UNIVERSITY OF
LIVERPOOL

Advanced Representation Learning for Dense Prediction Tasks in Medical Image Analysis

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

Sifan Song

October 2023

Abstract

Machine learning is a rapidly growing field of artificial intelligence that allows computers to learn and make predictions using human labels. However, traditional machine learning methods have many drawbacks, such as being time-consuming, inefficient, task-specific biased, and requiring a large amount of domain knowledge. A subfield of machine learning, representation learning, focuses on learning meaningful and useful features or representations from input data. It aims to automatically learn relevant features from raw data, saving time, increasing efficiency and generalization, and reducing reliance on expert knowledge. Recently, deep learning has further accelerated the development of representation learning. It leverages deep architectures to extract complex and abstract representations, resulting in significant outperformance in many areas.

In the field of computer vision, deep learning has made remarkable progress, particularly in high-level and real-world computer vision tasks. Since deep learning methods do not require handcrafted features and have the ability to understand complex visual information, they facilitate researchers to design automated systems that make accurate diagnoses and interpretations, especially in the field of medical image analysis. Deep learning has achieved state-of-the-art performance in many medical image analysis tasks, such as medical image regression/classification, generation and segmentation tasks. Compared to regression/classification tasks, medical image generation and segmentation tasks are more complex dense prediction tasks that understand semantic representations and generate pixel-level predictions.

This thesis focuses on designing representation learning methods to improve the performance of dense prediction tasks in the field of medical image analysis. With advances in imaging technology, more complex medical images become available for use in this field. In contrast to traditional machine learning algorithms, current deep learning-based representation learning methods provide an end-to-end approach to automatically extract representations without the need for manual feature engineering from the complex data. In the field of medical image analysis, there are three unique challenges requiring the design of advanced representation learning architectures, *i.e.*, limited labeled medical images, overfitting with limited data, and lack of interpretability. To address these challenges, we aim to design robust representation learning architectures for the two main directions of dense prediction tasks, namely medical image generation and segmentation.

For medical image generation, the specific topic that we focus on is chromosome straightening. This task involves generating a straightened chromosome image from a curved chromosome input. In addition, the challenges of this task include insufficient training images and corresponding ground truth, as well as the non-rigid nature of chromosomes, leading to distorted details and shapes after straightening. We first propose a study for the chromosome straightening task. We introduce a novel framework using image-to-image translation and demonstrate its efficacy and robustness in generating straightened chromosomes. The framework addresses the challenges of limited training data and outperforms existing studies. We then present a subsequent study to address the limitations of our previous framework, resulting in new state-of-the-art performance and better interpretability and generalization capability. We propose a new robust chromosome straightening framework, named Vit-Patch GAN, which instead learns the motion representation of chromosomes for straightening while retaining more details of shape and banding patterns.

For medical image segmentation, we focus on the fovea localization task, which is transferred from localization to small region segmentation. Accurate segmentation of the fovea region is crucial for monitoring and analyzing retinal diseases to prevent irreversible vision loss. This task also requires the incorporation of global features to effectively identify the fovea region and overcome hard cases associated with retinal diseases and non-standard fovea locations. We first propose a novel two-branch architecture, Bilateral-ViT, for fovea localization in retina image segmentation. This vision-transformer-based architecture incorporates global image context and blood vessel structure. It surpasses existing methods and achieves state-of-the-art results on two public datasets. We then propose a subsequent method to further improve the performance of fovea localization. We design a novel dual-stream deep learning architecture called Bilateral-Fuser. In contrast to our previous Bilateral-ViT, Bilateral-Fuser globally incorporates long-range connections from multiple cues, including fundus and vessel distribution. Moreover, with the newly designed Bilateral Token Incorporation module, Bilateral-Fuser learns anatomical-aware tokens, significantly reducing computational costs while achieving new state-of-the-art performance. Our comprehensive experiments also demonstrate that Bilateral-Fuser achieves better accuracy and robustness on both normal and diseased retina images, with excellent generalization capability.

Acknowledgements

First and foremost, I would like to express my sincere appreciation and deepest gratitude to my main supervisor, Dr. Jionglong Su, for his exceptional guidance and unwavering support throughout my Ph.D. journey. His encouragement and insightful feedback have continuously guided me in my pursuit of excellence. His dedication and commitment to my research have been invaluable, inspiring me both personally and professionally. I am truly grateful for his patience, understanding, and belief in my abilities. I am truly fortunate to have had him as my supervisor, and I will always cherish the knowledge and skills I have gained under his guidance.

I would also like to thank my co-supervisors, Prof. Jia Meng, Prof. Fei Ma, and Prof. Frans Coenen for their support and contribution to my Ph.D. research. I am very grateful for their willingness to share their knowledge and experience to enrich my study, as well as their patience in proofreading my papers. Furthermore, I would like to express my gratitude to Prof. S. Kevin Zhou for broadening my understanding of this field. I sincerely appreciate the academic and professional development opportunities he has provided me. Their support has been a tremendous source of motivation and inspiration for me.

I would like to thank my colleagues for insightful discussions and brainstorms, which have broadened my perspectives and refined my ideas. Their continuous support and friendship have made this challenging journey more manageable and enjoyable.

I would like to express my deepest gratitude to my parents for their support and encouragement throughout my academic journey. From the early stages of my education to the completion of my Ph.D., they have been there, providing the unwavering support I needed to pursue my dreams.

Finally, I would like to express my heartfelt appreciation to my girlfriend, Yining Wang, for her help, love and support. Whether it was during challenging moments or celebrating achievements, she has consistently been by my side. I am endlessly grateful for her encouragement and the motivation she brings to every aspect of my life.

Contents

Abstract **i**

Acknowledgements **iii**

Contents **vii**

List of Figures **x**

List of Tables **xii**

List of Abbreviations **xiii**

1 Introduction **1**

1.1 Overview 1

1.2 Challenges and Motivations 4

1.3 Contributions 6

1.4 Outlines 9

2 Literature Review **11**

2.1 Classical Architectures in CNN 12

2.2 Skip Connections 14

2.3 Attention Mechanisms 18

2.3.1 CNN-Based Attention 18

2.3.2 Multi-Head Self-Attention and Transformer 20

2.4 Conditional Generative Adversarial Network (cGAN) 24

3 A Novel Application of Image-to-Image Translation: Chromosome Straightening Framework by Learning from a Single Image **28**

3.1 Introduction and Problem Statement 28

3.2 Methodology 33

3.2.1 Data Augmentation Using a Single Image 33

3.2.2 Image-to-Image Translation for Straightening 36

3.3 Experiments and Results . . . 37
3.3.1 Chromosome Dataset . . . 37
3.3.2 Evaluation Metrics . . . 37
3.3.3 Implementation Details . . . 38
3.3.4 Results . . . 38
3.4 Limitation and Discussion . . . 42
3.4.1 Computation Time . . . 42
3.4.2 Failure Cases . . . 42
3.4.3 Potential Applications . . . 43
3.5 Conclusions and Future Work . . . 44

4 A Robust Framework of Chromosome Straightening with ViT-Patch GAN 45
4.1 Introduction and Problem Statement . . . 46
4.2 Methodology . . . 48
4.2.1 Network Architecture . . . 48
4.2.2 SL-matching Scheme . . . 50
4.3 Experiments . . . 50
4.3.1 Experiment Setup . . . 50
4.3.2 Comparison with State-of-the-Art Methods . . . 52
4.3.3 Ablation Study and Cross-Dataset Experiments . . . 53
4.4 Conclusion and Future Work . . . 54

5 Bilateral-ViT for robust fovea localization 56
5.1 Introduction and Problem Statement . . . 56
5.2 Related Work . . . 58
5.3 Methodology . . . 58
5.3.1 Network Architecture . . . 58
5.3.2 Implementation Details . . . 61
5.4 Experiments . . . 62
5.4.1 Fovea Localization on Normal and Diseased Images . . . 63
5.4.2 Comparison with State-of-the-Art Methods . . . 64
5.4.3 Ablation Study and Cross-Dataset Experiments . . . 64
5.5 Conclusion and Future Work . . . 66

6 Bilateral-Fuser: A Novel Multi-cue Fusion Architecture with Anatomical-aware Tokens for Fovea Localization 67
6.1 Introduction and Problem Statement . . . 68
6.2 Related Work . . . 70
6.2.1 Anatomical Structure-based Methods . . . 70
6.2.2 Deep Neural Networks in Fovea Localization . . . 71

6.3 Methodology . . . 72
6.3.1 Overall Architecture . . . 73
6.3.2 Bilateral Token Incorporation (BTI) modules . . . 74
6.4 Experiments . . . 76
6.4.1 Datasets and Network Configurations . . . 76
6.4.2 Implementation Details and Evaluation Metrics . . . 77
6.5 Results . . . 78
6.5.1 Comparison to State of the Art . . . 78
6.5.2 Fovea Localization on Normal and Diseased Images . . . 80
6.5.3 Comparison of Multi-Cue Fusion Architectures . . . 80
6.5.4 Performance of Cross-Dataset Experiments . . . 83
6.5.5 Ablation Study . . . 84
6.6 Conclusion and Future Work . . . 87

7 Conclusions and Future Work 89
7.1 Conclusions . . . 89
7.2 Future Work . . . 92
7.2.1 Designing Unified Frameworks . . . 92
7.2.2 Utilizing Unlabeled Data . . . 93

References 95

List of Figures

2.1 Architecture of AlexNet [77]. 12
2.2 Structure of residual block of ResNet [57]. 14
2.3 Architecture of DenseNet [64]. 15
2.4 Architecture of U-Net [118]. 16
2.5 Architecture of DeepLabV3+ [25]. 17
2.6 Architecture of a spatial transformer module [68]. 18
2.7 Architecture of a squeeze-and-excitation block [63]. 18
2.8 Architecture of a block of BAM [103]. 19
2.9 Architecture of CBAM [157]. 19
2.10 Architecture of Transformer [151]. 21
2.11 Architecture of Vision Transformer [41]. 22
2.12 Architecture of TransUNet [21]. 24
2.13 Framework of training a cGAN (Pix2Pix) [67]. 25
2.14 Framework of training CycleGAN [177]. 26

3.1 Karyotype of human chromosomes consisting of 22 autosomes pairs and a pair of sex chromosomes. 30
3.2 Seven types of images utilized in internal backbone extraction. (a) An example of original chromosomes; (b) an approximate central axis; (c) the smoothed central axis; (d) the smoothed central axis divided into 11 parts; (e) 10-point central axis; (f) the internal backbone; (g) the straightened internal backbone with the same length. 32
3.3 Examples of central axis extraction generated by thinning methods and our approach. 34
3.4 Examples of random data augmentation of a chromosome and corresponding internal backbone. 35

3.5 The overall process of the proposed framework for chromosome straightening. (a) The training processes of pix2pix or U-Net (the generator part of pix2pix), where X_B, Y_B are augmented backbones and chromosomes and $B \in \{1, ..., K\}$ where K is the number of augmented image pairs; X_{pred} is the predicted chromosome image through the generator, G_b. (b) The straightening process achieved by the optimal U-Net or generator G_b^*. $X_B^{'}$ and $X_{pred}^{'}$ are the vertical backbone and the straightened chromosome, respectively. . 35

3.6 Three examples of straightening results. From left to right: original images, the geometric method [120, 129], our framework using U-Net and pix2pix. Enlarged regions demonstrate marginally improved details of pix2pix over U-Net. 39

3.7 Training and validation accuracy curves of three CNN models for chromosome type classification (4-fold cross-validation). Shadow regions represent the range over four folds and solid lines represent mean accuracy. 42

3.8 Two examples of failure cases. From left to right: original images, the geometric method [120, 129], our framework using U-Net and pix2pix. . . . 43

3.9 Examples of synthesized results with a series of curved internal backbones (generated by our framework with the pix2pix model). 43

4.1 Overview of the architectures of the proposed framework, including (a) motion transformation generator and (b) our proposed ViT-Patch discriminator; (c) gives the structure comparison of a basic patch discriminator. . . . 48

4.2 Visual results of chromosome straightening generated by different methods. 52

4.3 Comparison of straightening performance (a) of ablation study and (b) on a large-scale dataset ($\mathtt{D}_{large}$). 53

4.4 Comparison of straightening performance based on different *test driving* images selected by random and our proposed SL-matching schemes. 54

5.1 The overall architecture of our proposed Bilateral-ViT network. 59

5.2 The structures of SIG blocks and MFF blocks. The subscript C denotes channel depths. C_{in}, C_{mid} and C_{out} represent channel depths of input, intermediate, and output feature maps for the MFF blocks, respectively. We set C_{mid} of three MFF blocks to small numbers, *i.e.*128, 64, 32, for improving the efficiency of multi-scale feature fusion. 60

5.3 Visual results of fovea localization predicted by different methods. 62

6.1 The overall architecture of our proposed Bilateral-Fuser network. 73

6.2 The structure of BTI module used in our Bilateral-Fuser. It contains TokenLearner, $T\times$ MHSA layers and TokenFuser. The h, w and c are height, width, and channel of the corresponding input features. The n represents the number of learned tokens. 74

6.3 Visualization of mean errors (Y-axis) of different multi-cue fusion models. X-axis is the computational cost (FLOPs). The red and blue markers are results on `PALM` and `Tisu` datasets, respectively. Numbers below the markers are corresponding mean errors. 81
6.4 Visual results of fovea localization predicted by different methods. 82
6.5 Visualizations generated from spatial attention maps, indicating the focus of TokenLearner in BTI module. These visual results have been resized and superimposed onto the the corresponding fundus. 85
6.6 Visualization of mean errors (Y-axis) of ablation studies for reducing and recovering tokens. X-axis is the computational cost (FLOPs). The red and blue markers are results on `PALM` and `Tisu` datasets, respectively. Numbers below the markers are corresponding mean errors. 87

List of Tables

2.1 The configuration of VGG networks, and conv(receptive field size)-(channel number) represents detailed parameters 13

3.1 LPIPS results on different chromosome datasets (mean ± std.). For LPIPS lower is more similar. 40
3.2 Comparison of averaged classification accuracy (4-fold cross-validation) . . . 41
3.3 Comparison of averaged AUC of chromosome type classification (4-fold cross-validation) 41

4.1 Comparison with existing studies, the best and second best results highlighted in bold and underlines. The DCA result is the average score of each experiment with cross-validation. 51

5.1 Comparison of performance on normal and diseased retinal images using the `Messidor` and `PALM` datasets. The best and second best results are highlighted in bold and italics respectively. 62
5.2 Comparison with existing studies using the `Messidor` and `PALM` datasets based on the R rule. The best and second best results are highlighted in bold and italics respectively. 63
5.3 **Top** and **Bottom**: Performance of the ablation study using the `Messidor` and `PALM` datasets respectively. VB refers to the vessel branch. The best and second best results are highlighted in bold and italics. 64
5.4 Performance of cross-dataset experiments. The models used here are exactly those in the **Bottom** of Table 5.3. They were constructed using `PALM` only and generated the following results on `Messidor`. The higher results based on the R, and the lower results based on distance errors, are better. VB refers to the vessel branch. The best and second best results are highlighted in bold and italics respectively. 65

6.1 Configuration Comparison of Bilateral-ViT models and the proposed Bilateral-Fuser. 77

6.2 Comparison with existing studies using the `Messidor` and `PALM` datasets based on the R rule. The best and second best results are highlighted in bold and italics respectively. 79

6.3 Comparison of performance on normal and diseased retinal images using the `Messidor` and `PALM` datasets. The best and second best results are highlighted in bold and italics respectively. 79

6.4 The performance of cross-dataset experiments. The models used here are exactly those selected in Fig. 6.3 (blue markers). **Top** and **Bottom:** The models trained on `Tisu` and tested on `PALM` and `Messidor`, respectively. The best and second best results are highlighted in bold and italics respectively. 84

6.5 Comparison of performance between different inputs for the main and satellite streams on `PALM` and `Tisu`. The best and second best results are highlighted in bold and italics. 84

List of Abbreviations

ML Machine Learning (Page 1)
AI Artificial Intelligence (Page 1)
KNN K-Nearest Neighbors (Page 1)
PCA Principal Components Analysis (Page 1)
SIFT Scale-Invariant Feature Transform (Page 1)
SVM Support Vector Machines (Page 1)
RL Representation Learning (Page 1)
DL Deep Learning (Page 2)
CNN Convolutional Neural Network (Page 2)
RNN Recurrent Neural Network (Page 2)
CV Computer Vision (Page 2)
NLP Natural Language Processing (Page 2)
ILSVRC ImageNet Large Scale Visual Recognition Challenge (Page 2)
MIA Medical Image Analysis (Page 2)
MRI Magnetic Resonance Imaging (Page 3)
CT Computed Tomography (Page 3)
PET Positron Emission Tomography (Page 3)
GAN Generative Adversarial Network (Page 4)
VAE Variational Autoencoder (Page 4)
MHSA Multi-Head Self Attention (Page 11)
ReLU Rectified Linear Unit (Page 12)
VGG Visual Geometry Group (Page 13)
ISBI International Symposium on Biomedical Imaging (Page 16)
STN Spatial Transformer Network (Page 18)
SENet Squeeze-and-Excitation Network (Page 19)
BAM Bottleneck Attention Module (Page 20)
CBAM Convolutional Block Attention Module (Page 20)
ViT Vision Transformer (Page 21)
cGAN Conditional Generative Adversarial Network (Page 24)
SSIM Structural Similarity Index (Page 37)

PSNR Peak-Signal-to-Noise Ratio (Page 37)
LPIPS Learned Perceptual Image Patch Similarity (Page 37)
FID Fréchet Inception Distance (Page 46)
SCSF Single Chromosome Straightening Framework (Page 47)
FOMM First Order Motion Model (Page 47)
PMEM PCA-based Motion Estimation Model (Page 47)
DCA Downstream Classification Accuracy (Page 51)
MFF Multi-scale Feature Fusion (Page 56)
OD Optic Discs (Page 58)
ROI Region of Interest (Page 58)
SIG Spatial Information Guidance (Page 60)
RSU ReSidual U-blocks (Page 61)
PALM Pathologic Myopia Challenge (Page 62)
BTI Bilateral Token Incorporation (Page 67)
SSL Self-Supervised Learning (Page 93)
MVM Masked Visual Modeling (Page 93)

Chapter 1

Introduction

1.1 Overview

Machine Learning (ML) is currently the most rapidly developing area of Artificial Intelligence (AI). ML enables computers to learn and make predictions utilizing limited human labels. It demonstrates the ability to process large amounts of unstructured data and address complex problems by designing mathematical models [124]. Many ML methods have been proposed, such as k-nearest neighbors (KNN) [5], principal components analysis (PCA) [105], scale-invariant feature transform (SIFT) [91] and support vector machines (SVM) [31], for making accurate predictions and decisions.

ML methods have been extensively applied in many domains to solve real-life problems, such as performing disease diagnosis (healthcare), personalizing product recommendations (recommender systems), optimizing pricing strategies (e-commerce), and translating languages (natural language processing) [124]. However, in the traditional ML domain, experts usually use hand-crafted features to perform the target task with their domain-specific knowledge. Although the manually selected features contain meaningful representations, these features cost significant time and effort from the experts to design an appropriate model for a given problem. Thus, traditional ML methods have many drawbacks, *e.g.*, time-consuming, inefficient, highly task-specific biased and requiring a large amount of domain knowledge.

Representation learning (RL), also known as feature learning, is a subfield of ML. It facilitates the learning of meaningful and useful features or representations from input data. The goal of representation learning is to transform data into more compact and expressive representations that capture underlying structures and patterns directly from the raw data. Compared to traditional ML methods, RL aims to automatically learn relevant features from raw data. Moreover, RL saves time, increases efficiency and generalization, and reduces reliance on expert knowledge.

In recent years, advances in deep learning (DL) have accelerated the development of RL. DL utilizes deep architectures, such as convolutional neural networks (CNNs) [77, 133], recurrent neural networks (RNNs) [93, 125] and Transformers [151, 41], to effectively extract more complex and abstract high-dimensional representations. One major advantage of DL-based representation learning is the elimination of the manual feature engineering process. DL can train a model in an end-to-end manner. Researchers typically only need to prepare input data, ground-truth, and specific objective functions based on the type of tasks (*e.g.*, regression/classification, segmentation or generation). During the DL training process, the designed model can automatically learn intrinsic patterns and relationships that may be difficult for human experts to identify manually. In addition, due to its significant superiority in handling large amounts of data in different modalities, DL is rapidly growing in many areas, such as computer vision (CV), natural language processing (NLP), autonomous vehicles, finance and gaming [80, 54, 71, 78, 19].

In particular, DL methods have evolved rapidly in the field of CV since the complexity and amount of digital data have increased greatly with advances in imaging technology. CV is a field that focuses on enabling computers to mimic the human visual system and gain a visual understanding of the world. In the early stages of CV, researchers focused only on low-level image processing problems, such as feature matching and edge detection [80]. Although the initial CNN (LeNet [81]) was applied to real-world handwritten digit recognition task, the successful application of CNN (AlexNet [77]) on the large-scale dataset, ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [121] gathered more of attention. Due to its ability to capture high-dimensional and hierarchical representations from large-scale datasets, DL has achieved remarkable progress in addressing high-level CV tasks, *e.g.*, image segmentation, generation, and object detection [80, 54, 19].

Since DL enables machines to discard handcrafted features and to understand complex and large amounts of visual information, it has facilitated researchers to focus on systems that reduce the need for expert labor while automatically making accurate interpretations, particularly in the field of medical image analysis (MIA). MIA is a field that develops advanced technologies to analyze medical images and clinical data for therapeutic, diagnostic, and research purposes [87]. Since 2013, automated diagnostic systems using deep features have been designed in MIA [17, 144, 107]. In this early stage, deep belief networks and stacked auto-encoders were utilized to analyze 3D brain magnetic resonance imaging (MRI) data for classification. Starting from 2015, the research focus of MIA has been clearly transferred to CNNs [130, 46]. U-Net [118] is the most popular architecture for addressing biomedical image segmentation. It contains an encoder and a decoder, and uses skip connections to preserve detailed features, which is a fundamental structure usually applied in recent DL architecture designs as well.

With advances in imaging technology, more complex medical images become available for use in the MIA field, such as optical microscope, computed tomography (CT), MRI and positron emission tomography (PET) images. These data, in 2D/3D formats, are important modalities for medical imaging and provide unique advantages for monitoring and diagnosing a variety of medical conditions. For example, CT is a type of 3D images utilizing X-ray technology. It provides detailed anatomical information and is valuable for detecting tumors and injuries in many parts, including the brain, chest and abdomen. MRI utilizes a strong magnetic field and radio waves to reveal detailed internal body structures. It provides the excellent demonstration of soft tissues and is therefore commonly used to evaluate organs such as the brain, heart, liver, and kidneys. Since CT exhibits strong contrast for dense structures and MRI provides exceptional soft tissue contrast, they can scan the same organ to provide mutual information or detect different diseases depending on the clinical scenario [14, 104]. These modern image acquisition techniques have produced more high-resolution and complex clinical images with better contrast and clarity, resulting in an increased demand for more experienced experts. However, it takes many years to train an experienced clinician, so the number of clinicians is currently growing at a slower rate than the number of patients [119, 29]. In this case, the demand for DL-based automated diagnosis systems has grown significantly.

Recent studies indicate that DL has achieved state-of-the-art performance on many tasks in MIA. The unique tasks in MIA are mainly categorized as three classes, *i.e.*, (1) regression/classification (rigid medical image registration, computer aided detection and diagnosis), (2) medical image generation (deformable registration, image reconstruction and image enhancement), and (3) medical image segmentation. Beyond these major classes, other medical tasks, such as landmark detection and report generation, are relatively infrequent [176, 128].

The regression/classification tasks aim to predict continuous numerical values or discrete predefined classes from input medical images. These values can be transformation parameters, coordinates or probability distribution of interested diseases, such as aligning medical images taken from different modalities [106] or detecting and diagnosing lung nodules in chest CT scans [88].

In contrast to regression/classification tasks, medical image generation and segmentation tasks require forming visual representations and generating dense predictions. The dense prediction tasks focus on generating predictions for each element, requiring efficient modeling of spatial relationships and processing of high-dimensional outputs. The medical image segmentation task that predicts each pixel can provide accurate boundaries and capture fine-grained details in the segmented regions, which is an important prerequisite [176]. For example, researchers have used DL technology to segment brain tumors in MRI scans [115]. DL models can accurately segment tumor regions, enabling clinicians to precisely measure tumor size, monitor its progression, and plan appropriate treatment strategies. The medical image generation task focuses on synthesizing new medical images that exhibit realistic and clinically relevant features. DL models can learn the underlying patterns and distributions of the input dataset to generate novel images. Researchers typically customize architectures and design novel loss functions from classical generative adversarial networks (GANs), variational autoencoders (VAEs) or image-to-image translation networks based on the target medical dataset to facilitate medical data augmentation, deformable registration across modalities, super-resolution and motion artifact reduction [79, 39, 83, 173, 146].

1.2 Challenges and Motivations

In this Ph.D. research, we look into designing RL/DL-based methods to improve the performance of dense prediction tasks in MIA field. Unlike traditional ML algorithms, RL/DL-based MIA methods simplify the process to an end-to-end approach that can automatically extract high-dimensional and hierarchical representations without the need for handcrafted features. However, fewer advanced DL and RL architectures have been designed and applied in the MIA field compared to the general CV domain due to the following three unique challenges:

1) The number of high-quality labeled medical images is limited. In DL, a substantial amount of labeled data is required to train a robust neural network. Advanced images, such as CT, MRI and PET, significantly increase the size of medical image datasets. While these datasets provide valuable resources for training and evaluating DL models, annotating these datasets is extremely time-consuming and expensive, especially in the MIA field where expert labeling is required. Although a large amount of clinical data exists, various data acquisition protocols and privacy concerns at different hospitals have led to isolated medical images, resulting in less available data and ground truth. For some rare diseases, the size of datasets, labels and experts are all insufficient. Even some common diseases, the training of experienced clinicians takes several years, much slower than the growth of medical data, so experienced clinicians are often inadequate to provide enough labels compared to the general CV domain.

2) With limited labeled medical data, training DL models faces the problem of overfitting. Overfitting occurs when a model is too specialized in the training data. With limited labeled data, DL models do not have enough examples to effectively learn the underlying patterns and may excessively rely on specific examples. In this case, the model may have poor generalization and reduced performance on unseen data.

3) Since the feature extraction and integration process of DL models is usually a black box, the generated results lack convincing evidence. Especially in MIA field, excellent interpretation is crucial. Due to transparency, explainability and ethics considerations, DL models are currently used only to provide auxiliary predictions for clinician's final diagnoses and treatments. Therefore, models with excellent interpretability are desirable for clinicians to make reliable diagnoses while saving time.

To address the above challenges, we aim to design robust representation learning architectures for MIA tasks. We shall look into two main directions of MIA, namely, medical image generation and segmentation. The reasons are three-fold. **First**, since medical image generation and segmentation tasks generate dense predictions, preparing labels for these tasks is much harder than regression tasks, leading to fewer labels for training. Therefore, it is important to improve the *training efficiency* of designed representation learning models, which can generate robust results based on a limited number of labels. **Second**, except for the limited labels, the architecture for image generation and segmentation basically consists of an encoder, which extracts features from the input, and a decoder, which reconstructs the features to the same size as the input. This complex architecture contains more parameters to be optimized, which may exacerbate the overfitting problem. Therefore, we aim to design architectures with excellent *generalization capability* which is essential for downstream clinical applications. **Third**, since both the semantic content and edges of dense predictions are important for diagnosis, these tasks require more interpretable evidence in order for clinicians to understand the factors that influence outcomes. In this case, the trust of clinicians and patients in the model predictions increases, resulting in better cooperation between human experts and computer-assisted intervention systems. Therefore, the *interpretability* of RL models applied to MIA tasks is crucial to develop and employ automated diagnosis AI systems.

We study a specific topic in each dense prediction task to design advanced representation learning models to improve *training efficiency*, *generalization capability* and *interpretability.* For medical image generation, we focus on the topic of chromosome straightening. In this topic, we need to generate a straightened chromosome based on the corresponding input curved chromosome image. In addition to the above challenges, the chromosome straightening task has two unique challenges compared to general image generation tasks. First, training images and corresponding ground-truth used to train a generative model are insufficient. Due to random mutation, structural rearrangement, and different laboratory conditions, it is almost impossible to find two visually identical chromosomes with the same dyeing condition but different curvatures under microscopes. Second, due to the non-rigid nature of chromosomes, distorted chromosome details and shapes after the straightening process leads to different semantic contents in the same position.

For medical image segmentation, we focus on the topic of fovea localization. We transfer this task from localization to small region segmentation as a dense prediction task. Accurate boundaries of the target region are more important for treatment planning, disease monitoring and diagnosis than for predicting localization points. The transferred task is also different from general medical image segmentation task such as colonic polyp segmentation and tumor segmentation. The fovea is an anatomical landmark of the retina, and is a small region with the dark appearance that is indistinguishable from the color intensity of the surrounding retinal tissue. Meanwhile, its local anatomical landmarks (*e.g.*, blood vessels) are also absent in the vicinity of the fovea. In this case, this task requires that the proposed model has an ability to incorporate global features for identifying fovea region with high efficacy, and to overcome the challenges of the occurrence of retinal diseases and non-standard fovea locations. Therefore, this fovea region segmentation task can reveal the representation learning capability of the designed model.

1.3 Contributions

To address the above challenges of efficient representation learning in MIA, we first present a novel work that outperforms existing methods in each dense prediction task (*i.e.*, Chapter 3 for medical image generation and Chapter 5 for medical image segmentation). Afterwards, we propose a subsequent study (*i.e.*, Chapter 4 for medical image generation and Chapter 6 for medical image segmentation) to address the limitations of the corresponding initial method to achieve state-of-the-art performance, better *training efficiency*, *generalization capability* and *interpretability* for each task. The main contributions of this thesis are summarized as follows,

- In Chapter 3, as the first work in medical image generation (*i.e.*, chromosome straightening), our study presents a novel framework for chromosomes straightening using image-to-image translation. The framework addresses the problem of input deficiency by proposing a pertinent augmentation approach to simultaneously increase the variability of curvatures from chromosomes and corresponding labels. We apply two effective image-to-image translation architectures, a U-shape network and conditional GAN (Pix2Pix), to show the efficacy and robustness of our straightening framework. We demonstrate that chromosomes straightened using our framework outperform original curved ones and chromosomes straightened using geometric algorithms in terms of accuracy of chromosome type classification.

- In Chapter 4, our subsequent work proposes a novel framework for robust chromosome straightening using Vit-Patch GAN. The framework consists of a self-learned generator and a Vision Transformer-based patch discriminator. The generator learns the motion representation of chromosomes for straightening, while the discriminator helps to retain more shape and banding pattern details in the straightened chromosomes. The proposed method addresses the challenges of chromosome straightening, such as unavailability of training images, distortion of chromosome details and shapes after straightening, and poor generalization capability. Our method achieves state-of-the-art performance in preserving chromosome details and has excellent generalization capability to a large size dataset.

- In Chapter 5, as the first work for medical image segmentation (*i.e.*, fovea localization), this paper introduces a novel approach, Bilateral-ViT, for robust fovea localization in retinal images. We use a vision-transformer-based architecture that incorporates global image context and blood vessel structure to achieve robust fovea localization. The proposed Bilateral-ViT outperforms existing methods and achieves state-of-the-art results on two public datasets (`Messidor` and `PALM`), especially in diseased retinal images. Its generalization capability is also demonstrated by cross-dataset experiments. The proposed approach has potential applications beyond retinal analysis and can be used in other areas of medical image segmentation.

- In Chapter 6, our subsequent work proposes a novel dual-stream deep learning architecture, called Bilateral-Fuser, for accurate fovea localization in retinal images. Bilateral-Fuser incorporates long-range connections of both global features of retinal images and anatomical landmarks (*i.e.*, vessel distributions) to achieve robust fovea localization. It also introduces a spatial attention mechanism in the dual-stream encoder to extract and fuse self-learned anatomical information, resulting in better interpretability. The architecture focuses more on features distributed along blood vessels and significantly reduces the computational cost by reducing token numbers. Comprehensive experiments demonstrate that the proposed Bilateral-Fuser achieves new state-of-the-art performance on two public datasets and one large-scale private dataset at only 25% of the computational cost of our previous architecture, Bilateral-ViT. Moreover, Bilateral-Fuser is more robust on both normal and diseased retina images and has better generalization capacity in cross-dataset experiments.

The aforementioned contributions have been published or peer-reviewed in the following journals and conference proceedings.

Published:

- Chapter 3:

  Song, S., Huang, D., Hu, Y., Yang, C., Meng, J., Ma, F., Coenen, F., Zhang, J. and Su, J., 2021, October. A novel application of image-to-image translation: chromosome straightening framework by learning from a single image. In 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-9). IEEE.

- Chapter 4:

  Song, S.*, Wang, J.*, Cheng, F.*, Cao, Q., Zuo, Y., Lei, Y., Yang, R., Yang, C., Coenen, F., Meng, J., Dang, K. and Su, J., 2022. A Robust Framework of Chromosome Straightening with ViT-Patch GAN. arXiv preprint arXiv:2203.02901. (This paper has been accepted by 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI).)

- Chapter 5:

  Song, S.*, Dang, K.*, Yu, Q., Wang, Z., Coenen, F., Su, J. and Ding, X., 2022, March. Bilateral-ViT for Robust Fovea Localization. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). (This paper is selected as the best-paper-award finalist.)

Under Peer-review:

- Chapter 6:

  Song, S., Wang, J., Wang, Z., Wang, S., Su, J., Ding, X. and Dang, K., 2023. Bilateral-Fuser: A Novel Multi-cue Fusion Architecture with Anatomical-aware Tokens for Fovea Localization. arXiv preprint arXiv:2302.06961.

1.4 Outlines

The rest of my thesis is organized as follows.

Chapter 2 describes the development and key designs of deep learning-based representation learning. These designs are fundamental structures utilized in the following chapters.

Chapter 3 presents a novel image-to-image translation-based chromosome straightening framework that transforms the task of straightening into learning mapping dependencies from a randomly augmented backbone to the corresponding chromosome. The framework allows the generation of straightened chromosomes from vertical backbones and outperforms geometric methods in more realistic images with uninterrupted banding patterns.

Chapter 4 proposes an advanced framework, named ViT-Patch GAN, for robust chromosome straightening, as a subsequent study of Chapter 3. The proposed framework includes a self-learned motion transformation generator and a vision-transformer-based patch discriminator, which together retain more shape details and banding patterns in the straightened chromosomes. This novel framework achieves state-of-the-art performance in preserving chromosome details and has excellent generalization for large datasets.

Chapter 5 proposes a novel two-branch architecture, Bilateral-ViT, which incorporates features of retinal images and vessel distributions for robust fovea localization in retinal images. The proposed method achieves state-of-the-art results on two public datasets, `Messidor` and `PALM`.

Chapter 6 proposes a novel dual-stream architecture, Bilateral-Fuser, for fovea localization. As a subsequent study of Chapter 5, Bilateral-Fuser incorporates long-range connections and global features using retina images and vessel distributions with self-learned anatomical-aware tokens. The proposed architecture achieves state-of-the-art performance on two public datasets and one large-scale private dataset, demonstrating excellent robustness, generalization capability and interpretability on both diseased retina images and hard cases.

Chapter 7 concludes the main results of my Ph.D. thesis and provides possible directions in future research.

Chapter 2

Literature Review

Deep learning-based representation learning methods have revolutionized the CV field. Compared to traditional machine learning algorithms, such as KNN [5], PCA [105], SIFT [91] and SVM [31], representation learning algorithms focus on automatically learning representations directly from raw data, eliminating or reducing the need for manual feature engineering. Modern representation learning models are typically deeper and have more complex architectures, such as deep neural networks, to learn hierarchical representations. These architectures can capture more compact and higher-level features, and are potentially more suited for downstream tasks [82]. Modern deep learning-based representation learning has gained a significant impact in almost all CV-related fields, such as image classification/recognition, object detection, image segmentation and generation. Deep learning has become an important approach to representation learning, pushing the boundaries of visual understanding and analysis in CV.

In this chapter, we focus on describing the development and key designs of deep learning-based representation learning methods. The first two sections review the classical architectures of CNN (Section 2.1) and the design of skip connections (Section 2.2). These sections are the fundamental structures widely used in our following chapters (Chapter 3 to 6). The next section (Section 2.3) describes two major popular branches of attention mechanisms, *i.e.*, CNN-based attention and multi-head self attention (MHSA). In particular, MHSA is exploited in Chapter 4, 5, 6 to capture global connectivity and improve representation learning performance. Moreover, the spatial attention, one type of CNN-based attentions, is applied in Chapter 6 to construct anatomical-aware tokens for MHSA while reducing the computational cost. The last section (Section 2.4) reviews some impor-

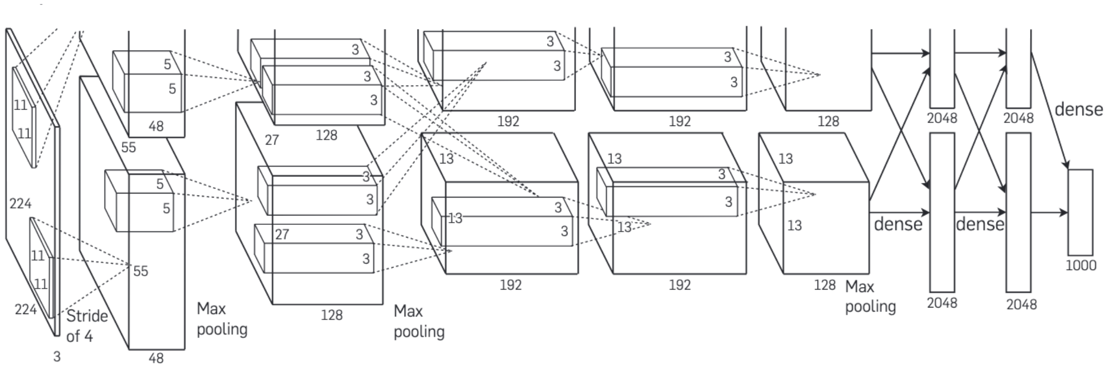


Figure 2.1: Architecture of AlexNet [77].

tant conditional generative adversarial networks for image generation. These methods are relevant to Chapter 3 and 4 for dense prediction (image generation) task in MIA.

2.1 Classical Architectures in CNN

In the early stages of CNN development, AlexNet [77] and VGG networks [133] are some of the most important architectures. AlexNet is the first popular CNN architecture to achieve significant improvements in image classification in the ILSVRC 2012. AlexNet achieves state-of-the-art performance of top-1 and top-5 on the ILSVRC 2010 and the ILSVRC 2012 datasets, surpassing previous models by a large margin. AlexNet attracted interest in the application of deep learning to image classification tasks, and then became a classical and widely used architecture.

As shown in Figure 2.1, AlexNet consists of five convolutional layers and three fully-connected layers. In addition to the combination of these layers, AlexNet has three key innovations, the application of Rectified Linear Unit (ReLU) nonlinearity, training on multiple GPUs, and the use of pooling layers, which prevent researchers from training big models on only a single GPU and allow the network to extract more complex features from the input images. AlexNet also introduces data augmentation, such as random cropping and horizontal flipping. Based on these varied data, AlexNet helps prevent overfitting and improves the robustness of the neural network. AlexNet also uses dropout regularization to randomly drop out neurons during training to reduce overfitting and improve generalization.

Table 2.1: The configuration of VGG networks, and conv(receptive field size)-(channel number) represents detailed parameters

	Block1[m]	Block2[m]	Block3[m]	Block4[m]	Block5[m]	FC Block[s]
VGG16	conv3-64	conv3-128	conv3-256	conv3-512	conv3-512	FC-4096
	conv3-64	conv3-128	conv3-256	conv3-512	conv3-512	FC-4096
			conv3-256	conv3-512	conv3-512	FC-1000
VGG19	conv3-64	conv3-128	conv3-256	conv3-512	conv3-512	FC-4096
	conv3-64	conv3-128	conv3-256	conv3-512	conv3-512	FC-4096
			conv3-256	conv3-512	conv3-512	FC-1000
			conv3-256	conv3-512	conv3-512	

[m] There is a maxpooling at the end of this block.
[s] There is a softmax at the end of this block.

VGG networks have become the standard CNN architecture and backbone, and achieved state-of-the-art performance in a range of computer vision tasks, including image classification and object detection [133, 117]. It is a deep CNN architecture with multiple layers from the Visual Geometry Group (VGG) at the University of Oxford. With reference to the detailed architecture given in Table 2.1, VGG16 consists of 16 layers, while VGG19 consists of 19 layers, and both networks use only ×3 with a stride of 1 convolutional filters throughout the network, while AlexNet uses 11×11 kernel with a stride of 4. Another innovations of VGG includes the use of maxpooling to downsample the feature maps and reduce the input dimensionality, and the use of fully connected layers at the end of the network to capture high-level representations of the input. In terms of accuracy, VGG16 and VGG19 outperform AlexNet on the validation and test sets of the ImageNet dataset, and they are the top models in the ILSVRC 2014 competition.

AlexNet and VGG networks are CNN architectures that learn deep representations of inputs, and are milestones in deep learning architectures that greatly surpassed the performance of machine learning algorithms. They are some of the most important CNN architectures that have made significant contributions to the field of CV [54]. Their key features and innovations, such as the use of convolutional layers with large/small receptive fields, normalization, pooling layers, and fully connected layers, remain fundamental components and have been widely adopted and modified in subsequent CNN models.

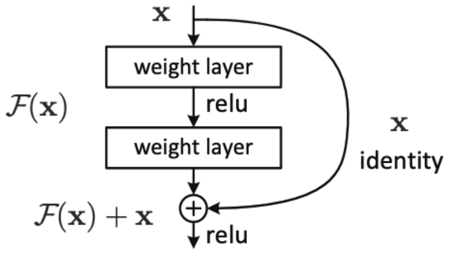


Figure 2.2: Structure of residual block of ResNet [57].

2.2 Skip Connections

Skip connections, also known as residual connections, are a type of connection in CNN models that allow the input to bypass one or more layers as a shortcut and be added or concatenated to the output of a subsequent layer. This design has significantly improved the performance of computer vision tasks, such as image recognition and image segmentation. Some of the most popular deep learning architectures utilize skip connections in both image recognition tasks (*e.g.*, ResNet [57] and DenseNet [64]) and dense prediction tasks (*e.g.*, U-Net [118] and DeepLab [22, 23, 24, 25]).

The residual network (ResNet) architecture [57] was first proposed and applied to image recognition tasks to address the challenge of training deeper neural networks. To address the problem of vanishing/exploding gradients, ResNet consists of a series of residual blocks, each of which contains several convolutional layers with skip connections between them. The residual features are the input features of each block and are added directly to the output features of convolutional layers (Figure 2.2). By using skip connections, ResNet can be trained more easily since it only needs to optimize the residual mappings instead of the original, unreferenced mappings. ResNet also introduces the use of global average pooling, which reduces the number of parameters in the model and improves its generalization performance.

Due to the design of the residual block, ResNet architecture is the first extremely deep CNN with both excellent robustness and generalization capability. In terms of number of layers, ResNet can be named as ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152. ResNet achieves state-of-the-art performance in the ILSVRC 2015 classification competition (1^{st} place). The ensemble of ResNet achieves a significant improvement over previous models. Furthermore, the ResNet has been further developed as a backbone

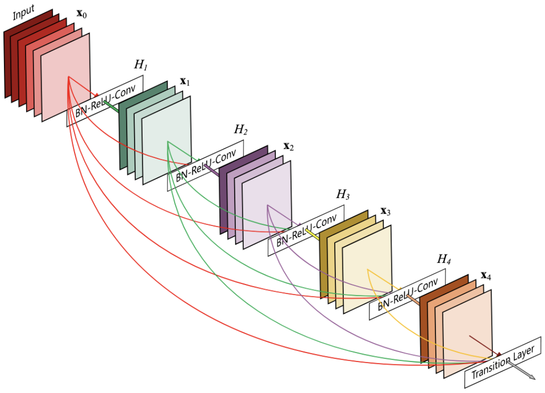


Figure 2.3: Architecture of DenseNet [64].

for various computer vision tasks and achieved state-of-the-art performance (1^{st} place in ImageNet localization, ImageNet detection, COCO segmentation and COCO detection of the same year).

DenseNet [64] is another important architecture using skip connections in image recognition field. DenseNet proposes dense connectivity pattern which makes it apart from other popular architectures (*e.g.*, ResNet). In DenseNet, each layer is directly connected and concatenated to every other layer in a feed-forward neural network (Figure 2.3). This dense connectivity improves feature reusability and enhances gradient flow throughout the network, resulting in improved information and gradient flow. The dense connections also alleviate the problem of vanishing gradients, enabling the model to effectively learn from both shallow and deep layers. DenseNet achieved state-of-the-art performance on many image recognition benchmark datasets, such as ImageNet, CIFAR-10, and CIFAR-100. Furthermore, DenseNet has also been applied to dense prediction tasks, such as semantic segmentation and object detection, demonstrating its versatility and effectiveness in different computer vision domains. However, one limitation of DenseNet is its relatively high memory consumption during training. This constraint may require memory management to train DenseNet on certain devices.

U-Net [118] is a notable model that uses skip connections and is originally designed for a dense prediction task (medical image segmentation). To generate segmentation results, the structure of U-Net consists of an encoder and a decoder for downsampling and upsampling of features. As shown in Figure 2.4, its skip connections are between the corresponding

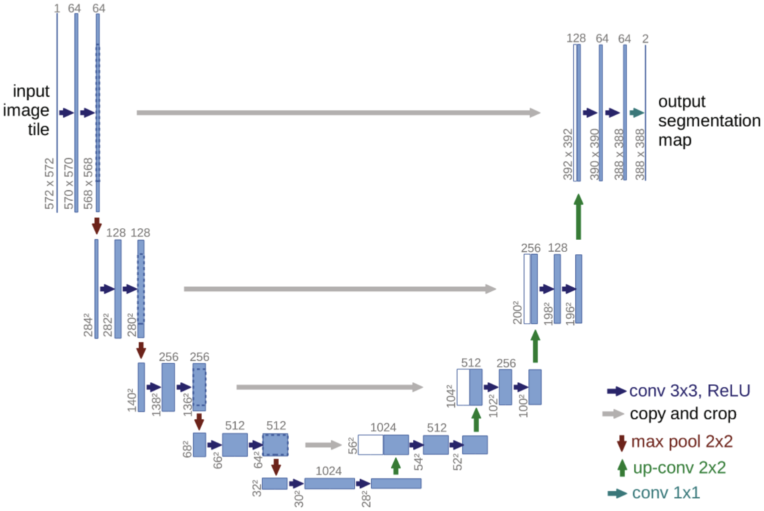


Figure 2.4: Architecture of U-Net [118].

layers of the encoder and decoder. The skipped features are concatenated (instead of element-wise summing) to the corresponding upsampled features, resulting in a doubling of the number of channels. Without skipped connections, the result of upsampling directly from the bottleneck does not recover detailed information. Skip connections enable the network to preserve spatial information and improve the accuracy of segmentation. By removing skip connections, features upsampled directly from the bottleneck structure are also decreased by the lack of spatial information.

The design of U-Net has achieved state-of-the-art results on various biomedical image segmentation tasks, including cell detection and organ segmentation. Using data augmentation, U-Net can be trained on datasets with only a small number of samples and has excellent generalization capabilities. At IEEE International Symposium on Biomedical Imaging (ISBI) 2015, U-Net took first place in the cell tracking challenge by a significant margin.

DeepLab architecture [22, 23, 24, 25] is another model that utilizes skip connections to improve the performance of semantic segmentation. From DeepLabv1 [22], the design of atrous convolution (also known as dilated convolution) is proposed to increase the receptive field of the network without increasing the number of parameters. DeepLabv3+ [25] uses a spatial pyramid pooling module to extract and combine multi-scale features. This module allows the model to combine high level semantic features with low level spatial details

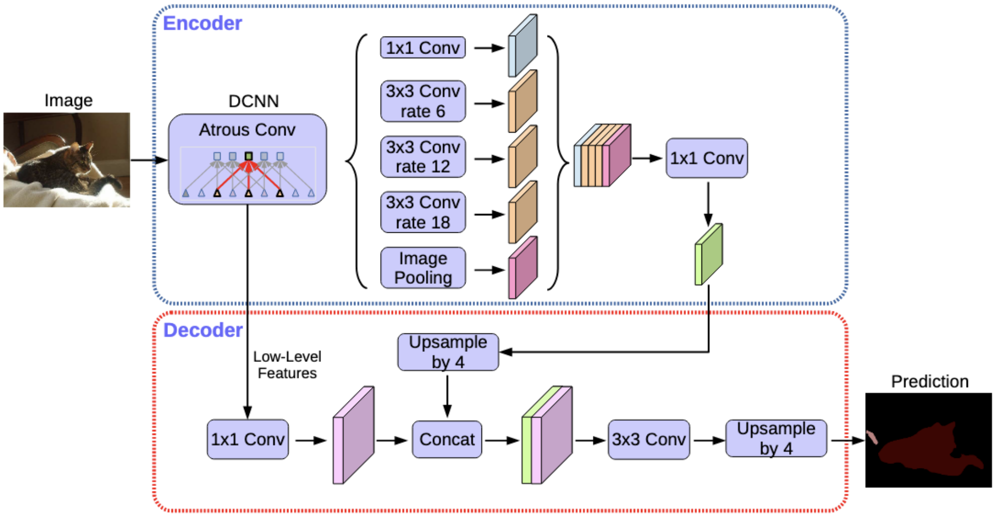


Figure 2.5: Architecture of DeepLabV3+ [25].

to improve the accuracy of segmentation. DeepLabv3+ also utilizes skip connections to pass these features from the encoder to the decoder to refine the segmentation output (Figure 2.5). It achieved state-of-the-art results on many semantic segmentation datasets, such as PASCAL VOC 2012 [43] and Cityscapes [30], outperforming other models of the time.

The structure of skip connections enables CNNs to learn deeper and more useful representations of the input. It has become a common design for current deep learning models, especially in computer vision tasks [19]. ResNet, U-Net, DenseNet and DeepLab architectures are some of the most popular and fundamental models that use skip connections to improve performance and address the vanishing gradient problem. By using shortcuts from input features to pass through one or more layers and be added/concatenated to the corresponding target features, skip connections enable networks to learn representations more easily and efficiently, and to obtain state-of-the-art results in different tasks. Therefore, the skip connections are widely utilized in the architectures proposed in the following chapters (Chapter 3 to 6) to improve model performance and training efficiency of both image generation and segmentation tasks.

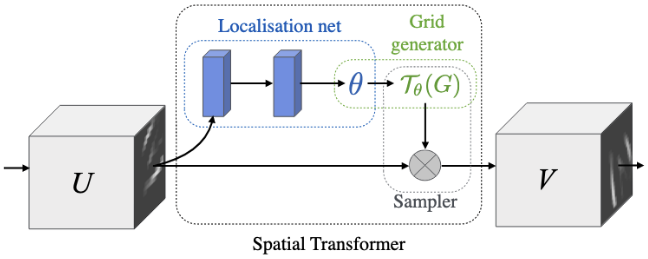


Figure 2.6: Architecture of a spatial transformer module [68].

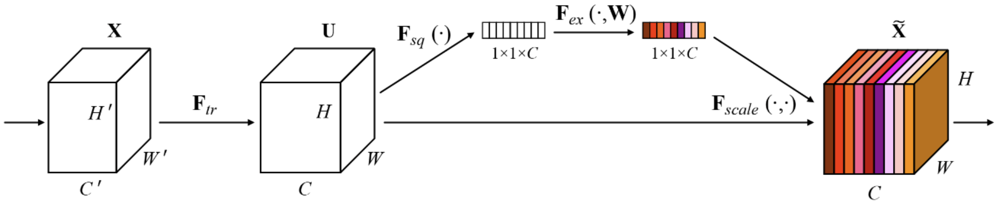


Figure 2.7: Architecture of a squeeze-and-excitation block [63].

2.3 Attention Mechanisms

2.3.1 CNN-Based Attention

CNN-based attention is a type of attention mechanisms that are used in various deep learning models to improve their performance in computer vision tasks. The attention mechanism enables neural networks to focus on important regions of the input image and assign higher weights to them. The weights of CNN-based attention are learnable parameters that give more focus on more relevant features and ignore irrelevant ones. The CNN-based attention mechanism is usually employed to learn spatial, channel and combined attention.

Spatial Transformer Network (STN) [68] is one of the most important models for applying attention mechanism in CNNs. As shown in Figure 2.6, STN consists of a localization network and a transformation network, both of which use convolutional layers to learn spatial transformations of the input. The localization network learns the parameters of an affine transformation, while the transformation network performs spatial transformations on the input, such as translation, rotation, scaling and warping. The main innovation of STN is the use of a differentiable method that enables the network to transform the in-

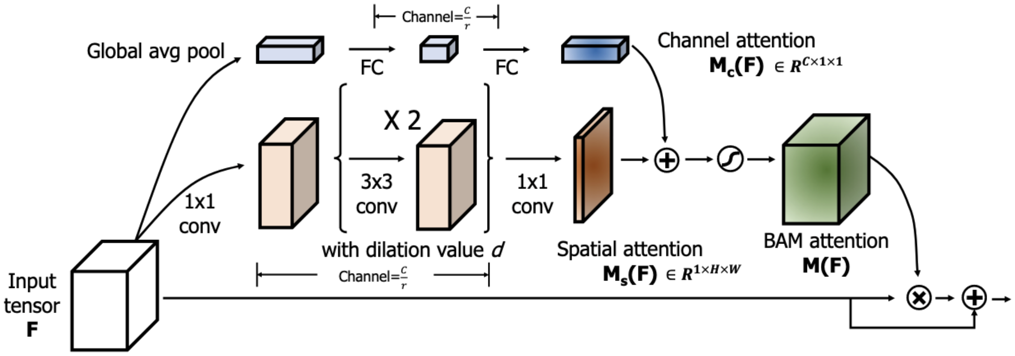


Figure 2.8: Architecture of a block of BAM [103].

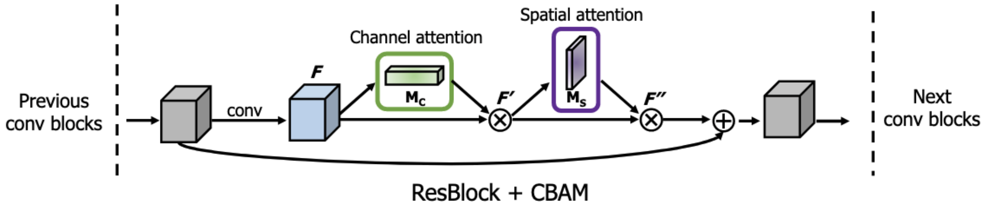


Figure 2.9: Architecture of CBAM [157].

teger as the focused region before classification prediction, thus improving computational efficiency. In addition, the training process of STN does not require any additional supervision or modification. STN demonstrates performance improvements in the accuracy for computer vision tasks. STN was trained on MNIST and CIFAR-10 datasets and it achieved higher accuracy on these datasets than models without the attention mechanism.

Squeeze-and-Excitation Network (SENet) [63] is a popular CNN model that exploits the channel attention mechanism. SENet consists of squeeze-and-excitation blocks, each of which contains a squeeze operation and an excitation operation (Figure 2.7). The squeeze operation reduces the dimensionality of the feature map, while the excitation operation learns the attention weight of each channel. The major contribution of SENet is the design of channel-wise attention which enables the network to selectively emphasize informative channels while suppressing irrelevant ones. The SE block can be considered as a plug-in that is directly applied to ResNets (*e.g.*, SE-ResNet-50 and SE-ResNet-152) as a backbone. SENet achieved outperformance on many tasks, especially winning first place in ILSVRC 2017 image classification competition.

Bottleneck Attention Module (BAM) [103] and Convolutional Block Attention Module (CBAM) [157] are also important CNN-based attention designs. Both structures are proposed in the same year (2018) and both effectively combine spatial attention and channel attention to improve the performance of various computer vision tasks, such as image classification and object detection. They are both designed as lightweight and general modules that can be seamlessly applied to CNN models.

The BAM is designed to generate spatial and channel attentions in parallel in the CNN architecture. As shown in Figure 2.8, in each BAM, the channel attention is generated by the global average pooling and FC layers, and the spatial attention is generated by the continuous standard/dilated convolutional layers. The spatial and channel attentions are then combined to generate a BAM attention of the same size as the input tensor. Finally, the BAM attention is multiplied with the input tensor and added to the skip-connected input tensor. BAM can be applied to every bottleneck before the pooling layer. In terms of performance, all backbone networks (*e.g.*, ResNet [57], ResNeXt [160], MobileNet [62] and SqueezeNet [66]) with integrated BAM showed improvements in both classification and object detection results.

Different from BAM, CBAM is designed to utilize spatial and channel attentions in a sequence order (Figure 2.9). It can be directly integrated into convolutional blocks, such as a residual block. The input feature first learns channel attention and subsequently utilizes spatial attention. In both attention modules, maxpooling and average pooling are utilized to extract features from different aspects. Comprehensive experiments and ablation studies show that CBAM (channel+spatial attentions) is superior to SENet (only channel attention) [157].

CNN-based attention has become a common design in deep learning models for improving performance and interpretability, especially in computer vision tasks. By enabling the network to focus on important regions (spatial, channel-wise or combined) of the input and assigning higher weights, CNN-based attention mechanisms enable networks to learn more effectively and achieve state-of-the-art results on various datasets.

2.3.2 Multi-Head Self-Attention and Transformer

Multi-Head Self-Attention (MHSA) and Transformer architectures [151] are powerful deep learning techniques that have made significant contributions in recent years for their ability to connect long-range dependencies and integrate global features. MHSA and Transformer

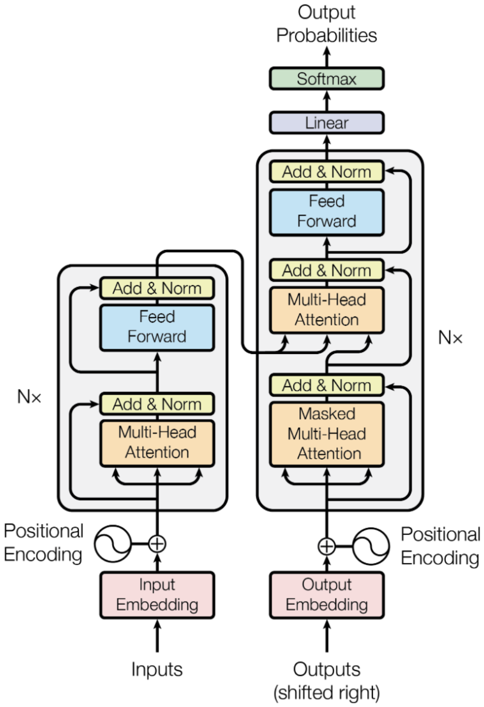


Figure 2.10: Architecture of Transformer [151].

are originally used in the field of NLP and then applied to computer vision tasks by Vision Transformer (ViT) [41]. Transformer architecture only utilizes self-attention blocks (*i.e.*, MHSA) to construct long-range dependencies of different parts of the input sequence to generate outputs. MHSA blocks enable the network to weigh the importance of different words simultaneously, addressing the challenge of poor performance in processing long sequential data.

Transformer architecture [151] is a pure attention-based neural network architecture that eliminates the need for RNNs and CNNs to model sequential data. As shown in Figure 2.10, Transformer includes an encoder and a decoder, both consisting of several attention layers. The encoder receives the input word embeddings and processes them in parallel. The decoder also incorporates the features of input sequences and captures the relevant information using MHSA mechanism. Each layer has two sub-layers: a multi-head self-attention layer and a fully connected feed-forward layer. In each MHSA layer, every word is projected as three vectors, *query*, *key* and *value*. Subsequently, the *query*

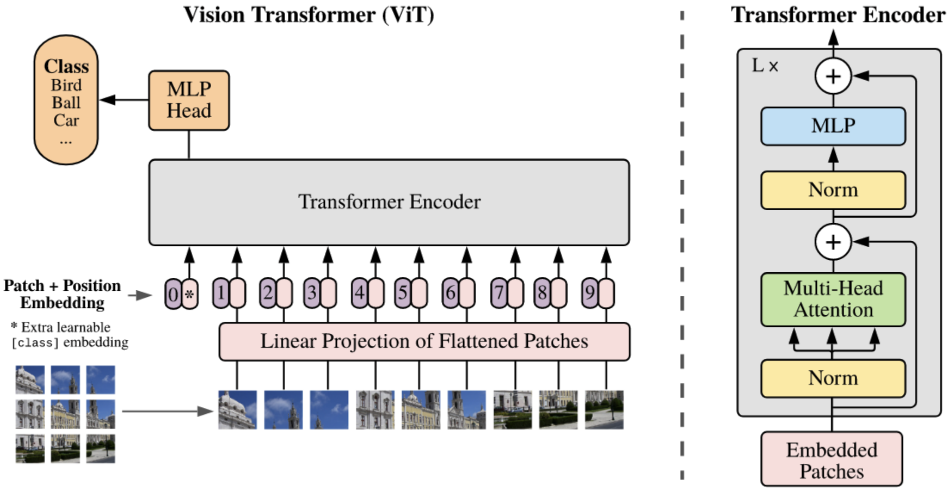


Figure 2.11: Architecture of Vision Transformer [41].

performs multiplication with each *key* (all pairs of positions) simultaneously to compute the similarity of the two embeddings as an attention score. This score is utilized to assign weights to the *value* of each embedding regardless of their distance. The calculation of self-attention is expressed as

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

where Q, K and V are packed *queries*, *keys* and *values* of embeddings for parallel computation. Transformer has achieved state-of-the-art results on various NLP tasks, including language modeling and machine translation. Experiments on two machine translation tasks demonstrate that Transformer outperforms traditional models by a large margin, while requiring significantly less training time due to parallel computation. This architecture represents significant outperformance relying only on the attention mechanism, and its success demonstrates that the MHSA mechanism is an important tool for language modeling.

MHSA and Transformer have also revolutionized the CV field. The ViT architecture [41] is the first pure Transformer model for image classification tasks. Prior to ViT, many studies had attempted to apply self-attention mechanisms to CNNs. However, due to the computational requirements, ResNet and its variants still achieve state-of-the-art performance. ViT applies the Transformer architecture which is originally utilized for NLP tasks directly to CV tasks with as few changes as possible, with the goal of eliminating model differences between the NLP and CV tasks. The proposed ViT utilizes a standard Transformer encoder (Figure 2.11), which uses an initial patch embedding layer to convert the patches of the input image into a sequence of vector representations (tokens). After a classification token and position embeddings are added, these tokens are processed by a series of Transformer blocks. The self-attention mechanism enables ViT to capture long-range dependencies between patches.

Due to incorporating global features and long-range connections from the entire input image, ViT has achieved state-of-the-art performance on many benchmark datasets, such as CIFAR10/100 [75], Oxford Flowers-102 [99], ImageNet-1K/21K [37]. Experimental results show that ViT, an architecture constructed on pure attention mechanism, achieves comparable or better accuracy than state-of-the-art convolutional networks, such as ResNet and EfficientNet [148].

The ViT architecture demonstrates the effectiveness of Transformers in image recognition at scale, and its success led to the exploration and adaptation for multi-modal architectures, where it can handle both visual and textual inputs. Recently, some models, such as GPT series of models [113, 114, 18] (proposed by OpenAI), BERT [38] and LaMDA [149] (proposed by Google), contain variants of the MHSA and Transformer architectures to achieve a breakthrough in large language models and state-of-the-art performance on various benchmarks. With ViT demonstrating the potential in bridging the gap between visual and textual inputs, many multi-modal architectures [111, 42, 162] are proposed to facilitate interactions between multiple modalities, aligning them and capturing their interdependencies thought tokens. The design of ViT opens up new possibilities for integrating visual and textual information in a unified framework, resulting in improved performance on multi-modal tasks.

In addition to image classification/recognition tasks, the architectures of MHSA and ViT have also been designed for a dense prediction (image segmentation) task. TransUNet [21] is the first segmentation work that combines the advantages of U-Net and ViT architectures. Similar to U-Net, TransUNet was originally applied to medical image segmentation tasks. As given in Figure 2.12, the proposed framework also includes an encoder and a decoder. The encoder utilizes a hybrid version of ViT, and the input image is first tokenized by a standard CNN architecture (*e.g.*, ResNet). The tokens are then fed into the Transformer to extract global connectivity and long-range dependencies,

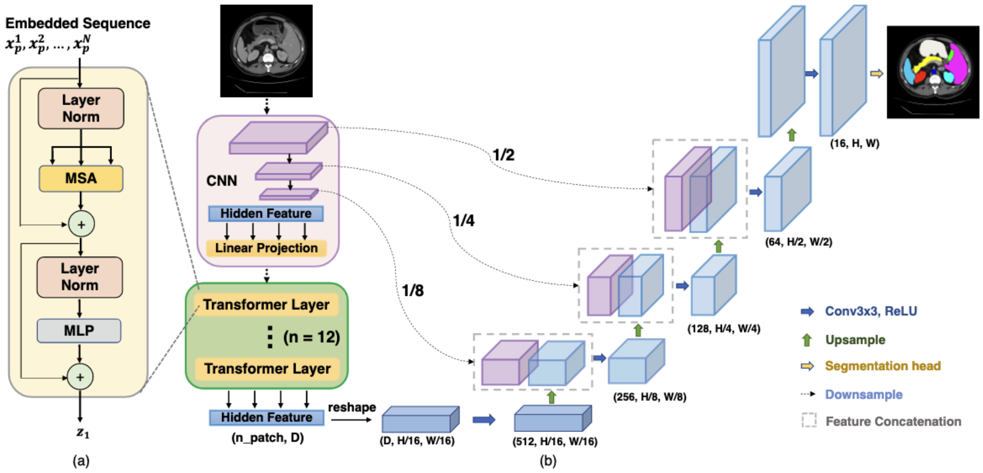


Figure 2.12: Architecture of TransUNet [21].

which is a limitation of U-Net due to its intrinsic locality of convolution operations. The decoder upsamples the encoded features from the bottleneck and concatenates them to the corresponding skip-connected CNN features. This hybrid design improves the utilization of low-level details while still utilizing Transformer architecture for modeling global context. Compared to pure CNN-based U-shape architectures (*e.g.*, U-Net [118] and Att-nUNet [100]), TransUNet achieves superior performance and more accurate segmentation results. TransUNet provides an alternative framework for medical image segmentation with significantly better performance and high efficacy of representation learning.

All in all, ViT and TransUNet are widely applied for recognition and dense prediction tasks in modern neural networks design. Due to the abovementioned advantages, the ViT and TransUNet architectures are utilized in Chapter 4, 5 and 6 to incorporate long-range dependencies across entire images. In Chapter 6, the MHSA mechanism is used to globally fuse and integrate features from multiple cues. Furthermore, since the computational cost of MHSA blocks is quadratic in the number of tokens, spatial attention is utilized to MHSA blocks, which significantly reduces the computational cost and improves the performance.

2.4 Conditional Generative Adversarial Network (cGAN)

Another dense prediction task, image generation, is a popular research field in deep learning that focuses on synthesizing new images with high fidelity. Significant progress achieved in

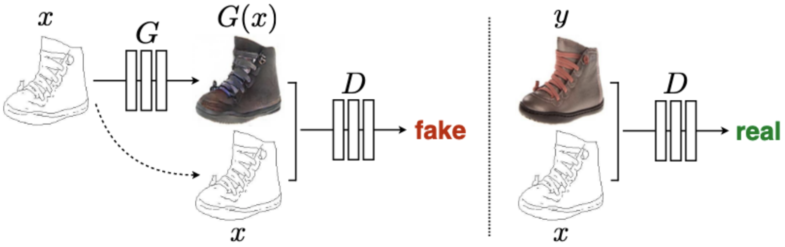


Figure 2.13: Framework of training a cGAN (Pix2Pix) [67].

this area with the development of various architectures and techniques to address the challenge of generating high quality images across different domains and styles [32]. DL-based image generation has a wide range of applications, such as generating new photorealistic images/faces, labelling street scenes, changing grey to color images and predicting future frames. It aims to synthesize new images and transformation/translation of existing images.

Various network architectures have been proposed to generate realistic and diverse images, such as generative adversarial network (GAN) [52] and variational autoencoder (VAE) [73]. These architectures use different strategies to learn the underlying data distribution and generate images with distinct features and styles. The objective of GANs is to train the generator to produce synthetic data that is indistinguishable from real data, while the discriminator is simultaneously trained to accurately classify between real and synthetic data. GAN consists of a generator and a discriminator. The generator utilizes random noise as input and generates images, while the discriminator aims to differentiate between real and generated images. These two components are trained in an adversarial manner, where the generator learns to improve its output based on the results of the discriminator. As a result, GAN learns to generate images that are visually indistinguishable from real images. Many studies has also been proposed to improve the performance of GAN, *e.g.*, DCGAN [112] and WGAN [7]. VAE [73] is another popular architecture for new image generation, and its objective is to learn a latent space representation that captures the underlying structure of the input data. VAE enables to control the generation of images by modeling the latent space. It consists of an encoder and a decoder. The encoder maps the input images to a latent space, while the decoder reconstructs images from the latent space. The latent space is constrained to a specific distribution, which allows it

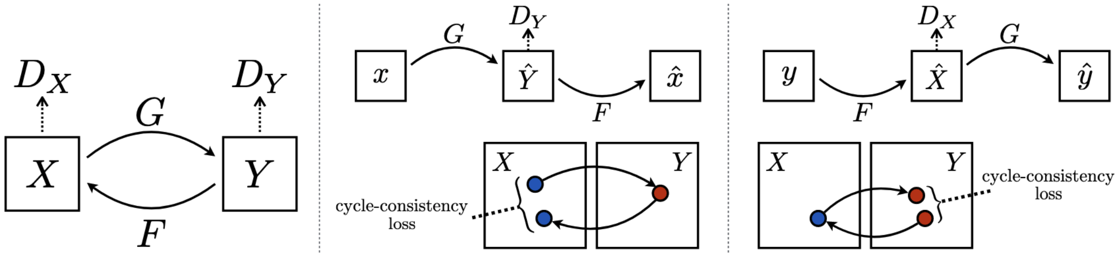


Figure 2.14: Framework of training CycleGAN [177].

to control image generation tasks, such as image inpainting and interpolation. However, some details may be lost when images are reconstructed from the latent space, leading to slightly blurred outputs compared to GANs.

Different from new image generation, image-to-image translation aims to generate images in the conditional setting. This type of tasks usually synthesize images conditioned on existing images, *e.g.*, generating photographs from sketches, changing gray to color images and transferring styles. Some of the most important designs are cGANs, such as Pix2Pix [67] and CycleGAN [177].

Pix2Pix [67] aims to learn a mapping between input images and corresponding output images based on a paired dataset. It consists of a generator and a discriminator. Unlike previous GANs, it utilizes a U-Net-like network as the generator and propose a PatchGAN classifier as the discriminator to focus more on the local details of style transfer. The generator takes an input image and generates the corresponding output image, while the patch discriminator differentiates the realism of the generated image (Figure 2.13). This discriminator extracts features at the patch level rather than evaluating the entire image as a whole. Pix2Pix assesses the realism of the generated image by classifying patches as real or fake, so it generates images are highly detailed. It demonstrates excellent capability to capture the global structure while preserving fine details. However, Pix2Pix requires a paired dataset to learn a desired mapping. Preparing such paired dataset can be time-consuming and labor-intensive, especially for certain specialized domains.

CycleGAN [177] is a cGAN model for unsupervised image-to-image translation tasks. It is able to translate images from one domain to another without the need for paired training data, which makes it suitable for many real-world applications where obtaining paired data is challenging. The key novelty of CycleGAN is the cycle-consistency loss. As shown in Figure 2.14, CycleGAN encourages that the output should be indistinguishable from the input ($x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$). This cycle-consistency loss ensures that if an image is translated from one domain to another and then back again, it should be reconstructed as close as possible to its original form. This constraint helps to reduce the space of possible mappings. The fundamental architecture of CycleGAN follows Pix2Pix, similar to the PatchGAN architecture. The discriminator provides feedback (real or fake) to the generator at the patch level, encouraging the capture of local details, styles, textures between image domains.

Since the cGAN-based models aim to learn the mapping between input and output images in different appearances or styles, it is utilized in Chapter 3 to transfer the chromosome straightening task to a mapping dependency learning task, thus enabling the input of a vertical chromosome backbone to generate the corresponding straightened chromosome. Compared to previous chromosome straightening studies based on geometric methods, this method generates continuous banding patterns and well-preserved details.

Chapter 3

A Novel Application of Image-to-Image Translation: Chromosome Straightening Framework by Learning from a Single Image

In our first work on chromosome straightening (medical image generation), we propose a framework based on the conditional generative adversarial network (cGAN), which can use a paired augmented dataset to train a converged model from only one curved chromosome image. We also propose a two-step strategy, including a novel chromosome backbone extraction approach and the following augmentation method, to prepare the augmented dataset. Compared to existing geometric approaches, our proposed method demonstrates superior straightening performance with uninterrupted banding patterns and well-preserved edge details.

3.1 Introduction and Problem Statement

In medical imaging, chromosome straightening plays a significant role in the pathological study of chromosomes and in the development of cytogenetic maps. Whereas different approaches exist for the straightening task, typically geometric algorithms are used whose outputs are characterized by jagged edges or fragments with discontinued banding patterns. To address the flaws in the geometric algorithms, we propose a novel framework based on image-to-image translation to learn a pertinent mapping dependence for synthesizing straightened chromosomes with uninterrupted banding patterns and preserved details. In addition, to avoid the pitfall of deficient input chromosomes, we construct an augmented dataset using only one single curved chromosome image for training models. Based on this framework, we apply two popular image-to-image translation architectures, U-shape networks and cGANs, to assess its efficacy. Experiments on a dataset comprised of 642 real-world chromosomes demonstrate the superiority of our framework, as compared to the geometric method in straightening performance, by rendering realistic and continued chromosome details. Furthermore, our straightened results improve the chromosome classification by 0.98%-1.39% mean accuracy.

There are 23 pairs of chromosomes in a normal human cell, comprised of 22 autosomes pairs (Type 1 to Type 22) and a pair of sex chromosomes (XX in females and XY in males). In the metaphase of cell division, the chromosomes become condensed and can be stained by the Giemsa banding technique [139] for observation under optical microscopes. The unique presence of light and dark regions (banding patterns) of different chromosome types are integrated into bars as cytogenetic maps. These banding patterns provide essential evidence for uncovering chromatin localization, genetic defects, and abnormal breakages [123]. For instance, human genetic diseases, such as cri-du-chat syndrome [60] and Pallister-Killian mosaic syndrome [74], can be diagnosed by identifying structural abnormalities in chromosomes.

With the advance in modern image acquisition techniques, digital images of chromosomes become fundamental to the construction of karyotypes (Fig. 3.1) and cytogenetic maps for studying structural features [10]. Because such tasks are labor-intensive and time-consuming, developing an automatic computer-assisted system has attracted significant research interest for the last 30 years. However, the condensed chromosomes are non-rigid with randomly varying degrees of curvatures along their lengths (Fig. 3.1). Such morphological features increase the difficulty of banding pattern analysis and abnormality identification.

An automatic karyotype construction system typically consists several steps, chromosome segmentation, straightening, classification and arrangement [135, 129, 102, 136, 169]. Straightened chromosomes have a higher accuracy of chromosome type classification [129]

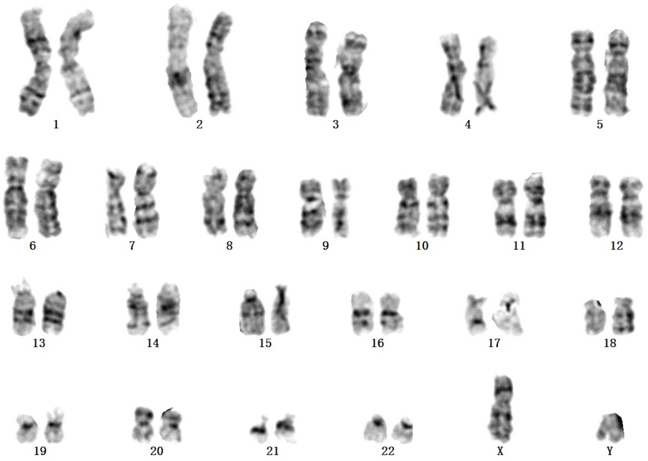


Figure 3.1: Karyotype of human chromosomes consisting of 22 autosomes pairs and a pair of sex chromosomes.

and they are pivotal in the development of cytogenetic maps [10]. The study of chromosome straightening first begins with cutting paper-based curved chromosome photo into pieces and arranging them into a straightened chromosome [142, 141]. To the best of our knowledge, based on digital images, current straightening approaches mainly utilize geometric algorithms which are broadly categorized by two approaches: (i) medial axis extraction and (ii) bending points localization. For the first approach, Barrett *et al.* [13] requires user interaction and manual labels. References [9, 69, 135] utilize thinning algorithms, such as morphological thinning [55] and Stentiford thinning [143]. However, such algorithms are not suitable for chromosomes with pronounced widths, resulting in many branches along their central axes when thinned [69, 135]. Additionally, when chromosome features are mapped or projected along straightened central axes, the jagged edges remain. The second approach involves analyzing bending points. For straightening, the chromosome is segmented by a single horizontal line from the potential bending point and its two arms are stitched in the vertical direction [120]. Sharma *et al.* [129] proposes an improved straightening method based on [120]. It fills the empty region between stitched arms by the mean pixel value at the same horizontal level as reconstructed banding patterns between stitched arms. However, this approach is also not suitable for the chromosomes whose arms are morphologically non-rigid, since the banding patterns of stitched arms are actually rotated rather than straightened along their central axes. Thus the reconstructed chromosomes contain distinct fragments with interrupted banding patterns, and the filled mean pixel value cannot restore realistic banding patterns. Moreover, it has poor performance with misidentifying bending points when there is more than one bending point in a chromosome.

To address the flaws in the geometric algorithms, we propose a novel framework based on image-to-image translation for synthesizing straightened chromosomes with preserved edges and unbroken banding patterns. Furthermore, we are the first to utilize deep learning and generative adversarial networks for straightening chromosomes.

Many studies have shown the success of image-to-image translation in diverse domains, examples including semantic segmentation [118], photo generation [164], and motion transfer [1, 89, 20]. U-Net [118] is one of the most popular and effective architectures. Its symmetrical contracting-expanding path structure and skip-connections are pivotal in the preservation of features. Its U-shape architecture has been modified for applications in many studies, such as a hybrid densely connected U-Net [86] and an architecture enhanced by multi-scale feature fusion [40]. Pix2pix is a milestone which boosts the performance of conditional generative adversarial networks based on image-to-image translation using a U-shape generator and a patch-wise discriminator [67].

Most applications of image-to-image translation require a large number of paired images. For example, a recent study [20] proposes an effective pipeline for translating human motions by synthesizing target bodies from pose extractions, and it is still trained using large-scale input frames with corresponding pose labels. Based on the mature field of pose detection, the pre-trained state-of-the-art pose detector is used to generate labels from a large number of frames of a given video. Chan *et al.* [20] subsequently trains deep learning models for mapping target body details from each body pose image.

In contrast, it is difficult to acquire sufficient training images and corresponding labels in the research of chromosome straightening. Due to random mutation, structural rearrangement, the non-rigid nature of chromosomes, and different laboratory conditions, it is almost impossible to find two visually identical chromosomes with the same curvature and dyeing condition under microscopes.

The challenge in this work is to straighten a curved chromosome using only a single chromosome image. Therefore, we propose a novel approach to first extract the internal backbone of the curved chromosome and subsequently increase the size of the chromosome dataset by random image augmentation. Instead of keypoint-based labels, we utilize stick figures as backbones which can retain more augmentation information. The other challenge of this research is to design a model that is able to render realistic and continued chromo-

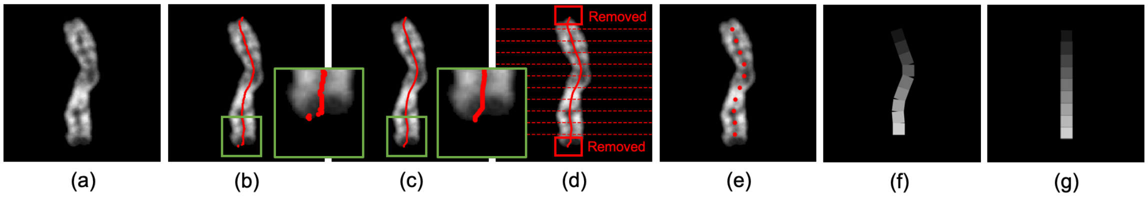


Figure 3.2: Seven types of images utilized in internal backbone extraction. (a) An example of original chromosomes; (b) an approximate central axis; (c) the smoothed central axis; (d) the smoothed central axis divided into 11 parts; (e) 10-point central axis; (f) the internal backbone; (g) the straightened internal backbone with the same length.

some details. At the same time, the straightening algorithm should not be affected by the non-rigid feature of chromosomes. Motivated by this, we innovatively apply image-to-image translation models to learn mapping dependencies from augmented internal backbones to corresponding chromosomes, resulting in high-quality outputs with preserved chromosome details. We also observe that the optimal generator of image-to-image translation models can complement banding patterns and edge details along with given internal backbones. Thus a straightened chromosome is synthesized when we feed a vertical backbone.

The key contributions of this research are three-fold. First, to address the deficiency of inputs, we propose a pertinent augmentation approach to increase the variability of curvatures from the given chromosome and corresponding label simultaneously. Second, using the augmented dataset, we apply two effective image-to-image translation architectures, U-shape networks and cGANs (pix2pix), which demonstrate the efficacy and robustness of our straightening framework. Third, in terms of the accuracy of chromosome type classification, we demonstrate that chromosomes straightened using our framework actually outperform the original curved chromosomes and the ones straightened using geometric algorithms.

The rest of this paper is organized as follows. In Section 3.2, the methodology is described in detail. In Section 3.3, we introduce the data preparation process and illustrate the comparison of straightening results. In Section 3.4, we discuss the limitations of the proposed approach and present some future research. Finally, we conclude our work in Section 3.5.

3.2 Methodology

In this section, we shall provide a detailed account of our framework. In Section 3.2.1, we propose an approach to generate augmented images and internal backbones from a single curved chromosome. In Section 3.2.2, we describe how the curved chromosome can be straightened by means of its backbone.

3.2.1 Data Augmentation Using a Single Image

For our framework, we propose a two-step strategy to construct an augmented dataset using only one curved chromosome image.

Algorithm 1 Chromosome internal backbone Extraction

Input: The digital image of a chromosome (C) whose width and height are W and H, respectively. The background of the image is black (0 pixel values).

Output: The internal backbone of the chromosome.

```
1: for each $h \in \{1, 2, ..., H\}$ do
2:     if the current row contains positive pixel values then
3:         find the first ($w_1$) and the last ($w_2$) positions whose pixel value is greater than 0;
4:         compute the central point $w_c^h = \frac{w_1^h + w_2^h}{2}$;
5:         record the $y$ axis values of the first and the last rows containing positive pixel values as $h_1$ and $h_2$, respectively.
6:     end if
7: end for
8: connect all $w_c^h$ to form an approximate central axis extending from $h_1$ to $h_2$;
9: smooth all $w_c^h$ by a moving average algorithm (11-pixel window length), to obtain $w_c^{\prime h}$;
10: divide the smoothed $w_c^{\prime h}$ equally into 11 parts (i.e. 12 points) by $y$ axis values in the range of $h_1$ to $h_2$;
11: remove the first and the last parts to obtain a 10-point central axis;
12: connect the adjacent splitting points by 33-pixel width sticks to obtain a 9-stick internal backbone;
13: generate a vertical 9-stick internal backbone with the same length between the the adjacent splitting points from Line 11.
```

Step 1. We construct the label of a curved chromosome (Fig. 3.2(a)) by extracting a pertinent internal backbone. The entire process is summarized in Algorithm 1. Considering the chromosome image to be comprised of rows of pixels, the centers of each row are connected to form an approximate central axis extending from top to bottom (Lines 1 to

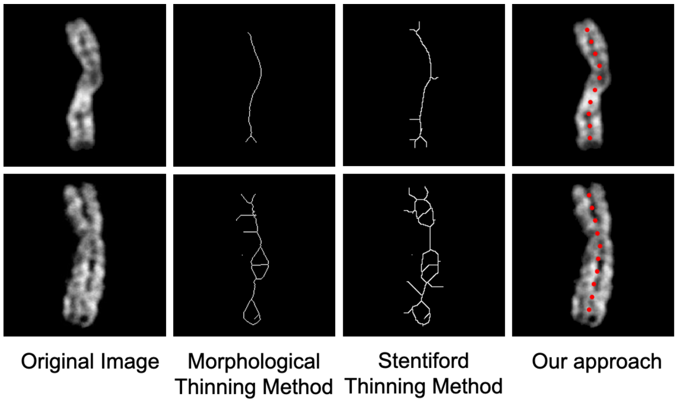


Figure 3.3: Examples of central axis extraction generated by thinning methods and our approach.

8 of Algorithm 1, Fig. 3.2(b)). To alleviate small-scale fluctuations generated in Line 8, this central axis is then smoothed by a moving average algorithm with an 11-pixel window length [152] (Line 9, Fig. 3.2(c)). We divide this smoothed central axis equally into 11 parts in the y axis. Since the first and the last parts may not be aligned in the same directions with both sides of the chromosome (red boxes), these two parts are subsequently removed (Lines 10 to 11, Fig. 3.2(d) to (e)). The remaining splitting points are connected by 33-pixel width sticks, and these 9 sticks are filled with pixel values in series of equal difference (23, 46, 69, 92, 115, 138, 161, 184, and 207) (Line 12, Fig. 3.2(f)). This stick figure contains the information of curvature, length, and orientation of the original chromosome. Finally, a vertical backbone is constructed with the same length of each stick (Line 13, Fig. 3.2(g)), and is fed into the fine-tuned image-to-image translation model for synthesizing the straightened chromosome.

Fig. 3.3 illustrates that the morphological and Stentiford thinning algorithms may cause branches and irregular rings when the chromosome features pronounced widths. Thus the previous work directed at chromosome straightening [9, 69, 135], composed of these thinning algorithms, cannot be utilized here. In contrast, our predicted 10-point central axis are approximately in accordance with the actual chromosome backbone.

Step 2. We improve the performance of deep learning models by generating more augmented chromosomes with different degrees of curvatures. We first apply random elastic deformation [150] and random rotation (from -45 to 45 degree) to the curved chromosome and its backbone simultaneously (Fig. 3.2(a) and (f)) until a sizeable number of aug-

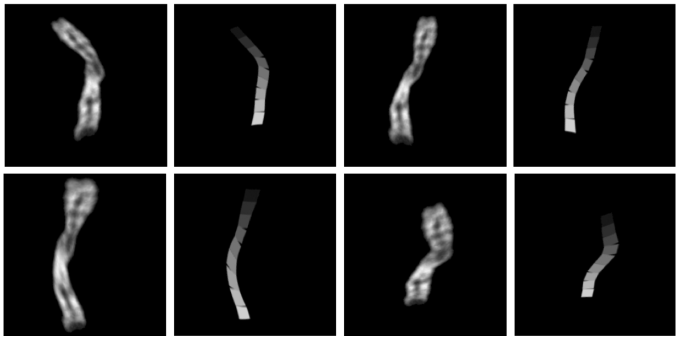
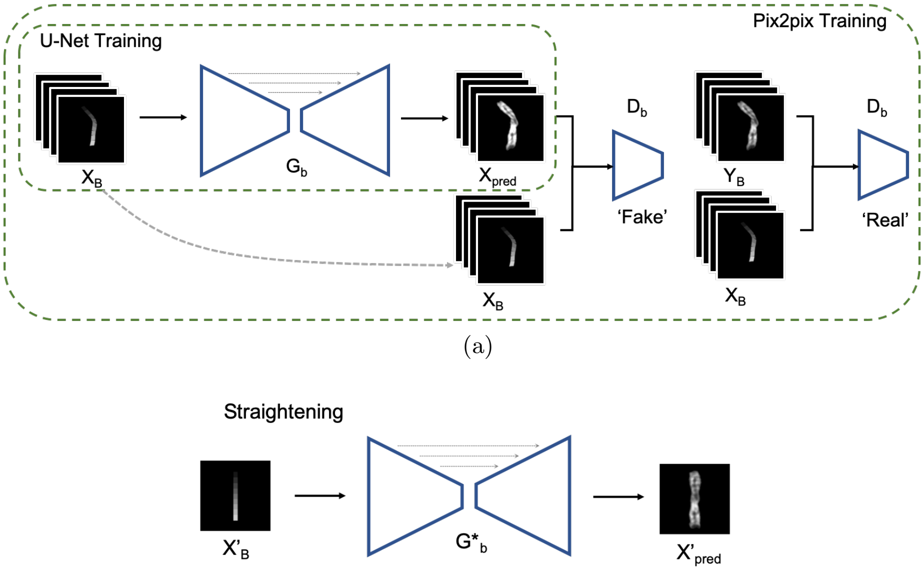


Figure 3.4: Examples of random data augmentation of a chromosome and corresponding internal backbone.



(a)

(b)

Figure 3.5: The overall process of the proposed framework for chromosome straightening. (a) The training processes of pix2pix or U-Net (the generator part of pix2pix), where X_B, Y_B are augmented backbones and chromosomes and $B \in \{1, ..., K\}$ where K is the number of augmented image pairs; X_{pred} is the predicted chromosome image through the generator, G_b. (b) The straightening process achieved by the optimal U-Net or generator G_b^*. $X_B^{'}$ and $X_{pred}^{'}$ are the vertical backbone and the straightened chromosome, respectively.

mented chromosomes and backbones (1000 pairs in this research) are obtained for training and validation (Fig. 3.4). Note that the setup of the elastic deformation algorithm [150] is $points = 3$ and $sigma = 18$ for 256×256 images, in order to generate plausible virtual curvatures. Since we utilize 33-pixel width sticks, rather than key points to label internal backbones, the detailed augmentation information, such as stretching, rotation and distortion, is retained and learned by the image-to-image translation models.

3.2.2 Image-to-Image Translation for Straightening

Since the objective of this study is to input a straightened backbone of a chromosome for synthesizing the corresponding chromosomes with preserved banding patterns, our novel image-to-image translation models are object specific. Therefore, it is essential to construct an augmented dataset for each image-to-image translation model. Utilizing the approach mentioned in Step 2, we generate 1000 augmented image pairs for each curved chromosome. The augmented dataset is then randomly split using a ratio of 9:1 for training and validation, respectively. Under our framework, we shall utilize two image-to-image translation models, U-Net and pix2pix (Fig. 3.5(a)). It should be noted that the U-Net utilized in this research is identical to the generator part of pix2pix. The training process of U-Net is a regular supervised learning method achieved by synthesized chromosomes and corresponding ground-truths. In pix2pix, a generator G_b synthesizes chromosomes from the augmented backbones to mislead D_b. Meanwhile, a discriminator D_b is trained for discerning "real" images from "fake" images yielded by the generator. The G_b and D_b is optimized with the objective function:

$$G_b^* = \arg\min_{G_b}\max_{D_b}\mathcal{L}_{cGAN}(G_b, D_b) + \lambda\mathcal{L}_{pix}(G_b) \tag{3.1}$$

where G_b^* represents the optimal generator; λ is a coefficient to balance two losses; $\mathcal{L}_{cGAN}(G_b, D_b)$ is the adversarial loss (Equation 2); and $\mathcal{L}_{pix}(G_b)$ is L1 distance to evaluate pixel-wise performance between generated images and ground-truths (Equation 3):

$$\mathcal{L}_{cGAN}(G_b, D_b) = \mathbb{E}_{x_B,z}[(D_b(x_B, G_b(x_B, z)) - 1)^2] + \mathbb{E}_{x_B,y_B}[(D_b(x_B, y_B))^2] \tag{3.2}$$

$$\mathcal{L}_{pix}(G_b) = \mathbb{E}_{x_B,y_B,z}[\|y_B - G(x_B, z)\|_1] \tag{3.3}$$

In the above: x_B and y_B represent augmented backbones and chromosomes, respectively; $B \in \{1, ..., K\}$ where K is the number of augmented pairs that we want; and z is the noise introduced in the generator.

To straighten the chromosome, we input its vertical backbone (Fig. 3.2(g)) into the optimal U-Net or optimal generator G_b^*, which will output the corresponding chromosome (Fig. 3.5(b)).

3.3 Experiments and Results

3.3.1 Chromosome Dataset

To test our framework on real-world images, we extract 642 low-resolution human chromosome images from karyotypes provided by a biomedical company. Images in this research have been cleaned so that connections between these images and their corresponding owners have been removed. Since the chromosomes with relatively long arms and noticeable curvatures require straightening (Figure 3.1), we collect Type 1 to 7 chromosomes in this research. We invert the color of these grey-scale images and center them in a 256 × 256 black background. As described in Section 3.2.1, 1000 augmented image pairs were obtained from each curved chromosome image before feeding into the U-Net and pix2pix models. It should be noted here that each augmented dataset is individually trained for straightening since our framework is object specific.

3.3.2 Evaluation Metrics

We apply two evaluation metrics to quantitatively measure the performance of these straightening methods. Due to the obvious morphological deformation between straightened results and original curved chromosomes, traditional similarity measurement metrics, such as Euclidean distance, structural similarity index (SSIM) [155] and peak-signal-to-noise ratio (PSNR) [61], designed for evaluating image quality degradation generated by image processing or compression, are not suitable for this task. Instead, Learned Perceptual Image Patch Similarity (LPIPS) [170] was used to evaluate straightening performance of different methods in this paper. The LPIPS is an emergent deep neural network-based method which is able to extract deep features of images for evaluating high-order structure similarity. Compared to the results of these traditional metrics, its results are more in accordance with human perceptual similarity judgment [170].

Apart from LPIPS, to ensure the details of straightened results are preserved in practice, we also assess the effectiveness of different straightening methods based on chromosome type classification. If the banding patterns and edge details of chromosomes are well preserved during straightening, the classification accuracy of straightened chromosomes should not decrease. In contrast, unpreserved details, such as broken bands, may not provide enough information for the classification model. The original images (642 curved chromosomes, Type 1 to 7) are randomly split using the ratio of 3:1 for 4-fold cross-validation. With a fixed random seed, this process is similarly carried out for the straightened chromosomes generated by different methods.

3.3.3 Implementation Details

Our experiments are implemented using PyTorch and run on two NVIDIA RTX 2080Ti GPUs. In each training process of chromosome straightening, the training and validation sets are split by a fixed random seed. The input image pairs are first normalized by default values (mean $\mu = 0.5$ and standard deviation $\sigma = 0.5$), and these results are fed into image-to-image translation models for learning the mapping dependence from backbones to chromosomes. Models are trained with an initial learning rate $lr = 0.00004$. The validation performance is checked three times per epoch, and the weights are saved when the best validation performance is updated. When the validation performance does not improve for 9 consecutive checks, the learning rate is reduced to 80% for fine-tuning. To avoid overfitting, the training process is terminated when there are 27 consecutive checks without updated validation performance. For each chromosome type classification model (Alexnet [76], ResNet50 [57] and DenseNet169 [64]), the training process is initialized with a learning rate of $lr = 0.00004$ and corresponding ImageNet pre-trained weights. We utilize 12 and 120 consecutive checks for fine-tuning and avoiding overfitting, respectively. Furthermore, we use identical random seeds, preprocessing and hyperparameter settings for 4-fold cross-validation of the chromosome type classification.

3.3.4 Results

Comparison of Straightening Performance

Although there are two categories of geometric methods (medial axis extraction [9, 69, 135] and bending points localization [120, 129]), we found that the morphological and Stentiford

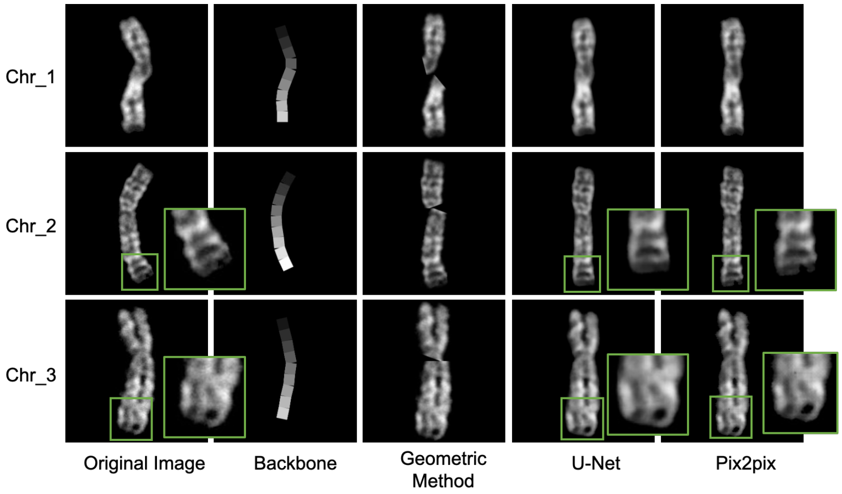


Figure 3.6: Three examples of straightening results. From left to right: original images, the geometric method [120, 129], our framework using U-Net and pix2pix. Enlarged regions demonstrate marginally improved details of pix2pix over U-Net.

thinning algorithms of medial axis extraction may cause many unexpected branches and irregular rings. Therefore, we investigated the performance of chromosome straightening using: (a) the geometric method (bending points localization) whose main component is used by [120, 129], and our image-to-image translation model based framework with (b) U-Net and (c) pix2pix models.

Fig. 3.6 gives three examples of the straightening results using the 642 curved chromosomes. The five columns correspond to: (i) the original unstraightened images, (ii) corresponding backbones extracted by our approach, (iii) outputs of the geometric method [120, 129], as well as the results from our framework with (iv) U-Net and (v) pix2pix, respectively. Although [129] additionally fills empty regions between stitched arms with the mean pixel values at the same horizontal level, the main problem of [120] whose results contain distinct segmented banding patterns between arms is still unresolved. In the third column of Fig. 3.6, we illustrate results of the straightening algorithm whose key part is used in [120, 129]. As examples in the third column of Chr_1 and Chr_2, the performance of the geometric method further deteriorates if there are curved arms and more than one bending point. Compared to these results, our framework demonstrates superiority both in translation consistency and in non-rigid straightening results (the fourth and fifth columns). The curvature of arms and the number of bending points do not decrease

Table 3.1: LPIPS results on different chromosome datasets (mean ± std.). For LPIPS lower is more similar.

	Original Images *vs.* Geometric Method	Original Images *vs.* U-Net	Original Images *vs.* Pix2pix	U-Net *vs.* Pix2pix
LPIPS	0.1621 ± 0.052	$\mathbf{0.1356 \pm 0.051}$	$\mathbf{0.1318 \pm 0.050}$	$\mathbf{0.0239 \pm 0.011}$

the performance of our framework because the image-to-image translation based framework relies on backbones rather than through morphological analysis. Since the provided chromosomes are low-resolution images, we notice that some straightened chromosomes (e.g. Chr_1) of U-Net and pix2pix have indistinguishable synthesized internal details and intensity. For many examples (enlarged area in Fig. 3.6), pix2pix marginally outperforms the U-Net model with more preserved edge details achieved by the patch-wise discriminator and adversarial training method. Since the chromosome images in this research are low-resolution (256 × 256), the ability to generate fine details using our framework with cGANs may become more obvious in high-resolution chromosome straightening and could be extended for use in the development of cytogenetic maps.

The average values and standard deviations (std.) of LPIPS are summarized in Table 3.1. Since LPIPS shows the perceptual distance between two images even there is obvious deformation, we quantify the similarity between curved chromosomes and straightened ones. We can observe that the straightening results of the pix2pix model under our framework achieves the best performance with a minimum LPIPS value (the third column of Table 3.1). The measurement of Original Images *vs.* U-Net and U-Net *vs.* Pix2pix indicates that the performance of U-Net is slightly worse than pix2pix due to the superior translation consistency of cGANs to U-shape neural networks. As a comparison, straightening results of the geometric method produced the highest LPIPS value, which may be caused by the broken banding patterns between stitched arms.

Comparison of Chromosome Type Classification Results on Different Straightened Datasets

We also performed experiments to determine if our proposed straightening framework enhanced the accuracy of the chromosome type classification. It is significant because the assessment of classification accuracy is an indispensable step in automatic karyotyping analysis [129, 171, 110]. Inaccurate straightened results may obscure the unique morphological features and banding patterns of different chromosome types.

Tables 3.2 and 3.3 give the comparisons between three standard state-of-the-art classi-

Table 3.2: Comparison of averaged classification accuracy (4-fold cross-validation)

Accuracy (%)	Alexnet	ResNet50	DenseNet169
Original Images (Baselines)	90.47	85.31	86.09
Geometric Method [120, 129]	78.44	70.16	73.59
U-Net	**91.51**	**85.65**	**87.65**
Pix2pix	**91.67**	**86.57**	**87.81**

Table 3.3: Comparison of averaged AUC of chromosome type classification (4-fold cross-validation)

AUC	Alexnet	ResNet50	DenseNet169
Original Images (Baselines)	0.9423	0.9163	0.9271
Geometric Method [120, 129]	0.8513	0.8317	0.8513
U-Net	**0.9487**	**0.9204**	**0.9301**
Pix2pix	**0.9510**	**0.9293**	**0.9311**

fication networks, AlexNet [76], ResNet50 [57] and DenseNet169 [64]. The accuracy scores and their Area Under Curve (AUC) are the mean value of 4-fold cross-validation results. We consider the scores trained by original curved chromosomes as baselines. We can see that wrongly identified bending points and stitched chromosome arms with discontinued banding patterns from the geometric method, reduce the classification results by a significant margin (-13.23% accuracy, -0.084 AUC on average). In contrast, our framework achieves top scores and marginally outperforms the baselines by 0.98% accuracy, 0.0045 AUC (U-Net) and 1.39% accuracy, 0.0085 mean AUC (pix2pix) on average. One possible reason is that the straightened and uninterrupted banding patterns help neural networks to learn uncurved and unrotated unique features of chromosomes. The superiority of our proposed framework suggests that it may benefit banding pattern identification and abnormality detection in the automatic pathological diagnosis of karyotypes. Fig. 3.7 depicts the mean accuracy curves of different training/validation sets of these three models. It illustrates that the chromosome type classification performance of datasets between original images, chromosomes generated by U-Net and pix2pix display similar trends, which is in accordance with the results of Table 3.2 and Table 3.3. This indicates the details of chromosomes are well preserved after straightening. In contrast, the chromosome type classification accuracy is severely affected by the discontinued banding patterns and unstraightened arms generated by the geometric method.

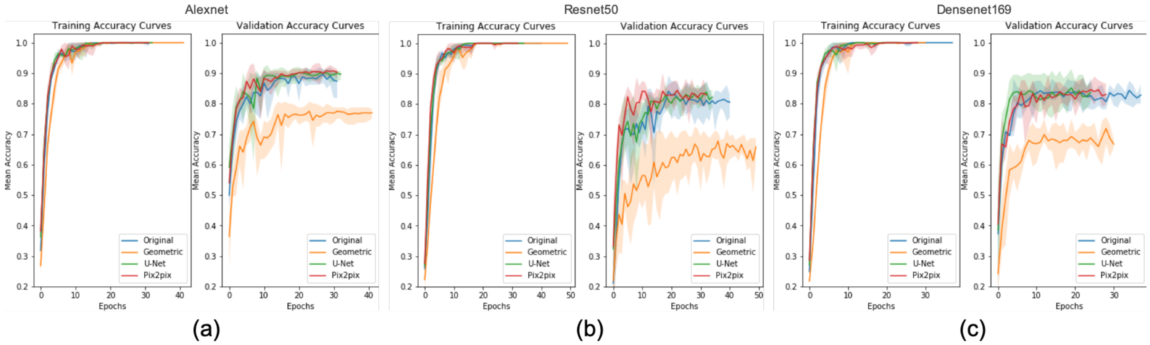


Figure 3.7: Training and validation accuracy curves of three CNN models for chromosome type classification (4-fold cross-validation). Shadow regions represent the range over four folds and solid lines represent mean accuracy.

3.4 Limitation and Discussion

3.4.1 Computation Time

To address the flaws, such as the broken banding patterns in geometric methods and random stretching in elastic deformation algorithms, we propose a chromosome straightening framework which is object specific. Therefore, it is time-consuming to train a separate straightening model for every curved chromosome. In future research, a generalized chromosome straightening model shall be designed. We would design an improved model for disentangling the information of internal backbones and banding patterns.

3.4.2 Failure Cases

Under our framework, we notice two types of failure cases. First, the straightening performance hinges on the accuracy of the central axes identified. When the curvature of a chromosome is too large, the extracted internal backbone may not be aligned in a similar direction with the original image (red arrows of Chr_4 in Fig. 3.8). In this case, the relation between the backbone and corresponding banding patterns are still preserved. As a result, that part may not be well straightened. Second, some irregular chromosomes may still cause small-scale fluctuations of backbones even after the moving average algorithm, resulting in blurred synthesized banding patterns and edge details (Chr_5 in Fig. 3.8). Because of this, high-quality labels of chromosomes are still deficient in the augmented dataset. A plausible direction would be an improvement of the backbone extraction method. A crowdsourcing

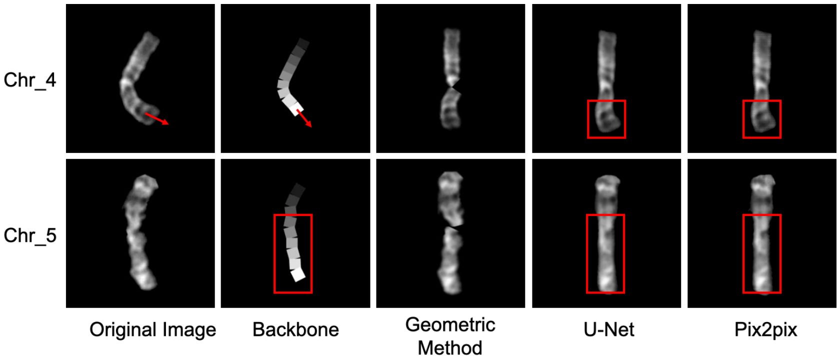


Figure 3.8: Two examples of failure cases. From left to right: original images, the geometric method [120, 129], our framework using U-Net and pix2pix.

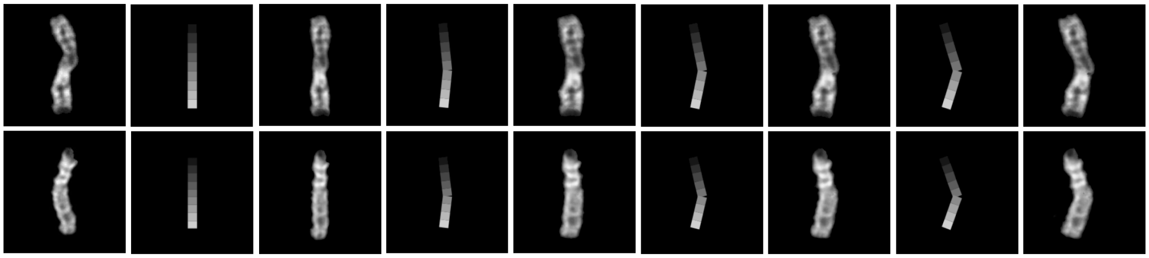


Figure 3.9: Examples of synthesized results with a series of curved internal backbones (generated by our framework with the pix2pix model).

database of labeled backbones could be established for developing a powerful deep learning based backbone detector of chromosomes.

3.4.3 Potential Applications

Since the results of our straightening framework demonstrate a higher classification accuracy, it is worthwhile to incorporate the framework into automatic karyotyping analysis and cytogenetic map construction. With the development of image-to-image translation research, many advanced modules and architectures, for example, attention-based GANs [167], may be integrated into our framework to further improve its efficacy and robustness.

Since our augmented datasets contain information concerning random deformation and rotation, we observe that fine-tuned generators not only have an ability to straighten chromosomes, but also can synthesize more chromosomes by inputting internal backbones with different curvatures (Fig. 3.9). Therefore, our framework demonstrates the potentiality for generating augmented chromosomes with highly preserved detail along with customized backbone images.

Compared to regular U-shape networks, cGANs have more potential in the application of high-resolution chromosome straightening with higher translation consistency. In the latest study, Artemov *et al.* [10] employs PhotoShop for straightening high-resolution chromosomes when developing cytogenetic maps, so an automatic high-resolution chromosome straightening framework is still in demand. Similar to the evolution from pix2pix to pix2pixHD [153], our straightening framework may also be further modified for high-resolution chromosome images.

3.5 Conclusions and Future Work

In this study, we propose a novel image-to-image translation based chromosome straightening framework which sets a new direction for object straightening. The framework transforms the task of straightening into the learning of mapping dependency from randomly augmented backbones to corresponding chromosomes. It allows straightened chromosomes to be generated from vertical backbones. The straightening performance of our framework is significantly better than the geometric approach with more realistic images of uninterrupted banding patterns. Under our framework, the average classification accuracy of U-Net and pix2pix evaluated by state-of-the-art classification models is higher than the baselines by 0.98% and 1.39%, respectively.

However, using this straightening framework it is still computationally expensive to train separate models for different curved chromosomes, the framework also may generate blurred results due to inaccurately identified internal backbones. Since the study of deep learning based chromosome straightening is at its infancy, many improvements can be made to our framework, such as a more accurate internal backbone extraction method, and a generalized architecture which is not object specific.

Chapter 4

A Robust Framework of Chromosome Straightening with ViT-Patch GAN

In the previous chapter, we propose a standard image-to-image translation framework for chromosome straightening. Compared to previous geometric chromosome straightening methods, the proposed framework and backbone extraction method achieve significantly better performance due to more realistic straightening results with continuous banding patterns and edge details. However, this work has two major limitations: (1) Since the model is trained by a dataset augmented from a single chromosome image, a separate model requires to be trained for each chromosome. The learned representations are only mapping dependencies from backbone images to their corresponding chromosome images. Therefore, this approach is time-consuming and has poor generalization for further straightening applied to other large datasets. (2) The internal backbone of chromosome requires to be extracted to construct an augmented dataset, so inaccurate extraction may lead to non-fully straightened or blurred results.

In this chapter, we propose an advanced chromosome straightening framework for more efficient representation learning in the direction of medical image generation. To address the above two limitations in our previous work, we design a novel and generalized chromosome straightening framework, named ViT-Patch GAN, which contains a generator and a discriminator. The generator can self-learn the motion representation of chromosomes with different degrees of curvature. With the help of the designed ViT-Patch discriminator, the straightening performance of the synthesized results of the generator is improved. We successfully convert this straightening task into a motion representation learning task. This confers two advantages. First, learning motion transformation rather than mapping dependencies from backbone to chromosome improves the generalization capability of model. A well-converged model has the potential to straighten all other chromosomes even in cross-dataset experiments, whereas in our previous work, we need to train a separate model for each chromosome images. Second, learning the motion representation avoids the step of extracting chromosome backbones, so the accuracy of the backbone extraction algorithm does not affect straightening results of this novel framework. Furthermore, this novel framework addresses other challenges of chromosome straightening task. It allows training on a small dataset and improves the preservation of straightening details.

4.1 Introduction and Problem Statement

Chromosomes carry the genetic information of humans. They exhibit non-rigid and non-articulated nature with varying degrees of curvature. Chromosome straightening is an important step for subsequent karyotype construction, pathological diagnosis and cytogenetic map development. However, robust chromosome straightening remains challenging, due to the unavailability of training images, distorted chromosome details and shapes after straightening, as well as poor generalization capability. In this paper, we propose a novel architecture, ViT-Patch GAN, consisting of a self-learned motion transformation generator and a Vision Transformer-based patch (ViT-Patch) discriminator. The generator learns the motion representation of chromosomes for straightening. With the help of the ViT-Patch discriminator, the straightened chromosomes retain more shape and banding pattern details. The experimental results show that the proposed method achieves better performance on Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS) and downstream chromosome classification accuracy, and demonstrates excellent generalization capability on a large dataset.

In a normal human cell, there are 22 pairs of autosomes (Type 1-22) and one pair of heterosomes (Type X & Y in male and two copies of Type X in female). By karyotype analysis, chromosome aberrations can be detected in the diagnosis of many genetic diseases, such as the Klinefelter syndrome [16] and specific cancers [4]. The banding patterns of chromosomes (unique light and dark stained bands) provide important evidence for the development of cytogenetic maps. Due to their non-rigid nature, condensed chromosomes exhibit varying degrees of curvature under the microscope. Therefore, chromosome straightening is an important upstream task for chromosome classification [171] and the subsequent karyotype construction and cytogenetic map development [10].

Chromosome straightening has been studied for a long time and its development may be described in three stages. (i) Printed images of bent chromosome are physically cut and rearranged for straightening [142, 141]. (ii) Geometric algorithms are extensively designed based on chromosome micrographs for automatic straightening, which consist of two main categories: extraction of the medial axis [9, 69, 135, 158] and finding bending points [120, 129, 169]. However, these geometric methods may extract inaccurate medial axes when chromosomes have large widths. The methods of bending point localization often use a cut line to separate chromosome arms, leading to discontinuous banding patterns. (iii) A Single Chromosome Straightening Framework (SCSF) [138] using conditional generative adversarial networks (cGAN) is proposed to straighten chromosomes with uninterrupted chromosome banding patterns. However, it requires a large number of input image pairs for training a model for each chromosome. ChrSNet [175] is proposed as a two-module framework with self-attention mechanism, but some edge details are not well-preserved. In addition, there are many studies on image deformation, such as medical image registration [3, 8]. The image registration technique requires two real-world chromosomes with the identical shape details but different curvatures, which is almost impossible to obtain. Thus it also fails in this chromosome straightening task. Recently, First Order Motion Model (FOMM) [131] and PCA-based Motion Estimation Model (PMEM) [132] are proposed by learning key point-based and region-based representations for motion transfer, respectively. However, they still require to train on sufficient image pairs for robust performance.

Challenges remain in the three stages of chromosome straightening. (1) *Lack of of training images.* It is almost impossible to take micrographs of two chromosomes with identical stained banding pattern but different curvatures due to the diversity of random mutations, chromosome condensation and laboratory conditions. Thus, it is challenging to train a robust deep learning-based model for straightening. (2) *Distorted chromosome details and shapes after straightening.* The condensed chromosomes are non-rigid with varying degrees of curvature. Straightened chromosome results require a high degree of preservation of consistent shape and details in the source image. Image registration and motion transfer methods tend to generate distorted results with driving image shapes. (3)

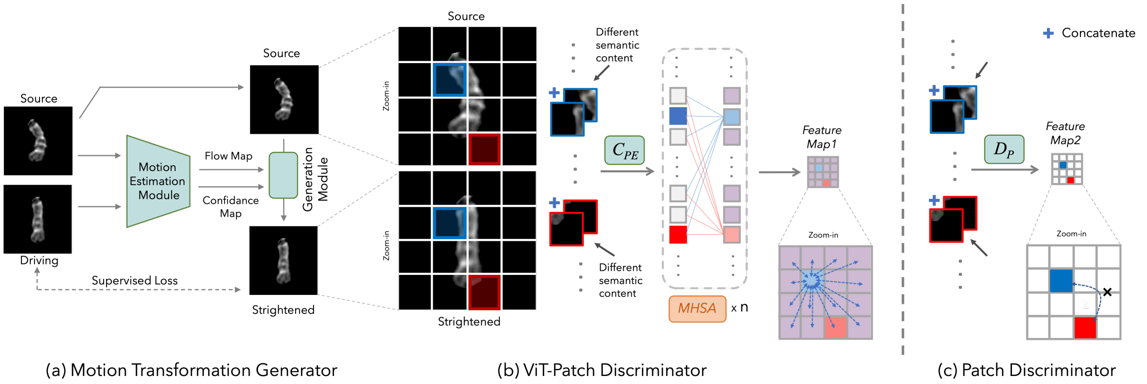


Figure 4.1: Overview of the architectures of the proposed framework, including (a) motion transformation generator and (b) our proposed ViT-Patch discriminator; (c) gives the structure comparison of a basic patch discriminator.

Poor generalization capability. The recent cGAN-based chromosome straightening framework [138] only learns mapping dependencies for each specific banding patterns, making it very time-consuming for large-scale applications.

The main contributions of this research address the above problems and are as follows:

- We propose a robust cGAN-based framework for chromosome straightening on a small dataset, by transfering the chromosome straightening task to the motion transformation task of non-rigid objects.

- We propose a novel architecture, ViT-Patch discriminator, to improve the detail preservation ability of our framework. Compared with existing methods, straightened chromosomes retain more shape and banding pattern details of the corresponding source images.

- Different from SCSF [138], our trained model demonstrates excellent generalization capability and is able to be applied to a large chromosome dataset for straightening.

4.2 Methodology

4.2.1 Network Architecture

Fig. 4.1 presents an overview of the proposed ViT-Patch GAN architecture. It is comprised of two parts, a motion transformation generator (Fig. 4.1-a) and a ViT-Patch discriminator

(Fig. 4.1-b). Compared to SCSF [138], ViT-Patch GAN requires a small dataset size for training (642 images). It straightens chromosomes with highly preserved details, and has high generalization capacity.

Generalized Framework of Chromosome Straightening. We apply PMEM [132] as the generator part since it learns motion representations through self-learned region estimation. We can consider the chromosome straightening task as a motion transformation task for non-rigid objects. In the training stage, the generator requires a *training source* and *training driving* images with the same chromosome but different curvatures. As shown in Fig. 4.1-a, using the Motion Estimation Module, the flow and confidence maps containing a combination of region and background transformations are then fed into the Generation Module to synthesize the straightened chromosome. Subsequently, the straightened chromosome is supervised with the *training driving* image by a supervised reconstruction loss ($\mathcal{L}_1$), consistent with PMEM [132]. One of the challenges of this task is the lack of *training driving* data. However, PMEM was trained on sufficient image pairs (video clips containing only the same object) [132]. On a small dataset, PMEM may inadequately transfer the shape of the driving image to the source image, leading to inaccurate straightening results.

To alleviate this drawback, we propose a ViT-Patch discriminator (Fig. 4.1-b) that encodes not only local details but also global feature connection. The final loss is the standard adversarial loss of the generator and discriminator in the cGAN-based framework [67]. During testing, the converged model straighten chromosomes by feeding *test source* and *test driving* images. It is worth noting that the *test driving* can be a different straight chromosome to the *test source* since the motion representation has been learned.

ViT-Patch Discriminator. Fig. 4.1-b and Fig. 4.1-c give the comparison between our proposed ViT-Patch discriminator and a basic patch discriminator. In a basic patch GAN [67], the semantic content of the corresponding patches at the same position between the source and generated images is generally the same. However, in this task, curvatures of a chromosome change after straightening. As a result, patches at the same position before and after straightening may contain significantly different chromosome patterns (*e.g.* the concatenated patches with black arrows in Fig. 4.1-b and Fig. 4.1-c).

In training a basic patch GAN, two patches at the same position are concatenated and fed into the discriminator network (D_P) that contains several consecutive convolutional blocks to output the *Feature Map2* for adversarial training (Fig. 4.1-c). Although adjacent patches overlap according to the receptive field, long-distance patches are independent and not informatively linked. This deficiency may lead to inaccuracy and limited performance in the chromosome straightening task. To address this, for the proposed ViT-Patch discriminator, patches at the same position are fed into a convolutional patch embedding layer (C_{PE}), and then processed by N Multi-Head Self-Attention (MHSA) Blocks as ViT [41] (4-16 used for ablation experiments). Afterwards, the encoded features contain long-range dependencies across the image to compensate for information loss. Thus the *Feature Map1* contains not only the local semantic content but also the information connectivity of the entire chromosome.

4.2.2 SL-matching Scheme

In addition to proposing a ViT-Patch discriminator to improve generalization, empirically, we found a large shape difference between the *test source* and *test driving* chromosome, leading to significant distortion. To alleviate this problem, we propose a Size-LPIPS matching scheme (SL-matching) to select a *test driving* with a similar size and shape for each *test source* image. The SL-matching scheme is comprised of two phases. In Phase 1 we perform a line-by-line scan of the chromosome image and record the midpoint of each line containing non-zero pixels. The resulting set of midpoints is first smoothed using a moving average algorithm, and its length calculated. Next, we perform a column-by-column scan, where we take the x-axis coordinates of all non-zero pixel columns as the width. The top three candidates whose lengths and widths are most similar to the corresponding *test source* are selected. In Phase 2 we calculate the perceptual score between a *test source* and each candidate chromosome. This score is the average result generated by LPIPS with AlexNet and VGG backbones [170]. Finally, the selected image has the largest size and perceptual similarity.

4.3 Experiments

4.3.1 Experiment Setup

Datasets and Implementation Details. A total of 16696 chromosome micrographs were used for the evaluation as provided by [138]. All data had been desensitized, and patients' personal information has been removed. Since 642 out of the 16696 chromosomes were straightened by SCSF [138], these 642 chromosomes were used as the dataset ($\mathtt{D}_{train}$) for training and testing the ViT-Patch GAN in a 4:1 ratio. In the training stage, the original chromosomes and corresponding straightened results in $\mathtt{D}_{train}$ were utilized as

Table 4.1: Comparison with existing studies, the best and second best results highlighted in bold and underlines. The DCA result is the average score of each experiment with cross-validation.

DCA (%)	FID↓	LPIPS$_A$↓	LPIPS$_V$↓	DCA$_{R34}$	DCA$_{R50}$	DCA$_{D169}$
Original	-	-	-	91.72	85.63	83.59
SCSF [138]	54.90	0.1272	0.0974	90.63	$\underline{86.87}$	85.31
FOMM [131]	53.71	0.1310	0.0982	$\underline{92.19}$	85.16	$\underline{88.28}$
PMEM [132]	$\underline{43.80}$	$\underline{0.1196}$	$\underline{0.0897}$	92.03	86.09	84.69
ViT-Patch	**42.21**	**0.1160**	**0.0874**	**93.91**	**90.16**	**89.06**

↓ represents that lower results are better.

training source and *training driving* images, respectively. Our clinical experts select 1200 real-world straight chromosomes of different lengths from all chromosome types as the dataset, $\mathtt{D}_{driving}$. In the testing stage, $\mathtt{D}_{driving}$ was used to select *test driving* for each *test source* utilizing the SL-matching scheme. A large dataset was employed in this study to assess generalization capacity: the remaining 14854 chromosome images provided by [138] ($\mathtt{D}_{large}$).

All experiments were implemented using one NVIDIA GeForce RTX 2080Ti GPU and coded using PyTorch. The Adam optimizer and a batch size of 1 were used. The total training epoch was 50 and the ratio of the training and test sets was 4:1 (5-fold cross-validation). The initial learning rates of the generator and discriminator were $5e^{-5}$ and $1e^{-5}$, respectively. We used MultiStepLR for the generator and discriminator with milestones 30 and 45. After completing cross validation, all 642 chromosomes of $\mathtt{D}_{train}$ were straightened. Subsequently, we performed downstream classification experiments (7 chromosome types) using initial learning rate of $4e^{-5}$ and ReduceLRonPlateau with patience 5 and early stopping protocol with patience 20. All input images were preprocessed to png format (256×256).

Evaluation Metrics. To quantitatively assess the quality of straightened chromosomes, we used the following three evaluation metrics: (i) Fréchet Inception Distance (FID) (ii) Learned Perceptual Image Patch Similarity (LPIPS) and (iii) Downstream Classification Accuracy (DCA). FID [59] assesses the quality of generated images by comparing the distribution between the real and generated images. LPIPS [170] estimates the perceptual similarity. Its results are closer to human judgment than many traditional metrics, such as $\mathcal{L}_2$, PSNR and SSIM, especially for deformed objects (between original and straightened chromosomes) [170]. LPIPS$_A$ and LPIPS$_V$ represent the scores generated by AlexNet and

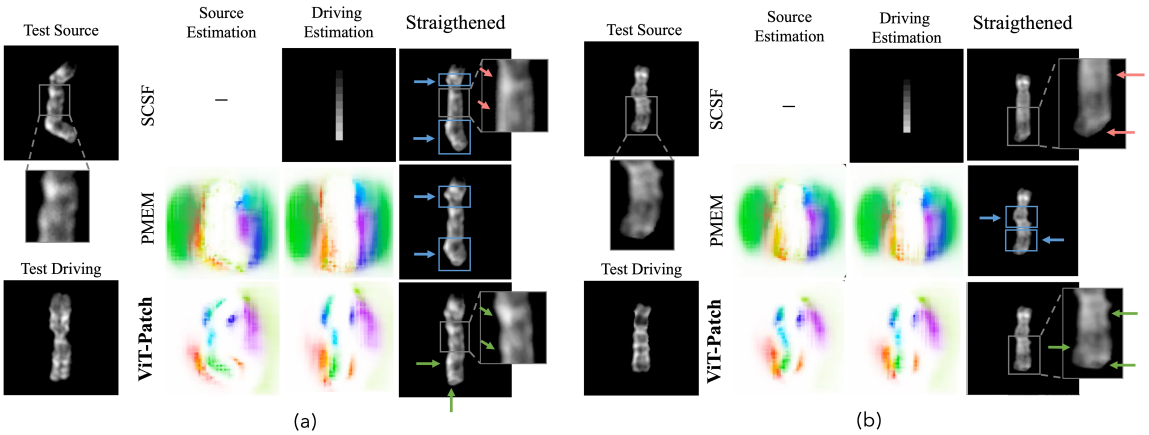


Figure 4.2: Visual results of chromosome straightening generated by different methods.

VGG backbones. Since chromosome classification is an important downstream task of chromosome straightening for subsequent karyotype construction and pathological diagnosis, the DCA can be used to assess the performance.

4.3.2 Comparison with State-of-the-Art Methods

Table 4.1 gives the quantitative assessment of chromosome straightening on $\mathtt{D}_{train}$. We compare the ViT-Patch GAN with state-of-the-art studies, SCSF [138], FOMM [131] and PMEM [132]. These results are the average scores using 5-fold cross-validation on $\mathtt{D}_{train}$. The FID value of ViT-Patch GAN decreases from 54.90 to 42.21, implying that the straightening quality of our method is closest to the real-world chromosome images. Moreover, ViT-Patch GAN achieves the best LPIPS scores on both AlexNet and VGG backbones, demonstrating the clear advantages of our proposed method in this task.

We perform downstream classification experiments using 642 chromosomes straightened by these state-of-the-art methods in exactly the same configuration. Each DCA is the average best result generated by 4-fold cross-validation classification experiments with commonly used models, ResNet34 (R34), ResNet50 (R50) and DenseNet169 (D169) [57, 64]. Although there is overfitting with the increasing depth of classification networks, we observe that ViT-Patch GAN significantly outperforms other methods by a large margin on DCA_{R34}, DCA_{R50} and DCA_{D169}. The proposed method achieves gains of 2.19%, 4.53% and 5.47% on DCA compared to the original bent chromosomes (baselines). Such results

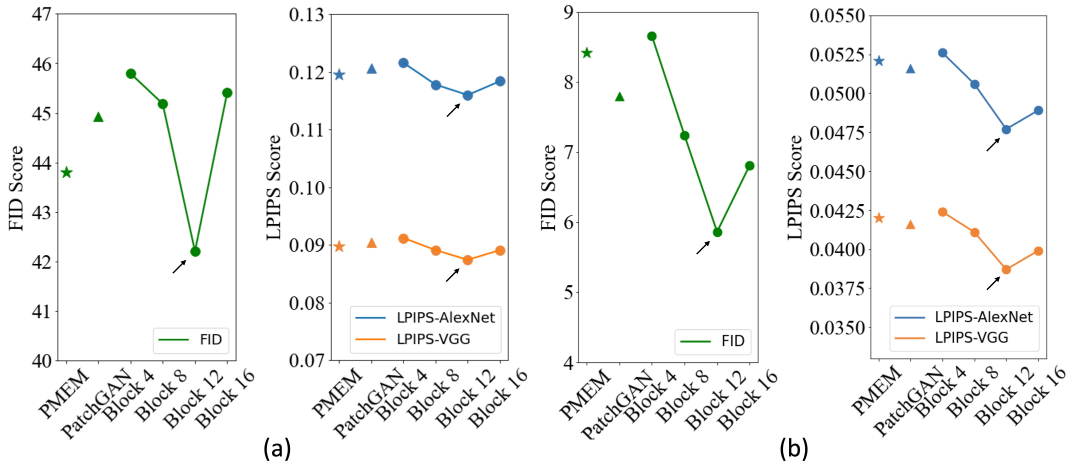


Figure 4.3: Comparison of straightening performance (a) of ablation study and (b) on a large-scale dataset ($\mathtt{D}_{large}$).

highlight that our proposed ViT-Patch GAN has an excellent reconstruction quality of straightened chromosomes.

Fig. 4.2 gives the comparison between two straightened chromosomes. Compared to SCSF and PMEM, ViT-Patch GAN achieves the most accurate straightening details (green arrows) with fewer reconstruction errors. These errors are generally of two types: (i) bending is not well recovered (blue arrows with boxes), and (ii) the chromosome shape and details are not well preserved (red arrows). This inadequate straightening is mainly caused by inaccurate motion representation. For SCSF, only a backbone responsible for straightening results in a lack of features. The estimated regions of PMEM are wide and mostly located on the background, resulting in the curvature being almost not straightened (blue arrows). In contrast, ViT-Patch GAN estimates more meaningful and accurate motion representation for straightening, resulting in highly reliable straightening results.

4.3.3 Ablation Study and Cross-Dataset Experiments

We conduct ablation experiments to compare the performance of different architectures. We also implement PatchGAN, which uses a basic patch discriminator with PMEM as the generator. Blocks 4-16 are experiments for our proposed ViT-Patch discriminator with a series number of MHSA blocks. Fig. 4.3-a demonstrates that the performance on all metrics increases from Blocks 4 to 12, and Blocks 12 of our method achieves the best

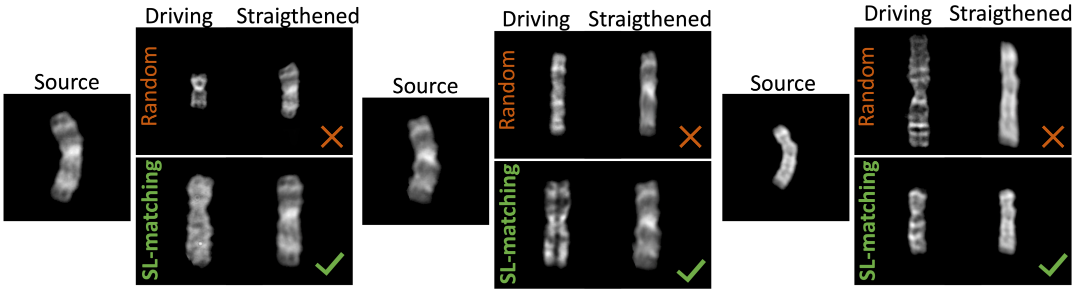


Figure 4.4: Comparison of straightening performance based on different *test driving* images selected by random and our proposed SL-matching schemes.

results (black arrows). In contrast, Blocks 16 may lead to overfitting, resulting in decreased performance.

We further conduct experiments with the exactly same ViT-Patch GAN model (Blocks 12), but only with different *test driving* images picked by random selection and the proposed SL-matching scheme. Fig. 4.4 demonstrates that the SL-matching scheme is important to drive source images for generating accurate straightening results.

We perform cross-dataset experiments to evaluate the generalization capability of ViT-Patch GAN (Fig. 4.3-b). The models (those used in Section 4.3.3) are trained on $\mathtt{D}_{train}$. The performance relationship between Fig. 4.3-b and Fig. 4.3-a is consistent. Our ViT-Patch GAN with Blocks 12 outperforms the others on all metrics, suggesting its robustness in the cross-dataset setting. Although our proposed chromosome straightening framework is trained on only 642 chromosomes ($\mathtt{D}_{train}$), the ViT-Patch GAN learns the global motion representation rather than only mapping dependencies of specific patterns (*i.e.*, SCSF). The converged model can further straighten chromosomes of all types on the large dataset, $\mathtt{D}_{large}$. Thus, it may be further utilized for large-scale applications of chromosome straightening.

4.4 Conclusion and Future Work

In this paper we propose a novel robust chromosome straightening framework, which includes a generator for learning chromosome motion representation and a ViT-Patch discriminator for generating more realistic straightened results. The ViT-Patch discriminator encodes both the local detail and long-range dependency. Qualitative and quantitative results demonstrate that the efficacy of our proposed method in retaining more chromosome shape and banding pattern details. Our framework also has excellent generalization capability in chromosome straightening for a large size dataset.

In our experiments, we note that this method achieves excellent performance on chromosomes with relatively long arms. Thus $\mathtt{D}_{driving}$ may be continuously updated based on more training and driving images to obtain better performance on short-armed chromosomes, thus training a more robust and generalized model. In this case, the potential of applying ViT-Patch GAN for large size datasets can be further improved.

Chapter 5

Bilateral-ViT for robust fovea localization

In our first work on fovea localization (medical image segmentation), we propose a novel multi-cue fusion architecture, named Bilateral-Vision-Transformer (Bilateral-ViT), which consists of two branches, a main branch that exploits long-range context and a vessel branches that encodes structure information from the blood vessel segmentation map. The encoded global context across the entire fundus image and the segmentation map are subsequently decoded by a customized Multi-scale Feature Fusion (MFF) module. Compared to previous studies that integrate features from the fundus image only, our proposed architecture focuses more on features distributed along anatomical structures (blood vessels) associated with fovea locations, achieving new state-of-the-art results on public datasets and significantly improving generalization in cross-dataset experiments.

5.1 Introduction and Problem Statement

Fovea is an important anatomical landmark of the retina. Detecting the location of the fovea is essential for the analysis of many retinal diseases. However, robust fovea localization remains a challenging problem, as the fovea region often appears fuzzy, and retina diseases may further obscure its appearance. This paper proposes a novel Vision Transformer (ViT) approach that integrates information both inside and outside the fovea region to achieve robust fovea localization. Our proposed network, named Bilateral-Vision-Transformer (Bilateral-ViT), consists of two network branches: a transformer-based main network branch for integrating global context across the entire fundus image and a vessel branch for explicitly incorporating the structure of blood vessels. The encoded features from both network branches are subsequently merged with a customized Multi-scale Feature Fusion module. Our comprehensive experiments demonstrate that the proposed approach is significantly more robust for diseased images and establishes the new state of the arts using the `Messidor` and `PALM` datasets.

The macula is the central region of the retina. The fovea is an important anatomical landmark located in the center of the macula, responsible for the most crucial part of a person's vision [156]. The severity of vision loss due to retinal diseases is usually related to the distance between the associated lesions and the fovea. Therefore, detecting the location of the fovea is essential for the analysis of many retinal diseases.

Despite its importance, robust fovea localization remains a challenging problem. The color contrast between the fovea region and its surrounding tissue is poor, leading to a fuzzy appearance. Furthermore, the fovea appearance may be obscured by lesions in the diseased retina; for example, geographic atrophy and hemorrhages significantly alter the fovea appearance. Such issues make it more difficult to perform localization based on the fovea appearance alone. Fortunately, anatomical structures outside the fovea region, such as blood vessels, are also helpful for localization [85, 6]. For this reason, we propose a novel Vision Transformer approach that integrates information both inside and outside the fovea region to achieve robust fovea localization.

Our proposed network, named Bilateral-ViT, consists of two network branches. We adopt a transformer-based U-net architecture [21] as the **main branch** for effectively integrating global context across the entire fundus image. In addition, we design a **vessel branch** that takes in a blood vessel segmentation map for explicitly incorporating the structure of blood vessels. Finally, the encoded features from both network branches are merged with a customized Multi-scale Feature Fusion module, leading to significantly improved performance. Thus, our key contributions are as follows:

- We propose a novel vision-transformer-based network architecture, that explicitly incorporates global image context and structure of blood vessels, for robust foveal localization.

- We demonstrate that the proposed approach is significantly more robust for challenging settings such as fovea localization in diseased retinas (over 9% improvements for specific evaluations). It also has a better generalization capability compared to the baseline methods, as shown in cross-dataset experiments.

- We establish the new state of the arts on both the `Messidor` and `PALM` datasets.

5.2 Related Work

Earlier work usually utilize hand-craft features to encode anatomical relationships among optic discs (OD), blood vessels, and fovea regions for fovea localization. Deka *et al.* [36] and Medhi *et al.* [92] generate the region of interest (ROI) using processed blood vessels for macula estimation. Certain methods utilize OD in the prediction of ROI and fovea center by selecting specific OD diameters [97], estimating OD orientations and minimum intensity values [127, 11]. Other applications use combined OD and blood vessels features to improve the performance of fovea localization [85, 6]. These methods generally perform less competitively than more recent deep-learning-based approaches.

Many deep learning-based methods formulate the fovea localization as a regression task [2, 94, 65, 159]. Some methods utilize retinal structures, such as OD and blood vessels, as constraints for inferring the location of the fovea. For example, Meyer *et al.* [94] adopt a pixel-wise distance regression approach for joint OD and fovea localization. Besides the regression-based approaches, Sedai *et al.* [126] propose a two-stage image segmentation framework for segmenting the image region around the fovea. Our work also belongs to the image segmentation paradigm [21, 126, 118, 109, 165]. Unlike all previous works, we customize the recent transformer-based segmentation network [21] to incorporate blood vessel information and demonstrate its superior performance compared to the existing approaches.

5.3 Methodology

5.3.1 Network Architecture

The overall architecture of Bilateral-ViT is illustrated in Fig. 5.1. The proposed Bilateral-ViT is based on a U-shape architecture with a vision transformer-based encoder (**the main branch**) for exploiting long-range contexts. In addition, we design a **vessel branch** to encode structure information from blood vessel segmentation maps. Finally, Multi-scale Feature Fusion (MFF) blocks are designed to effectively fuse data from the main and vessel branches.

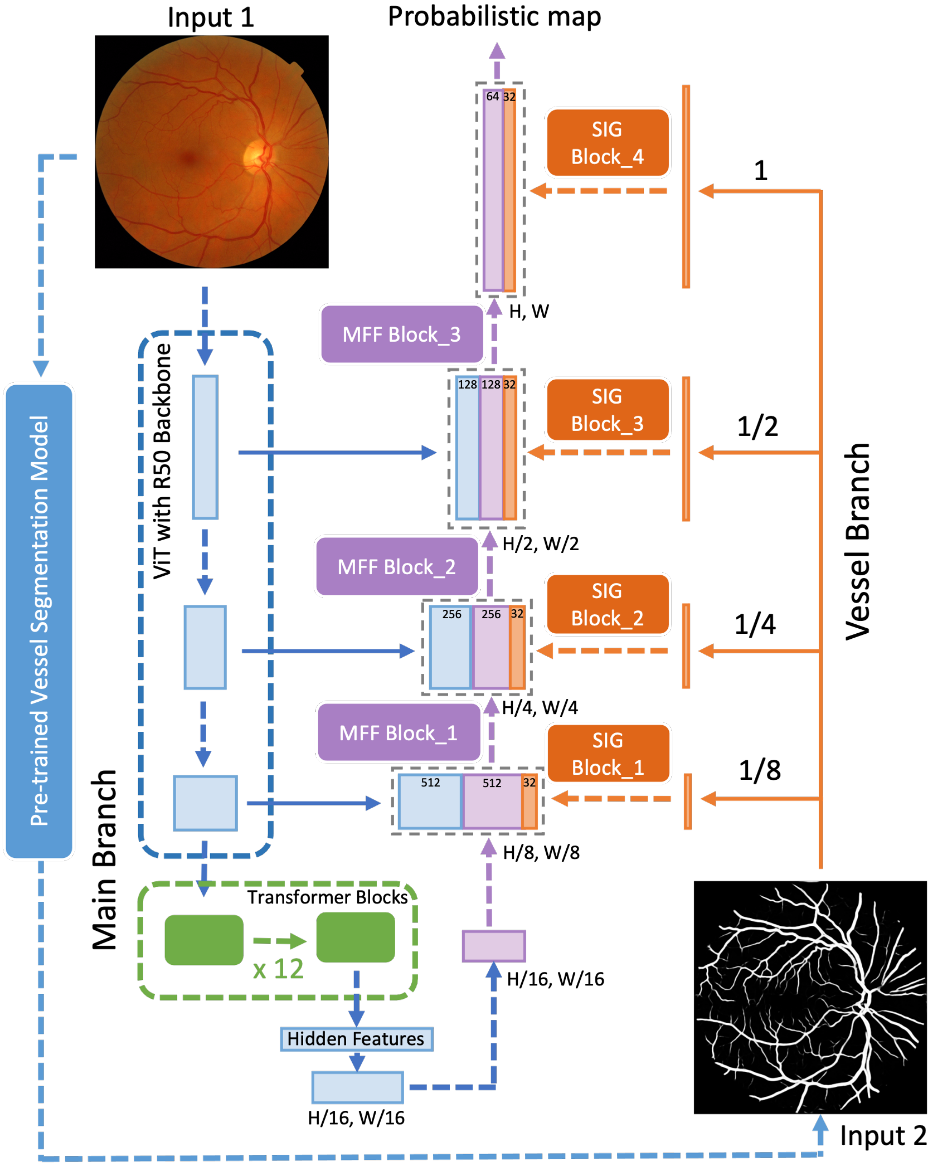


Figure 5.1: The overall architecture of our proposed Bilateral-ViT network.

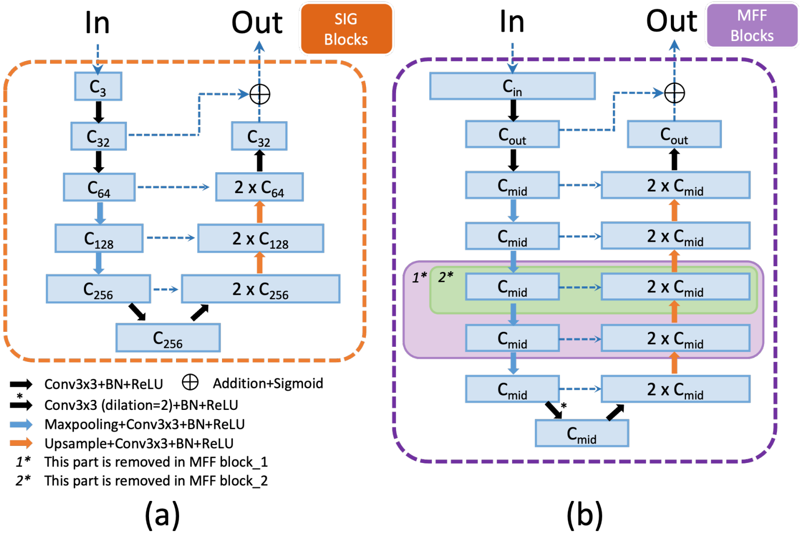


Figure 5.2: The structures of SIG blocks and MFF blocks. The subscript C denotes channel depths. C_{in}, C_{mid} and C_{out} represent channel depths of input, intermediate, and output feature maps for the MFF blocks, respectively. We set C_{mid} of three MFF blocks to small numbers, *i.e.*128, 64, 32, for improving the efficiency of multi-scale feature fusion.

Main Branch. We adopt the TransUNet [21] as the main branch due to its superior performance on other medical image segmentation tasks. In the main branch, we utilize a CNN-Transformer hybrid structure as the encoder. The CNN part is used as the initial feature extractor. It provides features at different scales for the skip connections to compensate for the information loss in the downsampling operation. The extracted features are then processed by 12 consecutive transformer blocks at the bottleneck of the UNet architecture. The transformer encodes the long-range dependencies of the input fundus image due to the multi-head self-attention structure. The output features of the last transformer block are then resized for later decoding operations.

Vessel Branch. In the vessel branch, we aim to exploit the structure information from the blood vessels. Unlike the main branch, where the input is a fundus image, we put in a vessel segmentation map generated by a pre-trained model. The pre-trained vessel segmentation model is built on the DRIVE dataset [140] with the TransUNet [21] architecture. Four identical Spatial Information Guidance (SIG) blocks are utilized in the vessel branch to extract multi-scale vessel-based features. The rescaled vessel segmentation maps are fed into the SIG blocks, the details of which are illustrated in Fig. 5.2-a. The design of the SIG blocks makes extensive use of customized ReSidual U-blocks (RSU). Qin *et al.* [109] indicate that the RSU block is superior in performance to other embedded structures (*e.g.*, plain convolution, residual-like, inception-like, and dense-like blocks), due to the enlarged receptive fields of the embedded U-shape architecture.

Multi-scale Feature Fusion (MFF) blocks. In contrast to the plain convolutional decoder blocks of the basic TransUNet, we use three Multi-scale Feature Fusion (MFF) blocks as the decoders for effective multi-scale feature fusion. The input to each MFF block is the concatenation of three types of features: (1) the multi-scale skip-connection features from the main branch, (ii) the hidden feature encoded by the last transformer block or the previous MFF block, (iii) the multi-scale SIG features from the vessel branch. The architecture of the MFF blocks is illustrated in Fig. 5.2-b, which is similar to one of the SIG blocks. From MFF block_1 to MFF block_3, we gradually increase the number of network layers in each MFF block. In this way, the later MFF blocks can capture more spatial context corresponding to larger feature maps. In the end, the concatenated feature maps of MFF block_3 and SIG block_4 are passed to two convolutional layers for outputting the fovea region score maps.

5.3.2 Implementation Details

We first remove the uninformative black background from the original fundus image, then pad and resize the cropped image region to a spatial resolution of 512×512. We perform intensity normalization and data augmentation on the input images of the main branch and the vessel branch. To train our Bilateral-ViT network, we generate circular fovea segmentation masks from the ground-truth fovea coordinates. During the testing phase, we apply the sigmoid function to network prediction for the probabilistic map. We then collect all pixels with significant probabilistic scores and calculate their median coordinates as the final fovea location coordinates.

All experiments were coded using PyTorch and conducted on one NVIDIA GeForce RTX TITAN GPU. The weights of convolutional and linear layers were initialized by Kaiming initialization protocol [58]. The initial learning rate was $1e^{-3}$ which gradually decays to $1e^{-7}$ over 200 epochs using the Cosine Annealing LR strategy. The optimizer was Adam [72] and the batch size 2. We employed a combination of dice loss and binary cross-entropy as the loss function.

Table 5.1: Comparison of performance on normal and diseased retinal images using the `Messidor` and `PALM` datasets. The best and second best results are highlighted in bold and italics respectively.

	1/8 R(%)		1/4 R(%)		1/2 R(%)		1R(%)		2R(%)	
`Messidor`	Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased
UNet (2015) [118]	82.65	79.00	95.15	93.33	97.76	95.00	97.95	95.33	97.95	95.33
U2 Net (2020) [109]	86.19	81.33	**98.51**	97.33	*99.63*	99.50	*99.63*	99.50	*99.63*	99.50
TransUNet (2021) [21]	*87.31*	**84.33**	*98.32*	*97.67*	**100.00**	*99.83*	**100.00**	*99.83*	**100.00**	*99.83*
Bilateral-ViT (**Proposed**)	**87.50**	*84.00*	**98.51**	**98.67**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

	1/8 R(%)		1/4 R(%)		1/2 R(%)		2/3 R(%)		1R(%)	
`PALM`	Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased
UNet (2015) [118]	57.45	9.43	74.47	18.87	76.60	41.51	76.60	50.94	76.60	64.15
U2 Net (2020) [109]	*70.21*	*11.32*	*93.62*	*28.30*	*95.74*	*60.38*	95.74	*77.36*	*97.87*	*84.91*
TransUNet (2021) [21]	**82.98**	5.66	**95.74**	18.87	**97.87**	43.40	*97.87*	52.83	*97.87*	75.47
Bilateral-ViT (**Proposed**)	**82.98**	**13.21**	**95.74**	**37.74**	**97.87**	**69.81**	**100.00**	**81.13**	**100.00**	**92.45**

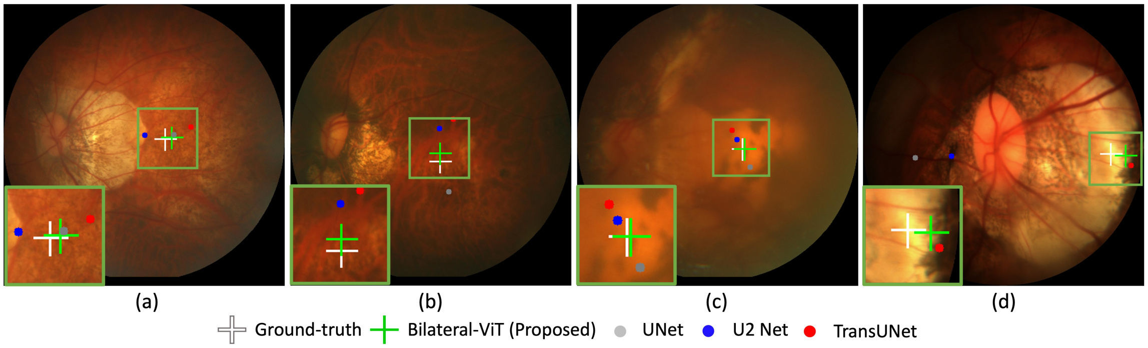


Figure 5.3: Visual results of fovea localization predicted by different methods.

5.4 Experiments

We performed experiments using the `Messidor` [34] and `PALM` [44] datasets. The `Messidor` dataset is for diabetic retinopathy analysis. It consists of 540 normal and 660 diseased retinas. We utilized 1136 images from this dataset with fovea locations provided by [48]. The `PALM` dataset was released for the Pathologic Myopia Challenge (PALM) 2019. It consists of 400 images annotated with fovea locations, in which 213 images are pathologic myopia, and the remaining 187 images are normal retinas. For fairness of comparison, we keep our data split identical to [159].

To evaluate the performance of fovea localization, we adopt the following evaluation protocol [48]: the fovea localization is considered successful when the Euclidean distance between the ground-truth and predicted fovea coordinates is no larger than a predefined threshold value, such as the optic disc radius R. For a comprehensive evaluation, accuracy

Table 5.2: Comparison with existing studies using the Messidor and PALM datasets based on the R rule. The best and second best results are highlighted in bold and italics respectively.

Messidor	1/8 R (%)	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
Gegundez-Arias *et al.*(2013) [48]	-	76.32	93.84	98.24	99.30
Aquino (2014) [6]	-	83.01	91.28	98.24	99.56
Dashtbozorg *et al.*(2016) [33]	-	66.50	93.75	98.87	-
Girard *et al.*(2016) [51]	-	-	94.00	98.00	-
Molina-Casado *et al.*(2017) [95]	-	-	96.08	98.58	99.50
Al-Bander *et al.*(2018) [2]	-	66.80	91.40	96.60	99.50
Meyer *et al.*(2018) [94]	70.33	94.01	97.71	99.74	-
GeethaRamani *et al.*(2018) [47]	-	85.00	94.08	99.33	-
Zheng *et al.*(2019) [174]	60.39	91.36	98.32	99.03	-
Huang *et al.*(2020) [65]	-	70.10	89.20	99.25	-
Xie *et al.*(2020) [159]	*83.81*	*98.15*	*99.74*	*99.82*	**100.00**
Bilateral-ViT (**Proposed**)	**85.65**	**98.59**	**100.00**	**100.00**	**100.00**

PALM	1/8 R (%)	1/4 R (%)	1/2 R (%)	2/3 R (%)	1R (%)
Xie *et al.*(2020) [159]	-	-	-	*87*	*94*
Bilateral-ViT (**Proposed**)	**46**	**65**	**83**	**90**	**96**

corresponding to different evaluation thresholds (for example, $2R$ indicating the predefined threshold values are set to twice the optic disc radius R) is usually reported.

5.4.1 Fovea Localization on Normal and Diseased Images

In Table 5.1, we evaluate the performance of normal and diseased cases separately. We reimplement several widely used segmentation networks as comparison baselines, such as UNet [118], U2 Net [109], and TransUNet [21]. Bilateral-ViT obtains 100% accuracy from $1/2R$ to $1R$ on all the Messidor images, and 100% accuracy from $2/3R$ to $1R$ on the normal PALM images. Thus demonstrating that the performance of Bilateral-ViT is highly reliable for normal fundus images.

For the diseased cases in the PALM dataset, Bilateral-ViT reaches 92.45% foveal localization accuracy for the threshold of $1R$ and significantly outperforms the second-best results by a large margin (7.54%). Fig. 5.3 provides some visual results of fovea localization on diseases images from the PALM dataset. Our Bilateral-ViT generates the most accurate predictions for the severely diseased images with large atrophic regions (see Fig. 5.3-a and Fig. 5.3-b), or the heavily blurred image (see Fig. 5.3-c). In Fig. 5.3-d where the fovea is close to the image border, the predicted fovea locations from baseline networks (UNet and

Table 5.3: **Top** and **Bottom**: Performance of the ablation study using the Messidor and PALM datasets respectively. VB refers to the vessel branch. The best and second best results are highlighted in bold and italics.

Messidor	1/8 R (%)	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
ViT+plain decoder (TransUNet [21])	**85.74**	97.98	*99.91*	*99.91*	*99.91*
ViT+VB+plain decoder	85.56	*98.33*	99.74	*99.91*	*99.91*
ViT+VB+MFF (**Proposed**)	*85.65*	**98.59**	**100.00**	**100.00**	**100.00**
ViT+VB (fundu as the input)+MFF	*85.65*	97.89	*99.91*	**100.00**	**100.00**

PALM	1/8 R (%)	1/4 R (%)	1/2 R (%)	2/3 R (%)	1R (%)
ViT+plain decoder (TransUNet [21])	42	55	69	74	86
ViT+VB+plain decoder	*45*	52	72	77	85
ViT+VB+MFF (**Proposed**)	**46**	**65**	**83**	**90**	**96**
ViT+VB (fundu as the input)+MFF	43	*58*	*82*	*89*	**96**

U2 Net) appear on the wrong side of the optic disc. However, TransUNet [21] and our method still perform well, potentially due to their long-range modeling capability. Such results highlight that our proposed Bilateral-ViT has a significant advantage for diseased cases.

5.4.2 Comparison with State-of-the-Art Methods

From Table 5.2, the Bilateral-ViT achieves state-of-the-art performance for all the evaluation settings. In particular, on the Messidor dataset, at $1/8R$, our network reaches the best accuracy of 85.65% with a gain of 1.84% compared to the second-best score (83.81%) [159]. It also reaches an accuracy of 100% at evaluation thresholds of $1/2R$, $1R$, and $2R$; in other words, the localization errors are at most $1/2R$ (approximately 19 pixels for an input image size of 512×512). PALM is a considerably more challenging dataset due to fewer images and complex diseased patterns. Our method achieved accuracies of 90% and 96% at $2/3R$ and $1R$, which is 3% and 2% better than the previous work [159], respectively.

5.4.3 Ablation Study and Cross-Dataset Experiments

We conducted a comprehensive set of ablation experiments to evaluate the effectiveness of different components (see Table 5.3):

- ViT+plain decoder: the TransUNet architecture [21] comprised of a vision transformer-based encoder and a plain decoder used as the comparison baseline.

Table 5.4: Performance of cross-dataset experiments. The models used here are exactly those in the **Bottom** of Table 5.3. They were constructed using `PALM` only and generated the following results on `Messidor`. The higher results based on the R, and the lower results based on distance errors, are better. VB refers to the vessel branch. The best and second best results are highlighted in bold and italics respectively.

Cross-Dataset	1/8 R(%)	1/4 R(%)	1/2 R(%)	1R(%)	2R(%)	Errors
Xie *et al.* [159]	-	-	-	95.26	-	22.84
ViT+plain decoder (TransUNet)	77.82	*95.95*	**98.59**	*99.03*	99.30	10.76
ViT+VB+plain decoder	*78.17*	95.69	98.24	98.77	99.12	11.38
ViT+VB+MFF (**Proposed**)	**81.78**	**96.48**	*98.42*	**99.38**	**100.00**	**8.57**
ViT+VB (fundu as the input)+MFF	77.02	94.28	97.62	98.68	*99.47*	*10.69*

- ViT+VB+plain decoder: we add the vessel branch (vessel segmentation mask as the input) to the baseline network.

- ViT+VB+MFF (**the proposed Bilateral-ViT**): we add the vessel branch (vessel segmentation mask as the input) and MFF blocks to the baseline network.

- ViT+VB (fundus as the input)+MFF: we add the vessel branch (fundu image as the input) and MFF blocks to the baseline network. This configuration compares the performance differences between fundus images and vessel segmentation maps as inputs to the vessel branch.

The performance of "ViT+plain decoder (TransUNet)" and "ViT+VB+plain decoder" are similar on both datasets; a possible reason is that the plain decoder does not have adequate capacity to fuse features from the vessel branch and transformer blocks. By further adding MFF blocks, the proposed Bilateral-ViT (ViT+VB+MFF) demonstrates superior performance, suggesting the significance of the customized MFF blocks. The performance of "ViT+VB+MFF' is much better than "ViT+VB (fundus as the input)+MFF", demonstrating the usefulness of the vessel segmentation map. On the other hand, we note that "ViT+VB (fundus as the input)+MFF" outperforms all the existing works, implying our network can achieve the state-of-the-art performance even without the input of a vessel segmentation map.

We conducted cross-dataset experiments to assess the generalization capability of the proposed Bilateral-ViT. The models were trained on the `PALM` dataset and tested on the `Messidor` dataset. From Table 5.4, the accuracy is 99.38% at $1R$, which is a 4.12% improvement over the best-reported result (95.26%). The average localization error for the original image resolution is 8.57 pixels compared to the previous best result of 22.84 pixels. In addition, the proposed Bilateral-ViT outperforms the baselines by a significant margin, especially for $1/8R$, demonstrating its robustness for the cross-dataset setting.

5.5 Conclusion and Future Work

This paper proposes a novel Vision Transformer (ViT) approach for robust fovea localization. It consists of a transformer-based main network branch for integrating global context and a vessel branch for explicitly incorporating the structure of blood vessels. The encoded features are subsequently merged with a customized Multi-scale Feature Fusion (MFF) module. Our experiments demonstrate that the proposed approach has a significant advantage in handling diseased images. It also has excellent generalization capability, as shown in the cross-dataset experiments. Thanks to the transformer-based feature encoder, the incorporation of blood vessel structure, and the carefully designed MFF module, our approach establishes the new state of the arts on both `Messidor` and `PALM` datasets.

Although the proposed Bilateral-ViT surpasses all previous studies, it has two major limitations. First, Bilateral-ViT has a large computational requirement since its encoder is a standard ResNet50-based hybrid ViT architecture. Second, features from fundus and vessel distribution are merged in the decoding stage, which may lead that these features are not global incorporated to predict the location of fovea in hard cases. These limitations will be addressed in Chapter 6.

Chapter 6

Bilateral-Fuser: A Novel Multi-cue Fusion Architecture with Anatomical-aware Tokens for Fovea Localization

In the previous chapter, the proposed Bilateral-ViT models long-range connectivity by proposing a two-branch segmentation architecture that exploits the multi-head self-attention mechanism of transformer networks. However, the high computational requirements and the limited receptive field when merging features remain problems for this approach.

To overcome these problems, we propose a novel dual-stream architecture, named Bilateral-Fuser, for multi-cue fusion in fundus images in this chapter. The proposed architecture utilizes a transformer-based structure that globally incorporates long-range connections from multiple cues, including fundus and vessel distribution. We transfer the feature merging process from decoder to encoder and design the Bilateral Token Incorporation (BTI) module in the Bilateral-Fuser architecture. The BTI module includes TokenLearner and TokenFuser for generating adaptive and learnable tokens and merge features of both cues with long-range dependencies. The proposed method achieves state-of-the-art results on two public datasets, Messidor and PALM, and one private dataset, surpassing previous methods in terms of accuracy, robustness and generalization. It also reduces the computational cost by using an attention mechanism to reduce the number of tokens. Moreover, we show that, due to the guidance of the input vessel segmentation map, the learned tokens are anatomical-aware and distributed along the vessel distribution.

6.1 Introduction and Problem Statement

Accurate localization of the fovea is a crucial step in analyzing retinal diseases since it helps prevent irreversible vision loss. Although current deep learning-based methods achieve better performance than traditional methods, they still face challenges such as inadequate utilization of anatomical landmarks, sensitivity to diseased retinal images, and various image conditions. In this paper, we propose a novel transformer-based architecture (Bilateral-Fuser) for multi-cue fusion. The Bilateral-Fuser explicitly incorporates long-range connections and global features using retina and vessel distributions to achieve robust fovea localization. We introduce a spatial attention mechanism in the dual-stream encoder to extract and fuse self-learned anatomical information. This design focuses more on features distributed along blood vessels and significantly reduces computational costs by reducing token numbers. Our comprehensive experiments demonstrate that the proposed architecture achieves state-of-the-art performance on two public datasets and one large-scale private dataset. Moreover, we show that the Bilateral-Fuser is more robust on both normal and diseased retina images and has better generalization capacity in cross-dataset experiments.

The fovea, an anatomical landmark of the retina, is responsible for sharp central vision at the center of the macula [156]. Accurate detection of the macula and fovea is a crucial prerequisite for the diagnosis of several retinal diseases, *e.g.*, diabetic maculopathy and age-related macular degeneration [6, 36, 92]. The severity of vision loss is often related to the distance between the fovea and associated abnormalities, such as hemorrhages and exudates [92].

Early detection of the fovea location is important to prevent the irreversible damage to vision [35, 50, 36]. A robust method of fovea localization is crucial for downstream tasks in automated fundus diagnosis. However, several challenges to fovea localization remain. First, the dark appearance of fovea is indistinguishable from the color intensity of the surrounding retinal tissue, and local anatomical landmarks (*e.g.*, blood vessels) are absent in the vicinity of the fovea [127, 11, 92]. Second, the accuracy of fovea localization may be affected by the occurrence of retinal diseases [85, 97, 47, 53]. For example, dark pathology caused by hemorrhages and microaneurysms may obscure the distinction between the fovea and retinal background. Bright lesions, such as exudates, may change the lightness of the fovea to bright rather than dark, leading to erroneous localization results. Third, poor light conditions and non-standard fovea locations during photography increase the difficulty of robust fovea localization [101, 159, 137]. Specifically, blurred and poorly illuminated photographs present challenges in estimating macula. For images where the optic disc (OD), rather than the macula, is centrally located, symmetry may lead to predictions opposite to the ground-truth. Therefore, a robust fovea localization method is necessary to model features of the entire image at a global scale.

Fortunately, other anatomical structures outside the fovea, such as blood vessels, are useful for localization [85, 97, 127, 11, 48, 49, 6, 36, 33, 51, 92, 95, 53, 137]. Previous works have utilized morphological methods to model the anatomical relationships between the fovea and blood vessels [85, 6, 36, 92, 53]. However, these morphological methods may fail when the image has rare fovea position and color intensity, as described above. Fovea localization is an important upstream task in clinical diagnosis, helping to diagnose maculopathy and abnormalities [6, 36]. Although recent works have employed deep learning methods to improve performance, they typically only utilize fundus images as input [126, 2, 47, 94, 101, 65, 159, 15]. These works are also implemented on datasets containing few challenging images, resulting in three main pitfalls: 1) inadequate exploitation of the anatomical structure outside the macula as only fundus images are used as input; 2) typical convolution-based architectures lacking incorporation of global features; and 3) sensitivity to challenging cases, such as rare fovea positions and severe lesions.

To address these challenges, we propose a novel architecture, Bilateral-Fuser, which is an updated version of our previous work [137] (achieving best-paper-award finalist in *ISBI2022*). Inspired by TransFuser[108], we design a dual-stream encoder to fuse multi-cue features and a decoder to generate result maps. To utilize the anatomical structure outside the macula, the encoder's inputs are images from two different cues (*i.e.*, fundus and vessel distribution). We fuse the multi-cue features of fundus and vessel in four transformer-based modules, named Bilateral Token Incorporation (BTI), in the encoder. This design allows the modeling of global features and long-range connections for fovea localization, ensuring robust performance even in challenging images. Unlike TransFuser, it directly reduces and recovers token numbers, applying average pooling and bilinear interpolation methods, respectively. Such operations may lead to information loss. Thus we avoid information loss by applying TokenLearner [122] in the BTI module. The attention mechanism of TokenLearner extracts self-learning spatial information from both cues. Our design effectively exploits structural features along the optic disk and vessel distribution, and the attention mechanism reduces the number of tokens in the BTI module, significantly reducing computational effort.

Our work makes the following key contributions:

- We propose a novel dual-stream architecture for multi-cue fusion. Compared to typical convolutional-based fusion, this transformer-based structure globally incorporates long-range connections from multiple cues.
- We introduce the BTI module with learnable tokens to improve the efficiency of transformer-based fusion. The adaptive learning of tokens significantly reduces the token number from 1024 to 64. The spatial attention mechanism of the learnable tokens focuses more on features along the vessel distribution, leading to robust fovea localization.
- The proposed Bilateral-Fuser achieves state-of-the-art performance on three datasets (`Messidor`, `PALM` and `Tisu`) at only 25% computational cost (62.11G FLOPs) compared to the best previous work [137] (249.89G FLOPs). It also offers better performance and generalization capability in challenging cases.

6.2 Related Work

6.2.1 Anatomical Structure-based Methods

Previous studies have typically relied on traditional image processing techniques to estimate fovea regions, as the approximate location of the macula is anatomically correlated with the optic disc (OD) and blood vessels [85, 97, 6, 36, 92, 53, 45]. The fovea center is located approximately 2.5 OD diameters from the center of the OD and is on the symmetric line of the main vessel branches that pass through the OD. These two features have been widely used for fovea localization [85, 97, 6, 92, 53, 45].

Some studies detect the fovea region based on OD location only. Narasimha *et al.* [97] propose a two-step approach that incorporates the distance from the OD center and the image intensity to update the region of interest (ROI), and then locates the fovea center. Sekhar *et al.* [127] use the spatial relationship to select a sector-shaped candidate ROI. The boundary of the sector is 30 degrees above and below the line through the center of the image and OD. They then use a threshold to filter the intensity to estimate the fovea region. Blood vessels in color fundus images are relatively darker structures compared with OD. Some works utilize only the extracted the skeleton image of vessels to estimate ROI containing the macula. Deka *et al.* [36] and Medhi *et al.* [92] divide the image into several horizontal strips and select the ROI with respect to the absence of blood vessels in the neighborhood of the macula. They then utilize thresholds to detect macula. Guo *et al.* [53] propose a morphological method to fit the segmented skeleton of major vessels using a parabola. The line of symmetry of the parabola is used to localize the fovea region.

The OD and vessels have been widely utilized in fovea localization due to their anatomical relationship. Asim *et al.* [11] estimate ROI based on the pre-detected OD location and minimum intensity values, and exclude the ROI near the vascular tree to improve the accuracy. Li *et al.* [85], Aquino *et al.* [6] and Fu *et al.* [45] also use a parabolic fit of major vessels to detect the orientation of the macula and use the anatomical relationship (*i.e.*, distance) between OD and fovea center to estimate the approximate location. The difference is that Fu *et al.* [45] use a deep learning method (U-Net [118]) rather than an image processing method to detect OD and vessels. However, these methods based on anatomical features may underperform when processing pathological images. Additionally, they generally perform less competitively than more recent deep learning-based approaches.

Some attempts have been made to localize the macula without relying on anatomical features. For example, GeethaRamani and Balasubramanian [47] propose an approach to segment the macula using an unsupervised clustering algorithm. Pachade *et al.* [101] directly select the square in the middle of the image as ROI and use a filter on intensity for fovea localization. However, as these methods do not consider anatomical features, they may fail when the illumination is different or the macula is not found in a standard location (*i.e.*, the center of the image).

6.2.2 Deep Neural Networks in Fovea Localization

Deep learning has demonstrated superiority over traditional image processing and morphological techniques in many fields of medical image analysis, such as classification, segmentation and object localization [145, 96, 109, 165, 163]. Regarding the task of fovea localization, existing studies can be broadly classified into two categories: regression and segmentation.

Many deep learning-based methods formulate fovea localization as a regression task. Al-Bander *et al.* [2] and Huang *et al.* [65] propose a two-step regression approach that first predicts ROI and subsequently feeds the ROI into neural networks to localize the fovea center. Meyer *et al.* [94] and Bhatkalkar *et al.* [15] adopt pixel-wise distance or heatmap regression approaches for joint OD and fovea localization. Xie *et al.* [159] propose a hierarchical regression network that employs a self-attention mechanism in fovea localization [28, 167]. The network predicts the fovea center through a three-stage localization architecture that crops features from coarse to fine.

In addition to regression, deep learning-based methods also employ the image segmentation paradigm. Tan *et al.* [147] design a single 7-layer convolutional network to point-wise predict the fovea region from input image patches. Sedai *et al.* [126] propose a two-stage image segmentation framework for segmenting the fovea region from coarse to fine. However, standard CNN-based architectures are limited by their fixed-size convolutional kernels, resulting in a lack of incorporation of long-range features. Consequently, these CNN-based architectures may fail when light conditions and fovea positions are abnormal, or when information on the OD and vessels is lacking due to lesions.

To overcome the issue of limited receptive field, our previous work models long-range connections by proposing a two-branch segmentation architecture (Bilateral-ViT [137]), which utilizes a multi-head self-attention (MHSA) mechanism of transformer networks [151, 41, 21]. The main branch of Bilateral-ViT consists of 12 consecutive MHSA layers in the bottleneck, constituting the global features for the decoder. An additional vessel branch is designed to extract multi-scale spatial information from the vessel segmentation map as the second input. The decoder of Bilateral-ViT simultaneously fuses multi-cue features between the fundus and blood vessel distribution, achieving the best-reported results on two public datasets, `Messidor` [34] and `PALM` [44]. However, the multi-scale convolutional operation in the decoder has two main limitations, (1) non-global multi-cue feature fusion and (2) computationally expensive. To overcome these limitations, we propose a novel architecture, named Bilateral-Fuser, which includes an encoder for global-connected multi-cue fusion and introduce adaptively learnable tokens to reduce computational amount.

6.3 Methodology

In this study, we propose a novel multi-cue fusion architecture, Bilateral-Fuser (Fig. 6.1), for accurate and robust fovea localization. Bilateral-Fuser utilizes a U-shape architecture, where the encoder is a dual-stream structure comprising a main stream, a satellite stream, and four intermediate Bilateral Token Incorporation (BTI) modules for exploiting and

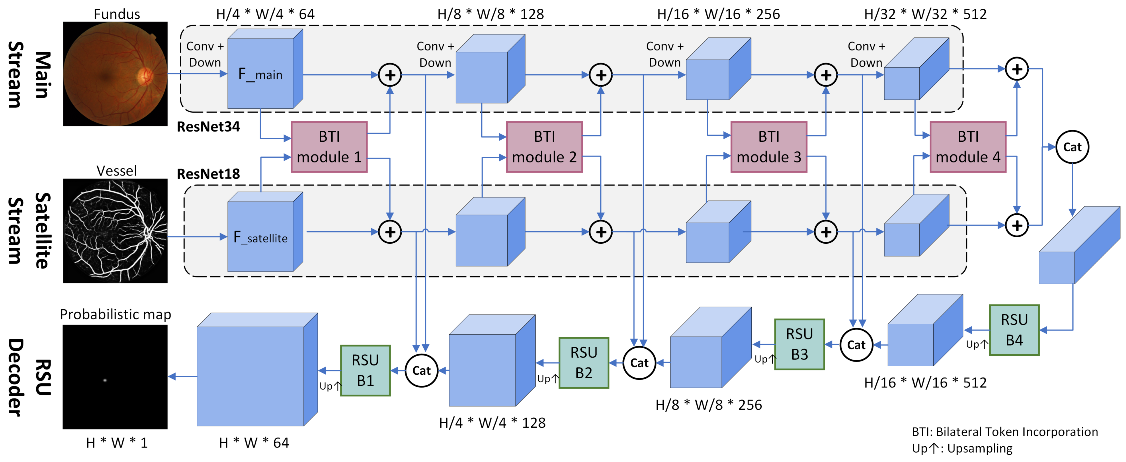


Figure 6.1: The overall architecture of our proposed Bilateral-Fuser network.

fusing global features from different cues. For the decoder, we employ several ReSidual U-blocks (RSU) [109] to effectively incorporate features from both the main and satellite streams.

6.3.1 Overall Architecture

The overall architecture of Bilateral-Fuser is illustrated in Fig. 6.1. In the encoder, the backbones of the main and satellite streams are ResNet34 and ResNet18, respectively. The main stream extracts detailed features from fundus images, while the satellite stream extracts anatomical structure information from the distribution of blood vessels. Unlike the main stream, which takes fundus images as input, the satellite stream takes a vessel segmentation map generated by a pre-trained model as input. This pre-trained vessel segmentation model is built on the DRIVE dataset [140] using the TransUNet [21] architecture, which is identical to that used in [137].

The dual-stream encoder with four intermediate modules for multi-cue fusion is inspired by PVT [154] and TransFuser [108]. Each stream's backbone is divided into four convolutional blocks, consisting of convolution and downsampling layers (Conv+Down). The resulting intermediate tensors (F_{main} and $F_{\text{satellite}}$) are then fed into the BTI module, which includes a TokenLearner, T consecutive Multi-Head Self-Attention (MHSA) layers, and a TokenFuser. The BTI module fuses multi-cue features and encodes long-range dependencies from both the fundus and vessel distribution. The output features are element-wise

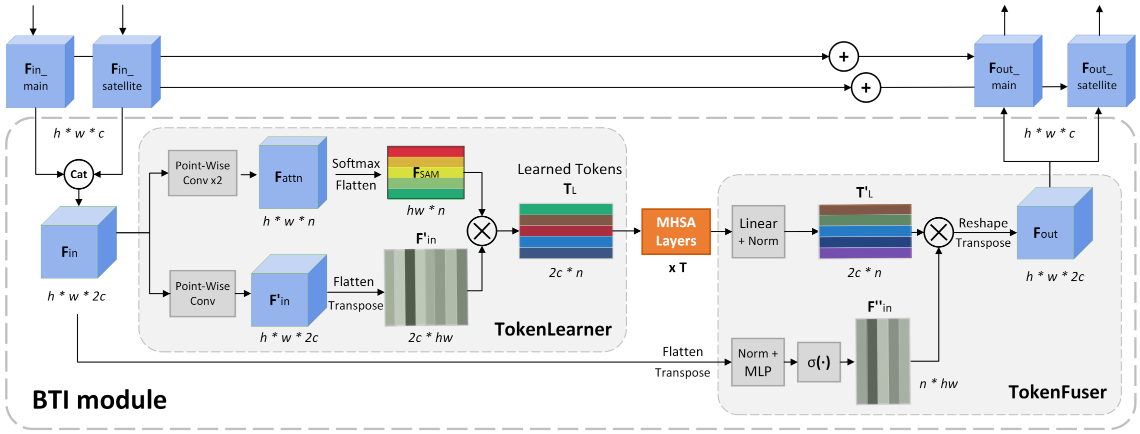


Figure 6.2: The structure of BTI module used in our Bilateral-Fuser. It contains TokenLearner, $T\times$ MHSA layers and TokenFuser. The h, w and c are height, width, and channel of the corresponding input features. The n represents the number of learned tokens.

summed with skip-connected features and fed into the next convolution and downsampling layers. In addition, these output features from the BTI module are also forwarded to RSU blocks for subsequent decoding operations.

Unlike the commonly used plain convolutional blocks in the basic UNet decoder, four customized ReSidual U-blocks (RSU) [109] are utilized in the decoder of Bilateral-Fuser for effective multi-scale feature incorporation. The design of the RSU blocks is identical to that used in our previous work [137]. As shown in Fig. 6.1, RSU B4 is the bottleneck between the encoder and decoder. The input to the other three RSU blocks is a concatenation of three types of features: (i) multi-scale skip-connection features from the main stream, (ii) multi-scale skip-connection features from the satellite stream, (iii) the hidden feature decoded by the previous RSU block. Qin *et al.* [109] demonstrate that the RSU block is superior in performance to other embedded structures (*e.g.*, plain convolution, residual-like, inception-like, and dense-like blocks), due to the enlarged receptive fields of the embedded U-shape architecture. Moreover, the superiority of the RSU structure as the decoder for incorporating multiple features has also been assessed by [137].

6.3.2 Bilateral Token Incorporation (BTI) modules

Standard transformer/MHSA-based architectures, such as Vision Transformer (ViT) [41] and TransUNet [21], typically split the input image into 2D windows (*e.g.*, 16 × 16 grid) to generate tokens. The tokenization output is then fed into subsequent MHSA layers to model long-range feature connectivity. However, these tokens are extracted individually from a fixed-size grid. Recent architectures with multiple transformer stages [154, 108] have several times more MHSA layers than the standard ViT. The naive tokens extracted from the grids may contain uninformative or irrelevant features for visual understanding, which is computationally expensive.

To alleviate these pitfalls, the Bilateral Token Incorporation (BTI) module is introduced in the Bilateral-Fuser architecture (Fig. 6.2). The BTI module includes the TokenLearner [122], which adaptively learns tokens using a spatial attention mechanism. After being processed by MHSA layers ($head = 8$ and $layer = 12$ for each BTI module), the tokens are remapped by TokenFuser [122] to the original input tensor dimensions (Fig. 6.2). Therefore, the BTI module has two main advantages: (1) generating adaptive and learnable tokens to reduce token numbers, and (2) merging and fusing features with long-range dependencies of both cues with high efficacy.

TokenLearner

Let $\mathbf{F}_{\text{in\_main}} \in \mathbb{R}^{h\times w\times c}$ and $\mathbf{F}_{\text{in\_satellite}} \in \mathbb{R}^{h\times w\times c}$ be the input tensors of the two streams, where h, w and c represent the height, width and channel of the corresponding BTI module. As shown in Fig. 6.2, the concatenated feature $\mathbf{F}_{\text{in}} \in \mathbb{R}^{h\times w\times 2c}$ is first fed into the TokenLearner. We customize the TokenLearner from [122] to generate $\mathbf{F}_{\text{attn}} \in \mathbb{R}^{h\times w\times n}$ using two consecutive point-wise convolutional layers to reduce the dimensionality, where n is the number of learned tokens. After applying the flatten and softmax functions, spatial attention maps with a dimension of $hw \times n$ are generated. Moreover, $\mathbf{F}_{\text{in}}$ is processed by a point-wise convolutional layer, which is then flattened and transposed to become $\mathbf{F}'_{\text{in}} \in \mathbb{R}^{2c\times hw}$. To adaptively learn tokens using the spatial attention mechanism, the tokenization function is given by:

$$\mathbf{T}_{\text{L}} = \mathbf{F}'_{\text{in}}\mathbf{F}_{\text{SAM}} \tag{6.1}$$

$$\mathbf{F}_{\text{SAM}} = \text{softmax}(\text{flatten}(\mathbf{F}_{\text{attn}})) \tag{6.2}$$

where the learned tokens are denoted as $\mathbf{T}_{\text{L}} \in \mathbb{R}^{2c\times n}$. Because of the spatial attention mechanism, the learned tokens are modeled using an informative combination of corresponding spatial locations. In comparison to the 1024 tokens used in ViT and TransUNet [41, 21], we only retain 8×8 tokens for each BTI module (given an input size of 512×512). Since the computation of MHSA is quadratic to the number of tokens, the computational cost is significantly decreased. Therefore, TokenLearner enables us to not only significantly reduce the number of tokens but also extract features related to the anatomical structure, *i.e.*, vessel distributions in this fovea localization task.

TokenFuser

As shown in Fig. 6.2, the resulting tokens from the the MHSA layers are recovered to their original tensor resolution ($h \times w \times 2c$) for further processing by Bilateral-Fuser. To fuse the information, we first utilize a fully-connected (linear) layer and output $\mathbf{F}'_{\mathrm{L}} \in \mathbb{R}^{2c \times n}$. By processing $\mathbf{F}_{\mathrm{in}}$ simultaneously, the output tensor $\mathbf{F}_{\mathrm{out}}$ is given by:

$$\mathbf{F}_{\mathrm{out}} = (\mathbf{T}'_{\mathrm{L}} \mathbf{F}''_{\mathrm{in}})^T \tag{6.3}$$

$$\mathbf{F}''_{\mathrm{in}} = \sigma(\mathrm{MLP}(\mathrm{flatten}(\mathbf{F}_{\mathrm{in}})^T)) \tag{6.4}$$

where tensors $\mathbf{F}_{\mathrm{in}}$, $\mathbf{F}_{\mathrm{out}} \in \mathbb{R}^{h \times w \times 2c}$ and $\mathbf{F}''_{\mathrm{in}} \in \mathbb{R}^{n \times hw}$. $\sigma(\cdot)$ is a sigmoid function and MLP represents two dense layers with an intermediate GeLU activation function. In this case, the modeled tokens $\mathbf{T}'_{\mathrm{L}}$ are remapped to $\mathbf{F}_{\mathrm{out}}$ which has the same resolution as the initial $\mathbf{F}_{\mathrm{in}}$. Following this, $\mathbf{F}_{\mathrm{out}}$ is equally split to $\mathbf{F}_{\mathrm{out\_main}}$&$\mathbf{F}_{\mathrm{out\_satellite}}$ and added element-wise to the skip-connected features (*i.e.*, $\mathbf{F}_{\mathrm{in\_main}}$&$\mathbf{F}_{\mathrm{in\_satellite}}$).

6.4 Experiments

6.4.1 Datasets and Network Configurations

We first conduct experiments using the `Messidor` [34] and `PALM` [44] datasets. The `Messidor` dataset was developed for analyzing diabetic retinopathy and comprises 540 normal and 660 diseased retinal images. For this dataset, we utilize 1136 images fovea locations provided by [48]. The `PALM` dataset was released for the Pathologic Myopia Challenge (PALM) 2019, which contains 400 images with fovea locations annotated. Of these, 213 images are pathologic myopia images, and the remaining 187 are normal retina images. For fairness of comparison, we follow the same data split as in the existing studies [159] and [137].

We also use a large-scale dataset (4103 images, named `Tisu`) which is collected from our cooperating hospital. All data have been desensitized with patients' personal informa-

Table 6.1: Configuration Comparison of Bilateral-ViT models and the proposed Bilateral-Fuser.

Methods	Tokens	Image Size	FLOPs	GPU/T[a]	GPU/I[b]
Bi-ViT [137]	32 × 32	512^2	249.89	16873	5459
Bi-ViT/Lit [137]	8 × 8	512^2	83.05	8653	3093
Bi-Fuser (**Ours**)	8 × 8	512^2	62.11	8083	2727

[a] The GPU usage when training (MiB)
[b] The GPU usage when inferencing (MiB)

tion removed. Compared to the `Messidor` and `PALM` datasets, `Tisu` is more challenging as it contains a larger number of fundus images with various abnormalities besides hemorrhages, microaneurysms, and exudates. The ground-truth fovea centers are determined by averaging the labels provided by three medical experts. The dataset is split into training and testing sets in a 4:1 ratio.

One of our main contributions is the considerable reduction in FLOPs and GPU usage of our proposed Bilateral-Fuser. Specifically, compared to the previous state-of-the-art Bilateral-ViT [137], our proposed Bilateral-Fuser consumes approximately 0.25 times FLOPs, 0.48 times GPU usage during training and 0.5 times GPU usage during inference. To evaluate whether the performance of Bilateral-ViT is caused by its large computational requirements, we have introduced a light version, Bilateral-ViT/Lit (*i.e.*, Bi-ViT/Lit in Table 6.1). The basic architecture of Bilateral-ViT has been maintained, but we have reduced the number of middle channels in every convolutional block by half and decreased the number of tokens from 32×32 to 8×8. In this case, Bilateral-ViT/Lit has comparable FLOPs and GPU usage to our Bilateral-Fuser for a fair evaluation.

6.4.2 Implementation Details and Evaluation Metrics

To preprocess the fundus images, we first remove the uninformative black background, and then pad and resize the cropped image region to 512 × 512. We simultaneously perform normalization and data augmentation on the input images of the main branch and the vessel branch. To train our Bilateral-ViT model, we generate circular fovea segmentation masks from the annotated fovea coordinates. During inference, we obtain a probabilistic map from the model's output by applying the sigmoid function. The final fovea location coordinates are obtained by calculating the median coordinates of all pixels in the map.

All experiments are implemented in PyTorch and conducted on one NVIDIA GeForce RTX TITAN GPU. We use the Adam optimizer [72] with a batch size of 2. The loss function is a combination of dice loss and binary cross-entropy. The experimental setup for the Bilateral-ViT architectures on `Messidor` and `PALM` datasets is the same as reported in [137]. For our proposed Bilateral-Fuser, we set the initial learning rate to $1e^{-3}$, which is gradually decayed to $1e^{-9}$ using the CosineAnnealingLR strategy over 300 epochs on `Messidor` and `PALM`. For the `Tisu` dataset, we set the initial learning rate to $6e^{-5}$ for all related experiments.

In accordance with the standard evaluation protocol, we adopt the following evaluation metrics to assess the performance of fovea localization [48, 94, 159, 137]: we consider the fovea localization to be successful if the Euclidean distance between the ground-truth and predicted fovea coordinates is no greater than a predefined threshold value, such as the radius of the optic disc R. To provide a comprehensive evaluation, we report the accuracy for different evaluation thresholds from $2R$ to $1/4R$ (*e.g.*, $2R$ indicating that the predefined threshold values are set to twice the radius of the optic disc R).

6.5 Results

6.5.1 Comparison to State of the Art

In Table 6.2, we compare the performance of Bilateral-Fuser with existing methods on the public dataset, `Messidor` and `PALM`. Methods are classified based on whether they use deep learning techniques and whether they incorporate multi-cue features. We observe that the traditional morphological methods [48, 49, 6, 33, 51, 95] rely on landmarks outside the macula, such as vessels or the optic disc. Bilateral-ViT [137] is the only previous deep learning-based method that incorporates fundus and vessel features. However, in most deep learning-based studies [2, 94, 47, 101, 65, 159, 15], only fundus images are used, resulting in poor incorporation of anatomical relationships throughout the entire image, leading to failure in more challenging cases.

The proposed Bilateral-Fuser, which combines a transformer-based multi-cue fusion encoder and adaptive learning tokens, outperforms all previous studies in terms of fovea localization accuracy on the `Messidor` and `PALM` datasets. Specifically, in Table 6.2, Bilateral-Fuser achieves the highest accuracy of 98.86% at $1/4R$, with gains of 0.71% and 3.53% compared to previous works [159] and [15], respectively. Our network also achieves better performance than Bilateral-ViT [137] and its light version. At evaluation thresholds of $1/2R$, $1R$, and $2R$, Bilateral-Fuser achieves 100% accuracy, indicating a localization error

Table 6.2: Comparison with existing studies using the `Messidor` and `PALM` datasets based on the R rule. The best and second best results are highlighted in bold and italics respectively.

`Messidor`	DL[a]	MF[b]	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
Gegundez-Arias *et al.*(2013) [48]	✗	✓	76.32	93.84	98.24	99.30
Giachetti *et al.*(2013) [49]	✗	✓	-	-	99.10	-
Aquino (2014) [6]	✗	✓	83.01	91.28	98.24	99.56
Dashtbozorg *et al.*(2016) [33]	✗	✓	66.50	93.75	98.87	99.58
Girard *et al.*(2016) [51]	✗	✓	-	94.00	98.00	-
Molina-Casado *et al.*(2017) [95]	✗	✓	-	96.08	98.58	99.50
Al-Bander *et al.*(2018) [2]	✓	✗	66.80	91.40	96.60	99.50
Meyer *et al.*(2018) [94]	✓	✗	94.01	97.71	99.74	-
GeethaRamani *et al.*(2018) [47]	✓	✗	85.00	94.08	99.33	-
Pachade *et al.*(2019) [101]	✓	✗	-	-	98.66	-
Huang *et al.*(2020) [65]	✓	✗	70.10	89.20	99.25	-
Xie *et al.*(2020) [159]	✓	✗	98.15	*99.74*	*99.82*	**100.00**
Bhatkalkar *et al.*(2021) [15]	✓	✗	95.33	*99.74*	**100.00**	-
Bi-ViT (2022) [137]	✓	✓	*98.59*	**100.00**	**100.00**	**100.00**
Bi-ViT/Lit (2022) [137]	✓	✓	98.50	**100.00**	**100.00**	**100.00**
Bi-Fuser (**Ours**)	✓	✓	**98.86**	**100.00**	**100.00**	**100.00**

`PALM`	DL[a]	MF[b]	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
Xie *et al.*(2020) [159]	✓	✗	-	-	94	-
Bi-ViT (2022) [137]	✓	✓	*65*	*83*	*96*	**98**
Bi-ViT/Lit (2022) [137]	✓	✓	55	80	94	*96*
Bi-Fuser (**Ours**)	✓	✓	**69**	**85**	**97**	**98**

[a] Whether the method is based on deep learning (DF).

[b] Whether the method is based on multi-cue features (MF), *e.g.*, fundus images, vessels or optical discs.

Table 6.3: Comparison of performance on normal and diseased retinal images using the `Messidor` and `PALM` datasets. The best and second best results are highlighted in bold and italics respectively.

`Messidor`	MF[a]	$\text{FLOPs}_{\downarrow}$	$\text{Err}_{\downarrow}$	1/4 R(%)		1/2 R(%)		1R(%)		2R(%)	
				Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased
UNet (2015) [118]	✗	193.31	12.39	95.15	93.33	97.76	95.00	97.95	95.33	97.95	95.33
U2 Net (2020) [109]	✗	151.00	7.31	98.51	97.33	*99.63*	99.50	*99.63*	99.50	*99.63*	99.50
TransUNet (2021) [21]	✗	168.73	7.61	98.32	97.67	**100.00**	*99.83*	**100.00**	*99.83*	**100.00**	*99.83*
Bi-ViT (2022) [137]	✓	249.89	*6.81*	98.51	**98.67**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Bi-ViT/Lit (2022) [137]	✓	*83.05*	**6.77**	*98.69*	*98.33*	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Bi-Fuser (**Ours**)	✓	**62.11**	**6.77**	**99.07**	**98.67**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

`PALM`	MF[a]	$\text{FLOPs}_{\downarrow}$	$\text{Err}_{\downarrow}$	1/4 R(%)		1/2 R(%)		1R(%)		2R(%)	
				Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased
UNet (2015) [118]	✗	193.31	149.30	74.47	18.87	76.60	41.51	76.60	64.15	78.72	73.58
U2 Net (2020) [109]	✗	151.00	62.62	*93.62*	28.30	*95.74*	60.38	*97.87*	84.91	*97.87*	**98.11**
TransUNet (2021) [21]	✗	168.73	104.38	**95.74**	18.87	**97.87**	43.40	*97.87*	75.47	*97.87*	84.91
Bi-ViT (2022) [137]	✓	249.89	*53.70*	**95.74**	*37.74*	**97.87**	*69.81*	**100.00**	*92.45*	**100.00**	*96.23*
Bi-ViT/Lit (2022) [137]	✓	*83.05*	62.47	87.23	26.42	93.62	67.92	*97.87*	90.57	*97.87*	94.34
Bi-Fuser (**Ours**)	✓	**62.11**	**48.72**	**95.74**	**45.28**	**97.87**	**73.58**	**100.00**	**94.34**	**100.00**	*96.23*

[a] Whether the method is based on multi-cue features (MF), *e.g.*, fundus images, vessels or optical discs.

of at most $1/2R$ (approximately 19 pixels for an input image size of 512×512).

The PALM dataset is more challenging, with a smaller number of images and complex diseased patterns. Our proposed Bilateral-Fuser demonstrates superiority over all other methods on this dataset in Table 6.2, achieving accuracies of 69% and 85% at $1/4R$ and $1/2R$, respectively, which are 4% and 2% better than Bilateral-ViT. In addition, our method achieves a 14% improvement ($1/4R$) over Bilateral-ViT/Lit, and a 3% improvement ($1R$) over both Bilateral-ViT/Lit and [159]. Therefore, our Bilateral-Fuser achieves state-of-the-art performance on both Messidor and PALM datasets with high computational efficiency.

6.5.2 Fovea Localization on Normal and Diseased Images

In Table 6.3, we separately evaluate the performance of fovea localization for normal and diseased cases in Messidor and PALM datasets to assess the robustness of our method. We compare our proposed Bilateral-Fuser to several widely used segmentation networks, including UNet [118], U2 Net [109], and a hybrid version of TransUNet with ResNet50 for patch embedding [21]. The models of Bilateral-ViT, Bilateral-ViT/Lit and our Bilateral-Fuser are identical to those used in Table 6.2.

In Table 6.3, our proposed Bilateral-Fuser achieves the lowest error (Err) with the smallest computational cost (FLOPs) on both datasets. Bilateral-ViT [137] and our proposed Bilateral-Fuser both obtain 100% accuracy from $1/2R$ to $2R$ on all the Messidor images, and achieve 100% accuracy of $1R$ and $2R$ on normal PALM images. Compared to existing methods, Bilateral-Fuser demonstrates superior performance on almost all metrics on Messidor. Although Bilateral-Fuser has only 0.25 times FLOPs (62.11G) compared to Bilateral-ViT (249.89G), it achieves the best performance on diseased images of PALM from $1/4R$ to $2R$, with up to 7.54% improvement compared to the other methods ($1/4R$, Diseased). Its improvement is significantly increased to 18.86% ($1/4R$, Diseased) compared to Bilateral-ViT/Lit (83.05G, *i.e.*, the network with the closest FLOPs to Bilateral-Fuser). Thus, our proposed Bilateral-Fuser is highly reliable in fovea localization on both normal and diseased fundus images with high efficacy.

6.5.3 Comparison of Multi-Cue Fusion Architectures

To comprehensively assess the performance of models with input features from multiple cues (fundus and vessel distributions), we implement a multi-cue fusion version for the

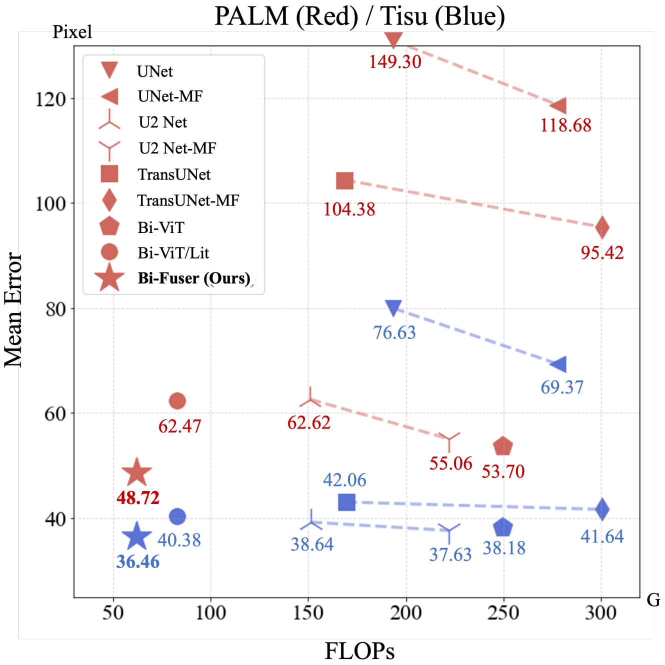


Figure 6.3: Visualization of mean errors (Y-axis) of different multi-cue fusion models. X-axis is the computational cost (FLOPs). The red and blue markers are results on PALM and Tisu datasets, respectively. Numbers below the markers are corresponding mean errors.

baseline models, UNet, U2 Net and TransUNet. We utilize two identical encoders, each with an input of a fundus image and a vessel map. The features are extracted independently and concatenated at the bottleneck for decoding (similar to [134, 108]). These modified baseline models are referred to as UNet-MF, U2 Net-MF and TransUNet-MF. The results in Table 6.3 show that architectures using multi-cue features outperform typical networks with fundus-only input. This is particularly evident in the more challenging PALM dataset, where the improvement is more pronounced. Moreover, the Tisu dataset is more complex than PALM, with more images (4103 *vs.* 400) and a wider range of disease types and severity. Therefore, the results on PALM and Tisu demonstrate the potential of architectures that can effectively handle complex datasets.

Fig. 6.3 shows a comparison of the described architectures (mean error against FLOPs) on PLAM (red markers) and Tisu (blue markers). Below each marker, we provide the corresponding mean error, and dashed lines connect each standard baseline model with its

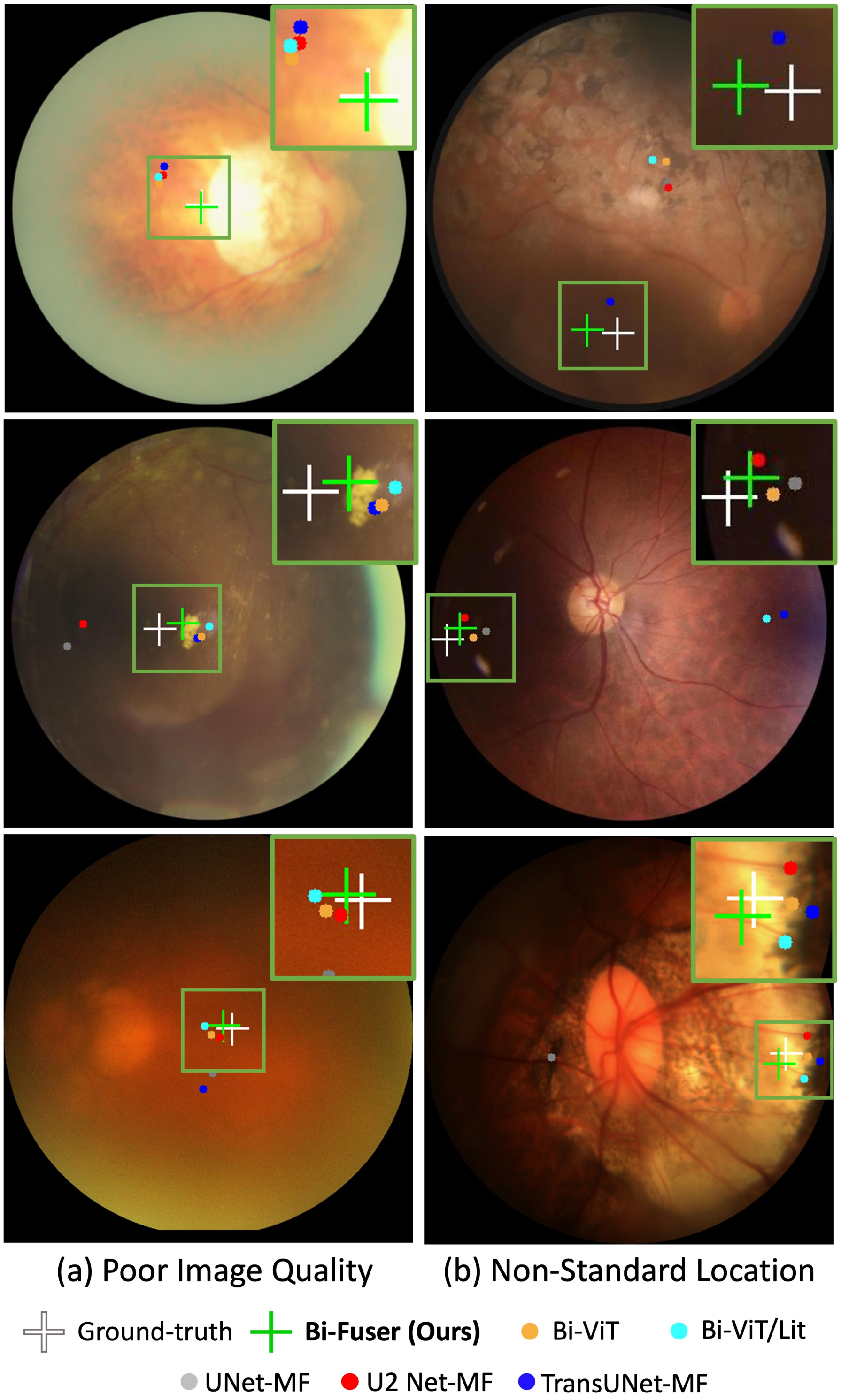


Figure 6.4: Visual results of fovea localization predicted by different methods.

multi-cue fusion architecture (MF). The multi-cue fusion versions (UNet-MF, U2 Net-MF, and TransUNet-MF) outperform their standard versions at a considerably higher computational cost due to the additional encoder. In Fig. 6.3, we can see that the performance comparison (MF versions *vs.* baseline models) is more apparent on PLAM since the dataset size of PLAM is much smaller than Tisu (400 *vs.* 4103). For all multi-cue fusion architectures, our proposed Bilateral-Fuser achieves the best results with the smallest errors on both PLAM (48.72 pixels) and Tisu (36.46 pixels). Moreover, it requires only 62.11G FLOPs, which is four times less than Bilateral-ViT (249.89G). Compared to the model with comparable FLOPs (Bilateral-ViT/Lit, 83.05G), Bilateral-Fuser demonstrates significant advantages of 13.75 and 3.92 on PLAM and Tisu, respectively.

Fig. 6.4 provides visual results of fovea localization on images with severe diseases from the PALM and Tisu datasets. These images in Fig. 6.4-a and Fig. 6.4-b suffer from poor image quality and non-standard fovea locations, respectively. Our Bilateral-Fuser generates the most accurate predictions for several challenging cases with poor lighting conditions and blurred appearance (Fig. 6.4-a). For another challenging types (Fig. 6.4-b), where the macula is close to the image boundary, the predictions of Bilateral-Fuser (green crosses) are closest to the ground-truth (white crosses). In contrast, fovea locations predicted by other architectures that cannot globally incorporate long-range multi-cue features may appear on the wrong side of the optic disc (Fig. 6.4-b). These results suggest that the Bilateral-Fuser architecture can adequately model fundus and vessel features of two streams, resulting in superior performance compared to other networks.

6.5.4 Performance of Cross-Dataset Experiments

We conduct cross-dataset experiments to assess the generalization capability of the proposed Bilateral-Fuser. In Table 6.4, models trained on Tisu (exactly the same ones in Fig. 6.3) are tested on Messidor and PALM datasets. In Table 6.4-**Top**, Bilateral-Fuser generally achieves similar accuracies as the other results from $1/4R$ to $2R$ on Messidor. On the more challenging dataset, PALM, Bilateral-Fuser achieves an improvement of 5.52 and 11.93 pixels in average localization error at the original image resolution compared to Bilateral-ViT (64.73 pixels) and its lighter version (71.14 pixels), respectively (Table 6.4-**Bottom**). Furthermore, Bilateral-Fuser outperforms the baselines (multi-cue fusion version) by a significant margin (at least 10.25 pixels), demonstrating its excellent generalization capability and robustness.

Table 6.4: The performance of cross-dataset experiments. The models used here are exactly those selected in Fig. 6.3 (blue markers). **Top** and **Bottom:** The models trained on `Tisu` and tested on `PALM` and `Messidor`, respectively. The best and second best results are highlighted in bold and italics respectively.

`Tisu`→`Messidor`	Err$_\downarrow$	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
UNet-MF (2015) [118]	8.89	97.45	99.38	*99.65*	*99.65*
U2 Net-MF (2020) [109]	8.48	97.10	*99.91*	**100.00**	**100.00**
TransUNet-MF (2021) [21]	8.25	97.45	99.82	**100.00**	**100.00**
Bi-ViT (2022) [137]	**7.30**	**98.59**	**100.00**	**100.00**	**100.00**
Bi-ViT/Lit (2022) [137]	*7.37*	*98.06*	*99.91*	**100.00**	**100.00**
Bi-Fuser (**Ours**)	7.62	**98.59**	*99.91*	**100.00**	**100.00**

`Tisu`→`PALM`	Err$_\downarrow$	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
UNet-MF (2015)	137.14	52.25	64.25	77.25	85.50
U2 Net-MF (2020)	69.46	55.00	**73.25**	90.00	97.25
TransUNet-MF (2021)	78.98	53.50	71.75	87.50	95.75
Bi-ViT (2022) [137]	*64.73*	**56.00**	*72.75*	90.00	*97.75*
Bi-ViT/Lit (2022) [137]	71.14	*55.50*	**73.25**	*90.75*	97.00
Bi-Fuser (**Ours**)	**59.21**	55.25	**73.25**	**93.50**	**98.50**

Table 6.5: Comparison of performance between different inputs for the main and satellite streams on `PALM` and `Tisu`. The best and second best results are highlighted in bold and italics.

`PALM`	Err$_\downarrow$	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
Fundus+Vessel (**Ours**)	**48.72**	**69**	**85**	**97**	**98**
Fundus-only	*54.46*	*64*	**85**	*95*	**98**
Vessel-only	72.25	57	*75*	92	*97*

`Tisu`	Err$_\downarrow$	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
Fundus+Vessel (**Ours**)	**36.46**	**52.32**	**75.49**	**92.93**	**97.44**
Fundus-only	*37.48*	**52.32**	*73.78*	*91.10*	*96.95*
Vessel-only	48.13	*33.66*	61.83	89.39	96.83

6.5.5 Ablation Study

Comparison of Inputs for Bilateral-Fuser

Table 6.5 compares the performance of Bilateral-Fuser when using different inputs. The standard input configuration uses fundus images and vessel maps as inputs for the main and satellite streams, respectively (Fundus+Vessel). These experiments achieve the best accuracy on all metrics, with the smallest mean error on both `PALM` (48.72 pixels) and `Tisu` (36.46 pixels). When using fundus images as the second input (Fundus+Fundus), the model's performance slightly degrades as feeding fundus images into the satellite stream does not provide the explicit anatomical structure for TokenLearner to learn where to focus

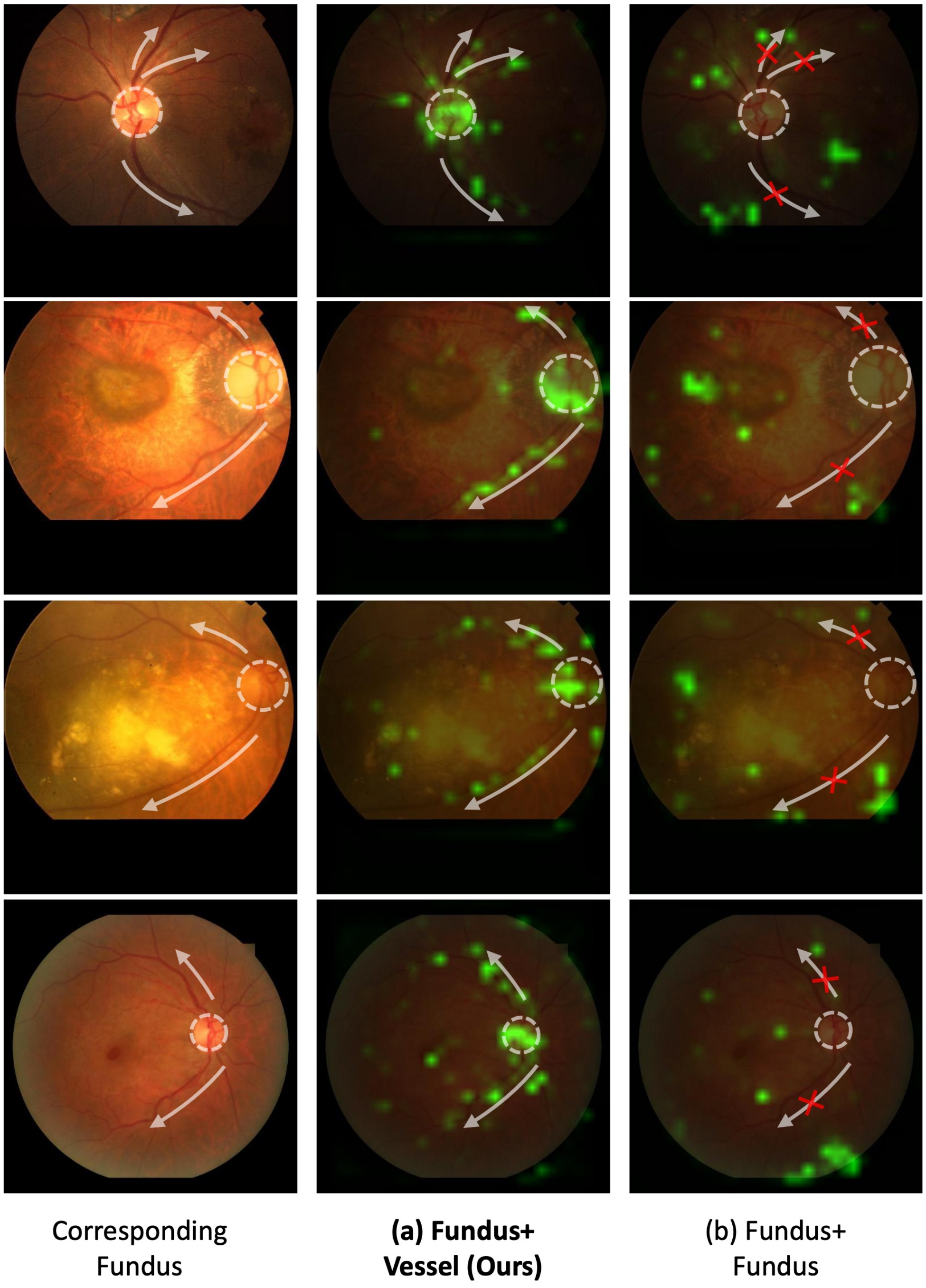


Figure 6.5: Visualizations generated from spatial attention maps, indicating the focus of TokenLearner in BTI module. These visual results have been resized and superimposed onto the the corresponding fundus.

its attention. Experiments using vessel maps as both inputs (Vessel+Vessel) lead to significant accuracy decreases on both `PALM` and `Tisu` by 23.53 and 11.67 pixels, respectively, indicating a severe loss of information.

As spatial attention maps reveal the self-learning features extracted by TokenLearners, we visually demonstrate these weight maps in Fig. 6.5. To visualize the attention of tokens in an element-wise manner, we maximize the probability values along the channels of spatial attention maps, and then normalize them. Finally, we superimpose these maps onto the corresponding fundus to compare their structural relationships.

Our experiments using Fundus+Vessel inputs show that the spatial attention maps focus on structural features along the optic disk and the direction of vessel branches (Fig. 6.5-a). This is feasible since these two structures have significant anatomical relationships with the fovea region [85, 6, 97, 127, 11]. In contrast, although more detailed information is available to the Fundus+Fundus experiments, these TokenLeaners, which are not guided by explicit anatomical structures, fail to learn features along the vessel distribution (Fig. 6.5-b). This leads to fewer tokens carrying useful features for fovea localization and may restrict the effectiveness of the intermediate BTI modules in Bilateral-Fuser, resulting in slight underperformance on `PALM` and `Tisu` datasets (Table 6.5). Therefore, the structural information provided by vessel maps as the second input is crucial for achieving accurate fovea localization.

Comparison of Methods for Reducing and Recovering Tokens

To evaluate the effectiveness of different components in reducing and recovering token numbers, we perform a comprehensive set of ablation experiments on the `PALM` and `Tisu` datasets. Instead of employing TokenLearner (TL) and TokenFuser (TF) with adaptively learnable parameters, we alternatively test more straightforward methods used in [108], average pooling (AvgPool) and bilinear interpolation (Interpolate), respectively.

In Fig. 6.6, we demonstrate a visualization of the mean error against the computational cost (FLOPs) on `PALM` (red) and `Tisu` (blue) datasets. Experiments using both TokenLearner and TokenFuser (TL+TF) achieve the best performance on `PALM` and `Tisu` with mean errors of 48.72 and 36.46 pixels, respectively. This is in contrast to using more straightforward methods such as average pooling (AvgPool) and bilinear interpolation (Interpolate), which may lead to a loss of information and reduced performance. In the proposed Bilateral-Fuser (TL+TF), total excess costs of FLOPs for the four BTI modules

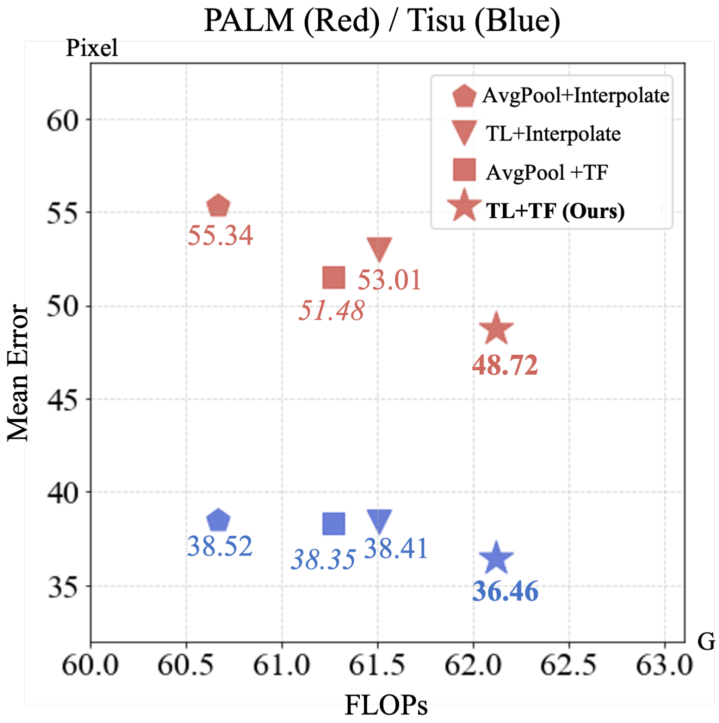


Figure 6.6: Visualization of mean errors (Y-axis) of ablation studies for reducing and recovering tokens. X-axis is the computational cost (FLOPs). The red and blue markers are results on `PALM` and `Tisu` datasets, respectively. Numbers below the markers are corresponding mean errors.

utilizing TokenLearner and TokenFuser are only **0.85G** and **0.61G**, respectively. This slight increase in computation leverages significant performance benefits, demonstrating the high efficacy of the adaptively learnable parameters of TokenLearner and TokenFuser in our architecture.

6.6 Conclusion and Future Work

Accurate detection of the macula and fovea is crucial for diagnosing retinal diseases. Although anatomical structures outside the fovea, such as the blood vessel distribution, are anatomically related to the fovea, relatively few recent deep learning approaches exploit them to improve the performance of fovea localization. In this paper, we propose a novel architecture, Bilateral-Fuser, which fuses features from the retina and corresponding vessel distribution with high efficacy for robust fovea localization. The Bilateral-Fuser contains a two-stream encoder for multi-cue fusion and a decoder for generating result maps. In addition, the Bilateral Token Incorporation (BTI) module in the encoder is designed to incorporate global anatomical features of inputs (both fundus and vessel images). Comprehensive experiments carried out demonstrate that the advantages of using Bilateral-Fuser with are more accurate localization results, insensitivity to diseased images, and low computational cost. Our proposed architecture achieves new state-of-the-art on two public datasets (`Messidor` and `PALM`) and one large-scale private dataset (`Tisu`) with metrics from $1/4\ R$ to $2R$. It also outperforms other methods on cross-dataset experiments with better generalization capacity.

Although our proposed Bilateral-Fuser outperforms all other previous studies, including Bilateral-ViT, the process of this fovea localization task is an end-to-end supervised training approach. The performance of the model remains highly dependent on the quality and quantity of labels. In this case, unlabeled fundus images are wasted. Therefore, we would attempt self-supervised learning methods to efficiently utilize these unlabeled data in future research. For example, we could design a novel contrastive learning framework with a multi-cue fusion module. The representations of numerous but unlabeled fundus images can be learned as pre-trained weights to further improve the localization performance of fovea by following downstream supervised training.

Chapter 7

Conclusions and Future Work

In this chapter, we shall present the conclusions of this Ph.D. thesis and discuss the key findings of our proposed studies for dense prediction tasks in the field of medical image analysis (Section 7.1). Additionally, we shall highlight two potential directions for future research (Section 7.2).

7.1 Conclusions

Machine learning is a rapidly developing area in artificial intelligence that enables computers to learn and make predictions using limited human labels. Many machine learning methods have been used for accurate predictions in various domains. Deep learning-based representation learning focuses on learning meaningful representations directly from raw data using deep architectures, eliminating manual feature engineering in machine learning area. In computer vision field, deep learning has made remarkable progress in high-level tasks, such as image segmentation, generation and detection.

Deep learning has also revolutionized the field of medical image analysis (MIA) by automating diagnosis and reducing the need for experts. MIA covers tasks such as regression/classification, medical image generation and segmentation. The regression/classification task predicts values or categories from medical images, while the dense prediction tasks (*i.e.*, medical image generation and segmentation) synthesize pixel-level predictions to aid in accurate medical analysis and treatment planning. While traditional machine learning algorithms in MIA rely on handcrafted features, deep learning-based representation learning methods provide an end-to-end approach without the need for manual feature engineering. However, unique challenges remain in the field of MIA, including limited labeled data, overfitting problems with limited data, and the need for interpretable results.

In this Ph.D. thesis, we propose advanced deep learning-based representation learning frameworks to address these challenges for dense prediction tasks in the field of MIA: medical image generation and segmentation. We explore a specific topic, namely chromosome straightening and fovea localization, for each direction of the dense prediction task. Overall, the contributions of this thesis include the introduction of novel frameworks and architectures that outperform existing methods, demonstrating improved training efficiency, generalization capability, and interpretability in their respective tasks. The research demonstrates state-of-the-art performance and highlights the potential applications of deep learning-based representation learning in MIA field. Specifically, for each topic, we first present a novel study and then propose its subsequent work with state-of-the-art performance at the time of publication of their respective papers. These methods are summarized as follows,

- In Chapter 3, we present a novel framework for chromosome straightening (medical image generation) using image-to-image translation. Our proposed method addresses the challenges of chromosome straightening, *i.e.*, the non-rigid nature of chromosomes and the difficulty in acquiring sufficient training image pairs. We propose a method to extract the internal backbone of curved chromosomes and increase the size of the dataset by random image augmentation. The backbone is composed of sticks with different gray values, which allows for more effective retention of augmentation information. The framework is then applied to two popular image-to-image translation architectures, namely a U-shape network and a conditional generative adversarial network, to synthesize straightened chromosomes with uninterrupted banding patterns and preserved details. Experiments conducted on a dataset of real-world chromosomes demonstrate that our proposed framework outperforms traditional geometric approaches in terms of straightening performance and the ability to generate realistic and continued chromosome details.

- In Chapter 4, we propose a novel architecture for robust chromosome straightening, named ViT-Patch GAN. In ViT-Patch GAN, a self-learned motion transformation generator and a Vision Transformer-based patch discriminator work together to improve the quality of the generated images. The experimental results demonstrate that the proposed method achieves better performance on various metrics compared to other existing methods. The proposed method addresses the limitations of our previous study (Chapter 3) and has three advantages. First, it requires a small dataset size for training, making it more efficient and cost-effective. Second, it retains more shape and banding pattern details of the corresponding source images, which is important for accurate analysis and diagnosis. Third, it has excellent generalization capacity since it can be applied to a large chromosome dataset for straightening. These advantages make the proposed method promising for robust chromosome straightening.

- In Chapter 5, we propose a novel method for robust fovea localization (medical image segmentation) based on the Vision Transformer architecture called Bilateral-ViT. Our method integrates global context and blood vessel structure information to achieve state-of-the-art performance on two public datasets, `Messidor` and `PALM`. The proposed Bilateral-ViT architecture includes a transformer-based main network branch and a vessel branch for explicitly incorporating the structure of blood vessels. The encoded features are subsequently merged with a customized Multi-scale Feature Fusion module. Our experiments demonstrate that the proposed method has significant advantages in handling diseased images and rare locations of the fovea. In cross-dataset experiments, our method significantly outperforms baselines and achieves better generalization capability than the best-reported method.

- In Chapter 6, we introduce a novel multi-cue fusion architecture, called Bilateral-Fuser, for fovea localization as a subsequent study of Chapter 5. The Bilateral-Fuser architecture incorporates anatomical-aware tokens to improve the robustness of fovea localization on both normal and diseased retina images. Our comprehensive experiments demonstrate that the Bilateral-Fuser architecture achieves state-of-the-art performance on multiple datasets, outperforming existing methods. We demonstrate the effectiveness of our approach on both normal and diseased retina images, showing that our method is more robust to variations in image quality and disease severity. Our experiments also show that the spatial attention of the self-learned tokens is focused on structural features along the optic disc and the direction of vessel branches. This suggests that our method is effective in capturing relevant anatomical landmarks for target region segmentation. In addition, our method has potential applications beyond retinal disease analysis, as this architecture can be applied to other medical imaging tasks. We believe that our method can contribute to the development of more accurate and efficient multi-cue/modal systems in the field of MIA.

7.2 Future Work

Although our proposed methods are superior to existing approaches, they still have the potential to improve the performance and generalization capability in future work. We believe that there are two possible directions, namely designing unified frameworks and utilizing unlabeled data, for the dense prediction tasks and their specific topics focused in this thesis (*i.e.*, chromosome straightening and fovea localization).

7.2.1 Designing Unified Frameworks

Recently, with the rapid development of deep learning, unified frameworks have gained attention in the CV field. The unified framework can handle multiple task types in a single architecture. For example, the previous object detection frameworks, such as Faster RCNN [117], SSD [90] and YOLO [116], while efficient, can only be utilized for a single specialized task. Therefore, some research is moving towards task unification while aiming at performance improvement. K-Net [172] is a framework that unifies semantic, instance and panoptic segmentation tasks by a set of learnable kernels. Each kernel is learned and updated to identify potential instances or semantic categories for pixels. Mask DINO [84] is designed based on DINO [168] with an additional mask prediction branch. In contrast to the original DINO only specialized for object detection, Mask DINO is also extended to semantic, instance and panoptic segmentation tasks. OneFormer [70] is a recently popular unified framework since it can be trained on universal data with a universal architecture, while achieving state-of-the-art performance on all semantic, instance and panoptic segmentation tasks. In addition, OneFormer is a multi-modal and prompt-based framework that can generate separate results for different task types with the corresponding input text prompts.

The reasons for designing a unified framework are three-fold. First, having a unified framework eliminates the requirement of developing and maintaining separate architectures for each task type. This simplifies the development process by reducing costs, such as the amount of codes, time and labor. Second, a unified framework improves data utilization efficiency by using shared model wights across task types, reducing memory and computational consumption compared to training multiple individual models. Third, since multiple task types are trained with universal data in a unified framework, the increased amount of data has the potential to benefit from shared representations and more informative features. This may cause improved performance for all tasks compared to training individual models independently.

For the specific tasks of medical image generation (*i.e.*, chromosome straightening) and medical image segmentation (*i.e.*, fovea localization), it is also worthwhile to design unified frameworks. Chromosome straightening plays an important role in the pathological study of chromosomes, development of cytogenetic maps and karyotype analysis [129, 10]. Straightened chromosomes can improve the performance of chromosome classification and abnormality detection [123]. Fovea is an important anatomical landmark of the retina. Accurate detection of the fovea region is essential for the analysis of many retinal diseases, such as diabetic maculopathy and age-related macular degeneration, to prevent irreversible vision loss [156, 6, 36, 92]. These two tasks are both intermediate steps in pathological studies and are important before the following steps in disease diagnosis. Therefore, newly designed framework may unify each task with its following tasks. Specifically, for chromosome straightening, a framework may be designed to combine chromosome straightening and chromosome type classification/abnormality detection. For fovea localization, a framework may be designed to combine the tasks of fovea region segmentation and disease identification in retina images. These unified frameworks are possible to be achieved by designing novel architectures based on Vision Transformer structure [41], as its class token containing holistic and semantic information may be used to recognize abnormalities, while other tokens containing more detailed information may be responsible for generating dense prediction results.

7.2.2 Utilizing Unlabeled Data

In contrast to the current research direction of proposing novel representation learning frameworks to effectively utilize limited labeled data, another direction may focus on extracting representations from large amounts of unlabeled data. For the specific tasks (chromosome straightening and fovea localization), it is valuable to research how to effectively utilize unlabeled data. There are three main reasons for this. First, unlabeled chromosome and retina images are abundant and readily available, while labeled images are insufficient and expensive to obtain. A framework based on self-supervised learning (SSL) can autonomously learn meaningful representations from unlabeled data without relying on explicit labels or annotations. Second, extracting representations from a large number of unlabeled images enables models to leverage more information and capture richer underlying patterns of the data. The pre-trained models can improve the efficiency of downstream supervised learning tasks. Third, training models using unlabeled images can result in better generalization performance. By learning robust representations, the models become more insensitive to variations in the input data, such as differences in illumination and anomalous conditions. The enhanced generalization may improve model performance on unseen or out-of-distribution data.

Since current topics can be considered as downstream tasks with dense predictions, a possible direction is to apply SSL-based methods to pre-train models on unlabeled images for performance improvement. Currently, popular SSL approaches can be classified into two main categories according to their mechanisms, *i.e.*, contrastive learning and masked visual modeling (MVM). Contrastive learning frameworks, such as SimCLR [26], Simsiam [27] and Barlow Twins [166], aim to learn useful representations by utilizing positive and negative pairs or only positive image pairs. They use objective functions (*e.g.*, InfoNCE or cosine similarity) to learn meaningful embedding spaces during pre-training. Data augmentation plays a crucial role in contrastive learning frameworks, improving the ability of model to generalize and capture robust representations. In contrast, MVM frameworks, such as SimMIM [161] and MAE [56], aim to predict occluded information by feeding images with random masked regions. Such models encode the context information from unmasked regions to infer the missing image content. By observing surrounding information, the model learns to understand the relationship between different visual regions and generates the missing dense context. Since contrastive learning extracts a single feature vector from the input image, while MVM extracts features responsible for generating masked regions, many studies have shown that representations extracted by MVM contain richer spatial and detailed information [98, 12, 98]. Therefore, MVM generally surpasses contrastive learning on detection and dense prediction tasks in transfer learning experiments. In this case, MVM can be utilized as a pre-trained method for these dense prediction tasks in the field of medical image analysis. By extracting more meaningful and informative representations from unlabeled chromosome and retina images, the pre-trained models can be loaded to our frameworks in this thesis to facilitate the model training process, improve the efficiency of image utilization, and achieve better robustness and generalization capability.

Bibliography

[1] Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. In *Computer Graphics Forum*, volume 38, pages 219–233. Wiley Online Library, 2019.

[2] Baidaa Al-Bander, Waleed Al-Nuaimy, Bryan M Williams, and Yalin Zheng. Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. *Biomedical Signal Processing and Control*, 2018.

[3] Fakhre Alam and Sami Ur Rahman. Challenges and solutions in multimodal medical image subregion detection and registration. *Journal of medical imaging and radiation sciences*, 50(1):24–30, 2019.

[4] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray. Chromosome aberrations in solid tumors. *Nature genetics*, 34(4):369–376, 2003.

[5] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[6] Arturo Aquino. Establishing the macular grading grid by means of fovea centre detection using anatomical-based and visual-based features. *Computers in biology and medicine*, 2014.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[8] Nitin Arora, Mamta Martolia, and Alaknanda Ashok. A comparative study of the image registration process on the multimodal medical images. *Asia-pacific Journal of Convergent Research Interchange*, 3(1):1–17, 2017.

[9] Tanvi Arora, Renu Dhir, and Manish Mahajan. An algorithm to straighten the bent human chromosomes. In *2017 Fourth International Conference on Image Information Processing (ICIIP)*, pages 1–6. IEEE, 2017.

[10] Gleb N Artemov, Vladimir N Stegniy, Maria V Sharakhova, and Igor V Sharakhov. The development of cytogenetic maps for malaria mosquitoes. *Insects*, 9(3):121, 2018.

[11] Khawaja Muhammad Asim, A Basit, and Abdul Jalil. Detection and localization of fovea in human retinal fundus images. In *2012 International Conference on Emerging Technologies*, 2012.

[12] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[13] SD Barrett and CR De Carvalho. A software tool to straighten curved chromosome images. *Chromosome Research*, 11(1):83–88, 2003.

[14] Thomas Beyer, Luc Bidaut, John Dickson, Marc Kachelriess, Fabian Kiessling, Rainer Leitgeb, Jingfei Ma, Lalith Kumar Shiyam Sundar, Benjamin Theek, and Osama Mawlawi. What scans we will read: imaging instrumentation trends in clinical oncology. *Cancer Imaging*, 20(1):1–38, 2020.

[15] Bhargav J Bhatkalkar, S Vighnesh Nayak, Sathvik V Shenoy, and R Vijaya Arjunan. Fundusposnet: A deep learning driven heatmap regression model for the joint localization of optic disc and fovea centers in color fundus images. *IEEE Access*, 9:159071–159080, 2021.

[16] Anders Bojesen and Claus H Gravholt. Klinefelter syndrome in clinical practice. *Nature Clinical Practice Urology*, 4(4):192–204, 2007.

[17] Tom Brosch, Roger Tam, and Alzheimer's Disease Neuroimaging Initiative. Manifold learning of brain mris by deep learning. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 633–640. Springer, 2013.

[18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[19] Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

[20] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5933–5942, 2019.

[21] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[22] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[23] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[24] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[25] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[27] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[28] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

[29] Niteesh K Choudhry, Robert H Fletcher, and Stephen B Soumerai. Systematic review: the relationship between clinical experience and quality of health care. *Annals of Internal medicine*, 142(4):260–273, 2005.

[30] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[31] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[32] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[33] Behdad Dashtbozorg, Jiong Zhang, Fan Huang, and Bart M ter Haar Romeny. Automatic optic disc and fovea detection in retinal images using super-elliptical convergence index filters. In *International Conference on Image Analysis and Recognition*, 2016.

[34] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 2014.

[35] K Sai Deepak and Jayanthi Sivaswamy. Automatic assessment of macular edema from color retinal images. *IEEE Transactions on medical imaging*, 31(3):766–776, 2011.

[36] Dharitri Deka, Jyoti Prakash Medhi, and SR Nirmala. Detection of macula and fovea for disease analysis in color fundus images. In *International Conference on Recent Trends in Information Systems (ReTIS)*, 2015.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[39] Neel Dey, Jo Schlemper, Seyed Sadegh Mohseni Salehi, Bo Zhou, Guido Gerig, and Michal Sofka. Contrareg: Contrastive learning of multi-modality unsupervised deformable image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–77. Springer, 2022.

[40] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2157–2167, 2020.

[41] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[42] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

[43] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[44] Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunović, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, and Xiulan Zhang. Palm: Pathologic myopia challenge, 2019.

[45] Yinghua Fu, Ge Zhang, Jiang Li, Dongyan Pan, Yongxiong Wang, and Dawei Zhang. Fovea localization by blood vessel vector in abnormal fundus images. *Pattern Recognition*, 129:108711, 2022.

[46] Xinting Gao, Stephen Lin, and Tien Yin Wong. Automatic feature learning to grade nuclear cataracts based on deep learning. *IEEE Transactions on Biomedical Engineering*, 62(11):2693–2701, 2015.

[47] R GeethaRamani and Lakshmi Balasubramanian. Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening. *Computer methods and programs in biomedicine*, 2018.

[48] Manuel E Gegundez-Arias, Diego Marin, Jose M Bravo, and Angel Suero. Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques. *Computerized Medical Imaging and Graphics*, 2013.

[49] Andrea Giachetti, Lucia Ballerini, Emanuele Trucco, and Peter J Wilson. The use of radial symmetry to localize retinal landmarks. *Computerized Medical Imaging and Graphics*, 2013.

[50] Luca Giancardo, Fabrice Meriaudeau, Thomas P Karnowski, Yaqin Li, Seema Garg, Kenneth W Tobin Jr, and Edward Chaum. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Medical image analysis*, 16(1):216–226, 2012.

[51] Fantin Girard, Conrad Kavalec, Sébastien Grenier, Houssem Ben Tahar, and Farida Cheriet. Simultaneous macula detection and optic disc boundary segmentation in retinal fundus images. In *Medical Imaging 2016: Image Processing*, 2016.

[52] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[53] Xiaoxin Guo, Han Wang, Xinfeng Lu, Xiaoying Hu, Songtian Che, and Yinan Lu. Robust fovea localization based on symmetry measure. *IEEE Journal of Biomedical and Health Informatics*, 24(8):2315–2326, 2020.

[54] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[55] Zicheng Guo and Richard W Hall. Parallel thinning with two-subiteration algorithms. *Communications of the ACM*, 32(3):359–373, 1989.

[56] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[59] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[60] Christine Hills, James H Moller, Marsha Finkelstein, Jamie Lohr, and Lisa Schimmenti. Cri du chat syndrome and congenital heart disease: a review of previously reported cases and presentation of an additional 21 cases from the pediatric cardiac care consortium. *Pediatrics*, 117(5):e924–e927, 2006.

[61] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.

[62] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[63] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[64] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[65] Yijin Huang, Zhiquan Zhong, Jin Yuan, and Xiaoying Tang. Efficient and robust optic disc detection and fovea localization using region proposal network and cascaded network. *Biomedical Signal Processing and Control*, 2020.

[66] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[67] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[68] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[69] Sahar Jahani and Seyed Kamaledin Setarehdan. Centromere and length detection in artificially straightened highly curved human chromosomes. *International journal of Biological engineering*, 2(5):56–61, 2012.

[70] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.

[71] Niels Justesen, Philip Bontrager, Julian Togelius, and Sebastian Risi. Deep learning for video game playing. *IEEE Transactions on Games*, 12(1):1–20, 2019.

[72] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

[73] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014.

[74] Anna Kostanecka, Lindsey B Close, Kosuke Izumi, Ian D Krantz, and Mary Pipan. Developmental and behavioral characteristics of individuals with pallister–killian syndrome. *American Journal of Medical Genetics Part A*, 158(12):3018–3025, 2012.

[75] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[76] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[77] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[78] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2020.

[79] Gihyun Kwon, Chihye Han, and Dae-shik Kim. Generation of 3d brain mri using auto-encoding generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 118–126. Springer, 2019.

[80] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[81] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[82] Andy Lee. Comparing deep neural networks and traditional vision algorithms in mobile robotics. *Swarthmore University*, 2015.

[83] Feihong Li, Wei Huang, Mingyuan Luo, Peng Zhang, and Yufei Zha. A new vaegan model to synthesize arterial spin labeling images from structural mri. *Displays*, 70:102079, 2021.

[84] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.

[85] Huiqi Li and Opas Chutatape. Automated feature extraction in color retinal images by a model based approach. *IEEE Transactions on biomedical engineering*, 2004.

[86] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[87] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[88] Bo Liu, Wenhao Chi, Xinran Li, Peng Li, Wenhua Liang, Haiping Liu, Wei Wang, and Jianxing He. Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect. *Journal of cancer research and clinical oncology*, 146:153–185, 2020.

[89] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019.

[90] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[91] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. IEEE, 1999.

[92] Jyoti Prakash Medhi and Samarendra Dandapat. An effective fovea detection and automatic assessment of diabetic maculopathy in color fundus images. *Computers in biology and medicine*, 2016.

[93] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.

[94] Maria Ines Meyer, Adrian Galdran, Ana Maria Mendonça, and Aurélio Campilho. A pixel-wise distance regression approach for joint retinal optical disc and fovea detection. In *MICCAI*, 2018.

[95] José M Molina-Casado, Enrique J Carmona, and Julián García-Feijoó. Fast detection of the main anatomical structures in digital retinal images based on intra-and inter-structure relational knowledge. *Computer methods and programs in biomedicine*, 2017.

[96] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-garadi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, 53(3):1655–1720, 2020.

[97] Harihar Narasimha-Iyer, Ali Can, Badrinath Roysam, V Stewart, Howard L Tanenbaum, Anna Majerovics, and Hanumant Singh. Robust detection and classification of longitudinal changes in color retinal fundus images for monitoring diabetic retinopathy. *IEEE transactions on biomedical engineering*, 2006.

[98] Duy-Kien Nguyen, Vaibhav Aggarwal, Yanghao Li, Martin R Oswald, Alexander Kirillov, Cees GM Snoek, and Xinlei Chen. R-mae: Regions meet masked autoencoders. *arXiv preprint arXiv:2306.05411*, 2023.

[99] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[100] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[101] Samiksha Pachade, Prasanna Porwal, and Manesh Kokare. A novel method to detect fovea from color fundus images. In *Computing, Communication and Signal Processing*. Springer, 2019.

[102] Esteban Pardo, José Mário T Morgado, and Norberto Malpica. Semantic segmentation of mfish images using convolutional networks. *Cytometry Part A*, 93(6):620–627, 2018.

[103] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[104] Ramesh Paudyal, Akash D Shah, Oguz Akin, Richard KG Do, Amaresha Shridhar Konar, Vaios Hatzoglou, Usman Mahmood, Nancy Lee, Richard J Wong, Suchandrima Banerjee, et al. Artificial intelligence in ct and mr imaging for oncological applications. *Cancers*, 15(9):2573, 2023.

[105] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[106] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. Comir: Contrastive multimodal image representation for registration. *Advances in neural information processing systems*, 33:18433–18444, 2020.

[107] Sergey M Plis, Devon R Hjelm, Ruslan Salakhutdinov, Elena A Allen, Henry J Bockholt, Jeffrey D Long, Hans J Johnson, Jane S Paulsen, Jessica A Turner, and Vince D Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229, 2014.

[108] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021.

[109] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106, 2020.

[110] Yulei Qin, Juan Wen, Hao Zheng, Xiaolin Huang, Jie Yang, Ning Song, Yue-Min Zhu, Lingqian Wu, and Guang-Zhong Yang. Varifocal-net: A chromosome classification approach using deep convolutional networks. *IEEE transactions on medical imaging*, 38(11):2569–2581, 2019.

[111] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al.

Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[112] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[113] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[114] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[115] Ramin Ranjbarzadeh, Annalina Caputo, Erfan Babaee Tirkolaee, Saeid Jafarzadeh Ghoushchi, and Malika Bendechache. Brain tumor segmentation of mri images: A comprehensive review on the application of artificial intelligence tools. *Computers in Biology and Medicine*, page 106405, 2022.

[116] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[117] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[118] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[119] William Rosenberg and Anna Donald. Evidence based medicine: an approach to clinical problem-solving. *Bmj*, 310(6987):1122–1126, 1995.

[120] Mehrsan Javan Roshtkhari and Seyed Kamaledin Setarehdan. A novel algorithm for straightening highly curved images of human chromosome. *Pattern recognition letters*, 29(9):1208–1217, 2008.

[121] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al.

Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[122] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34:12786–12797, 2021.

[123] AO Saidzhafarova, GN Artemov, TV Karamysheva, NB Rubtsov, and VN Stegnii. Molecular cytogenetic analysis of dna from pericentric heterochromatin of chromosome 2l of malaria mosquito anopheles beklemishevi (culicidae, diptera). *Russian journal of genetics*, 45(1):49–53, 2009.

[124] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

[125] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[126] Suman Sedai, Ruwan Tennakoon, Pallab Roy, Khoa Cao, and Rahil Garnavi. Multi-stage segmentation of the fovea in retinal fundus images using fully convolutional neural networks. In *IEEE 14th International Symposium on Biomedical Imaging (ISBI)*, 2017.

[127] S Sekhar, Waleed Al-Nuaimy, and Asoke K Nandi. Automated localisation of optic disk and fovea in retinal fundus images. In *2008 16th European Signal Processing Conference*, 2008.

[128] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 2023.

[129] Monika Sharma, Oindrila Saha, Anand Sriraman, Ramya Hebbalaguppe, Lovekesh Vig, and Shirish Karande. Crowdsourcing for chromosome segmentation and deep classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–41, 2017.

[130] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *Information Processing in*

Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24, pages 588–599. Springer, 2015.

[131] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

[132] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021.

[133] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015.

[134] Ibrahim Sobh, Loay Amin, Sherif Abdelkarim, Khaled Elmadawy, Mahmoud Saeed, Omar Abdeltawab, Mostafa Gamal, and Ahmad El Sallab. End-to-end multi-modal sensors fusion system for urban automated driving. 2018.

[135] Devaraj Somasundaram and VR Vijay Kumar. Straightening of highly curved human chromosome for cytogenetic analysis. *Measurement*, 47:880–892, 2014.

[136] Sifan Song, Tianming Bai, Yanxin Zhao, Wenbo Zhang, Chunxiao Yang, Jia Meng, Fei Ma, and Jionglong Su. A new convolutional neural network architecture for automatic segmentation of overlapping human chromosomes. *Neural Processing Letters*, pages 1–17, 2021.

[137] Sifan Song, Kang Dang, Qinji Yu, Zilong Wang, Frans Coenen, Jionglong Su, and Xiaowei Ding. Bilateral-vit for robust fovea localization. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[138] Sifan Song, Daiyun Huang, Yalun Hu, Chunxiao Yang, Jia Meng, Fei Ma, Frans Coenen, Jiaming Zhang, and Jionglong Su. A novel application of image-to-image translation: Chromosome straightening framework by learning from a single image. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–9. IEEE, 2021.

[139] Michael R Speicher and Nigel P Carter. The new cytogenetics: blurring the boundaries with molecular biology. *Nature reviews genetics*, 6(10):782–792, 2005.

[140] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 2004.

[141] VN Stegnii and VM Kabanova. Chromosome analysis of anopheles atroparvus and anopheles maculipennis (diptera, culicidae). *Zoologicheskii zhurnal*, 1978.

[142] VN Stegniĭ and MV Sharakhova. Systemic reorganization of the architechtonics of polytene chromosomes in onto-and phylogenesis of malaria mosquitoes. structural features regional of chromosomal adhesion to the nuclear membrane. *Genetika*, 27(5):828–835, 1991.

[143] FWM Stentiford and RG Mortimer. Some new heuristics for thinning binary handprinted characters for ocr. *IEEE transactions on systems, man, and cybernetics*, (1):81–84, 1983.

[144] Heung-Il Suk and Dinggang Shen. Deep learning-based feature representation for ad/mci classification. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 583–590. Springer, 2013.

[145] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.

[146] Daiki Tamada, Hiroshi Onishi, and Utaroh Motosugi. Motion artifact reduction in abdominal mr imaging using the u-net network. In *Proceedings of the ICMRM and Scientific Meeting of KSMRM, Seoul, Korea*, 2018.

[147] Jen Hong Tan, U Rajendra Acharya, Sulatha V Bhandary, Kuang Chua Chua, and Sobha Sivaprasad. Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. *Journal of Computational Science*, 20:70–79, 2017.

[148] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[149] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[150] Gijs van Tulder. elasticdeform. `https://github.com/gvtulder/elasticdeform`, 2019.

[151] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[152] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[153] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[154] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[155] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[156] JJ Weiter, GL Wing, CL Trempe, and MA Mainster. Visual acuity related to retinal distance from the fovea in macular disease. *Annals of ophthalmology*, 1984.

[157] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[158] Ning Xie, Xu Li, Kang Li, Yang Yang, and Heng Tao Shen. Statistical karyotype analysis using cnn and geometric optimization. *IEEE Access*, 7:179445–179453, 2019.

[159] Ruitao Xie, Jingxin Liu, Rui Cao, Connor S Qiu, Jiang Duan, Jon Garibaldi, and Guoping Qiu. End-to-end fovea localisation in colour fundus images with a hierarchical deep regression network. *IEEE Transactions on Medical Imaging*, 2020.

[160] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[161] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[162] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.

[163] Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in oncology*, 11:638182, 2021.

[164] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision (ECCV)*, pages 517–532. Springer, 2016.

[165] Qinji Yu, Kang Dang, Nima Tajbakhsh, Demetri Terzopoulos, and Xiaowei Ding. A location-sensitive local prototype network for few-shot medical image segmentation. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021.

[166] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[167] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[168] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[169] Jiping Zhang, Wenjing Hu, Shuyuan Li, Yaofeng Wen, Yong Bao, Hefeng Huang, Chenming Xu, and Dahong Qian. Chromosome classification and straightening based on an interleaved and multi-task network. *IEEE Journal of Biomedical and Health Informatics*, 2021.

[170] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[171] Wenbo Zhang, Sifan Song, Tianming Bai, Yanxin Zhao, Fei Ma, Jionglong Su, and Limin Yu. Chromosome classification with convolutional neural network based deep learning. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE, 2018.

[172] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.

[173] Can Zhao, Aaron Carass, Blake E Dewey, Jonghye Woo, Jiwon Oh, Peter A Calabresi, Daniel S Reich, Pascal Sati, Dzung L Pham, and Jerry L Prince. A deep learning based anti-aliasing self super-resolution algorithm for mri. In *Medical Image*

Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, pages 100–108. Springer, 2018.

[174] Shaohua Zheng, Youxing Zhu, Lin Pan, and Ting Zhou. New simplified fovea and optic disc localization method for retinal images. *Journal of Medical Imaging and Health Informatics*, 2019.

[175] Sunyi Zheng, Jingxiong Li, Zhongyi Shui, Chenglu Zhu, Yunlong Zhang, Pingyi Chen, and Lin Yang. Chrsnet: Chromosome straightening using self-attention guided networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*, pages 119–128. Springer, 2022.

[176] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[177] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.