., 2022). Similar to the PointNet-based methods, these modified methods focus on better aggregation of neighbouring features to learn global features. Techniques used are similar to those used in PointNet-based methods.

2.2.1.3. Graph-based methods

The first graph-based method is superpoint graph (SPG) (Landrieu and Simonovsky, 2018). SPG segments a point cloud in three steps: geometrically homogeneous partitioning of the original point cloud space, embedding original points as superpoints, and segmenting the superpoint graph with graph neural networks. Follow-up studies have made substantial improvements on these three steps. For example, supervised oversegmentation (Landrieu and Boussaha, 2019), graph embedding module (Zhiheng and Ning, 2019) and attention mechanisms (C.-Q. Huang et al., 2022; L. Wang et al., 2019; Wen et al., 2021; Zhiheng and Ning, 2019) have been used for better partition, embedding and feature aggregation, respectively.

2.2.1.4. 3D CNN methods

Huang and You (Jing Huang and Suya You, 2016) are the first researchers to represent a raw point cloud in voxels and segment voxels with standard 3D CNN. The subsequent study can be grouped into two themes. The first (main) theme is to reduce computational consumption (Meng et al., 2019), in which research is focused primarily on the development and utilisation of sparse discretization representations (Choy et al., 2019; Graham et al., 2018; Park et al., 2023; Tang et al., 2020; You et al., 2020). The second theme is similar to other types of methods, i.e. improving the ability of 3D CNN to aggregate features (Yuhong Chen et al., 2022).

Chapter 2: Literature Review

2.2.2. Image-based methods

Image-based methods first project a 3D point cloud as 2D image(s) and then use 2D networks to perform segmentation. Initially, researchers projected point clouds from multiple views (Boulch et al., 2017; Lawin et al., 2017; Tatarchenko et al., 2018). This requires the processing of multiple images to segment a point cloud from a single scan, which is less efficient. Therefore, projecting point clouds as spherical panoramic images, with the scanning device as the centroid, has become the dominant method in subsequent studies (Milioto et al., 2019; B. Wu et al., 2018; Wu et al., 2019).

2.2.3. Performance comparison of 3D and image-based methods on Semantic3D

The performances of the existing methods on Semantic3D are summarised in Table 2-2.

Table 2-2: Quantitative comparison of existing methods on Semantic3D (Reduced-8) (%).

		Time (s)	Params (M)	mIoU	OA	man- made	natural.	high veg	low veg	buildings	hard scape	Scanning art	cars
3D methods	RF MSSF (Thomas et al., 2018)	1643.75	-	62.7	90.3	87.6	80.3	81.8	36.4	92.2	24.1	42.6	56.6
	ShellNet (Zhang et al., 2019)	3000	0.48	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
	OctreeNet (F. Wang et al., 2020)	184.84	-	59.1	89.9	90.7	82.0	82.4	39.3	90.0	10.9	31.2	46.0
	GACNet (L. Wang et al., 2019)	1380	-	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
	SPGraph (Landrieu and Simonovsky, 2018)	3000	0.25	73.2	94.0	97.4	92.6	87.9	44.0	83.2	31.0	63.5	76.2
	KPConv (Thomas et al., 2019)	600	14.9	74.6	92.9	90.9	82.2	84.2	47.9	94.9	40.0	77.3	79.7
	RandLA-Net (Q. Hu et al., 2020)	-	0.95	77.4	94.8	95.6	91.4	86.6	51.5	95.7	51.5	69.8	76.8
	SCF-Net (Fan et al., 2021)	563.6	-	77.6	94.7	97.1	91.8	86.3	51.2	95.3	50.5	67.9	80.7
	RFCR (Gong et al., 2021)	-	-	77.8	94.3	94.2	89.1	85.7	54.4	95.0	43.8	76.2	83.7
Image- based methods	DeePr3SS (Lawin et al., 2017)	-	134	58.5	88.9	85.6	83.2	74.2	32.4	89.7	18.5	25.1	59.2
	SnapNet (Boulch et al., 2018)	3600	29	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4

Since most studies reported their performances on Semantic3D's reduced-8 test set, this test set is also used in this thesis. The reduced-8 test set is popular because it contains only 3% of the data points of the standard test set, which allows less efficient methods to complete the test within a reasonable period of time. The accuracies of most 3D methods are much higher than those of image-based methods. This has led most researchers to focus on improving 3D methods. There are only two image-based methods (based on multi-view projection), the performances of which on Semantic3D are reported. Their performances in terms of accuracy and efficiency are far inferior to those of the state-of-the-art (SOTA) 3D methods.

2.3. Reflections

Although the SOTA 3D methods are far more accurate and efficient than image-based methods according to Table 2-2, it is still argued that image-based methods have great potential for the following reasons.

Firstly, an image-based method with the spherical projection is much more efficient than 3D methods, but has not been used by researchers for semantic segmentation of TLS point clouds. For example, SqueezeSeg took only approximately 8.7 milliseconds to process a single point cloud (B. Wu et al., 2018). In contrast, improving efficiency of 3D methods is extremely difficult. This is because aggregating local features to learn global features in 3D methods has to rely on nearest neighbour search or memory-hungry index structure. The computational consumption of either technique grows exponentially with the number of processed points. This limits the use of 3D methods for efficient semantic segmentation of high-density TLS point clouds in practical applications.

Chapter 2: Literature Review

Secondly, it is thought that image-based methods still have great potential in terms of accuracy. This judgement is inspired by the fact that image segmentation networks can achieve up to 86% mIoU on the Cityscapes image dataset (Cordts et al., 2016), which is also collected in urban scenes and includes even finer classes (more classes) to be segmented. Despite this, existing image-based methods have two notable weaknesses. First, they do not use SOTA image segmentation networks. Second, they often use all the possible features available for segmentation and do not select more suitable image features for segmentation. Therefore, this thesis begins with a detailed investigation (Chapter 3) to address those two issues. The study in the subsequent chapters (Chapters 4-7) is based on new findings and reflections emerging during the course of the research, as outlined below. The experiments in Chapter 3 suggest that image-based methods cannot effectively learn geometric features from existing image features. Therefore, novel locally enhanced image-based geometric representations are developed in Chapter 4. The experiments in Chapter 4 show that using the appropriate three-channel combination can make better use of the pre-trained model to achieve higher segmentation accuracy. This motivates the study of dimension reduction methods in Chapters 5 and 6. Finally, considering the very high resolution (VHR) of the spherical panoramic images projected from the TLS point clouds, in Chapter 7, a novel semantic segmentation framework is developed to improve the segmentation accuracy of VHR images.

Chapter 3: Manual feature selection

This chapter is based on the published paper: Cai, Y., Huang, H., Wang, K., Zhang, C., Fan, L., Guo, F., 2021. Selecting Optimal Combination of Data Channels for Semantic Segmentation in City Information Modelling (CIM). *Remote Sens.* 13, 1367.

<https://doi.org/10.3390/rs13071367>

Note: The research presented in this chapter improves the segmentation accuracy of image-based methods by using state-of-the-art image segmentation network and manually selecting the best feature combination.

3.1. Introduction

Over the last decade, the concept of city information modelling (CIM) has received a growing interest in many fields, such as surveying engineering and civil engineering (Stojanovski, 2018). Generally, CIM provides valuable benefits for stakeholders, including enhancing the public management process and establishing an intelligent digital platform to store, control, and understand big data. Xu et al., (2014) suggested that geographic information systems (GIS) and building information modelling (BIM) can be integrated to facilitate and achieve the CIM concept. GIS models are utilized to represent graphical and geometrical information, while BIM models are applied to characterize semantic and topological information. Nonetheless, issues of model accuracy and timely information update are challenging (Lu and Lee, 2017). Furthermore, it is challenging to automatically identify the discrepancies between the as-built and as-planned models, which would cause significant delays, for example, in responding to project modification management (Golparvar-Fard et al., 2011; Kim et al., 2020).

A popularly used technique to create the as-built model is the 3D reconstruction, which has been developed to present the latest as-is information for infrastructures and the city. To acquire the point cloud data, laser scanning technologies such as light detection and ranging (LiDAR), terrestrial laser scanning system (TLS), and aerial laser scanning system (ALS) have been usually adopted. Many studies have presented that the main advantages of the TLS technologies include high point density (about one billion points per scan) and high geometric accuracy (up to millimetres) (Badenko et al., 2019; Bernat, 2014). Therefore, TLS is more appropriate for CIM applications

(requires high accuracy and density data). In addition, unlike the data collected in the typical remote sensing applications (e.g., satellite images and ALS), the TLS can collect image information of the scene immediately after completing the laser scanning. With the coordinate transformation matrix (usually provided by the manufacturer), the colour information (RGB) can be directly mapped to the corresponding laser point. In this case, the data obtained through TLS usually has seven aligned channels of data: RGB from the camera sensor and XYZ and I (intensity).

After acquiring raw point clouds that can provide accurate geometric information for CIM, the semantic segmentation technique is usually adopted to obtain the semantic information from the raw point cloud. In addition to the seven channels in the raw point cloud, additional channels can be derived to describe the scene. The widely used two types are the depth channel and the normal vector channel generated by XYZ. However, in practice, not all channels can bring a positive improvement to semantic segmentation. Several studies in the remote sensing application have indicated the importance of selecting an optimal combination of data channels regarding multispectral datasets. For instance, Xie et al., (2018) presented a novel hyperspectral band approach to select an optimal band for image classification based on clustering-based selection methods. Their results indicated that the proposed method was more effective and able to generate better band selection results. Li et al., (2018) utilized discrete particle swarm optimization to model the various errors (i.e., reconstruction, imaging, and demosaicing errors) associated with spectral reconstruction for optimal channel combination. The optimization results reduced the time in the computational process. Abdalla et al., (2019) developed a robust DL method to group the RGB channels for automatic colour calibration for plants. Bhuiyan et al.,

Chapter 3: Manual feature selection

(2020) experimented with testing the optimal three-channel combination in model prediction using very high spatial resolution (VHSR) multispectral (MS) satellite images. Their findings emphasized the importance of considering input MS channels and the careful selection of optimal channels of DL network predictions for mapping applications. Park et al., (2020) presented a novel image prioritization method to select the limited channel based on cloud coverage for nanosatellite application. By reducing the channels, they achieved an extremely low computational power and light network on a nanosatellite.

The abovementioned studies have provided insightful guidance for optimal channel combinations for image channels. However, these researches mainly investigated the optimal combination of channels in land-use mapping, agricultural, and disaster monitoring, focusing on the region highlight field (e.g., icewedge polygons). There is no agreement on the optimal combination of channels that should be used for CIM applications in urban scenes. For example, Pierdicca et al., (2020) presented the deep learning (DL) framework using 12 channels as input: XYZ coordinates, X'Y'Z' normalized coordinates, colour features (HSV channels), normal features (in X, Y, and Z direction) for cultural heritage point cloud segmentation. Alshawabkeh, (2020) developed a novel dataset to evaluate the feasibility of combined LiDAR data and images for object segmentation by integrating RGBD channels (i.e., colour and depth information). In the joint 3D object detection and semantic segmentation, Meyer et al., (2019) used RGB together with aligned LiDAR information (point's range, height, azimuth angle, intensity, and indication of occupation) as the input of their networks. Lawin et al., (2017) transformed the XYZ channels into depth and normal information

and particularly investigated the improvements in 3D semantic segmentation by using the depth, colour, and normal information.

Thus, the present chapter aims to explore a simple optimal combination of data channels based on their semantic segmentation performance in the urban scenario. To more objectively evaluate the gain from the combination of channels, the performance of various channel combinations will be tested on different published encoder-to-decoder segmentation networks in this study. Objectives are set to accomplish the aim as follows: (1) To determine the optimal group of channels in terms of its overall accuracy (OA) and mean intersection over union (mIoU); and (2) to empirically verify the robustness of the optimal channel combination across different networks.

The remainder of this chapter is organized as follows. Section 3.2 will introduce the selected benchmark dataset and the proposed framework and experiment arrangement for the optimal channel combination selection. Then, the performance of various channel combinations on different networks is summarized in Section 3.3. Findings are drawn in Section 3.4 and Section 3.5.

3.2. Materials and Methodology

3.2.1. Paradigms for semantic segmentation

According to the comprehensive survey proposed by Guo et al., (2021), point cloud semantic segmentation approaches in the DL framework can be divided into three paradigms: Projection-based, point-based, and discretization-based. The projection-based methods usually project a 3D point cloud into 2D images, including multi-view and spherical images. The point-based methods directly work on irregular point clouds

Chapter 3: Manual feature selection

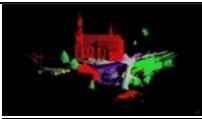
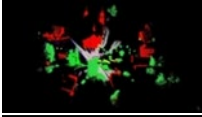
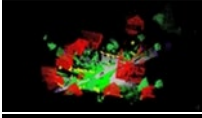

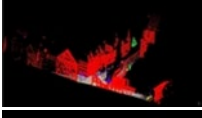

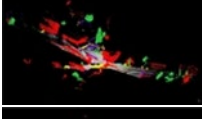
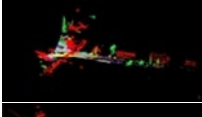
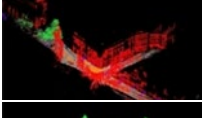
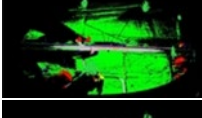
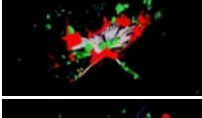
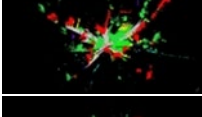
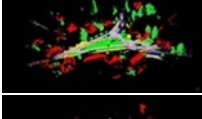
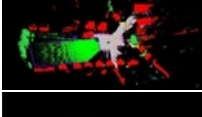
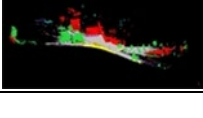
by applying dedicated local feature convolutions. The discretization-based methods usually convert a point cloud into volumetric rasterization to create an ordered grid of point clouds.

The point-based and discretization-based approach is directly processed on the 3D data, which is extremely time-consuming or memory-costly in sampling training and inferencing. For example, in the work of RandLA-Net (Q. Hu et al., 2020), they evaluate the time consumption of recent representative works on Sequence 08 of the SemanticKITTI with 81,920 total number of points, where the best test result was 442 points/s. On the contrary, the SnapNet (Boulch et al., 2018) test 30 M points used even worse arithmetic. The average processing time is about 32 min, and the corresponding process speed is about 15,625 points/s, which is 35 times faster than the point-based method. Meanwhile, for the CIM application, the total number of points is up to 10^8 per scan. Therefore, the point-based and discretization-based approaches are not efficient enough in terms of time. On the other hand, the performance of multi-view segmentation methods is dependent on viewpoint selection and occlusions. Therefore, in this chapter, spherical image-based semantic segmentation is adopted.

3.2.2. Study Materials

The online large-scale point cloud segmentation benchmark dataset Semantic3D is used in this case study (Hackel et al., 2017). This benchmark dataset contains 15 annotated point clouds representing different city scenes, where the points are labelled as eight classes (i.e., 1: Man-made terrain, 2: Natural terrain, 3: High vegetation, 4: Low vegetation, 5: Buildings, 6: Hard scape, 7: Scanning artefacts, 8: Cars). Each point cloud is obtained by a separate scanning. The basic information of 15 labelled point clouds is summarized in Table 3-1.

Table 3-1: Summary of basic information of 15 labelled point clouds.

Index	Preview	Name	Number of Points	Description	Propose
1		bildstein1	29302501	church in bildstein	Train
2		bildstein3	23765246	church in bildstein	Test
3		bildstein5	24671679	church in bildstein	Train
4		domfountain1	35494386	cathedral in feldkirch	Train
5		domfountain2	35188343	cathedral in feldkirch	Test
6		domfountain3	35049972	cathedral in feldkirch	Train
7		untermaederbrunnen1	16658648	fountain in balgach	Train
8		untermaederbrunnen3	19767991	fountain in balgach	Test
9		neugasse	50109087	neugasse in st. gallen	Test
10		sg27_1	161044280	railroad tracks	Train
11		sg27_2	248351425	town square	Train
12		sg27_4	280994028	village	Test
13		sg27_5	218269204	crossing	Train
14		sg27_9	222908898	soccer field	Train
15		sg28_4	258719795	town	Train

Chapter 3: Manual feature selection

3.2.3. Methodology

The way to select the optimal group of data channels for semantic segmentation consists of two parts: Data pre-processing and two-step verification. In the data pre-processing stage, which is shown in Figure 3-1, the first step is to convert the data of different channels in the point clouds into a panoramic (PAN) image separately. The next step is to slip the PAN image into subsets. The data in the panoramic form usually have a large resolution. For example, PAN image resolution for a normal scale single laser scan station with around thirty million laser points can be higher than 3000 x 7200. Such a large resolution requires a high graphic memory size for the hardware. The PAN form data needs to be split into pieces with smaller sizes according to the hardware performance. The PAN images are augmented by random cropping with 512 x 512 and random horizontal flipping.

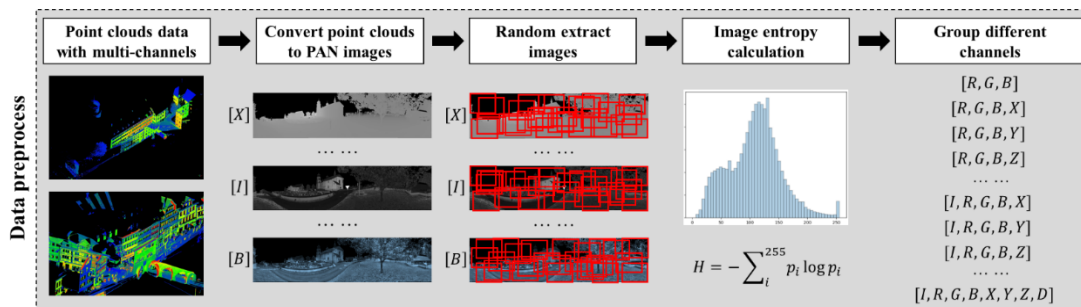


Figure 3-1: Detailed steps of the data preprocess.

Additionally, for the laser data in CIM, the “invalid” data often occurs. When the emitted laser beam points to the sky and does not return, there would be no valid coordinates and intensity, as a result, “zero” appears in the dataset. Therefore, to accelerate the convergence speed of neural network training, the proportion of such anomalous data in the comprehensive data is required to be adjusted.

Before grouping the data from different channels into different combinations, the image entropy (H) for each channel should be calculated (Gull and Skilling, 1984). Entropy is a statistical measure of randomness that can be used to characterize the texture or the contained information of the input image. The entropy of an image can be calculated by the first order from its histogram which provides the occurrence frequency (or probability) of all different grey levels in the image. The first-order image entropy is calculated as follows, where p_i is the probability of grey level i :

$$H = - \sum_i^{255} p_i \log p_i \quad (3-1)$$

Using the entropy, those possible channels that are richer in information can be roughly determined. Therefore, in the subsequent channel grouping in this chapter, some meaningless combinations are targeted and filtered out to reduce the time for choosing the optimal channel combination. In the channel grouping, the R, G, and B channels from the image are integrated with the I (intensity) channel acquired by the laser scanner to investigate the effect of intensity on semantic segmentation results. Alternatively, the R, G, and B channels from the image can be combined with the X, Y, and Z channels from the laser scanner, respectively, to compare the performance gained from the different channels. After that, the datasets for semantic segmentation are prepared, and all the images with appropriate sizes are stored according to the predefined combinations.

In selecting the optimal channel combination, a two-step verification strategy is applied to speed up identifying potential optimal combinations. First, networks with fewer parameters are applied to quickly estimate the potential optimal channel combinations. Then, networks with a deeper structure are adopted to verify the robustness of the optimal channel combinations. If the results show a high consistency

Chapter 3: Manual feature selection

across all the different networks, a reliable basis can be achieved for further subsequent substitutions or changes to the neural networks.

The encoder-to-decoder architecture for semantic segmentation is applied in this research. The encoder generates the feature maps for the input image, while the decoder uses the learned deconvolution layers to recover the image to the original size from the feature maps. The encoder-to-decoder structure can achieve better performance in reducing the information loss problem than those of the fully convolutional structure (Ronneberger et al., 2015). In addition, the structure of encoder-to-decoder is more flexible, as the encoder and decoder can be chosen from the commonly used neural network structures, respectively. For example, the encoder can be chosen from the ResNet (He et al., 2016), MobileNetV2 (Sandler et al., 2018), Xception (Chollet, 2017), Inception-ResNet-v2 (Szegedy et al., 2017) and HRCNet (Xu et al., 2020). The performance of different neural networks with varying complexity is evaluated in terms of overall accuracy (OA) and mean intersection over union (mIoU).

3.2.4. Experiment arrangement

It is necessary to ensure that the test data is similar to the data used for network training (Hand, 2008). Therefore, the selection of test data is based on the following reasons. First, it is noticed that point clouds 1-3, point clouds 4-6, and point clouds 8-9 are collected from three city scenes, respectively. Hence, a random point cloud from each scene is selected as the test data (i.e., point clouds 2, 5, and 8). Since the remaining 6 point clouds (i.e., point clouds 9-15) are collected from six different city scenes, to keep the test-to-train ratio similar to the previous selection (around 1/3), two point clouds (i.e., point clouds 9 and 12) are randomly selected as the test data. Therefore, a

total of five labelled point clouds were selected for testing, and the remaining ten were used to train the semantic segmentation networks, as shown in Table 3-1.

The pre-processing of the dataset follows the proposed method demonstrated in Section 3.2, where the size of the input images is taken as 512x512 to contain enough context information for semantic segmentation. Before deciding the combination of channels, it is necessary to check the image entropy first to avoid the combination with very little information.

As indicated in Figure 3-2, among the 15 scans provided by the Semantic 3D, the entropy values of RGB tend to be consistent. All of them remain in the top three, followed by intensity, but the performance is not stable for the other four channels (X, Y, Z, D). Therefore, the RGB channels from the image sensor dominate the subsequent channel combinations. Moreover, to verify the improvement of the data from the laser scanning on the semantic segmentation, the remaining channels are combined with RBG separately.

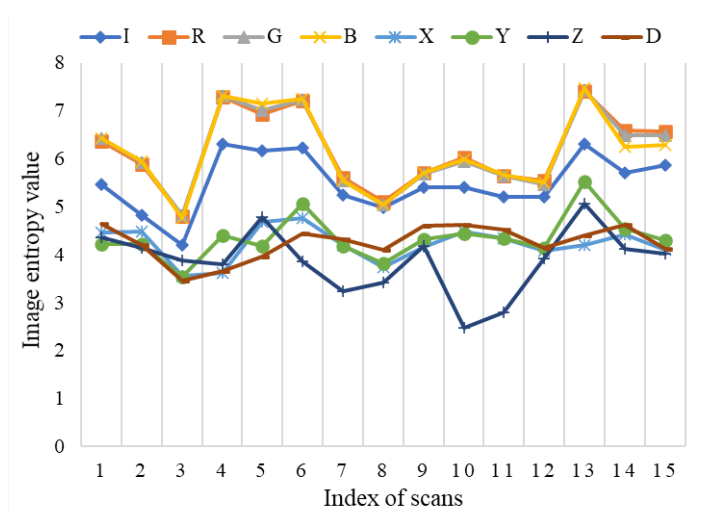


Figure 3-2: The entropy of different channel data.

Chapter 3: Manual feature selection

As shown in Table 3-2, a total of 13 combinations of channels are investigated in this research. These combinations are designed to investigate the effect of channels X, Y, Z, D, and intensity on the segmentation performance. Nine popular networks are used in this study, which includes two basic U-net with different depths, seven networks having the same decoder (i.e., DeepLab v3+) (Chen et al., 2018b), and different backbones (i.e., ResNet18, ResNet50, ResNet101, MobileNetV2, Xception, Inception-ResNet-v2, HRCNet). All the structures of networks are the same as the original implementation. Finally, the cross-entropy loss is used in this study.

Table 3-2: Combinations of channels.

Index	1	2	3	4	5	6	7
Combination	8 Channels	RGB	XYZD	IXYZD	IRGB	IRGBX	IRGBY
Index	8	9	10	11	12	13	-
Combination	IRGBZ	IRGBD	RGBX	RGBY	RGBZ	RGBD	-

The experiment is carried out on a PC with a processor of AMD Ryzen 9 3950X, RAM of 64 GB, and two GPUs of NVIDIA GeForce GTX 2080Ti. In addition, MATLAB 2020b is used for programming on the operating system of Windows 10. For a fair comparison through the whole experiment process, all the training used the same training protocol, which is a widely used strategy in deep learning research (Jingdong Wang et al., 2021; Zhao et al., 2018, 2017). More specifically, the SGD optimizer with a base learning rate of 0.05, a momentum of 0.9, and a weight decay of 0.001 was adopted in this study. The step learning rate policy was applied, which drops the learning rate by a factor of 0.1 every 10 epochs. For data augmentation, random image extraction and random horizontal flipping were applied (as described in the data process step). The total number of augmented images was 384 K, which were divided into 50 groups for training (50 epochs). Due to the limited physical memory on GPU cards, the “batchsize” was set as 16 (a total of 24 K iterations), and synchronized batch

normalization across GPU cards was adopted during training. Similar to (Jingdong Wang et al., 2021; Zhao et al., 2018, 2017), by applying random data augmentation and batch normalization, all the networks used in this study are considered to be resistant to overfitting.

3.3. Results

Figure 3-3 and Figure 3-4 demonstrate the mIoU and OA performance of the 13 combinations using nine networks. It is found that only the intensity channel brings a stable improvement of the segmentation performance. As shown in Table 3-3, the intensity channel improves mIoU and OA by an average of 3.24% and 2.01%, respectively. In contrast, it is found that the X, Y, and Z channels impair the segmentation performance. Table 3-3 shows that the X, Y, and Z channels reduce the mIoU by 2.84%, 2.97%, and 0.63%, respectively, and reduce the OA by 2.69%, 4.05%, and 3.46%, respectively. Finally, it is found that the effect of D channel depends on the criteria used for performance evaluation. More specifically, an additional channel of distance improves the mIoU by 3.09%, while reducing the OA by 2.0%. Since mIoU represents the average of the segmentation accuracy of each class, which indicates that the D channel is beneficial for the segmentation of imbalanced classes (classes with less data, i.e., difficult for segmentation).

Chapter 3: Manual feature selection

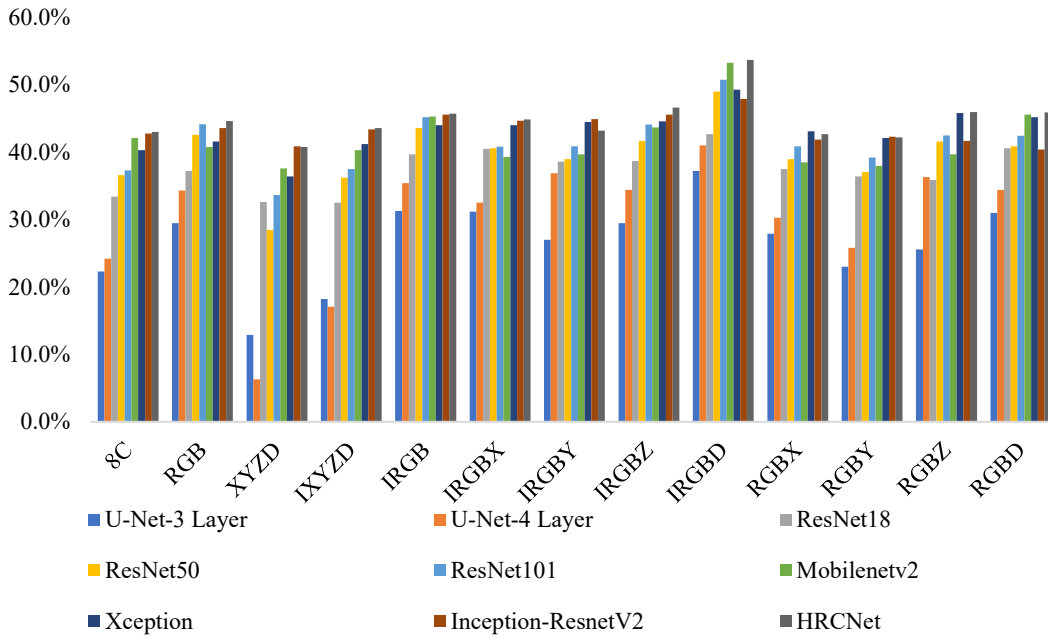


Figure 3-3: Mean intersection over union (mIoU) on test point clouds.

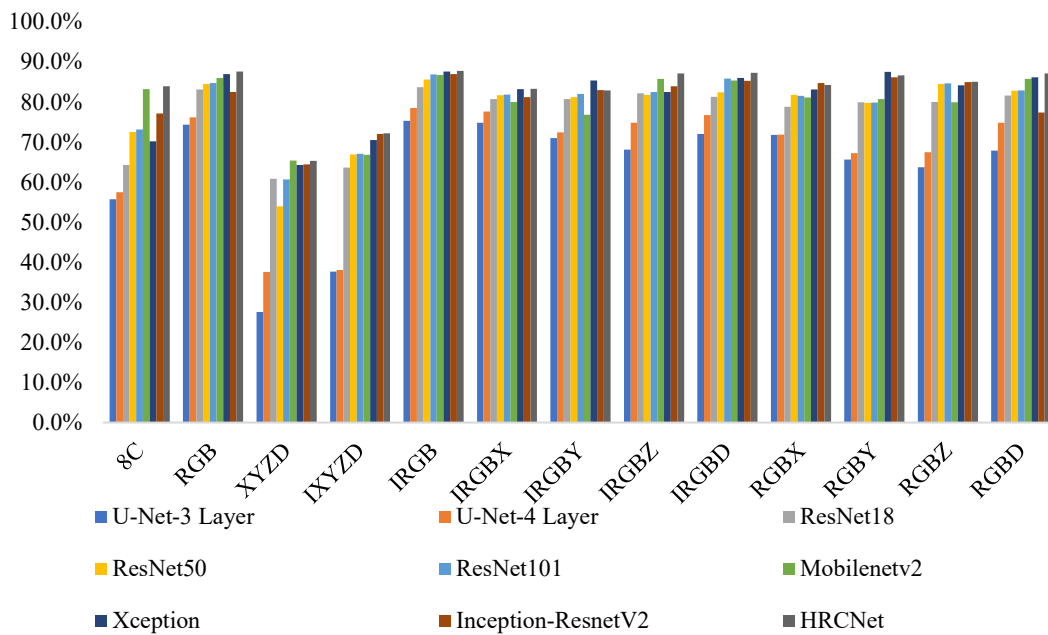


Figure 3-4: Overall accuracy (OA) on test point clouds.

Table 3-3: Average improvement by adding different channels.

Base channels	Additional channels	Improvement on mIoU	Improvement on OA
RGB	+ Intensity	1.95%	1.43%
XYZD	+ Intensity	4.47%	6.06%
RGBX	+ Intensity	1.87%	0.63%
RGBY	+ Intensity	3.17%	0.21%
RGBZ	+ Intensity	1.50%	1.54%
RGBD	+ Intensity	6.46%	1.74%
Average		3.24%	2.01%
RGB	+ X	-3.69%	-1.52%
IRGB	+ X	-1.90%	-3.85%
Average		-2.84%	-2.69%
RGB	+ Y	-3.62%	-3.60%
IRGB	+ Y	-2.31%	-4.43%
Average		-2.97%	-4.05%
RGB	+ Z	-0.47%	-3.47%
IRGB	+ Z	-0.72%	-3.36%
Average		-0.63%	-3.46%
RGB	+ D	0.77%	-2.00%
IRGB	+ D	5.47%	-1.88%
Average		3.08%	-2.00%

3.4. Discussion

Based on the aforementioned results, it is inferred that the combination of IRGBD channels provides the best mIoU performance, while the combination of IRGB channels provides the best OA performance. These inferences are confirmed in Table 3-4 and Table 3-5, where the highest value of mIoU and OA for each network is highlighted as green and yellow, respectively. It is observed that the optimal combination of channels is the same for all networks, which shows the robustness of the optimal combination.

Chapter 3: Manual feature selection

Table 3-4: The mIoU of seven networks regarding different combinations of channels (the highest mIoU for each network is marked as green).

	8C	RGB	XYZD	IXYZD	IRGB	IRGBX	IRGBY
U-Net-3 Layer	22.3%	29.5%	12.9%	18.2%	31.3%	31.2%	27.0%
U-Net-4 Layer	24.2%	34.3%	6.3%	17.1%	35.4%	32.5%	36.9%
ResNet18	33.4%	37.2%	32.6%	32.5%	39.7%	40.5%	38.6%
ResNet50	36.6%	42.6%	28.5%	36.2%	43.6%	40.6%	39.0%
ResNet101	37.3%	44.1%	33.7%	37.5%	45.2%	40.8%	40.9%
Mobilenetv2	42.1%	40.8%	37.6%	40.3%	45.3%	39.3%	39.7%
Xception	40.3%	41.6%	36.4%	41.2%	44.0%	44.0%	44.5%
Inception-ResnetV2	42.8%	43.6%	40.9%	43.4%	45.6%	44.7%	44.9%
HRCNet	43.0%	44.6%	40.8%	43.6%	45.7%	44.8%	43.2%
	IRGBZ	IRGBD	RGBX	RGBY	RGBZ	RGBD	-
U-Net-3 Layer	29.5%	37.2%	27.9%	23.0%	25.6%	31.0%	-
U-Net-4 Layer	34.4%	41.0%	30.3%	25.8%	36.3%	34.4%	-
ResNet18	38.7%	42.7%	37.5%	36.4%	35.9%	40.6%	-
ResNet50	41.7%	49.0%	39.0%	37.1%	41.6%	40.9%	-
ResNet101	44.1%	50.8%	40.9%	39.2%	42.5%	42.4%	-
Mobilenetv2	43.7%	53.3%	38.5%	38.0%	39.7%	45.6%	-
Xception	44.6%	49.3%	43.1%	42.1%	45.8%	45.2%	-
Inception-ResnetV2	45.6%	47.9%	41.9%	42.3%	41.7%	40.4%	-
HRCNet	46.6%	53.7%	42.7%	42.2%	46.0%	45.9%	-

Table 3-5: The OA of seven networks regarding different combinations of channels (the highest OA for each network is marked as yellow).

	8C	RGB	XYZD	IXYZD	IRGB	IRGBX	IRGBY
U-Net-3 Layer	55.7%	74.3%	27.6%	37.7%	75.3%	74.8%	71.0%
U-Net-4 Layer	57.5%	76.2%	37.6%	38.1%	78.5%	77.6%	72.4%
ResNet18	64.3%	83.1%	60.8%	63.6%	83.7%	80.7%	80.7%
ResNet50	72.5%	84.5%	54.0%	66.9%	85.6%	81.7%	81.2%
ResNet101	73.2%	84.7%	60.7%	67.1%	86.9%	81.9%	82.0%
Mobilenetv2	83.2%	86.0%	65.4%	66.8%	86.7%	80.0%	76.8%
Xception	70.2%	87.0%	64.3%	70.5%	87.6%	83.2%	85.4%
Inception-ResnetV2	77.1%	82.5%	64.4%	72.0%	87.0%	81.2%	83.0%
HRCNet	83.9%	87.6%	65.3%	72.2%	87.8%	83.3%	82.9%
	IRGBZ	IRGBD	RGBX	RGBY	RGBZ	RGBD	-
U-Net-3 Layer	68.1%	72.0%	71.8%	65.6%	63.7%	67.9%	-
U-Net-4 Layer	74.8%	76.7%	71.9%	67.2%	67.5%	74.8%	-
ResNet18	82.2%	81.3%	78.8%	79.9%	80.0%	81.6%	-
ResNet50	81.8%	82.4%	81.8%	79.8%	84.5%	82.8%	-
ResNet101	82.5%	85.8%	81.5%	79.9%	84.6%	82.9%	-
Mobilenetv2	85.8%	85.4%	81.1%	80.7%	79.9%	85.8%	-
Xception	82.5%	86.0%	83.1%	87.5%	84.2%	86.2%	-
Inception-ResnetV2	83.9%	85.3%	84.7%	86.2%	85.0%	77.4%	-
HRCNet	87.1%	87.3%	84.3%	86.6%	85.1%	87.1%	-

In the meantime, by ranking the mIoU and OA of all the 13 channel combinations for seven networks, as shown in Table 3-6 and Table 3-7, it is found that the worst channel combination also presents a high consistency across the seven networks, but the consistency decreases for other combinations ranked in the middle. This indicates that the channel combinations with respect to extreme cases are more consistent than others.

Table 3-6: Ranking of the mIoU performance of 13 channel combinations for seven networks.

	1	2	3	4	5	6	7
U-Net-3 Layer	IRGBD	IRGB	IRGBX	RGBD	RGB	IRGBZ	RGBX
U-Net-4 Layer	IRGBD	IRGBY	RGBZ	IRGB	IRGBZ	RGBD	RGB
ResNet18	IRGBD	RGBD	IRGBX	IRGB	IRGBZ	IRGBY	RGBX
ResNet50	IRGBD	IRGB	RGB	IRGBZ	RGBZ	RGBD	IRGBX
ResNet101	IRGBD	IRGB	RGB	IRGBZ	RGBZ	RGBD	IRGBY
Mobilenetv2	IRGBD	RGBD	IRGB	IRGBZ	8C	RGB	IXYZD
Xception	IRGBD	RGBZ	RGBD	IRGBZ	IRGBY	IRGB	IRGBX
Inception-ResnetV2	IRGBD	IRGB	IRGBZ	IRGBY	IRGBX	RGB	IXYZD
HRCNet	IRGBD	IRGBZ	RGBZ	RGBD	IRGB	IRGBX	RGB
	8	9	10	11	12	13	-
U-Net-3 Layer	IRGBY	RGBZ	RGBY	8C	IXYZD	XYZD	-
U-Net-4 Layer	IRGBX	RGBX	RGBY	8C	IXYZD	XYZD	-
ResNet18	RGB	RGBY	RGBZ	8C	XYZD	IXYZD	-
ResNet50	IRGBY	RGBX	RGBY	8C	IXYZD	XYZD	-
ResNet101	RGBX	IRGBX	RGBY	IXYZD	8C	XYZD	-
Mobilenetv2	IRGBY	RGBZ	IRGBX	RGBX	RGBY	XYZD	-
Xception	RGBX	RGBY	RGB	IXYZD	8C	XYZD	-
Inception-ResnetV2	8C	RGBY	RGBX	RGBZ	XYZD	RGBD	-
HRCNet	IXYZD	IRGBY	8C	RGBX	RGBY	XYZD	-

Table 3-7: Ranking of the OA performance of 13 channel combinations for seven networks.

	1	2	3	4	5	6	7
U-Net-3 Layer	IRGB	IRGBX	RGB	IRGBD	RGBX	IRGBY	IRGBZ
U-Net-4 Layer	IRGB	IRGBX	IRGBD	RGB	IRGBZ	RGBD	IRGBY
ResNet18	IRGB	RGB	IRGBZ	RGBD	IRGBD	IRGBX	IRGBY
ResNet50	IRGB	RGB	RGBZ	RGBD	IRGBD	IRGBZ	RGBX
ResNet101	IRGB	IRGBD	RGB	RGBZ	RGBD	IRGBZ	IRGBY
Mobilenetv2	IRGB	RGB	IRGBZ	RGBD	IRGBD	8C	RGBX
Xception	IRGB	IRGBY	RGB	RGBD	IRGBD	IRGBY	IRGBZ
Inception-ResnetV2	IRGB	IRGBY	IRGBD	IRGBZ	IRGBX	IRGBZ	IRGBY
HRCNet	IRGB	RGB	IRGBD	RGBD	IRGBZ	IRGBY	IRGBZ
	8	9	10	11	12	13	-
U-Net-3 Layer	RGBD	IRGBY	IRGBZ	8C	IXYZD	XYZD	-
U-Net-4 Layer	IRGBX	IRGBZ	IRGBY	8C	IXYZD	XYZD	-
ResNet18	IRGBZ	IRGBY	IRGBX	8C	IXYZD	XYZD	-
ResNet50	IRGBX	IRGBY	IRGBY	8C	IXYZD	XYZD	-
ResNet101	IRGBX	IRGBX	IRGBY	8C	IXYZD	XYZD	-
Mobilenetv2	IRGBY	IRGBX	IRGBZ	IRGBY	IXYZD	XYZD	-
Xception	IRGBX	IRGBX	IRGBZ	IXYZD	8C	XYZD	-
Inception-ResnetV2	RGB	IRGBX	IRGBD	8C	IXYZD	XYZD	-
HRCNet	IRGBX	8C	IRGBX	IRGBY	IXYZD	XYZD	-

Moreover, it is noticed that the simple mixture of all the available channels (i.e., column 8C in Table 3-4 and Table 3-5) always results in a worse performance compared to that of combinations with fewer channels. To explore this thoroughly, for

Chapter 3: Manual feature selection

channel combinations 8C, RGB, IRGB, and IRGBD, the training curves for networks with the Inception-ResnetV2 backbone are plotted in Figure 3-5, and two test images are used to obtain the feature maps and corresponding segmentation results for comparison, as demonstrated in Figure 3-6 and Figure 3-7.

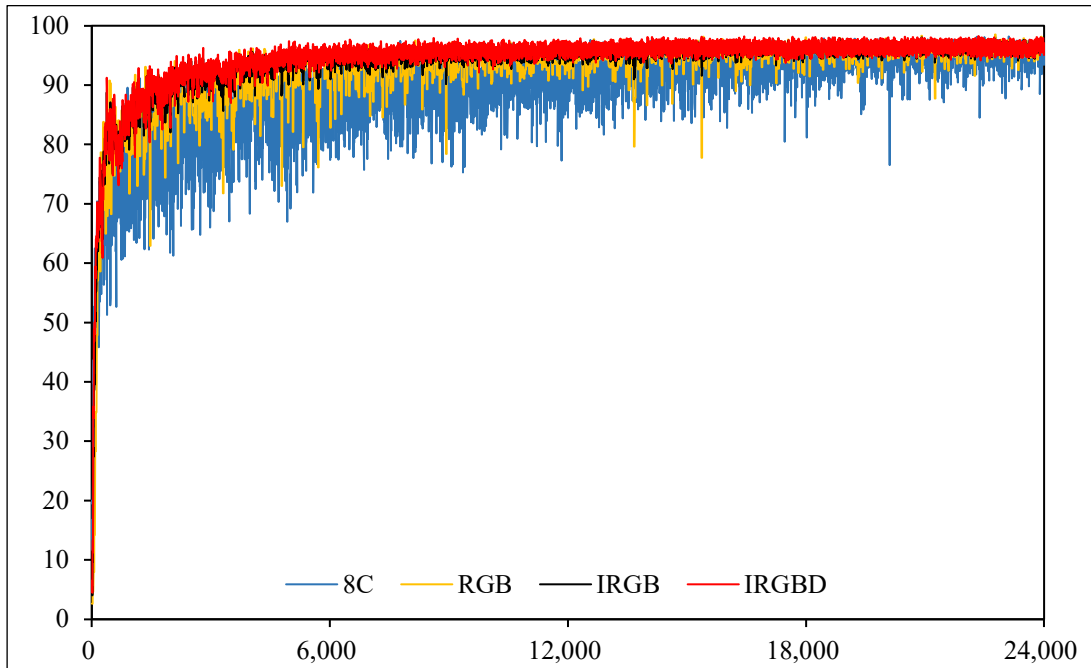


Figure 3-5: Training accuracy for combinations of 8C (all the channels), RGB (color), IRGB (intensity and color), and IRGBD (intensity, color and depth) using networks of Inception-ResnetV2 backbone.

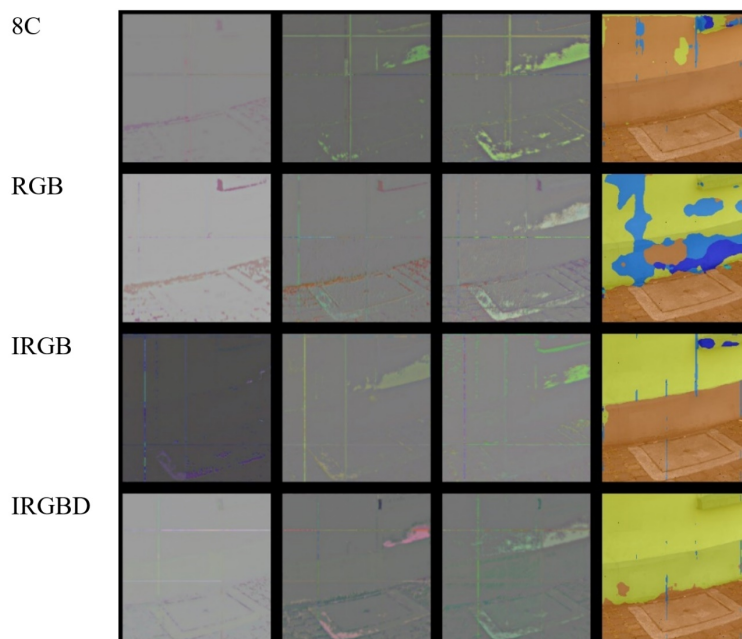


Figure 3-6: Feature maps and segmentation results for four combinations for the building-road joint image.

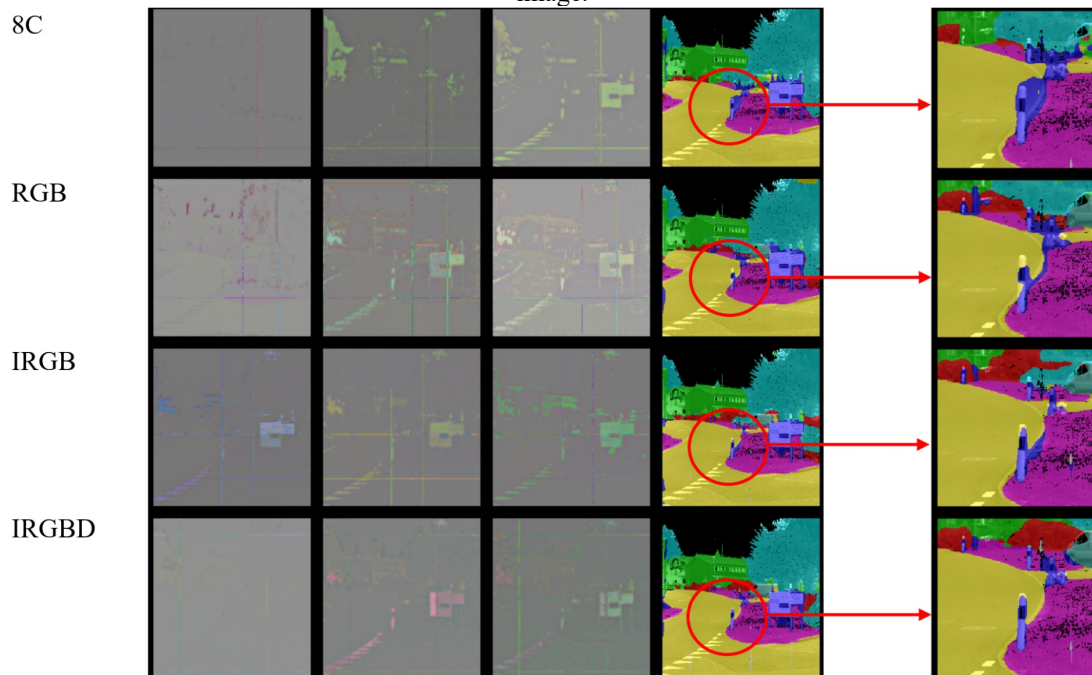


Figure 3-7: Feature maps and segmentation results for four combinations in the street view image.

From Figure 3-5, it is observed that the training process of combination of 8C converges much slower than others, which might indicate that the network struggled to learn the "correct" feature when there is a mixture of "useful" and "useless" data input. Taking the segmentation results in Figure 3-6 as an example, compared to the result of RGB combination, the additional I channel (i.e., IRGB) does help remove the mislabelled pixels in the wall region, but it also causes the mislabelling of the whole bottom part of the wall. The segmentation result is even worse for the 8C combination, which completely fails to distinguish the building and the road. A similar situation occurs for the street view test, as shown in Figure 3-7. Compared to the segmentation results for the RGB combination, the 8C combination causes a large mislabelling area around the road sign. Both test image results show that the IRGBD combination yields the best segmentation results.

Chapter 3: Manual feature selection

The average time of single training for nine networks is summarized in Figure 3-8, where the average time of Xception (17.2 h) is two times more than that of ResNet18 (7.5 h). Moreover, since the channel analysis requires a series of comparative tests to ascertain the optimal channel combination, the differences in training time between the networks are magnified. For example, the total channel analysis time for ResNet18 and Xception are 97.4 and 193.6 h, respectively. Since the previous investigation shows a high consistency of optimal channel combination across different networks, the efficiency can be improved significantly by conducting the channel analysis on a small network before training on more sophisticated networks. In addition, the total inference time (including PAN image generation, inference, back projection) is around 170 k points/s.

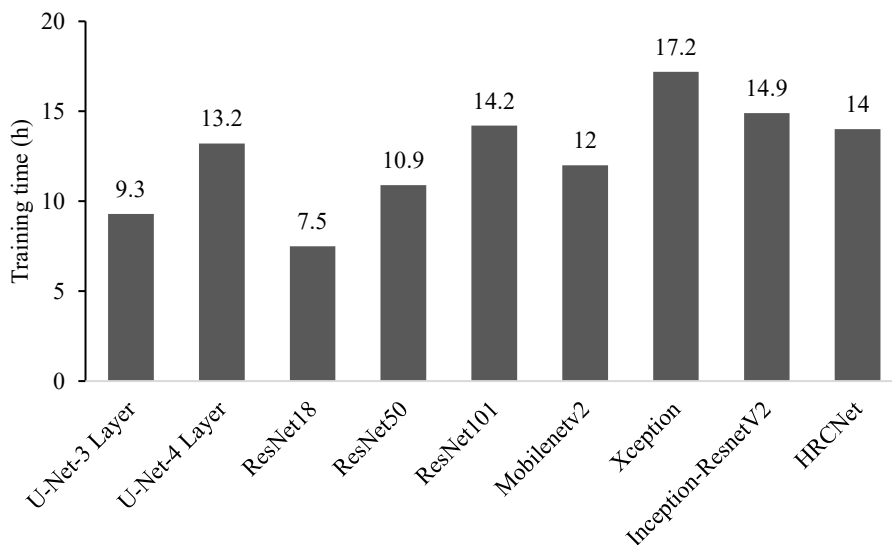


Figure 3-8: Summary of the average time of single training for nine network structures.

Finally, since the IRGBD channel combination and HRCNet got the best performance (mIoU is more critical than OA) in the previous testing, they were selected to evaluate the performance on the Semantic3D (reduced-8) test dataset. The reason to choose the reduced-8 rather than the high density test dataset is that previous methods (especially

point-based methods) are often tested on the reduced-8 test dataset as they cannot handle high density point clouds efficiently. The complete training dataset (15 point clouds) was used in this stage, and the training protocol remains the same as mentioned in Section 3.2.4. The quantitative segmentation results are summarized in Table 3-8 below, where XJTLU outperforms previous best image-based methods by 4.4% regarding mIoU, and even outperforms several recently published point/discretization-based methods, which show the effectiveness of the proposed methods.

Table 3-8: Quantitative results of different approaches on Semantic3D (reduced-8). Accessed on 16 March 2021 (the overperformed methods are marked in grey).

		mIoU (%)	OA (%)	man-made	natural terrain	high veg	low veg	buildings	hard-scape	scanning art	cars
Point/ discretization -based methods	SEGCloud (Tchapmi et al., 2017)	61.3	88.1	83.9	66.0	86.0	40.5	91.1	30.9	27.5	64.3
	RF MSSF (Thomas et al., 2018)	62.7	90.3	87.6	80.3	81.8	36.4	92.2	24.1	42.6	56.6
	Edge-Con (Contreras and Denzler, 2019)	59.5	87.9	84.5	70.9	76.6	26.1	91.4	18.6	56.5	51.4
	ShellNet (Zhang et al., 2019)	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
	OctreeNet (F. Wang et al., 2020)	59.1	89.9	90.7	82.0	82.4	39.3	90.0	10.9	31.2	46.0
	GACNet (L. Wang et al., 2019)	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
	RandLA-Net (Q. Hu et al., 2020)	77.4	94.8	95.6	91.4	86.6	51.5	95.7	51.5	69.8	76.8
Projection -based methods	DeePr3SS (Lawin et al., 2017)	58.5	88.9	85.6	83.2	74.2	32.4	89.7	18.5	25.1	59.2
	SnapNet (Boulch et al., 2018)	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4
	XJTLU (This study)	63.5	89.4	85.4	74.4	74.6	31.9	93.0	25.2	41.5	82.0

3.5. Summary

With the development of CIM, there is an increasing demand for high precision semantic segmentation information. Data fusion is an emerging method to improve the segmentation performance. However, without a selection of effective data fusion sources, extra effort is required in both data collection and processing. Therefore, an efficient data fusion approach is proposed in this chapter by exploring the optimal combination of data channels. The analysis on the performance of different

Chapter 3: Manual feature selection

combinations of data channels is applied to obtain the optimal combination by adopting various neural networks. The robustness of the optimal combination is proved using a case study, which demonstrates the feasibility of the proposed data fusion channel selection. The findings can be utilized to achieve a significant improvement on efficiency by adopting a simple structured network for the channel analysis before applying a more complex network. In addition, the case study demonstrates that, without adopting this framework, a simple mixture of available data sources impairs the segmentation performance, which shows the necessity of channel selection in data fusion. Finally, using the selected channel combination and network, this study achieved the best performance among image-based methods and outperformed several recent point/discretization-based methods.

Although the feasibility of the proposed method has been investigated on 2D convolutional neural networks, other types of networks exist that could be used for semantic segmentation in CIM, such as vision transformer (Wu et al., 2020) and point-based network (Q. Hu et al., 2020). Therefore, future work will focus on the investigation of the robustness of the optimal combination of data sources among different types of networks.

Chapter 4: Locally enhanced image-based geometric features

This chapter is based on the published paper: Cai, Y., Fan, L., Atkinson, P.M., Zhang, C., 2022a. Semantic Segmentation of Terrestrial Laser Scanning Point Clouds Using Locally Enhanced Image-Based Geometric Representations. IEEE Trans. Geosci. Remote Sens. 60, 1–15. <https://doi.org/10.1109/TGRS.2022.3161982>

Note: The research presented in this chapter improves the segmentation accuracy of image-based methods by developing novel image-based geometric features.

4.1. Introduction

The rapid development of three-dimensional (3D) data acquisition technologies has led to various types of sensors, such as terrestrial laser scanning (TLS) devices, RGB-D cameras and LiDAR (R. Zhang et al., 2018). Among these instruments, TLS stands out for its ability to quickly acquire large-volume (hundreds of millions of points per scan) and high-precision (millimetre level) point cloud data and is, therefore, used widely in applications where high-quality point cloud data are required. These may include, but are not limited to, 3D building reconstruction (Cai et al., 2021a; Cai and Fan, 2021; Cao et al., 2021; Fan and Cai, 2021; Huang et al., 2021), vegetation and forest assessments (Fan et al., 2014; Liu et al., 2019; Safaie et al., 2021; Zheng et al., 2021), and cultural heritage management (Guo et al., 2021; Montuori et al., 2014).

In addition to the high-precision geometric information provided by TLS point clouds, semantic segmentation is often required as the basis for more complex purposes in the aforementioned applications. The goal of semantic segmentation of point clouds is mainly to annotate each data point with a semantic label, which is often based on the geometry, the reflection intensity and sometimes the colour information provided by the data point itself and its neighbours. This can be achieved via traditional supervised classification methods (Guo et al., 2021; Vosselman et al., 2017; Weinmann et al., 2015) or deep learning approaches (Cai et al., 2021b; Charles et al., 2017; Jaritz et al., 2019; Landrieu and Simonovsky, 2018; Qi et al., 2017). Compared to traditional classification methods using handcrafted features (e.g., support vector machines, random forests and conditional random fields), deep learning methods are becoming increasingly popular because they can automatically learn the feature representations needed for segmentation from raw data, avoid complex feature design, and typically

Chapter 4: Locally enhanced image-based geometric features

result in higher segmentation accuracy (Q. Hu et al., 2020; Pan et al., 2019; R. Zhang et al., 2018).

Existing point cloud segmentation methods can be categorized into three major groups based on the form of the input data: point-based, voxel-based and image-based methods. The pioneering work on point-based methods is PointNet (Charles et al., 2017), which used shared Multi-Layer Perceptrons (MLPs) to learn pre-point features and used symmetrical pooling functions to learn global features. On the basis of PointNet, many other point-based networks have been proposed in recent years, which can be subdivided into pointwise MLP methods, graph-based methods, point convolution methods, and RNN-based methods (Guo et al., 2021). This class of algorithm can typically achieve high accuracy, and the state-of-the-art (SOTA) method is the RFCR (Gong et al., 2021) in this category, which achieved an Overall Accuracy (OA) of 94.3% and a mean Intersection over Union (mIoU) of 77.8% on the Semantic3D (reduced-8) (Gong et al., 2021; Hackel et al., 2017). However, while point-based methods are focused on increasing the segmentation accuracy of point clouds, their high computational cost makes them too costly for practical application to large-scale TLS point clouds. For example, for a use case where the processing time was revealed (Hackel et al., 2017), it ranges from 10 to 50 minutes to process 4-point clouds containing 80 million points in Semantic3D (reduced-8).

For the second class of voxel-based methods (Choy et al., 2019; Graham et al., 2018; Meng et al., 2019; Silberman et al., 2012; F. Wang et al., 2020), they first convert the point cloud into a dense/sparse discrete voxel representation and then apply the 3D convolutional neural network (CNN). Since 3D convolutional networks are extremely

Chapter 4: Locally enhanced image-based geometric features

computationally intensive and consume significant amounts of Graphics Processing Unit (GPU) memory, such methods have to make careful trade-offs in terms of segmentation accuracy and processing time. From the published performance of these methods on various benchmark datasets (Armeni et al., 2016; Geiger et al., 2013, 2012; Hackel et al., 2017; Silberman et al., 2012), such methods are not only less accurate than the first type of method, but also very slow in processing and, therefore, are considered unsuitable for processing large-scale TLS point cloud data.

The image-based methods utilize 2D convolutional neural networks (CNNs) to segment multi-channel images generated from point cloud data. There are two approaches for image generation. The first approach (Boulch et al., 2018; Lawin et al., 2017; Tatarchenko et al., 2018) projects point cloud data from multiple virtual camera views onto a plane, while the second approach (Cai et al., 2021b; Milioto et al., 2019; B. Wu et al., 2018; Wu et al., 2019) projects the point cloud data as a panoramic image centred at the scanner. The second approach is more efficient than the multi-view ones because processing is limited to only one panoramic image for each point cloud obtained (Boulch et al., 2018; Cai et al., 2021b). Coupled with the use of 2D CNNs (much more efficient than those networks used in point-based and voxel-based methods), the panoramic images offer an extremely fast approach to segmenting point cloud data. For example, the SOTA image-based method (Cai et al., 2021b) takes only 5.13s to process the Semantic3D (reduced-8) (Hackel et al., 2017) test dataset. However, it was noticed that its segmentation accuracy (Cai et al., 2021b) was relatively low compared to the SOTA point-based method RFCR (Gong et al., 2021), achieving an OA of only 89.4% and a mIoU of 63.5% on Semantic3D (reduced-8). Therefore, image-based methods are ideal for processing large-scale TLS point cloud

data, but such methods available in the literature suffer from the problems elaborated in the next paragraph, which also form the likely basis for any further improvements in their segmentation accuracy.

Three types of information of TLS point clouds can be considered for semantic segmentation (i.e., geometric information (coordinates and their derivatives), intensity and RGB if images were taken). In the existing image-based methods, it was noticed that combinations of feature channels considered (Boulch et al., 2018; Cai et al., 2021b; Lawin et al., 2017; Tatarchenko et al., 2018) always included the RGB information, without which the segmentation accuracy degraded significantly. This is not surprising as the true colours include rich information about the objects to be segmented. However, this means that those methods are highly reliant on the RGB information and cannot effectively handle the cases where the RGB information is missing (no images taken) or is mismatched to point clouds due to moving objects in the scene or the imperfect matching between images and point clouds taken separately. In addition, the geometric information was either not considered or not used in an effective way. In contrast, point-based and voxel-based methods perform well for point clouds with only coordinate information (Q. Hu et al., 2020; Landrieu and Simonovsky, 2018; Meng et al., 2019; Thomas et al., 2019), indicating that geometric features are valuable for point cloud semantic segmentation. Hence, it is reasonable to speculate that the application scope and segmentation accuracy of image-based approaches can be improved further if the geometric information contained in the point cloud is utilized effectively.

Therefore, under the umbrella of image-based methods, this study aims to improve and

Chapter 4: Locally enhanced image-based geometric features

generalize this class of methods by considering the characterization of the geometric information of scenes/objects in the panoramic images derived from coordinates of point cloud data. The increase in accuracy relates to the semantic segmentation while the generalization refers to cases where the RGB information is missing in the point cloud data. To this end, an image enhancement method is proposed to characterize the local geometric features in the images. Based on the enhanced images, this research proposes a new combination of feature channels without the RGB information. In the CNN used for extracting the semantic information in this study, the Atrous Spatial Pyramid Pooling (ASPP) module (Chen et al., 2018a) is considered to aggregate multi-scale high-level features from HRNet (Jingdong Wang et al., 2021). In past studies (Chen et al., 2018a, 2018b, 2017), the aggregation was typically executed using coarse-resolution feature maps. However, in this study, the finest-resolution feature maps in HRNet are used for the aggregation, the outputs of which are concatenated with multiple low-level features for segmentation.

The main contributions of this research are the establishment of a new image enhancement method for characterizing effectively the local geometric features in the panoramic images derived from point clouds, and the finding that the utilization of those local geometric features can increase the segmentation accuracy of image-based methods. The approach proposed in this study offers a better alternative channel combination to replace those involving the RGB channels, which is very useful for cases where the RGB information is absent or inaccurate

4.2. Methodology

The methodology considered in this research involves the following key steps. Firstly, the information (e.g., intensity and XYZ coordinates) contained in the unstructured point cloud data was projected into a multichannel panoramic image using the transformation relationship between the Cartesian coordinate system and the spherical coordinate system. Secondly, the local-based enhancement was applied to the panoramic image channels that contain geometric information such as XYZ coordinates and range. Lastly, semantic information was extracted from the panoramic image using a pre-trained customized CNN, and back-projected to the raw point cloud data to obtain semantically segmented point cloud. More detailed descriptions of these steps are provided in Sections 4.2.3-4.2.6.

4.2.1. Study data

The large-scale Semantic3D dataset (Hackel et al., 2017) was used to demonstrate and evaluate the proposed method, which contains a total of 30 labelled TLS point clouds collected at 10 different scenes. Point cloud data were labelled into eight classes, namely: made terrain, natural terrain, high vegetation, low vegetation, buildings, hard scape, scanning artefacts and cars. The ground reference labels for 15 training point clouds are available from the dataset supplier. The online evaluation frequency of test set results is limited to once every three days. Therefore, except for Section 4.3.4 where the test set was used for comparison with the state-of-the-art results, all other experiments were conducted on the training set. More specifically, for Sections 4.3.2-4.3.3, the performance of the proposed method was evaluated by employing 5-fold cross-validation on the Semantic3D training dataset.

Chapter 4: Locally enhanced image-based geometric features

4.2.2. Segmentation accuracy metrics

To evaluate the segmentation performance, the same evaluation metrics as used in the Semantic3D online evaluation were used in this study, i.e., OA and mIoU. The OA metric is the ratio of correctly classified points (regardless of class) to the total number of points. The mIoU metric is the mean IoU of all classes. For class i , the IoU metric is the ratio of correctly classified pixels to the total number of ground reference data and predicted pixels in that class. The formulae for the aforementioned metrics are shown in Equations 4-1, 4-2 and 4-3.

$$OA = \frac{TP}{\text{Total number of points}} \quad (4-1)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (4-2)$$

$$mIoU = \frac{\sum_{i=1}^N IoU_i}{N} \quad (4-3)$$

where TP, FN, FP, i , N represent the true positive, false negative, false positive points classified, index of class and total number of classes, respectively.

In general, OA provides a quick and computationally inexpensive estimate of the percentage of correctly classified points, while mIoU provides a measurement of accuracy that not only penalizes false positives, but also increases the penalty against segmentation errors in small classes. Since the numbers of points contained in the eight classes of the Semantic3D benchmark dataset are highly imbalanced (shown in Figure 4-1), mIoU is considered more critical in this research.

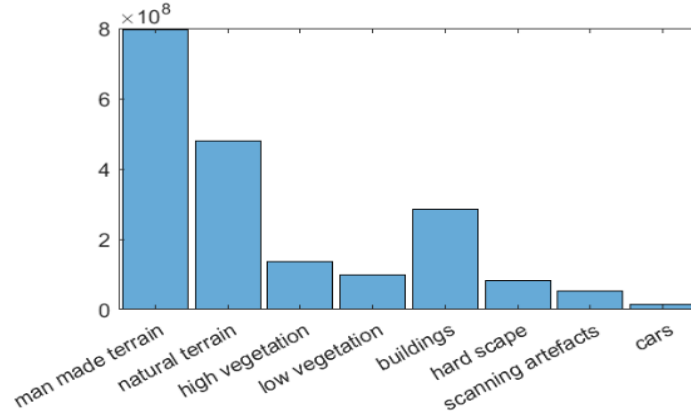


Figure 4-1: The distribution of the classes of points in Semantic3D dataset.

4.2.3. Point cloud to image projection

Many terrestrial laser scanners collect point cloud data through vertically rotating optics that are mounted on a horizontally rotating base. Since their rotational steps are usually fixed throughout a single scan, the point cloud data obtained would theoretically have fixed inclination and azimuthal resolutions. These two resolutions are typically the same. In other words, if the point cloud data are considered as vectors originating from the origin (i.e., the scanner's optical centre), these vectors will be uniformly distributed in a spherical space centred at the origin. Therefore, TLS point clouds are inherently suitable to be projected into spherical coordinate systems. Based on this, the following method for point cloud to image projection was used in this study, which is demonstrated using the example shown in Figure 4-2.a. Firstly, the Cartesian coordinates of the point cloud data were transformed into spherical coordinates using Equations 4-4, 4-5 and 4-6.

$$range (r) = \sqrt{x^2 + y^2 + z^2} \quad (4-4)$$

$$inclination (\theta) = arccos \frac{z}{r} \quad (4-5)$$

$$azimuth (\varphi) = arctan \frac{y}{x} \quad (4-6)$$

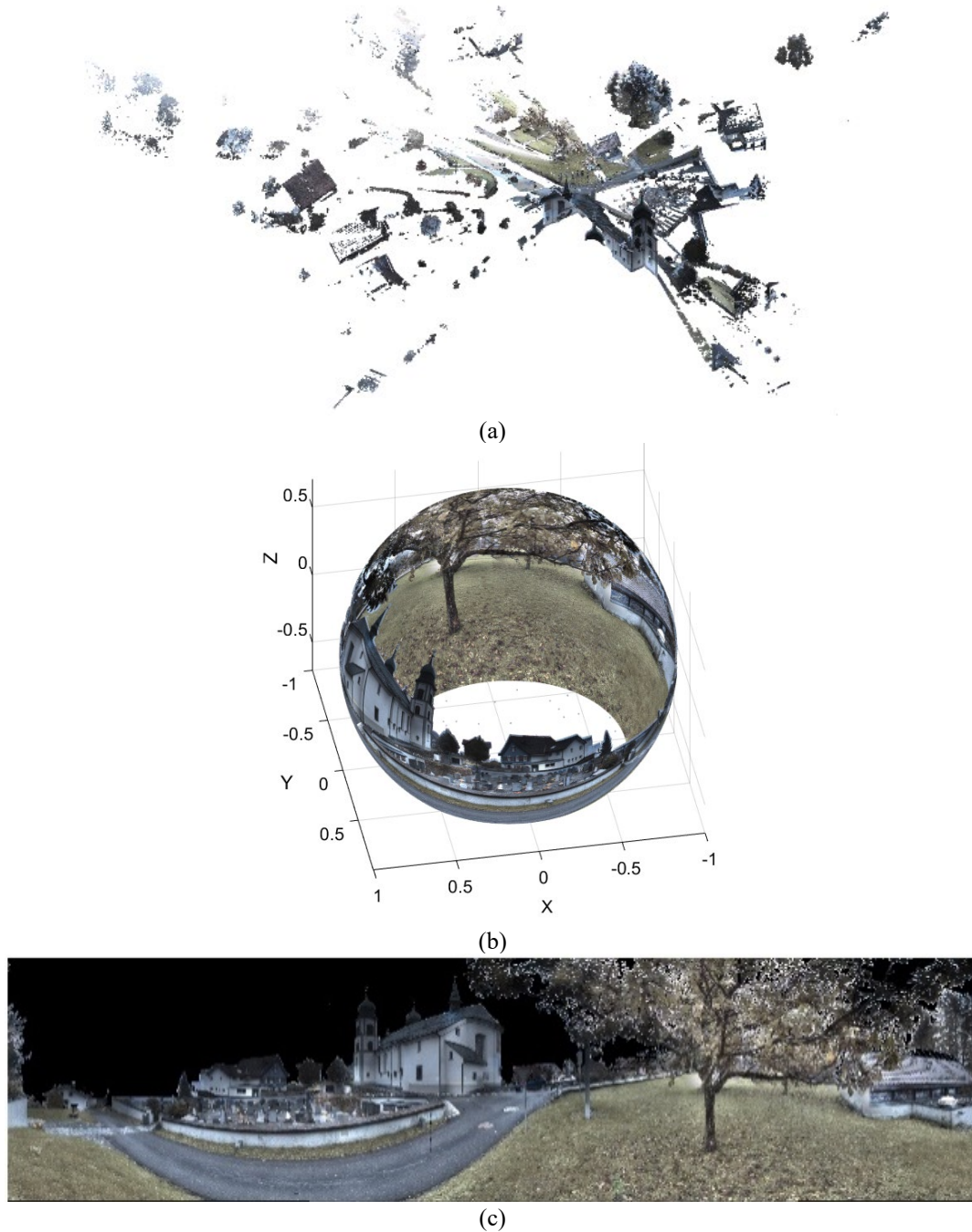


Figure 4-2: Key stages in the projection process: (a). The raw input point cloud, (b). All points scaled to a spherical surface at a distance of 1 from the origin (i.e., the center of the scanner), (c). The panoramic image rasterized from the spherical surface.

Secondly, the position of each data point in the unit spherical surface (i.e., "continuous" spherical image) is determined by its inclination θ and azimuth φ , as shown in Figure 4-2.b. Thirdly, by using a specific angular resolution ω to discretize the "continuous" spherical image, a rasterized spherical image is obtained. To ensure the image continuity, the image angular resolution should be slightly larger than the

Chapter 4: Locally enhanced image-based geometric features

scanner angular resolution. Finally, by mapping the available information (e.g., RGB, intensity, range) to the rasterized spherical image and splitting it from a certain azimuth (e.g., 180° used in the subsequent experiments), the multichannel panoramic image is obtained (e.g., the RGB panoramic image in Figure 4-2.c). More specifically, for a data point of the inclination θ and the azimuth φ in the spherical coordinate system, its pixel location in the panoramic image is determined using Equation 4-7.

$$\left(\left\lceil \frac{90 - \theta}{\omega} \right\rceil, \left\lceil \frac{180 - \varphi}{\omega} \right\rceil \right) \quad (4-7)$$

where the former element represents the row location for the inclination θ , the latter element represents the column location for the azimuth φ , ω is the angular resolution, $\lceil x \rceil$ rounds x to the nearest integer greater than or equal to x .

Because of the fine angular resolutions of laser scanners, the resolution of the projected panoramic image could be ultra-high. For example, the equivalent panoramic image size of the point cloud captured using the RTC360's finest resolution is 8333×20334 pixels.

During the point cloud to image projection, it is often the case that a single image pixel contains multiple data points. In this case, the pixel values in the panoramic feature image (e.g., RGB image) were taken as the average values of multiple data points, while the pixel values (labelled classes) in the labelled panoramic image (labelled image used for training) were taken as the ones corresponding to the rarest class to increase network segmentation accuracy regarding the imbalanced class (typically, the class with fewer data is harder to segment).

Chapter 4: Locally enhanced image-based geometric features

4.2.4. Enhancement of image-based geometric features

As shown in Figure 4-3.a, the panoramic RGB image is relatively clear. However, objects in the grayscale images obtained by projecting the XYZ coordinates and the range information were not shown clearly, such as the panoramic image of the Z coordinate shown in Figure 4-3.b. Due to this phenomenon, existing image-based methods (Boulch et al., 2018; Cai et al., 2021b; Lawin et al., 2017) rely mainly on the RGB information, and this type of grayscale images was usually used as auxiliary information only.

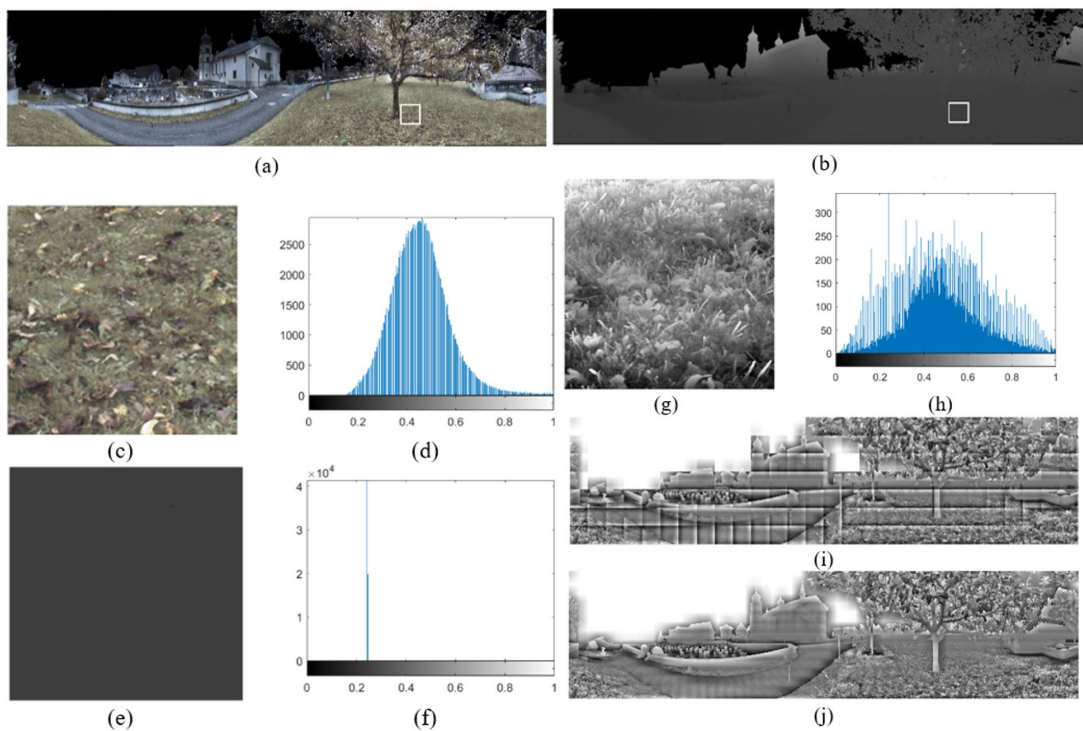


Figure 4-3: Illustrations of image enhancement effects: (a). The panoramic image projected from RGB channels, (b). The panoramic image projected from Z coordinate, (c). A local RGB image extracted from the box in (a), (d). The distribution histogram of the pixel values in (c), (e). The local Z coordinate image extracted from the box in (b), (f). The distribution histogram of the pixel values in (e), (g). The enhanced local Z coordinate image. (h) The distribution histogram of the pixel values in (g), (i). The enhanced Z coordinate image without overlapping. (j). The enhanced Z coordinate image with overlapping.

By comparing the pixel value distribution histograms (Figure 4-3.d and Figure 4-3.f) of the RGB image (Figure 4-3.c) and Z -coordinate image (Figure 4-3.e) for the same local area (area within the 256×256 white box in Figure 4-3.a and Figure 4-3.b), it was

Chapter 4: Locally enhanced image-based geometric features

found that the distribution of grayscale values of the Z coordinate image was extremely concentrated compared to the RGB image. This is due to the fact that the range of variation in the coordinates of adjacent local data points is relatively small compared to that of the whole dataset. Based on this observation and the fact that CNNs are good at learning local features rather than global ones, the proposed enhancement method is local-based and its detailed description is presented as follows.

Firstly, for a given local area, the grayscale values are redistributed so that their histogram conforms to the Rayleigh Distribution defined in Equation 4-8.

$$f(z) = \frac{z}{\sigma^2} e^{\left(-\frac{z^2}{2\sigma^2}\right)}, z \geq 0 \quad (4-8)$$

where the value of σ is taken as 0.4 so that the expected value of mean grayscale values is 0.5. After this local enhancement was applied, the "hidden" geometrical features in Figure 4-3.e are revealed clearly in Figure 4-3.g, and the corresponding redistributed histogram is shown in Figure 4-3.h. Intuitively, the enhanced Z coordinate image (Figure 4-3.g) contains many detailed geometric features that are distinct from the RGB image in Figure 4-3.c.

In the above example, the local enhancement method essentially magnifies the Z coordinate differences within the local area. However, if there is a general trend for the values within adjacent local areas, applying the local enhancement method individually to each area will result in discontinuous pixel values at the edges of the local areas. For example, the Z -values of the grass area on the right side of Figure 4-3.a gradually increases from the bottom to the top. If the local enhancement method is applied without overlap (the sizes of the local areas are taken as 256*256 pixels) between two adjacent local areas, the bottom pixels of the top local area (e.g., Figure

Chapter 4: Locally enhanced image-based geometric features

4-3.g) are set close to black and the top pixels of the bottom local area (i.e., the local area right below the area representing by Figure 4-3.g) are set close to white. This leads to those horizontal edge discontinuities on the right side of Figure 4-3.i. This phenomenon is the reason for choosing the Rayleigh distribution instead of a uniform distribution in this research. In general, an image with a uniformly distributed histogram will contain the most information (Gonzalez et al., 2009). However, adopting the uniformly distributed histogram means that more points will be distributed close to the two extremes (i.e., zero or one), which will exacerbate the discontinuity at the edges.

To minimize the edge discontinuity, an overlapped local enhancement was used in this study. More specifically, the panoramic image was firstly divided into square areas of the same size that overlap each other by one-eighth of the edge length, and the local enhancement method was applied to each square area. During this process, symmetric padding was used to fill in the blank areas when the actual image area was insufficient. Finally, for the overlapping part, the pixel values were taken as the average of the values of the overlapped pixels. The Z coordinate image enhanced using this method is shown in Figure 4-3.j, where the size of the local square area was taken as 256×256 pixels (same as for Figure 4-3.i) for this example. It can be observed that the edge discontinuity was effectively mitigated by the overlapping strategy. It should be noticed that the size of the local area has a significant effect on the final enhanced image, and the selection of a proper size is demonstrated in Section 4.3.2.

4.2.5. Semantic segmentation network structure

To obtain the semantic information from the fine-resolution panoramic images, a customized CNN was adopted in this research, which consists of two parts: a backbone

and a segmentation head. The entire network structure is shown in Figure 4-4, which is named as HR-EHNet to indicate that it is designed for the segmentation of fine-resolution enhanced panoramic images.

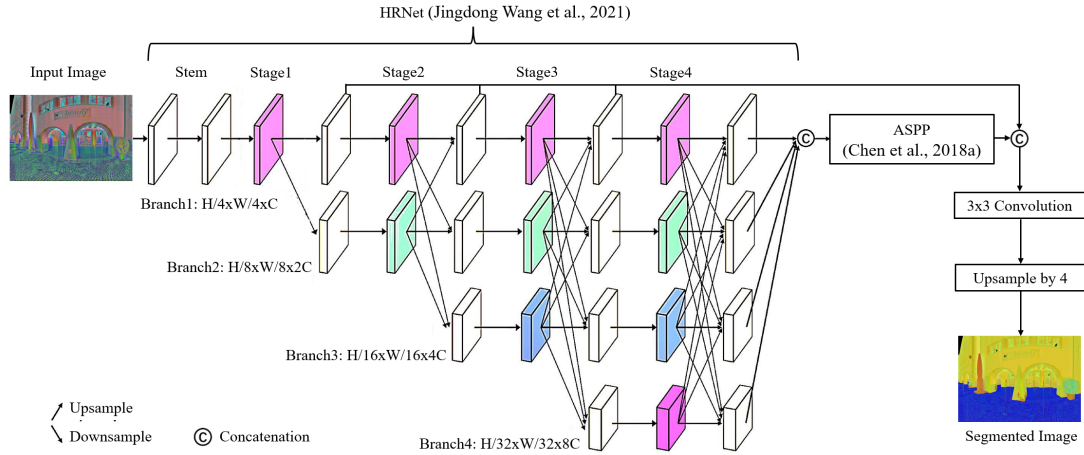


Figure 4-4: Illustration of the HR-EHNet network structure: upsampling and downsampling were implemented by bilinear interpolation and strided 3x3 convolution, respectively; The colored blocks that represent multiple residual convolution operations were performed.

The backbone part is responsible for extracting features from the input images (Jingdong Wang et al., 2021). Although there are various backbone structures available (Chollet, 2017; Gao et al., 2021; J. Hu et al., 2020; Sandler et al., 2018; Szegedy et al., 2017; Jingdong Wang et al., 2021), only HRNet was designed for processing fine-resolution images (Jingdong Wang et al., 2021), which has widely been adopted for excellent semantic segmentation results (Borse et al., 2021; Xu et al., 2020; Yu et al., 2021; Yuan et al., 2020). As such, it was adopted in this study. More specifically, the HRNet_W48 version (larger version) was adopted, where the number 48 indicates the network width of the finest resolution branch. The basic network structure of HRNet is depicted in Figure 4-4. Different from most used single-branch backbones (He et al., 2016), the HRNet has four parallel branches corresponding to four downsample levels (4, 8, 16, and 32, respectively). As for the width of the network (i.e., the number of feature map channels/ the number of convolutional kernels), HRNet adopts a scheme

Chapter 4: Locally enhanced image-based geometric features

where the number of channels is doubled accordingly whenever the resolution of a feature map decreases (Jingdong Wang et al., 2021). Compared to single-branch backbones, HRNet significantly increases the network depth (i.e., the number of convolutional layers) with respect to fine-resolution features, and meanwhile retains coarse-resolution features to provide global contextual information. Since a deeper network structure extends the receptive field and enhances the discrimination of each pixel, the fine-resolution segmentation task could benefit from the deep fine-resolution branch in HRNet.

The segmentation head is responsible for interpreting the extracted features from the backbone to assign an appropriate label to each pixel. The ASPP segmentation head was adopted in this study, which was first proposed by (Chen et al., 2018a) and adopted widely by others (Cai et al., 2021b; Chen et al., 2018b, 2017; Takikawa et al., 2019). The ASPP module employs several parallel atrous (dilated) convolutions with different dilation rates to extract semantic information from different spatial scales (Chen et al., 2018b). The commonly used output stride for the ASPP module is 16 or 8 (16 most commonly in the literature), which means that its input resolution corresponds to a downsampling level of 16 or 8, respectively. This is because most of the backbones are single-branch structures, which generate only high-level features at a relatively high downsampling level. This is not the case for HRNet. Therefore, the ASPP module is attached to the end of the first branch (corresponding to a downsampling level of 4) to take advantage of the fine-resolution features in HR-Net. It was ascertained in previous research (Chen et al., 2018b, 2017) that the proper dilation rate combination for ASPP with an output stride of 16 includes 6, 12 and 18, which should be multiplied by 2 (i.e., 12, 24 and 36) when an output stride of 8 was

Chapter 4: Locally enhanced image-based geometric features

used. Hence, for an output stride of 4, the dilation rate combination is taken as 24, 48 and 72 in this research. Finally, similar to the DeeplabV3+ (Chen et al., 2018b), the output of ASPP is concatenated with three groups of low-level features (corresponding to the outputs of the first three stages of the first branch) for the final segmentation.

4.2.6. Pretraining of network and transfer learning

For image semantic segmentation, it is a consensus that a higher segmentation accuracy can be obtained using pre-trained networks (Ling Shao et al., 2015; Shahin Shamsabadi et al., 2020). This step was also employed in this research where the Cityscapes dataset (Cordts et al., 2016) was used for network pretraining. Similar to Semantic3D, Cityscapes was focused on semantic segmentation in urban scenes and was collected mainly in Europe. Cityscapes contains 5,000 finely labeled fine-resolution RGB images, which were originally divided into 2975, 500, and 1525 images for training, validation and testing, respectively (Cordts et al., 2016). However, since it is beneficial to use a larger dataset for the pretraining, all the training and validation images were used as the training set in this study. Pixels in these images are labelled into 30 classes. Compared to Semantic3D, Cityscapes covers a wider range of urban scenes, has a greater variety of annotations, and suffers from a greater class imbalance.

The training protocol for conducting pre-training followed previous research (Chen et al., 2018a; Jingdong Wang et al., 2021; Zhao et al., 2018, 2017). The stochastic gradient descent with momentum (SGDM) optimizer was adopted. The base learning rate, the momentum and the weight decay were set to 0.01, 0.9, and 0.0005, respectively. The poly learning rate policy was used for dropping the learning rate, where the power was set to 0.9. The focal loss function (Lin et al., 2020) was adopted

Chapter 4: Locally enhanced image-based geometric features

to address the issues of imbalanced classes. The size of the input images was set as 512*1024 pixels. The images were augmented by random cropping, random resize (0.5~2) and random horizontal flipping. Finally, HR-EHNet was trained for 180,000 iterations with a mini-batch size of 8 and synchronized batch normalization.

Since HR-EHNet was pre-trained using the RGB images of Cityscapes, the number of convolutional kernel channels in the first convolutional layer was three, which accepts only three-channel images as its input. However, subsequent experiments in Section 4.3.2-4.3.3 need to use input images with various numbers of channels for comparison. Therefore, in those experiments, the first convolutional layer of the pre-trained HR-EHNet was replaced by a new convolutional layer where its kernel channel number is equal to the number of input features. Meanwhile, the channel number of the convolutional kernels in the last two convolutional layers of the pre-trained HR-EHNet corresponds to the total number of classes (i.e., 19) for Cityscapes. This was replaced by new convolution layers with kernels of 8 channels to accommodate the number of classes in Sementic3D. The weights in these convolution layers were initialized randomly. When HR-EHNet was fine-tuned using the images generated from Semantic3D, almost the same training protocols as those in pre-training were used, except that the iteration numbers were reduced to 60,000 and 75,000 for the training with five-fold cross-validation and for completing the training with the training set, respectively.

4.3. Experiment and results

4.3.1. Information loss from point clouds to images

One of the most frequently quoted drawbacks of image-based approaches is the inevitable information loss during the process where point cloud data are projected to

images (Guo et al., 2021). However, based on the literature surveyed in this research, no previous studies have quantitatively evaluated the information loss in that process. Therefore, a quantitative analysis of information loss was carried out for the projection method proposed in the first place. In this study, the degree of information loss was quantified by comparing the labelling information of the Semantic3D training dataset before and after a complete projection process (i.e., point cloud to image, followed by image to point cloud), in which OA and mIoU were used as the evaluation metrics. Following the projection process described in Section 4.2.3 and using a set of angular resolutions equal to $1/n$ degree (where n equals 1, 2, 3, ... 50), the corresponding OA and mIoU were recorded and shown in Figure 4-5.

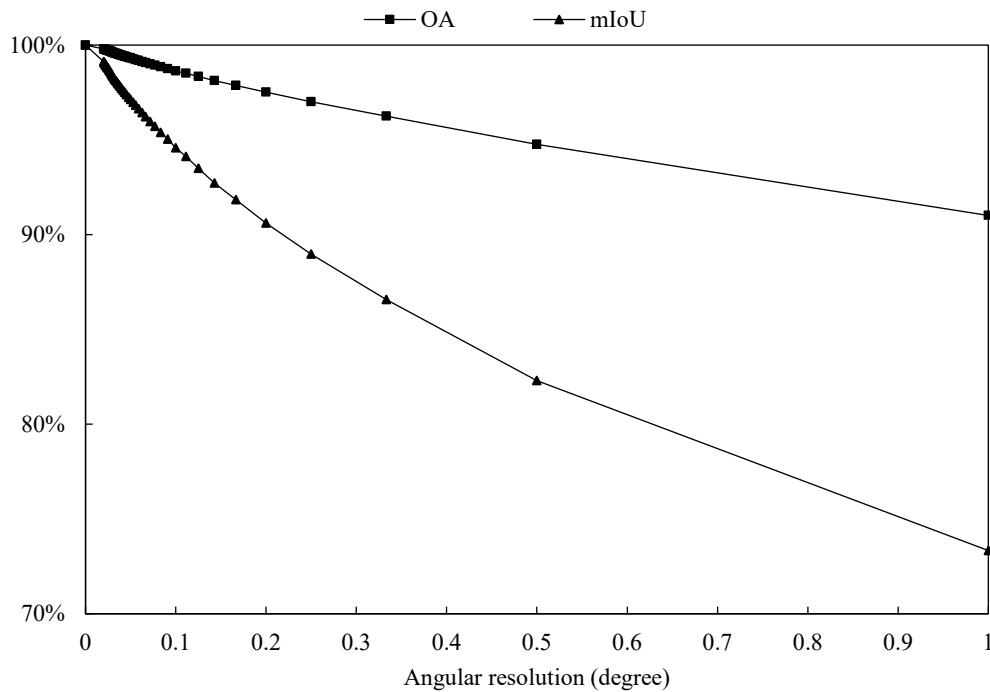


Figure 4-5: Plot of accuracy (OA and mIoU) versus angular resolution.

It is observed in Figure 4-5 that OA and mIoU decreased gradually with increasing angular resolution, and the decreasing rate of mIoU was much higher. Although there was almost no information loss when extremely small angular resolutions were used (e.g., OA = 0.998, mIoU = 0.991 for the angular resolutions of $1/50$ degree), this will

Chapter 4: Locally enhanced image-based geometric features

result in excessive computational demands for subsequent image processing and leave many noisy blank pixels in projected images (e.g., Figure 4-6.a). The angular resolution of 1/20 degree (i.e., an image size of 3600*7200 pixels) was used in this research to perform the point cloud-image projection, as it can provide visually clean projected images (e.g., Figure 4-6.b) with a relatively low information loss (OA = 0.993, mIoU = 0.97).

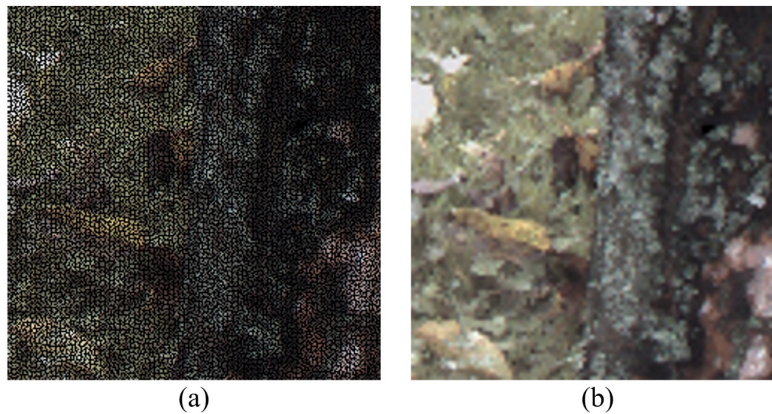


Figure 4-6: Effects of an excessive angular resolution on the projected image: (a). Many black empty pixels for an angular resolution of 1/50 degree, (b). A continuous image without empty pixels for an angular resolution of 1/20 degree.

4.3.2. Effect of local enhancement area on the segmentation results

As mentioned in Section 4.2.4, the size of the local square area used during enhancement has an impact on the enhanced images produced. For example, the enhanced images of the Z coordinate using a local area of 128*128, 32*32, and 8*8 pixels were shown in Figure 4-7.a, Figure 4-7.b and Figure 4-7.c, respectively, in addition to that using a size of 256*256 pixels in Figure 4-3.j. It is seen that there are notable differences in the enhanced images when different local patch sizes are used.

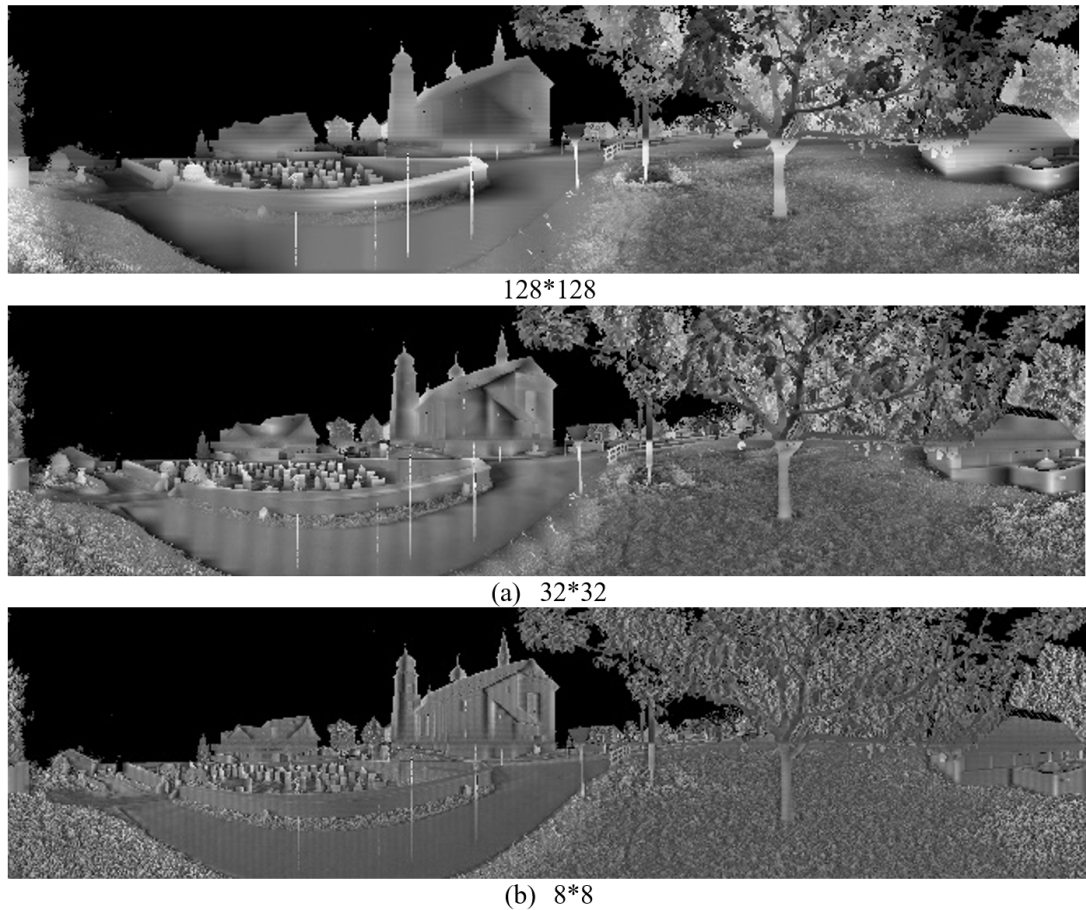


Figure 4-7: Impacts of the local enhancement area on the enhancement results: (a). 128*128 pixels, (b). 32*32 pixels, (c). 8*8 pixels.

To determine an appropriate local patch size for enhancement and to test its effects on the segmentation results, an experiment was conducted for eight local patch sizes (8*8, 16*16... 1024*1024, i.e., $2^{3\sim 10} * 2^{3\sim 10}$). Four groups of original grayscale images were used in this experiment, which were projected from XYZ coordinates and range (D) in Semantic3D, respectively. Each group contain 15 images (corresponding to 15 training point clouds) with a size of 3600*7200 pixels (i.e., an angular resolution of 1/20 degree). These original images were enhanced using each of the eight different sizes, leading to a total of 32 groups of enhanced images. The pre-trained HR-EHNet was fine-tuned on these 36 groups of single-channel images, respectively. The segmentation performances of each group are shown in Figure 4-8.

Chapter 4: Locally enhanced image-based geometric features

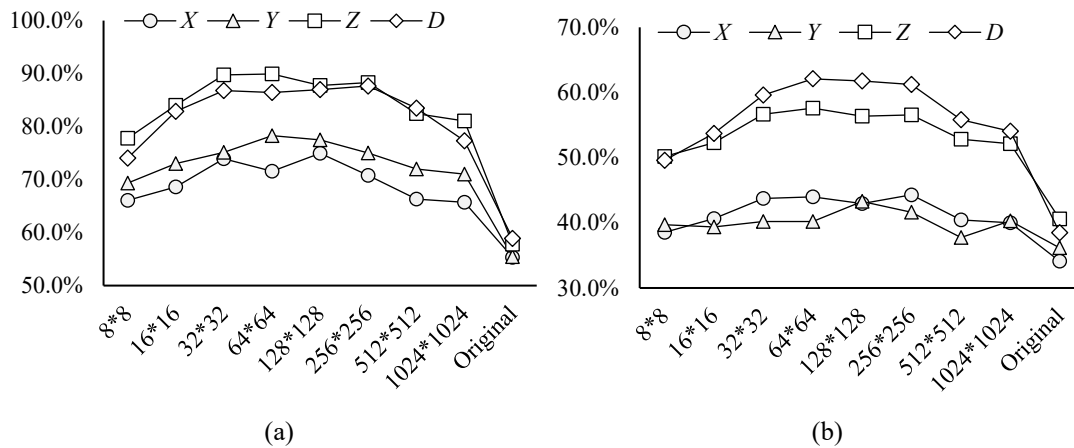


Figure 4-8: Impacts of the local enhancement area on the enhancement results: (a). OA, (b). mIoU.

From Figure 4-8, it is seen that the segmentation accuracy of the network was significantly increased by using the image enhancement in this experiment. However, it was also noticed that the accuracies of the networks trained using the enhanced images derived from X or Y coordinates were considerably less than those based on Z coordinates and range in terms of both OA and mIoU metrics. Therefore, these two types of information (i.e., X and Y coordinates) were not considered in the subsequent sections (i.e., Section 4.3.3 to the end of this chapter). In addition, it is observed that for the images derived from Z coordinates and range, the OA and mIoU metrics were relatively similar when the image enhancement was performed using local area sizes from $32*32$ pixels to $256*256$ pixels. This suggests that the local area size for the image enhancement does not require careful adjustments as long as it is within that range. Nevertheless, since it can be seen from Figure 4-8.b that the images obtained from Z coordinates and range with a local area size of $64*64$ pixels produced the highest mIoU index, this size was selected for this research.

4.3.3. Selecting combinations of feature channels

In this section, various combinations of the channels were tested, including the enhanced Z coordinate images (Z_e), enhanced range images (D_e), and intensity images (I) where the raw intensity values of Semantic3D dataset were used without any corrections. This is followed by tests on conventional combinations involving RGB channels ($IRGBD$ and $IRGB$) that were demonstrated to be relatively accurate channel combinations in previous studies (Cai et al., 2021b). In addition, the combinations of Z_e and D_e with $IRGB$ and $IRGBD$ were tested. A total number of eight combinations of channels were investigated in this research. The test results are shown in Table 4-1.

Table 4-1: quantitative results of different channel combinations on the semantic3d training set (five-fold cross-validation).

Channels	Index	mIoU (%)	OA (%)	man-made	natural	high veg	low veg	buildings	hard scape	scanning art	cars
Z_eD_e	1	68.4	89.3	85.3	75.0	81.9	41.3	95.3	33.7	42.2	92.5
IZ_e	2	66.4	90.1	86.5	75.1	68.3	45.3	93.4	26.8	49.2	86.3
ID_e	3	64.5	88.7	85.5	73.9	71.8	24.0	93.6	27.8	51.5	88.1
IZ_eD_e	4	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
$IRGB$	5	63.8	90.0	85.2	76.5	80.5	39.6	92.7	31.4	33.7	71.0
$IRGBD$	6	66.0	90.4	85.4	74.4	74.6	31.9	93.0	45.2	41.5	82.0
$IRGBZ_eD_e$	7	68.8	90.9	86.5	78.7	83.7	40.6	95.2	41.3	41.9	82.5
$IRGBDZ_eD_e$	8	68.7	90.6	86.4	76.9	81.8	51.0	94.8	36.9	43.5	78.0

Based on the first four sets of experiments, it is observed that using I , Z_e , and D_e together is more accurate than any combination of two of them. In addition, it is clear that the segmentation accuracy achieved by the IZ_eD_e combination was significantly higher than those achieved by the $IRGB$ and the $IRGBD$ combinations. For comparisons of the segmentation accuracy with respect to each class, the segmentation accuracy of IZ_eD_e was found to be higher than the other two combinations ($IRGB$ and $IRGBD$) in most of the classes, especially in recognizing high vegetation and low vegetation. It was also found that the integration of Z_eD_e to $IRGB$ or $IRGBD$ significantly increased their segmentation accuracy in comparison to $IRGB$ or $IRGBD$

Chapter 4: Locally enhanced image-based geometric features

alone, but both cases failed to exceed the segmentation accuracy (mIoU and OA) achieved by the combination IZ_eD_e . However, it was also observed that IZ_eD_e did not perform best for some individual classes. The likely reasons are presented in the following. An individual channel may be favourable to the segmentation of a particular class. However, when multiple channels are combined, their interactions also play an important role in the segmentation accuracy of that particular class. In other words, the network will take into account the trade-off between the contribution of each channel (similar to a weighted average effect) to achieve a higher overall segmentation accuracy for all classes. Consequently, the accuracy of the segmented results of individual classes with or without the use of a particular channel may vary from one to another.

4.3.4. Final performance of HR-EHNet

Based on the experimental results in Table 4-1, the channel combination IZ_eD_e was selected as the final input to HR-EHNet, which happened to be a three-channel image. This means that for this particular combination, the first convolutional layer of the pre-trained HR-EHNet is unnecessarily replaced with a randomly initialized one. According to previous work (Pan et al., 2019), the operation of replacing the first convolutional layer could reduce the segmentation accuracy. Therefore, an experiment was conducted to determine whether to retain the pre-trained first convolutional layer in the final version of HR-EHNet. More specifically, the fourth experiment in Table 4-1 was repeated on the condition that the first pre-trained convolutional layer of HR-EHNet was retained. The corresponding segmentation results are summarized in Table 4-2. As expected, the strategy of retaining the first pre-trained convolutional layer is beneficial for segmentation accuracy in mIoU and, therefore, adopted in the final version of HR-EHNet. All the prerequisites for performing the final training of HR-

EHNet have now been determined. Therefore, the pre-trained HR-EHNet was fine-tuned with the complete training set (i.e., 15 images with IZ_eD_e feature channels and a size of 3600*7200 pixels) for 75,000 iterations according to the training protocols described in Section 4.2.6.

Table 4-2: Impacts of retaining or replacing the first layer of the pre-trained network on the segmentation results when IZ_eD_e were used as the input channels (five-fold cross-validation).

First layer	mIoU	OA	man-made	natural	high veg	low veg	buildings	hard scape	scanning art	cars
Replaced	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
Remain	73.1	91.6	85.5	76.1	89.3	57.3	95.1	46.8	46.8	88.2

The performance of HR-EHNet was evaluated on the Semantic3D (reduced-8) test dataset, which contains four point clouds. The four pseudo colour images of IZ_eD_e and the corresponding segmentation results are illustrated in Figure 4-9. Through visual inspection, it is observed that the majority of the objects are correctly segmented and that most of the mislabels are concentrated at the edges where different objects intersect. These two-dimensional segmentation results were projected onto each data point in the point clouds to produce the segmented point clouds, which were uploaded to the online evaluation system of Semantic3D. The evaluation results have been made publicly available in the Semantic3D website under the name HR-EHNet (IZ_eD_e). The quantitative results of HR-EHNet and the recently published methods on Semantic3D (reduced-8) are summarized in Table 4-3. Without RGB channels, HR-EHNet significantly outperforms the best outcomes of the previous image-based methods by 2.7% (OA) and 10.7% (mIoU), and meanwhile performed better than most of the point-based methods. It is also noted that HR-EHNet achieved the best segmentation accuracy with respect to high vegetation and cars among all the published methods.

Chapter 4: Locally enhanced image-based geometric features

Table 4-3: Quantitative results (%) of different approaches on Semantic3D (reduced-8).

		Time (s)	Params (M)	mIoU	OA	man- made	natural.	high veg	low veg	buildings	hard scape	Scanning art	cars
Point- based Methods	RF MSSF (Thomas et al., 2018)	1643.75	-	62.7	90.3	87.6	80.3	81.8	36.4	92.2	24.1	42.6	56.6
	ShellNet (Zhang et al., 2019)	3000	0.48	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
	OctreeNet (F. Wang et al., 2020)	184.84	-	59.1	89.9	90.7	82.0	82.4	39.3	90.0	10.9	31.2	46.0
	GACNet (L. Wang et al., 2019)	1380	-	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
	SPGraph (Landrieu and Simonovsky, 2018)	3000	0.25	73.2	94.0	97.4	92.6	87.9	44.0	83.2	31.0	63.5	76.2
	KPConv (Thomas et al., 2019)	600	14.9	74.6	92.9	90.9	82.2	84.2	47.9	94.9	40.0	77.3	79.7
	RandLA-Net (Q. Hu et al., 2020)	-	0.95	77.4	94.8	95.6	91.4	86.6	51.5	95.7	51.5	69.8	76.8
	RFCR (Gong et al., 2021)	-	-	77.8	94.3	94.2	89.1	85.7	54.4	95.0	43.8	76.2	83.7
Projection- based Methods	DeePr3SS (Lawin et al., 2017)	-	134	58.5	88.9	85.6	83.2	74.2	32.4	89.7	18.5	25.1	59.2
	SnapNet (Boulch et al., 2018)	3600	29	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4
	XJTLU (Cai et al., 2021b)	5.13	70.6	63.5	89.4	85.4	74.4	74.6	31.9	93.0	25.2	41.5	82.0
	HR-EHNet (This study)	11.72	73.6	74.2	92.1	85.1	75.5	89.6	55.9	95.5	50.8	48.3	92.5

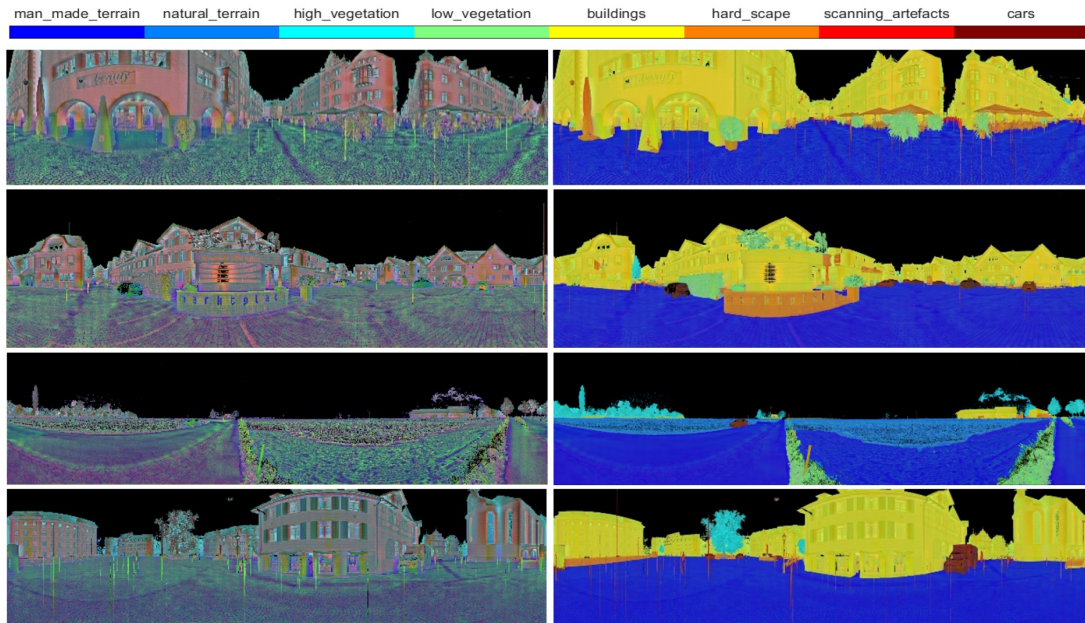


Figure 4-9: (a). The pseudo color images of $IZ_e D_e$ feature channels for the point clouds in Semantic3D (reduced-8) test set, (b). The corresponding segmentation results (The legend is only for the visualization of the segmentation results in (b)).

The time spent on each step of HR-EHNet is recorded in Table 4-4. The data used in this test is the Semantic3D (reduced-8) test dataset, where the four-point clouds contain a total of 78.7 million data points. The inference was conducted with an AMD 3700X @3.6GHz CPU and an NVIDIA RTX2080Ti GPU. The total processing time was 11.72s, which was much faster than the other methods in Table 4-3 except XJTLU (Cai et al., 2021a). As shown in Table 4-4, HR-EHNet is slower than XJTLU because of the additional image enhancement step used.

Table 4-4: The times taken by each step of HR-EHNet to process the Semantic3D (reduced-8) test dataset.

	Time (s)	% of total time
Point cloud-image projection	0.17	1.5%
Enhancement	6.89	58.8%
Inference with neural network	4.55	38.8%
Image-point Cloud projection	0.11	1.0%
Total time	11.72	-

4.4. Discussion

The core idea of HR-EHNet is to provide CNNs with distinguishable local geometric characteristics by enhancements of images derived from point cloud data. In this research, local image enhancement was implemented by a hand-crafted algorithm. Although the image enhancement method proposed was experimentally demonstrated to be effective and insensitive to the local patch size, it consumed more than half of the processing time as shown in Table 4-4. Considering that image enhancement is a relatively simple task in comparison to image segmentation, it is worth investigating how to reduce its processing time in the future. For example, one potential solution is to use the current image enhancement results as the target images to train a relatively simple neural network.

In this research, not all possible channel combinations were tested and as such there is

Chapter 4: Locally enhanced image-based geometric features

no guarantee that IZ_eD_e is the best among all possible channel combinations. This is because the computational effort required would be enormous and the focus of this research was not on screening the optimal channel combinations. As such, developing an efficient way to identify optimal channel combinations is highly desirable in future research. Nevertheless, the results in this research showed that the channel combination IZ_eD_e represents a promising choice.

The experimental results in Section 4.3.3 show that adding additional information (e.g., RGB or RGBD) to IZ_eD_e had a negative impact on the overall results (i.e., mIoU and OA). The primary reason for this phenomenon is the low reliability of the RGB images as mentioned Section 4.1. For example, the RGB and the Z_e images of the same scene are shown in Figure 4-10.a and Figure 4-10.b, respectively. The RGB image shows a cyclist that does not exist in the Z_e image because the acquisition was not done simultaneously. The experimental results in Table 4-1 indicate that such false RGB information is an obstacle for neural networks to learn correct features.

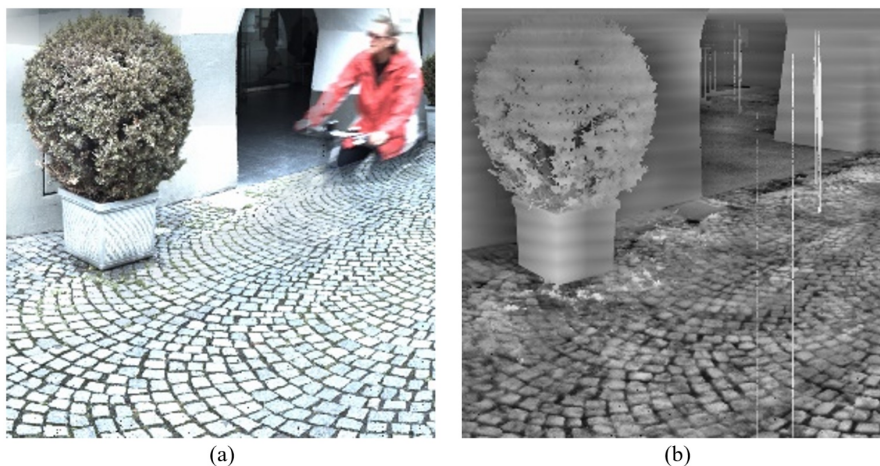


Figure 4-10: Incorrect RGB information in TLS point cloud data: (a). The RGB image contains the cyclist that were not scanned by TLS, (b). The enhanced Z image for the same scene.

Chapter 4: Locally enhanced image-based geometric features

More evidence is shown in Figure 4-11, where Figure 4-11.b shows the classes predicted using the pseudo colour images of $IZ_e D_e$ in Figure 4-11.a, and Figure 4-11.d shows the classes predicted using $IRGBZ_e D_e$ (i.e., $IZ_e D_e$ in Figure 4-11.a and the RGB images in Figure 4-11.c). For Scene 1, it is seen that the vehicle in Figure 4-11.b was correctly segmented using only $IZ_e D_e$. However, when the erroneous RGB information was added, chaotic segmentation results (Figure 4-11.d) were obtained. A similar situation occurred for the vase in Scene 2. Nevertheless, Table 4-1 shows that for some particular classes (i.e., buildings, hardscape, man-made and natural), combining RGB information with $IZ_e D_e$ improved their segmentation accuracies. This is because the accuracy of the RGB information is uncertain. When the RGB information of a particular class is accurate, it may be beneficial to include the RGB information for the segmentation of that particular class. For example, although the vehicle was segmented incorrectly in Scene 1 in Figure 4-11.d due to the erroneous RGB information, the vegetation at the windows was segmented correctly due to the correct and high contrast RGB information. However, when the RGB information of two adjacent classes does not show clear contrast, the inclusion of it may be problematic for the segmentation as demonstrated in the next paragraph. Future research may address the quality of RGB information from two perspectives. The most straightforward solution is to design a TLS strategy that simultaneously collects point cloud and imagery data to reduce as much false information as possible. The second possible solution is to design a neural network structure to enhance its ability to discriminate correct information from redundant/false information.

Chapter 4: Locally enhanced image-based geometric features

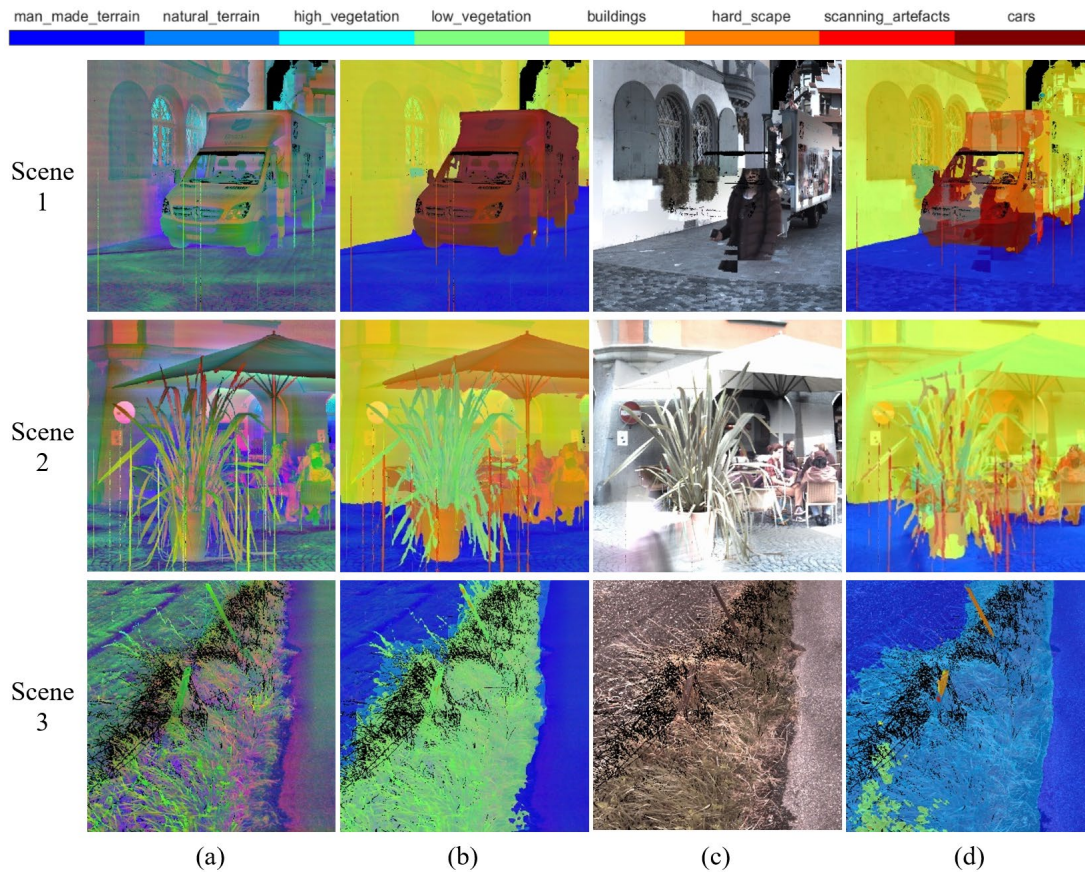


Figure 4-11: Comparisons between $IZ_e D_e$ and $IRGBZ_e D_e$ on segmentation results for three scenes: (a). The pseudo color images of $IZ_e D_e$, (b). The segmentation results using the corresponding $IZ_e D_e$ images, (c) The RGB images, (d). The segmentation results using the feature combination of $IRGBZ_e D_e$.

Compared with other segmentation methods, HR-EHNet was found to performance excellently in the recognition of plants and vehicles. This finding suggests its potential application to applications such as forest classification and autonomous driving. The possible reason for lower accuracies in the other methods for these two types of objects is that the use of the RGB information may cause confusion to neural networks if plants or vehicles have similar colours (i.e., spectra) to their surrounding objects. In contrast, HR-EHNet performs semantic segmentation mainly via the geometric features in the enhanced images, which are independent of colour and can replace RGB images. For example, although the RGB colours of the vegetation in Scene 2 and Scene 3 seem to be accurate visually, the segmentation results obtained after adding the RGB

Chapter 4: Locally enhanced image-based geometric features

information became worse due to its similarity to the surrounding objects. Furthermore, this characteristic is presumed to have significant advantages in terms of resistance to adversarial attacks, which may offer better security for certain applications such as autonomous driving. There are many studies demonstrating that deep learning relying on RGB images is vulnerable to colour perturbation attacks (Duan et al., 2020; Eykholt et al., 2018; Shahin Shamsabadi et al., 2020). For example, shining light from a laser pointer on a stop sign may cause neural networks to fail to recognize the stop sign, which poses a significant safety challenge for autonomous driving (Duan et al., 2020). Thus, it may be beneficial to extend the idea in this research to point cloud data that are used typically for a wider range of applications, including autonomous driving.

At present, there is only one TLS point cloud dataset (i.e., Semantic3D) publicly available for evaluating algorithms. Although Semantic3D is a large point cloud dataset in terms of the number of data points, it is small when it is processed as an image dataset (i.e., project each point cloud as a panoramic image), in comparison to image datasets such as the Cityscapes and Mapillary Vistas datasets (Cordts et al., 2016; Hackel et al., 2017; Zhang et al., 2021). Therefore, establishing a larger point cloud dataset would be extremely beneficial to the development of relevant research fields. It is also thought interesting to explore the feasibility of few-shot learning using the relatively small existing point cloud dataset.

Only 15 labelled panoramic images can be derived from the Semantic3D training set, which is not sufficient to support the decent training of HR-EHNet from scratch. Therefore, the Cityscapes dataset - that was taken in similar urban scenes and semantically labelled - was used for the network pre-training in this study. However,

Chapter 4: Locally enhanced image-based geometric features

such semantically labelled images are often not publicly available. In contrast, unlabelled image datasets can readily be obtained for various application scenarios, through online resources and/or field acquisitions. Therefore, it is interesting to investigate how to effectively use techniques such as self-supervised learning (He et al., 2020; Pathak et al., 2016; Z. Wu et al., 2018; Zhan et al., 2019; Zhang et al., 2017) to pre-train networks using unlabelled images.

4.5. Summary

In this chapter, a novel image enhancement method was proposed to characterize effectively the local geometric features in the panoramic images derived from TLS point cloud data. The enhanced images (i.e., enhanced Z -coordinates Z_e , and enhanced range D_e) alone and in various combinations of other popular feature channels (i.e., intensity I , RGB, range D) were used in a pre-trained CNN to assess the potential for semantic segmentation of the Semantic 3D datasets. It was found that compared with the commonly used channel combinations I RGB or I RGBD, the proposed combination I Z_e D_e produced more accurate semantic segmentation predictions. By fine-tuning the customized pre-trained HR-EHNet with the channel combination I Z_e D_e , an OA of 92.1% and a mIoU of 74.2% were obtained on the Semantic3D (reduced-8) test dataset, which substantially outperformed the other image-based methods. This suggests that effective utilization of local geometric features in images can increase the segmentation accuracy of image-based methods. This study also offers a better alternative channel combination to replace those involving the RGB channels, which may be extremely useful for cases where the RGB information is absent or inaccurate.

Chapter 5: Automatic feature extraction

This chapter is based on the published paper: Cai, Y., Fan, L., Zhang, C., 2022b. Semantic Segmentation of Multispectral Images via Linear Compression of Bands: An Experiment Using RIT-18. Remote Sens. 14, 2673. <https://doi.org/10.3390/rs14112673>

Note: The experiments in Chapter 4 (Table 4-2) show that using the pre-trained weights intact (i.e., retaining the weights of the first layer) can improve segmentation accuracy. This requires the number of channels of the input images to match that of the pre-trained images i.e. for multichannel images it is required to reduce the dimensionality to 3. There are two types of dimension reduction methods. One is feature extraction and the other is feature selection. This chapter develops a learnable feature extraction method. The effectiveness of the proposed method is demonstrated in this chapter with a multispectral semantic segmentation dataset named RIT-18. The test results on Semantic3D can be found in Chapter 6.

5.1. Introduction

Semantic segmentation of images is a fundamental task in computer vision, in which a label is assigned to each pixel. For remotely sensed imagery, its semantic segmentation (known as pixel-based classification previously) is the basis for many applications, such as forest monitoring, cloud detection, and land-use planning (Dechesne et al., 2017; Dong et al., 2018; Goldblatt et al., 2018; Marmanis et al., 2018). There are sensors that can capture images with more than three bands (e.g., multispectral and hyperspectral images). Compared to the three bands obtained by RGB cameras, the additional spectral information of multispectral images could be used to, potentially, achieve a higher segmentation accuracy. However, semantic segmentation of multispectral images is challenging due to the limited training samples and high-dimensional features.

Compared to RGB sensors with fixed spectral bands, the spectral bands between different multispectral sensors are usually different. This means that the multispectral image datasets obtained by different multispectral sensors are unique. Coupled with the fact that data annotation is very time-consuming, the total amount of annotated data available for a specific multispectral semantic segmentation task is typically very limited. Therefore, relative lightweight networks (Boulch et al., 2018; Kemker et al., 2018; Lawin et al., 2017; Mateo-García et al., 2020; Saxena et al., 2020) were often used for multispectral semantic segmentation to avoid overfitting. However, it is a consensus that deeper and wider neural networks that have been pre-trained on large-scale datasets can achieve a higher segmentation accuracy compared to lighter neural networks without pre-training. Many neural networks developed for RGB images have become deeper and wider and achieved excellent performances on many challenging

benchmark datasets. Hence, the accuracy of multispectral semantic segmentation may be improved significantly if the multispectral images can be tailored properly to fit the pre-trained state-of-the-art networks. Due to the peak phenomenon (Kallepalli et al., 2014; Sima and Dougherty, 2008; Theodoridis and Koutroumbas, 2001), direct feeding of multispectral images to networks developed originally for three-channel images often leads to poor segmentation accuracy (Bhuiyan et al., 2020; Cai et al., 2022a, 2021b). The most direct solution to this problem is to reduce the input image dimension to three.

There are mainly two types of methods for image dimensionality reduction, i.e., feature extraction (Bandos et al., 2009; Belkin and Niyogi, 2003; Bhatti et al., 2022; Farrell and Mersereau, 2005; Fauvel et al., 2009; Jing Wang and Chein-I Chang, 2006; Roweis and Saul, 2000; Tenenbaum et al., 2000) and band selection (Bhuiyan et al., 2020; Hu et al., 2019; Roy et al., 2021b; Su et al., 2011; Sun et al., 2020; W. Sun et al., 2022; Zhu et al., 2021). The representative feature extraction methods include principle component analysis-based methods (Farrell and Mersereau, 2005; Fauvel et al., 2009; Uddin et al., 2021; Xiuping Jia and Richards, 1999; Zabalza et al., 2014), independent component analysis-based methods (Jing Wang and Chein-I Chang, 2006), linear discriminant analysis-based methods (Bandos et al., 2009; Du, 2007; Wang et al., 2017), and locality preserving projection-based methods (Deng et al., 2018; Li et al., 2012). Most of these methods perform linear transformations of the original spectral bands to optimize their objectives. One common objective in the feature extraction methods is to retain as much information as possible in the processed images (in which the dimension is reduced). Existing feature extraction methods are generally fast in processing, but the changes in the original spectral reflectance may

Chapter 5: Automatic feature extraction

cause difficulties in physical interpretations and hinder the applications where physical spectral measurements are required. In addition, since the optimization objectives of those methods are not on the segmentation accuracy of the neural networks, they do not guarantee a decent performance in segmentation.

Band selection refers to the selection of a subset of spectrum bands from the original image. Depending on the usage of labelled images during the selection process, band selection can be categorized as unsupervised (S. Huang et al., 2022; Jia et al., 2022, 2016; Martínez-UsóMartinez-Uso et al., 2007; Q. Wang et al., 2019, 2018; B. Xu et al., 2021) and supervised (Cao et al., 2016; Chang et al., 1999; Demir and Ertürk, 2008; Feng et al., 2017; Guo et al., 2006; Huang and He, 2005; Keshava, 2004; Yang et al., 2011) methods. The former one is to select the most representative bands based on their statistical characteristics, such as dissimilarity, information entropy, information divergence, or correlation. The latter one typically uses the labelled images to select a band combination that maximizes class separability. In general, the band selection methods require an iterative trial of band combinations, which is more computationally intensive than the feature extraction methods. As for the advantages, unsupervised band selection methods are more popular among scholars as they do not require labelled data and have more application scenarios. Meanwhile, the supervised band selection methods have been proven to achieve better segmentation accuracy than other types of existing methods.

Motivated by the aforementioned limitations of the current methods and inspired initially by the fact that adjacent bands of the multispectral/hyperspectral images are usually relatively similar, a hot-pluggable head structure for the linear compression

(referred to as LC-Net) of bands is proposed in this study for a supervised feature extraction, which directly optimizes the segmentation accuracy. The main advantages of LC-Net are its compatibility with existing networks and faster training speed while maintaining similar accuracy compared to the supervised grid search (SGS). The structure of LC-Net can be kept as simple as possible, which adds negligible computational costs to the training and inference processes. The effectiveness of LC-Net was tested using three different networks on the RIT-18 benchmark dataset (Kemker et al., 2018) in this study.

5.2. Materials and Methods

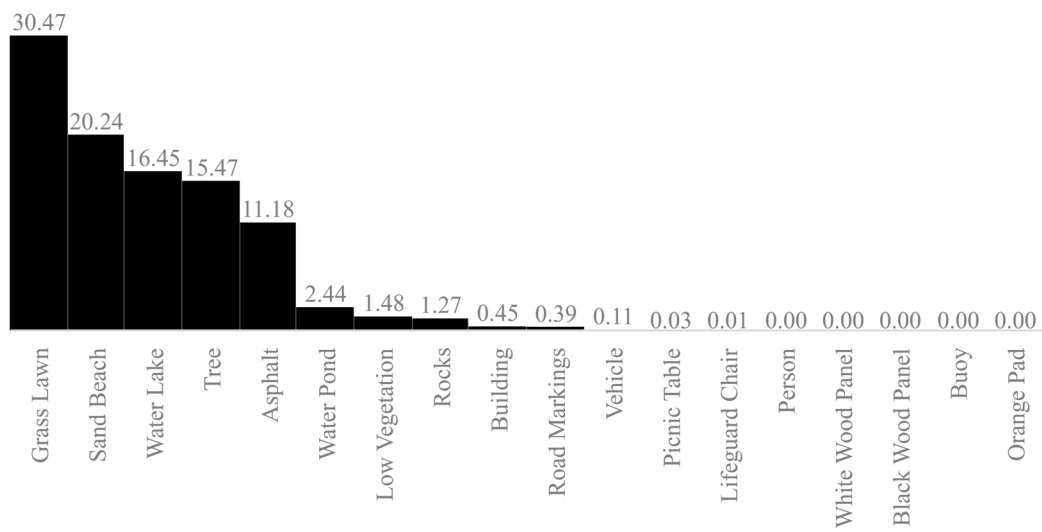
5.2.1. Study Data

The RIT-18 (Kemker et al., 2018) multispectral dataset used in this study is a six-band multispectral dataset that includes visible RGB bands (band 3 for R, band 2 for G, and band 1 for B) and three near-infrared bands (bands 4–6). It contains a training image with a resolution of 9393×5642 and a test image with a resolution of 8833×6918 . Each pixel of the images in RIT-18 is assigned to one of the eighteen classes. A comparison between RIT-18 and other commonly used publicly available multispectral datasets is summarized in Table 5-1. It shows that RIT-18 has the largest number of labelled classes and the finest ground sample distance (GSD) compared to ISPRS Vaihingen and Potsdam (Rottensteiner et al., 2012), Zurich Summer (Volpi and Ferrari, 2015), and L8 SPARCS (Hughes and Hayes, 2014). Together with the highly unbalanced classes, as shown in Figure 5-1, they make RIT-18 a very challenging dataset. In addition, as a 6-band dataset, RIT-18 has a total of 20 combinations of three bands, which is neither too many (120 possible combinations), as in L8 SPARCS, nor too few, as in other datasets. Therefore, as a challenging but computationally affordable dataset, RIT-18 was considered in this study.

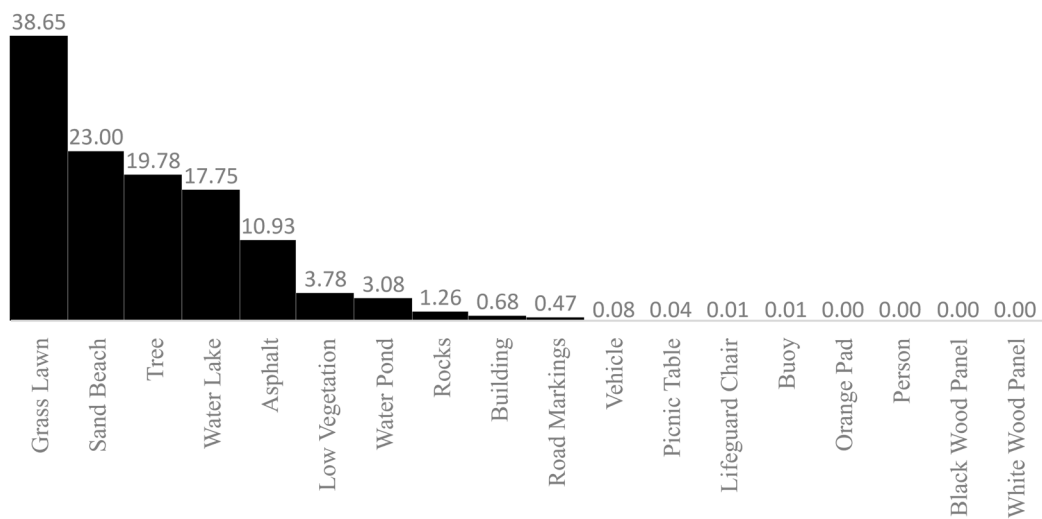
Chapter 5: Automatic feature extraction

Table 5-1: Summary of five commonly used multispectral datasets.

Dataset	Sensor(s)	GSD (m)	Number of classes	Number of bands	Distribution of bands
Vaihingen	Green/Red/IR	0.09	5	3	Green, Red, IR
Potsdam	VNIR	0.05	5	4	Blue, Green, Red, NIR
Zurich Summer	QuickBird	0.61	8	4	Blue, Green, Red, NIR
L8 SPARCS	Landsat 8	30	5	10	Coastal, Green, Red, NIR, SWIR-1, SWIR-2, Pan, Cirrus, TIRS
RIT-18	VNIR	0.047	18	6	Blue, Green, Red, NIR-1, NIR-2, NIR-3



(a)



(b)

Figure 5-1: Percentage of each class in RIT-18: (a) training image (%), (b) test image (%).

To avoid any confusion, it is worth mentioning that there are 18 classes in the training images but only 16 classes in the test images, which explains why 16 classes are shown in Section 5.3 and Section 5.4.

5.2.2. LC-Net (contains modifications that do not affect the results in this chapter)

Despite the rich spectral information contained in the multispectral images, these additional bands cause new challenges, not only in terms of more GPU memory and computational consumption but also in the peak phenomenon (Kallepalli et al., 2014; Sima and Dougherty, 2008; Theodoridis and Koutroumbas, 2001). This phenomenon shows that the use of additional features (e.g., spectral bands) introduces complexity to the classifier and increases the number of parameters and training data needed to achieve the same classification accuracy (using fewer features). Directly using more features may lead to worse classification results (Hughes, 1968; Theodoridis and Koutroumbas, 2001). To address this problem, LC-Net is proposed in this study to reduce the number of input channels for the subsequent network at the initial stage. This prevents the subsequent network from receiving too many features while allowing the network to choose its preferred features.

The LC-Net is defined as a 1×1 group convolution layer with $n_{per\ group}$ bands per group at the initial stage of the networks. Based on the original band number $n_{original}$, $n_{per\ group}$ is determined as follows:

$$n_{per\ group} = \text{round up to the nearest integer}\left(\frac{n_{original}}{3}\right) \quad (5-1)$$

For cases where the number of original bands is not divisible by three, n_{blank} blank band(s) is added to the original multispectral image, n_{blank} is calculated as:

$$n_{blank} = 3 \times n_{per\ group} - n_{original} \quad (5-2)$$

Chapter 5: Automatic feature extraction

As adding a blank band does not provide any new features to the model, it does not introduce additional complexity and, thus, does not suffer from the peaking phenomenon. This setup will compress $n_{per\ group}$ adjacent input bands in sequence without any overlap. The operation of LC-Net for RIT-18 dataset is shown in Figure 5-2. More specifically, each band was multiplied by a weight, and, after that, two adjacent bands (i.e., bands 1 and 2, bands 3 and 4, and bands 5 and 6 for the RIT-18 dataset) were added together to form three new bands. The weights applied to the bands were randomly initialized and iteratively updated during the training.

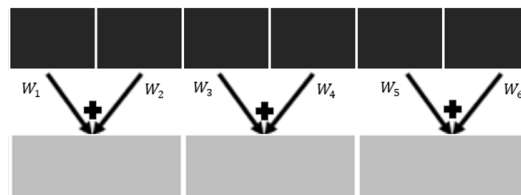


Figure 5-2: Linear combinations of two adjacent bands without any overlap for the RIT-18 dataset.

In addition, for a more comprehensive understanding, the performance of combining non-adjacent bands was also tested. The implementation details are shown in Section 5.2.5. Unless stated otherwise, the LC-Net referred to in the following sections is based on the combination of adjacent bands.

As shown in Figure 5-3, the feature maps obtained from LC-Net can be fed to any existing networks for subsequent processing after passing through a batch normalization layer. There is no non-linear activation layer used after the batch normalization layer, not only because it is not consistent with the linear compression that LC-Net aims to achieve but also because the application of activation functions to low-dimensional features (3 dimensions in this case) can degrade network performance, as demonstrated by many studies (Chollet, 2017; Sandler et al., 2018; X. Zhang et al., 2018).

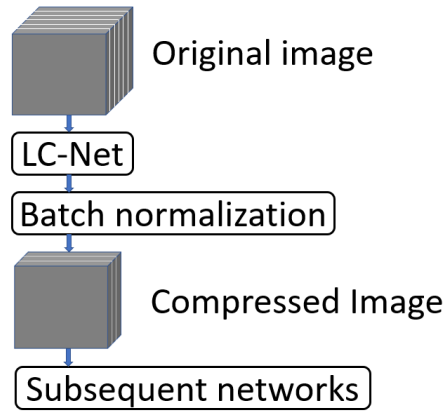


Figure 5-3: Apply LC-Net to the networks.

The subsequent networks shown in Figure 5-3 include an encoder and a decoder. For the encoder (i.e., backbone), the networks used in this study consist of ResNet50 (He et al., 2016), HRNet-w18 (Jingdong Wang et al., 2021), and Swin-tiny (Liu et al., 2021a), which are popular convolutional neural networks and vision transformers. All the backbones are pre-trained using the ImageNet dataset (Russakovsky et al., 2015). For the decoder, the same decoder structure of Fully Convolutional Networks (Shelhamer et al., 2017) is used for all the networks adopted in this study. The detailed network structures faithfully follow their original implementations and are described in detail in Section 5.2.3.

5.2.3. Network Structure

The backbones (i.e., ResNet (He et al., 2016), HRNet (Jingdong Wang et al., 2021), and Swin (Liu et al., 2021a)) used in this study can be considered three milestones in the network's structure design. For example, the residual connection proposed by ResNet (He et al., 2016) has become a standard paradigm for subsequent network designs. The basic block design of ResNet (He et al., 2016) is shown in Figure 5-4, which consists of a stack of convolution layers, a Batch Normalization (BN) layer, and a Rectified Linear Unit (ReLU). The complete ResNet (He et al., 2016) is constructed

Chapter 5: Automatic feature extraction

by connecting the stem and the multiple basic blocks in a series. The detailed architecture specifications of ResNet-50 are presented in Table 5-2. The overall design follows two rules: One is to apply the same hyperparameters (the width and filter size) to the blocks of the same spatial resolution. The other is that when the spatial resolution is reduced by half, the width of the block doubles. The second rule ensures that all blocks have approximately the same computational complexity in terms of floating-point operations (FLOPs). Similar to residual connection, these two rules were applied to most network designs, including Swin (Liu et al., 2021a) and HRNet (Jingdong Wang et al., 2021).

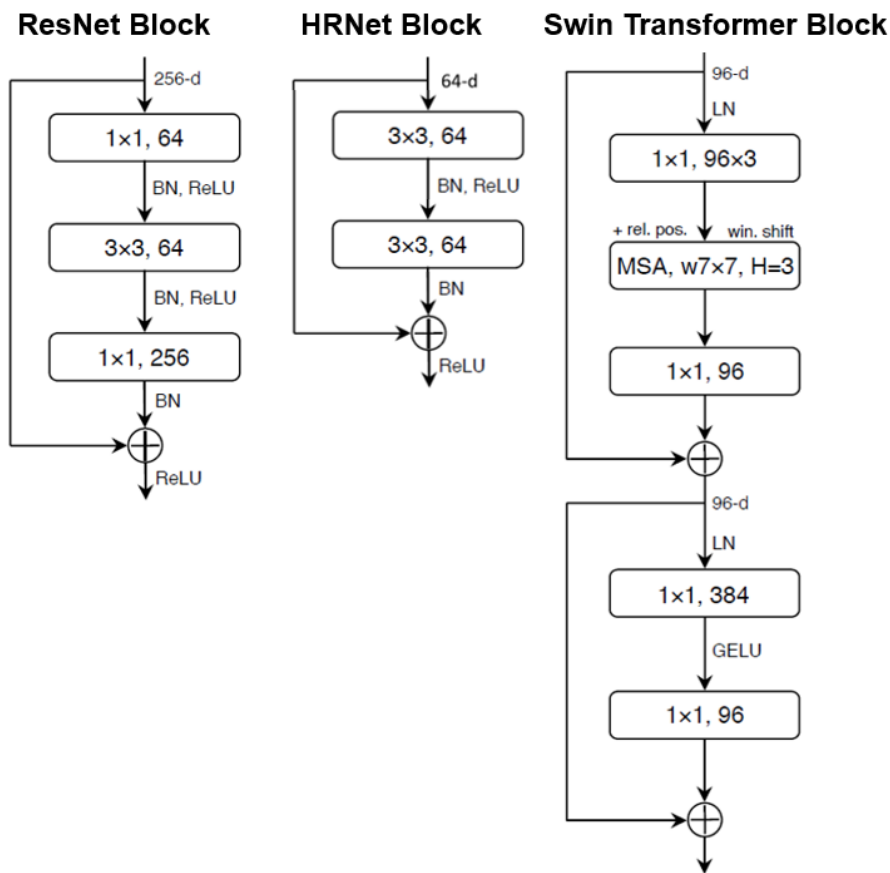


Figure 5-4: Basic block designs for ResNet, HRNet, and Swin.

As shown in Table 5-2, the macro design of Swin (Liu et al., 2021a) is similar to that of ResNet (He et al., 2016). Both are single-branch structures. As the network depth increases, the spatial resolution gradually decreases, and the width increases. Their main differences are within the basic blocks, as shown in Figure 5-4 and Table 5-2. The basic block of Swin is built by a Multi-head Self-Attention (MSA) module with shifted windows (win. shift) and relative position bias (rel. pos), followed by a two Multi-Layer Perceptron (MLP) with a Gaussian Error Linear Unit (GELU) in between. For clarity, the MLP layers in Swin (Liu et al., 2021a) are noted as a “1 × 1 convolution” in Figure 5-4 since they are equivalent.

Table 5-2: Detailed architecture specifications of ResNet50 and Swin-tiny, where the bracket indicates a residual block, and the number outside the brackets is the number of stacked blocks for the stage.

	Output Size	ResNet50	Swin-Tiny
Stem	$\frac{H}{4} \times \frac{W}{4}$	$7 \times 7, 64, \text{stride } 2$ $3 \times 3 \text{ max pool, stride } 2$	$4 \times 4, 96, \text{stride } 4$
Resolution 1	$\frac{H}{4} \times \frac{W}{4}$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 96 \times 3 \\ MSA, w7 \times 7, H = 3, rel. pos. \\ 1 \times 1, 96 \\ 1 \times 1, 384 \\ 1 \times 1, 96 \end{bmatrix} \times 2$
Resolution 2	$\frac{H}{8} \times \frac{W}{8}$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 192 \times 3 \\ MSA, w7 \times 7, H = 6, rel. pos. \\ 1 \times 1, 192 \\ 1 \times 1, 768 \\ 1 \times 1, 192 \end{bmatrix} \times 2$
Resolution 3	$\frac{H}{16} \times \frac{W}{16}$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 384 \times 3 \\ MSA, w7 \times 7, H = 12, rel. pos. \\ 1 \times 1, 384 \\ 1 \times 1, 1536 \\ 1 \times 1, 384 \end{bmatrix} \times 6$
Resolution 4	$\frac{H}{32} \times \frac{W}{32}$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 768 \times 3 \\ MSA, w7 \times 7, H = 24, rel. pos. \\ 1 \times 1, 768 \\ 1 \times 1, 3072 \\ 1 \times 1, 768 \end{bmatrix} \times 2$
FLOPs		4.1×10^9	4.4×10^9
Parameters		25.6×10^6	28.3×10^6

The main innovation of HRNet (Jingdong Wang et al., 2021) is the use of four parallel branches of the same depth, corresponding to four down-sampling levels (4, 8, 16, and 32). Compared to single-branch backbones, HRNet (Jingdong Wang et al., 2021)

Chapter 5: Automatic feature extraction

significantly increases the network depth with respect to fine-resolution features. This proves to be beneficial for pixel-level image processing tasks (Borse et al., 2021; Xu et al., 2020; Yu et al., 2021). The basic block and detailed architecture specifications of HRNet-w18 are shown in Figure 5-4 and Table 5-3, respectively.

Table 5-3: Detailed architecture specifications of HRNet-w18, where the bracket indicate a residual block, and the number outside the brackets is the number of stacked blocks for the stage.

	Output Size	Stem	Stage 1	Stage 2	Stage 3	Stage 4
Resolution 1	$\frac{H}{4} \times \frac{W}{4}$	$3 \times 3, 64,$ stride 2 $3 \times 3, 64,$ stride 2	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 3, 18 \\ 3 \times 3, 18 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 3, 18 \\ 3 \times 3, 18 \end{bmatrix} \times 16$	$\begin{bmatrix} 3 \times 3, 18 \\ 3 \times 3, 18 \end{bmatrix} \times 12$
Resolution 2	$\frac{H}{8} \times \frac{W}{8}$			$\begin{bmatrix} 3 \times 3, 36 \\ 3 \times 3, 36 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 3, 36 \\ 3 \times 3, 36 \end{bmatrix} \times 16$	$\begin{bmatrix} 3 \times 3, 36 \\ 3 \times 3, 36 \end{bmatrix} \times 12$
Resolution 3	$\frac{H}{16} \times \frac{W}{16}$				$\begin{bmatrix} 3 \times 3, 72 \\ 3 \times 3, 72 \end{bmatrix} \times 16$	$\begin{bmatrix} 3 \times 3, 72 \\ 3 \times 3, 72 \end{bmatrix} \times 12$
Resolution 4	$\frac{H}{32} \times \frac{W}{32}$					$\begin{bmatrix} 3 \times 3, 144 \\ 3 \times 3, 144 \end{bmatrix} \times 12$
FLOPs				4.3×10^9		
Parameters				21.3×10^6		

5.2.4. Training Setting

The experiment was carried out on a PC with a processor of AMD Ryzen 9 3950X, RAM of 64 GB, and two GPUs of NVIDIA GeForce GTX 3090. In addition, the PyTorch framework (MMSegmentation Contributors, 2020) in Ubuntu (20.04) was used for programming the experiment. For a fair comparison of the whole experiment process, all the training used the same training protocol, which is a strategy used widely in the research of deep learning (Dosovitskiy et al., 2020; Liu et al., 2021a; Xie et al., 2021). More specifically, the AdamW optimizer was adopted with the following setup: a base learning rate of 0.00006, a weight decay of 0.01, a linear learning rate decay, and a linear warmup of 1500 iterations.

For the data augmentation, the images with a size of 512×512 were extracted randomly, in addition to the application of a random horizontal and vertical flipping. Due to the limited physical memory of GPU cards, the batch size was set to be 16, and synchronized batch normalization across the GPU cards was adopted during the training. The total number of training iterations was 15,000. Similar to (Jingdong Wang et al., 2021; Zhao et al., 2018, 2017), by applying the random data augmentation and the batch normalization, all the networks used in this study are considered resistant to overfitting.

5.2.5. Comparisons

To check the performance (i.e., segmentation accuracy and efficiency) of using LC-Net, the proposed approach was compared with three commonly used approaches.

The first one is the Direct Feeding (DF) approach, which is to modify the first convolutional layer (typically 3 kernel channels) in the networks (i.e., ResNet50, HRNet-w18, and Swin-tiny) so that the number of kernel channels is identical to the number of bands of the input image. In the case of RIT-18 (6 bands), the modified first convolutional layer has 6 kernel channels, the initial parameters of which are randomly allocated.

The second approach is Principal Component Analysis (PCA). More specifically, the singular value decomposition method is adopted for extracting the first three principal components. The extracted 3-band images are used as the network input.

Chapter 5: Automatic feature extraction

The third approach is a Supervised Grid Search (SGS), in which the optimal combination of the 3-band is determined by trialling all possible combinations. As the most fundamental band selection method, it ensures the selection of the optimal band combination. To facilitate fairer comparisons between those approaches, the parameters in the first convolutional layer of the pre-trained backbones were randomly reset in the DF approach, the SGS approach, and the LC-Net approach.

Further to the aforementioned three approaches, two alternative means of band compressions are also investigated in this study, which is detailed in the next two paragraphs.

The band compression using adjacent bands (i.e., Figure 5-2) in LC-Net is based on the assumption that the neighbouring bands are often more similar to each other and, therefore, less effective information may be lost in this way. To test whether this hypothesis is necessary, a compression of non-adjacent bands (e.g., bands 1 and 3, bands 2 and 5, and bands 4 and 6) in the same means was also tested to explore the differences.

In addition to the compressions of two bands into one to form a 3-band image using RIT-18, another test is considered, in which all bands are compressed to form each band of the 3-band input image (i.e., all bands are compressed three times, and a new band is produced each time). This is referred to as the Linear Combination of All Bands (LCAB). Since the number of output bands is not constrained by the number of input bands, the LCAB approach is more flexible than LC-Net. However, since each band

of the output image is a linear combination of all input bands, LCAB potentially suffers from a greater influence of the mixed spectral information received.

5.3. Results

The results of the PCA, DF, and LC-Net approaches are summarized in Table 5-4. The segmentation accuracy using the PCA approach was considerably lower than those of the other two approaches and, therefore, was not analysed in detail. It was observed that the use of LC-Net brought consistent improvements in the final segmentation performance. In comparison to those using the DF approach, an average improvement of 12.1% in overall accuracy (OA) was achieved by adding LC-Net. Meanwhile, for the more critical accuracy metric, i.e., the mean accuracy (MA), the use of LC-Net led to an average improvement of 14.0% compared to those obtained using the DF approach. The processing time introduced by the extra learnable parameters (six weights in these cases) was found to be negligible.

Table 5-4: Summary of the key performance of PCA, DF, LC-NET, and CoinNet (%).

Class	ResNet50			HRNet-w18			Swin-Tiny			CoinNet
	PCA	DF	LC-Net	PCA	DF	LC-Net	PCA	DF	LC-Net	-
Road Markings	0.0	33.6	40.6	0.0	3.2	73.3	0.0	13.3	63.1	85.1
Tree	72.7	91.5	90.1	8.8	78.1	85.4	80.2	82.1	88.3	77.6
Building	13.3	54.4	69.2	0.0	58.0	62.2	0.0	61.3	65.0	52.3
Vehicle	0.0	50.5	53.2	0.0	54.6	55.7	0.0	38.5	49.5	59.8
Person	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lifeguard Chair	0.0	82.1	66.3	0.0	32.2	79.5	0.0	39.1	99.5	0.0
Picnic Table	0.0	4.0	9.9	0.0	22.6	22.7	0.0	32.4	12.6	0.0
Orange Pad	0.0	0.0	0.0	0.0	95.8	0.0	0.0	83.1	0.0	0.0
Buoy	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
Rocks	1.2	84.3	93.3	5.3	73.1	91.5	4.0	88.0	90.3	84.8
Low Vegetation	0.0	1.7	4.9	0.6	11.4	5.6	0.0	2.9	19.0	4.1
Grass/Lawn	86.2	95.2	95.4	97.4	97.5	95.1	84.1	94.9	95.5	96.7
Sand/Beach	92.0	10.0	93.4	86.2	94.0	94.0	96.3	76.1	95.9	92.1
Water/Lake	58.2	96.9	97.6	89.3	95.9	98.1	93.4	98.8	94.3	98.4
Water/Pond	13.2	14.2	95.9	0.0	7.0	98.0	43.0	63.3	98.2	92.7
Asphalt	77.6	51.0	91.0	72.9	42.7	92.9	78.6	53.4	90.9	90.4
Overall Accuracy	73.8	68.7	90.7	69.5	82.4	90.4	81.2	84.6	91.0	88.8
Mean Accuracy	25.9	41.8	56.3	22.5	44.1	59.6	30.0	48.0	60.1	52.1

Chapter 5: Automatic feature extraction

It is worth mentioning that the best semantic segmentation results obtained in this study (i.e., Swin-tiny+ LC-Net) outperformed the traditional machine learning and deep learning approaches reported in the literature (Kemker et al., 2017; Pan et al., 2019; Saxena et al., 2020). For example, the best results (those fine-tuned without additional data) so far on RIT-18 were achieved by CoinNet (Pan et al., 2019), which were exceeded by 2.2% and 8.0% in terms of OA and MA, respectively, using the proposed approach.

To check the effectiveness of the results using LC-Net, the accuracy of the segmentation results using a trial selection of all possible combinations (20 in total) of the three bands was also investigated in this study. Table 5-5 shows the segmentation accuracies of the networks used with different combinations of the three input bands.

Table 5-5: Comparison of accuracies and computational cost using PCA, DF, LC-Net, and SGS (%). The numbers shown in the “Input Method” column for SGS indicate which three bands were used as the input images to the subsequent networks.

Input Method	ResNet50			HRNet-w18			Swin-Tiny			
	OA	MA	Training hours	OA	MA	Training hours	OA	MA	Training hours	
PCA	73.8	25.9	3.6	69.5	22.5	3.9	81.2	30	4	
DF	68.7	41.8	3.6	82.4	44.1	3.9	84.6	48	4	
LC-Net	90.7	56.3	3.6	90.4	59.6	3.9	91.0	60.1	4	
SGS	123	72.6	39.7	72.3	43.0		72.8	43.5		
	124	88.7	53.2	88.4	56.5		88.9	57.0		
	125	86.7	51.4	86.4	54.7		87.0	55.2		
	126	88.6	53.1	88.3	56.4		88.9	56.9		
	134	85.4	53.2	85.1	56.5		85.7	57.0		
	135	78.8	43.3	78.4	46.6		79.1	47.1		
	136	74.5	43.8	74.1	47.1		74.7	47.6		
	145	88.8	51.2	88.5	54.5		89.1	55.0		
	146	86.4	52.8	86.1	56.1		86.6	56.6		
	156	89.3	54.4	89.8	57.7		89.6	58.2		
	234	86.3	51.9	72	85.9	55.2	77	86.5	55.7	80
	235	78.5	51.5		78.2	54.8		78.7	55.3	
	236	81.0	44.1		80.7	47.3		81.3	47.8	
	245	87.8	53.3		87.4	56.6		88.1	57.1	
	246	85.4	49.6		85.1	52.9		85.7	53.4	
	256	88.2	51.4		87.9	54.7		88.5	55.2	
	345	89.3	50.7		88.9	54.0		89.5	54.5	
346	89.1	53.7		88.8	57.0		89.4	57.5		
356	89.4	55.7		89.0	57.8		89.8	59.3		
456	47.5	33.5		47.2	36.8		47.8	37.3		

It was observed that the accuracies achieved under each combination varied. For the three networks considered, the accuracy of the most accurate combination was on average 1.30% lower and 1.07% lower than those of LC-Net for the OA metric and the MA metric, respectively. In addition to this, LC-Net requires much less time than the band selection method because it only needs to be trained once. A visual comparison between DF, SGS, and LC-Net for Swin-tiny is shown in Figure 5-5. It is observed that the results of DF are notably worse than the other two methods.

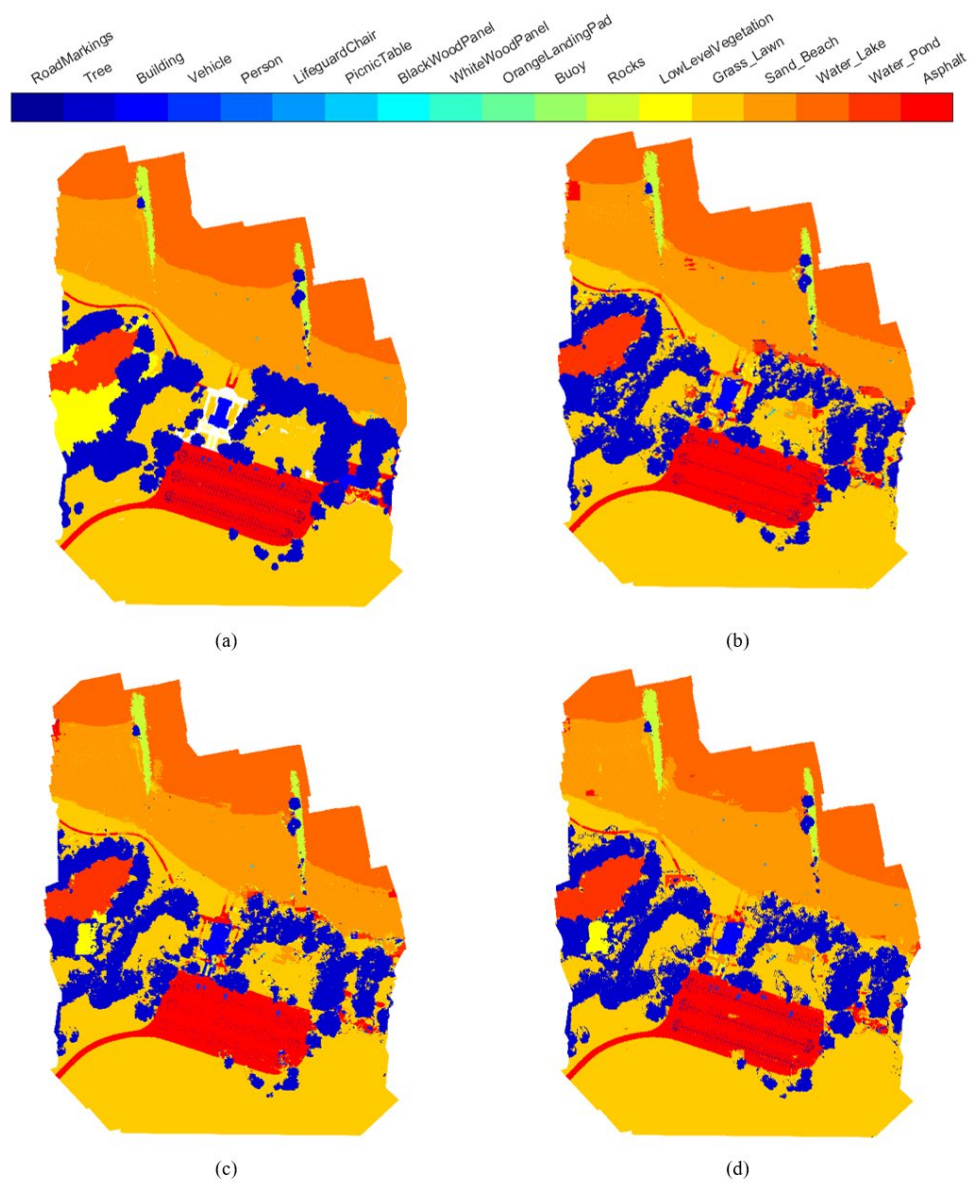


Figure 5-5: RIT-18 segmentation results for Swin-tiny on the test image: (a) ground truth, (b) segmentation map using the DF approach, (c) segmentation map using the SGS approach, (d) segmentation map using LC-Net.

Chapter 5: Automatic feature extraction

The aforementioned results are based on a particular compression of two adjacent bands into one (i.e., one from bands 1 and 2, one from bands 3 and 4, and one from bands 5 and 6). No adjacent bands were also considered for the band compression. Table 5-6 shows the comparisons between the performances of LC-Net using adjacent bands or non-adjacent bands. Very small differences were observed in the overall accuracies. In addition, the segmentation accuracies of LCAB are shown for comparison to LC-Net in Table 5-6. Significant degradation in the segmentation accuracy was observed.

Table 5-6: Performance of LC-Net and LCAB (%).

Methods	The Formation of 3 Input Bands to the Networks from the Original 6 Bands	ResNet50		HRNet-w18		Swin-Tiny	
		OA	MA	OA	MA	OA	MA
LC-Net	(12), (34), (56)	90.7	56.3	90.4	59.6	91.0	60.1
LC-Net (non-adjacent)	(13), (25), (46)	90.8	56.2	90.2	59.4	91.2	59.0
LCAB	(1-6), (1-6), (1-6)	84.1	53.4	87.1	54.2	88.6	55.0

5.4. Discussion

The segmentation results of ResNet50+ LC-Net, HRNet-w18+ LC-Net, and Swin-tiny+ LC-Net are shown in Figure 5-6. Compared to the ground truth (Figure 5-6.a), it was seen that the majority of the pixels were correctly labelled using any of those networks. To understand the likely causes for the mislabelled pixels by the networks, the spectral information (i.e., the RGB images of bands 1–3 and the pseudo colour images of bands 4–6) of RIT-18 is shown in Figure 5-7. After the visual comparisons between Figure 5-6 and Figure 5-7, it was found that segmentation errors appeared mainly in three types of areas: the object edges, shaded areas, and changes in terrain. In contrast to extensive attention received in the field of semantic segmentation (Marmanis et al., 2018; Yuan et al., 2020) for improving the segmentation accuracy at the object edges, the latter two error sources have not received much attention. To address these issues, it is necessary to identify the causes of the segmentation errors in

those regions. A relatively large mislabelled local area was marked with a black dashed box in both Figure 5-6 and Figure 5-7, where a continuous area of low-level vegetation was segmented as other classes (mainly tree or grass/lawn). As shown in Figure 5-7, the area is a valley where the surrounding trees cast shadows, which resulted in abrupt changes in the spectral characteristics of the low-level vegetation. Another common reason worth considering for the low segmentation accuracy of a particular class in the test set is the under-representation of that class in the training set. For RIT-18, the proportion of low-level vegetation in the training data ranks seventh among all classes. The eighth, ninth, and tenth classes in the training data are rocks, buildings, and road signs, with 1.27%, 0.45%, and 0.39%, respectively. Their segmentation accuracy (Swin-tiny + LC-Net) on the test set was 90.3%, 65.0%, and 63.1%, respectively, which are all much higher than that of low-level vegetation. Therefore, the lack of training data may not be the reason for the low segmentation accuracy of low-level vegetation, and the abrupt changes in the spectral characteristics are likely to be the cause of the mis-segmentation. Similarly, the shadows of trees on the sandy beach surface seemed to cause a mix of mislabelled pixels as enveloped by a solid purple box. Therefore, future studies may consider pre-identifying areas with shadows and topographic changes and treating them separately. This may require additional labelling of such areas or additional data, such as digital elevation models and solar information about the target area.

Chapter 5: Automatic feature extraction

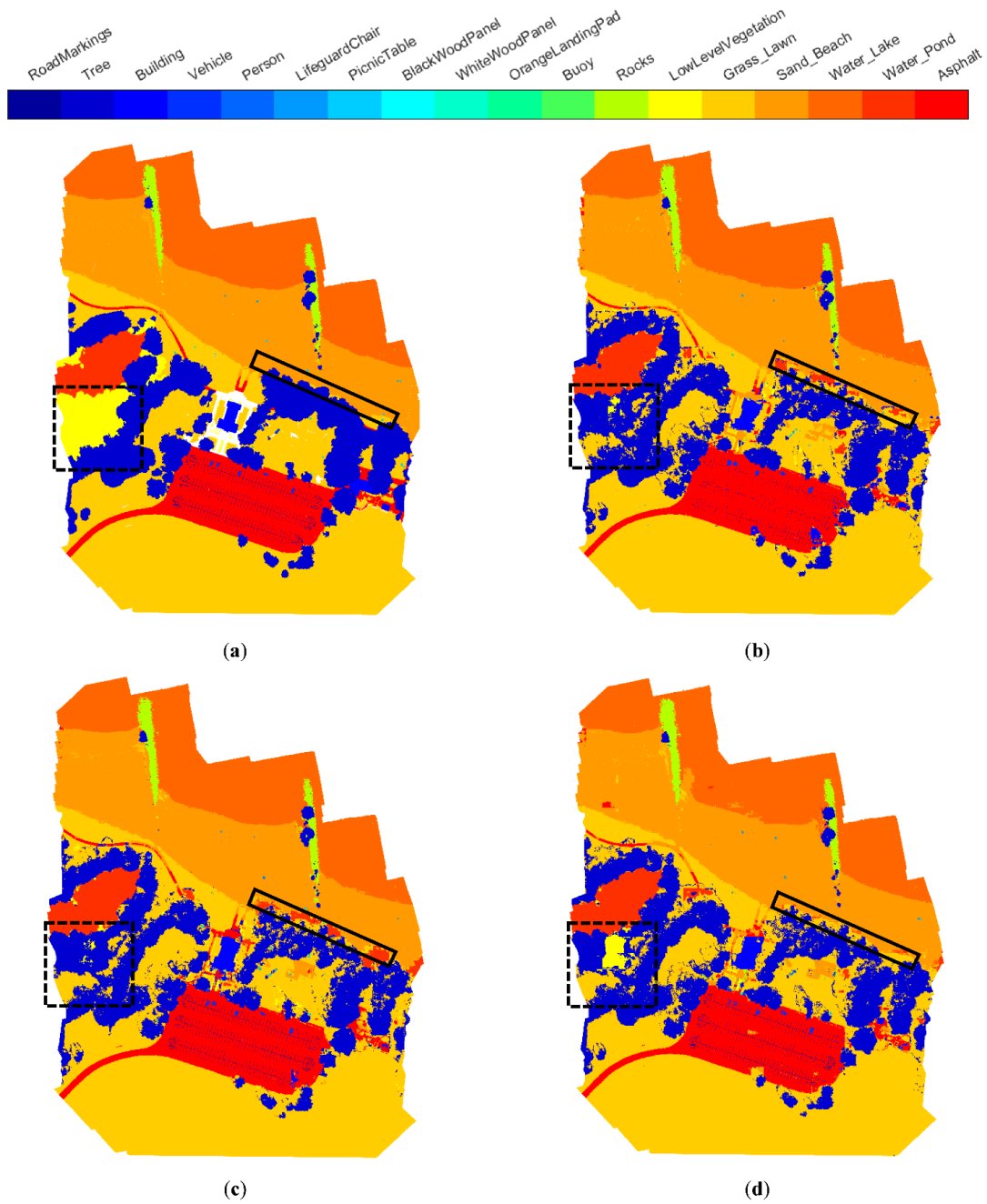
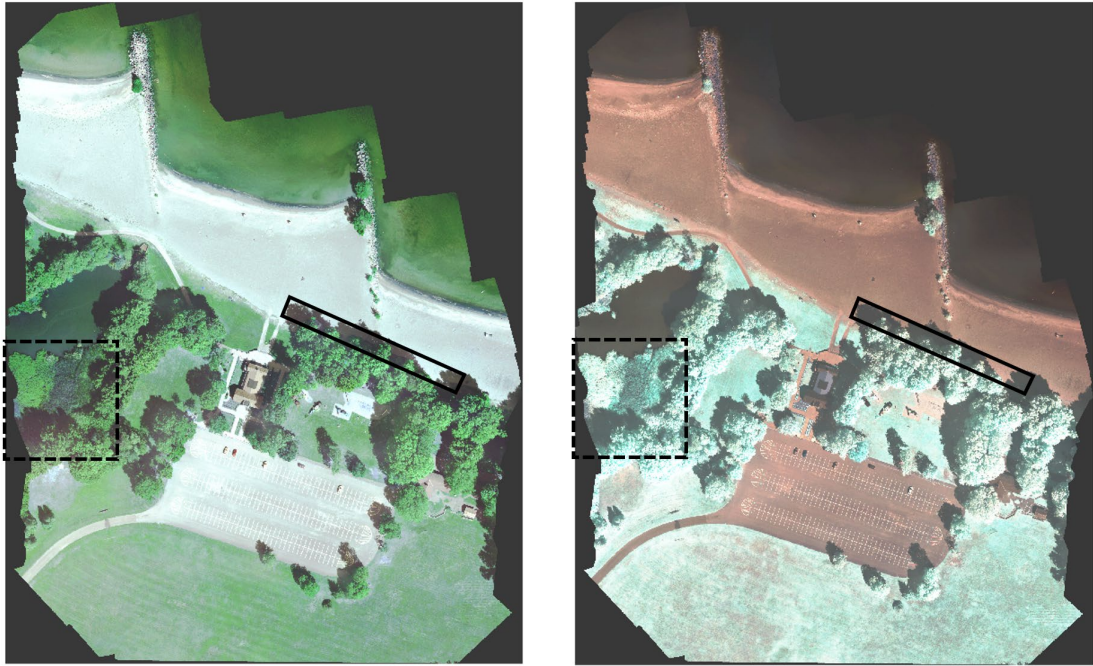


Figure 5-6: RIT-18 segmentation results of the test image: (a) ground truth, (b) segmentation map from ResNet50+ LC-Net, (c) segmentation map from HRNet-w18+ LC-Net, (d) segmentation map from Swin-tiny+ LC-Net.



(a)

(b)

Figure 5-7: (a) RGB image (bands 3, 2, and 1) of the RIT-18 test image; (b) pseudo color image of bands 4–6 of the RIT-18 test image.

While the results demonstrate that adding LC-Net to the networks can improve the overall segmentation performance, not every segmented class can benefit from LC-Net, as shown in Figure 5-8. The likely reason for this is that LC-Net forces the neural networks to perform a relatively aggressive band compression in the initial stage, which inevitably causes a certain amount of information loss. Since the weights of LC-Net were randomly initialized and the weights in the backbone were also different, the lost information may, by chance, include the one that is important for a specific class. This conjecture can be inferred to some extent by the information in Table 5-7, where the final weights of LC-Net associated with the three network structures were recorded and showed no specific pattern. Therefore, future studies may focus on finding more effective band compression methods. For example, more complex (e.g., a 3×3 convolutional kernel), nonlinear (e.g., adding the activation function after the LC-Net), and multilayer band compression methods might be able to compress more effective information into the same number of bands.

LC-Net compresses the number of bands in RIT-18 to half of its original number. However, the optimal level of the band compression (with respect to the number of bands) may not be exactly half of the original number. As such, future research may focus on how to set the degree of the band compression as a learnable variable to improve the network performance further.

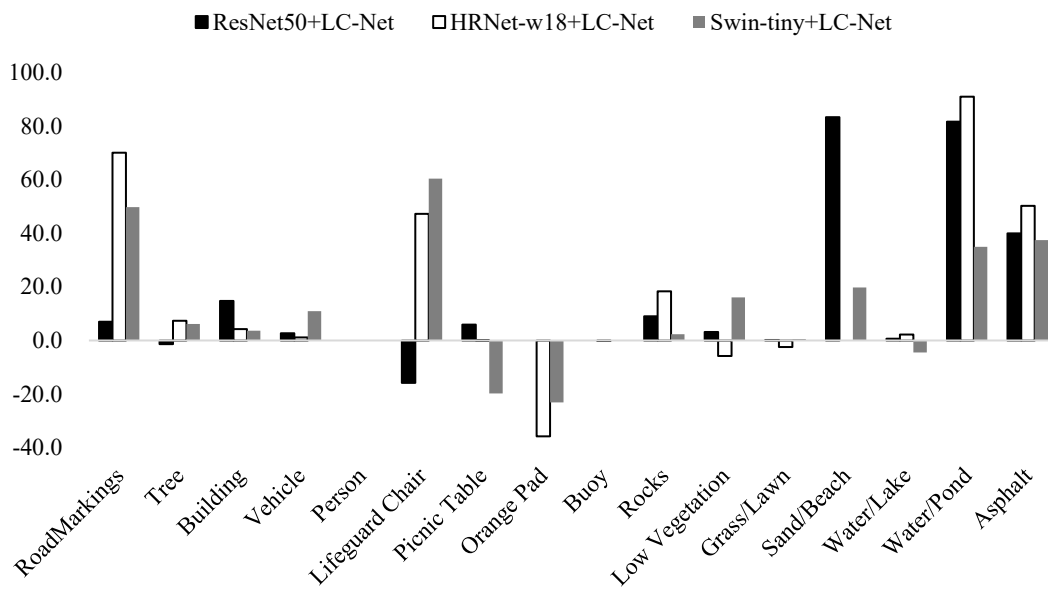


Figure 5-8: The effect of the addition of LC-Net on the segmentation accuracy of different classes against the DF approach (%).

Table 5-7: Final weights in LC-Net used in the three networks considered.

	Final Weights in LC-Net		
	ResNet50 + LC-Net	HRNet-w18 + LC-Net	Swin-tiny + LC-Net
Band 1	0.048	-0.066	0.253
Band 2	-0.893	0.201	-0.159
Band 3	0.927	-0.464	0.220
Band 4	1.100	-0.722	0.470
Band 5	-0.931	-0.555	0.530
Band 6	0.460	-0.417	0.348


5.5. Summary

Neural networks have been proven to be powerful tools for semantic segmentation with respect to RGB images. However, the application of those networks to multispectral images that have many more bands usually requires a tedious and time-consuming trial band selection process. To solve this problem, a simple LC-Net is proposed in this study to automatically reduce the number of bands to fit that required by networks via an embedded learning process at almost no cost. It was found that the accuracies (in terms of OA and MA) of the semantic segmentation on RIT-18 were significantly improved (12.1% and 14.03%, respectively) when LC-Net was added to the networks considered. In comparison to the SGS of the optimal combination of three bands (from 20 possible combinations for RIT-18), the segmentation accuracies using LC-Net were found to be slightly higher than those of the optimal combination of three bands obtained through the time-consuming exhaustive selection. Meanwhile, the computational cost of LC-Net was much less than the trial selection process.

Chapter 6: Automatic feature selection

This section is based on a journal article being prepared for submission. I declare that this paper are purely my own work as the first authorship. The ideas and experiments were developed by myself, discussed and approved by my supervisors (Lei Fan is the corresponding author).

Yuanzhi Cai 

Lei Fa 

Note: This chapter develops a task-adaptive feature selection method. The performance of the proposed method is compared with various dimension reduction methods including the one proposed in Chapter 5.

6.1. Introduction

Multichannel image refers to image data containing more than three channels (typically less than ten channels). It is also known as multispectral image when only spectral channels are included. Semantic segmentation of multichannel image is the basis for many remote sensing applications, such as cloud detection (S. Chen et al., 2022; L. Peng et al., 2022; Zhang et al., 2022), land use/ land cover classification (Aryal et al., 2022; Song et al., 2021) and forest monitoring (Anees and Aryal, 2014; Rajbhandari et al., 2019).

Although additional channels provide more information for semantic segmentation, it is not desirable to use extra channels than are necessary for many practical applications. Firstly, there is always a cost associated with the acquisition of additional channels. For example, the quest for additional spectral channel(s) by multispectral sensors not only entails higher financial costs, but often leads to compromises in spatial resolutions. Secondly, higher segmentation accuracy can be achieved by excluding unnecessary channels that are impacted by noise and false information, and excessive less relevant channels (Cai et al., 2022a, 2021b, 2020; Chang, 2022; Chang and Ma, 2022; Lo et al., 2023; Q. Wang et al., 2020).

For some widely recognised classes of objects (e.g., clouds and water bodies), their preferred channel(s) for segmentation have been studied for decades. However, these studies are insufficient to meet the challenges in demand for semantic segmentation. On one hand, it is often required to segment multiple classes simultaneously. On the other hand, more refined object classes and new data are constantly introduced for

segmentation. Therefore, an automated channel selection method that can adapt to the needs of different tasks is highly desirable.

Currently, the main type of imagery data investigated in the field of channel selection is hyperspectral images. In this context, a channel selection task often needs to select dozen(s) of channels out of one to two hundred candidate channels. Due to the fine spectral resolution of a hyperspectral image, its neighbouring channels often contain features that are very similar to each other. Consequently, it is possible to achieve semantic segmentation accuracy in the satisfactory range using hyperspectral images with removal of redundant channels and can sometimes achieve even higher accuracy than using the original hyperspectral channels (S. Huang et al., 2022; Jia et al., 2022; Sun and Du, 2019; Q. Wang et al., 2018; Zhai et al., 2019). Coupled with the fact that the removal of redundant channels does not rely on labelling information, unsupervised methods have become the dominant research direction for channel selection. In unsupervised methods, channel selection is typically performed by using one or a combination of ranking, clustering and search strategies to optimise various criteria, such as entropy (Gong et al., 2016; Sotoca et al., 2007; H. Sun et al., 2022), variants of PCA (Chang et al., 1999; Lee et al., 1990; Roger, 1994), and minimising similarity (Datta et al., 2015; Du and Yang, 2008; Jia et al., 2016; Rodriguez and Laio, 2014). These methods select channels based on the characteristics of the input data themselves and do not take into account the preferences of classes to be segmented. In other words, they are not task-adaptive.

There are also supervised channel selection methods that make use of label information. The major difference between supervised and unsupervised methods lies in their

Chapter 6: Automatic feature selection

optimisation criteria. More specifically, additional criteria such as prediction accuracy (Archibald and Fann, 2007), mutual information criteria (Feng et al., 2015, 2014; Guo et al., 2006; Peng et al., 2005) and Fisher score (Gu et al., 2012) became available to supervised methods due to the presence of label information. Although existing supervised methods are task-adaptive, they often suffer from the following common deficiencies. Firstly, most methods use criteria other than prediction accuracy, which diverts the optimisation process from obtaining the highest prediction accuracy and is likely to lead to selecting channel combinations with sub-optimal accuracy. Secondly, a majority (sometimes all) of excluded channels in most channel selection strategies were determined by evaluating individual channels. Since “the m best features are not the best m features” (Cover, 1974; Jain et al., 2000; Peng et al., 2005), using such selection strategies is not the best solution. Finally, most of the existing supervised methods require training multiple classifiers and/or training classifier(s) multiple times, which leads to low computational efficiency.

As an attempt to resolve the aforementioned issues associated with existing supervised methods, a novel one-shot task-adaptive (OSTA) channel selection method is proposed in this study. OSTA has the following characteristics: a) directly optimising for segmentation accuracy, b) considering channel interactions (i.e., no channel is excluded individually), c) integrating channel selection and network fine-tuning within a relatively efficient and predictable timeframe. All these characteristics are realised by formulating the channel selection as a pruning process for a supernet. As such OSTA consists of three stages, namely the training stage of the supernet, the pruning stage, and the fine-tuning stage. In this study, the effectiveness and the efficiency of OSTA are tested using four datasets, including an eight-band cloud detection dataset named L7

Irish (Hughes and Hayes, 2014), a ten-band cloud detection dataset named L8 Biome (Foga et al., 2017), a six-band very-high resolution dataset named RIT-18 (Kemker et al., 2018) and an eight-channel image dataset that is transformed from a terrestrial laser scanning (TLS) point cloud dataset named Semantic3D (Hackel et al., 2017). The major contributions of this chapter are:

- (1). Development of a novel channel selection method (i.e., OSTA) to overcome the following main issues with the existing methods. They are not dedicated to optimising segmentation accuracy, do not take full account of channel interactions and require repeated training of classifier(s).
- (2). Comprehensive evaluation of channel selection methods by exhaustive testing of the accuracy performance of 3-channel combinations on four benchmark datasets.
- (3). In addition to channel selection, this study also leads to some interesting findings:
 - a) there are channel combinations that are robust to the network initialisation;
 - b) the coastal aerosol band has been neglected in the past research for cloud detection but turns out to be an importance channel for cloud detection according to this study;
 - c) training with channel combinations other than the selected one can improve the semantic segmentation accuracy.

6.2. Methodology

6.2.1. OSTA

The overall training strategy of OSTA is shown in Figure 6-1. It starts with a supernet training stage that accounts for 15% of the total training iterations. At this first stage, the objective of the supernet training is to perform semantic segmentation using any combination of channels, and during the training the learning rate increases linearly from zero to the target value. Subsequently, the channel combinations used for training

Chapter 6: Automatic feature selection

are progressively pruned according to their validation accuracies until only one combination remains. This pruning process accounts for 35% of the total training iterations. In the final stage, the remaining 50% of the total training iterations are used to fine-tune the semantic segmentation network (SSN) for the selected channel combination. The poly learning rate policy is used in the latter two stages.

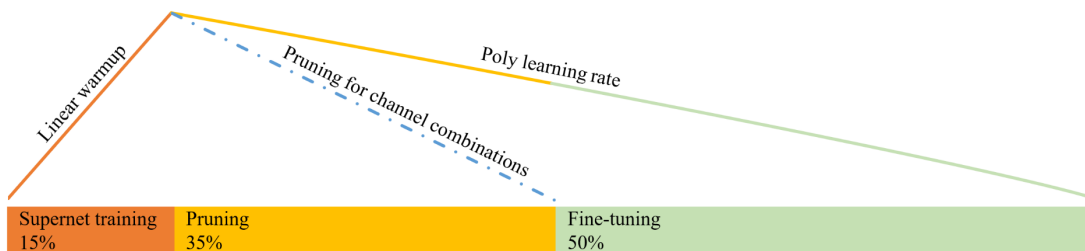


Figure 6-1: Percentages of training iterations of the three stages in OSTA and the learning rate schedule.

The essence of OSTA is to select the channel combination with the best validation accuracy. This can ensure two points: a) a good generalisation performance (i.e., accuracy) of the selected channel combination, b) the combination of “best m features” is taken into account as none of the channels is removed individually. More detailed explanations of these three stages are as follows.

6.2.1.1. Supernet training

The supernet has two parts: an input layer that is pruned later, which treats each channel combination as one of its input channels (ICs), and a subsequent weight-sharing SSN which takes individual inputs from each input channel (IC) for semantic segmentation. The key to a successful supernet training is to train each candidate channel (also known as path or branch) fairly, to which many studies have been devoted (Chu et al., 2021, 2020; T. Huang et al., 2022; C. Peng et al., 2022; Su et al., 2021). Fortunately, fair training of candidate channels can be easily achieved in OSTA

by randomly sampling ICs during the training. Therefore, pruning from a supernet is inherently suitable for implementing channel selection.

6.2.1.2. Pruning

To minimise the additional amount of computation required for pruning, a discrete forward-only pruning strategy that is based on validation accuracy is proposed. More specifically, the training of the SSN is paused uniformly for n times during the pruning stage, where n is the number of ICs to be pruned. At each pause, the segmentation accuracy of the SSN on the validation set is tested for each remaining IC. The IC with the worst validation accuracy is removed in subsequent training. By repeating the training of the SSN and the pruning of an IC in sequence, the selected channel combination (SCC) (i.e., selected IC) is obtained at the end of the pruning stage. Since only one layer in the supernet needs to be pruned, the validation accuracy can serve as the equivalent of gradients in OSTA.

6.2.1.3. Fine-tuning

This stage is to fine-tune the SSN on the SCC. Since validation accuracy is not required for this stage, the validation set can be merged into the training set to fine-tune the SSN.

6.2.2. Establishment of benchmarks

6.2.2.1. Benchmark data

Four semantic segmentation datasets are used in this study, including L7 Irish (Hughes and Hayes, 2014), L8 Biome (Foga et al., 2017), RIT-18 (Kemker et al., 2018) and Semantic3D (Hackel et al., 2017), and they are used as follows.

Chapter 6: Automatic feature selection

L7 Irish and L8 Biome are processed in the same way as both datasets contain multispectral satellite images and are labelled with the same 4 classes (i.e., shadow, clear sky, thin cloud and thick cloud). Each dataset is partitioned evenly into a training set (75%) and a test set (25%). More specifically, the last of every four images is used for testing, according to the order of the images on their official website. As for the label configurations, two commonly used ones (L. Peng et al., 2022; Yang et al., 2022) are tested in this study. In the first configuration, classes of shadow and clear sky are merged into a single class of background. Datasets with this configuration (i.e., background, thin cloud and thick cloud) are denoted as L7 Irish 3C and L8 Biome 3C. The second configuration is established upon the first one, where classes of thin cloud and thick cloud are further merged into a single class of cloud. Datasets with this configuration (i.e., background and cloud) are denoted as L7 Irish 2C and L8 Biome 2C. For L7 Irish and L8 Biome, the bands used in this study are blue (B), green (G), red (R), near-infrared (NIR), short-wave infrared (SWIR), thermal low gain (TLG), thermal high gain (THG) and mid-infrared (MIR), and coastal aerosol (CA), B, G, R, NIR, SWIR1, SWIR2, cirrus (C), thermal 1 (T1) and thermal 2 (T2), respectively.

RIT-18 is a six-band (B, G, R, NIR1, NIR2, NIR3) image dataset having 18 labelled classes. The original validation set of RIT-18 is used as the test set in this study.

Semantic3D is originally a point cloud dataset having 8 labelled classes, which is transformed into an eight-channel image dataset in this study. The original training set of Semantic3D is partitioned into a training set and a test set as in the previous study (Cai et al., 2021b). Each point cloud data is transformed into an eight-channel image using spherical projection and the enhancement method proposed in (Cai et al., 2022a).

Specifically, the following feature channels are included: R, G, B, intensity (I), z-coordinate image (Z), depth image (D), enhanced z-coordinate image (Ze), and enhanced depth image (De).

The key characteristics of benchmark data used are summarised in Table 6-1. Since the original images are too large to be fed into the network, they are cropped into smaller images without overlaps. The crop sizes are shown in Table 6-1. To calculate the validation accuracy, the training set is partitioned into a sub-training set and a sub-validation set with a similar class distribution in the first two stages of OSTA.

Table 6-1: Summary of benchmark data used.

Dataset	No. classes	No. channels	No. and size of image(s)		Crop size	No. cropped images		
			Training	Testing		Sub-training	Sub-validation	Testing
L7 Irish 2C	2	8	155, $\approx 8000 \times 8000$	51, $\approx 8000 \times 8000$	1024 \times 1024	10012	100	3246
L7 Irish 3C	3	8	155, $\approx 8000 \times 8000$	51, $\approx 8000 \times 8000$	1024 \times 1024	10012	100	3246
L8 Biome 2C	2	10	72, $\approx 8000 \times 8000$	24, $\approx 8000 \times 8000$	1024 \times 1024	5041	50	1555
L8 Biome 3C	3	10	72, $\approx 8000 \times 8000$	24, $\approx 8000 \times 8000$	1024 \times 1024	5041	50	1555
RIT-18	18	6	1, 9393 \times 5642	1, 8833 \times 6918	1024 \times 1024	48	12	63
Semantic3D	8	8	10, 3600 \times 7200	5, 3600 \times 7200	1024 \times 2048	128	32	80

6.2.2.2. Benchmark methods

There are three types of benchmark methods used in this study. They are briefly introduced in this section.

The first one is the direct feeding (DF), which uses the original multichannel image as the input data for the SSN. DF is considered as a baseline method.

The second type of methods are channel selection methods, which select the “best channel combination” as input data for the SSN. Different channel selection methods will result in different “best channel combination” depending on the selection criteria and selection strategy used. To provide the most comprehensive comparison, this study

Chapter 6: Automatic feature selection

exhaustively tested all possible channel combinations using supervised grid search (SGS). Although the SGS test results have implicitly included all the segmentation accuracies that can be achieved using existing channel selection methods, this study also explicitly compared three recent channel selection methods, including BS-Nets (Cai et al., 2020), ONR (Q. Wang et al., 2020) and DARecNet-BS (Roy et al., 2021a).

The third type of methods are feature extraction methods, which use the extracted new feature channels as input data for the SSN. Although feature extraction methods have been criticized for not preserving the original channels, it is still interesting to test the segmentation accuracy they can obtain. Two feature extraction methods are used for benchmarking in this study, namely principal component analysis (PCA) and LC-Net (Cai et al., 2022b). The former is a classical feature extraction method, while the latter is a task-adaptive feature extraction method that performs a linear compression of the input data dimensions by learning.

6.2.2.3. Training setting

For a fair comparison, the same SSN is used for all tested methods. Specifically, the backbone and decoder of SSN are ConvNeXt-T (Z. Liu et al., 2022) and UperNet (Xiao et al., 2018), respectively, given their proven performance (Cai et al., 2023). Unless otherwise specified, ConvNeXt-T is initialised with the ImageNet pre-trained weights. The training strategy used is the same as the original implementation (Z. Liu et al., 2022) except for the following points. The total number of training iterations is set as 10k. The sizes of the input patches used for the network training are set to 1/4 of the crop size in Table 6-1, i.e., 512×512 for L7 Irish, L8 Biome and RIT-18, and 512×1024 for Semantic3D. It is worth noting that the segmentation accuracy can significantly be improved by adjusting the histogram of the benchmark datasets to

match the distribution of the pre-training dataset, which is implemented in this study. Finally, the accuracy evaluation for the Semantic3D data is based on the accuracy of the image results.

Except for DF, both channel selection and feature extraction methods require a predetermined dimension reduction target. In this study, the dimension reduction target of three was tested, i.e., the number of channels of the original multichannel image was reduced to three. The rationale for testing this setup is as follows. Firstly, for two of the benchmark data used (i.e., RIT-18 and Semantic3D), previous studies (Cai et al., 2022a, 2022b) showed that a better segmentation accuracy was achieved using a properly selected three channels than that using more channels. Secondly, higher accuracy can be obtained when fine-tuning with data similar to the pre-training data. Coupled with the fact that most state-of-the-art computer vision models are pre-trained on RGB datasets, it is often desirable to reduce the dimension of multichannel images to 3 in practical applications. For example, compressing four-channel images (i.e., red, green, blue and thermal) into three-channel images (i.e., $0.5 \times (\text{red} + \text{thermal})$, $0.5 \times (\text{green} + \text{thermal})$, $0.5 \times (\text{blue} + \text{thermal})$) has become a popular fusion method in the field of crack detection (F. Liu et al., 2022; Pozzer et al., 2022; Jun Yang et al., 2019).

6.2.3. Evaluation metrics

In total, six evaluation metrics are used in this study to measure the performance of OSTA, as shown in Table 6-2. Mean accuracy (mA) and mean intersection over union (mIoU) are used as the major accuracy metrics for RIT-18 and other three datasets (i.e., L7 Irish, L8 Biome, and Semantic3D), respectively. In addition, two novel accuracy metrics are proposed in this study, namely combination accuracy percentile (CAP) and

Chapter 6: Automatic feature selection

difference in combination accuracy (DCA). CAP measures the ranking of OSTA test accuracy against the accuracies of all combinations tested in SGS. This metric represents the percentage of SGS combinations that their test accuracies are exceeded by OSTA. DCA measures the difference between the test accuracy of “SCC in OSTA” (SCC_{OSTA}) and the test accuracy of SCC_{OSTA} in SGS. This metric describes the impact of the first two training stages of OSTA on the test accuracy of SCC_{OSTA} . This indicator can help better understand the role of the different stages of OSTA

The measure of efficiency is based on the ratio of consumptions (time and memory) with OSTA to consumptions without OSTA (i.e., direct training SSN). This allows the evaluation metrics to show the additional consumption caused by channel selection in percentage.

Table 6-2: Evaluation metrics used in OSTA.

Accuracy	Mean accuracy (mA)	$\frac{1}{C} \sum_{c=1}^C \frac{TP_c}{\text{Number of points}_c}$
	Mean intersection over union (mIoU)	$\frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FP_c + FN_c}$
	Combination accuracy percentile (CAP)	Percentage of SGS-derived combinations with test accuracies lower than OSTA accuracy. For an OSTA accuracy that is not identical to the SGS accuracy, its CAP is obtained by linear interpolation based on the two nearest larger and smaller SGS accuracies.
	Difference in combination accuracy (DAC)	Accuracy of SCC_{OSTA} – Accuracy of SCC_{OSTA} in SGS
Efficiency	Ratio of additional time (RAT)	$\frac{\text{Time of training OSTA}}{\text{Time of direct training SSN}} - 100\%$
	Ratio of additional memory (RAM)	$\frac{\text{Memory of training OSTA}}{\text{Memory of direct training SSN}} - 100\%$

Where TP_c , FP_c , and FN_c represent the true positive, false positive and false negatives of class c , respectively.

6.3. Experiments and results

6.3.1. Semantic segmentation on benchmarks

The qualitative semantic segmentation results are visualised in Figure 6-2, where the horizontal and vertical coordinates show the relative accuracy (CAP) and absolute

accuracy (mIoU/mA), respectively. Each data point regarding the SGS method represents the segmentation accuracy of each three-channel combination tested.

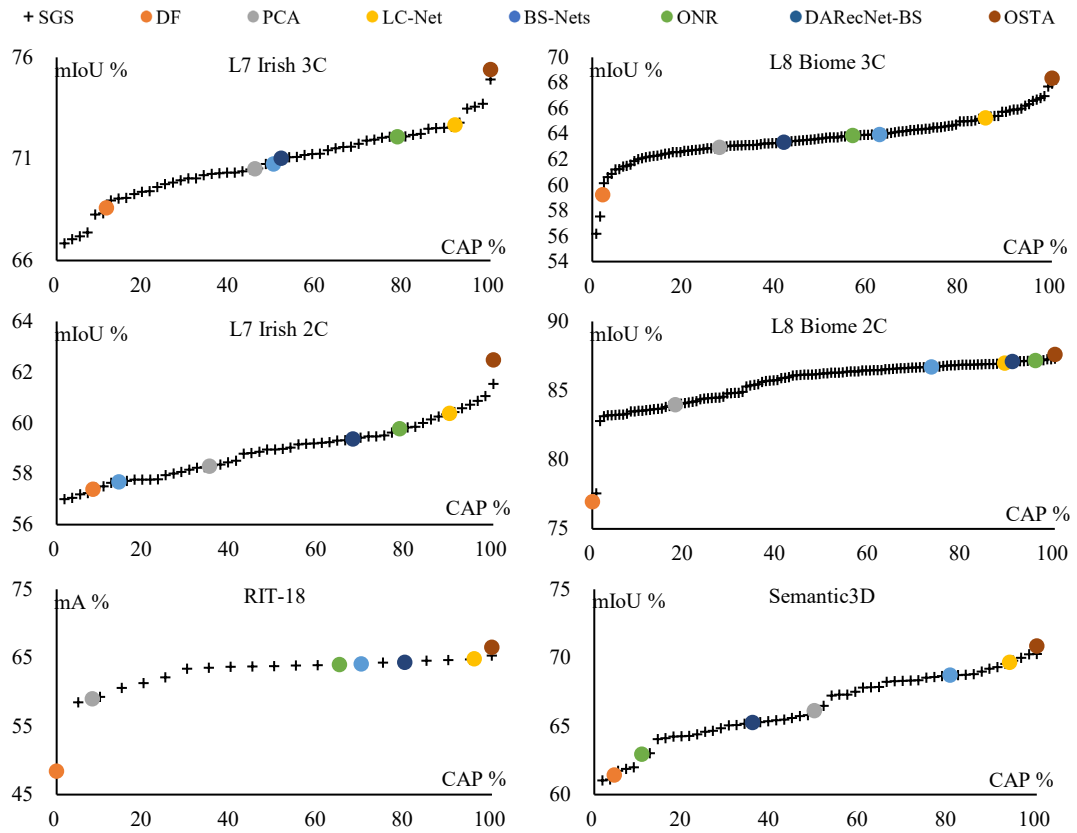


Figure 6-2: Semantic segmentation results on benchmark data.

DF is the worst one among all benchmark methods. For all benchmark data, a randomly selected 3-channel combination would likely achieve higher accuracy than DF. This demonstrates the value of band selection for semantic segmentation of multichannel images. The reason for this is that fine-tuning with data that has the same channel number as the pre-training data can result in higher accuracy, as mentioned in Section 6.2.2.3.

It was noticed that the CAPs of existing dimension reduction methods (i.e., PCA, LC-Net, BS-Net, ONR and DARecNet) were unstable across benchmark data. This indicates that they are short of task adaptability. In contrast, the proposed OSTA

Chapter 6: Automatic feature selection

achieved the highest accuracy (i.e., saturated CAP of 100%) in all tests, which proved its effectiveness. It was surprising that OSTA was able to outperform the highest SGS accuracy. Since existing channel selection methods train the SSN with the SCC alone, the highest SGS accuracy is usually considered to be the upper limit achievable. Investigating the reason why OSTA can exceed this upper limit is one of the focuses of subsequent ablation study (Section 6.3.3).

The quantitative segmentation accuracies are summarised in Table 6-3, together with the index of the SCC and the DCA where applicable. It was found that even for the same object (i.e., cloud), SGS selected different channel combinations for different granularity of annotation (L7 Irish 3C verse L7 Irish 2C and L8 Biome 3C verse L8 Biome 2C). This demonstrates the importance of task adaptive capacity for channel selection methods. In addition, it was noticed that the SCC were different for OSTA and SGS. This observation was further investigated in Section 6.3.3

Table 6-3: Summary of the semantic segmentation results on benchmark data.

		L7 Irish 3C				L8 Biome 3C				RIT-18			
		Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	Index	mA %	CAP %	DCA %
DF		-	57.39	8.33	-	-	59.24	2.21	-	-	48.41	0	-
Feature extraction	PCA	-	58.31	35.04	-	-	62.96	27.64	-	-	58.99	8.21	-
	LC-Net	-	60.38	89.96	-	-	65.27	85.56	-	-	64.86	95.96	-
Channel selection	BS-Nets	42	57.68	14.29	0	30	63.97	62.50	0	4	64.09	70.00	0
	ONR	36	59.78	78.57	0	81	63.89	56.67	0	7	63.99	65.00	0
	DARecNet	32	59.38	67.86	0	82	63.35	41.67	0	3	64.33	80.00	0
	SGS	50	61.54	100.00	0	3	67.94	100.00	0	19	65.32	100.00	0
	OSTA	56	62.49	100.00	+1.42	9	68.38	100.00	+1.69	19	66.53	100.00	+1.21
		L7 Irish 2C				L8 Biome 2C				Semantic3D			
		Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %
DF		-	68.59	11.45	-	-	76.97	0	-	-	61.44	4.47	-
Feature extraction	PCA	-	70.52	45.71	-	-	83.99	17.92	-	-	66.13	49.68	-
	LC-Net	-	72.67	91.84	-	-	87.00	89.17	-	-	69.67	93.79	-
Channel selection	BS-Nets	42	70.76	50.00	0	30	86.72	73.33	0	34	68.72	80.36	0
	ONR	36	72.10	78.57	0	81	87.19	95.83	0	27	62.96	10.71	0
	DARecNet	32	71.04	51.79	0	82	87.12	90.83	0	43	65.26	35.71	0
	SGS	56	74.91	100.00	0	106	87.32	100.00	0	56	70.27	100.00	0
	OSTA	50	75.40	100.00	+1.69	23	87.63	100.00	+0.42	36	70.86	100.00	+1.08

6.3.2. Efficiency of OSTA

The efficiency performances of OSTA are shown in Table 6-4. It was observed that the RAM of OSTA was equal to zero. This was because the calculations required for

the proposed pruning method were the same as the forward calculations for training and did not require additional GPU memory.

The calculation for estimating RATs and the measured RATs are shown in Table 6-4. The number of equivalent pruning iterations required for L7 Irish, L8 Biome, RIT-18 and Semantic3D was calculated to be 398.8%, 196.7%, 6.27% and 127.6% of the number of total training iterations, respectively. The final estimated RAT values were much smaller than these values (i.e., 83.3%, 196.7%, 1.35% and 32.54%, respectively) because a pruning iteration took much less time than a training iteration. The estimated RATs were close to, but slightly lower than the actual measured values. This is because parallelism was not perfect during the actual calculation. Nevertheless, this proves that the time consumption of OSTA was predictable.

Table 6-4: Efficiency metrics and calculation of OSTA.

		L7 Irish	L8 Biome	RIT-18	Semantic3D
RAM		0	0	0	0
RAT	Patches to be processed in each epoch of sub-validation set	$100 \times 4 = 400$	$50 \times 4 = 200$	$12 \times 4 = 48$	$32 \times 4 = 128$
	Total number of 3-channel combinations	56	120	20	56
	Total number of epochs of sub-validation set during pruning	$56 + 55 \dots + 2 = 1595$	$120 + 119 \dots + 2 = 7259$	$20 + 19 \dots + 2 = 209$	$56 + 55 \dots + 2 = 1595$
	Total equivalent pruning iterations	$400 \times 1595 / 16 = 39875$	$200 \times 7259 / 16 = 90737.5$	$48 \times 209 / 16 = 627$	$128 \times 1595 / 16 = 12,760$
	Ratio between equivalent pruning and training iterations	398.8%	907.4%	6.27%	127.6%
	Ratio between time of a pruning and a training iteration (measured)	20.9%	21.7%	21.5%	25.5%
	Estimated RAT	$398.8\% \times 20.9\% = 83.3\%$	$907.4\% \times 21.7\% = 196.7\%$	$6.27\% \times 21.5\% = 1.35\%$	$127.6\% \times 25.5\% = 32.54\%$
	Measured RAT	85.5%	198.1%	1.71%	37.7%

6.3.3. Ablation study

6.3.3.1. Replacing for pruning stage

The pruning had two components, including the pruning criteria and the pruning strategy. The effect of using different pruning criteria on the final accuracy was first tested. The test results are shown in Table 6-5. The highest accuracies were achieved when using the validation accuracy as the pruning criterion. In addition, these SCCs

Chapter 6: Automatic feature selection

all achieved higher accuracies in Tests 1 to 3 than they did in the SGS (i.e., positive DCAs). This suggested that the pruning criteria were not the reason for obtaining a positive DCA.

Table 6-5: Results of replacing pruning criteria.

	Pruning criteria	L7 Irish 3C				L8 Biome 3C				RIT-18			
		Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %
OSTA	Validation accuracy	56	62.49	100.00	+1.42	9	68.38	100.00	+1.69	19	66.53	100.00	+1.21
Test 1	Train accuracy	34	60.75	94.90	+0.40	16	66.73	96.94	+0.81	14	65.45	100.00	+1.57
Test 2	PCA	42	58.53	41.14	+0.85	90	64.56	76.33	+0.92	8	65.00	97.17	+1.23
Test 3	Entropy	49	58.23	31.92	+0.77	65	65.72	89.11	+0.37	20	64.01	66.00	+1.89
		L7 Irish 2C				L8 Biome 2C				Semantic3D			
		Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %
OSTA	Validation accuracy	50	75.40	100.00	+1.69	23	87.63	100.00	+0.42	36	70.86	100.00	+1.08
Test 1	Train accuracy	53	73.65	98.02	+1.16	28	87.40	90.76	+0.49	21	69.61	93.32	+0.62
Test 2	PCA	42	71.47	63.17	+0.71	90	83.59	11.67	+0.33	51	67.70	60.04	+0.46
Test 3	Entropy	49	71.98	73.75	+0.39	65	85.45	35.42	+0.75	1	63.12	12.69	+1.25

In the second ablation experiment, a very aggressive pruning strategy was tested. It ranked the validation accuracy for all channel combinations once at the end of the supernet training (i.e., stage 1) and the combination with the highest accuracy was selected for subsequent fine-tuning. The original iterations used for pruning were merged into the fine-tuning stage in this strategy. The results in Table 6-6 suggested that an overaggressive pruning strategy could lead to a remarkable drop in the final accuracy. Nevertheless, relatively large positive DCAs were still obtained in this ablation experiment, suggesting that the main cause for the positive DCA would not be in the pruning stage.

Table 6-6: Results of replacing pruning strategy.

	Pruning strategy	L7 Irish 3C				L8 Biome 3C				RIT-18			
		Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %
OSTA	Discrete forward-only pruning strategy	56	62.49	100	+1.42	9	68.38	100.00	+1.69	19	66.53	100	+1.21
Test 4	Ranking at the end of stage 1	13	60.56	92.52	+0.73	16	66.51	95.56	+0.59	16	65.24	99.30	+0.95
		L7 Irish 2C				L8 Biome 2C				Semantic3D			
		Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %
OSTA	Discrete forward-only pruning strategy	50	75.40	100	+1.69	23	87.63	100.00	+0.42	36	70.86	100	+1.08
Test 4	Ranking at the end of stage 1	39	73.95	98.57	+1.37	32	87.38	100.00	+0.38	11	69.95	96.16	+0.75

6.3.3.2. Replacing for supernet training stage

The results of ablation on the supernet training stage are shown in Table 6-7. When the linear warmup was removed, the poly learning rate policy was used for all stages of OSTA. Meanwhile, when the supernet training was removed, OSTA was started directly from the pruning stage. From the results of Test 5-7, it is clear that the entire supernet training stage was the key to achieving the positive DCA.

Table 6-7: Results of removing linear warmup and/or supernet training stage.

	Removing		L7 Irish 3C				L8 Biome 3C				RIT-18			
	Linear warmup	Supernet training	Index	mIoU	% CAP	% DCA	Index	mIoU	% CAP	% DCA	Index	mA	% CAP	% DCA
OSTA	N	N	56	62.49	100.00	+1.42	9	68.38	100.00	+1.69	19	66.53	100.00	+1.21
Test 5	Y	N	56	62.36	100.00	+1.29	2	68.31	100.00	+1.50	19	66.25	100.00	+0.93
Test 6	N	Y	53	61.06	98.13	+0.63	23	67.49	98.90	+0.51	19	65.77	100.00	+0.45
Test 7	Y	Y	26	60.57	92.63	-0.02	1	66.49	95.49	+0.28	10	64.69	92.00	+0.04
			L7 Irish 2C				L8 Biome 2C				Semantic3D			
			Index	mIoU	% CAP	% DCA	Index	mIoU	% CAP	% DCA	Index	mIoU	% CAP	% DCA
OSTA	N	N	50	75.40	100.00	+1.69	23	87.63	100.00	+0.42	36	70.86	100.00	+1.08
Test 5	Y	N	33	74.13	98.84	+0.56	2	87.54	100.00	+0.26	36	70.66	100.00	+0.87
Test 6	N	Y	33	73.70	98.09	+0.13	81	87.41	100.00	+0.22	56	70.59	100.00	+0.32
Test 7	Y	Y	12	73.44	94.54	+0.04	96	87.34	100.00	+0.14	55	70.11	97.29	-0.14

Based on this finding, an additional set of tests was conducted. The selected combinations by OSTA were used to directly fine-tune the trained supernet. The results in Table 6-8 confirmed that using the trained supernet as a pre-trained model can improve DCA, and indicated that the pruning stage can further boost DCA.

Table 6-8: Results of directly fine-tuning the trained supernet with selected combinations by OSTA.

	Direct fine-tune the trained supernet on channel combination	L7 Irish 3C				L8 Biome 3C				RIT-18			
		Index	mIoU	% CAP	% DCA	Index	mIoU	% CAP	% DCA	Index	mA	% CAP	% DCA
OSTA		56	62.49	100.00	+1.42	9	68.38	100.00	+1.69	19	66.53	100.00	+1.21
Test 8	Selected by OSTA	56	62.06	100.00	+1.09	9	68.13	100.00	+1.44	19	66.05	100.00	+0.73
		L7 Irish 2C				L8 Biome 2C				Semantic3D			
		Index	mIoU	% CAP	% DCA	Index	mIoU	% CAP	% DCA	Index	mIoU	% CAP	% DCA
OSTA		50	75.40	100.00	+1.69	23	87.63	100.00	+0.42	36	70.86	100.00	+1.08
Test 8	Selected by OSTA	50	74.87	99.94	+1.16	23	87.43	100.00	+0.22	36	70.48	100.00	+0.70

6.3.3.3. Replacing for pretrained weights in SSN

As shown in Table 6-7, pruning from different initial conditions may select different channel combinations, which can be interpreted from the following perspective. For a given dataset and SSN, the “best” channel combination may not be fixed for different

Chapter 6: Automatic feature selection

initial values of parameters used. Therefore, two additional network initialization methods were tested in this ablation study, including a SSN fine-tuning on the Cityscapes (Cordts et al., 2016) (based on parameters trained on ImageNet) and a randomly initialized SSN. Due to the tremendous amount of work required for benchmarking, only the Semantic3D data was tested. Semantic3D was chosen because it contains scenes similar to that in the Cityscapes. It is of interest to test the effect of using a dataset from similar scenes to pre-train the SSN on segmentation accuracy.

Table 6-9: Results of replacing for pretrained weights in SSN.

		ImageNet				Cityscapes				Random				
		Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	Index	mIoU %	CAP %	DCA %	
DF		-	61.44	4.47	-	-	64.58	17.86	-	-	45.77	86.31	-	
Feature extraction	PCA	-	66.13	49.68	-	-	63.85	14.17	-	-	35.46	9.69	-	
	LC-Net	-	69.67	93.79	-	-	70.30	96.76	-	-	46.79	91.84	-	
Channel selection	BS-Nets	34	68.72	80.36	0	34	69.63	87.50	0	34	38.33	33.93	0	
	ONR	27	62.96	10.71	0	27	64.08	16.07	0	27	33.08	7.14	0	
	DARecNet	43	65.26	35.71	0	43	67.41	53.57	0	43	45.79	87.50	0	
	Top 10 in SGS		56	70.27	100.00	0	36	71.19	100.00	0	26	52.70	100.00	0
			55	70.25	98.21	0	46	70.88	98.21	0	46	50.08	98.21	0
			46	69.98	96.43	0	55	70.17	96.43	0	11	49.62	96.43	0
			36	69.78	94.64	0	56	70.13	94.64	0	36	47.13	94.64	0
			52	69.55	92.86	0	21	70.02	92.86	0	25	46.83	92.86	0
			25	69.30	91.07	0	19	69.86	91.07	0	6	46.76	91.07	0
			11	69.20	89.29	0	14	69.82	89.29	0	33	46.56	89.29	0
			21	68.99	87.50	0	34	69.63	87.50	0	43	45.79	87.50	0
			19	68.80	85.71	0	5	69.57	85.71	0	56	45.76	85.71	0
			44	68.76	83.93	0	39	69.46	83.93	0	21	45.34	83.93	0
OSTA	36	70.86	100.00	+1.08	46	71.39	100.00	+0.51	26	55.97	100.00	+3.27		

The results are summarised in Table 6-9, which supported the speculation that using different pre-training parameters would lead to changes in the “best” channel combination. Nevertheless, OSTA achieved the highest accuracy for all the initialisation cases tested, again proving its task adaptive capability. In addition, almost all methods achieved better absolute segmentation accuracies (mIoU) using the cityscapes pre-trained SSN than using the ImageNet pre-trained SSN. This illustrates the importance of similarity between the fine-tuning and pre-training data for the fine-tuning accuracy.

6.4. Discussion

6.4.1. Most robust combination

It was found that some channel combinations from SGS in Table 6-9 achieved promising CAPs regardless of the initialisation values used for the parameters, which are summarised in Table 6-10. Among these combinations, the combination 46 achieved excellent CAPs for all cases with an average value of 97.62%. The channel combination with such a characteristic is referred to as the most robust combination (MRC). Designing selection criteria for MRC is meaningful future work. The detailed segmentation results for the channel combinations in Table 6-10 are shown in Table 6-11, which might provide some inspiration for readers. For example, by changing the dimension used for sorting (Table 6-12), it was seen that combinations 46 and 36 dominated the highest segmentation accuracy for the two most difficult classes, i.e., hard scape and scanning artefacts, respectively.

Table 6-10: Summary of recurring channel combinations in the Top 10 of SGS.

Index	Channels			CAP %				
				ImageNet	Cityscapes	Random	Average	Standard deviation
11	R	B	De	89.29	67.86	96.43	84.52	14.87
36	G	Ze	De	94.64	100.00	94.64	96.43	3.09
46	B	Ze	De	96.43	98.21	98.21	97.62	1.03
55	Z	Ze	De	98.21	96.43	75.00	89.88	12.92
56	D	Ze	De	100.00	94.64	85.71	93.45	7.22

Table 6-11: Detailed segmentation results for combinations in Table 6-10.

Pretrained weights	Index	CAP %	mIoU %	IoU %							
				Man-made	Natural	High Veg.	Low Veg.	Buildings	Hard scape	Scanning artefacts	Cars
ImageNet	11	89.29	69.20	91.81	83.67	88.17	62.04	89.27	31.47	36.17	71.01
	36	94.64	69.78	92.49	83.91	88.84	63.85	88.00	30.57	39.21	71.16
	46	96.43	69.98	92.83	83.60	87.80	66.66	88.40	35.08	35.06	70.43
	55	98.21	70.25	92.78	84.83	89.38	65.61	88.57	31.00	39.22	70.64
	56	100.00	70.27	92.66	86.71	90.19	70.23	88.00	34.64	24.16	75.23
Cityscapes	11	67.86	68.70	90.07	81.15	86.29	60.17	87.41	31.69	40.73	72.08
	36	100.00	71.19	92.43	83.66	88.94	62.38	89.93	36.05	43.35	72.78
	46	98.21	70.88	92.42	83.86	88.46	61.47	89.96	36.42	43.01	71.46
	55	96.43	70.17	92.97	83.46	87.87	62.56	87.60	34.49	39.17	73.21
	56	94.64	70.13	92.31	85.42	89.45	61.93	90.20	36.26	33.69	71.75
Random	11	96.43	49.62	82.28	47.90	84.62	33.08	77.34	14.67	28.70	28.39
	36	94.64	47.13	62.99	33.83	85.35	30.75	76.38	20.30	31.6	35.88
	46	98.21	50.08	76.69	51.08	82.90	31.51	78.51	21.12	27.46	31.36
	55	75.00	44.89	77.68	33.78	76.35	23.78	72.02	20.27	27.87	27.35
	56	85.71	45.76	81.61	44.43	69.63	26.98	75.74	20.34	24.71	22.61

Table 6-12: Replication of Table 6-11, re-arranged in the vertical direction according to the combination used. The bolded one indicate the best accuracy for that class were achieved by using corresponding combination.

Index	Pretrained weights	CAP %	mIoU %	IoU %							
				Man-made	Natural	High Veg.	Low Veg.	Buildings	Hard scape	Scanning artefacts	Cars
11	ImageNet	89.29	69.20	91.81	83.67	88.17	62.04	89.27	31.47	36.17	71.01
	Cityscapes	67.86	68.70	90.07	81.15	86.29	60.17	87.41	31.69	40.73	72.08
	Random	96.43	49.62	82.28	47.90	84.62	33.08	77.34	14.67	28.70	28.39
36	ImageNet	94.64	69.78	92.49	83.91	88.84	63.85	88.00	30.57	39.21	71.16
	Cityscapes	100.00	71.19	92.43	83.66	88.94	62.38	89.93	36.05	43.35	72.78
	Random	94.64	47.13	62.99	33.83	85.35	30.75	76.38	20.30	31.60	35.88
46	ImageNet	96.43	69.98	92.83	83.60	87.80	66.66	88.40	35.08	35.06	70.43
	Cityscapes	98.21	70.88	92.42	83.86	88.46	61.47	89.96	36.42	43.01	71.46
	Random	98.21	50.08	76.69	51.08	82.90	31.51	78.51	21.12	27.46	31.36
55	ImageNet	98.21	70.25	92.78	84.83	89.38	65.61	88.57	31.00	39.22	70.64
	Cityscapes	96.43	70.17	92.97	83.46	87.87	62.56	87.60	34.49	39.17	73.21
	Random	75.00	44.89	77.68	33.78	76.35	23.78	72.02	20.27	27.87	27.35
56	ImageNet	100.00	70.27	92.66	86.71	90.19	70.23	88.00	34.64	24.16	75.23
	Cityscapes	94.64	70.13	92.31	85.42	89.45	61.93	90.20	36.26	33.69	71.75
	Random	85.71	45.76	81.61	44.43	69.63	26.98	75.74	20.34	24.71	22.61

6.4.2. Coastal aerosol band for cloud detection

The CA band has always been ignored in the field of cloud detection. When L8 Biome was established, the CA band was not used to label clouds, as stated in (Foga et al., 2017) that “Band 1 (coastal aerosol) was never used”. However, this study suggested that CA might be an important channel for cloud detection. As shown in Table 6-13, the CA band was frequently present in the top 10 of all SGS channel combinations for the L8 Biome benchmark data. In particular, 8 of those top 10 combinations included the CA band for segmenting clouds into thin and thick clouds. In the future, it would be of interest to use methods such as Bradley-Terry model (Stein et al., 2005) to statistically analyse the error matrices generated by different channel combinations. New index might be developed for cloud detection. In addition, it is conjecture that for other remote sensing tasks, “CA band” could also exist.

Chapter 6: Automatic feature selection

Table 6-13: Top 10 in SGS for cloud detection benchmark data.

	L7 Irish 3C				L7 Irish 2C				L8 Biome 3C				L8 Biome 2C			
	Index	Band 1	Band 2	Band 3	Index	Band 1	Band 2	Band 3	Index	Band 1	Band 2	Band 3	Index	Band 1	Band 2	Band 3
Top 10 in SGS	50	NIR	TLG	THG	56	TLG	THG	MIR	3	CA	B	NIR	106	NIR	SWIR2	T1
	56	TLG	THG	MIR	50	NIR	TLG	THG	10	CA	G	NIR	22	CA	NIR	SWIR1
	5	B	G	THG	33	G	SWIR	MIR	23	CA	NIR	SWIR2	107	NIR	SWIR2	T2
	25	G	R	THG	12	B	NIR	SWIR	2	CA	B	R	23	CA	NIR	SWIR2
	26	G	R	MIR	27	G	NIR	SWIR	9	CA	G	R	96	R	SWIR2	T1
	53	SWIR	TLG	THG	39	R	NIR	THG	22	CA	NIR	SWIR1	81	G	SWIR2	T1
	34	G	TLG	THG	26	G	R	MIR	74	G	NIR	T1	103	NIR	SWIR1	T1
	6	B	G	MIR	38	R	NIR	TLG	1	CA	B	G	60	B	SWIR2	T1
	15	B	NIR	MIR	53	SWIR	TLG	THG	37	B	G	R	27	CA	SWIR1	SWIR2
	39	R	NIR	THG	34	G	TLG	THG	16	CA	R	NIR	57	B	SWIR1	T1

6.4.3. Training with channel combinations other than the selected one

It was found that both the supernet training and the pruning stages boosted DCA. These two stages shared common operations in that they both used channel combinations other than SCC to train the SSN. Training in this way could force the SSN to learn to use channel-invariant features for semantic segmentation. This is the reason why SCC in OSTA can achieve higher accuracy than training SSN with SCC alone (i.e., in SGS). This may provide insight for the development of new methods for model pre-training and/or data augmentation. In addition, it is conjectured that this mechanism can be used to reduce the domain-shift problem caused by the different spectrums used in image sensors. Integration of this mechanism with existing methods (Aryal and Neupane, 2023; Z. Li et al., 2022; Yan et al., 2020) may lead to better solutions for domain adaptation.

6.4.4. Limitation of OSTA

The major limitation of OSTA is that it still requires the prerequisite setting of dimension reduction target. Therefore, developing a channel selection method that can automatically determine the optimal number of channels is meaningful for future work.

A complementary future work is to develop pre-trained models based on multichannel image datasets. This will not only benefit the task of channel selection, but will greatly facilitate the advancement of the field of multichannel image processing.

6.5. Summary

This study confirmed that it was always desirable to use fewer feature channels to achieve higher semantic segmentation accuracy. Although many channel selection methods have been developed to achieve this aim, they have several limitations. Limitations affecting the semantic segmentation accuracy come from the use of selection criteria other than segmentation accuracy and the use of evaluation of individual channels to select channel combinations. Meanwhile, the limitation affecting efficiency comes from the repetitive training of classifier(s). A one-shot task-adaptive (OSTA) channel selection method was proposed in this study to determine a channel combination with the-state-of-art accuracy of semantic segmentation, in comparison to the existing methods. OSTA was based on the concept of pruning from a supernet, which integrated the channel selection and SSN training processes, thus avoiding repetitive training of SSN. The limitations affecting accuracy were addressed by using the semantic segmentation accuracy of different channel combinations on the validation set as the pruning criterion in OSTA. The effectiveness and efficiency of OSTA were tested using four datasets, including L7 Irish, L8 Biome, RIT-18 and Semantic3D. OSTA achieved the highest semantic segmentation accuracies in all benchmark tests. Compared to a single training session of SSN, OSTA did not require extra memory footprint, and took a minimum of 1.71% to a maximum of 198.1% extra time (predictable) to select the best 3-channel combination for four datasets tested.

Chapter 6: Automatic feature selection

To the best knowledge of the authors, OSTA was found to be the first channel selection method that produced a semantic segmentation accuracy exceeding the highest accuracy obtained by exhaustive tests of channel combinations. Experiments suggested that this was because training the SSN with extra channel combinations could improve the semantic segmentation accuracy. This mechanism can potentially be used to develop new pre-training/data augmentation methods.

Experiments also revealed that in addition to the “best channel combinations”, there was a most robust channel combination that achieved excellent accuracy performance regardless of the network parameter initialisation method used. It is recommended that future work be devoted to design selection criteria for this type of channel combination.

It was also interesting to find out that the coastal aerosol band was important for cloud detection. New cloud detection methods could be developed in the future based on this finding.

Chapter 7: Stacking-based semantic segmentation framework

This chapter is based on the published paper: Cai, Y., Fan, L., Fang, Y., 2023a. SBSS: Stacking-Based Semantic Segmentation Framework for Very High-Resolution Remote Sensing Image. IEEE Trans. Geosci. Remote Sens. 61, 1–14. <https://doi.org/10.1109/TGRS.2023.3234549>

Note: This chapter develops a novel framework to improve the semantic segmentation accuracy for images. The different methods developed in this thesis are also integrated in this chapter and its performance is tested on Semantic3D.

7.1. Introduction

Semantic segmentation is a fundamental task for many remote sensing applications, such as land cover classification, cloud detection and urban scene understanding (Cai et al., 2022a, 2021b; Yang Chen et al., 2022a, 2022b; Ding et al., 2020; Hansch and Hellwich, 2021; Tokarczyk et al., 2015; Wei and Hansch, 2022). With the rapid development of imaging technology, the resolution of the acquired images has significantly improved. This trend is also reflected in the publicly available semantic segmentation datasets. For example, the image resolutions are approximately 480p (480×367) in the PASCAL VOC2012 dataset (Everingham et al., 2012), 2k (2048×1024) in the Cityscapes (2016) dataset (Cordts et al., 2016), and 4k (4096×2160 to 3840×2160) in the recently released UAVid (2020) dataset (Lyu et al., 2020). For a fixed imaging distance, higher camera resolution means finer spatial resolution of an image acquired. The rich spatial details in Very High Resolution (VHR) images provide an opportunity for more accurate semantic segmentation of a target scene. However, the use of VHR images poses a new challenge to the semantic segmentation task, i.e., the simultaneous segmentation of objects with large scale discrepancies. This is caused by the fact that the fine spatial resolution of VHR images enables segmentation of objects at smaller scales.

Extensive research has been conducted to address this challenge, where deep learning has become the dominant approach. A typical semantic segmentation neural network consists of two components: the encoder and the decoder. The encoder is responsible for extracting features from an input image at multiple down-sampling scales. Subsequently, by interpreting these extracted features, the decoder assigns an

appropriate label to each pixel in the image. Various designs have been proposed for these two components to achieve better segmentation performance.

Encoders can broadly be divided into single branch networks and multi-branch networks, according to their macro designs. The representative single-branch encoder networks include VGGNet (Simonyan and Zisserman, 2015), ResNet (He et al., 2016), Xception (Chollet, 2017), MobileNetV2 (Sandler et al., 2018), Swin (Liu et al., 2021b) and ConvNeXt (Z. Liu et al., 2022), which extract features at progressively reduced scales in a tandem fashion. The representative multi-branch networks include BiSeNet (Yu et al., 2018) and HRNet (Jingdong Wang et al., 2021). BiSeNet uses the spatial path and the context path to extract high-resolution spatial details and global contextual information, respectively. HRNet uses four parallel branches to extract high-level features at four down-sampling scales simultaneously.

A critical aspect of the decoder design is to expand the receptive field to model long-range dependencies without reducing the spatial resolution. Based on the mechanisms for long-range dependency modelling, decoders can be classified as convolution-based networks and dot-product attention-based networks. The well-known convolution-based ones include U-Net (Ronneberger et al., 2015), Spatial Pyramid Pooling (SPP) (He et al., 2015), Pyramid Pooling Module (PPM) (Zhao et al., 2017) and Atrous Spatial Pyramid Pooling (ASPP) (Chen et al., 2018a, 2018b, 2017), which rely mainly on the pyramid-like structure. Meanwhile, the success of dot-product attention-based networks is due to their global modelling ability. However, the computational cost of the dot-product attention mechanism increases quadratically with the size of the features used (Dosovitskiy et al., 2020). Therefore, efficient use of this mechanism is

Chapter 7: Stacking-based semantic segmentation framework

a main focus of the relevant previous research, which includes Dual Attention Network (DANet) (Li et al., 2021), Disentangled Non-Local Neural Networks (DNLNet) (Yin et al., 2020), Object Context network (OCRNet) (Yuan et al., 2020), Attentive Bilateral Contextual network (ABCNet) (Li et al., 2021) and Multiattention Network (MANet) (R. Li et al., 2022).

Although there are various designs of segmentation networks, they share one common characteristic, which is that features can only be extracted on a predefined set of scales. Due to computational constraints, it is impractical to extract features at too many scales (currently up to four scales) in a segmentation network. However, there is no guarantee that those pre-defined scales are the optimal ones for a given application scenario. To enable a segmentation network to analyse images at a wider range of scales, it is the common practice to use a test-time data augmentation called the Multi Scale (MS) test. The MS resizes the original images to various scales and feeds them into a segmentation network. The output segmentation maps are then often assembled by average voting. In addition, the MS is typically used in conjunction with training-time multi-scale data augmentation. When both methods are used, the entire segmentation framework falls under an ensemble learning technique called bootstrap aggregating (Bagging) (Breiman, 1996). More specifically, images at different scales are used as the bagging samples in this process. The role of bagging is to reduce the variance of errors among multiple predictions using different bagging samples (Breiman, 2001; Bühlmann and Yu, 2002). In other words, the MS works best if the prediction error for each class is randomly distributed over the scales used for the image resizing. The MS has become the default method for testing the best performance of a segmentation

network. However, it is worth investigating whether there is a better method to fuse the segmented maps resulting from input images of different resizing scales.

It is common sense that the size distribution of objects of different classes in an image is similar to that in reality, despite the effects of perspective. For example, larger objects in reality are usually also larger in images (e.g., buildings often occupy a large area in street view images). Meanwhile, studies have shown that the size of the effective receptive field for a network is limited (Ding et al., 2022; Kim et al., 2021; Luo et al., 2017). Therefore, it is hypothesised that for a given segmentation network and a dataset, each class may have its preferred resizing scale for segmentation. More specifically, the classes that typically have large objects may prefer to be shrunk so that they can be fitted into the effective receptive field and segmented as a whole, while the classes that typically have small objects may prefer to be zoomed in to avoid becoming indistinguishable after being downsampled by the segmentation network. In other words, the prediction errors for each class may have biases related to the resizing scales.

Based on this hypothesis and inspired by previous studies (Leblanc and Tibshirani, 1996; Wolpert, 1992), a Stacking-Based Semantic Segmentation (SBSS) framework was proposed in this study to reduce the error associated with the resizing scales. In the SBSS framework proposed, a segmentation map obtained at the smaller scale is gradually corrected by a learnable Error Correction Module (ECM) using a segmentation map obtained at a larger resizing scale. This process starts with an initial segmentation map at the smallest scale considered and is repeated multiple times (each time, a larger scale was used). The computational complexity of the SBSS framework

Chapter 7: Stacking-based semantic segmentation framework

is flexible, which can be altered by assigning different Error Correction Schemes (ECS). In particular, two ECS were proposed in this study, namely ECS-MS and ECS-SS, which have similar Floating-point operations (Flops) to the MS test and the Single-Scale (SS) test, respectively. The ECS-MS is designed for the applications requiring the highest possible segmentation accuracy. Meanwhile ECS-SS is designed for applications where Graphics Processing Unit (GPU) memory is limited. The effectiveness of the SBSS framework was demonstrated on four datasets, including Cityscapes, UAVid, LoveDA and Potsdam, which cover a variety of scenarios (e.g., urban and rural) and acquisition perspectives (e.g., street levels, inclined drones and aerial views).

7.2. SBSS framework

7.2.1. Overview of SBSS framework

For n selected scales (sorted from the smallest to the largest), the workflow of the SBSS framework is illustrated in Figure 7-1 where $i = 1 \sim n$. The explanations of the abbreviations used are listed in Table 7-1. To implement SBSS, an initial segmentation map (Y_1) is required, which is obtained using the following two steps: (1) an original input image is resized to the smallest scale considered; (2) the resized input image (X_1) is fed to a segmentation network to obtain the initial segmentation map.

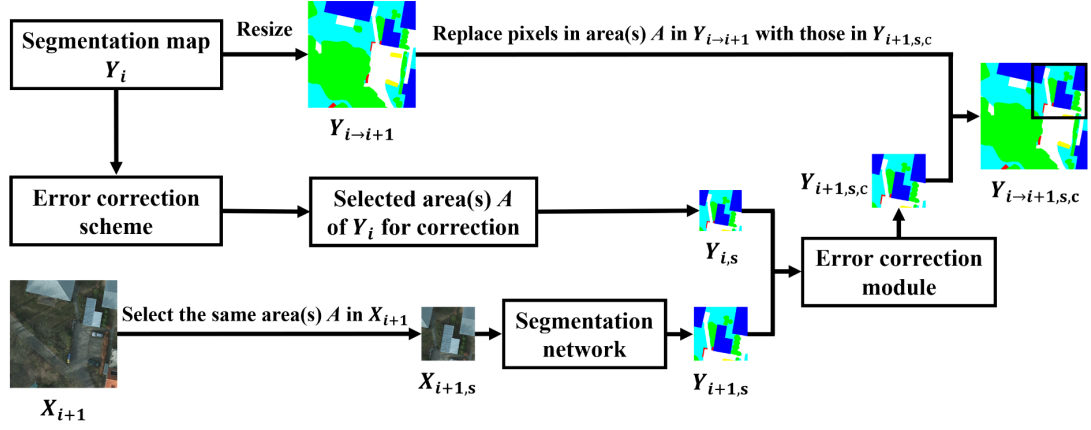


Figure 7-1: Workflow of one iteration of the SBSS framework, in which Y_i and X_{i+1} are the input maps; $Y_{i \to i+1,s,c}$ is the output map of one iteration and the segmentation map (i.e., a new Y_i) for the next iteration; the loop is ended when $i + 1 = n$ (i.e., the number of scales used in SBSS); the area(s) A are determined by the error correction scheme, and are also applied to X_{i+1} and $Y_{i \to i+1}$.

Table 7-1: The explanations of the abbreviations used in Figure 7-1.

Abbreviations	Explanations
Y_i	The segmentation map of the i^{th} scale.
$Y_{i \to i+1}$	The segmentation map resized from the i^{th} to the $(i + 1)^{\text{th}}$ scale.
$Y_{i,s}$	The selected area in Y_i for error correction.
X_{i+1}	The input image that is resized to the $(i + 1)^{\text{th}}$ scale.
$X_{i+1,s}$	The selected area from X_{i+1} for error correction.
$Y_{i+1,s}$	The segmentation map obtained using $X_{i+1,s}$ as the input.
$Y_{i+1,s,c}$	The corrected map $Y_{i+1,s}$ of the selected area.
$Y_{i \to i+1,s,c}$	The corrected map $Y_{i \to i+1}$.

The segmentation map (Y_i) at the beginning of the workflow is used in two parallel processes. In one process, Y_i is simply resized to a next scale to obtain the resized map $Y_{i \to i+1}$. In the other process, Y_i is fed into an error correction scheme (detailed in Section 7.2.3) to determine the area(s) where error correction is required (for ease of demonstration, only one local area is shown in Figure 7-1). Once the local area is identified, it is used to crop the local segmentation information $Y_{i,s}$ from Y_i , and meanwhile to crop the local image $X_{i+1,s}$ from the input image (X_{i+1}) that is resized to the $i+1$ scale. The resized input image of the local area is also fed into the segmentation network to obtain its corresponding segmentation information $Y_{i+1,s}$. The two segmentation maps (i.e., $Y_{i,s}$ and $Y_{i+1,s}$) of the selected area are processed in an error

Chapter 7: Stacking-based semantic segmentation framework

correction module (detailed in Section 7.2.2), which results in a corrected map $Y_{i+1,s,c}$ of the selected area. The segmentation information in $Y_{i+1,s,c}$ is used to replace that in the corresponding pixels in $Y_{i \rightarrow i+1}$, which produces an updated segmentation map $Y_{i \rightarrow i+1,s,c}$. If additional rounds of the iteration process are considered, $Y_{i \rightarrow i+1,s,c}$ is essentially the segmentation map (Y_i) used at the beginning of the workflow in the next round. Otherwise, it is the output (i.e., $Y_{n-1 \rightarrow n,s,c}$) of the last round. The final segmentation map is obtained by resizing $Y_{n-1 \rightarrow n,s,c}$ to the size of the original input image.

There are three main components in the SBSS framework, including a segmentation network, ECS and ECM. The segmentation network can be any existing network that provides a reasonably good initial segmentation result. The ECS is also flexible. For example, one can choose to correct the entire segmentation map or only select areas within it. All of these allow SBSS to be applied to various application scenarios with different demands. More detailed descriptions of the ECM and the two proposed ECS (ECS-MS and ECS-SS) are given in Section 7.2.2 and Section 7.2.3 respectively.

7.2.2. Error correction module

The proposed error correction module is shown in Figure 7-2. The segmentation map ($Y_{i,s}$) at a lower scale is first resized to a higher scale to obtain the resized map $Y_{i \rightarrow i+1,s}$. $Y_{i \rightarrow i+1,s}$ is concatenated with the segmentation map $Y_{i+1,s}$ using the input image at a higher scale. Subsequently, they (i.e., $Y_{i+1,s}$ and $Y_{i \rightarrow i+1,s}$) are fed into an Error Correction Network (ECN) to obtain the initial corrected segmentation map ($Y_{i+1,s,c,initial}$). Finally, an Adaptive Confidence Threshold (ACT) is used to replace

the corresponding pixels in $Y_{i \rightarrow i+1,s}$ with the more confident pixels in $Y_{i+1,s,c,initial}$ to obtain corrected segmentation information $Y_{i+1,s,c}$.

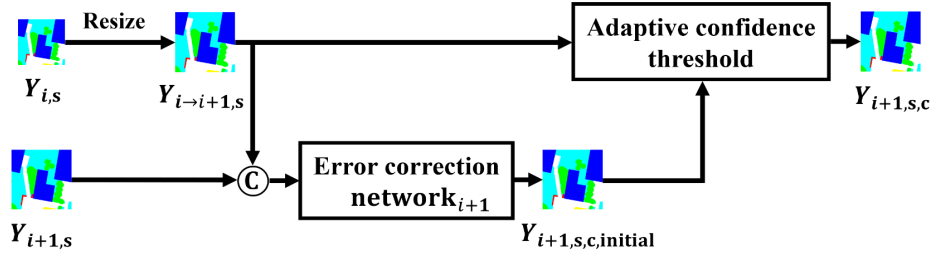


Figure 7-2: Error correction module.

7.2.2.1. Error Correction Network:

The detailed structure of the ECN is presented in Figure 7-3. For a dataset having C classes, its segmentation map also has C channels. Each channel of the segmentation map records the segmentation probability of its corresponding class. Therefore, concatenating the two segmentation maps will result in a feature map with a channel number of $2C$, which is used as the input to the ECN. The input features are processed through the stem block and two residual blocks. The resulting feature map (96 channels) is then compressed by a pointwise convolution layer (with C kernels) to output the initial corrected segmentation map ($Y_{i+1,s,c,initial}$). The weights of the ECN are not shared across scales (i.e., the ECN is trained separately for each scale).

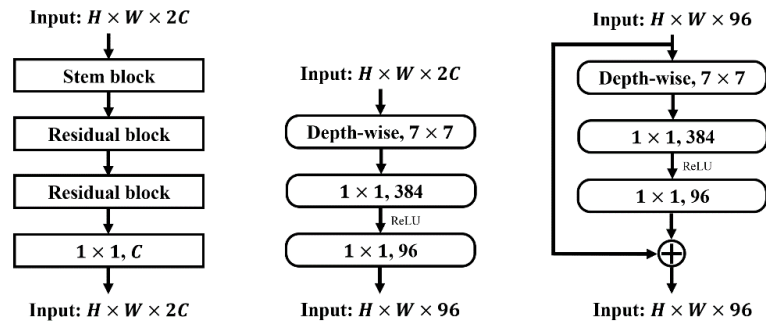


Figure 7-3: Structures of the error correction network (left), the stem block (middle), and the residual block (right).

Chapter 7: Stacking-based semantic segmentation framework

The structure of the residual block is designed in reference to the basic block of ConvNeXt (Z. Liu et al., 2022). Similar to ConvNeXt's study, it is found that better results were achieved using blocks with large convolutional kernels than those with small 3x3 kernels (e.g., ResNet block). The other minor modification is the removal of the normalisation layer. This is because a small decrease in segmentation accuracy was observed when using the normalisation layer in this study.

7.2.2.2. Adaptive confidence threshold

The objective of using the ACT is to select pixels that have low confidence levels in $Y_{i \rightarrow i+1, s}$ but high confidence levels in the initial corrected segmentation map ($Y_{i+1, s, c, \text{initial}}$). The ACT is implemented as follows. For each pixel in a segmentation map, its value in each channel represents the confidence level of belonging to the corresponding class, and the values in all channels are summed to one. Thus, the confidence level of each pixel in $Y_{i+1, s, c, \text{initial}}$ is its maximum value over all channels. The confidence map for $Y_{i+1, s, c, \text{initial}}$ and $Y_{i \rightarrow i+1, s}$ are denoted as $Y_{i+1, s, c, \text{initial}}^{\text{confidence}}$ and $Y_{i \rightarrow i+1, s}^{\text{confidence}}$, respectively. An Adaptive Confidence (AD) map is produced to match the objective of the adaptive confidence threshold, as shown in Equation 7-1 where “.*” represents the element-wise multiplication.

$$AD = (1 - Y_{i \rightarrow i+1, s}^{\text{confidence}}) .* Y_{i+1, s, c, \text{initial}}^{\text{confidence}} \quad (7-1)$$

The ACT was set to be the median pixel value in the AD map. Finally, the regions in the AD map that exceed the threshold are recorded and the results within that region in $Y_{i \rightarrow i+1, s}$ are replaced with the results in $Y_{i+1, s, c, \text{initial}}$ to obtain $Y_{i+1, s, c}$.

7.2.3. Error correction scheme

In the SBSS framework, apart from the segmentation network used, ECS also has a significant impact on the overall computational load. The commonly used metrics for

quantifying computational load include Flops and GPU memory footprints. For a given network structure, the main factors affecting Flops and GPU memory footprints are the total number and the size of the input patches, respectively. Therefore, the focus of developing an ECS is on the selection of areas to be corrected and on the choice of a patch size that does not exceed the GPU memory limit. With reference to the Flops required for the two commonly used test methods (i.e., MS and SS), two ECS (ECS-MS and ECS-SS) are proposed in this study. It is worth noting that the SBSS framework using ECS-MS and ECS-SS are abbreviated as SBSS-MS and SBSS-SS in subsequent sections. Moreover, for image patch extraction, the non-overlapping sliding window approach is used in this study.

7.2.3.1. Error correction scheme with Flops at the multi-scale test level

The MS has widely been adopted to obtain the highest possible segmentation accuracy. The commonly used set of scales is {0.5, 0.75, 1.0, 1.25, 1.5, 1.75}(Jingdong Wang et al., 2021; Junjue Wang et al., 2021). Under this setting, the original image is resized using each of those six scales before being processed by a segmentation network. The total size of the image patches that need to be processed by MS is shown in Table 7-2, which is almost nine times that of the original image.

Table 7-2: Comparison of the total size of the images to be processed by MS and SBSS-MS.

Scales		0.5	0.75	1.0	1.25	1.5	1.75	Sum
MS	Selection percentage of non-overlapping patches	100%	100%	100%	100%	100%	100%	-
	Ratio of the total size of the patches to the original image	25%	56%	100%	156%	225%	306%	869%
ECS-MS	Selection percentage of non-overlapping patches	100%	100%	100%	100%	100%	0%	-
	Ratio of the total size of the patches to the original image	25%	56%	100%	156%	225%	0%	563%

The proposed ECS-MS also processes all the image patches at each scale used, but with less scales. As shown in Table 7-2, the scale 1.75 is discarded in ECS-MS, which is to compensate for the additional Flops for ECM. Since the Flops of ECM are quite

Chapter 7: Stacking-based semantic segmentation framework

small compared to that of the segmentation network, the current ECS-MS setup is conservative. The exact Flops for ECS-MS and MS are given in Section 7.3.5.

7.2.3.2. Error correction scheme with Flops at the single-scale test level

The SS is often used when computational resources are limited, which only analyses the original image (i.e., at scale 1.0). Based on such considerations, the ECS-SS is designed in this study. The ECS-SS analyses images at four scales: {0.25, 0.5, 1.0, 1.5}. To keep the flops consumed by ECS-SS similar to SS, ECS-SS is unable to analyse all the image patches at four scales. Therefore, a selection strategy is designed for ECS-SS to analyse only part of the image patches. As shown in Table 7-3, for the two smaller scales (i.e., 0.25 and 0.5), all image patches are selected because of their relatively small total size compared to the original image. While for the latter two scales (i.e., 1.0 and 1.5), only part of the image patches is selected. The selection is based on the confidence map at that scale (i.e., $Y_{i \rightarrow i+1, s}^{\text{confidence}}$). Patches with relatively low confidence accumulations are selected for analysis. With this setup, SBSS-SS allows the use of images at a wider range of scales for analysis while keeping the total size of the images to be processed at 75% of that of SS. Similar to ECS-MS, the Flops saved by using ECS-SS are compensation for the extra Flops involved in ECM, and the exact Flops are provided in Section 7.3.5.

Table 7-3: Comparison of the total size of the images to be processed by SS and SBSS-SS.

Scales		0.25	0.5	1.0	1.5	Sum
SS	Selection percentage of non-overlapping patches	0%	0%	100%	0%	-
	Ratio of the total size of the patches to the original image	0%	0%	100%	0%	100%
ECS-SS	Selection percentage of non-overlapping patches	100%	100%	25%	8.33%	-
	Ratio of the total size of the patches to the original image	6.25%	25%	25%	18.75%	75%

7.3. Experiments and results

7.3.1. Datasets and implementation details

7.3.1.1. Datasets

The effectiveness of the proposed SBSS framework was tested on four datasets, including Cityscapes, UAVid, LoveDA and Potsdam. The key characteristics of these datasets are summarised in Table 7-4. The partition of the training, validation and test sets for the first three datasets follows their original implementations. For the Potsdam dataset, the RGB images with IDs of 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15 and 7_13 were used as the testing set (the same set of images was also used as the validation set in this study), while the remaining 24 RGB images were used for training.

Table 7-4: Summary of four datasets used.

Dataset	Number of images			Number of classes	Image resolution
	Training set	Validation set	Test set		
Cityscapes	2975	500	1525	19	2048×1024
UAVid	200	70	150	8	4096×2160 or 3840×2160
LoveDA	2522	1669	1796	7	1024×1024
Potsdam	24	-	14	6	6000×6000

7.3.1.2. Training setting

The training settings used in this study are summarized in Table 7-5. Most of these settings were consistent across three datasets used. The crop size used was different as it was set to be proportional to the original image in the dataset. For the UAVid dataset which has two different image sizes, all images were resized to 4096 × 2160 for ease of processing. The total training iterations for LoveDA and Potsdam were significantly less than those of the other two datasets, due to the relatively small sizes of LoveDA and Potsdam.

Chapter 7: Stacking-based semantic segmentation framework

Table 7-5: Training setting for the segmentation network.

Dataset	Cityscapes	UAVid	LoveDA	Potsdam
Patch size	1024×512	1024×540	512×512	750×750
Total training iterations	80 k	80 k	15 k	15 k
Pretraining dataset	ImageNet-1k			
Optimizer	Stochastic Gradient Descent (SGD)			
Initial learning rate	0.01			
Learning rate schedule	Poly learning rate policy with a power of 0.9			
Momentum	0.9			
Weight decay	0.0005			
Batch size	16			
Loss function	Cross entropy			
Data augmentation	Random cropping, random resize (0.25~2), random horizontal flipping, photo metric distortion			

After the segmentation networks had been trained, they were used to generate segmentation maps for each scale required to train the ECN. The ECN was trained using settings similar to those in Table 7-5. The differences include: no pretraining, no random scaling and photometric distortion being used, the total number of training iterations, and the initial learning rate that was reduced to one tenth of that in Table 7-5.

7.3.1.3. Evaluation metrics

The segmentation accuracy was evaluated using mean Intersection over Union (mIoU) in this study. Based on the confusion matrix, the mIoU is computed as:

$$mIoU = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FP_c + FN_c} \quad (7-2)$$

Where TP_c , FP_c , and FN_c represent the true positive, false positive and false negatives of class c , respectively.

7.3.2. Scale related segmentation error

The study is based on the hypothesis that different classes have their preferences for the resizing scale used. Extensive experiments were conducted in this study to confirm the validity of this hypothesis, which is presented in this section.

Chapter 7: Stacking-based semantic segmentation framework

Table 7-6: Input scales that achieve the highest segmentation accuracy for different classes.

Method	Backbone	Cityscapes (Patch size of 1024×512)				UAVid (Patch size of 1024×540)			
		Road	Sidewalk	Person	Bicycle	Building	Tree	Static car	Human
FCN	HRNet-w18	0.75	0.75	1.50	1.50	0.75	0.50	0.75	1.75
BiSeNetV1	ResNet50	0.75	0.75	1.75	1.50	0.75	0.50	1.00	1.75
PSPNet	ResNet50	0.75	0.75	1.50	1.50	0.50	0.50	1.00	1.75
DeepLabV3+	ResNet50	0.75	0.75	1.50	1.25	0.50	0.50	0.50	1.50
DANet	ResNet50	0.75	0.75	1.75	1.50	0.50	0.75	1.00	1.50
GCNet	ResNet50	0.75	0.75	1.50	1.50	0.75	0.50	1.00	1.50
DNLNet	ResNet50	0.75	0.75	1.50	1.50	0.75	0.75	1.00	1.25
UperNet	ResNet50	0.75	0.75	1.50	1.50	0.75	0.50	1.00	1.75
UperNet	Swin-T	0.75	0.75	1.75	1.50	0.50	0.75	1.00	1.75
UperNet	ConvNeXt-T	0.75	0.75	1.75	1.50	0.75	0.50	0.75	1.25

Method	Backbone	LoveDA (Patch size of 512×512)				Potsdam (Patch size of 750×750)			
		Agricultural	Water	Forest	Barren	Building	Impervious	Tree	Car
FCN	HRNet-w18	0.50	0.50	1.00	1.00	0.75	1.00	1.00	1.50
BiSeNetV1	ResNet50	0.75	0.75	1.25	1.00	1.00	1.00	1.25	1.50
PSPNet	ResNet50	0.50	0.75	1.25	1.00	0.75	1.00	1.25	1.50
DeepLabV3+	ResNet50	0.50	0.50	1.25	1.00	0.75	1.00	1.25	1.25
DANet	ResNet50	0.50	0.75	1.50	1.00	0.75	1.00	1.00	1.25
GCNet	ResNet50	0.50	0.75	1.25	1.00	0.75	1.00	1.00	1.25
DNLNet	ResNet50	0.50	0.75	1.25	1.00	0.75	1.00	1.25	1.25
UperNet	ResNet50	0.50	0.75	1.25	1.00	0.75	1.00	1.00	1.50
UperNet	Swin-T	0.50	0.75	1.50	1.00	0.75	1.25	1.00	1.25
UperNet	ConvNeXt-T	0.50	0.75	1.75	1.00	0.75	1.00	1.25	1.25

In total, ten segmentation networks of similar sizes were tested in this study, including HRNet (Jingdong Wang et al., 2021), BiSeNetV1 (Yu et al., 2018), PSPNet (Zhao et al., 2017), DeepLabV3+ (Chen et al., 2018b), DANet (Li et al., 2021), GCNet (Cao et al., 2019), DNLNet (Yin et al., 2020), UperNet (Xiao et al., 2018), Swin (Liu et al., 2021b), and ConvNeXt (Z. Liu et al., 2022). These networks are representative works in the field of semantic segmentation. In the experiments, these networks were trained on the training sets of the four datasets. For each dataset, the input images in the validation set were resized to a set of scales $\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$, generating six new validation sets at these scales. The SS tests were performed on these newly generated validation sets. The scale corresponding to the highest segmentation accuracy obtained for each class was recorded. To facilitate the presentation of the

Chapter 7: Stacking-based semantic segmentation framework

experimental results in reasonably sized tables, four classes were randomly selected in each dataset and their preferred resizing scales are presented in Table 7-6. It was observed that the preferred resizing scale for a class is usually the opposite of the size of the image area occupied by that class. For example, the classes that usually occupy larger areas in the image prefer to be segmented using smaller resizing scales. These experimental results proved the validity of the hypothesis of this study.

7.3.3. Segmentation network choice

In the SBSS framework, the role of the segmentation network is to provide the raw segmentation map at multiple scales. A more accurate raw segmentation map is beneficial to improve the final segmentation accuracy of the SBSS framework. The segmentation accuracies of those ten networks (presented in Section 7.3.2) using SS on the validation set are summarised in Table 7-7. The highest segmentation accuracies were achieved by ConvNeXt-T in all tests, and was therefore chosen as the segmentation network for the SBSS framework in this study.

Table 7-7: Segmentation accuracy (mIoU) on validation sets using single scale tests (%).

Method	Backbone	Cityscapes		UAVid		LoveDA		Potsdam	
		Patch size		Patch size		Patch size		Patch size	
		1024 ×512	512 ×256	1024 ×540	512 ×270	512 ×512	256 ×256	750 ×750	375 ×375
FCN	HRNet-w18	75.73	68.85	73.75	72.40	51.20	49.60	85.49	76.69
BiSeNetV1	ResNet50	75.06	58.55	73.19	71.60	49.36	45.90	84.70	81.78
PSPNet	ResNet50	77.90	72.68	73.42	72.03	51.49	49.74	85.85	84.26
DeepLabV3+	ResNet50	78.66	74.35	73.65	72.40	50.71	48.72	85.73	84.22
DANet	ResNet50	78.64	74.36	73.63	72.54	51.36	50.28	86.06	84.98
GCNet	ResNet50	77.68	73.49	73.33	72.22	50.80	49.64	85.82	84.70
DNLNet	ResNet50	78.31	74.50	73.47	72.33	51.25	50.27	85.61	84.65
UperNet	ResNet50	77.65	71.92	73.94	72.34	51.04	48.75	85.62	83.62
UperNet	Swin-T	77.47	74.99	74.06	72.57	52.42	50.30	86.07	85.01
UperNet	ConvNeXt-T	78.84	75.52	74.26	72.68	52.52	50.47	86.41	85.12

7.3.4. Ablation study on the test setting

To evaluate the effectiveness of the different components within the SBSS framework, extensive ablation experiments were conducted in this study. The experimental setup

Chapter 7: Stacking-based semantic segmentation framework

and the corresponding results are presented in Table 7-8. For each dataset, experiments were conducted with two input patch sizes. The larger one represented the input size typically used in previous studies. The other input size was half of the larger one, which was used to simulate the scenario of limited GPU memories.

Table 7-8: The quantitative results of the ablation studies on validation sets of four datasets.

Method	Cityscapes		UAVid		LoveDA		Potsdam	
	Patch size	mIoU (%)	Patch size	mIoU (%)	Patch size	mIoU (%)	Patch size	mIoU (%)
MS		81.32		75.76		53.18		87.34
ECS-MS + ACT	1024	81.64	1024	76.14	512	53.91	750	87.43
ECS-MS + ECN	×512	82.82	×540	76.53	×512	55.77	×750	87.61
SBSS-MS (ECS-MS + ACT + ECN)		83.05		76.78		56.28		87.68
SS		75.52		72.34		50.47		85.12
ECS-SS + ACT	512	76.15	512	72.73	256	51.03	375	85.45
ECS-SS + ECN	×256	78.11	×270	73.57	×256	51.66	×375	86.15
SBSS-SS (ECS-SS + ACT + ECN)		78.52		73.85		51.98		86.35

As shown in Table 7-8, using either the ACT or the ECN alone improved the segmentation accuracy by an average of 0.43% or 1.40% respectively, which justified the design of the ECM. In addition, an average improvement of 1.55% and 1.81% in segmentation accuracy was achieved by using SBSS-MS and SBSS-SS, respectively.

To visually validate the effectiveness of the proposed SBSS framework, a comparison of the segmentation results generated by MS and SBSS-MS is shown in Figure 7-4. It can be observed that the segmentation results obtained using SBSS-MS had less visually fragmented areas compared to the MS (e.g., the areas within the red box in the example images of UAVid, LoveDA and Potsdam). In the meantime, the segmentation example from Cityscapes showed that SBSS-MS was also able to segment objects that were completely missed by MS.

Chapter 7: Stacking-based semantic segmentation framework

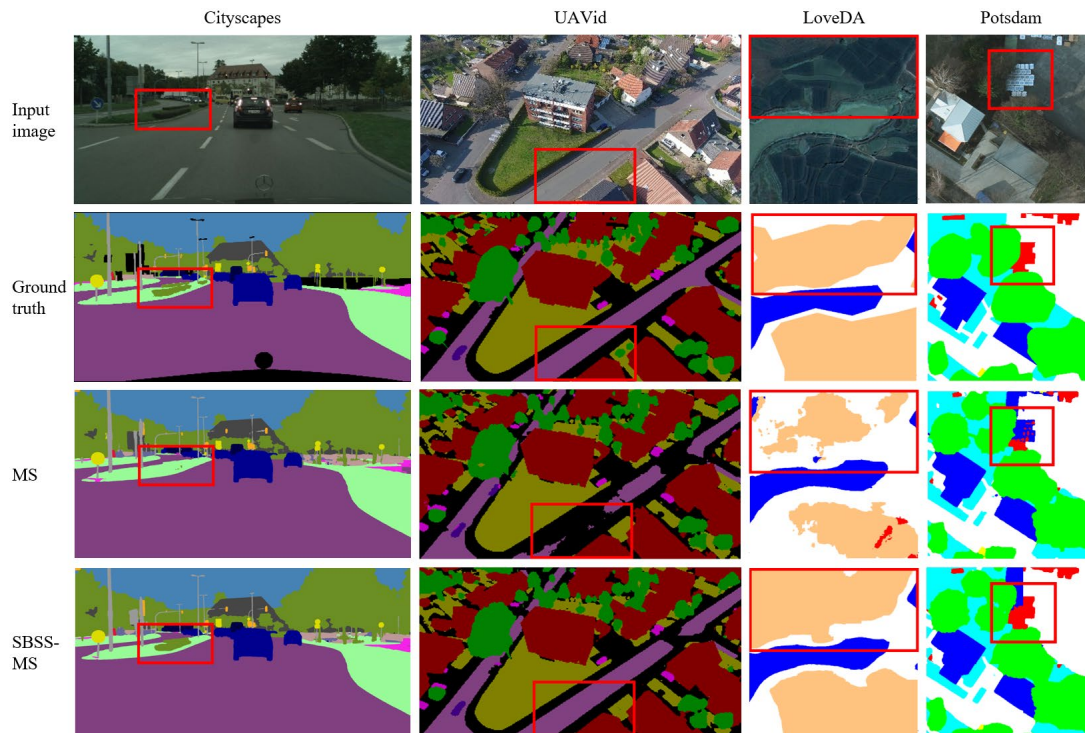


Figure 7-4: Qualitative comparisons between MS and SBSS-MS on the Cityscapes, UAVid, LoveDA and Potsdam validation sets.

7.3.5. Complexity and the speed of the SBSS framework

For a comprehensive comparison of the efficiency of SBSS with the other methods, experiments were conducted on the Cityscapes validation set, the results of which are recorded in Table 7-9. All models in the experiments were implemented using the PyTorch framework. The training time was measured with four NVIDIA GTX 3090 GPUs. The inference speed was measured in terms of the number of tasks (original images rather than single patches) per second, and calculated as the average value of 500 tests on a single NVIDIA GTX 3090 GPU.

For the multi-scale test level comparisons, the SBSS-MS achieved both the highest speed (0.85 Task/s) and the highest accuracy (mIoU of 83.05%).

Chapter 7: Stacking-based semantic segmentation framework

Table 7-9: Comparison of the efficiency of the SBSS framework with other methods on cityscapes validation set.

Method	Backbone	Test method	Patch size	mIoU (%)	Parameters (M)	Flops per patch (G)	Flops per task (G)	Training Time (h)	Task/s
FCN	HRNet-w18	MS	1024×512	79.47	48.98	356.91	13919.49	9.88	0.44
BiSeNetV1	ResNet50	MS	1024×512	79.31	59.24	197.91	7718.49	10.73	0.81
PSPNet	ResNet50	MS	1024×512	80.55	48.98	356.91	13919.49	15.79	0.54
DeepLabV3+	ResNet50	MS	1024×512	81.18	43.59	352.72	13756.08	16.32	0.49
DANet	ResNet50	MS	1024×512	81.09	49.85	398.30	15533.70	15.18	0.48
GCNet	ResNet50	MS	1024×512	80.40	49.63	395.46	15422.94	19.35	0.53
DNLNet	ResNet50	MS	1024×512	80.73	50.02	399.76	15590.64	14.53	0.47
UperNet	ResNet50	MS	1024×512	80.17	66.42	473.65	18472.35	14.17	0.50
UperNet	Swin-T	MS	1024×512	79.99	59.84	469.04	18292.56	14.25	0.48
UperNet	ConvNeXt-T	MS	1024×512	81.32	60.14	467.15	18218.85	18.25	0.53
SBSS-MS	ConvNeXt-T	ECS-MS	1024×512	83.05	60.99	490.51	11281.76	22.18	0.85
FCN	HRNet-w18	SS	1024×512	75.73	48.98	356.91	1427.64	9.88	4.31
BiSeNetV1	ResNet50	SS	1024×512	75.06	59.24	197.91	791.64	10.73	7.91
PSPNet	ResNet50	SS	1024×512	77.90	48.98	356.91	1427.64	15.79	5.25
DeepLabV3+	ResNet50	SS	1024×512	78.66	43.59	352.72	1410.88	16.32	4.74
DANet	ResNet50	SS	1024×512	78.64	49.85	398.30	1593.20	15.18	4.72
GCNet	ResNet50	SS	1024×512	77.68	49.63	395.46	1581.84	19.35	5.14
DNLNet	ResNet50	SS	1024×512	78.31	50.02	399.76	1599.04	14.53	4.55
UperNet	ResNet50	SS	1024×512	77.65	66.42	473.65	1894.60	14.17	4.89
UperNet	Swin-T	SS	1024×512	77.47	59.84	469.04	1876.16	14.25	4.63
UperNet	ConvNeXt-T	SS	1024×512	78.84	60.14	467.15	1868.60	18.25	5.12
SBSS-SS	ConvNeXt-T	ECS-SS	512×256	78.52	60.78	123.79	1485.54	21.53	6.50

In the single scale test level comparisons, the input patch size of the other methods was set to four times the size of the SBSS-SS. The rationale for this setting is as follows. It was noticed that the segmentation accuracy decreased when using smaller patches, as shown in Table 7-7. Meanwhile, Table 7-8 shows that SBSS-SS improved the segmentation accuracy compared to SS. In addition, the memory footprint was proportional to the patch size. For example, with the same network, a memory footprint with a patch size of 512 x 256 is a quarter of the one that uses a patch size of 1024 x 512. Therefore, it is meaningful to test whether SBSS-SS using a smaller patch size can achieve similar accuracy to other methods using a larger patch size. The results in Table 7-9 show that this can be achieved with SBSS-SS, which achieves mIoU (78.52%) that is merely (0.32%) lower than the highest one (78.84%).

Chapter 7: Stacking-based semantic segmentation framework

7.3.6. Quantitative results on Cityscapes, UAVid, LoveDA and Potsdam test sets

To further confirm the effectiveness of the proposed SBSS, an experimental comparison between the SBSS and other state-of-the-art methods was conducted on the test set of each dataset considered. The input patch sizes used for the different datasets were the same as those listed in Table 7-8. Apart from the Potsdam dataset that was evaluated offline, the segmentation results were submitted to the online servers dedicated for other dataset for evaluation, and the performance results are summarised in Table 7-10, Table 7-11, Table 7-12 and Table 7-13.

Because of the strict limitations on the test frequency in the Cityscapes test server, only the segmentation results from the proposed methods (SBSS-MS and SBSS-SS) were submitted for evaluation. The performance results of the other methods on the Cityscapes dataset were taken directly from their original publications. It is worth mentioning that since these methods were obtained with different training sets, backbones, training and test setups, it is probably not very rigorous to simply compare the results in Table 7-10. Nevertheless, those results show that SBSS-MS achieved the highest segmentation accuracy, which confirms the effectiveness of the proposed method.

Table 7-10: Quantitative comparison results on the cityscapes test set. the input patch sizes used in SBSS-MS and SBSS-SS are 1024×512 and 512×256 respectively.

Method	Backbone	Trained on	Test method	mIoU (%)
PSPNet	ResNet101	Train	MS	78.4
BiSeNetV1	ResNet101	Train & Val	SS	78.9
PSANet	ResNet101	Train & Val	MS	80.1
DenseASPP	DenseNet201	Train & Val	MS	80.6
SETR	ViT-L	Train & Val	MS	81.1
Segmenter	ViT-L	Train & Val	MS	81.3
DANet	ResNet101	Train & Val	MS	81.5
HRNet	HRNet-w48	Train & Val	MS	81.6
EANet	ResNet101	Train & Val	MS	81.7
OCR	ResNet101	Train & Val	MS	81.8
DNL	ResNet101	Train & Val	MS	82.0
SegFormer	MiT-B5	Train & Val	MS	82.2
SBSS-SS	ConvNeXt-T	Train & Val	ECS-SS	80.3
SBSS-MS	ConvNeXt-T	Train & Val	ECS-MS	82.6

The UAVid and the LoveDA datasets are less restricted in terms of the test frequency in the test servers. The Potsdam dataset can be tested offline. As such, for fairer comparisons, the segmentation results from the applications of all the methods (including the proposed one and the others) to these three datasets were obtained using the same settings for training and testing. The training and test settings were the same as those used in Section 7.3.1.2 and Section 7.3.4, respectively. The proposed SBSS-MS achieved the highest segmentation accuracy at the multi-scale test level on all three datasets. At the same time, SBSS-SS achieved comparable segmentation accuracy to the other methods using a smaller patch size (i.e., smaller memory footprint) at the single scale test level on all three datasets.

Chapter 7: Stacking-based semantic segmentation framework

Table 7-11: Quantitative comparison results on the uavid test set (%).

Method	Backbone	Test method	Patch size	mIoU	Building	Static Car	Tree	Moving Car	Clutter	Road	Human	Vegetation
FCN	HRNet-w18	MS	1024×540	71.11	89.85	69.08	81.86	78.03	71.80	83.89	28.08	66.30
BiSeNetV1	ResNet50	MS	1024×540	69.03	88.57	65.03	81.47	73.18	69.64	82.20	26.55	65.57
PSPNet	ResNet50	MS	1024×540	69.76	88.96	63.84	81.34	75.84	70.61	82.81	29.37	65.27
DeepLabV3+	ResNet50	MS	1024×540	71.06	89.44	71.10	81.65	77.23	71.01	82.81	29.43	65.84
DANet	ResNet50	MS	1024×540	70.22	89.36	66.43	81.46	75.53	70.96	82.83	29.60	65.59
GCNet	ResNet50	MS	1024×540	69.62	89.24	63.80	81.25	74.87	70.69	82.69	29.34	65.06
DNLNet	ResNet50	MS	1024×540	69.67	88.89	63.74	81.58	75.58	70.42	82.74	28.84	65.58
UperNet	ResNet50	MS	1024×540	70.87	89.35	68.66	81.68	77.30	71.00	83.10	30.20	65.67
UperNet	Swin-tiny	MS	1024×540	70.43	89.17	65.11	81.33	78.16	70.43	82.71	29.66	65.63
UperNet	ConvNeXt-T	MS	1024×540	71.22	89.81	69.57	81.78	77.92	71.14	82.57	30.83	66.11
SBSS-MS	ConvNeXt-T	ECS-MS	1024×540	72.99	91.05	76.00	82.50	78.20	73.06	83.57	31.95	67.59
FCN	HRNet-w18	SS	1024×540	69.34	88.90	66.29	80.45	75.94	69.62	81.91	27.82	63.78
BiSeNetV1	ResNet50	SS	1024×540	67.31	87.32	63.11	80.06	70.26	67.47	80.40	26.62	63.25
PSPNet	ResNet50	SS	1024×540	68.23	88.09	61.47	80.13	73.98	69.04	81.19	28.65	63.25
DeepLabV3+	ResNet50	SS	1024×540	69.57	88.56	68.21	80.43	75.98	69.27	81.22	29.17	63.69
DANet	ResNet50	SS	1024×540	68.61	88.48	63.27	80.31	73.38	69.47	81.56	28.70	63.75
GCNet	ResNet50	SS	1024×540	67.95	88.27	61.03	79.84	72.52	69.00	81.19	29.07	62.71
DNLNet	ResNet50	SS	1024×540	68.14	87.92	60.57	80.35	73.50	68.80	81.47	28.89	63.63
UperNet	ResNet50	SS	1024×540	69.24	88.35	66.18	80.31	75.60	69.23	81.53	29.67	63.03
UperNet	Swin-tiny	SS	1024×540	69.13	88.44	64.04	80.33	76.14	69.14	81.74	29.33	63.89
UperNet	ConvNeXt-T	SS	1024×540	70.05	89.22	67.64	80.91	76.07	69.84	81.34	30.59	64.77
SBSS-MS	ConvNeXt-T	ECS-SS	512×270	70.00	88.19	70.50	81.37	76.08	68.80	82.31	27.50	65.25

Table 7-12: Quantitative comparison results on the loveda test set (%).

Method	Backbone	Test method	Patch size	mIoU	Background	Building	Road	Water	Barren	Forest	Agricultural
FCN	HRNet-w18	MS	512×512	52.74	45.33	59.60	56.26	80.62	17.81	48.92	60.64
BiSeNetV1	ResNet50	MS	512×512	50.46	44.73	55.36	55.52	77.85	14.07	45.80	59.87
PSPNet	ResNet50	MS	512×512	52.43	45.42	57.50	58.96	79.24	17.98	48.66	59.21
DeepLabV3+	ResNet50	MS	512×512	52.55	44.99	56.88	59.35	79.19	18.41	48.83	60.19
DANet	ResNet50	MS	512×512	50.92	44.05	54.15	54.97	77.62	19.33	47.12	59.17
GCNet	ResNet50	MS	512×512	52.76	45.80	58.30	57.94	79.55	18.47	48.50	60.76
DNLNet	ResNet50	MS	512×512	52.59	45.33	57.13	57.59	79.60	19.01	48.23	61.27
UperNet	ResNet50	MS	512×512	52.30	45.44	57.32	59.17	79.16	18.00	47.66	59.38
UperNet	Swin-tiny	MS	512×512	53.22	46.29	58.66	58.86	80.91	17.88	47.88	62.03
UperNet	ConvNeXt-T	MS	512×512	53.57	46.51	60.26	59.95	80.53	17.12	48.14	62.50
SBSS-MS	ConvNeXt-T	ECS-MS	512×512	54.50	46.31	62.35	58.66	82.06	19.59	49.48	63.07
FCN	HRNet-w18	SS	512×512	51.08	43.78	57.56	54.33	78.59	16.95	47.13	59.24
BiSeNetV1	ResNet50	SS	512×512	48.67	42.50	53.38	53.62	76.93	14.03	42.79	57.42
PSPNet	ResNet50	SS	512×512	50.63	44.15	54.72	56.54	76.81	17.49	47.13	57.54
DeepLabV3+	ResNet50	SS	512×512	50.54	43.35	54.40	56.96	76.61	17.65	47.08	57.69
DANet	ResNet50	SS	512×512	49.18	42.18	42.18	58.29	56.23	20.36	48.79	61.78
GCNet	ResNet50	SS	512×512	50.82	44.47	55.55	55.63	77.35	17.54	46.64	58.56
DNLNet	ResNet50	SS	512×512	50.56	43.85	54.22	54.88	77.04	17.51	46.68	59.76
UperNet	ResNet50	SS	512×512	50.27	43.75	54.81	56.58	76.93	17.20	45.75	56.87
UperNet	Swin-tiny	SS	512×512	51.63	44.85	55.96	56.54	80.06	17.87	45.58	60.62
UperNet	ConvNeXt-T	SS	512×512	52.19	44.89	62.05	59.16	79.75	16.74	47.22	55.55
SBSS-MS	ConvNeXt-T	ECS-SS	256×256	52.15	44.69	58.88	58.32	79.14	16.52	46.68	60.82

Table 7-13: Quantitative comparison results on the potsdam test set (%).

Method	Backbone	Test method	Patch size	mIoU	Impervious surface	Building	Low vegetation	Tree	Car
FCN	HRNet-w18	MS	750×750	86.79	88.10	94.06	78.85	80.78	92.17
BiSeNetV1	ResNet50	MS	750×750	86.37	87.74	93.53	78.43	80.24	91.90
PSPNet	ResNet50	MS	750×750	85.81	87.26	93.37	77.14	79.59	91.68
DeepLabV3+	ResNet50	MS	750×750	86.85	88.20	94.16	78.37	80.79	92.75
DANet	ResNet50	MS	750×750	86.85	87.87	93.76	78.53	80.99	93.09
GCNet	ResNet50	MS	750×750	86.91	88.35	93.83	78.52	80.95	92.90
DNLNet	ResNet50	MS	750×750	86.76	88.18	93.75	78.08	80.89	92.91
UperNet	ResNet50	MS	750×750	86.84	88.27	93.87	78.52	80.96	92.56
UperNet	Swin-tiny	MS	750×750	87.01	88.52	94.31	79.06	81.16	92.01
UperNet	ConvNeXt-T	MS	750×750	87.34	88.82	94.58	79.27	81.60	92.44
SBSS-MS	ConvNeXt-T	ECS-MS	750×750	87.68	88.95	94.88	79.42	81.82	93.34
FCN	HRNet-w18	SS	750×750	85.49	87.30	92.97	77.32	79.45	90.43
BiSeNetV1	ResNet50	SS	750×750	84.66	86.23	92.34	76.54	78.11	90.06
PSPNet	ResNet50	SS	750×750	85.81	87.26	93.37	77.14	79.59	91.68
DeepLabV3+	ResNet50	SS	750×750	85.68	87.32	93.43	76.83	79.33	91.51
DANet	ResNet50	SS	750×750	86.01	87.49	93.48	77.16	79.86	92.05
GCNet	ResNet50	SS	750×750	85.82	87.47	93.24	77.04	79.67	91.70
DNLNet	ResNet50	SS	750×750	85.61	87.21	93.10	76.35	79.54	91.84
UperNet	ResNet50	SS	750×750	85.62	87.28	92.90	77.09	79.64	91.17
UperNet	Swin-tiny	SS	750×750	86.07	87.88	93.75	77.92	79.99	90.81
UperNet	ConvNeXt-T	SS	750×750	86.41	88.07	94.04	78.15	80.52	91.27
SBSS-SS	ConvNeXt-T	ECS-SS	375×375	86.35	87.93	93.82	78.06	80.55	91.40

7.3.7. Qualitative Analysis of the Segmentation Results

As introduced in Section 7.1, the objective of the proposed SBSS is to reduce the error associated with the resizing scales. While the efficiency and the effectiveness of SBSS were demonstrated in Section 7.3.4, Section 7.3.5 and Section 7.3.6, it is informative to appreciate what kind of errors associated with the resizing scales were corrected by SBSS. The class building was chosen for analysis because the segmentation of buildings is crucial for many applications and this class happens to be present in all four datasets used.

The top four plots of Figure 7-5 show the segmentation accuracy (IoU) of buildings for the four datasets when they were tested with SS at different resizing scales. It was observed that the segmentation accuracies of buildings in the Cityscapes, UAVid and Potsdam datasets were generally higher when smaller resizing scales were used. This

Chapter 7: Stacking-based semantic segmentation framework

is consistent with the fact that these datasets were collected in urban scenes, where buildings are expected to be relatively large objects. However, as the LoveDA dataset includes many images of rural scenes where buildings are comparatively smaller objects than forests and water bodies, a larger resizing scale was favourable.

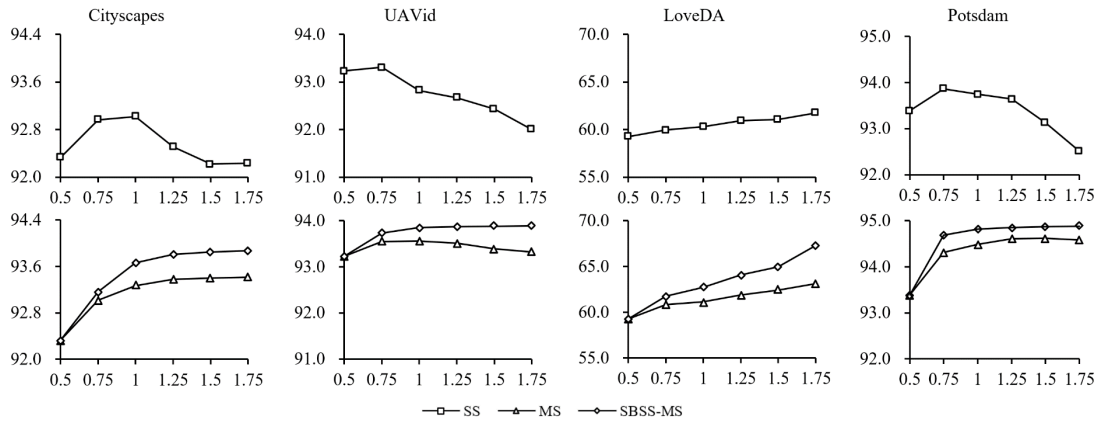


Figure 7-5: Segmentation accuracy (IoU %) of buildings using different resizing scales. The horizontal coordinates (of the top 4 plots) for SS refer to the resizing scale used. The horizontal coordinates (of the bottom 4 plots) for MS and SBSS-MS represent the utilisation of all the scales that are equal to and smaller than the current scale (e.g., the coordinate value 1 represents the case where the scales 0.5, 0.75 and 1 were all used).

In addition, the segmentation accuracies of the MS and SBSS-MS were tested as follows. The scale started with the smallest one (i.e., 0.5), followed by a stepwise increase (with an increment of 0.25 each time) until the largest scale of 1.75 was reached. In each test, all the scales that are equal to or smaller than a particular scale were used. For example, for the scale 1, the following scales 0.5, 0.75 and 1 were used. The results are shown in the bottom four plots of Figure 7-5, suggesting that the accuracy of MS did not always increase when more and larger scales were used. For example, the accuracy of MS on the UAVid and Potsdam datasets decreased after using scales larger than 0.75 and 1.5, respectively. To investigate the causes of this phenomenon, the segmentation results of a UAVid image using SS at six resizing scales and those for MS and SBSS-MS using the scales from 0.5 to 1.75 are plotted in Figure 7-6.

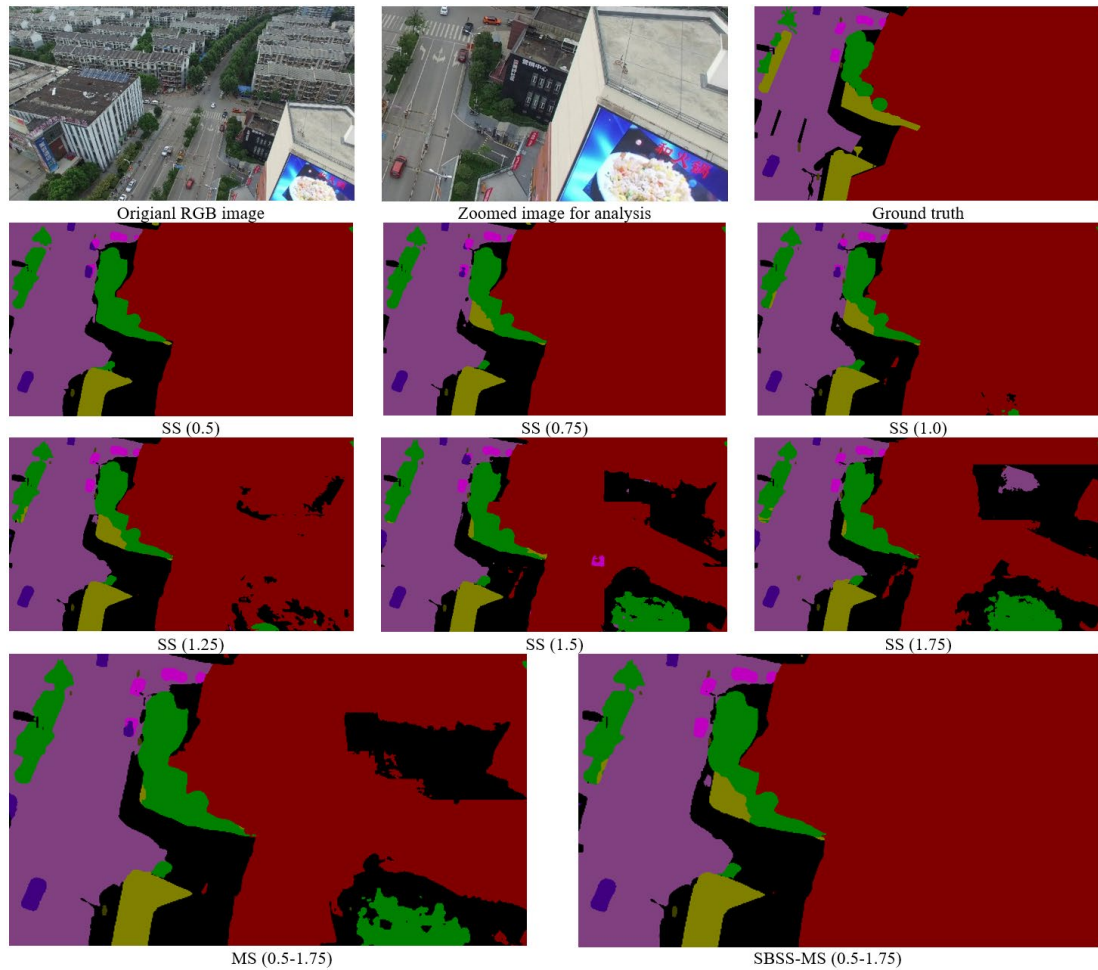


Figure 7-6: Visual comparisons between SS, MS and SBSS-MS on the UAVid validation set. The numbers in brackets represent the resizing scales used.

As shown in Figure 7-6, when the scale increased, SS yielded more fragmented and erroneous results for the large building on the right. The likely reasons for this are presented in the following. First, the use of a large resizing scale introduced additional difficulties for the global context information modelling. Second, the complex building facade in Figure 7-6 made correct segmentation more dependent on modelling global information rather than local one. Since MS never learns at which scale the segmentation results are more reliable for the building, the final results obtained using the average voting inevitably inherit some of the errors in the segmentation results at relatively large scales. In contrast, SBSS-MS preserved the correct segmentation results for buildings in a more intelligent way.

Chapter 7: Stacking-based semantic segmentation framework

Apart from this type of erroneous segmentation of a large area of the building, there is another common type of error that occurs at the edges of the building. To verify whether SBSS-MS can handle this type of error, a comparison was made between results generated by SBSS-MS using two different sets of resizing scales (i.e., 0.5-1.0 and 0.5-1.75), and the results are shown in Figure 7-7. It shows that using the SS results at three larger scales (i.e., 1.25, 1.5 and 1.75), SBSS-MS improved the segmentation accuracy at the building edges.

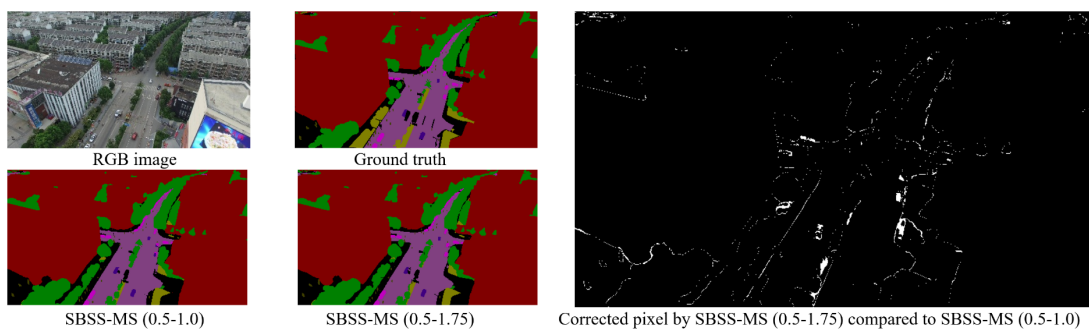


Figure 7-7: Visual comparisons between SBSS-MS using different set of resizing scales on the UAVid validation set.

Also, it is worth mentioning that although building was used as the example class for the demonstration, the aforementioned two kinds of improvements are expected to be applicable to the other classes. For example, SBSS-MS corrected the car that was mis-segmented in MS (the top left part in Figure 7-6). The segmentation improvement at the object edges in Figure 7-7 was also valid for other classes such as tree and road.

7.3.8. Exploring the potential for higher accuracy

The computational complexity of SBSS-MS has been strictly limited in the previous sections (i.e., Section 7.3.4, Section 7.3.5 and Section 7.3.6) to facilitate a fair comparison with the other methods. However, it is of interest to discover the accuracy that can be achieved by SBSS-MS when adequate computational resources are available.

The results in Table 7-7 show that the segmentation accuracy decreased considerably when smaller patch sizes were used. Therefore, this study tested the case of directly performing the segmentation on the entire image in the first place. As the memory of the GPU used is limited, only the LoveDA dataset, which has a relatively small original image size, was tested. In addition, on the basis of using the whole image for segmentation, the case of using more resizing scales (i.e., 0.5-2.0) was tested.

The quantitative results of the test are shown in Table 7-14, suggesting that using the whole image for segmentation did improve the segmentation accuracy slightly (0.21% in mIoU), but using more scales was a more effective way (1.09% in mIoU) in comparison. The segmentation results for the second and third settings in Table 7-14 show that by using larger resizing scales SBSS-MS improved the segmentation accuracy for all classes. Meanwhile, Figure 7-8 suggests that the improvement was mainly at the edges of the objects.

Table 7-14: Quantitative comparison for SBSS-MS using different input methods on the LoveDA test set (%).

Method	Patch size	Scales used	mIoU	Background	Building	Road	Water	Barren	Forest	Agricultural
SBSS-MS	512×512	0.5-1.5	54.50	46.31	62.35	58.66	82.06	19.59	49.48	63.07
SBSS-MS	Entire image	0.5-1.5	54.71	46.92	62.20	58.17	82.24	19.78	49.41	64.22
SBSS-MS	Entire image	0.5-2.0	55.59	48.30	62.85	58.22	83.06	21.02	49.76	65.93

Chapter 7: Stacking-based semantic segmentation framework

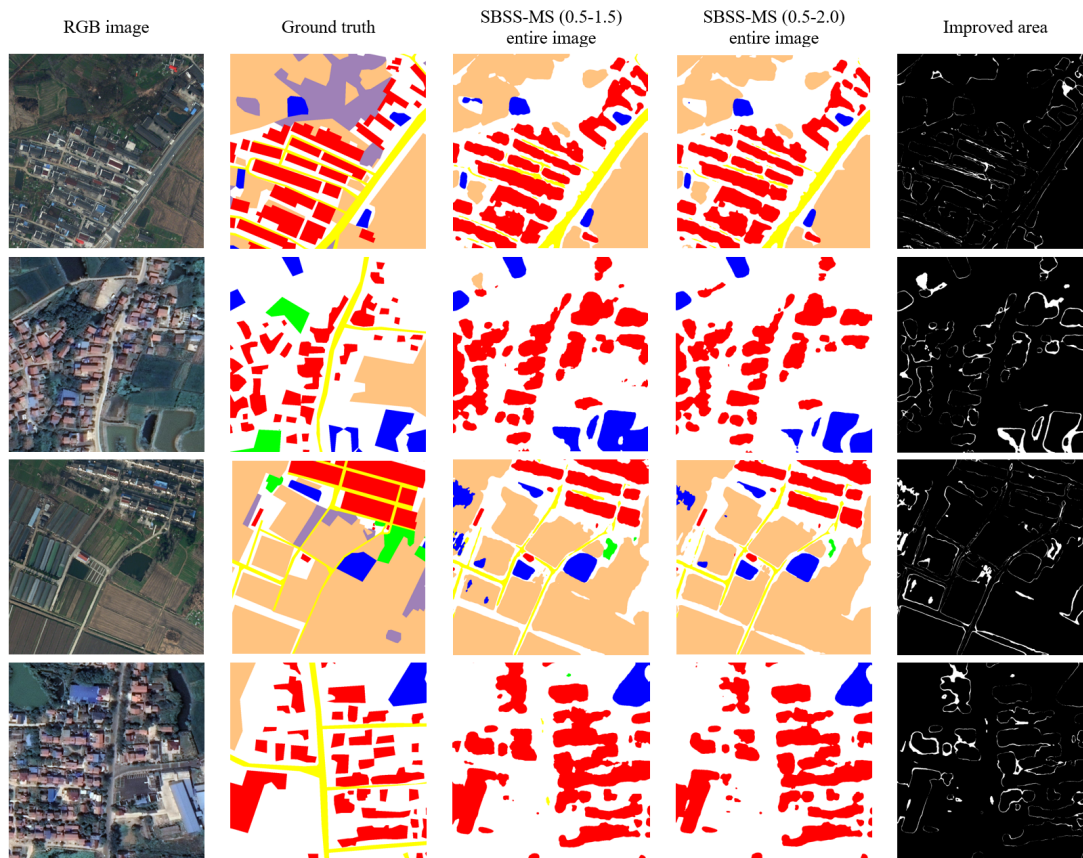


Figure 7-8: Qualitative comparisons between SBSS-MS with different setting on the LoveDA validation set.

7.3.9. Integrating developed methods for the semantic segmentation of TLS point clouds (additional results to the published version)

The semantic segmentation accuracy that can be achieved by integrating the methods developed in this thesis was tested in this section. More specifically, the SBSS-MS developed in this chapter was used along with the OSTA developed in Chapter 6. The training strategies used are summarised in Table 7-15. The combination of features selected by the integration method is blue, enhanced z-coordinate image and enhanced depth image. The quantitative comparison between the integration method and other methods is shown in Table 7-16. The integrated method achieved the highest accuracy among image-based methods and is very close to the accuracy of the SOTA 3D

methods. The processing time of the integrated method was less than one tenth of that of the SOTA 3D method (SCF-Net).

Table 7-15: Training setup for the integrated method.

Input feature channel	Red; Green; Blue; Intensity; Z-coordinate image; Depth image; Enhanced z-coordinate image; Enhanced depth image
Dimension reduction method	OSTA
Scales used in SBSS-MS	0.5, 0.75, 1.0, 1.25, 1.5, 1.75
Patch size	1200×600
Total training iterations	80 k
Pretraining dataset	Cityscapes and ImageNet-1k
Optimizer	AdamW
Initial learning rate	10 ⁻⁴
Learning rate schedule	Poly learning rate policy with a power of 1.0
Minimum Learning rate	Zero
Warmup ratio	10 ⁻⁶
Weight decay	0.05
Batch size	16
Loss function	Cross entropy
Data augmentation	Random cropping, random resize (0.5~2), random horizontal flipping, photo metric distortion

Table 7-16: Quantitative comparison of different methods on Semantic3D (Reduced-8) (%).

	Time (s)	Params (M)	mIoU	OA	man-made	natural	high veg	low veg	buildings	hard scape	Scanning art	cars	
3D methods	RF MSSF (Thomas et al., 2018)	1643.75	-	62.7	90.3	87.6	80.3	81.8	36.4	92.2	24.1	42.6	56.6
	ShellNet (Zhang et al., 2019)	3000	0.48	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
	OctreeNet (F. Wang et al., 2020)	184.84	-	59.1	89.9	90.7	82.0	82.4	39.3	90.0	10.9	31.2	46.0
	GACNet (L. Wang et al., 2019)	1380	-	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
	SPGraph (Landrieu and Simonovsky, 2018)	3000	0.25	73.2	94.0	97.4	92.6	87.9	44.0	83.2	31.0	63.5	76.2
	KPCnv (Thomas et al., 2019)	600	14.9	74.6	92.9	90.9	82.2	84.2	47.9	94.9	40.0	77.3	79.7
	RandLA-Net (Q. Hu et al., 2020)	-	0.95	77.4	94.8	95.6	91.4	86.6	51.5	95.7	51.5	69.8	76.8
	SCF-Net (Fan et al., 2021)	563.6	-	77.6	94.7	97.1	91.8	86.3	51.2	95.3	50.5	67.9	80.7
RFCR (Gong et al., 2021)	-	-	77.8	94.3	94.2	89.1	85.7	54.4	95.0	43.8	76.2	83.7	
Image-based methods	DeePr3SS (Lawin et al., 2017)	-	134	58.5	88.9	85.6	83.2	74.2	32.4	89.7	18.5	25.1	59.2
	SnapNet (Boulch et al., 2018)	3600	29	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4
	XJTLU (Cai et al., 2021b)	5.13	70.6	63.5	89.4	85.4	74.4	74.6	31.9	93.0	25.2	41.5	82.0
	HR-EHNet (Cai et al., 2022a)	11.72	73.6	74.2	92.1	85.1	75.5	89.6	55.9	95.5	50.8	48.3	92.5
	Integration of developed methods	52.64	61.0	76.6	93.8	87.3	76.3	88.1	59.6	95.1	54.4	59.2	92.9

7.4. Future work

In this study, the final segmentation result is obtained by fusing the segmentation results at a predefined set of resizing. However, the results in Table 7-6 indicate that the optimal scale for each class varies from case to case. Therefore, the performance of SBSS could be further improved by developing algorithms that can adaptively select a set of scales for analyses. Similarly, it can be speculated that such adaptive algorithms would also be useful for choosing how many patches to analyse at each scale. In the course of this study, it was noticed that the error distribution has a pattern not only in the dimension of resizing scales, but also in the spatial dimension. For example, errors are more likely to occur at the edge of a patch. This characteristic could be taken into consideration in future study.

In addition to an improvement of the SBSS framework itself, future research could consider combining the SBSS framework with other existing methods such as edge-aware segmentation (Jung et al., 2022; A. Li et al., 2022; Marmanis et al., 2018; Zheng et al., 2020) or object-based segmentation (Yuan et al., 2020; C. Zhang et al., 2018). When visually inspecting the differences between ground truth, MS and SBSS-MS segmentation results (e.g., Figure 7-4 and Figure 7-8), it was found that although SBSS-MS could obtain more accurate segmentation results than MS, there are still cases where an object is segmented into pieces. Therefore, integration with those studies (e.g., (Jung et al., 2022; A. Li et al., 2022; Marmanis et al., 2018; Yuan et al., 2020; C. Zhang et al., 2018; Zheng et al., 2020)) that specifically address this issue may further improve segmentation accuracy.

7.5. Summary

This study experimentally demonstrated that different classes in images have their preferred resizing scales for semantic segmentation. On this basis, the SBSS framework was proposed, which uses a learnable ECM to fuse segmentation results that are more likely to be correct at each resizing scale, and an ECS to control the computational complexity. Extensive experiments were conducted on the four benchmark datasets considered, i.e., Cityscapes, UAVid, LoveDA and Potsdam datasets. The results show that SBSS achieved promising performances in the various scenarios considered. Specifically, SBSS-MS achieved a higher segmentation accuracy with less Flops, faster speed, and similar memory footprint compared to MS. Meanwhile, SBSS-SS achieved a similar segmentation accuracy with a quarter of the memory footprint, similar Flops and speed compared to SS. In the future, more sophisticated ECS and ECM can be proposed to further improve the performance of SBSS or to adapt it to specific application requirements.

Chapter 8: Conclusion

Terrestrial laser scanning has widely been used for high precision 3D large-scale scene recording. Semantic segmentation of TLS point clouds is the basis of intelligent development of many applications. However, existing point cloud semantic segmentation methods are either inferior in accuracy (image-based methods) or in efficiency (3D methods). This thesis improves the semantic segmentation accuracy of image-based methods on TLS point clouds while preserving relatively high efficiency by developing image-based geometric features, an automatic feature selection method, and a stacking-based semantic segmentation framework. The image-based semantic segmentation method with the improvement techniques developed can achieve an accuracy comparable to the state-of-the-art 3D methods and only requires less than a tenth of the processing time of the fastest 3D method (Fan et al., 2021).

8.1. Key results

This section presents the key results with respect to each of the research objectives and questions described in Section 1.2, as well as additional important results.

Objective 1. Select the optimal feature combination from commonly used image-based features for semantic segmentation of TLS point clouds.

Key results 1-1: It was found the segmentation accuracies achieved using appropriate feature combinations were significantly higher than using all available features. This highlights the importance of feature selection for image-based semantic segmentation of TLS point clouds.

Key results 1-2: It was found that the optimal feature combination was robust to different network structures. Based on this, an efficient manual feature selection

Chapter 8: Conclusion

method was developed in Chapter 3, which uses a lightweight network for feature selection before fine-tuning a more complex network.

Key results 1-3a: It was found that different optimal feature combinations might exist when different accuracy metrics (i.e., IRGB for OA and IRGBD for mIoU) were used. This indicates that the optimal feature combination has its corresponding preconditions.

Key results 1-4a: It was noticed that existing image-based methods relied heavily on colour information for semantic segmentation. This indicates that existing image-based methods do not fully exploit the geometric information.

Objective 2. Develop novel image-based geometric features to improve segmentation accuracy of TLS point clouds.

Key results 2-1: Locally enhanced image-based geometric features were developed in Chapter 4. It was observed that using the image-based geometric features together with the optimal feature combinations selected in Chapter 3 could significantly improve the segmentation accuracy compared to using these optimal feature combinations alone.

Key results 2-2: It was also shown that using only image-based geometric features and intensity feature (i.e., $IZ_e D_e$) achieved the highest segmentation accuracy among all the feature combinations tested. To the best of the author's knowledge, this is the first image-based method that can achieve even higher segmentation accuracy without using colour information.

Key results 2-3: It was noticed that using a 3-feature combination as input had a unique advantage in that the segmentation accuracy can be further improved by retaining the weights within the first layer of the pre-trained model.

Objective 3. Develop novel dimension reduction methods to transform multichannel images into 3-channel images (i.e., containing 3 features) to better utilise model(s) pre-trained on large-scale RGB image datasets.

Key results 3-1: For the multispectral dataset RIT-18, which has 18 classes and relatively reliable feature channels, using appropriate 3-feature combinations can still achieve significantly higher accuracies than that using all features. This was concluded from a fair comparison that the parameters of the first layer of the pre-trained models were randomly initialized for all tests. Given the Key results 2-3 and Key result 3-1, together with the fact that the existing models are all pre-trained with 3-channel RGB datasets (mainly ImageNet), it is reasonable to set the goal of dimension reduction as obtaining three feature dimension.

Key results 3-2: An end-to-end feature extraction method LC-Net was developed in Chapter 5, which learns to compress adjacent features through iterative training to obtain a combination of three new features. The semantic segmentation accuracy of LC-Net is comparable to the optimal 3-channel combination for RIT-18 obtained by exhaustive trial and error.

Key results 3-3: A one-shot task-adaptive channel selection method (OSTA) was developed in Chapter 6, which formulates channel selection as a pruning process for a supernet. The outcomes of six groups of experiments (L7Irish3C, L7Irish2C, L8Biome3C, L8Biome2C, RIT-18 and Semantic3D) demonstrated the effectiveness and efficiency of OSTA. OSTA achieved the highest segmentation accuracies in all tests (62.49% (mIoU), 75.40% (mIoU), 68.38% (mIoU), 87.63% (mIoU), 66.53% (mA) and 70.86% (mIoU), respectively). It even exceeded the highest accuracies of exhaustive tests (61.54% (mIoU), 74.91% (mIoU), 67.94% (mIoU), 87.32% (mIoU), 65.32% (mA) and 70.27% (mIoU), respectively), where

Chapter 8: Conclusion

all possible channel combinations were tested. All of this can be accomplished within a predictable and relatively efficient timeframe, ranging from 101.71% to 298.1% times the time required to train the segmentation network alone.

Key results 3-4: For all the tests, OSTA achieved semantic segmentation accuracies higher than LC-Net.

Key results 3-5a: Training the semantic segmentation network with extra feature combinations in the early stage can improve the final accuracy.

Key results 3-6a: The optimal feature combination can be influenced by the parameter initialization method used. However, it was found that there were robust feature combinations that performed well with all initialization methods tested.

Key results 3-7a: The coastal aerosol band has been neglected in the past research for cloud detection but turns out to be an important channel for cloud detection according to this study.

Objective 4. Develop a novel image semantic segmentation framework to improve segmentation accuracy.

Key results 4-1: A stacking-based semantic segmentation framework (SBSS) was developed in Chapter 7, which can improve the segmentation accuracy by learning the preferred resizing scales for different object classes.

Key results 4-2: By integrating the methods developed in this thesis, the semantic segmentation accuracy of image-based methods on TLS point clouds has been raised to an unprecedented level. The improved image-based method achieved a 76.6% mIoU and 93.8% OA on the Semantic3D benchmark (reduced8) with a total processing time of only 52.64 s.

8.2. Future work

Based on the research presented in this thesis, the following key recommendations are made for future work.

Recommendation 1: It is recommended to establish a much larger dataset than the existing one (Semantic3D) to facilitate the development of this field.

Recommendation 2: It is recommended to develop TLS technology that acquires all information simultaneously to avoid inconsistent representation of objects by different modalities.

Recommendation 3: The aim of this thesis is to improve the accuracy of image-based methods while maintaining high efficiency. It is also worth investigating from the opposite direction, i.e. to improve the efficiency of the 3D method while maintaining high accuracy. The key to successfully conducting this study may lie in finding the fundamental differences between image-based and 3D methods.

Recommendation 4: It is recommended to develop pre-trained models that are trained using multichannel images. It is expected that this will substantially improve the accuracy for downstream tasks using multichannel images.

Recommendation 5: On the basis of the realization of Recommendation 4, it is recommended to develop feature selection methods that can automatically determine the optimal number of features.

References

- Abdalla, A., Cen, H., Abdel-Rahman, E., Wan, L., He, Y., 2019. Color Calibration of Proximal Sensing RGB Images of Oilseed Rape Canopy via Deep Learning Combined with K-Means Algorithm. *Remote Sens.* 11, 3001. <https://doi.org/10.3390/rs11243001>
- Alshawabkeh, Y., 2020. Linear feature extraction from point cloud using color information. *Herit. Sci.* 8, 28. <https://doi.org/10.1186/s40494-020-00371-6>
- Anees, A., Aryal, J., 2014. A Statistical Framework for Near-Real Time Detection of Beetle Infestation in Pine Forests Using MODIS Data. *IEEE Geosci. Remote Sens. Lett.* 11, 1717–1721. <https://doi.org/10.1109/LGRS.2014.2306712>
- Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2018. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1437–1451. <https://doi.org/10.1109/TPAMI.2017.2711011>
- Archibald, R., Fann, G., 2007. Feature Selection and Classification of Hyperspectral Images With Support Vector Machines. *IEEE Geosci. Remote Sens. Lett.* 4, 674–677. <https://doi.org/10.1109/LGRS.2007.905116>
- Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1534–1543. <https://doi.org/10.1109/CVPR.2016.170>
- Aryal, J., Neupane, B., 2023. Multi-Scale Feature Map Aggregation and Supervised Domain Adaptation of Fully Convolutional Networks for Urban Building Footprint Extraction. *Remote Sens.* 15, 488. <https://doi.org/10.3390/rs15020488>

References

- Aryal, J., Sitaula, C., Aryal, S., 2022. NDVI Threshold-Based Urban Green Space Mapping from Sentinel-2A at the Local Governmental Area (LGA) Level of Victoria, Australia. *Land* 11, 351. <https://doi.org/10.3390/land11030351>
- Badenko, V., Fedotov, A., Zotov, D., Lytkin, S., Volgin, D., Garg, R.D., Min, L., 2019. Scan-to-bim methodology adapted for different application, in: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. International Society for Photogrammetry and Remote Sensing, pp. 1–7. <https://doi.org/10.5194/isprs-archives-XLII-5-W2-1-2019>
- Bandos, T.V., Bruzzone, L., Camps-Valls, G., 2009. Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis. *IEEE Trans. Geosci. Remote Sens.* 47, 862–873. <https://doi.org/10.1109/TGRS.2008.2005729>
- Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences, in: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 9296–9306. <https://doi.org/10.1109/ICCV.2019.00939>
- Belkin, M., Niyogi, P., 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* 15, 1373–1396. <https://doi.org/10.1162/089976603321780317>
- Bernat, M., 2014. STUDIES ON THE USE OF TERRESTRIAL LASER SCANNING IN THE MAINTENANCE OF BUILDINGS BELONGING TO THE CULTURAL HERITAGE, in: *International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM*. pp. 307–318. <https://doi.org/10.5593/SGEM2014/B23/S10.039>

- Bhatti, U.A., Yu, Z., Chanussot, J., Zeeshan, Z., Yuan, L., Luo, W., Nawaz, S.A., Bhatti, M.A., Ain, Q. ul, Mehmood, A., 2022. Local Similarity-Based Spatial–Spectral Fusion Hyperspectral Image Classification With Deep CNN and Gabor Filtering. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. <https://doi.org/10.1109/TGRS.2021.3090410>
- Bhuiyan, M.A.E., Witharana, C., Liljedahl, A.K., Jones, B.M., Daanen, R., Epstein, H.E., Kent, K., Griffin, C.G., Agnew, A., 2020. Understanding the Effects of Optimal Combination of Spectral Bands on Deep Learning Model Predictions: A Case Study Based on Permafrost Tundra Landform Mapping Using High Resolution Multispectral Satellite Imagery. *J. Imaging* 6, 97. <https://doi.org/10.3390/jimaging6090097>
- Borse, S., Wang, Y., Zhang, Y., Porikli, F., 2021. InverseForm: A Loss Function for Structured Boundary-Aware Segmentation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 5897–5907. <https://doi.org/10.1109/CVPR46437.2021.00584>
- Boulch, A., Guerry, J., Le Saux, B., Audebert, N., 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Comput. Graph.* 71, 189–198. <https://doi.org/10.1016/j.cag.2017.11.010>
- Boulch, A., Le Saux, B., Audebert, N., 2017. Unstructured point cloud semantic labeling using deep segmentation networks, in: Eurographics Workshop on 3D Object Retrieval, EG 3DOR. pp. 17–24. <https://doi.org/10.2312/3dor.20171047>
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24, 123–140. <https://doi.org/10.1007/BF00058655>

References

- Bühlmann, P., Yu, B., 2002. Analyzing bagging. *Ann. Stat.* 30, 927–961.
<https://doi.org/10.1214/aos/1031689014>
- Cai, Y., Fan, L., 2021. An Efficient Approach to Automatic Construction of 3D Watertight Geometry of Buildings Using Point Clouds. *Remote Sens.* 13, 1947.
<https://doi.org/10.3390/rs13101947>
- Cai, Y., Fan, L., Atkinson, P.M., Zhang, C., 2022a. Semantic Segmentation of Terrestrial Laser Scanning Point Clouds Using Locally Enhanced Image-Based Geometric Representations. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15.
<https://doi.org/10.1109/TGRS.2022.3161982>
- Cai, Y., Fan, L., Fang, Y., 2023. SBSS: Stacking-Based Semantic Segmentation Framework for Very High-Resolution Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.* 61, 1–14. <https://doi.org/10.1109/TGRS.2023.3234549>
- Cai, Y., Fan, L., Zhang, C., 2022b. Semantic Segmentation of Multispectral Images via Linear Compression of Bands: An Experiment Using RIT-18. *Remote Sens.* 14, 2673. <https://doi.org/10.3390/rs14112673>
- Cai, Y., Fan, L., Zhang, C., 2021a. An overview of constructing geometric models of buildings using point clouds, in: Zhang, Y., Zhang, D. (Eds.), 2021 International Conference on Image, Video Processing, and Artificial Intelligence. SPIE, p. 26.
<https://doi.org/10.1117/12.2611685>
- Cai, Y., Huang, H., Wang, K., Zhang, C., Fan, L., Guo, F., 2021b. Selecting Optimal Combination of Data Channels for Semantic Segmentation in City Information Modelling (CIM). *Remote Sens.* 13, 1367. <https://doi.org/10.3390/rs13071367>
- Cai, Y., Liu, X., Cai, Z., 2020. BS-Nets: An End-to-End Framework for Band Selection of Hyperspectral Image. *IEEE Trans. Geosci. Remote Sens.* 58, 1969–1984. <https://doi.org/10.1109/TGRS.2019.2951433>

- Cao, X., Xiong, T., Jiao, L., 2016. Supervised Band Selection Using Local Spatial Information for Hyperspectral Image. *IEEE Geosci. Remote Sens. Lett.* 13, 1–5. <https://doi.org/10.1109/LGRS.2015.2511186>
- Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2019. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, pp. 1971–1980. <https://doi.org/10.1109/ICCVW.2019.00246>
- Cao, Z., Chen, D., Peethambaran, J., Zhang, Z., Xia, S., Zhang, L., 2021. Tunnel Reconstruction With Block Level Precision by Combining Data-Driven Segmentation and Model-Driven Assembly. *IEEE Trans. Geosci. Remote Sens.* 59, 8853–8872. <https://doi.org/10.1109/TGRS.2020.3046624>
- Chang, C.-I., 2022. Band Sampling for Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* 60, 1–24. <https://doi.org/10.1109/TGRS.2021.3102861>
- Chang, C.-I., Ma, K.Y., 2022. Band Sampling of Kernel Constrained Energy Minimization Using Training Samples for Hyperspectral Mixed Pixel Classification. *IEEE Trans. Geosci. Remote Sens.* 60, 1–21. <https://doi.org/10.1109/TGRS.2022.3146803>
- Chang, C.I., Chein-I Chang, Qian Du, Tzu-Lung Sun, Althouse, M.L.G., 1999. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* 37, 2631–2641. <https://doi.org/10.1109/36.803411>
- Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J., 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 77–85. <https://doi.org/10.1109/CVPR.2017.16>

References

- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018a. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. <https://doi.org/10.1109/TPAMI.2017.2699184>
- Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
- Chen, L.-Z., Li, X.-Y., Fan, D.-P., Wang, K., Lu, S.-P., Cheng, M.-M., 2019. LSA-Net: Feature Learning on Point Sets by Local Spatial Aware Layer. *arXiv*.
- Chen, S., Li, W., Cao, Y., Lu, X., 2022. Combining the Convolution and Transformer for Classification of Smoke-Like Scenes in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–19. <https://doi.org/10.1109/TGRS.2022.3208120>
- Chen, Yuhong, Peng, W., Tang, K., Khan, A., Wei, G., Fang, M., 2022. PyraPVConv: Efficient 3D Point Cloud Perception with Pyramid Voxel Convolution and Sharable Attention. *Comput. Intell. Neurosci.* 2022, 1–9. <https://doi.org/10.1155/2022/2286818>
- Chen, Yang, Weng, Q., Tang, L., Liu, Q., Fan, R., 2022a. An Automatic Cloud Detection Neural Network for High-Resolution Remote Sensing Imagery With Cloud–Snow Coexistence. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. <https://doi.org/10.1109/LGRS.2021.3102970>

- Chen, Yang, Weng, Q., Tang, L., Zhang, X., Bilal, M., Li, Q., 2022b. Thick Clouds Removing From Multitemporal Landsat Images Using Spatiotemporal Neural Networks. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. <https://doi.org/10.1109/TGRS.2020.3043980>
- Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Institute of Electrical and Electronics Engineers Inc., pp. 1800–1807. <https://doi.org/10.1109/CVPR.2017.195>
- Choy, C., Gwak, J., Savarese, S., 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks, in: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3070–3079. <https://doi.org/10.1109/CVPR.2019.00319>
- Chu, X., Zhang, B., Xu, R., 2021. FairNAS: Rethinking Evaluation Fairness of Weight Sharing Neural Architecture Search, in: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 12219–12228. <https://doi.org/10.1109/ICCV48922.2021.01202>
- Chu, X., Zhou, T., Zhang, B., Li, J., 2020. Fair DARTS: Eliminating Unfair Advantages in Differentiable Architecture Search, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Science and Business Media Deutschland GmbH, pp. 465–480. https://doi.org/10.1007/978-3-030-58555-6_28
- Contreras, J., Denzler, J., 2019. Edge-Convolution Point Net for Semantic Segmentation of Large-Scale Point Clouds, in: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, pp. 5236–5239. <https://doi.org/10.1109/IGARSS.2019.8899303>

References

- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3213–3223. <https://doi.org/10.1109/CVPR.2016.350>
- Cover, T.M., 1974. The Best Two Independent Measurements Are Not the Two Best. IEEE Trans. Syst. Man. Cybern. SMC-4, 116–117. <https://doi.org/10.1109/TSMC.1974.5408535>
- Datta, A., Ghosh, S., Ghosh, A., 2015. Combination of Clustering and Ranking Techniques for Unsupervised Band Selection of Hyperspectral Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 2814–2823. <https://doi.org/10.1109/JSTARS.2015.2428276>
- Dechesne, C., Mallet, C., Le Bris, A., Gouet-Brunet, V., 2017. SEMANTIC SEGMENTATION OF FOREST STANDS OF PURE SPECIES AS A GLOBAL OPTIMIZATION PROBLEM. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-1/W1, 141–148. <https://doi.org/10.5194/isprs-annals-IV-1-W1-141-2017>
- Demir, B., Ertürk, S., 2008. Phase correlation based redundancy removal in feature weighting band selection for hyperspectral images. Int. J. Remote Sens. 29, 1801–1807. <https://doi.org/10.1080/01431160701802471>
- Deng, Y.-J., Li, H.-C., Pan, L., Shao, L.-Y., Du, Q., Emery, W.J., 2018. Modified Tensor Locality Preserving Projection for Dimensionality Reduction of Hyperspectral Images. IEEE Geosci. Remote Sens. Lett. 15, 277–281. <https://doi.org/10.1109/LGRS.2017.2786223>
- Ding, L., Zhang, J., Bruzzone, L., 2020. Semantic Segmentation of Large-Size VHR Remote Sensing Images Using a Two-Stage Multiscale Training Architecture.

- IEEE Trans. Geosci. Remote Sens. 58, 5367–5376.
<https://doi.org/10.1109/TGRS.2020.2964675>
- Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., Sun, J., 2022. Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
<https://doi.org/10.48550/arxiv.2203.06717>
- Dong, Z., Yang, B., Hu, P., Scherer, S., 2018. An efficient global energy optimization approach for robust 3D plane segmentation of point clouds. ISPRS J. Photogramm. Remote Sens. 137, 112–133.
<https://doi.org/10.1016/j.isprsjprs.2018.01.013>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations (ICLR2021).
- Du, Q., 2007. Modified Fisher's Linear Discriminant Analysis for Hyperspectral Imagery. IEEE Geosci. Remote Sens. Lett. 4, 503–507.
<https://doi.org/10.1109/LGRS.2007.900751>
- Du, Q., Yang, H., 2008. Similarity-Based Unsupervised Band Selection for Hyperspectral Image Analysis. IEEE Geosci. Remote Sens. Lett. 5, 564–568.
<https://doi.org/10.1109/LGRS.2008.2000619>
- Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A.K., Yang, Y., 2020. Adversarial camouflage: Hiding physical-world attacks with natural styles, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp. 997–1005.
<https://doi.org/10.1109/CVPR42600.2020.00108>

References

- Engelmann, F., Kontogianni, T., Hermans, A., Leibe, B., 2017. Exploring Spatial Context for 3D Semantic Segmentation of Point Clouds, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, pp. 716–724. <https://doi.org/10.1109/ICCVW.2017.90>
- Engelmann, F., Kontogianni, T., Schult, J., Leibe, B., 2019. Know What Your Neighbors Do: 3D Semantic Segmentation of Point Clouds, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp. 395–409. https://doi.org/10.1007/978-3-030-11015-4_29
- Everingham, M., Gool, L. Van, Williams, C., Winn, J., Zisserman, A., Aytar, Y., Eslami, A., 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012): Part I – Classification Challenge. PASCAL VOC Challenge.
- Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D., 2018. Robust Physical-World Attacks on Deep Learning Visual Classification, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1625–1634. <https://doi.org/10.1109/CVPR.2018.00175>
- Fan, L., Cai, Y., 2021. An Efficient Filtering Approach for Removing Outdoor Point Cloud Data of Manhattan-World Buildings. *Remote Sens.* 13, 3796. <https://doi.org/10.3390/rs13193796>
- Fan, L., Powrie, W., Smethurst, J., Atkinson, P.M., Einstein, H., 2014. The effect of short ground vegetation on terrestrial laser scans at a local scale. *ISPRS J. Photogramm. Remote Sens.* 95, 42–52. <https://doi.org/10.1016/j.isprsjprs.2014.06.003>

- Fan, S., Dong, Q., Zhu, F., Lv, Y., Ye, P., Wang, F.-Y., 2021. SCF-Net: Learning Spatial Contextual Features for Large-Scale Point Cloud Segmentation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 14499–14508. <https://doi.org/10.1109/CVPR46437.2021.01427>
- Farrell, M.D., Mersereau, R.M., 2005. On the Impact of PCA Dimension Reduction for Hyperspectral Detection of Difficult Targets. *IEEE Geosci. Remote Sens. Lett.* 2, 192–195. <https://doi.org/10.1109/LGRS.2005.846011>
- Fauvel, M., Chanussot, J., Benediktsson, J.A., 2009. Kernel Principal Component Analysis for the Classification of Hyperspectral Remote Sensing Data over Urban Areas. *EURASIP J. Adv. Signal Process.* 2009, 783194. <https://doi.org/10.1155/2009/783194>
- Feng, J., Jiao, L., Liu, F., Sun, T., Zhang, X., 2015. Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy. *IEEE Trans. Geosci. Remote Sens.* 53, 2956–2969. <https://doi.org/10.1109/TGRS.2014.2367022>
- Feng, J., Jiao, L.C., Zhang, X., Sun, T., 2014. Hyperspectral band selection based on trivariate mutual information and clonal selection. *IEEE Trans. Geosci. Remote Sens.* 52, 4092–4115. <https://doi.org/10.1109/TGRS.2013.2279591>
- Feng, S., Itoh, Y., Parente, M., Duarte, M.F., 2017. Hyperspectral Band Selection From Statistical Wavelet Models. *IEEE Trans. Geosci. Remote Sens.* 55, 2111–2123. <https://doi.org/10.1109/TGRS.2016.2636850>
- Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Joseph Hughes, M., Laue, B., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* 194, 379–390.

References

- Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., Torr, P., 2021. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 652–662. <https://doi.org/10.1109/TPAMI.2019.2938758>
- Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *Int. J. Rob. Res.* 32, 1231–1237. <https://doi.org/10.1177/0278364913491297>
- Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3354–3361. <https://doi.org/10.1109/CVPR.2012.6248074>
- Goldblatt, R., Stuhlmacher, M.F., Tellman, B., Clinton, N., Hanson, G., Georgescu, M., Wang, C., Serrano-Candela, F., Khandelwal, A.K., Cheng, W.H., Balling, R.C., 2018. Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover. *Remote Sens. Environ.* 205, 253–275. <https://doi.org/10.1016/j.rse.2017.11.026>
- Golparvar-Fard, M., Bohn, J., Teizer, J., Savarese, S., Peña-Mora, F., 2011. Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques. *Autom. Constr.* 20, 1143–1155. <https://doi.org/10.1016/j.autcon.2011.04.016>
- Gong, J., Xu, J., Tan, X., Song, H., Qu, Y., Xie, Y., Ma, L., 2021. Omni-supervised Point Cloud Segmentation via Gradual Receptive Field Component Reasoning, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 11668–11677. <https://doi.org/10.1109/CVPR46437.2021.01150>

- Gong, M., Zhang, M., Yuan, Y., 2016. Unsupervised Band Selection Based on Evolutionary Multiobjective Optimization for Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* 54, 544–557. <https://doi.org/10.1109/TGRS.2015.2461653>
- Gonzalez, R.C., Woods, R.E., Masters, B.R., 2009. Digital Image Processing, Third Edition. *J. Biomed. Opt.* 14, 029901. <https://doi.org/10.1117/1.3115362>
- Graham, B., Engelcke, M., Maaten, L. van der, 2018. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 9224–9232. <https://doi.org/10.1109/CVPR.2018.00961>
- Gu, Q., Li, Z., Han, J., 2012. Generalized Fisher Score for Feature Selection. *Proc. 27th Conf. Uncertain. Artif. Intell. UAI 2011* 266–273. <https://doi.org/10.48550/arxiv.1202.3725>
- Gull, S.F., Skilling, J., 1984. Maximum entropy method in image processing. *IEE Proc. F Commun. Radar Signal Process.* 131, 646. <https://doi.org/10.1049/ip-f-1.1984.0099>
- Guo, B., Gunn, S.R.S.R., Damper, R.I.I., Nelson, J.D.B.D.B., 2006. Band Selection for Hyperspectral Image Classification Using Mutual Information. *IEEE Geosci. Remote Sens. Lett.* 3, 522–526. <https://doi.org/10.1109/LGRS.2006.878240>
- Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2021. Deep Learning for 3D Point Clouds: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 4338–4364. <https://doi.org/10.1109/TPAMI.2020.3005434>
- Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. SEMANTIC3D.NET: A NEW LARGE-SCALE POINT CLOUD CLASSIFICATION BENCHMARK, in: *ISPRS Annals of the Photogrammetry*,

References

- Remote Sensing and Spatial Information Sciences. Copernicus GmbH, pp. 91–98. <https://doi.org/10.5194/isprs-annals-IV-1-W1-91-2017>
- Hand, D.J., 2008. Data Clustering: Theory, Algorithms, and Applications by Guojun Gan, Chaoqun Ma, Jianhong Wu. *Int. Stat. Rev.* 76, 141–141. https://doi.org/10.1111/j.1751-5823.2007.00039_2.x
- Hansch, R., Hellwich, O., 2021. Fusion of Multispectral LiDAR, Hyperspectral, and RGB Data for Urban Land Cover Classification. *IEEE Geosci. Remote Sens. Lett.* 18, 366–370. <https://doi.org/10.1109/LGRS.2020.2972955>
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum Contrast for Unsupervised Visual Representation Learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 9726–9735. <https://doi.org/10.1109/CVPR42600.2020.00975>
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1904–1916. <https://doi.org/10.1109/TPAMI.2015.2389824>
- Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2011–2023. <https://doi.org/10.1109/TPAMI.2019.2913372>
- Hu, P., Liu, X., Cai, Y., Cai, Z., 2019. Band Selection of Hyperspectral Images Using Multiobjective Optimization-Based Sparse Self-Representation. *IEEE Geosci. Remote Sens. Lett.* 16, 452–456. <https://doi.org/10.1109/LGRS.2018.2872540>

- Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2021. Learning Semantic Segmentation of Large-Scale Point Clouds with Random Sampling. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1–1. <https://doi.org/10.1109/TPAMI.2021.3083288>
- Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 11105–11114. <https://doi.org/10.1109/CVPR42600.2020.01112>
- Hua, B.-S., Tran, M.-K., Yeung, S.-K., 2018. Pointwise Convolutional Neural Networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 984–993. <https://doi.org/10.1109/CVPR.2018.00109>
- Huang, C.-Q., Jiang, F., Huang, Q.-H., Wang, X.-Z., Han, Z.-M., Huang, W.-Y., 2022. Dual-Graph Attention Convolution Network for 3-D Point Cloud Classification. *IEEE Trans. Neural Networks Learn. Syst.* 1–13. <https://doi.org/10.1109/TNNLS.2022.3162301>
- Huang, H., Zhang, C., Hammad, A., 2021. Effective Scanning Range Estimation for Using TLS in Construction Projects. *J. Constr. Eng. Manag.* 147, 04021106. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0002127](https://doi.org/10.1061/(ASCE)CO.1943-7862.0002127)
- Huang, Q., Wang, W., Neumann, U., 2018. Recurrent Slice Networks for 3D Segmentation of Point Clouds, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2626–2635. <https://doi.org/10.1109/CVPR.2018.00278>

References

- Huang, R., He, M., 2005. Band Selection Based on Feature Weighting for Classification of Hyperspectral Data. *IEEE Geosci. Remote Sens. Lett.* 2, 156–159. <https://doi.org/10.1109/LGRS.2005.844658>
- Huang, S., Zhang, H., Pizurica, A., 2022. A Structural Subspace Clustering Approach for Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. <https://doi.org/10.1109/TGRS.2021.3102422>
- Huang, T., You, S., Wang, F., Qian, C., Zhang, C., Wang, X., Xu, C., 2022. GreedyNASv2: Greedier Search with a Greedy Path Filter, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 11892–11901. <https://doi.org/10.1109/CVPR52688.2022.01160>
- Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* 14, 55–63. <https://doi.org/10.1109/TIT.1968.1054102>
- Hughes, M., Hayes, D., 2014. Automated Detection of Cloud and Cloud Shadow in Single-Date Landsat Imagery Using Neural Networks and Spatial Post-Processing. *Remote Sens.* 6, 4907–4926. <https://doi.org/10.3390/rs6064907>
- Jain, A.K., Duin, P.W., Jianchang Mao, 2000. Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 4–37. <https://doi.org/10.1109/34.824819>
- Jaritz, M., Gu, J., Su, H., 2019. Multi-View PointNet for 3D Scene Understanding, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, pp. 3995–4003. <https://doi.org/10.1109/ICCVW.2019.00494>
- Jeppesen, J.H., Jacobsen, R.H., Inceoglu, F., Toftegaard, T.S., 2019. A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sens. Environ.* 229, 247–259. <https://doi.org/10.1016/j.rse.2019.03.039>

- Jia, S., Tang, G., Zhu, J., Li, Q., 2016. A Novel Ranking-Based Clustering Approach for Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* 54, 88–102. <https://doi.org/10.1109/TGRS.2015.2450759>
- Jia, S., Yuan, Y., Li, N., Liao, J., Huang, Q., Jia, X., Xu, M., 2022. A Multiscale Superpixel-Level Group Clustering Framework for Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* 60, 1–18. <https://doi.org/10.1109/TGRS.2022.3150361>
- Jiang, M., Wu, Y., Zhao, T., Zhao, Z., Lu, C., 2018. PointSIFT: A SIFT-like Network Module for 3D Point Cloud Semantic Segmentation. *arXiv*.
- Jing Huang, Suya You, 2016. Point cloud labeling using 3D Convolutional Neural Network, in: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 2670–2675. <https://doi.org/10.1109/ICPR.2016.7900038>
- Jing Wang, Chein-I Chang, 2006. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. *IEEE Trans. Geosci. Remote Sens.* 44, 1586–1600. <https://doi.org/10.1109/TGRS.2005.863297>
- Jung, H., Choi, H.-S., Kang, M., 2022. Boundary Enhancement Semantic Segmentation for Building Extraction From Remote Sensed Image. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12. <https://doi.org/10.1109/TGRS.2021.3108781>
- Kallepalli, A., Kumar, A., Khoshelham, K., 2014. Entropy based determination of optimal principal components of Airborne Prism Experiment (APEX) imaging spectrometer data for improved land cover classification. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* XL–8, 781–786. <https://doi.org/10.5194/isprsarchives-XL-8-781-2014>

References

- Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* 145, 60–77. <https://doi.org/10.1016/j.isprsjprs.2018.04.014>
- Kemker, R., Salvaggio, C., Kanan, C., 2017. High-Resolution Multispectral Dataset for Semantic Segmentation.
- Keshava, N., 2004. Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries. *IEEE Trans. Geosci. Remote Sens.* 42, 1552–1565. <https://doi.org/10.1109/TGRS.2004.830549>
- Kim, B.J., Choi, H., Jang, H., Lee, D.G., Jeong, W., Kim, S.W., 2021. Dead Pixel Test Using Effective Receptive Field. <https://doi.org/10.48550/arxiv.2108.13576>
- Kim, Seungho, Kim, Sangyong, Lee, D.-E., 2020. 3D Point Cloud and BIM-Based Reconstruction for Evaluation of Project by As-Planned and As-Built. *Remote Sens.* 12, 1457. <https://doi.org/10.3390/rs12091457>
- Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J., 2022. Stratified Transformer for 3D Point Cloud Segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 8490–8499. <https://doi.org/10.1109/CVPR52688.2022.00831>
- Landrieu, L., Boussaha, M., 2019. Point Cloud Oversegmentation With Graph-Structured Deep Metric Learning, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 7432–7441. <https://doi.org/10.1109/CVPR.2019.00762>
- Landrieu, L., Simonovsky, M., 2018. Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs, in: 2018 IEEE/CVF Conference on Computer Vision

- and Pattern Recognition. IEEE, pp. 4558–4567.
<https://doi.org/10.1109/CVPR.2018.00479>
- Lawin, F.J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F.S., Felsberg, M., 2017. Deep Projective 3D Semantic Segmentation, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 95–107. https://doi.org/10.1007/978-3-319-64689-3_8
- Leblanc, M., Tibshirani, R., 1996. Combining Estimates in Regression and Classification. *J. Am. Stat. Assoc.* 91, 1641–1650.
<https://doi.org/10.1080/01621459.1996.10476733>
- Lee, J.B., Woodyatt, A.S., Berman, M., 1990. Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform. *IEEE Trans. Geosci. Remote Sens.* 28, 295–304. <https://doi.org/10.1109/36.54356>
- Li, A., Jiao, L., Zhu, H., Li, L., Liu, F., 2022. Multitask Semantic Boundary Awareness Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. <https://doi.org/10.1109/TGRS.2021.3050885>
- Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M., 2022. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13.
<https://doi.org/10.1109/TGRS.2021.3093977>
- Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., Atkinson, P.M., 2021. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* 181, 84–98. <https://doi.org/10.1016/j.isprsjprs.2021.09.005>

References

- Li, W., Prasad, S., Fowler, J.E., Bruce, L.M., 2012. Locality-Preserving Dimensionality Reduction and Classification for Hyperspectral Image Analysis. *IEEE Trans. Geosci. Remote Sens.* 50, 1185–1198. <https://doi.org/10.1109/TGRS.2011.2165957>
- Li, Y., Li, X., Zhang, Z., Shuang, F., Lin, Q., Jiang, J., 2022. DenseKPNET: Dense Kernel Point Convolutional Neural Networks for Point Cloud Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. <https://doi.org/10.1109/TGRS.2022.3162582>
- Li, Y., Majumder, A., Zhang, H., Gopi, M., 2018. Optimized Multi-Spectral Filter Array Based Imaging of Natural Scenes. *Sensors* 18, 1172. <https://doi.org/10.3390/s18041172>
- Li, Z., Liu, M., Chen, Y., Xu, Y., Li, W., Du, Q., 2022. Deep Cross-Domain Few-Shot Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* 60, 1–18. <https://doi.org/10.1109/TGRS.2021.3057066>
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2020. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>
- Lin, Y., Vosselman, G., Cao, Y., Yang, M.Y., 2021. Local and global encoder network for semantic segmentation of Airborne laser scanning point clouds. *ISPRS J. Photogramm. Remote Sens.* 176, 151–168. <https://doi.org/10.1016/j.isprsjprs.2021.04.016>
- Ling Shao, Fan Zhu, Xuelong Li, 2015. Transfer Learning for Visual Categorization: A Survey. *IEEE Trans. Neural Networks Learn. Syst.* 26, 1019–1034. <https://doi.org/10.1109/TNNLS.2014.2330900>

- Liu, F., Liu, J., Wang, L., 2022. Asphalt Pavement Crack Detection Based on Convolutional Neural Network and Infrared Thermography. *IEEE Trans. Intell. Transp. Syst.* 23, 22145–22155. <https://doi.org/10.1109/TITS.2022.3142393>
- Liu, J., Wang, T., Skidmore, A.K., Jones, S., Heurich, M., Beudert, B., Premier, J., 2019. Comparison of terrestrial LiDAR and digital hemispherical photography for estimating leaf angle distribution in European broadleaf beech forests. *ISPRS J. Photogramm. Remote Sens.* 158, 76–89. <https://doi.org/10.1016/j.isprsjprs.2019.09.015>
- Liu, Z., Hu, H., Cao, Y., Zhang, Z., Tong, X., 2020. A Closer Look at Local Aggregation Operators in Point Cloud Analysis, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Science and Business Media Deutschland GmbH, pp. 326–342. https://doi.org/10.1007/978-3-030-58592-1_20
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021a. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 *IEEE/CVF Int. Conf. Comput. Vis.* 9992–10002. <https://doi.org/10.1109/ICCV48922.2021.00986>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 *IEEE/CVF Int. Conf. Comput. Vis.* 9992–10002. <https://doi.org/10.1109/ICCV48922.2021.00986>
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 11966–11976. <https://doi.org/10.1109/CVPR52688.2022.01167>

References

- Lo, Y., Fu, L., Lu, T., Huang, H., Kong, L., Xu, Y., Zhang, C., 2023. Medium-Sized Lake Water Quality Parameters Retrieval Using Multispectral UAV Image and Machine Learning Algorithms: A Case Study of the Yuandang Lake, China. *Drones* 7, 244. <https://doi.org/10.3390/drones7040244>
- Lu, Q., Lee, S., 2017. Image-Based Technologies for Constructing As-Is Building Information Models for Existing Buildings. *J. Comput. Civ. Eng.* 31, 04017005. [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000652](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000652)
- Luo, W., Li, Y., Urtasun, R., Zemel, R., 2017. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 4905–4913.
- Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., Yang, M.Y., 2020. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* 165, 108–119. <https://doi.org/10.1016/j.isprsjprs.2020.05.009>
- Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* 135, 158–172. <https://doi.org/10.1016/j.isprsjprs.2017.11.009>
- Martínez-Usó Martínez-Uso, A., Pla, F., Sotoca, J.M., García-Sevilla, P., 2007. Clustering-Based Hyperspectral Band Selection Using Information Measures. *IEEE Trans. Geosci. Remote Sens.* 45, 4158–4171. <https://doi.org/10.1109/TGRS.2007.904951>
- Mateo-García, G., Laparra, V., López-Puigdollers, D., Gómez-Chova, L., 2020. Transferring deep learning models for cloud detection between Landsat-8 and Proba-V. *ISPRS J. Photogramm. Remote Sens.* 160, 1–17.

- Meng, H.-Y., Gao, L., Lai, Y.-K., Manocha, D., 2019. VV-Net: Voxel VAE Net With Group Convolutions for Point Cloud Segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 8499–8507. <https://doi.org/10.1109/ICCV.2019.00859>
- Meyer, G.P., Charland, J., Hegde, D., Laddha, A., Vallespi-Gonzalez, C., 2019. Sensor Fusion for Joint 3D Object Detection and Semantic Segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 1230–1237. <https://doi.org/10.1109/CVPRW.2019.00162>
- Milioto, A., Vizzo, I., Behley, J., Stachniss, C., 2019. RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 4213–4220. <https://doi.org/10.1109/IROS40897.2019.8967762>
- MMSegmentation Contributors, 2020. OpenMMLab Semantic Segmentation Toolbox and Benchmark [WWW Document]. <https://github.com/open-mmlab/msegmentation>.
- Montuori, A., Luzi, G., Stramondo, S., Casula, G., Bignami, C., Bonali, E., Bianchi, M.G., Crosetto, M., 2014. Combined use of ground-based systems for Cultural Heritage conservation monitoring, in: 2014 IEEE Geoscience and Remote Sensing Symposium. IEEE, pp. 4086–4089. <https://doi.org/10.1109/IGARSS.2014.6947384>
- Munoz, D., Bagnell, J.A., Vandapel, N., Hebert, M., 2009. Contextual classification with functional Max-Margin Markov Networks, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 975–982. <https://doi.org/10.1109/CVPR.2009.5206590>

References

- Pan, B., Shi, Z., Xu, X., Shi, T., Zhang, N., Zhu, X., 2019. CoinNet: Copy Initialization Network for Multispectral Imagery Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* 16, 816–820. <https://doi.org/10.1109/LGRS.2018.2880756>
- Park, J., Kim, C., Kim, S., Jo, K., 2023. PCSCNet: Fast 3D semantic segmentation of LiDAR point cloud for autonomous car using point convolution and sparse convolution network. *Expert Syst. Appl.* 212, 118815. <https://doi.org/10.1016/j.eswa.2022.118815>
- Park, J.H., Inamori, T., Hamaguchi, R., Otsuki, K., Kim, J.E., Yamaoka, K., 2020. RGB Image Prioritization Using Convolutional Neural Network on a Microprocessor for Nanosatellites. *Remote Sens.* 12, 3941. <https://doi.org/10.3390/rs12233941>
- Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context Encoders: Feature Learning by Inpainting, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2536–2544. <https://doi.org/10.1109/CVPR.2016.278>
- Peng, C., Myronenko, A., Hatamizadeh, A., Nath, V., Siddiquee, M.M.R., He, Y., Xu, D., Chellappa, R., Yang, D., 2022. HyperSegNAS: Bridging One-Shot Neural Architecture Search with 3D Medical Image Segmentation using HyperNet, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 20709–20719. <https://doi.org/10.1109/CVPR52688.2022.02008>
- Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>

- Peng, L., Chen, X., Chen, J., Zhao, W., Cao, X., 2022. Understanding the Role of Receptive Field of Convolutional Neural Network for Cloud Detection in Landsat 8 OLI Imagery. *IEEE Trans. Geosci. Remote Sens.* 60, 1–17. <https://doi.org/10.1109/TGRS.2022.3150083>
- Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E.S., Frontoni, E., Lingua, A.M., 2020. Point Cloud Semantic Segmentation Using a Deep Learning Framework for Cultural Heritage. *Remote Sens.* 12, 1005. <https://doi.org/10.3390/rs12061005>
- Pozzer, S., De Souza, M.P.V., Hena, B., Hesam, S., Rezaiye, R.K., Rezazadeh Azar, E., Lopez, F., Maldague, X., 2022. Effect of different imaging modalities on the performance of a CNN: An experimental study on damage segmentation in infrared, visible, and fused images of concrete structures. *NDT E Int.* 132, 102709. <https://doi.org/10.1016/j.ndteint.2022.102709>
- Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space, in: *Advances in Neural Information Processing Systems*. pp. 5100–5109.
- Rajbhandari, S., Aryal, J., Osborn, J., Lucieer, A., Musk, R., 2019. Leveraging Machine Learning to Extend Ontology-Driven Geographic Object-Based Image Analysis (O-GEOBIA): A Case Study in Forest-Type Mapping. *Remote Sens.* 11, 503. <https://doi.org/10.3390/rs11050503>
- Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. *Science (80-.)*. 344, 1492–1496. <https://doi.org/10.1126/science.1242072>
- Roger, R.E., 1994. A faster way to compute the noise-adjusted principal components transform matrix. *IEEE Trans. Geosci. Remote Sens.* 32, 1194–1196. <https://doi.org/10.1109/36.338369>

References

- Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
- Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. THE ISPRS BENCHMARK ON URBAN OBJECT CLASSIFICATION AND 3D BUILDING RECONSTRUCTION. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* I-3, 293–298. <https://doi.org/10.5194/isprsannals-I-3-293-2012>
- Roweis, S.T., Saul, L.K., 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* (80-.). 290, 2323–2326. <https://doi.org/10.1126/science.290.5500.2323>
- Roy, S.K., Das, S., Song, T., Chanda, B., 2021a. DARecNet-BS: Unsupervised Dual-Attention Reconstruction Network for Hyperspectral Band Selection. *IEEE Geosci. Remote Sens. Lett.* 18, 2152–2156. <https://doi.org/10.1109/LGRS.2020.3013235>
- Roy, S.K., Manna, S., Song, T., Bruzzone, L., 2021b. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* 59, 7831–7843. <https://doi.org/10.1109/TGRS.2020.3043267>
- Roynard, X., Deschaud, J.-E., Goulette, F., 2018. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *Int. J. Rob. Res.* 37, 545–557. <https://doi.org/10.1177/0278364918767506>

- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- Safaie, A.H., Rastiveis, H., Shams, A., Sarasua, W.A., Li, J., 2021. Automated street tree inventory using mobile LiDAR point clouds based on Hough transform and active contours. *ISPRS J. Photogramm. Remote Sens.* 174, 19–34. <https://doi.org/10.1016/j.isprsjprs.2021.01.026>
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
- Saxena, N., Babu, N.K., Raman, B., 2020. Semantic segmentation of multispectral images using res-seg-net model, in: Proceedings - 14th IEEE International Conference on Semantic Computing, ICSC 2020. Institute of Electrical and Electronics Engineers Inc., pp. 154–157. <https://doi.org/10.1109/ICSC.2020.00030>
- Shahin Shamsabadi, A., Sanchez-Matilla, R., Cavallaro, A., 2020. ColorFool: Semantic Adversarial Colorization, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1148–1157. <https://doi.org/10.1109/CVPR42600.2020.00123>
- Shelhamer, E., Long, J., Darrell, T., 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 640–651. <https://doi.org/10.1109/TPAMI.2016.2572683>

References

- Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor Segmentation and Support Inference from RGBD Images. pp. 746–760. https://doi.org/10.1007/978-3-642-33715-4_54
- Sima, C., Dougherty, E.R., 2008. The peaking phenomenon in the presence of feature-selection. *Pattern Recognit. Lett.* 29, 1667–1674. <https://doi.org/10.1016/j.patrec.2008.04.010>
- Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR. <https://doi.org/10.48550/arxiv.1409.1556>
- Song, Y., Aryal, J., Tan, L., Jin, L., Gao, Z., Wang, Y., 2021. Comparison of changes in vegetation and land cover types between Shenzhen and Bangkok. *L. Degrad. Dev.* 32, 1192–1204. <https://doi.org/10.1002/ldr.3788>
- Sotoca, J.M., Pla, F., Sánchez, J.S., 2007. Band selection in multispectral images by minimization of dependent information. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 37, 258–267. <https://doi.org/10.1109/TSMCC.2006.876055>
- Stein, A., Aryal, J., Gort, G., 2005. Use of the Bradley-Terry model to quantify association in remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* 43, 852–856. <https://doi.org/10.1109/TGRS.2005.843569>
- Stojanovski, T., 2018. City Information Modelling (CIM) and Urban Design Morphological Structure, Design Elements and Programming Classes in CIM. *Comput. a better tomorrow - Proc. 36th eCAADe Conf. - Vol. 1 1*, 507–516.
- Su, H., Yang, H., Du, Q., Sheng, Y., 2011. Semisupervised Band Clustering for Dimensionality Reduction of Hyperspectral Imagery. *IEEE Geosci. Remote Sens. Lett.* 8, 1135–1139. <https://doi.org/10.1109/LGRS.2011.2158185>

- Su, X., You, S., Wang, F., Qian, C., Zhang, C., Xu, C., 2021. BCNet: Searching for Network Width with Bilaterally Coupled Network, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2175–2184. <https://doi.org/10.1109/CVPR46437.2021.00221>
- Sun, H., Ren, J., Zhao, H., Yuen, P., Tschannerl, J., 2022. Novel Gumbel-Softmax Trick Enabled Concrete Autoencoder With Entropy Constraints for Unsupervised Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. <https://doi.org/10.1109/TGRS.2021.3075663>
- Sun, W., Du, Q., 2019. Hyperspectral Band Selection: A Review. *IEEE Geosci. Remote Sens. Mag.* 7, 118–139. <https://doi.org/10.1109/MGRS.2019.2911100>
- Sun, W., Peng, J., Yang, G., Du, Q., 2020. Correntropy-Based Sparse Spectral Clustering for Hyperspectral Band Selection. *IEEE Geosci. Remote Sens. Lett.* 17, 484–488. <https://doi.org/10.1109/LGRS.2019.2924934>
- Sun, W., Yang, G., Peng, J., Meng, X., He, K., Li, W., Li, H.-C., Du, Q., 2022. A Multiscale Spectral Features Graph Fusion Method for Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12. <https://doi.org/10.1109/TGRS.2021.3102246>
- Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning, in: 31st AAAI Conference on Artificial Intelligence, AAAI 2017. pp. 4278–4284.
- Takikawa, T., Acuna, D., Jampani, V., Fidler, S., 2019. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 5228–5237. <https://doi.org/10.1109/ICCV.2019.00533>

References

- Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K., Li, J., 2020. Toronto-3D: A Large-scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 797–806. <https://doi.org/10.1109/CVPRW50498.2020.00109>
- Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S., 2020. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 12373 LNCS, 685–702. https://doi.org/10.1007/978-3-030-58604-1_41/FIGURES/7
- Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.-Y., 2018. Tangent Convolutions for Dense Prediction in 3D, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3887–3896. <https://doi.org/10.1109/CVPR.2018.00409>
- Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S., 2017. SEGCloud: Semantic Segmentation of 3D Point Clouds, in: 2017 International Conference on 3D Vision (3DV). IEEE, pp. 537–547. <https://doi.org/10.1109/3DV.2017.00067>
- Tenenbaum, J.B., Silva, V. de, Langford, J.C., 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* (80-.). 290, 2319–2323. <https://doi.org/10.1126/science.290.5500.2319>
- Theodoridis, S., Koutroumbas, K., 2001. Pattern recognition and neural networks, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Berlin, Heidelberg, pp. 169–195. https://doi.org/10.1007/3-540-44673-7_8

- Thomas, H., Goulette, F., Deschaud, J.-E., Marcotegui, B., LeGall, Y., 2018. Semantic Classification of 3D Point Clouds with Multiscale Spherical Neighborhoods, in: 2018 International Conference on 3D Vision (3DV). IEEE, pp. 390–398. <https://doi.org/10.1109/3DV.2018.00052>
- Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L., 2019. KPConv: Flexible and Deformable Convolution for Point Clouds, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 6410–6419. <https://doi.org/10.1109/ICCV.2019.00651>
- Tokarczyk, P., Wegner, J.D., Walk, S., Schindler, K., 2015. Features, Color Spaces, and Boosting: New Insights on Semantic Classification of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 53, 280–295. <https://doi.org/10.1109/TGRS.2014.2321423>
- Uddin, M.P., Mamun, M. Al, Hossain, M.A., 2021. PCA-based Feature Reduction for Hyperspectral Remote Sensing Image Classification. *IETE Tech. Rev.* 38, 377–396. <https://doi.org/10.1080/02564602.2020.1740615>
- Vallet, B., Brédif, M., Serna, A., Marcotegui, B., Paparoditis, N., 2015. TerraMobilita/iQmulus urban point cloud analysis benchmark. *Comput. Graph.* 49, 126–133. <https://doi.org/10.1016/j.cag.2015.03.004>
- Varney, N., Asari, V.K., Graehling, Q., 2020. DALES: A Large-scale Aerial LiDAR Data Set for Semantic Segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 717–726. <https://doi.org/10.1109/CVPRW50498.2020.00101>
- Volpi, M., Ferrari, V., 2015. Semantic segmentation of urban scenes by learning local class interactions, in: 2015 IEEE Conference on Computer Vision and Pattern

References

- Recognition Workshops (CVPRW). IEEE, pp. 1–9.
<https://doi.org/10.1109/CVPRW.2015.7301377>
- Vosselman, G., Coenen, M., Rottensteiner, F., 2017. Contextual segment-based classification of airborne laser scanner data. *ISPRS J. Photogramm. Remote Sens.* 128, 354–371.
- Wang, F., Zhuang, Y., Gu, H., Hu, H., 2020. OctreeNet: A Novel Sparse 3-D Convolutional Neural Network for Real-Time 3-D Outdoor Scene Analysis. *IEEE Trans. Autom. Sci. Eng.* 17, 735–747.
<https://doi.org/10.1109/TASE.2019.2942068>
- Wang, Jingdong, Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2021. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 3349–3364. <https://doi.org/10.1109/TPAMI.2020.2983686>
- Wang, Junjue, Zheng, Z., Ma, A., Lu, X., Zhong, Y., 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation, in: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. <https://doi.org/10.48550/arxiv.2110.08733>
- Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J., 2019. Graph Attention Convolution for Point Cloud Semantic Segmentation, in: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10288–10297.
<https://doi.org/10.1109/CVPR.2019.01054>
- Wang, Q., Li, Q., Li, X., 2019. Hyperspectral Band Selection via Adaptive Subspace Partition Strategy. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 12, 4940–4950. <https://doi.org/10.1109/JSTARS.2019.2941454>

- Wang, Q., Meng, Z., Li, X., 2017. Locality Adaptive Discriminant Analysis for Spectral–Spatial Classification of Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* 14, 2077–2081. <https://doi.org/10.1109/LGRS.2017.2751559>
- Wang, Q., Zhang, F., Li, X., 2020. Hyperspectral Band Selection via Optimal Neighborhood Reconstruction. *IEEE Trans. Geosci. Remote Sens.* 58, 8465–8476. <https://doi.org/10.1109/TGRS.2020.2987955>
- Wang, Q., Zhang, F., Li, X., 2018. Optimal Clustering Framework for Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* 56, 1–13. <https://doi.org/10.1109/TGRS.2018.2828161>
- Wang, S., Suo, S., Ma, W.-C., Pokrovsky, A., Urtasun, R., 2018. Deep Parametric Continuous Convolutional Neural Networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2589–2597. <https://doi.org/10.1109/CVPR.2018.00274>
- Wei, P., Hansch, R., 2022. Random Ferns for Semantic Segmentation of PolSAR Images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12. <https://doi.org/10.1109/TGRS.2021.3131418>
- Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS J. Photogramm. Remote Sens.* 105, 286–304. <https://doi.org/10.1016/j.isprsjprs.2015.01.016>
- Wen, C., Li, X., Yao, X., Peng, L., Chi, T., 2021. Airborne LiDAR point cloud classification with global-local graph attention convolution neural network. *ISPRS J. Photogramm. Remote Sens.* 173, 181–194.
- Wen, X., Han, Z., Liu, X., Liu, Y.-S., 2020. Point2SpatialCapsule: Aggregating Features and Spatial Relationships of Local Regions on Point Clouds Using

References

- Spatial-Aware Capsules. *IEEE Trans. Image Process.* 29, 8855–8869.
<https://doi.org/10.1109/TIP.2020.3019925>
- Wolpert, D.H., 1992. Stacked generalization. *Neural Networks* 5, 241–259.
[https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- Wu, B., Wan, A., Yue, X., Keutzer, K., 2018. SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud, in: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 1887–1893.
<https://doi.org/10.1109/ICRA.2018.8462926>
- Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P., 2020. Visual Transformers: Token-based Image Representation and Processing for Computer Vision.
<https://doi.org/10.48550/arxiv.2006.03677>
- Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K., 2019. SqueezeSegV2: Improved Model Structure and Unsupervised Domain Adaptation for Road-Object Segmentation from a LiDAR Point Cloud, in: 2019 International Conference on Robotics and Automation (ICRA). IEEE, pp. 4376–4382.
<https://doi.org/10.1109/ICRA.2019.8793495>
- Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised Feature Learning via Non-parametric Instance Discrimination, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3733–3742.
<https://doi.org/10.1109/CVPR.2018.00393>
- Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified Perceptual Parsing for Scene Understanding, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

- Bioinformatics). Springer Verlag, pp. 432–448. https://doi.org/10.1007/978-3-030-01228-1_26
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* 34 pre-proceedings (NeurIPS 2021).
- Xie, F., Li, F., Lei, C., Ke, L., 2018. Representative Band Selection for Hyperspectral Image Classification. *ISPRS Int. J. Geo-Information* 7, 338. <https://doi.org/10.3390/ijgi7090338>
- Xiuping Jia, Richards, J.A., 1999. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *IEEE Trans. Geosci. Remote Sens.* 37, 538–542. <https://doi.org/10.1109/36.739109>
- Xu, B., Li, X., Hou, W., Wang, Y., Wei, Y., 2021. A Similarity-Based Ranking Method for Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* 59, 9585–9599. <https://doi.org/10.1109/TGRS.2020.3048138>
- Xu, M., Ding, R., Zhao, H., Qi, X., 2021. PAConv: Position Adaptive Convolution with Dynamic Kernel Assembling on Point Clouds, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3172–3181. <https://doi.org/10.1109/CVPR46437.2021.00319>
- Xu, X., Ding, L., Luo, H., Ma, L., 2014. From Building Information Modeling to City Information Modeling. *J. Inf. Technol. Constr.* 19.
- Xu, Z., Zhang, W., Zhang, T., Li, J., 2020. HRCNet: High-Resolution Context Extraction Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* 13, 71. <https://doi.org/10.3390/rs13010071>

References

- Yan, K., Hu, Q., Wang, H., Huang, X., Li, L., Ji, S., 2022. Continuous Mapping Convolution for Large-Scale Point Clouds Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. <https://doi.org/10.1109/LGRS.2021.3107006>
- Yan, L., Fan, B., Liu, H., Huo, C., Xiang, S., Pan, C., 2020. Triplet Adversarial Domain Adaptation for Pixel-Level Classification of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 58, 3558–3573. <https://doi.org/10.1109/TGRS.2019.2958123>
- Yang, H., Du, Q., Su, H., Sheng, Y., 2011. An Efficient Method for Supervised Hyperspectral Band Selection. *IEEE Geosci. Remote Sens. Lett.* 8, 138–142. <https://doi.org/10.1109/LGRS.2010.2053516>
- Yang, Jun, Wang, W., Lin, G., Li, Q., Sun, Yeqing, Sun, Yixuan, 2019. Infrared Thermal Imaging-Based Crack Detection Using Deep Learning. *IEEE Access* 7, 182060–182077. <https://doi.org/10.1109/ACCESS.2019.2958264>
- Yang, Jiancheng, Zhang, Q., Ni, B., Li, L., Liu, J., Zhou, M., Tian, Q., 2019. Modeling Point Clouds With Self-Attention and Gumbel Subset Sampling, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3318–3327. <https://doi.org/10.1109/CVPR.2019.00344>
- Yang, Z., Yan, Z., Sun, X., Diao, W., Yang, Y., Li, X., 2022. Category Correlation and Adaptive Knowledge Distillation for Compact Cloud Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–18. <https://doi.org/10.1109/TGRS.2022.3174910>
- Ye, X., Li, J., Huang, H., Du, L., Zhang, X., 2018. 3D recurrent neural networks with context fusion for point cloud semantic segmentation, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence*

- and Lecture Notes in Bioinformatics). Springer Verlag, pp. 415–430.
https://doi.org/10.1007/978-3-030-01234-2_25
- Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., Hu, H., 2020. Disentangled Non-local Neural Networks, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Science and Business Media Deutschland GmbH, pp. 191–207. https://doi.org/10.1007/978-3-030-58555-6_12
- You, Y., Lou, Y., Liu, Q., Tai, Y.-W., Ma, L., Lu, C., Wang, W., 2020. Pointwise Rotation-Invariant Network with Adaptive Sampling and 3D Spherical Voxel Convolution. Proc. AAAI Conf. Artif. Intell. 34, 12717–12724.
<https://doi.org/10.1609/aaai.v34i07.6965>
- Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp. 334–349.
https://doi.org/10.1007/978-3-030-01261-8_20
- Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J., 2021. Lite-HRNet: A Lightweight High-Resolution Network, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 10435–10445.
<https://doi.org/10.1109/CVPR46437.2021.01030>
- Yuan, Y., Chen, X., Wang, J., 2020. Object-Contextual Representations for Semantic Segmentation, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Cham, pp. 173–190. https://doi.org/10.1007/978-3-030-58539-6_11

References

- Zabalza, J., Ren, J., Yang, M., Zhang, Y., Wang, J., Marshall, S., Han, J., 2014. Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing. *ISPRS J. Photogramm. Remote Sens.* 93, 112–122. <https://doi.org/10.1016/j.isprsjprs.2014.04.006>
- Zhai, H., Zhang, H., Zhang, L., Li, P., 2019. Laplacian-Regularized Low-Rank Subspace Clustering for Hyperspectral Image Band Selection. *IEEE Trans. Geosci. Remote Sens.* 57, 1723–1740. <https://doi.org/10.1109/TGRS.2018.2868796>
- Zhan, X., Pan, X., Liu, Z., Lin, D., Loy, C.C., 2019. Self-Supervised Learning via Conditional Motion Propagation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1881–1889. <https://doi.org/10.1109/CVPR.2019.00198>
- Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sens. Environ.* 216, 57–70. <https://doi.org/10.1016/j.rse.2018.06.034>
- Zhang, J., Wu, J., Wang, H., Wang, Y., Li, Y., 2022. Cloud Detection Method Using CNN Based on Cascaded Feature Attention and Channel Attention. *IEEE Trans. Geosci. Remote Sens.* 60, 1–17. <https://doi.org/10.1109/TGRS.2021.3120752>
- Zhang, R., Isola, P., Efros, A.A., 2017. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 645–654. <https://doi.org/10.1109/CVPR.2017.76>

- Zhang, R., Li, G., Li, M., Wang, L., 2018. Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning. *ISPRS J. Photogramm. Remote Sens.* 143, 85–96.
- Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 6848–6856. <https://doi.org/10.1109/CVPR.2018.00716>
- Zhang, Y., Sidibé, D., Morel, O., Mériaudeau, F., 2021. Deep multimodal fusion for semantic image segmentation: A survey, *Image and Vision Computing.* <https://doi.org/10.1016/j.imavis.2020.104042>
- Zhang, Z., Hua, B.-S., Yeung, S.-K., 2019. ShellNet: Efficient Point Cloud Convolutional Neural Networks Using Concentric Shells Statistics, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 1607–1616. <https://doi.org/10.1109/ICCV.2019.00169>
- Zhao, C., Zhou, W., Lu, L., Zhao, Q., 2019. Pooling Scores of Neighboring Points for Improved 3D Point Cloud Segmentation, in: 2019 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 1475–1479. <https://doi.org/10.1109/ICIP.2019.8803048>
- Zhao, H., Jiang, L., Fu, C.W., Jia, J., 2019. Pointweb: Enhancing local neighborhood features for point cloud processing, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp. 5560–5568. <https://doi.org/10.1109/CVPR.2019.00571>
- Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 6230–6239. <https://doi.org/10.1109/CVPR.2017.660>

References

- Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J., 2018. PSANet: Point-wise Spatial Attention Network for Scene Parsing, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 270–286. https://doi.org/10.1007/978-3-030-01240-3_17
- Zhao, Y., Birdal, T., Deng, H., Tombari, F., 2019. 3D Point Capsule Networks, in: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1009–1018. <https://doi.org/10.1109/CVPR.2019.00110>
- Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L., 2021. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers, in: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6877–6886. <https://doi.org/10.1109/CVPR46437.2021.00681>
- Zheng, X., Huan, L., Xia, G.-S., Gong, J., 2020. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. *ISPRS J. Photogramm. Remote Sens.* 170, 15–28. <https://doi.org/10.1016/j.isprsjprs.2020.09.019>
- Zhiheng, K., Ning, L., 2019. PyramNet: Point Cloud Pyramid Attention Network and Graph Embedding Module for Classification and Segmentation.
- Zhu, M., Jiao, L., Liu, F., Yang, S., Wang, J., 2021. Residual Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* 59, 449–462. <https://doi.org/10.1109/TGRS.2020.2994057>

Publications

Conference proceedings:

Cai, Y., Fan, L., Zhang, C., 2021a. An overview of constructing geometric models of buildings using point clouds, in: Zhang, Y., Zhang, D. (Eds.), 2021 International Conference on Image, Video Processing, and Artificial Intelligence. SPIE, p. 26.
<https://doi.org/10.1117/12.2611685>

Journal articles:

Cai, Y., Huang, H., Wang, K., Zhang, C., Fan, L., Guo, F., 2021. Selecting Optimal Combination of Data Channels for Semantic Segmentation in City Information Modelling (CIM). *Remote Sens.* 13, 1367. <https://doi.org/10.3390/rs13071367>

Cai, Y., Fan, L., 2021. An Efficient Approach to Automatic Construction of 3D Watertight Geometry of Buildings Using Point Clouds. *Remote Sens.* 13, 1947.
<https://doi.org/10.3390/rs13101947>

Fan, L., Cai, Y., 2021. An Efficient Filtering Approach for Removing Outdoor Point Cloud Data of Manhattan-World Buildings. 13, 1947.
<https://doi.org/10.3390/rs13193796>

Cai, Y., Fan, L., Atkinson, P.M., Zhang, C., 2022. Semantic Segmentation of Terrestrial Laser Scanning Point Clouds Using Locally Enhanced Image-Based Geometric Representations. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15.
<https://doi.org/10.1109/TGRS.2022.3161982>

Publications

Cai, Y., Fan, L., Zhang, C., 2022. Semantic Segmentation of Multispectral Images via Linear Compression of Bands: An Experiment Using RIT-18. *Remote Sens.* 14, 2673.

<https://doi.org/10.3390/rs14112673>

Gong, P., Cai, Y., Zhou, Z., Zhang, C., Chen, B., Sharples, S., 2022. Investigating spatial impact on indoor personal thermal comfort. *J. Build. Eng.* 45, 103536.

<https://doi.org/10.1016/j.jobe.2021.103536>

Cai, Y., Fan, L., Fang, Y., 2023. SBSS: Stacking-Based Semantic Segmentation Framework for Very High-Resolution Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.* 61, 1–14. <https://doi.org/10.1109/TGRS.2023.3234549>

Cheng, B., Chen, S., Fan, L., Li, Y., Cai, Y., Liu, Z., 2023. Windows and Doors Extraction from Point Cloud Data Combining Semantic Features and Material Characteristics. *Build.* 2023, Vol. 13, Page 507 13, 507.

<https://doi.org/10.3390/BUILDINGS13020507>

Fang, Y., Cai, Y., Fan, L., 2023. SDRCNN: A single-scale dense residual connected convolutional neural network for pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* <https://doi.org/10.1109/JSTARS.2023.3292320>

IEEE COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Semantic segmentation of terrestrial laser scanning point clouds using locally enhanced image-based geometric representations
Cai, Yuanzhi; Fan, Lei; Atkinson, Peter; Zhang, Cheng
Transactions on Geoscience and Remote Sensing

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the [IEEE PSPB Operations Manual](#).
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Yuanzhi Cai

Signature

22-03-2022

Date (dd-mm-yyyy)

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality,

authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use. The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers.** Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.

Please direct all questions about IEEE copyright policy to:

IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

IEEE COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

SBSS: Stacking-Based Semantic Segmentation Framework for Very High Resolution Remote Sensing Image

Cai, Yuanzhi; Fan, Lei; Fang, Yuan

Transactions on Geoscience and Remote Sensing

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the [IEEE PSPB Operations Manual](#).
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Yuanzhi Cai

Signature

04-01-2023

Date (dd-mm-yyyy)

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine

whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use. The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers.** Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.

Please direct all questions about IEEE copyright policy to:

IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966