Learning to Segment with Uncertainty: Mask-Guided Feature Refinement and Multi-Path Fusion for High-Resolution Remote Sensing Images

Jinpeng Wang,Xinhao Wu,Yuru Wang

Abstract

Semantic segmentation of high resolution remote sensing imagery plays a crucial role in land use monitoring and urban planning. Although deep learning-based methods have achieved significant progress and can obtain satisfactory segmentation results, they suffer from limitations in robustness and reliability, and tend to produce overconfident predictions that often overlook the inherent uncertainty in the segmentation process. This poses challenges for their deployment in safety-critical applications such as land detection and disaster assessment. Moreover, despite the widespread adoption of encoder-decoder architectures, existing methods still struggle to fully exploit the high-dimensional features extracted by encoders. To address these issues, this study proposes an Uncertainty Mask Feature Refinement Module and a Multi-Path Feature Interaction structure. Specifically, the proposed method computes total uncertainty through Monte Carlo sampling and learns a binary mask by applying learnable perturbations to the original data, thereby identifying regions that contribute significantly to output uncertainty. This approach quantifies the inherent uncertainty in the data and refines feature representations in high-uncertainty regions. After the encoder encodes the refined features, each decoder layer not only aggregates features from the corresponding encoder layer but also incorporates features from deeper encoding levels, achieving fine-grained segmentation results through multi-path feature interaction. We conducted comprehensive experiments on two well-known high-resolution remote sensing benchmark datasets (ISPRS Vaihingen and ISPRS Potsdam). The results demonstrate that the proposed method outperforms state-of-the-art approaches in segmentation accuracy while explicitly constructing uncertainty estimates, providing new insights for developing more robust and reliable remote sensing image segmentation models.

1. Introduction

Remote sensing technology, as a core means of acquiring information about the Earth's surface, plays an irreplaceable role in numerous fields such as urban planning, land resource management, environmental monitoring, and national defense security [1, 10]. With the development of satellite and aerial platforms, the spatial resolution of high-resolution (HR) remote sensing imagery has reached the sub-meter or even centimeter level, which has greatly propelled the demand for pixel-level, fine-grained feature identification and scene understanding [2, 3, 34, 35]. Remote Sensing Image Semantic Segmentation (RSISS), as an important branch of the computer vision field, aims to assign a predefined semantic label (e.g., building, road, vegetation, etc.) to each pixel in an image, serving as a key and fundamental task for achieving intelligent interpretation of remote sensing imagery [2, 3].

In recent years, benefiting from the immense success of Convolutional Neural Networks (CNNs) in natural image semantic segmentation tasks, the field of RSISS has also achieved breakthrough progress [1, 2, 10]. Early works such as Fully Convolutional Networks (FCN) [4], U-Net [5], SegNet[11], and the DeepLab series [14, 6] laid the foundation for end-to-end pixel-level classification. Driven by the powerful feature representation capabilities of CNNs (e.g., ResNet [15]), researchers have adapted and improved these fundamental architectures significantly to address the characteristics of remote sensing imagery, for example, U-Net-based ResUNet-a [16], Pan-Net [36], and Xu-Net [37]; SegNet-based improvements [38, 39]; and various applications of DeepLab in remote sensing [31, 32, 33]. Concurrently, to tackle the properties of remote sensing images, techniques such as multi-scale fusion (e.g., PSPNet [12], RefineNet [13]), atrous convolution [6, 14], and attention mechanisms [3] have been introduced. Attention mechanisms, in particular, such as the Dual Attention Network (DANet)[7], SENet [45], CBAM [46], and specialized attention modules for remote sensing [17, 18, 19, 47, 48], have significantly enhanced the model's ability to discriminate complex features [2]. However,

directly applying these advanced models to HR-RS images still presents numerous challenges.

The semantic segmentation of HR-RS images faces unique challenges, primarily reflected in several aspects: First, the multi-scale variance and small object problem. There is an enormous disparity in the size of ground objects, ranging from large-scale parcels to minuscule vehicles or trees. Target features are easily lost during the continuous down-sampling process, and the precise localization of small objects is extremely demanding [2, 3, 28, 29, 30]. Second, spectral complexity of ground objects, with severe phenomena of "intra-class variation" (same object, different spectra) and "inter-class similarity" (different objects, same spectrum). This makes it difficult to effectively distinguish feature categories based on spectral features alone; consequently, multi-modal data fusion (e.g., DSM or LiDAR) [43, 44] has become an important research direction. Furthermore, blurred boundaries and imprecise localization are a core difficulty [3, 9]. Because models lose spatial details during feature extraction in the encoder stage, the segmentation maps restored by the decoder are often not sharp or precise at object boundaries. To address this, researchers have explored techniques like Conditional Random Fields (CRFs) post-processing [40, 41] and edge-aware loss functions [42, 49], but these still struggle to meet the strict requirements for fine-grained boundaries in HR-RS applications. Although deep learning models[27] have surpassed traditional methods in segmentation performance, balancing accuracy and computational efficiency, especially when handling boundary regions and complex multi-scale objects, remains an urgent problem to be solved [1,2, 9].

Beyond the structural and data-specific challenges mentioned above, model "trustworthiness," encompassing reliability and explainability, is becoming an increasingly severe challenge, especially in high-stakes remote sensing applications such as disaster response and security monitoring[53]. Standard deep learning models tend to produce "overconfident" predictions [53, 52] and cannot effectively quantify their own predictive uncertainty. This has spurred research into Uncertainty Quantification (UQ) for semantic segmentation[54]. Existing methods include Bayesian deep learning (e.g., Bayesian SegNet[50]), which models network weights [55], as well as simpler, more scalable methods like Deep Ensembles [51]. Understanding the types of uncertainty—whether aleatoric (stemming from data noise) or epistemic (stemming from model knowledge gaps) [55]—is crucial for building robust systems. Furthermore, as models become increasingly complex, Explainable AI (XAI) methods are becoming essential for understanding the basis of model decisions. Techniques like gradient attribution (e.g., Grad-CAM [56]) or perturbation-based analysis[57] are used to generate saliency maps. However, how to attribute uncertainty itself (Uncertainty Attribution) [58, 59], and how to design uncertainty-aware loss functions [53] or uncertainty-guided learning frameworks [54], remain active areas of research. This highlights the gap between high-performance models and deployable, trustworthy systems.

To address the aforementioned challenges, we propose a model based on uncertainty attribution. It constructs an uncertainty mask, semantically aligning the uncertainty score with the mask, and then applies perturbation and progressive refinement specifically to the identified high-uncertainty regions. Furthermore, we optimize the single-path U-Net network, transforming it into a multi-path architecture. This new design transmits the results from each encoder layer to corresponding higher levels in the decoder, thereby reducing the semantic gap between high-level and low-level features.

To summarize, our main contributions are as follows:

•We propose a novel uncertainty attribution model. This model constructs an uncertainty mask, semantically aligning uncertainty scores with the mask to locate high-uncertainty regions. We further design a perturbation and progressive refinement mechanism specifically targeting these identified regions, significantly improving model robustness on ambiguous and difficult samples.

•To address the semantic gap between high-level semantics and low-level details in HR-RS images, we optimize the U-Net [5] architecture. We modify the traditional single-path U-Net into a multi-path architecture and design new skip connections that transmit results from each encoder layer to corresponding higher levels in the decoder. This design effectively reduces the semantic gap and promotes superior feature fusion.

•We train and test our model on the benchmark datasets ISPRS Vaihingen and ISPRS Potsdam. Our method surpasses existing state-of-the-art (SOTA) approaches on key metrics such as F1-Score, while maintaining high inference efficiency.

2. Related Work

2.1 Uncertainty Attribution

In the context of model trustworthiness, Uncertainty Attribution (UA) [60, 61, 62, 63] has emerged as a crucial research direction that focuses on identifying and localizing input features or regions that make significant contributions to a model's predictive uncertainty. This differs from conventional explainable AI (XAI) methods, which primarily interpret the predictions themselves rather than the sources of uncertainty. Existing UA approaches can be broadly categorized into two families. Gradient-based methods [64, 65, 62] backpropagate the output uncertainty to the input to generate attribution maps; they are computationally efficient but often overly

sensitive to small input perturbations, which leads to noisy or imprecise explanations [62]. Perturbation-based methods [66, 67] systematically modify (e.g., occlude or mask) parts of the input and observe the resulting change in uncertainty to assess feature importance, yet classical instantiations (such as [66]) typically require exponentially many perturbation samples, making them extremely expensive[68]. To overcome these challenges, recent studies have shifted toward more advanced strategies. For example, CLUE [60] and its variants [69, 70] use generative models to alter uncertain inputs, producing " improved " images with minimal uncertainty and attributing by contrasting the differences. However, a key limitation is their reliance on pretrained generative models, which are often difficult to obtain in specialized domains such as remote sensing. Another promising avenue is to formulate UA as an optimization problem [71, 72, 73], in which continuous optimization is used to efficiently search for the smallest or most informative perturbation regions that maximally reduce uncertainty—that is, to learn a mask —closely aligning with the methodology proposed in this work.

2.2 U-Net

U-Net [5] is one of the most influential foundational architectures in semantic segmentation, laying the groundwork for end-to-end p ixel-level classification alongside FCN [4] and SegNet [74]. Its core design is a symmetric encoder – decoder structure containing a contracting path that captures contextual semantics and an expanding path that enables precise localization. U-Net's key innovation lies in its skip connections, which concatenate high-resolution, detail-rich feature maps from shallow encoder layers with upsampled, semantics-rich feature maps from deeper decoder layers. This design allows the network to exploit both high-level semantics and low-level details simultaneously, delivering outstanding performance on segmentation tasks that demand sharp boundaries — especially in biomedical imagery and high-resolution remote sensing (HR-RS). Consequently, U-Net has become a standard architecture for remote-sensing segmentation and has inspired numerous variants tailored to remote-sensing characteristics. However, the classic skip-connection scheme introduces a central challenge: the semantic gap caused by the substantial semantic disparity between shallow encoder features and deep decoder features. As UNet++ points out, directly fusing such semantically dissimilar feature maps complicates optimization and can reduce segmentation accuracy. To mitigate this issue, researchers have explored multiple improvements, such as the nested, dense skip pathways proposed by UNet++, and hybrid architectures like UNetFormer that incorporate Transformer-based decoders to better fuse global and local contextual information.

3. Methodology

Our proposed methodology addresses the significant challenges of high-resolution remote sensing (HR-RS) image segmentation — namely, high intra-class variance, low inter-class separability, and ambiguity at object boundaries — by jointly optimizing both the input data representation and the network architecture itself. We present a two-stage framework, illustrated in Fig. X, designed to operate in sequence.

First, we introduce an Uncertainty-Guided Data Refinement module. This pre-processing stage actively identifies and enhances the most ambiguous regions of an image (e.g., blurred boundaries, shadows) before they are fed into the main network. Second, we propose the FTUNetformer++, a novel multi-path U-Net architecture. This network is specifically designed to process the refined data, utilizing dense, cross-scale feature interaction to effectively bridge the "semantic gap" between shallow, high-resolution details and deep, semantic-rich features.

3.1 Uncertainty-Guided Data Refinement

Deep segmentation models often exhibit overconfidence, particularly in HR-RS imagery where data-inherent uncertainty is high. While traditional Uncertainty Attribution (UA) methods exist, they are often computationally prohibitive (requiring many inferences) or produce noisy, pixel-level attributions that lack semantic context. To overcome this, we reformulate UA as an efficient optimization problem. Our goal is to actively learn a semantically-aware mask that identifies input regions which, when subjected to a specific perturbation, maximally reduce the model's overall predictive uncertainty.

3.1.1 Uncertainty Mask Alignment

Our method is founded on a key hypothesis: applying a smooth perturbation (e.g., blurring) to regions where the model is genuinely uncertain (ambiguous boundaries, low-contrast objects) will help the model resolve ambiguity and thus decrease its predictive entropy. Conversely, perturbing high-confidence, well-defined regions will introduce noise and increase entropy.

We aim to find an optimal binary mask M that minimizes the model's predictive uncertainty U when applied as a perturbation selector. This uncertainty U is quantified as the Average Predictive Entropy, estimated from N stochastic forward passes (e.g., using Deep Ensembles or MC-Dropout) on the perturbed input. The optimization objective is defined as:

$$M^* = \underset{M}{\operatorname{argmin}}\, U\left((1-M)\odot x + M \odot \operatorname{Blur}(x,\phi)\right) + \lambda \|M\|_1$$

Here, Blur is not a fixed operation but a Gaussian blur function with a learnable standard deviation φ, serving as our targeted perturbation source. The term U represents the model's predictive uncertainty. The L_1 norm M_1 is a regularization term, controlled by λ, which encourages the mask Mt o be sparse, focusing only on the most critical uncertainty-inducing regions.

To make this optimization tractable and ensure the mask is semantically meaningful, we do not optimize M in the pixel-space. Instead, we parameterize the mask M as a weighted combination of K semantic region proposals M_k, which are obtained from a pre-trained segmentation model: M = \sum_{k=1}^K w_k M_k. This aligns the attribution process with human-understandable concepts.

To optimize the binary weights w_k \in \{0, 1\}, we model them as samples from a Bernoulli distribution, w_k \sim Ber. We then employ the Gumbel-Sigmoid reparameterization trick, which provides a continuous and differentiable approximation, enabling us to learn the optimal mask parameters p_k efficiently via standard gradient descent.

3.1.2Uncertainty-Guided Refinement

After the alignment process converges, the resulting mask M^ highlights the regions that confuse the model the most. We then use this mask as a guide to refine the original image x, creating a new input x_{\text{refined}} that is inherently easier for the network to segment.

First, we use a Difference-of-Gaussians (DoG) operation, a computationally efficient and effective method, to extract informative multi-scale boundary and texture features: text{DoG}(x) = x - G_\φ(x). This isolates the high-frequency details.Next, we modulate these boundary features using the learned uncertainty mask M and a learnable scaling factor α. This modulated signal is then added back to the original image:

$$x_{\text{refined}} = x + \alpha \cdot M^* \odot (x - G_\phi(x))$$

This operation functions as a targeted data augmentation. It selectively amplifies the boundary details only in the model's most ambiguous regions, while leaving high-confidence areas (where $M^*=0$) unchanged. This pre-processing step effectively reduces data-level uncertainty and prepares the input for more robust feature extraction by the subsequent segmentation network.

3.2 Multi-Path U-Net Architecture (FTUNetformer++)

To fully leverage the refined input data, we must address the primary architectural challenge in segmentation: the "semantic gap." Standard U-Net encoders produce shallow features with precise spatial detail but weak semantics, and deep features with rich semantics but poor localization. The simple, one-to-one skip connections in a standard U-Net are often insufficient to optimally fuse these heterogeneous representations.

To solve this, we designed FTUNetformer++, a deep-supervision-based codec architecture that replaces simple skip connections with a dynamic, multi-path feature interaction mechanism.

3.2.1Encoder

The encoder employs a Swin Transformer as its backbone. Unlike CNNs, the Swin Transformer's windowed self-attention mechanism excels at capturing long-range dependencies and global contextual information across multiple scales. This is critical for HR-RS segmentation, where objects (e.g., large buildings, roads) can span large portions of the image, requiring a global understanding of context.

3.2.2Multi-Path Decoder

Our core architectural innovation lies in the decoder. Instead of a single skip connection per level, we design a dense, cross-scale fusion strategy where each decoder stage integrates information from all deeper decoder layers and its corresponding encoder layer.

As shown in Fig. X, the decoder feature D_i at layer i is computed by concatenating its corresponding encoder feature E_i with the upsampled outputs of all deeper decoder layers D_j (where j > i). This aggregated feature set is then processed by a sophisticated fusion module φ:

$$D_i = \phi\left(\operatorname{Concat}\left(E_i, \{\operatorname{Up}(D_j)\}_{j>i}\right)\right)$$

where φ is a fusion module (containing 1x1 convolution, a Weighted-Sum (WS) module, and a GLTB block), \text{Up}(\cdot) is the upsampling operation, and \text{Concat} denotes feature concatenation.

This multi-path architecture allows high-level semantic information (like class concepts) to fully interact with shallow spatial details (like textures and boundaries) at an early stage of decoding. The decoder blocks themselves use Global-Local Transformer Blocks (GLTBs) to simultaneously model local structures and global context.

3.2.3Joint Supervision

To ensure all parts of this deep, multi-path network are optimized effectively, we employ a joint supervision strategy. In addition to the main loss L_{\text{main}} computed at the final segmentation output, we introduce an auxiliary branch part-way through the decoder (e.g., between the middle and bottom GLTBs).

This branch, managed by an Auxiliary Head (AH), produces an intermediate, lower-resolution segmentation map, and a corresponding L_{\text{aux}} is computed. The final loss for the network is a weighted sum:

$$L_{\text{total}} = L_{\text{main}} + \beta \cdot L_{\text{aux}}$$

This deep supervision forces the intermediate layers to learn discriminative and semantically meaningful features, providing a richer gradient signal that stabilizes training and encourages the collaborative optimization of both shallow and deep features throughout the network.

4. Experimental Results And Analysis

4.1 Datasets

ISPRS Vaihingen: This dataset consists of 33 aerial images captured over Vaihingen, Germany, with an average size of 2494 × 2064 pixels and a ground sampling distance (GSD) of 9 cm. Each image contains three spectral bands: near-infrared (NIR), red (R), and green (G). Fine-grained annotations are provided for six classes: impervious surface (ImSurf), building, low vegetation (LowVeg), tree, car, and clutter/background. Following the standard split, we use 16 images for training and the remaining 17 for testing.

ISPRS Potsdam: This dataset comprises 38 aerial images collected over Potsdam, Germany, each measuring 6000 × 6000 pixels with a GSD of 5 cm, offering higher resolution than Vaihingen. The imagery covers four bands: red (R), green (G), blue (B), and near-infrared (NIR). The annotation categories match those of Vaihingen. We train on 24 images and reserve 14 for testing.

4.2 Evaluation Metrics

In our evaluation, we report Overall Accuracy (OA) to measure global pixel-level classification accuracy, while Mean F1 Score (mF1) and Mean Intersection over Union (mIoU) jointly characterize performance in high-resolution remote sensing segmentation by assessing the precision–recall balance and the overlap between predicted and ground-truth regions.

4.3 Quantitative Analysis

Our model is implemented using the PyTorch framework, and all experiments are conducted on an NVIDIA RTX 4090 GPU. During training, we crop input images into 512 × 512 patches and apply standard data augmentation, including random flipping, rotation, and scale jittering.

When both proposed components are combined, the results in Tables 1 and 2 show that FTUNetFormer++&Refine—our full model—achieves the highest performance on both datasets (85.36% mIoU on Vaihingen and 87.66% mIoU on Potsdam). These gains validate the effectiveness of the proposed uncertainty-guided data refinement mechanism, which strengthens the boundary features of high-uncertainty regions before training and enables better handling of ambiguous samples. Our FTUNetFormer++&Refine therefore delivers the best overall performance.

Table 1. Performance Comparison(The second data group)

	Params	Per-Class IOU(%)					mF1	OA	mIOU
		ImSurf	Building	LowVeg	Tree	Car			
FTUNetFormer	96M	94.31	**93.31**	**74.64**	77.25	82.33	91.31	93.53	84.37
Wide FTUNetFormer	105M	94.48	92.70	74.27	82.30	81.58	91.75	93.72	85.07
FTUNetFormer++	103M	94.43	92.49	74.32	82.29	82.48	91.84	93.68	85.21
FTUNetFormer&Refine	96M	94.32	92.79	74.28	82.36	81.01	91.71	93.73	84.95
Wide FTUNetFormer&Refine	105M	94.46	92.61	74.31	82.41	81.77	91.78	93.74	85.11
FTUNetFormer++&Refine	103M	**94.53**	92.81	74.51	**82.52**	**82.51**	**91.92**	**93.81**	**85.36**

Table 2. Performance Comparison.

	Params	Per-Class IOU(%)					mF1	OA	mIOU
		ImSurf	Building	LowVeg	Tree	Car			
FTUNetFormer	96M	89.14	94.22	78.81	80.58	92.37	93.07	91.63	87.02
Wide FTUNetFormer	105M	89.35	94.16	78.52	80.86	93.03	93.03	91.80	87.19
FTUNetFormer++	103M	89.51	93.96	79.06	80.91	**93.78**	93.18	91.95	87.44
FTUNetFormer&Refine	96M	89.27	93.59	79.24	80.94	93.06	93.06	91.87	87.22
Wide FTUNetFormer&Refine	105M	89.36	94.32	79.00	80.91	93.32	93.14	91.94	87.38
FTUNetFormer++&Refine	103M	**90.16**	**94.31**	**79.50**	**81.11**	93.22	**93.30**	**92.30**	**87.66**

Notably, our FTUNetFormer++&Refine (103M parameters) outperforms the wider FTUNetFormer (105M parameters) on both datasets while using fewer parameters. This indicates that our architectural improvements are more effective than merely widening the baseline model.

4.4 Ablation Studies

Our approach comprises two key innovations: (A) the optimization-based "uncertainty mask alignment" and "data refinement" mechanism in Section 3.1, governed by the hyperparameter α; and (B) the "multi-level U-Net" (ML-UNet) architecture in Section 3.2 designed to narrow the semantic gap. We conduct ablation studies to validate each component individually.

Across all trials, mIoU remains high, indicating that the data refinement module is robust to the choice of α. On Vaihingen, a smaller α = 0.3 yields the best mIoU of 85.21%, whereas on Potsdam the stronger refinement α =

0.9 performs best (87.39%), with $\alpha = 0.5$ (87.29%) staying highly competitive. This suggests that the optimal refinement strength can vary across datasets.

4.5 Qualitative Visualization

To illustrate the advantages of our approach, we present visual results for challenging scenes from the Vaihingen and Potsdam datasets in Figures 3 and 4.

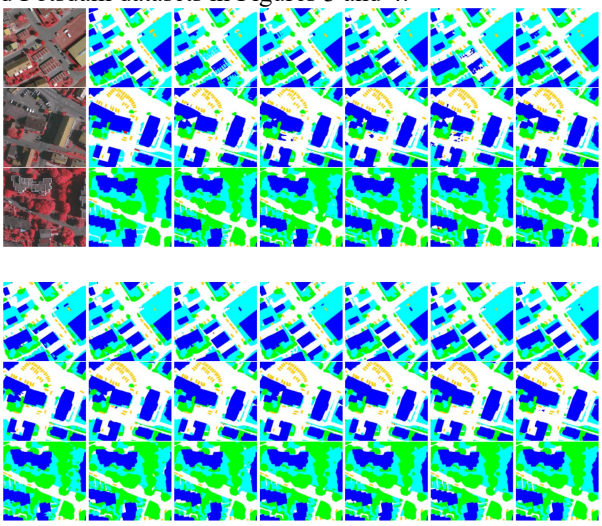

Figure 3. Potsdam Visualization Results

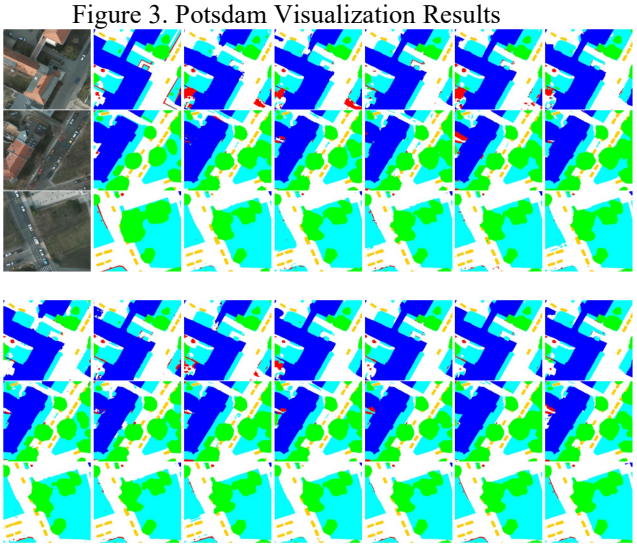

Figure 4. Vaihingen Visualization Results

As shown in Fig. 3, the baseline FTUNetFormer exhibits noticeable "sticking" and "fragmentation" near building edges and small objects such as cars. For example, it misclassifies vehicle shadows as background or yields incomplete building boundaries. In contrast, our method markedly alleviates these issues. Thanks to the uncertainty attribution in Section 3.1, the model strengthens features around these "difficult boundaries" before training, enabling sharper and more complete building contours at inference time. Meanwhile, the multi-level U-Net in Section 3.2 preserves richer low-level details, allowing precise localization and segmentation of small vehicles, closely matching the ground truth.

In Fig. 4, the baseline struggles to distinguish between low vegetation (LowVeg) and impervious surfaces (ImSurf), which share similar spectral characteristics. By proactively identifying and refining these high-uncertainty regions during training, our model demonstrates greater robustness, reduces class confusion, and produces semantically more consistent segmentation maps.

5. Conclusion

We address blurred boundaries, multi-scale fusion challenges, and overconfident predictions in high-resolution remote sensing segmentation with a framework centered on uncertainty attribution. By casting uncertainty mask alignment as an optimization task, we learn pixel-level perturbation masks and parameters that pinpoint and refine the model's most confounding regions, while a multi-level U-Net with dense multi-path skip connections mitigates the semantic gap. Experiments demonstrate state-of-the-art mIoU on ISPRS Vaihingen and Potsdam, and visualizations further highlight gains on fuzzy borders, small objects, and easily confused classes. Looking ahead, we plan to embed the uncertainty-guided refinement mechanism into end-to-end training and extend the framework to multimodal remote sensing.

References

[1] FirstName Alpher, Frobnication. IEEE TPAMI, 12(1):234–778, 2002.

[2] FirstName Alpher and FirstName Fotheringham-Smythe. Frobnication revisited. Journal of Foo, 13(1):234–778, 2003.

[3] Ma, Y. and Gulimila, K., "A review of image semantic segmentation methods in high-resolution remote sensing image interpretation", Journal of Computer Science and Exploration, vol. 17, no. 7, pp. 1526-1548, 2023.

[4] Long J , Shelhamer E , Darrell T .Fully Convolutional Networks for Semantic Segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.DOI:10.1109/CVPR.2015.7298965.

[5] Ronneberger O , Fischer P , Brox T .U-Net: Convolutional Networks for Biomedical Image Segmentation[J].Springer, Cham, 2015.DOI:10.1007/978-3-662-54345-0_3.

[6] Chen L C , Zhu Y , Papandreou G ,et al.Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[J].Springer, Cham, 2018.DOI:10.1007/978-3-030-01234-2_49.

[7] Fu J , Liu J , Tian H ,et al.Dual Attention Network for Scene Segmentation[J].IEEE, 2020.DOI:10.1109/CVPR.2019.00326.

[8] Liu Z , Lin Y , Cao Y ,et al.Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[J]. 2021.DOI:10.48550/arXiv.2103.14030.
[9] Ez-Zahouani B , El Kharki O ,Kanga Idé, S,et al.Determination of Segmentation Parameters for Object-Based Remote Sensing Image Analysis from Conventional to Recent Approaches: A Review[J].International Journal of Geoinformatics, 2023.DOI:10.52939/ijg.v19i1.2497.
[10] Zhu X X , Tuia D , Mou L ,et al.Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources[J].IEEE Geoscience & Remote Sensing Magazine, 2018, 5(4):8-36.DOI:10.1109/MGRS.2017.2762307.
[11] Badrinarayanan V , Kendall A , Cipolla R .SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017:1-1.DOI:10.1109/TPAMI.2016.2644615.
[12] Zhao H , Shi J , Qi X ,et al.Pyramid Scene Parsing Network[J].IEEE Computer Society, 2016.DOI:10.1109/CVPR.2017.660.
[13] Lin G , Milan A , Shen C ,et al.RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation[J].IEEE, 2017.DOI:10.1109/CVPR.2017.549.
[14] Chen L C , Papandreou G , Kokkinos I ,et al.Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J].Computer Science, 2014(4):357-361.DOI:10.1080/17476938708814211.
[15] He K , Zhang X , Ren S ,et al.Deep Residual Learning for Image Recognition[J].IEEE, 2016.DOI:10.1109/CVPR.2016.90.
[16] Diakogiannis F I , Waldner F , Caccetta P ,et al.ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data[J].Elsevier, 2020.DOI:10.1016/J.ISPRSJPRS.2020.01.013.
[17] Panboonyuen T , Jitkajornwanich K , Lawawirojwong S ,et al.Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning[J].Remote Sensing[2025-11-13].DOI:10.20944/preprints201812.0090.v3.
[18] L. Mou, Y. Hua and X. X. Zhu, "Relation Matters: Relational Context-Aware Fully Convolutional Network for Semantic Segmentation of High-Resolution Aerial Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 11, pp. 7557-7569, Nov. 2020, doi: 10.1109/TGRS.2020.2979552.
[19] Li R , Zheng S , Zhang C ,et al.Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images[J].IEEE Transactions on Geoscience and Remoter Sensing, 2022, 60(000):13.DOI:10.1109/TGRS.2021.3093977.
[20] Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. Remote Sens. 2022, 14, 1956. https://doi.org/10.3390/rs14091956.
[21] Zhang C , Jiang W , Zhang Y ,et al.Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery[J].IEEE Transactions on Geoscience and Remote Sensing, 2022, PP:1-1.DOI:10.1109/tgrs.2022.3144894.
[22] Chen J , Lu Y , Yu Q ,et al.TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation[J]. 2021.DOI:10.48550/arXiv.2102.04306.
[23] Cao H , Wang Y , Chen J ,et al.Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation[J]. 2021.DOI:10.48550/arXiv.2105.05537.
[24] Xie E , Wang W , Yu Z ,et al.SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers[J]. 2021.DOI:10.48550/arXiv.2105.15203.
[25] Dosovitskiy A , Beyer L , Kolesnikov A ,et al.An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. 2020.DOI:10.48550/arXiv.2010.11929.
[26] Zheng S , Lu J , Zhao H ,et al.Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers[J]. 2020.DOI:10.48550/arXiv.2012.15840.
[27] Wang J , Sun K , Cheng T ,et al.Deep High-Resolution Representation Learning for Visual Recognition[J].Institute of Electrical and Electronics Engineers (IEEE), 2021(10).DOI:10.1109/TPAMI.2020.2983686.
[28] Chen X .Cascaded Multi-scale Structure with Self-smoothing Atrous Convolution for Semantic Segmentation[J].IEEE Transactions on Geoscience and Remote Sensing, 2021.DOI:10.1109/TGRS.2021.3088902.
[29] B J Z A , B D Z A , B B S A ,et al.Multi-source collaborative enhanced for remote sensing images semantic segmentation[J].Neurocomputing, 2022.DOI:10.1016/j.neucom.2022.04.045.
[30] Luo, Y.; Wang, J.; Yang, X.; Yu, Z.; Tan, Z. Pixel Representation Augmented through Cross-Attention for High-Resolution Remote Sensing Imagery Segmentation. Remote Sens. 2022, 14, 5415. https://doi.org/10.3390/rs14215415.
[31] He H , Yang D , Wang S ,et al.Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss[J].Remote Sensing, 2019, 11(9):1015.DOI:10.3390/rs11091015.
[32] Zhan Z , Zhang X , Liu Y ,et al.Vegetation Land Use/Land Cover Extraction From High-Resolution Satellite Images Based on Adaptive Context Inference[J].IEEE Access, 2020, 8:21036-21051.DOI:10.1109/ACCESS.2020.2969812.
[33] Venugopal N .Automatic Semantic Segmentation with DeepLab Dilated Learning Network for Change Detection in Remote Sensing Images[J].Neural Processing Letters, 2020.DOI:10.1007/s11063-019-10174-x.
[34] Kattenborn T , Eichel J , Fassnacht F E .Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery[J].Scientific Reports, 2019(1).DOI:10.1038/S41598-019-53797-9.
[35] Kussul N , Lavreniuk M , Skakun S ,et al.Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data[J].IEEE Geoscience and Remote Sensing Letters, 2017, PP(99):1-5.DOI:10.1109/LGRS.2017.2681128.
[36] Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building Extraction from High-Resolution Aerial

Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms. Remote Sens. 2019, 11, 917. https://doi.org/10.3390/rs11080917.
[37] Xu, Y., Feng, M. R., Pi, J. T., et al., "Remote sensing image segmentation method based on deep learning model", Journal of Computer Applications, vol. 39, no. 10, pp.2905-2914,2019.doi:10.11772/j.issn.1001-9081.2019030529.
[38] Yang, J. Y., Zhou, Z. X., Du, Z. R., et al., "Extraction of rural construction land from high-resolution remote sensing images based on SegNet semantic model", Transactions of the Chinese Society of Agricultural Engineering, vol. 35,no.5,pp.259-266,2019.doi:10.11975/j.issn.1002-6819.2019.05.031.
[39] Zhang, C. S., Ge, Y. W., and Jiang, X., "Building extraction from high-resolution remote sensing images based on sparse constraint SegNet", Journal of Xi'an University of Science and Technology, vol. 40, no. 3, pp. 441-448, 2020. doi:10.3969/j.issn.1672-9315.2020.03.010.
[40] Li Z , Wang R , Zhang W ,et al.Multiscale Features Supported DeepLabV3+ Optimization Scheme for Accurate Water Semantic Segmentation[J].IEEE Access, 2019:1-1.DOI:10.1109/ACCESS.2019.2949635.
[41] Zhu Q , Li Z , Zhang Y ,et al.Building Extraction from High Spatial Resolution Remote Sensing Images via Multiscale-Aware and Segmentation-Prior Conditional Random Fields[J].Remote Sensing, 2020, 12(23):3983.DOI:10.3390/rs12233983.
[42] A X Z , A L H , B G S X A ,et al.Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss - ScienceDirect[J].ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 170:15-28.DOI:10.1016/j.isprsjprs.2020.09.019.
[43] Peng C , Li Y , Jiao L ,et al.Densely Based Multi-Scale and Multi-Modal Fully Convolutional Networks for High-Resolution Remote-Sensing Image Semantic Segmentation[J].IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(8):2612-2626.DOI:10.1109/JSTARS.2019.2906387.
[44] Zhang M , Hu X , Zhao L ,et al.Translation-aware Semantic Segmentation via Conditional Least Square Generative Adversarial Networks[J].Journal of Applied Remote Sensing, 2017, 11(4).DOI:10.1117/1.JRS.11.042622.
[45] Hu J , Shen L , Sun G ,et al.Squeeze-and-Excitation Networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).DOI:10.1109/TPAMI.2019.2913372.
[46] Woo S , Park J , Lee J Y ,et al.CBAM: Convolutional Block Attention Module[J].Springer, Cham, 2018.DOI:10.1007/978-3-030-01234-2_1.
[47] Zhao D , Wang C , Gao Y ,et al.Semantic Segmentation of Remote Sensing Image Based on Regional Self-Attention Mechanism[J].IEEE geoscience and remote sensing letters, 2022(19-).DOI:10.1109/LGRS.2021.3071624.
[48] Li C ,Xin_Li, Xia R ,et al.Hierarchical Self-Attention Embedded Neural Network With Dense Connection for Remote-Sensing Image Semantic Segmentation[J].IEEE Access, 2021.DOI:10.1109/ACCESS.2021.3111899.
[49] Marmanis D , Schindler K , Wegner J D ,et al.Classification With an Edge: Improving Semantic Image Segmentation with Boundary Detection[J].Isprs Journal of Photogrammetry & Remote Sensing, 2017, 135.DOI:10.1016/j.isprsjprs.2017.11.009.
[50] Kendall A , Badrinarayanan V , Cipolla R .Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding[J].Computer Science, 2015.DOI:10.48550/arXiv.1511.02680.
[51] Lakshminarayanan B , Pritzel A , Blundell C .Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles[J]. 2016.DOI:10.48550/arXiv.1612.01474.
[52] Singh R, Principe J C. Quantifying model uncertainty for semantic segmentation using operators in the RKHS[EB/OL]. arXiv e-prints, 2022. doi:10.48550/arXiv.2211.01999.
[53] Landgraf S , Hillemann M , Wursthorn K ,et al.Uncertainty-aware Cross-Entropy for Semantic Segmentation[J].ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, 2024, 10(2).DOI:10.5194/isprs-annals-X-2-2024-129-2024.
[54] Wang Z , Li Y , Guo Y ,et al.Data-Uncertainty Guided Multi-Phase Learning for Semi-Supervised Object Detection[J]. 2021.DOI:10.48550/arXiv.2103.16368.
[55] Kendall A , Gal Y .What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?[J]. 2017.DOI:10.48550/arXiv.1703.04977.
[56] Selvaraju R R , Cogswell M , Das A ,et al.Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization[J].International Journal of Computer Vision, 2020, 128(2):336-359.DOI:10.1007/s11263-019-01228-7.
[57] Fong R , Patrick M , Vedaldi A .Understanding Deep Networks via Extremal Perturbations and Smooth Masks[J].IEEE, 2019.DOI:10.1109/ICCV.2019.00304.
[58] Wang H , Joshi D , Wang S ,et al.Gradient-based Uncertainty Attribution for Explainable Bayesian Deep Learning[J].ArXiv, 2023, abs/2304.04824.DOI:10.48550/arXiv.2304.04824.
[59] Wang H , Biswas B A , Ji Q .Optimization-Based Uncertainty Attribution Via Learning Informative Perturbations[C]//European Conference on Computer Vision.Springer, Cham, 2025.DOI:10.1007/978-3-031-73229-4_14.
[60] Javier Antorán, Bhatt U , Adel T ,et al.Getting a CLUE: A Method for Explaining Uncertainty Estimates[J]. 2020.DOI:10.48550/arXiv.2006.06848.
[61] Perez I , Skalski P , Barns-Graham A E ,et al.Attribution of Predictive Uncertainties in Classification Models (Supplementary material)[J]. 2022.
[62] Wang, H., Joshi, D., Wang, S., Ji, Q.: Gradient-based uncertainty attribution for explainable bayesian deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12044-12053 (2023).
[63] Wang H , Joshi D , Wang S ,et al.Gradient-based Uncertainty Attribution for Explainable Bayesian Deep Learning[J].ArXiv, 2023, abs/2304.04824.DOI:10.48550/arXiv.2304.04824.
[64] Smilkov D , Thorat N , Kim B ,et al.SmoothGrad: removing noise by adding noise[J]. 2017.DOI:10.48550/arXiv.1706.03825.

[65] Sundararajan M , Taly A , Yan Q .Axiomatic Attribution for Deep Networks[J]. 2017.DOI:10.48550/arXiv.1703.01365.
[66] Petsiuk V , Das A , Saenko K .RISE: Randomized Input Sampling for Explanation of Black-box Models[J]. 2018.DOI:10.48550/arXiv.1806.07421.
[67] Ribeiro M T , Singh S , Guestrin C ."Why Should I Trust You?": Explaining the Predictions of Any Classifier[J].ACM, 2016.DOI:10.1145/2939672.2939778.
[68] Watson D , O'Hara J , Tax N ,et al.Explaining Predictive Uncertainty with Information Theoretic Shapley Values[J].ArXiv, 2023, abs/2306.05724.DOI:10.48550/arXiv.2306.05724.
[69] Ley D , Bhatt U , Weller A .\\delta$-CLUE: Diverse Sets of Explanations for Uncertainty Estimates[J]. 2021.DOI:10.48550/arXiv.2104.06323.
[70] Ley D , Bhatt U , Weller A .Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates[J]. 2021.DOI:10.48550/arXiv.2112.02646.
[71] Dabkowski P , Gal Y .Real Time Image Saliency for Black Box Classifiers[J]. 2017.DOI:10.48550/arXiv.1705.07857.
[72] Fong R , Patrick M , Vedaldi A .Understanding Deep Networks via Extremal Perturbations and Smooth Masks[J].IEEE, 2019.DOI:10.1109/ICCV.2019.00304.
[73] Yang Q , Zhu X , Fwu J K ,et al.MFPP: Morphological Fragmental Perturbation Pyramid for Black-Box Model Explanations[J].2021, 000(9411940).DOI:10.1109/ICPR48806.2021.9413046.
[74] Badrinarayanan V , Kendall A , Cipolla R .SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017:1-1.DOI:10.1109/TPAMI.2016.2644615.