BiCoR-Seg: Bidirectional Co-Refinement Framework for High-Resolution Remote Sensing Image Segmentation

Jinghao Shi
China University of Geosciences (Wuhan)
Wuhan 430070, Hubei, China
shijinghao@cug.edu.cn

Jianing Song
China University of Geosciences (Wuhan)
Wuhan 430070, Hubei, China
songjn@cug.edu.cn

Abstract

High-resolution remote sensing image semantic segmentation (HRSS) is a fundamental yet critical task in the field of Earth observation. However, it has long faced the challenges of high inter-class similarity and large intra-class variability. Existing approaches often struggle to effectively inject abstract yet strongly discriminative semantic knowledge into pixel-level feature learning, leading to blurred boundaries and class confusion in complex scenes. To address these challenges, we propose Bidirectional Co-Refinement Framework for HRSS (BiCoR-Seg). Specifically, we design a Heatmap-driven Bidirectional Information Synergy Module (HBIS), which establishes a bidirectional information flow between feature maps and class embeddings by generating class-level heatmaps. Based on HBIS, we further introduce a hierarchical supervision strategy, where the interpretable heatmaps generated by each HBIS module are directly utilized as low-resolution segmentation predictions for supervision, thereby enhancing the discriminative capacity of shallow features. In addition, to further improve the discriminability of the embedding representations, we propose a cross-layer class embedding Fisher Discriminative Loss to enforce intra-class compactness and enlarge inter-class separability. Extensive experiments on the LoveDA, Vaihingen, and Potsdam datasets demonstrate that BiCoR-Seg achieves outstanding segmentation performance while offering stronger interpretability. The released code is available at https://github.com/ShiJinghao566/BiCoR-Seg.

Keywords: ***Remote Sensing, Image segmentation, Class embedding, Heatmap-driven, Bidirectional information synergy***

1. Introduction

High-resolution remote sensing image semantic segmentation (HRSS), as a fundamental task in the field of Earth observation [59], holds significant application value in various real-world scenarios such as urban planning, land-use monitoring, ecological environment assessment, and disaster emergency response [28, 29]. Its objective is to assign precise semantic labels to each pixel, thereby enabling the automatic identification of ground object categories and the fine depiction of spatial structures.

However, due to the diversity of ground object categories, significant variations in scale, complex textures, and drastic illumination changes in remote sensing images [22, 28], HRSS has long faced dual challenges of large intra-class variation and high inter-class similarity [36]. These characteristics make it difficult for models to learn feature representations that are highly discriminative and generalizable [42], leading to common problems in complex scenes, such as blurred boundaries, category confusion, and fragmented regions.

Researchers have conducted extensive explorations. Early convolutional neural network (CNN)-based methods [23, 27, 32] effectively extracted multi-scale local features through encoder–decoder architectures. However, due to the inherently limited receptive field of convolution operations, these methods struggled to capture global contextual information, resulting in insufficient capability to distinguish land-cover categories that depend on large-scale spatial layouts, such as roads and buildings. Subsequent works introduced dilated convolutions [48], attention mechanisms [1], and Transformer architectures [37] to enhance global perception by modeling long-range dependencies. Although these approaches improved feature representation capacity, they still relied on implicit semantic learning. Discriminative, category-level knowledge (e.g., "buildings" should exhibit rigid boundaries, "water bodies" should have smooth surfaces) remained buried within massive network parameters. Without an explicit mechanism to guide the alignment between pixel-level features and high-level semantics, models were prone to misclassification in regions with ambiguous features [21, 41].

Recent studies have attempted to incorporate class prototype mechanisms [30, 50] to explicitly model seman-

tic knowledge, enabling interactions between features and class representations within the semantic space. These methods have, to some extent, enhanced category separability and semantic consistency. However, such interactions are often unidirectional, as class embeddings influence feature aggregation during the decoding phase but cannot receive feedback from the pixel level for further refinement. Moreover, existing class embeddings are typically static and lack adaptability to specific image content, while their semantic spaces often lack explicit constraints. Therefore, establishing a tight and interpretable bidirectional optimization mechanism between semantic embeddings and visual features has become a key challenge for further improving HRSS performance.

To address the aforementioned issues, we propose the Bidirectional Co-Refinement Framework for HRSS (BiCoR-Seg). We posit that precise segmentation relies on establishing an explicit interaction mechanism between pixel-level features and category semantics during the forward propagation process of the network. Therefore, BiCoR-Seg abandons the traditional one-way decoding paradigm and introduces a heatmap-based bidirectional information synergy module (HBIS). In each interaction, HBIS first computes the similarity between the category embeddings and the feature map to generate class heatmaps. Subsequently, HBIS leverages these heatmaps to guide the category embeddings in perceiving the feature distributions, and updates their representations by selecting top-K high-response pixels. The enhanced category semantics then feed back into the feature maps by generating category-specific affine modulation parameters to adjust feature distributions. Furthermore, we design a dual supervision mechanism that employs a hierarchical heatmap loss to supervise the spatial localization accuracy and applies a Fisher discriminative loss to enforce intra-class compactness and inter-class separability of the category embedding space. This iterative interaction mechanism progressively enhances the discriminative power of features, enabling the network to simultaneously capture global semantic consistency and local boundary details, thereby significantly improving segmentation accuracy and generalization in complex scenes.

Overall, the contributions of this paper are as follows:

- We propose the BiCoR-Seg framework, which achieves explicit semantic interaction and layer-wise collaborative refinement between feature and class embeddings, providing a novel and interpretable modeling paradigm for remote sensing image segmentation.
- We design the HBIS module, which enables bidirectional information interaction driven by heatmaps to realize mutual guidance and semantic co-construction between features and class embeddings, thereby establishing an explicit and interpretable collaborative mechanism between the class embeddings and feature maps.
- We introduce hierarchical heatmap deep supervision and a Fisher discriminant loss to enhance inter-layer semantic consistency and explicitly constrain the spatial structural distribution of class embeddings, significantly improving the discriminability of class embeddings and the separability of features.
- Our BiCoR-Seg achieves highly competitive performance across three mainstream datasets, fully demonstrating the effectiveness and superiority of the proposed framework in complex scenarios.

2. Related Work

2.1. Deep Learning Based Remote Sensing Image Segmentation

The introduction of deep learning has significantly advanced the development of semantic segmentation in remote sensing imagery. Early methods were primarily based on convolutional neural networks (CNNs). Representative of this stage, Fully Convolutional Networks (FCNs) [27] achieved pixel-level end-to-end prediction for the first time. Subsequent models such as U-Net [32] and its variants, including UNet++ [57] and Attention U-Net [31], effectively integrated high-level semantics with low-level spatial details through encoder–decoder architectures and skip connections. To further enhance multi-scale perception, PSPNet [54] and the DeepLab series [5–7] introduced pyramid pooling and atrous (dilated) convolution structures. ConDSeg [24] improved boundary recognition through a contrastive-driven feature enhancement mechanism, while HRNet [35] maintained high-resolution feature representations via parallel multi-resolution branches, demonstrating outstanding performance in fine-grained object segmentation.

To overcome the limitation of convolution's local receptive field in global context modeling, researchers incorporated attention mechanisms [1] and Transformer [37] architectures into segmentation tasks. Models such as SETR [55] and Segmenter [34] demonstrated the feasibility of pure Transformer-based designs; meanwhile, Swin Transformer [25] and SegFormer [45] achieved a balance between performance and efficiency through hierarchical structures and lightweight decoders. In the remote sensing domain, methods like AerialFormer [17] and UNetFormer [40] combine the strengths of convolutional networks and Transformers, achieving a balance between local detail preservation and global semantic consistency. Hybrid architectures such as TransUNet [4], LeViT-UNet [46], and Swin-UNet [2] further strengthen cross-level feature interaction, exhibiting superior performance in multi-scale object recognition and complex scene segmentation. Collectively, these studies constitute the technological foundation

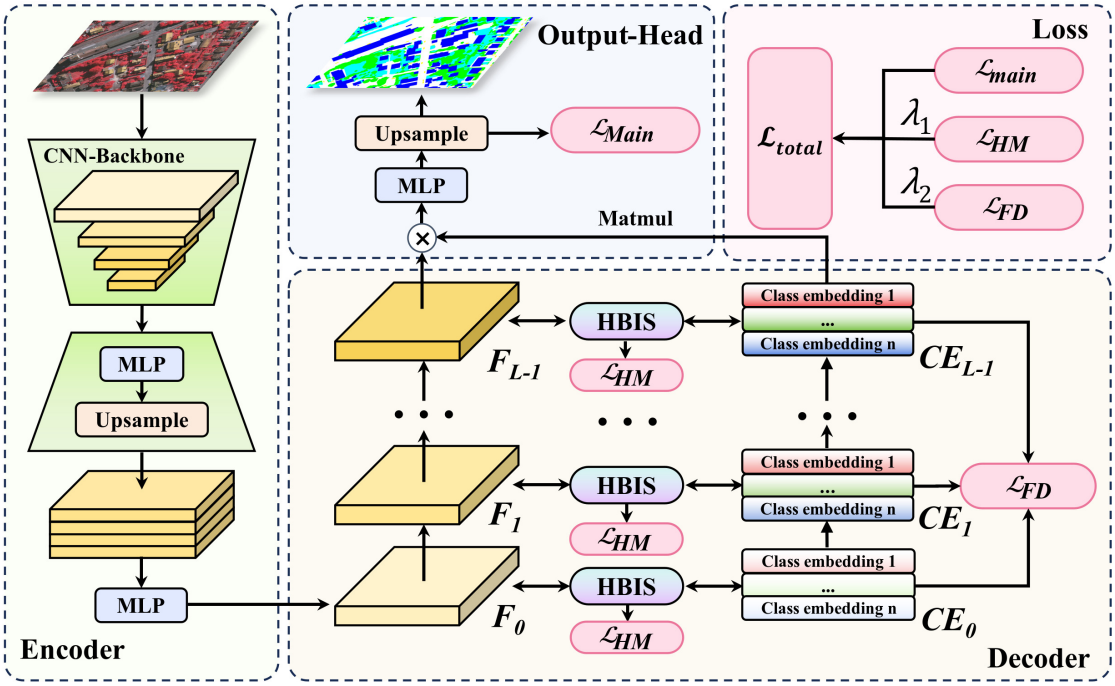

Figure 1: Overall architecture of the proposed BiCoR-Seg framework. The proposed framework consists of an encoder for multi-scale feature extraction, a decoder composed of multiple cascaded HBIS modules, and an output head.

of modern remote sensing segmentation models.

2.2. Explicit Semantic Modeling Methods

Recent studies have begun exploring explicitly semantic-guided paradigms to more directly integrate category-level semantic knowledge into the segmentation process. The core idea originates from the query-based object detector DETR [58], which introduces a set of learnable class prototypes or class queries. These methods maintain a dedicated embedding vector for each semantic category, serving as a carrier of high-level semantic concepts that interact with pixel-level features during the network's decoding stage. In the field of general segmentation, MaskFormer [11] and its subsequent improvement Mask2Former [10] have achieved remarkable success by reformulating the segmentation task as a query-based mask classification problem. K-Net [53] further extends this idea by employing a set of dynamically updated kernels for instance segmentation, while SegViT [50] utilizes learnable class queries to compute attention maps for generating explicit semantic masks. In the remote sensing domain, the Prototypical Contrastive Network (PCN) [30] enhances intra-class compactness in binary segmentation tasks through prototype contrastive constraints, whereas CenterSeg [52] further designs a classifier based on category centers and multiple prototypes.

However, in existing methods, the interaction between semantics and visual features is typically unidirectional and shallow, preventing high-level semantic knowledge from effectively guiding low-level feature learning. Moreover, most prior class embedding approaches are static, meaning that the class prototypes remain unchanged during inference, independent of the image context, and lack hierarchical adaptability or explicit regularization within the embedding space. In contrast to the above works, our proposed BiCoR-Seg framework establishes a bidirectional and iterative information flow across all decoder stages via the HBIS module, leveraging class heatmaps to ensure that semantic knowledge deeply and continuously guides the entire feature learning process. Furthermore, we introduce hierarchical class heatmap supervision and a Fisher discriminant loss, which jointly impose direct and effective constraints on the cooperative refinement between features and embeddings from both spatial localization and semantic distribution perspectives.

3. Method

3.1. Overall Architecture

To address the issues of large intra-class variance and small inter-class differences in high-resolution remote sensing image semantic segmentation (HRSS), which often result in recognition ambiguity and blurred boundaries, and to enable interactive fusion between class embeddings and pixel-level features during the decoding stage, we pro-

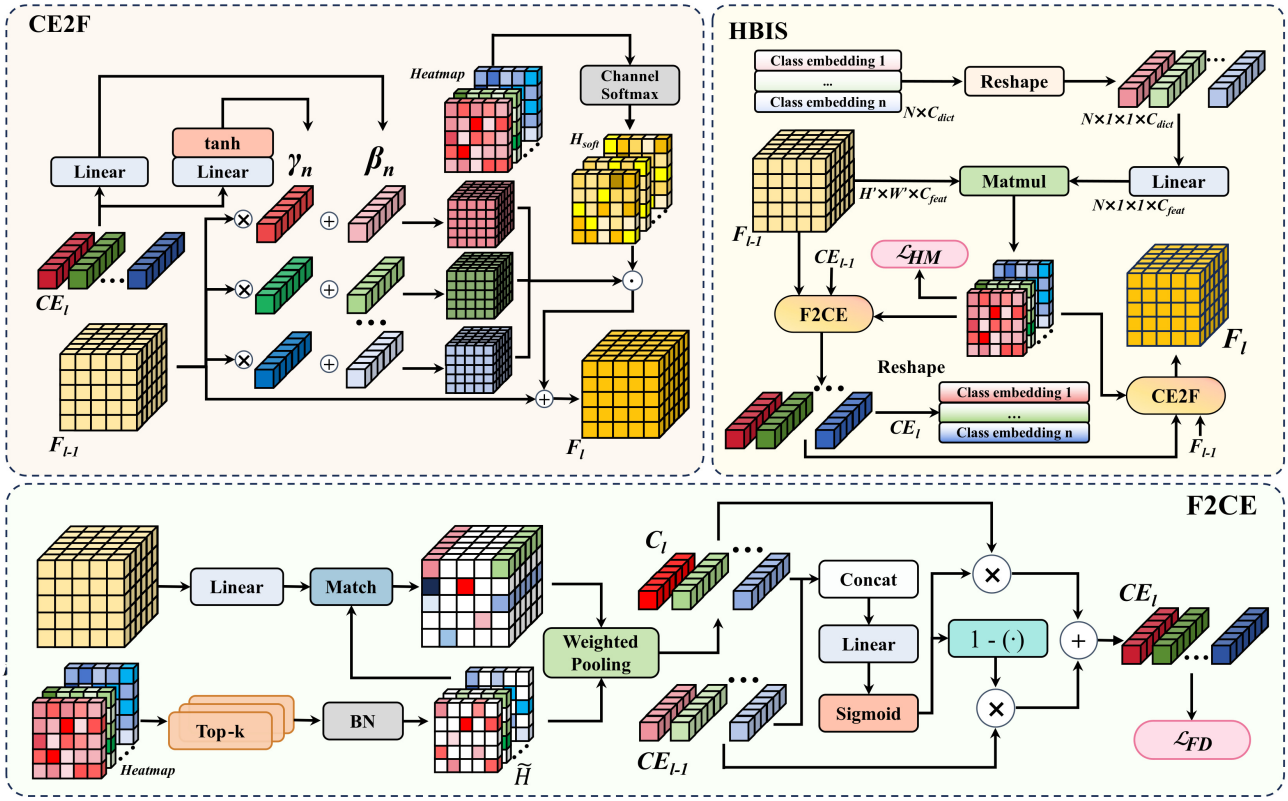

Figure 2: Architecture of the proposed HBIS module. The HBIS module is composed of two information pathways: Feature-to-Class Embedding (F2CE) and Class Embedding-to-Feature (CE2F). Three types of heatmaps are utilized in the bidirectional interaction: the raw class heatmap, the normalized heatmap $\tilde{H}$ (with a BN layer for normalization), and the Softmax-normalized heatmap H^{soft} for semantic fusion.

pose a Bidirectional Cooperative Refinement framework for HRSS(BiCoR-Seg). This framework progressively injects discriminative semantic concepts into the pixel feature learning process, thereby achieving deep integration between semantic embeddings and visual representations.

As illustrated in Fig. 1, the proposed framework consists of three main components: an encoder for extracting multi-scale visual features, an iterative decoder that performs cooperative refinement between features and semantics, and an output head that produces the final segmentation results. Specifically, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the encoder first extracts multi-scale features and fuses them through an aggregation module composed of MLPs and upsampling operations, resulting in a high-resolution feature map F_0 and an initial class embedding CE_0. Subsequently, F_0 and CE_0 are refined through an L-layer decoder constructed with HBIS modules, which perform bidirectional cooperative refinement between pixel features and class embeddings. Let $CE_l = \{CE_{l,n}\}_{n=1}^{N}$ denote the set of class embeddings output by the l-th decoder layer, where $CE_{l,n}$ represents the embedding vector of the n-th class at layer l, and N is the total number of semantic classes. Finally, the output head performs matrix multiplication between the final feature map F_L and the class embeddings CE_L, and the result is upsampled to obtain the final pixel-level probability prediction map P.

In the following sections, we focus on the core component of the framework, the Heatmap-driven Bidirectional Information Synergy module (HBIS), and provide a detailed description of the hierarchical heatmap supervision loss and Fisher discriminant loss that are specifically designed to optimize this architecture.

3.2. Heatmap-driven Bidirectional Information Synergy Module (HBIS)

Although current explicit semantic modeling methods introduce class-prototype embeddings, their interactions are often unidirectional: the class embeddings only participate in feature aggregation at the end of decoding, lacking positive guidance for shallow feature learning. Moreover, the updating process of class embeddings is static and independent of the image content, preventing high-level semantic knowledge from actively influencing the entire feature learning process.

To address this issue, we design HBIS, as illustrated in Fig. 2, to establishes an interpretable and effective bidirectional communication between class embeddings and pixel-level features. In the following subsection, we first introduce the core component, namely the generation of class heatmaps, and then, based on this foundation, elaborate on two enhancement pathways within HBIS: the feature-to-class embedding update path (F2CE) and the class embedding-to-feature enhancement path (CE2F).

Class Heatmap Generation. To initiate bidirectional communication, it is first necessary to establish a correspondence between abstract class concepts and their spatial locations in the image. Therefore, we design a class heatmap that projects each class embedding into the image space, generating a spatial confidence map that reveals the potential distribution of each class across the image.

We dynamically activate the functionality of the class embeddings, enabling them to serve as query vectors. Specifically, the class embeddings from the previous layer, denoted as CE_{l-1}, are transformed through a learnable linear layer $\text{Linear}(\cdot)$ into a set of vectors whose dimensions match the channel dimension of the feature map C_{feat}. By computing the dot-product similarity between each pixel feature and all class embeddings, we obtain the response intensity of the n-th class at the spatial location (x, y):

$$H_{l,n}(x,y) = \sigma\big(F_{l-1}(x,y) \cdot \text{Linear}(CE_{l-1,n})^\top\big), \quad (1)$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function. The output value $H_{l,n}(x,y) \in [0,1]$ represents the confidence that the pixel belongs to the n-th class. Consequently, a class heatmap for the current layer is formed as $H_l \in \mathbb{R}^{H' \times W' \times N}$, whose resolution is consistent with that of the feature map and is utilized for subsequent bidirectional information collaboration.

Feature Map to Class Embedding Updating (F2CE). The core objective of the F2CE pathway is to endow class embeddings with image adaptivity, thereby progressively refining generic semantic priors into contextual representations tightly related to the content of the current image. At layer l, this pathway aggregates, for each category, the most relevant visual features from the feature map F_{l-1}. The input class embedding CE_{l-1} thus incorporates the image-specific contextual information and is updated into a more discriminative CE_{l-1}.

This procedure consists of two steps. First is region-guided feature pooling: for each class, we use its heatmap $H_{l,n}$ to select the Top-K high-response pixels and aggregate them into a region $\Omega_{l,n}$. Using the normalized heatmap values as weights, the pixel features within these regions are projected by a linear layer, $\text{Linear}(\cdot)$, into the class-embedding dimensionality C_{class}, followed by weighted pooling. This ensures that the information aggregation concentrates on the most relevant visual cues while suppressing the interference from irrelevant areas. Finally, we obtain the contextual feature vector for class n at the current layer, $C_{l,n}$, which aggregates the most representative pixel information of class n in the current image, as formulated below

$$\tilde{H}_{l,n}(x,y) = \begin{cases} \dfrac{H_{l,n}(x,y)}{\sum\limits_{(u,v)\in\Omega_n} H_{l,n}(u,v) + \varepsilon}, & (x,y) \in \Omega_n, \\ 0, & (x,y) \notin \Omega_n, \end{cases} \quad (2)$$

$$C_{l,n} = \sum_{(x,y)\in\Omega_n} \tilde{H}_{l,n}(x,y) \cdot \text{Linear}(F_{l-1}(x,y)). \quad (3)$$

Here, $\tilde{H}_{l,n}(x,y)$ denotes the normalized class heatmap value of the n-th class at pixel position (x,y) in the l-th layer.

Directly adding or replacing the contextual vector C_l with the class embedding CE_{l-1} may lead to instability during training. Therefore, we introduce a gated update mechanism to achieve stable and adaptive fusion. Specifically, we compute an adaptive gating vector G_l by concatenating the previous class embedding and the current contextual feature vector, and then passing the result through a learnable linear layer $\text{Linear}(\cdot)$ to dynamically determine the ratio between information preservation and absorption:

$$G_l = \sigma\big(\text{Linear}([CE_{l-1} \| C_l])\big), \quad (4)$$

where $\|$ denotes vector concatenation, and $G_l \in \mathbb{R}^N$. Using this gating vector G_l, we perform weighted fusion between historical semantic information and current contextual information to generate updated class embeddings. For the n-th class embedding at the l-th layer, the update rule is defined as:

$$CE_{l,n} = (1 - G_{l,n}) \cdot CE_{l-1,n} + G_{l,n} \cdot C_{l,n}. \quad (5)$$

Here, $G_{l,n}$ represents the gating coefficient of G_l for the n-th class. This mechanism enables each class embedding to maintain semantic stability while adapting dynamically to the specific image content.

Class Embedding to Feature Map Updating (CE2F). When the class embeddings are updated through the F2CE pathway, they carry more accurate and context-aware semantic knowledge. The goal of the CE2F pathway is to write the semantic knowledge learned by CE_l back to the pixel-level feature maps, guiding and enhancing those ambiguous features to become more category-discriminative.

This process also consists of two steps. First is the class-specific feature modulation. We assume that objects of different classes should have distinct feature distributions. Therefore, using the updated class embeddings CE_l, we

learn a pair of affine transformation parameters for each class $CE_{l,n}$, namely the channel-wise scaling factor γ_n and bias β_n. These affine parameters are generated as follows:

$$\gamma_n = 1 + \tanh\big(\text{Linear}(CE_{l,n})\big), \tag{6}$$

$$\beta_n = \text{Linear}(CE_{l,n}), \tag{7}$$

where $\tanh(\cdot)$ denotes the hyperbolic tangent activation function. Both γ_n and β_n belong to $\mathbb{R}^{C_{\text{feat}}}$.These parameters can be viewed as the encapsulation of the n-th semantic concept in the feature space. The original features are modulated by these class-specific affine parameters to form the modulated features $\tilde{F}_{l,n}$. For each pixel (x, y), the updated feature vector $\tilde{F}_{l,n}(x, y)$ is computed as:

$$\tilde{F}_{l,n}(x, y) = \gamma_n \odot F_{l-1}(x, y) + \beta_n, \tag{8}$$

where $F_{l-1}(x, y) \in \mathbb{R}^{C_{\text{feat}}}$ represents the feature vector at position (x, y), and $\odot$ denotes channel-wise multiplication. Next comes the heatmap-guided modulation feature fusion. The degree to which each pixel should be influenced by a specific class-modulated feature is controlled by the class heatmap. We apply a Softmax normalization along the class dimension of the heatmap H_l to obtain the per-pixel class assignment probability H_l^{soft}:

$$H_{l,n}^{\text{soft}}(x, y) = \frac{\exp\big(H_{l,n}(x, y)\big)}{\sum_{k=1}^{N} \exp\big(H_{l,k}(x, y)\big)}, \tag{9}$$

In this context, $H_{l,n}^{\text{soft}}(x, y)$ denotes the Softmax-normalized heatmap value of the n-th class at pixel position (x, y) in the l-th layer, representing the probability that this pixel belongs to class n. Then, we use these probabilities as weights to fuse the modulated features from all classes and add them back to the original features to form the enhanced representation:

$$F_l(x, y) = \alpha F_{l-1}(x, y) + (1-\alpha) \sum_{n=1}^{N} H_{l,n}^{\text{soft}}(x, y) \cdot \tilde{F}_{l,n}(x, y), \tag{10}$$

where α is a learnable parameter, and $H_{l,n}^{\text{soft}}(x, y)$ denotes the probability that the pixel at position (x, y) belongs to the n-th class. Through the CE2F pathway, each pixel-level feature representation is directly guided by the most probable class-specific semantic knowledge, effectively mitigating issues of inter-class similarity and intra-class ambiguity.

For the hierarchical heatmap supervision loss $\mathcal{L}_{HM}$, the effectiveness of each HBIS module highly depends on the quality of the generated class heatmaps. To ensure that heatmaps at shallow layers can learn meaningful semantic localization ability, we regard them as low-resolution segmentation predictions and apply deep supervision. At each decoder layer l, the heatmap H_l is upsampled to the ground-truth size and constrained with standard segmentation losses:

$$\mathcal{L}_{HM} = \sum_{l=1}^{L} \mathcal{L}_{CE}(\text{Up}(H_l), Y) + \mathcal{L}_{Dice}(\text{Up}(H_l), Y), \tag{11}$$

where $\text{Up}(\cdot)$ denotes the upsampling operation that resizes the prediction map to the same spatial resolution as the ground truth Y. This deep supervision strategy effectively alleviates the gradient vanishing problem and strengthens the model's capability of learning discriminative spatial features in shallow layers.

3.3. Multi-level Cooperative Supervision

To ensure that BiCoR-Seg can learn highly discriminative features and semantic representations, we design a composite multi-level loss function. The total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda_1 \mathcal{L}_{HM} + \lambda_2 \mathcal{L}_{FD}, \tag{12}$$

$$\mathcal{L}_{main} = \mathcal{L}_{CE}(P, Y) + \mathcal{L}_{Dice}(P, Y). \tag{13}$$

Although HBIS achieves semantic alignment at the spatial level, class embeddings may still overlap in semantic space, that is, embeddings of different classes become too close, leading to ambiguous class discrimination. To address this, we introduce a Fisher Discriminant Constraint $\mathcal{L}_{FD}$, which explicitly enhances inter-class separability and intra-class compactness from a geometric perspective of the feature space. For the class embedding set at the l-th layer, $CE_l = \{CE_{l,n}\}_{n=1}^{N}$, we compute the within-class scatter $S_w^{(l)}$ and between-class scatter $S_b^{(l)}$ within a mini-batch as follows:

$$S_w^{(l)} = \frac{1}{BN} \sum_{n=1}^{N} \sum_{b=1}^{B} \| CE_{l,n}^{(b)} - \mu_n^{(l)} \|_2^2, \tag{14}$$

$$S_b^{(l)} = \frac{1}{N} \sum_{n=1}^{N} \| \mu_n^{(l)} - \mu^{(l)} \|_2^2, \tag{15}$$

where $\mu_n^{(l)}$ and $\mu^{(l)}$ denote the class center and the overall mean at the l-th layer, respectively:

$$\mu_n^{(l)} = \frac{1}{B} \sum_{b=1}^{B} CE_{l,n}^{(b)}, \tag{16}$$

$$\mu^{(l)} = \frac{1}{N} \sum_{n=1}^{N} \mu_n^{(l)}. \tag{17}$$

The Fisher Discriminant loss $\mathcal{L}_{FD}$ aims to minimize the ratio between the within-class scatter and the between-class

Table 1: Comparison with other methods on the LoveDA dataset. "Agri." denotes Agriculture.

Method	Venue	Background.	Building	Road	Water	Barren	Forest	Agri.	mIoU
FCN [27]	CVPR 2015	42.6	49.5	48.1	73.1	11.8	43.5	58.3	46.7
U-Net [32]	MICCAI 2015	43.1	52.7	52.8	73.1	10.3	43.0	59.9	47.8
PSPNet [54]	ISPRS 2016	44.4	52.1	53.5	76.5	9.7	44.1	57.9	48.3
LinkNet [3]	CVPR 2017	43.6	52.1	52.5	76.9	12.2	45.1	57.3	48.5
UNet++ [57]	MICCAI 2018	42.9	52.6	52.8	74.5	11.4	44.4	58.8	48.2
DeepLabv3+ [8]	ECCV 2018	43.0	50.9	52.0	74.4	10.4	44.2	58.5	47.6
SemanticFPN [20]	CVPR 2019	42.0	51.5	53.4	74.7	11.2	44.6	58.7	48.2
FarSeg [56]	CVPR 2020	43.4	51.8	53.3	76.1	10.8	43.2	58.6	48.2
SegFormer [45]	NeurIPS 2021	42.2	56.4	50.7	78.5	17.2	45.2	53.8	49.1
BANet [39]	GRSL 2022	43.7	51.5	51.1	76.9	16.6	44.9	62.5	49.6
UNetFormer [40]	ISPRS 2022	44.7	58.8	54.9	79.6	20.1	46.0	62.5	52.4
RSSFormer [47]	TIP 2023	**52.4**	60.7	55.2	76.3	18.7	45.4	58.3	52.4
TransUNet [4]	MIA 2024	43.0	56.1	53.7	78.0	9.3	44.9	56.9	48.9
Hi-ResNet [9]	JSTARS 2024	46.7	58.3	55.9	80.1	17.0	46.7	62.7	52.5
SFA-Net [18]	RS 2024	48.4	60.3	59.1	81.9	24.1	46.2	64.0	54.9
AFENet [16]	TGRS 2025	47.4	59.2	**59.2**	81.6	21.4	48.7	66.4	54.8
DMA-Net [13]	RS 2025	49.3	60.1	58.7	**82.3**	17.3	47.1	64.8	54.2
Ours	–	48.1	**60.9**	58.7	80.8	**24.4**	**48.7**	**66.8**	**55.5**

scatter, encouraging embeddings of the same category to be more compact and those of different categories to be farther apart in the feature space. It is formulated as:

$$\mathcal{L}_{FD}^{(l)} = \frac{S_w^{(l)}}{S_b^{(l)} + \epsilon}. \tag{18}$$

We apply this loss to the output CE_l of each decoder layer and compute a weighted summation to ensure that a highly discriminative semantic space is maintained from shallow to deep layers:

$$\mathcal{L}_{FD} = \sum_{l=1}^{L} \mathcal{L}_{FD}^{(l)}. \tag{19}$$

Through the cooperative effect of the hierarchical heatmap supervision loss $\mathcal{L}_{HM}$ and the Fisher Discriminant constraint $\mathcal{L}_{FD}$, BiCoR-Seg achieves a comprehensive optimization objective, which significantly enhances segmentation accuracy and class discrimination capability, while exhibiting superior generalization stability in complex scenarios.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct extensive experiments and evaluations on three challenging high-resolution remote sensing image segmentation benchmark datasets. To ensure reproducibility and fair comparison, all datasets are divided following the standard protocols provided by the open-source remote sensing segmentation framework GeoSeg [43].

LoveDA (Land-cover Domain Adaptive) [38]. A large-scale dataset containing both urban and rural scenes, with a spatial resolution of 0.3 m and covering seven land-cover categories. This dataset exhibits diverse scenes and severe class imbalance, posing high demands on the model's domain adaptation capability and robustness. It contains 2522 images for training, 1669 images for validation, and 1796 images for testing. Since the ground-truth labels of the test set are not publicly available, we follow the official evaluation protocol and submit our predictions to the online evaluation server to obtain the final results.

ISPRS Vaihingen. [33] This dataset consists of 33 high-resolution aerial images with a ground sampling distance (GSD) of 9 cm. It contains six classes in total: five fore-

Table 2: Comparison with other methods on the ISPRS Vaihingen dataset. "Imp. surf." and "Low veg." stand for Impervious surfaces and Low vegetation, respectively.

Method	Venue	Imp. surf.	Building	Low veg.	Tree	Car	mIoU	OA	mF1
U-Net [32]	MICCAI 2015	84.33	86.48	73.13	83.89	40.82	60.92	82.02	73.73
DANet [15]	CVPR 2019	91.13	94.82	83.47	88.92	62.98	74.52	89.52	84.27
TreeUNet [49]	ISPRS 2019	92.50	94.90	83.60	89.60	85.90	80.92	90.40	89.30
SegViT [50]	NeurIPS 2022	91.97	95.26	82.24	90.84	80.68	79.35	90.50	88.20
CTFNet [44]	GRSL 2023	90.69	94.35	81.66	87.27	82.72	77.84	88.60	87.34
SFA-Net [18]	RS 2024	93.50	96.30	85.40	90.20	**90.70**	84.06	-	91.20
LSRFormer [51]	TGRS 2024	93.84	**96.40**	**85.82**	90.70	90.79	84.56	91.90	91.50
AFENet [16]	TGRS 2025	96.90	95.72	85.07	90.64	89.37	84.55	91.67	91.54
Ours	–	**97.36**	96.11	85.78	**91.08**	89.40	**85.38**	**92.10**	**91.94**

Table 3: Comparison with other methods on the ISPRS Potsdam dataset.

Method	Venue	Imp. surf.	Building	Low veg.	Tree	Car	mIoU	OA	mF1
U-Net [32]	MICCAI 2015	86.54	88.73	72.09	79.49	46.67	60.46	85.29	75.64
DANet [15]	CVPR 2019	91.00	95.60	86.10	87.60	84.30	80.30	89.10	88.90
TreeUNet [49]	ISPRS 2019	93.10	97.30	86.80	87.10	95.80	85.50	90.70	92.00
SegViT [50]	NeurIPS 2022	92.45	97.01	89.77	89.98	95.23	86.85	91.20	92.89
CTFNet [44]	GRSL 2023	91.48	96.30	86.04	87.23	92.48	83.20	89.38	90.70
LSRFormer [51]	TGRS 2024	94.08	97.33	88.63	89.82	97.05	87.80	91.90	93.40
SFA-Net [18]	RS 2024	**95.00**	**97.50**	88.30	89.60	97.10	87.60	-	93.50
AFENet [16]	TGRS 2025	94.57	96.86	88.17	89.78	96.83	87.50	92.00	93.24
Ours	–	94.82	97.38	**91.83**	**91.40**	**97.52**	**89.87**	**92.10**	**94.59**

ground categories (impervious surfaces, buildings, low vegetation, trees, and cars) and one background class. Following the GeoSeg split protocol, 15 images are used for training, one image (ID 30) for validation, and 17 images (IDs 2, 4, 6, etc.) for testing [40]. All images are cropped into non-overlapping patches of 1024×1024 pixels for training and inference.

ISPRS Potsdam. [33] This dataset includes 38 aerial images with even higher resolution, each with a size of 6000×6000 pixels. The category definitions are consistent with those of the Vaihingen dataset. Following the official split, after removing one mislabeled image (ID 7_10), we use 22 images for training, one image (ID 2_10) for validation, and 14 images (IDs 2_13, 2_14, 3_13, etc.) for testing [40]. All images are also cropped into patches of 1024×1024 pixels before being fed into the model.

To quantitatively evaluate the segmentation performance, we adopt three widely used metrics in semantic segmentation tasks, following previous studies [12, 14, 27]: mean Intersection-over-Union (mIoU), Overall Accuracy (OA), and mean F1-Score (mF1).

4.2. Implementation Details

All experiments are conducted on an NVIDIA RTX 4090 GPU. We adopt ConvNeXt-B [26] as the backbone network and initialize it with weights pre-trained on ImageNet. Two HBIS layers are inserted into the decoder stage. In each layer, Top-K is employed as a hard selection operator to select the most discriminative regions for semantic aggregation, with the parameter set to 0.02. To ensure training stability, this selection process is treated as a stop-gradient operation, while gradients are propagated only through the subsequent weighted aggregation with respect to the feature representations.

Table 4: Ablation study on key components. The checkmark indicates the module or loss is enabled.

Modules				mIoU	Params (M)	FLOPs (G)	FPS
Baseline	**HBIS**	$\mathcal{L}_{FD}$	$\mathcal{L}_{HM}$	**LoveDA**			
✓				54.20	93.5	93.63	66.79
✓	✓			55.15	89.5	87.80	65.17
✓	✓	✓		55.21			
✓	✓		✓	55.36			
✓	✓	✓	✓	**55.49**			

We use the Adam optimizer [19] with an initial learning rate of 0.8×10^{-4} and apply a cosine annealing schedule to dynamically adjust the learning rate. The batch size is set to 8. The weights of the composite loss are empirically set to $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$.

4.3. Comparison with Other Methods

On the most challenging LoveDA dataset, compared with publicly reported results under similar-scale backbones (e.g., SegFormer with MiT-B5-level capacity), our model achieves an mIoU of 55.5%,,reaching a new SOTA level, as shown in Table 1. The model performs particularly well on the "Agriculture" and "Forest" categories, with IoUs of 66.8% and 48.7%, respectively, indicating that our approach effectively handles large-scale and texture-complex natural scenes.

On the ISPRS Vaihingen dataset, our model achieves an mF1 score of 91.94%, as shown in Table 2. The advantage is particularly evident for the "Imp. surf" and "Tree" classes, where the F1 scores reach 97.36% and 91.08%, respectively, fully demonstrating that BiCoR-Seg excels at modeling fine structures and complex boundaries in high-resolution aerial imagery.

On the ISPRS Potsdam dataset, our model also delivers strong results, with an mF1 of 94.59%, as shown in Table 3. It achieves notable improvements in distinguishing the highly confusable "Low veg." and "Tree" categories. Meanwhile, for the small-object class "Car", the model attains an F1 score of 97.52%, showcasing robust multi-scale perception and fine-grained segmentation capability in dense urban scenes.

In summary, BiCoR-Seg achieves performance surpassing existing methods across three datasets with distinct characteristics, providing strong evidence of the effectiveness of our proposed approach. This demonstrates that our model can dynamically and precisely inject high-level semantic knowledge into pixel-level feature learning, thereby significantly enhancing segmentation accuracy and generalization capability in complex remote sensing scenarios.

Table 5: Ablation on the number of HBIS layers.

Layers	mIoU		Params(M)
	LoveDA	**Vaihingen**	
$l = 1$	55.29	84.54	89.0
$l = 2$	**55.49**	**85.37**	89.5
$l = 3$	55.28	84.67	90.0

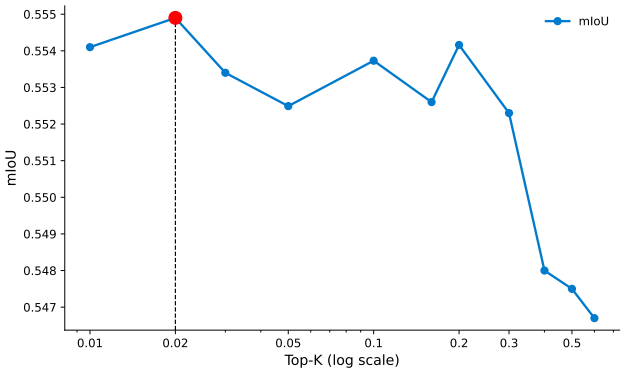

Figure 3: Effect of the Top-K threshold on model performance on the LoveDA dataset.

4.4. Ablation Study

Ablation on Key Components. To systematically validate the effectiveness of the core components in our framework, we conduct ablation studies on the HBIS module and the multi-level cooperative supervision losses. We set a model employing a conventional *cross-attention* mechanism as the baseline. Specifically, we replace the HBIS module in the decoder with a standard cross-attention layer, which also enables interactions between class embeddings and pixel feature maps, but lacks our heatmap-guided, *explicit bidirectional co-refinement* mechanism. The baseline model is optimized only with the main segmentation loss $\mathcal{L}_{main}$.

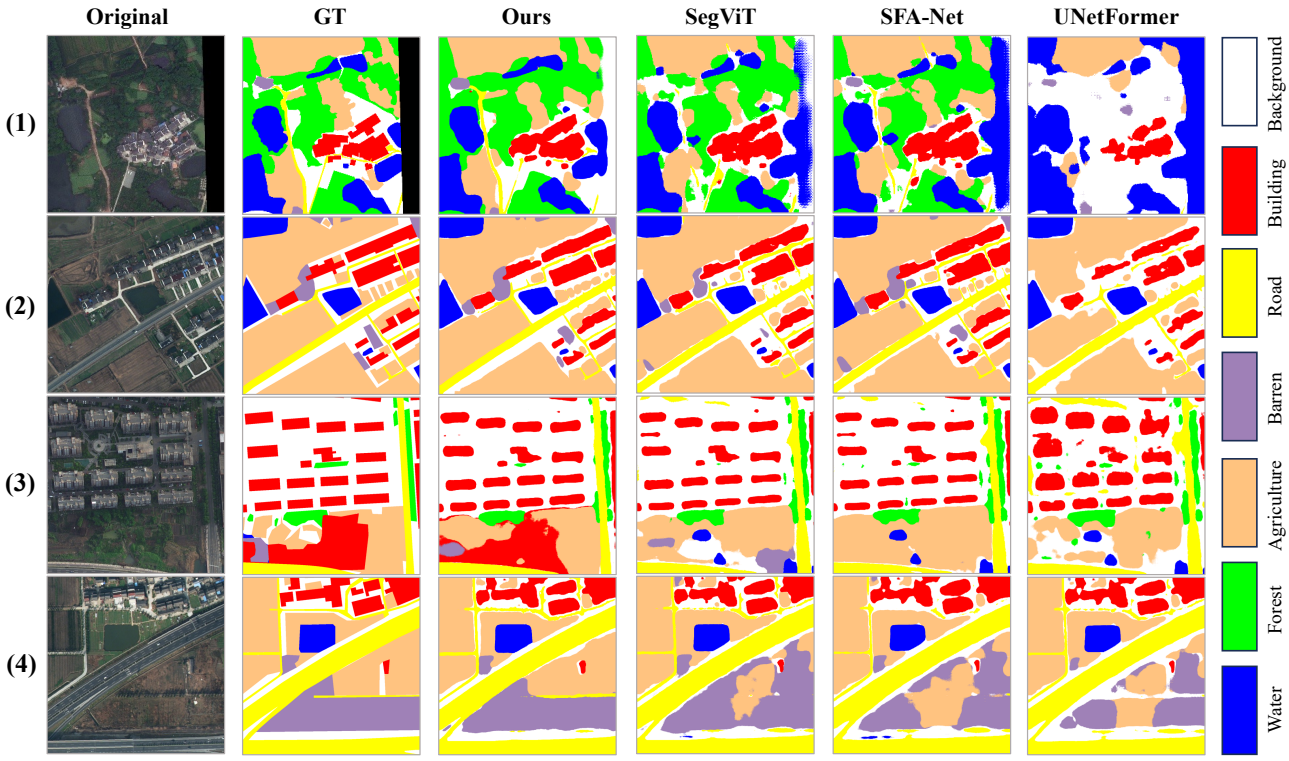

Figure 4: Qualitative comparison of segmentation results on challenging remote sensing scenes. BiCoR-Seg produces more complete and semantically consistent segmentation maps compared with other representative methods.

Quantitative Analysis As shown in Table 4, simply replacing the cross-attention module in the baseline with our HBIS module yields a notable improvement: the mIoU on LoveDA increases from $\mathbf{54.20}\%$ to $\mathbf{55.15}\%$. This gain strongly demonstrates that, compared with conventional attention-based interaction, our bidirectional co-refinement mechanism more effectively establishes tight connections between high-level semantics and pixel features. We further introduce two auxiliary losses. Using either the hierarchical heatmap supervision or the Fisher discriminant loss alone consistently improves performance. When both are applied together, the model achieves the best results, reaching $\mathbf{55.49}\%$ mIoU on LoveDA. These results indicate that the spatial alignment supervision provided by $\mathcal{L}_{HM}$ and the semantic discriminability constraint from $\mathcal{L}_{FD}$ are complementary, and their combination offers the most comprehensive optimization objective for BiCoR-Seg.

In terms of computational complexity and inference efficiency (Table 4), replacing the conventional cross-attention with the proposed HBIS module introduces no significant overhead. Specifically, FLOPs are reduced from 93.63G to 87.80G, respectively, while the inference speed only slightly decreases from 66.79 FPS to 65.17 FPS. These results demonstrate that the cooperative refinement mechanism achieves consistent performance gains with a limited and controllable impact on inference efficiency.

Number of HBIS Modules L. The number of cascaded HBIS modules L in the decoder determines the depth of semantic–feature interaction and refinement. We evaluate the model performance with different values of L on the LoveDA and Vaihingen datasets. As shown in Table 5, when L=1, the performance is relatively low, indicating that a single interaction is insufficient for adequate semantic alignment. When L increases to 2, the performance improves markedly. However, when L is further increased to 3, a slight degradation is observed, which may be attributed to information redundancy caused by overly deep interactions. Therefore, we set L=2 in all experiments to strike the best balance between performance and computational efficiency.

Effect of the Top-K Threshold. In the F2CE pathway, we employ a Top-K mechanism to select high-response pixels for updating class embeddings. The size of the Top-K threshold determines the range of visual feature used for class embedding updates. A too-small threshold may lead to insufficient information, whereas an excessively large one may introduce noise. We evaluate the influence of different

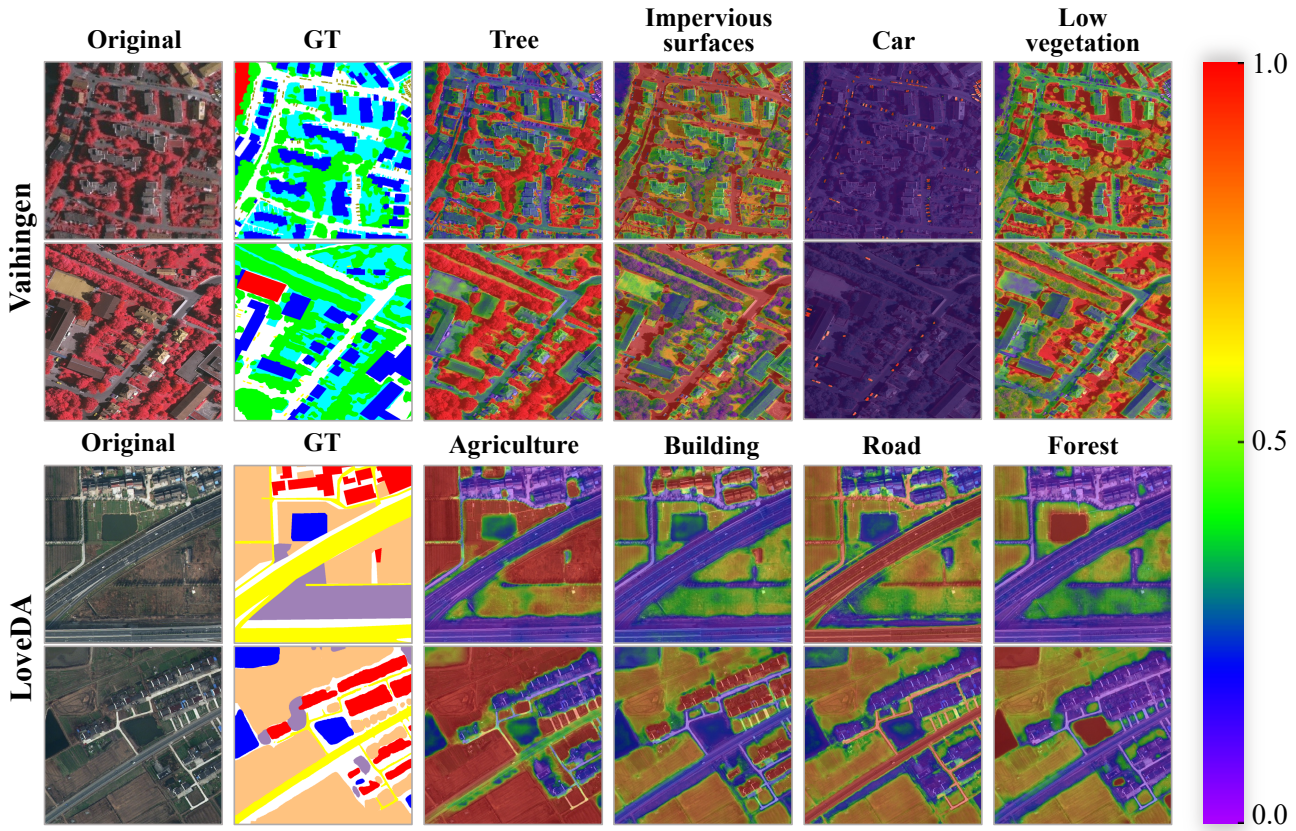

Figure 5: Visualization of class heatmaps generated by BiCoR-Seg for representative categories. The HBIS module effectively highlights category-specific regions while suppressing irrelevant background noise.

threshold values on model performance using the LoveDA dataset, and the results are shown in Fig. 3.

Experimental results show that when the Top-K threshold is set to 0.02, the model achieves the best performance, reaching an mIoU of 55.49%. This indicates that selecting the top 2% of pixels with the highest responses to update class embeddings captures the most representative visual information while effectively filtering out irrelevant regions.

We visualize the segmentation results of BiCoR-Seg and several representative methods on challenging scenes, as shown in Fig. 4. It can be clearly observed that our method significantly outperforms other segmentation approaches. For instance, in the fourth example, the comparative methods confuse the purple *bare soil* with the orange *agriculture*, while our method generates complete and highly discriminative segmentation masks, demonstrating stronger spatial awareness.

To further illustrate the interpretability of BiCoR-Seg, particularly the working mechanism of the HBIS module, we visualize the class heatmaps generated for representative categories, as shown in Fig. 5. The heatmaps explicitly reveal the spatial regions that the model focuses on during segmentation decision-making. It can be seen that the heatmaps accurately highlight the pixel regions corresponding to each semantic category and effectively suppress background noise. For the *building* category, the heatmap precisely emphasizes rooftop regions, while for small objects such as *vehicles*, the heatmap also achieves accurate localization. These observations demonstrate that, through iterative collaborative refinement, BiCoR-Seg is capable of purifying its internal semantic representations and focusing its attention on the most discriminative ground-object regions. This provides interpretable evidence for its superior segmentation performance, visually confirming that our proposed HBIS mechanism enables high-level semantic concepts to be progressively and precisely aligned with pixel-level visual evidence, thereby significantly improving final segmentation accuracy.

5. Conclusion

In this paper, we propose a Bidirectional Collaborative Refinement Framework (BiCoR-Seg) to address the preva-

lent issues of inter-class similarity, intra-class variability, and semantic ambiguity in high-resolution remote sensing image semantic segmentation. The framework establishes an explicit semantic association between feature representations and class embeddings through a heatmap-driven bidirectional information collaboration mechanism, enabling the model to achieve progressive optimization of features and semantics across layers. Experimental results demonstrate that BiCoR-Seg achieves superior performance on multiple benchmark datasets, significantly improving segmentation accuracy and boundary sharpness in complex scenes. Moreover, quantitative analyses show that the class-specific heatmaps accurately reflect the model's discriminative rationale, thereby validating the effectiveness of the bidirectional refinement mechanism.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1, 2

[2] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2

[3] A. Chaurasia and E. Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017. 7

[4] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. 2, 7

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arxiv 2017. *arXiv preprint arXiv:1706.05587*, 2:1, 2019. 2

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7

[9] Y. Chen, P. Fang, X. Zhong, J. Yu, X. Zhang, and T. Li. Hiresnet: Edge detail enhancement for high-resolution remote sensing segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 7

[10] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[11] B. Cheng, A. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 3

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 8

[13] C. Deng, H. Liang, X. Qin, and S. Wang. Dma-net: Dynamic morphology-aware segmentation network for remote sensing images. *Remote Sensing*, 17(14):2354, 2025. 7

[14] M. Everingham. The pascal visual object classes challenge,(voc2007) results. *http://pascallin. ecs. soton. ac. uk/challenges/VOC/voc2007/index. html.*, 2007. 8

[15] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 8

[16] F. Gao, M. Fu, J. Cao, J. Dong, and Q. Du. Adaptive frequency enhancement network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 7, 8

[17] T. Hanyu, K. Yamazaki, M. Tran, R. A. McCann, H. Liao, C. Rainwater, M. Adkins, J. Cothren, and N. Le. Aerialformer: Multi-resolution transformer for aerial image segmentation. *Remote Sensing*, 16(16):2930, 2024. 2

[18] G. Hwang, J. Jeong, and S. J. Lee. Sfa-net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sensing*, 16(17):3278, 2024. 7, 8

[19] D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 9

[20] A. Kirillov, R. Girshick, K. He, and P. Dollar. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[21] M. Lei, S. Li, Y. Wu, H. Hu, Y. Zhou, X. Zheng, G. Ding, S. Du, Z. Wu, and Y. Gao. Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception, 2025. 1

[22] M. Lei and X. Liu. Solo-net: A sparser but wiser method for small object detection in remote-sensing images. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2023. 1

[23] M. Lei, H. Wu, X. Lv, and L. Jiang. Ddranet: A dynamic density-region-aware network for crowd counting. *IEEE Signal Processing Letters*, 2024. 1

[24] M. Lei, H. Wu, X. Lv, and X. Wang. Condseg: A general medical image segmentation framework via contrast-driven feature enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4571–4579, 2025. 2

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 8

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2, 7, 8

[28] Z. Luo, J. Pan, Y. Hu, L. Deng, Y. Li, C. Qi, and X. Wang. Rs-dseg: semantic segmentation of high-resolution remote sensing images based on a diffusion model component with unsupervised pretraining. *Scientific Reports*, 14(1):18609, 2024. 1

[29] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019. 1

[30] K. Nogueira, M. M. Faita-Pinheiro, A. P. M. Ramos, W. N. Gonçalves, J. M. Junior, and 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 (WACV)*, pages 8366–8376, January 2024. 1, 3

[31] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2, 7, 8

[33] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf. The ISPRS benchmark on urban object classification and 3d building reconstruction. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume I-3, pages 293–298, 2012. 7, 8

[34] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[35] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[36] L. Sun, H. Zou, J. Wei, X. Cao, S. He, M. Li, and S. Liu. Semantic segmentation of high-resolution remote sensing images based on sparse self-attention and feature alignment. *Remote Sensing*, 15(6):1598, 2023. 1

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2

[38] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 7

[39] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 7

[40] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. 2, 7, 8

[41] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7303–7313, 2021. 1

[42] D. Wu, Z. Guo, A. Li, C. Yu, N. Sang, and C. Gao. Semantic segmentation via pixel-to-center similarity calculation. *CAAI Transactions on Intelligence Technology*, 9(1):87–100, 2024. 1

[43] G. Wu and Z. Guo. Geoseg: A computer vision package for automatic building segmentation and outline extraction. *CoRR*, abs/1809.03175, 2018. 7

[44] H. Wu, P. Huang, M. Zhang, and W. Tang. Ctfnet: Cnn-transformer fusion network for remote-sensing image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2023. 8

[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 2, 7

[46] G. Xu, X. Zhang, X. He, and X. Wu. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 42–53. Springer, 2023. 2

[47] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang. Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Transactions on Image Processing*, 32:1052–1064, 2023. 7

[48] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1

[49] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li. Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 156:1–13, 2019. 8

[50] B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, et al. Segvit: Semantic segmentation with plain vision transformers. *Advances in Neural Information Processing Systems*, 35:4971–4982, 2022. 1, 3, 8

[51] R. Zhang, Q. Zhang, and G. Zhang. Lsrformer: Efficient transformer supply convolutional neural networks with global information for aerial image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. 8

[52] W. Zhang, M. Ma, Y. Jiang, R. Lian, Z. Wu, K. Cui, and X. Ma. Center-guided classifier for semantic segmentation of remote sensing images. *arXiv preprint arXiv:2503.16963*, 2025. 3

[53] W. Zhang, J. Pang, K. Chen, and C. C. Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 3

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 7

[55] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[56] Z. Zheng, Y. Zhong, J. Wang, and A. Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4096–4105, 2020. 7

[57] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018. 2, 7

[58] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[59] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017. 1