Neurocomputing 705 (2026) 134891

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Data-centric benchmark for label noise estimation and ranking in remote sensing Binary building segmentation

Keiller Nogueira[a,1,*], Codruț-Andrei Diaconu[b,1], Dávid Kerekes[c,1], Jakob Gawlikowski[b], Cédric Léonard[b], Nassim Ait Ali Braham[b], June Moh Goo[d], Zichao Zeng[d], Zhipeng Liu[e], Pallavi Jain[f], Andrea Nascetti[c], Ronny Hänsch[g]

[a] School of Computer Science and Informatics, University of Liverpool, Liverpool, L69 7ZX, UK
[b] German Aerospace Center (DLR), Remote Sensing Technology Institute, Wessling, 82234, Germany
[c] Division of Geoinformatics, KTH Royal Institute of Technology, Stockholm, 114 28, Sweden
[d] University College London, London, WC1E 6BT, UK
[e] University of Exeter, Exeter, EX4 4QJ, UK
[f] Inria, Montpellier, 34000, France
[g] German Aerospace Center (DLR), Microwaves and Radar Institute, Wessling, 82234, Germany

ARTICLE INFO

Keywords:
Label noise
Data-centric machine learning
Land-cover classification
Building segmentation
Semantic segmentation
Remote sensing

ABSTRACT

High-quality pixel-level annotations are essential for the semantic segmentation of remote sensing imagery. However, such labels are expensive to obtain and are often affected by noise due to the labor-intensive and time-consuming nature of pixel-wise annotation, which makes it challenging for human annotators to label every pixel accurately. Annotation errors can significantly degrade the performance and robustness of modern segmentation models, motivating the need for reliable mechanisms to identify and quantify noisy training samples. This paper introduces a novel data-centric benchmark, together with a new, publicly available binary building segmentation dataset. This specific task serves as a representative and practically relevant testbed that enables controlled experimentation with different annotation perturbations. Furthermore, we also introduce two techniques for identifying, quantifying, and ranking training samples according to their level of label noise in remote sensing semantic segmentation. The proposed methods leverage complementary strategies based on model uncertainty, prediction consistency, and representation analysis, and consistently outperform established baselines across a range of experimental settings. The outcomes of this work are publicly available at https://github.com/keillernogueira/label_noise_segmentation.

1. Introduction

Semantic segmentation, the task of assigning a semantic category to each pixel in an image, is a core problem in remote sensing and underpins numerous downstream applications, including land-cover mapping [1], environmental monitoring [2], and urban analysis [3]. Recent advances in deep learning have led to substantial performance gains, particularly when models are trained on large-scale annotated datasets [4,5]. However, the effectiveness of these models is critically dependent on the quality (as well as quantity) of the training labels [6,7].

Producing highly accurate pixel-level annotations is both time-consuming and expensive, and even expert-generated annotations are inherently error-prone. As a result, many large-scale benchmark datasets are created using more scalable alternatives, such as pre-existing geographic maps, automated labeling pipelines, or crowd-sourced annotations [8,9]. While these approaches facilitate the creation of large and diverse datasets, they also introduce substantial levels of label noise due to different factors such as limited annotator expertise, annotation ambiguity, or systematic errors in automated pipelines [10,11]. Recent studies estimate that real-world datasets can exhibit noise ratios ranging from 8.0% to 38.5% [12]. This issue is particularly problematic for deep learning models, which are known to memorize noisy labels, potentially leading to overfitting and degraded generalization performance [12,13].

Towards tackling this, label noise has been extensively studied in image classification, where each sample is typically associated with a single

* Corresponding author.
Email address: keiller.nogueira@liverpool.ac.uk (K. Nogueira).

[1] Equal contribution.

https://doi.org/10.1016/j.neucom.2026.134891
Received 28 February 2026; Received in revised form 10 August 2026; Accepted 23 August 2026
Available online 24 August 2026
0925-2312/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

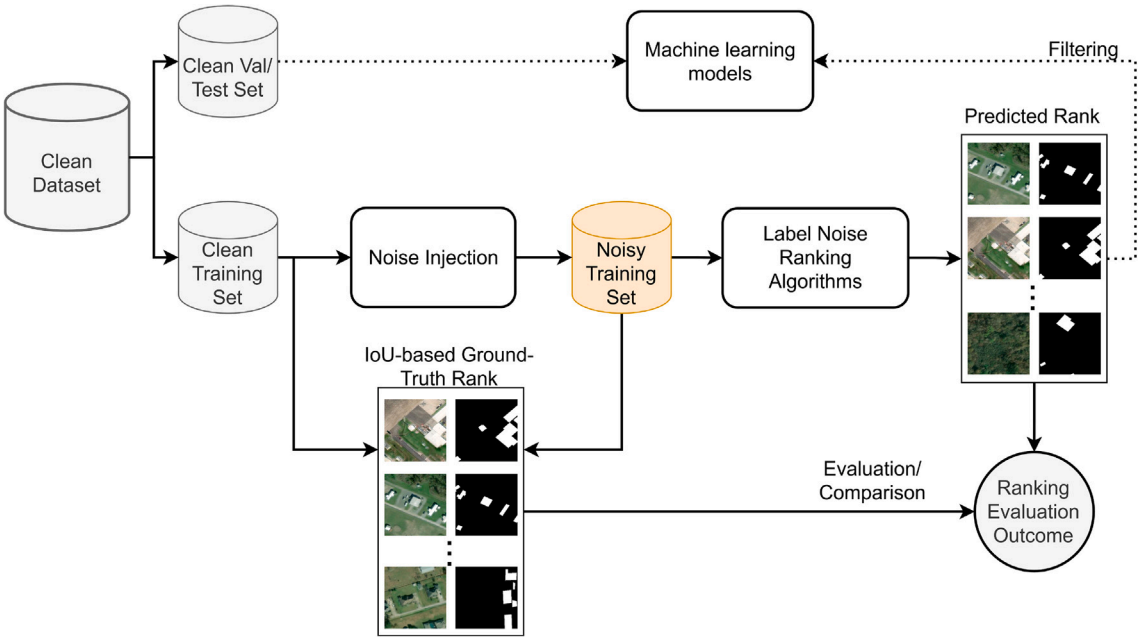


Fig. 1. Overview of the proposed benchmark. A well-curated, clean dataset is first partitioned into training and validation/test sets. A noise injection algorithm is then applied to the training set to artificially introduce label noise. By comparing the noisy and clean annotations, a ground-truth ranking of training samples is generated based on pixel-wise Intersection-over-Union (IoU). The resulting noisy training set can subsequently be processed by a noise estimation method, which assigns each image a relative label-noise score, enabling a ranking from least to most affected by annotation errors. The predicted ranking can be evaluated against the ground-truth ranking using ranking-based metrics, such as Kendall's τ [21]. Finally, the estimated ranking can also be used to prioritize or select higher-quality samples under different data selection criteria or annotation budgets, enabling more robust model training and improved generalization, which may be assessed on the original clean validation/test sets.

label, which may be correct or incorrect [14]. In contrast, semantic segmentation poses a more complex challenge: label noise is not binary but exists on a spectrum. Within a single image, some regions of a segmentation mask may be accurate, while others are erroneous, spatially misaligned, or semantically ambiguous. In this scenario, the structured and heterogeneous nature of noise makes its identification and quantification significantly more difficult, yet also more informative, as partially correct annotations may still contain valuable training signals.

Despite its importance, systematic approaches for identifying and quantifying label noise in semantic segmentation, particularly in remote sensing, remain underexplored. Existing work often focuses on modifying model architectures, loss functions, or regularization strategies to improve robustness to noisy labels [15–17]. While effective in specific settings, such methods are frequently model-dependent or rely on additional assumptions, such as access to clean validation data, limiting their general applicability. Other approaches leverage model confidence, frequently combined with specific learning strategies, such as semi- and weakly supervised learning, to down-weight or remove suspicious examples during training [18–20]. While effective, these methods are typically tightly coupled to specific models or training strategies and pipelines, complicating direct comparisons across different models and datasets.

To address these challenges, this paper proposes a new Data-Centric benchmark that reframes label noise estimation in semantic segmentation as a ranking problem. Rather than making a binary decision about whether an annotation is clean or noisy, the proposed task explicitly acknowledges the continuous nature of segmentation noise by ranking images from least to most affected by pixel-wise label errors. Such rankings, derived from model confidence and prediction-label agreement, can be used to prioritize higher-quality samples for training, data curation, or targeted manual relabeling under limited annotation budgets.

This benchmark, the central outcome of the Data-Centric Land Cover Classification Challenge organized as part of the 2024 Workshop on Machine Vision for Earth Observation and Environment Monitoring (MVEO), consists of: (i) a new, real-world, publicly available dataset comprising 5000 training samples and 1298 validation/test samples, and (ii) the two top-performing approaches from the challenge, grounded in Data-Centric and Confident Learning principles, which employ complementary strategies to estimate annotation reliability from model predictions and learned representations. An overview of the proposed benchmark can be seen in Fig. 1. Overall, by providing a unified benchmark and comparative analysis, this work aims to advance research on noise-aware learning, Confident Learning, and Data-Centric methodologies for remote sensing semantic segmentation.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the proposed dataset, Section 4 introduces the methods, Section 5 describes the experimental setup, Section 6 presents the results and discussion, and Section 7 concludes the paper.

2. Related work

Label noise has long been recognized as a major challenge in supervised learning, and a substantial body of work has focused on developing models that are robust to noisy annotations [7,12]. In the context of remote sensing image segmentation, most existing studies address label noise implicitly by modifying learning objectives, architectures, or training strategies to reduce sensitivity to annotation errors [15,22]. More recently, however, explicit identification and estimation of label noise have emerged as an important research direction, particularly within the paradigm of data-centric and confident learning [6]. Some recent works exploring this direction provide initial assessments of multiple data-centric strategies for noisy label detection, but are limited to image classification only [23]. In any case, rather than solely improving model robustness, these approaches aim to assess the reliability of training data itself, enabling data curation, prioritization, or targeted relabeling [24,25]. Overall, existing methods for handling noisy labels can be categorized into model- and data-recalibration approaches.

Model re-calibration methods focus on learning robust predictors directly from noisy data by adapting loss functions, regularization schemes, or training dynamics. Examples include noise-robust loss formulations that combine standard cross-entropy with complementary objectives to mitigate memorization of noisy labels [15,16], as well as approaches that introduce explicit regularization terms to prevent the network from overfitting to noisy labels [17]. While effective under moderate noise levels, these methods often rely on strong assumptions about the noise distribution and tend to degrade when annotation noise becomes severe or highly structured [12].

In contrast, data re-calibration methods explicitly target the training data by identifying, filtering, reweighting, or correcting unreliable annotations. Several approaches estimate label confidence using model uncertainty, prediction consistency, or ensemble disagreement, and subsequently down-weight or remove suspicious samples or regions during training [18,24,26,27]. Semi-supervised and co-training strategies have also been explored to iteratively identify trustworthy labels and suppress noisy pixels [19,20,28]. Despite their effectiveness, these approaches are often tightly coupled to specific models or training pipelines, making their behavior difficult to compare across methods and datasets.

While these approaches offer valuable insights into the behavior of noise detection methods in remote sensing settings, most of these existing methods are frequently model-dependent or rely on additional assumptions, such as access to clean validation data or weak and dense annotations, limiting their general applicability. Moreover, they typically focus on handling noisy labels within a specific model or training pipeline, making direct comparisons difficult. Consequently, to date, there are no standardized benchmarks and evaluation protocols for assessing label-noise identification methods in remote sensing semantic segmentation, particularly from a ranking-based perspective.

The present work addresses this gap by introducing a new benchmark for label noise estimation in remote sensing semantic segmentation, in which training samples are ranked according to their level of label noise. For pixel-level label noise in semantic segmentation, annotation errors exhibit different structural properties and are typically more spatially correlated, making the problem fundamentally more complex. By framing noise estimation as a continuous ranking task rather than a binary classification problem, this benchmark provides a unified evaluation framework for comparing diverse data-centric strategies in remote sensing semantic segmentation.

3. Dataset

In this section, we introduce: (i) the new dataset designed for experiments on label noise in remote sensing semantic segmentation, (ii) the noisy-label manipulations applied to generate controlled noisy training data, and (iii) an evaluation process that enables a systematic comparison of label-noise detection and ranking approaches.

3.1. Dataset

The proposed dataset builds on the high-resolution SpaceNet8 dataset [29], which comprises paired pre- and post-flooding RGB images acquired by WorldView-3 over regions in East Louisiana (USA) and Germany, with spatial resolutions ranging from 0.3 m to 0.8 m per pixel. Each image is accompanied by a pixel-level annotation mask containing four semantic classes: flooded buildings, non-flooded buildings, flooded roads, and non-flooded roads. These annotations were manually created and subjected to rigorous quality control, and can therefore be assumed to contain minimal label noise.

In this work, only the pre-flooding images and building-related classes are considered, resulting in a binary segmentation task that distinguishes buildings from the background. The original images are subdivided into patches of size 256×256 pixels. From the resulting patches, 5000 samples were randomly selected for training, while the remaining 1298 samples were reserved for validation and testing. Representative examples of the dataset and their corresponding clean annotations are shown in Fig. 2.

To enable a controlled evaluation of label-noise identification and ranking, synthetic noise is injected into the segmentation masks after patch extraction. This allows access to both clean reference labels and their noisy counterparts, which are required to quantify the accuracy of noise-level estimation. The noise generation process is described next.

3.2. Noise synthesis

To evaluate noise-level estimation, we synthetically introduced label noise into the segmentation masks of the training set. For each sample, one of the following seven noise types was randomly applied:

1. **Global shrink/expansion:** uniformly scales all building masks by a random factor in the range $[0.3, 2.5]$.

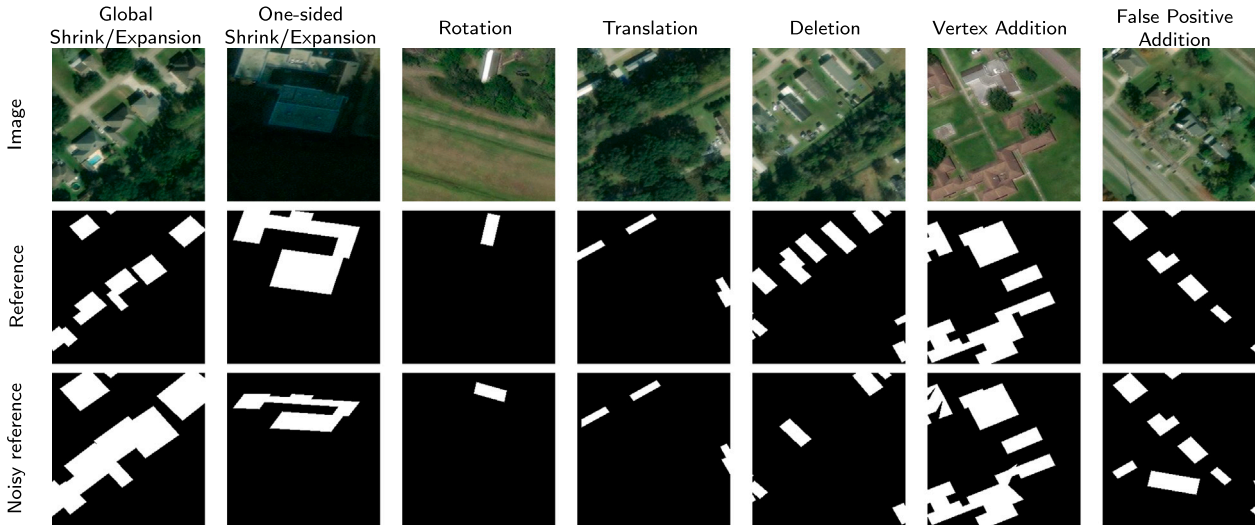


Fig. 2. Samples from the SpaceNet8 dataset [29]. The first row shows the RGB image; the second row presents the reference segmentation masks; and the third row shows the segmentation masks corrupted with synthetic noise. Each column displays a different type of noise. White pixels represent the building class whereas black pixels represent the background.

2. **One-sided shrink/expansion:** scales all building masks along a single axis (either vertical or horizontal) by a random factor in the range $[0.3, 2.5]$.
3. **Moderate rotation:** rotates the building masks by a randomly sampled angle in the range $[-90°, 90°]$.
4. **Small translation:** shifts the building masks by a random offset of -20 to 20 pixels along both the x and y axes.
5. **Deletion:** randomly removes a subset of building masks, with a deletion ratio sampled from $[0.3, 0.6]$.
6. **Vertex addition:** modifies building shapes by adding 2 to 10 vertices to the polygon contours of the buildings.
7. **False positive addition:** inserts 1 to 5 synthetic building shapes, cropped from other images, into the mask to simulate false positives.

These noise types were designed to reflect a range of realistic annotation errors, including geometric distortions, partial omissions, and spurious insertions. Fig. 2 shows representative examples of the synthesized label noise.

3.3. Evaluation process

To construct the ground-truth ranking used for evaluation, we compute the pixel-wise Intersection-over-Union (IoU) between the clean segmentation masks and their noisy counterparts for each training patch. The resulting IoU scores are then sorted to generate a reference ranking, in which samples with higher IoU (i.e., lower noise levels) are ranked above those with lower IoU. This ranking serves as the ground truth for evaluating the accuracy of label-noise estimation and ranking methods.

After generating such ground-truth rankings, the quality of the predicted rankings can be assessed using any standard ranking-based metrics, such as Kendall's τ and Spearman's rank correlation coefficient, among others [21].

4. Methodology

This section summarizes the two top-performing approaches submitted to the challenge from which the proposed benchmark emerged. Both methods address the task of estimating the relative level of label noise in remote sensing semantic segmentation data by producing a ranking of training samples according to their susceptibility to annotation errors.

Formally, given a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times C}$ denotes a remote sensing image and $y_i \in \mathbb{R}^{H\times W}$ is the corresponding (potentially noisy) building segmentation label, the objective is not to correct or filter individual pixels, but to assign each image x_i a noise-related score that reflects the overall reliability of its associated label map y_i. These resulting scores induce a ranking over the dataset, which can be directly compared to the ground-truth ranking and further exploited for data curation, sample selection, or annotation prioritization. The following subsections describe the core principles and implementation details of the two approaches.

4.1. Augmented ensemble ranking

This approach is based on the RefineNet architecture [30] and utilizes pretrained weights from a model trained on the INRIA building footprint dataset [31], as provided by [32]. Preliminary results showed that the pretrained RefineNet model performed poorly on the given task. Therefore, the model was fine-tuned on the provided dataset.

To mitigate the effect of label noise and encourage generalization, we applied strong data augmentations to each training pair (x_i, y_i). Concretely, we sampled an augmentation operator $\mathcal{A}$ and formed $(\tilde{x}_i, \tilde{y}_i) = \mathcal{A}(x_i, y_i)$, where geometric transformations (horizontal flips, 90° rotations, and small affine rotations) were applied identically to the image and the label map. For geometric warps, images were resampled using bilinear interpolation, while label maps were resampled using nearest-neighbor interpolation to preserve discrete class IDs; pixels mapped outside the valid field of view were padded with a constant value (background, i.e., $\tilde{y} = 0$). In addition, appearance transformations (brightness/contrast, CLAHE, sharpening) were applied to the image only.

Given the binary segmentation nature of the dataset, the network outputs a single-channel building probability per-pixel $p_\theta(\tilde{x}_i)(u) \in (0, 1)$, with pixel-location $u \in \Omega := \{1, .., H\} \times \{1, \dots, W\}$, and the corresponding background probability $1 - p_\theta(\tilde{x}_i)(u)$. We optimize the pixel-wise binary cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\Omega|}\sum_{u\in\Omega}\Big[\tilde{y}_i(u)\log p_\theta(\tilde{x}_i)(u) + \big(1-\tilde{y}_i(u)\big)\log\big(1-p_\theta(\tilde{x}_i)(u)\big)\Big], \tag{1}$$

where Ω is the set of pixel locations and $\tilde{y}_i(u) \in \{0, 1\}$ denotes the augmented building mask at pixel u. The model was fine-tuned for 30 epochs.

To improve model robustness, we trained an ensemble $\mathcal{E} = \{f_{\theta^{(1)}}, f_{\theta^{(2)}}, \dots, f_{\theta^{(K)}}\}$ with $K = 10$ models. Stochasticity in the training process, such as different random initializations, augmentation sequences, and data shuffling, led to diversity across ensemble members. Final predictions $\hat{y}$ are then obtained using majority voting at the pixel level:

$$\hat{y}(x) = \text{mode}\left(\left\{f_{\theta^{(k)}}(x)\right\}_{k=1}^{K}\right) \tag{2}$$

To estimate the noise level of individual labels, we measured the differences between the model's predictions and the provided annotations. Concretely, for each image, we compared the predicted building mask to the given noisy building mask using IoU. We then converted this agreement into a noise score (i.e., 1 - IoU) and sorted all images from lowest to highest estimated noisiness to form the final ranking.

4.2. Regularized variance ranking

This solution employed a pretrained ScaleMAE[33] encoder and a UperNet [34] decoder trained from scratch for the task. The decoder was first trained using ADAM [35] and a cross-entropy loss on the SpaceNet2 [36] dataset with different seeds, yielding an ensemble of $K = 8$ base networks. The provided data was split into four folds for cross-validation (CV), and two networks per CV split were fine-tuned until convergence. To reduce overfitting on potentially noisy labels, fine-tuning was continued by progressively increasing the L2 regularization coefficient of the ADAM optimizer until a peak in the F1 score on the holdout set was observed.

For each image, the IoU between the predicted labels $\hat{y}_{i,k}$ and the noisy labels y_i was computed. The best-performing prediction $\text{IoU}_i = \max_k(\text{IoU}_{i,k})$ from the ensemble was selected for each image. The images were then ranked according to their score S_i calculated by the following formula:

$$S_i = \text{IoU}_i - (0.5 - \text{IoU}_i) \times \text{avg}(\text{var}_k(\hat{y}_{i,k}))$$

Here, the variance $\text{var}_k(\hat{y}_{i,k})$ is computed pixel-wise across all eight model predictions and then averaged over all pixels in the image. This approach penalizes images with high IoU and high variance, whereas it assigns greater weight to images with high variance and low IoU. The rationale behind this is that high-variance regions in the predictions, particularly those with low IoU, are likely to contain label noise.

5. Experimental setup

5.1. Baselines

The introduced techniques are compared against two traditional baseline models: (i) CleanLab [24], which estimates label noise by comparing ground-truth annotations with out-of-sample model predictions using a softmax-based confidence metric (called softmin). To obtain

out-of-sample predictions (i.e., predictions from models that did not observe the corresponding samples during training), a 5-fold cross-validation protocol is applied to the noisy training set. In each iteration, a U-Net model [37] is trained and used to generate predictions for one held-out subset. (ii) Uncertainty Quantification [26,27], which employs meta-classifiers to quantify label uncertainty and identify pixel-wise annotation noise within training samples. As this approach also relies on out-of-sample predictions, we reuse the same cross-validated networks trained for CleanLab to ensure a fair comparison.

5.2. Experimental protocol

Two different experimental protocols were designed to evaluate the proposed approaches and baselines.

The first protocol corresponds to the original challenge setting and assesses the agreement between predicted and ground-truth rankings using ranking-based metrics, including Kendall's τ and Spearman's rank correlation coefficient [21].

The second protocol evaluates the practical impact of label noise on remote sensing semantic segmentation and examines how noise identification and filtering can mitigate its effects. To this end, a U-Net [37] and a SegFormer [38] model are trained using only the top-ranked 25%, 50%, and 75% training patches selected by each method. Validation and testing are conducted on the remaining 1298 clean samples, as described in Section 3.1. In this case, model performance is measured using the F1-score, a widely adopted metric for semantic segmentation [39].

5.3. Implementation details

To evaluate the impact of label noise and the effectiveness of noise-aware data selection, we employ two widely used semantic segmentation architectures: U-Net [37] with a ResNet-18 [40] backbone, and SegFormer [38]. The former is a fully convolutional model, while the latter is based on a transformer architecture.

All proposed methods are implemented using PyTorch [41]. During training, we use the following hyperparameters: 100 training epochs, AdamW [42] as the optimizer, a learning rate of 0.001, and a batch size of 128 for U-Net [37] and 64 for Segformer [38], reflecting GPU memory constraints.

6. Results and discussion

This section presents and discusses the results obtained by the two winning solutions of the proposed challenge. We first analyze the ranking performance using the first experimental protocol, which directly evaluates the accuracy of label-noise estimation. We then examine the practical impact of label noise on downstream semantic segmentation performance using the second experimental protocol. Finally, we analyze the impact of the distinct noise types on the methods.

6.1. Ranking results

This section presents the ranking results obtained by the introduced approaches and the tested baselines. Precisely, Table 1 summarizes the rank-based evaluation results using Kendall's τ and Spearman's correlation coefficient. In addition to the introduced approaches and the baseline methods described previously, a *Random* ranking experiment is included to provide context for the obtained results. In this setting, samples are assigned random scores between 0 and 1.0, serving as a lower-bound reference for comparison.

As expected, all evaluated methods outperform the Random baseline by a substantial margin, confirming that they are able to capture meaningful information related to label noise. More importantly, both introduced approaches significantly outperform the baseline methods, demonstrating their effectiveness in estimating and ranking the relative level of pixel-wise annotation noise in remote sensing semantic segmentation data.

Among all methods, *Augmented Ensemble Ranking* achieves the strongest overall performance, attaining a Kendall's τ of 0.61 and a Spearman's rank correlation of 0.77. This is followed by *Regularized Variance Ranking*, which achieves a Kendall's τ of 0.56 and a Spearman's rank correlation of 0.73. These results indicate that the proposed strategies produce outcomes that are well aligned with the ground-truth ranking, with particularly strong agreement in terms of rank correlation.

Table 1
Results with respect to the produced rankings.

Method	Kendall's τ	Spearman's
Random (lower bound)	0.0015	0.0021
CleanLab [24]	0.1717	0.2383
Uncertainty Quantification [26,27]	0.1262	0.1383
Augmented Ensemble Ranking	0.6104	0.7709
Regularized Variance Ranking	0.5683	0.7312

6.2. Noise effect results

This section evaluates the practical impact of label noise on remote sensing semantic segmentation and examines how noise identification and filtering can mitigate its effects. Specifically, Tables 2 and 3 present the obtained results of the U-Net [37] and SegFormer models [38] when trained using different proportions of ranked training samples. In addition to the proposed approaches and the baseline methods described earlier, three reference experiments are included to better contextualize the results: (i) *Random*, which serves as a lower-bound baseline, where training samples are randomly selected from the noisy training set without any noise-aware filtering, (ii) *Noisy Labels (oracle)*, where the ground-truth ranking is used to select training samples, providing an upper bound on performance achievable when the actual noise levels are known, and (iii) *Random Clean Labels*, which represents an idealized upper bound in which clean samples are randomly selected for training.

Several consistent observations can be drawn from both Tables. First, training models using a reduced subset of less-noisy samples (e.g., the top 50% ranked patches) consistently yields better performance than training on the full noisy dataset, a practice that is still common in the literature. This highlights the detrimental impact of label noise on

Table 2
Results (% F1-score) of the U-Net model [37] trained on subsets of varying sizes (25%, 50%, 75%, and 100%).

Method	25%	50%	75%	100%
Random (lower bound)	72.31	75.25	75.29	–
CleanLab [24]	74.51	77.39	77.78	–
Uncertainty Quantification [26,27]	75.35	77.30	77.85	–
Augmented Ensemble Ranking	79.62	80.34	79.02	–
Regularized Variance Ranking	78.93	79.56	79.09	–
Noisy Labels (oracle)	80.27	80.98	80.10	78.26
Random Clean (upper bound)	82.08	83.15	84.53	84.93

Table 3
Results (% F1-score) of the SegFormer model [38] trained on subsets of varying sizes (25%, 50%, 75%, and 100%).

Method	25%	50%	75%	100%
Random (lower bound)	73.49	74.64	75.90	–
CleanLab [24]	74.62	76.35	78.77	–
Uncertainty Quantification [26,27]	74.35	76.20	78.39	–
Augmented Ensemble Ranking	79.47	79.37	78.23	–
Regularized Variance Ranking	78.52	79.39	78.74	–
Noisy Labels (oracle)	79.81	79.66	79.97	78.32
Random Clean (upper bound)	80.20	82.50	83.12	84.72

segmentation performance and underscores the importance of selective data usage. Second, the results indicate the existence of a noise threshold beyond which incorporating additional training samples (despite increasing set size) leads to performance degradation due to the accumulation of noisy labels. This trend is observed across both architectures, suggesting that the effect is model-agnostic and primarily driven by data quality. Finally, across all training regimes and for both U-Net and SegFormer, the proposed methods consistently outperform the baseline approaches. This demonstrates their effectiveness in identifying and quantifying label noise, as well as their practical utility for guiding data selection strategies that lead to improved segmentation performance.

6.3. Noise type ablation

In this subsection, we further analyze the influence of different noise types to better understand their effects on the proposed ranking strategies. To this end, we compare the ground-truth ranking with: (i) a combined ranking obtained by averaging the rank positions of each sample assigned by the two proposed methods, (ii) the ranking generated by both baseline approaches. The comparison is then analyzed separately for each noise type introduced in the dataset.

Fig. 3 reports the results in terms of Kendall's τ. The combined ranking produced by the proposed methods consistently outperformed both baseline approaches across all considered noise types. Particularly, false-positive additions and false-negative deletions have the least impact on the combined ranking. This indicates that both proposed methods are particularly effective at identifying and quantifying these extremely common and impactful annotation errors. Among the baselines, CleanLab [24] achieved its strongest performance for false-positive additions, although it remained below the combined ranking. In contrast, its performance deteriorated substantially for deletion perturbations, yielding the lowest correlation observed among all evaluated settings. Similarly, the Uncertainty Quantification [26,27] failed to reliably identify both noise types, producing very weak correlations. These results indicate that the baselines struggle to reliably identify certain classes of annotation errors, whereas the proposed methods maintain robust performance across a wide range of perturbation scenarios.

Conversely, noise types involving annotation shrinking and expansion resulted in the lowest agreement between the ground-truth and the combined ranking. A similar trend can be observed for the baseline methods, particularly for the Uncertainty Quantification approach [26,27], which achieved very weak correlations under these settings. We attribute this behavior to the ambiguity introduced at object boundaries, where small geometric distortions can lead to large pixel-wise discrepancies while largely preserving the semantic identity of the object, making such noise patterns more challenging to detect and rank accurately.

7. Conclusions

In this paper, we introduce a novel Data-Centric benchmark that reframes label noise estimation in semantic segmentation as a ranking problem. Rather than making a binary decision about whether an annotation is clean or noisy, the proposed task explicitly acknowledges the continuous nature of segmentation noise by ranking images from least to most affected by pixel-wise label errors. This formulation enables a more nuanced and practically relevant assessment of annotation reliability in large-scale remote sensing datasets.

As part of this benchmark, we introduce a new set of noisy building segmentation labels, derived from the SpaceNet8 [29] dataset. The resulting dataset incorporates multiple realistic noise types and is made publicly available together with clean, noise-free ground-truth label masks, enabling controlled and reproducible evaluation of label-noise estimation methods. In addition, we present and analyze the two winning approaches submitted to the Data-Centric Land Cover Classification Challenge from which the proposed benchmark emerged. Both methods rely on ensembles of neural networks and exploit discrepancies between predicted and ground-truth segmentation masks as proxies for estimating annotation noise, albeit through complementary strategies.

Experimental results demonstrate the effectiveness of the proposed techniques, which consistently outperform traditional baselines across multiple evaluation protocols. A key empirical finding is that training models on a reduced subset of less-noisy samples (such as the top 50% ranked patches) often yields superior performance compared to training on the full noisy dataset, a practice that remains common in the literature. This observation highlights the detrimental impact of label noise on segmentation performance and underscores the importance of selective data usage. Moreover, it shows that prioritizing higher-quality samples can not only improve model accuracy but also substantially reduce training time and computational costs.

Overall, this work and its findings open several promising research directions, including: (i) more effective prioritization of high-quality samples for training, data curation, or targeted manual relabeling under

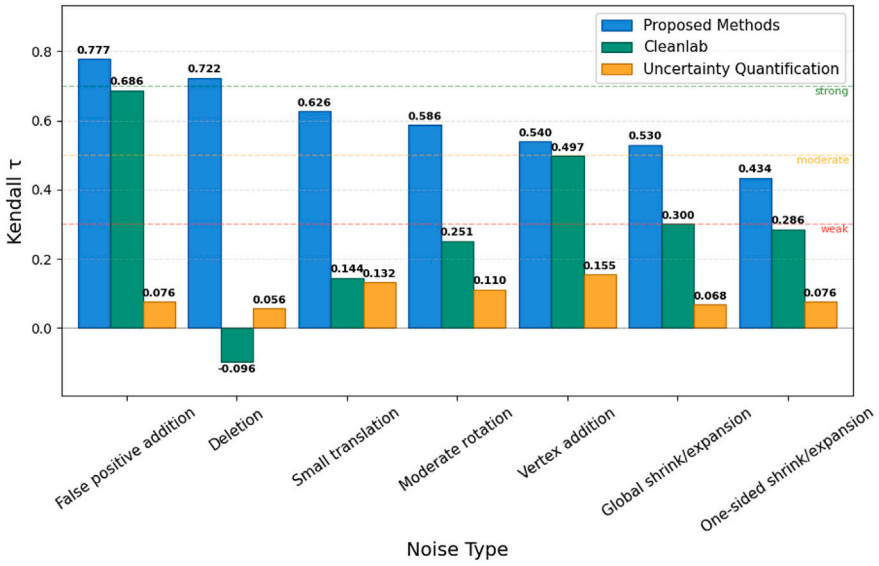


Fig. 3. Comparison, in terms of Kendall's τ, between the reference ranking, the average rank position assigned to each sample across the two proposed approaches, and the employed baselines, stratified by noise type.

limited annotation budgets, and (ii) the advancement of noise-aware learning, Confident Learning, and Data-Centric methodologies for remote sensing semantic segmentation through the provision of a unified benchmark and comparative evaluation framework.

For future work, several directions can be explored. While the synthetic perturbations considered in this work enable controlled and reproducible experiments, they do not fully capture the complexity of annotation errors encountered in real-world remote sensing datasets. In practice, label noise may exhibit systematic biases, spatial dependencies, annotator-specific patterns, or domain-dependent artifacts that are difficult to model explicitly. Nevertheless, the use of synthetic noise provides an important advantage: the exact type and location of annotation errors are known, allowing a rigorous and quantitative evaluation of label-noise detection performance. In contrast, evaluations on genuinely noisy datasets are challenging because reliable ground-truth information about annotation errors is typically unavailable, making an objective assessment of detection accuracy difficult. Therefore, the presented benchmark should primarily be interpreted as a controlled comparative framework for analyzing the behavior of different methods under well-defined perturbation scenarios. Investigating the transferability of these findings to real-world noisy datasets remains an important direction for future work. Other directions for future work include extending the benchmark to additional semantic classes and remote sensing modalities and further investigating the relationship between noise characteristics and model uncertainty in semantic segmentation tasks.

CRediT authorship contribution statement

Keiller Nogueira: Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation, Conceptualization. **Codruț-Andrei Diaconu:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Dávid Kerekes:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Jakob Gawlikowski:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Cédric Léonard:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Nassim Ait Ali Braham:** Writing – review & editing, Writing – original draft, Visualization, Methodology. **June Moh Goo:** Data curation. **Zichao Zeng:** Data curation. **Zhipeng Liu:** Data curation. **Pallavi Jain:** Data curation. **Andrea Nascetti:** Writing – review & editing, Writing – original draft. **Ronny Hänsch:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work of CAD, JG, CL, and NB was supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE (at) FZJ partition.

Data availability

Research link provided.

References

[1] Z. Fan, T. Zhan, Z. Gao, R. Li, Y. Liu, L. Zhang, Z. Jin, S. Xu, Land cover classification of resources survey remote sensing images based on segmentation model, IEEE Access 10 (2022) 56267–56281.
[2] K. Nogueira, M.M. Faita-Pinheiro, A.P.M. Ramos, W.N. Gonçalves, J.M. Junior, J.A. Dos Santos, Prototypical contrastive network for imbalanced aerial image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 8366–8376.
[3] P. Jia, C. Chen, D. Zhang, Y. Sang, L. Zhang, Semantic segmentation of deep learning remote sensing images based on band combination principle: application in urban planning and land use, Comput. Commun. 217 (2024) 97–106.
[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3523–3542.
[5] T. Zhou, W. Xia, F. Zhang, B. Chang, W. Wang, Y. Yuan, E. Konukoglu, D. Cremers, Image segmentation in foundation model era: a survey, arXiv preprint arXiv:2408.12957, 2024.
[6] R. Roscher, M. Russwurm, C. Gevaert, M. Kampffmeyer, J.A. Dos Santos, M. Vakalopoulou, R. Hänsch, S. Hansen, K. Nogueira, J. Prexl, et al., Better, not just more: data-centric machine learning for earth observation, IEEE Geosci. Remote Sens. Mag. 12 (4) (2024) 335–355.
[7] Y. Gong, G. Liu, Y. Xue, R. Li, L. Meng, A survey on dataset quality in machine learning, Inf. Softw. Technol. 162 (2023) 107268.
[8] M. Usmani, M. Napolitano, F. Bovolo, Towards global scale segmentation with OpenStreetMap and remote sensing, ISPRS Open J. Photogramm. Remote Sens. 8 (2023) 100031.
[9] X. Huang, S. Wang, D. Yang, T. Hu, M. Chen, M. Zhang, G. Zhang, F. Biljecki, T. Lu, L. Zou, et al., Crowdsourcing geospatial data for earth and human observations: a review, J. Remote Sens. 4 (2024) 0105.
[10] J. Lin, T. Yu, Z.J. Wang, Rethinking crowdsourcing annotation: partial annotation with salient labels for multilabel aerial image classification, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–12.
[11] M. Usmani, F. Bovolo, M. Napolitano, Remote sensing and deep learning to understand noisy openstreetmap, Remote Sens. 15 (18) (2023) 4639.
[12] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: a survey, IEEE Trans. Neural Netw. Learn. Syst. 34 (11) (2022) 8135–8153.
[13] M. Kimhi, O. Kerem, E. Grad, E. Rivlin, C. Baskin, Noisy annotations in semantic segmentation, arXiv preprint arXiv:2406.10891, 2024.
[14] T. Burgert, M. Ravanbakhsh, B. Demir, On the effects of different types of label noise in multi-label remote sensing image classification, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.
[15] M. Xi, J. Li, Z. He, M. Yu, F. Qin, NRN-RSSEG: a deep neural network model for combating label noise in semantic segmentation of remote sensing images, Remote Sens. 15 (1) (2022) 108.
[16] Q. Zhang, S. Pan, F. Min, Y. Wu, Z. Zhang, H. Luo, Noisy supervised deep learning for remote sensing image segmentation using electronic maps, IEEE Geosci. Remote Sens. Lett. 20 (2023) 1–5.
[17] Z. Wu, B. Hou, X. Guo, B. Ren, Z. Li, S. Wang, L. Jiao, CCNR: cross-regional context and noise regularization for SAR image segmentation, Int. J. Appl. Earth Obs. Geoinf. 121 (2023) 103363.
[18] C. Liu, C.M. Albrecht, Y. Wang, X.X. Zhu, CromSS: cross-modal pretraining with noisy labels for remote sensing image segmentation, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–17.
[19] M. Otsu, J. Zhou, B. Han, A. Accad, S. Pan, A robust co-training framework to handle noisy labels for remote sensing image segmentation, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–14.
[20] X. Xu, D. Zou, J. Zou, S. Xia, W. Sheng, MulMatch-CL: a semi-supervised teacher-student framework for robust crop segmentation in UAV imagery, IEEE Access 13 (2025) 122914–122927.
[21] S. Gupta, M. Kutlu, V. Khetan, M. Lease, Correlation, prediction and ranking of evaluation metrics in information retrieval, in: European Conference on Information Retrieval, Springer, 2019, pp. 636–651.
[22] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, A.J. Plaza, Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise, IEEE Trans. Geosci. Remote Sens. 59 (10) (2020) 8798–8811.
[23] F. Kröber, G. Hoxha, R. Roscher, An assessment of data-centric methods for label noise identification in remote sensing data sets, arXiv:2603.16835, 2026.
[24] V. Lad, J. Mueller, Estimating label quality and errors in semantic segmentation data via any model, arXiv preprint arXiv:2307.05080, 2023.
[25] C. Northcutt, L. Jiang, I. Chuang, Confident learning: estimating uncertainty in dataset labels, J. Artif. Intell. Res. 70 (2021) 1373–1411.
[26] C. Aybar, D. Montero, G. Mateo-García, L. Gómez-Chova, Lessons learned from CloudSEN12 dataset: identifying incorrect annotations in cloud semantic segmentation datasets, in: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2023, pp. 892–895.
[27] M. Rottmann, M. Reese, Automated detection of label errors in semantic segmentation datasets via deep learning and uncertainty quantification, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3214–3223.
[28] P. Li, P. Purkait, T. Ajanthan, M. Abdolshah, R. Garg, H. Husain, C. Xu, S. Gould, W. Ouyang, A. Van Den Hengel, Semi-supervised semantic segmentation under label noise via diverse learning groups, in: IEEE/CVF International Conference on Computer Vision, 2023, pp. 1229–1238.
[29] R. Hänsch, J. Arndt, D. Lunga, M. Gibb, T. Pedelose, A. Boedihardjo, D. Petrie, T.M. Bacastow, SpaceNet 8-the detection of flooded roads and buildings, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1472–1480.
[30] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: multi-path refinement networks for high-resolution semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1925–1934.
[31] E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez, Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark, in: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2017, pp. 3226–3229.
[32] F.A. Palnak, building-footprint-segmentation, https://github.com/fuzailpalnak/building-footprint-segmentation.
[33] C.J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, T. Darrell, Scale-mae: a scale-aware masked autoencoder for multiscale

geospatial representation learning, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2023, pp. 4065–4076.
[34] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 418–434.
[35] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980, 2017.
[36] A.V. Etten, D. Lindenbaum, T.M. Bacastow, SpaceNet: a remote sensing dataset and challenge series, arXiv:1807.01232, 2019.
[37] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[38] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, in: International Conference on Neural Information Processing Systems, 2021.
[39] K.R. Hasan, A.B. Tuli, M.A.-M. Khan, S.-H. Kee, M.A. Samad, A.-A. Nahid, Deep-learning-based semantic segmentation for remote sensing: a bibliometric literature review, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17 (2023) 1390–1418.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[42] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101, 2017.

Author biography

Keiller Nogueira received the BSc degree in Computer Science from Universidade Federal de Viçosa (Brazil) in 2012. He received the MSc and PhD degrees in Computer Science from the Universidade Federal de Minas Gerais, Brazil, in 2015 and 2019, respectively. Since 2024, he has been a lecturer in the School of Computer Science and Informatics at the University of Liverpool, UK. Before that, from 2019 to 2024, he was a lecturer in Data Science at the University of Stirling, UK. He has published several high-quality articles in leading journals and conferences. His research interests include Deep and Machine Learning, Pattern Recognition, Image Processing, Computer Vision, and Remote Sensing.

Codruț-Andrei Diaconu received the B.Sc. degree in computer engineering from the Gheorghe Asachi Technical University of Iași, Romania, and the M.Sc. degree in computing science (Data Science and Systems Complexity) from the University of Groningen, The Netherlands. He is currently pursuing a Ph.D. degree at the German Aerospace Center (DLR) in collaboration with the Technical University of Munich (TUM), Germany. His research focuses on deep learning for Earth observation and cryosphere remote sensing, with a particular interest in multi-modal glacier monitoring.

Dávid Kerekes is a PhD student in Geoinformatics at KTH Royal Institute of Technology, focusing on the adoption of deep learning models for Earth Observation data analysis. He holds a Master's degree in Artificial Intelligence from KU Leuven and has actively participated in several international geospatial AI challenges, including SpaceNet7, xView3, and Embed2Scale. Through these experiences, he has developed expertise in best practices for applying deep learning to large-scale EO datasets.

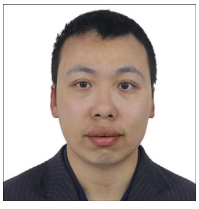
Jakob Gawlikowski received a B.Sc. and an M.Sc. degree in Mathematics from the Technical University of Munich (TUM), Germany in 2016 and 2019, respectively. He obtained his PhD at the German Aerospace Center (DLR) in collaboration with TUM. Since 2025, he has been with the DLR Remote Sensing Technology Institute, where he currently leads the Trustworthy AI4EO team. His research focuses on the trustworthiness of machine learning approaches, with particular emphasis on data quality, uncertainty quantification (UQ), explainable AI (xAI), and multi-modal learning setups.

Cédric Léonard received an engineering degree in electronics from the National Institute for Applied Sciences (INSA) Rennes, France, and an M.Sc. degree in computer science (information technology) from Åbo Akademi University, Finland. He is currently pursuing a PhD at the German Aerospace Center (DLR) in collaboration with the Technical University of Munich (TUM), Germany. His research activities focus on Deep Learning for Earth Observation and Synthetic Aperture Radar (SAR), with particular interest in efficient AI approaches for onboard and edge computing, particularly FPGA-based implementations.

Nassim Ait Ali Braham received his M.Sc. degree in computer science from Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria, in 2019 and an M.Sc. degree in Artificial Intelligence and Data Science from Université Paris Dauphine-PSL, Paris, France, in 2020. He is pursuing a Ph.D. degree at the German Aerospace Center, Wessling, Germany, and at the Technical University of Munich, Munich, Germany. In 2019, he spent six months as a research intern at the LIRIS-CNRS laboratory, Lyon, France. In 2020, he spent six months at the LAMSADE-CNRS laboratory, PSL Research University, Paris, France. In 2023, he spent six months as a visiting researcher at INRIA THOTH, Grenoble, France. His research interests include deep learning, computer vision, self-supervised learning, and remote sensing.

June Moh Goo received the BEng degree in Engineering Mathematics from the University of Bristol, United Kingdom, in 2019, and the M.Sc. degree in Data Science from the University of Bath, United Kingdom, in 2020. He also undertook postgraduate studies in Machine Learning at University College London (UCL), UK. He is currently working toward the Ph.D. degree in Civil, Environmental, and Geomatic Engineering at University College London, UK. He is an active member of the Photogrammetry and 3D Imaging Group and is supported by the Engineering and Physical Sciences Research Council through an industrial CASE studentship with Ordnance Survey. His research interests include LiDAR super-resolution, domain adaptation for 3D point clouds, and label-efficient learning for 3D data.

Zichao Zeng is currently pursuing a PhD degree at University College London (UCL), UK. His research focuses on improving interactive pedestrian navigation using Computer Vision techniques, with particular emphasis on vision-based geolocalization. His doctoral research is funded by UK Research and Innovation (UKRI) through the Engineering and Physical Sciences Research Council (EPSRC), in collaboration with Ordnance Survey (OS), and supported by the UCL International Scholar Doctoral Training Award (ISAD). His research interests include Computer Vision, Visual Localization, Multimodal Deep Learning, and Robot Vision.



Zhipeng Liu received the B.Eng. degree in software engineering from Harbin Institute of Technology, Weihai, China, in 2020, and the M.Sc. degree in advanced computer science from the University of Exeter, Exeter, U.K., in 2022, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, remote sensing, and disaster monitoring.

Pallavi Jain received the M.Sc. and Ph.D. degrees in Computer Science from Technological University Dublin, Ireland, in 2018 and 2023, respectively. Since 2023, she has been a postdoctoral researcher at UMR TETIS, in collaboration with INRIA. Her research interests include computer vision, self-supervised learning, cross-modal learning, and interpretable machine learning, with a particular emphasis on remote sensing applications.

Andrea Nascetti is an Associate Professor at KTH Royal Institute of Technology, working at the intersection of computer vision and remote sensing developing ML and DL methods using multi-modal satellite time series for biomass estimation, 3D mapping, and wildfire monitoring. He received his M.Sc. and Ph.D. degrees in Geodesy and Geomatics from Sapienza University of Rome. His recent research investigates geospatial foundation model capabilities through the development of large-scale benchmarks such as PANGAEA-Bench. He also led the international data challenge (i.e., BioMassters), establishing a widely used benchmark for forest biomass estimation.

Ronny Hänsch received his Diploma and Ph.D. in computer science from TU Berlin in 2007 and 2014. He is a senior scientist at the DLR Microwave and Radar Institute, leading the Machine Learning Team in the SAR Technology Department. He has served as chair of the GRSS IADF Technical Committee, editor of the GRSS eNewsletter, and co-chair of the ISPRS Working Group on Image Orientation and Sensor Fusion. He is editor-in-chief of IEEE GRSL and an associate editor of the ISPRS Journal. He has led multiple remote sensing competitions, including the IEEE Data Fusion Contests and the SpaceNet 8 and 9 Challenges.