*remote sensing*

MDPI

*Article*

# CRECA-Net: Class Representation-Enhanced Class-Aware Network for Semantic Segmentation of High-Resolution Remote Sensing Images

**Ruolan Liu, Bingcai Chen *, Lin Yu and Shaodong Zhang**

School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China; rl@mail.dlut.edu.cn (R.L.)

* Correspondence: china@dlut.edu.cn

## Highlights

**What are the main findings?**

- We propose CRECA-Net, a novel class-aware segmentation network that integrates the CPR module and the DA loss to enhance segmentation performance on high-resolution remote sensing imagery characterized by complex backgrounds and high inter-class similarity.
- The CPR module incorporates pixel selection, confidence-aware contribution weighting, and inter-class prototype separation constraints to generate reliable and discriminative class representations for subsequent CLCA modules, while the DA loss adaptively emphasizes hard samples during training to improve overall segmentation performance.

**What are the implications of the main findings?**

- CRECA-Net provides a high-accuracy framework for the fine-grained interpretation of high-resolution remote sensing scenes, supporting applications such as urban planning and land-use analysis.
- The findings demonstrate that improving class representation quality is essential for achieving superior segmentation performance in class-aware networks, offering valuable insights for the design of segmentation models in complex remote sensing scenarios.

## Abstract

High-resolution remote sensing (RS) images exhibit complex backgrounds, large intra-class variability, and low inter-class differences, posing substantial challenges for semantic segmentation. Although existing class-level contextual modeling methods partially alleviate these issues, they often overlook the importance of accurate and discriminative class representations and fail to effectively handle hard samples during training. To address these limitations, we propose CRECA-Net, a class representation-enhanced class-aware network designed from two complementary perspectives: class prototype refinement and difficulty-aware learning. Specifically, we introduce a class prototype refinement (CPR) module that improves class representations through pixel selection, confidence-aware contribution weighting, and an inter-class prototype separation loss, yielding more reliable and discriminative class centers. In addition, class-level context aggregation (CLCA) modules capture pixel-to-class prototype correlations via cross-attention to inject class-aware semantics into decoder features, thereby reducing interference from cluttered backgrounds and visually similar categories. Furthermore, a difficulty-aware (DA) loss dynamically estimates pixel-wise difficulty and redistributes the loss weights within each image, gradually shifting the learning focus from easy to hard samples while maintaining training stability. Extensive experiments on two benchmark RS segmentation datasets demonstrate that CRECA-Net consistently outperforms state-of-the-art methods across multiple evaluation metrics.

check for updates

Academic Editor: Zhenwei Shi

Received: 10 December 2025
Revised: 6 March 2026
Accepted: 9 March 2026
Published: 21 March 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

**Keywords:** class prototype refinement; class-level context modeling; difficulty-aware loss; remote sensing image; semantic segmentation

## 1. Introduction

Advances in remote sensing platforms and Earth observation technologies have markedly improved the spatial resolution of remote sensing (RS) imagery, enabling finer and more reliable observations of the Earth's surface. Semantic segmentation, which assigns a semantic label to each pixel, has become a core technique for fine-grained RS scene understanding. It supports precise identification and delineation of geographic objects and serves as a fundamental tool in a wide range of applications, including road extraction [1–3], traffic monitoring [4,5], urban planning [6–8], land cover classification [9–11], and disaster assessment [12,13].

Driven by rapid advances in deep learning and the increasing availability of large-scale, annotated datasets, automatic segmentation of RS imagery has achieved remarkable success in recent years. However, despite these advances, accurately segmenting geospatial objects in high-resolution RS images remains a significant challenge, primarily due to several inherent characteristics of RS data.

***Complex Background:*** RS images often exhibit highly complex and heterogeneous backgrounds, where target objects are frequently embedded in or obscured by large, cluttered surroundings. This complexity often leads to higher false-positive rates in segmentation. In certain datasets, irrelevant background regions may occupy more than 90% of the total image area [14].

***Large Intra-Class Variance and Small Inter-Class Variance:*** Although fine spatial resolution provides richer visual cues, such as detailed textures, shapes, and subtle color variations, it also exacerbates intra-class diversity while reducing inter-class separability. This issue is further intensified by structural ambiguity and boundary uncertainty arising from complex geographic layouts and the top-down imaging perspective. Consequently, distinguishing visually similar and easily confusable categories becomes particularly challenging in high-resolution RS imagery. For instance, objects of the same category (e.g., cars) may differ considerably in size, color, and orientation even within a single scene, while similar categories (e.g., trees vs. low vegetation or rooftops vs. ground surfaces) are often difficult to distinguish due to highly similar spectral or textural properties. As illustrated in Figure 1, these factors collectively contribute to large intra-class variance (yellow arrows) and small inter-class variance (blue arrows), making purely appearance-based feature discrimination inadequate. To address these challenges, numerous studies [15–18] have attempted to improve segmentation accuracy by aggregating contextual information.

Existing context aggregation methods in semantic segmentation can be broadly categorized into two paradigms: spatial and relational context aggregation.

Spatial context aggregation methods commonly expand the receptive field by capturing multi-scale spatial contexts. Representative approaches such as PSPNet [15] and the DeepLab series [16,19,20] exploit pyramid pooling and dilated convolutions to aggregate contextual information from multiple spatial scales. Although effective, these methods rely on fixed isotropic receptive windows that aggregate contextual information in a category-agnostic manner, thereby indiscriminately integrating category-relevant and irrelevant regions. As a result, the aggregated context becomes susceptible to interference from background noise and visually similar yet semantically different categories, a problem that is particularly pronounced in RS imagery characterized by cluttered scenes and subtle inter-class differences.

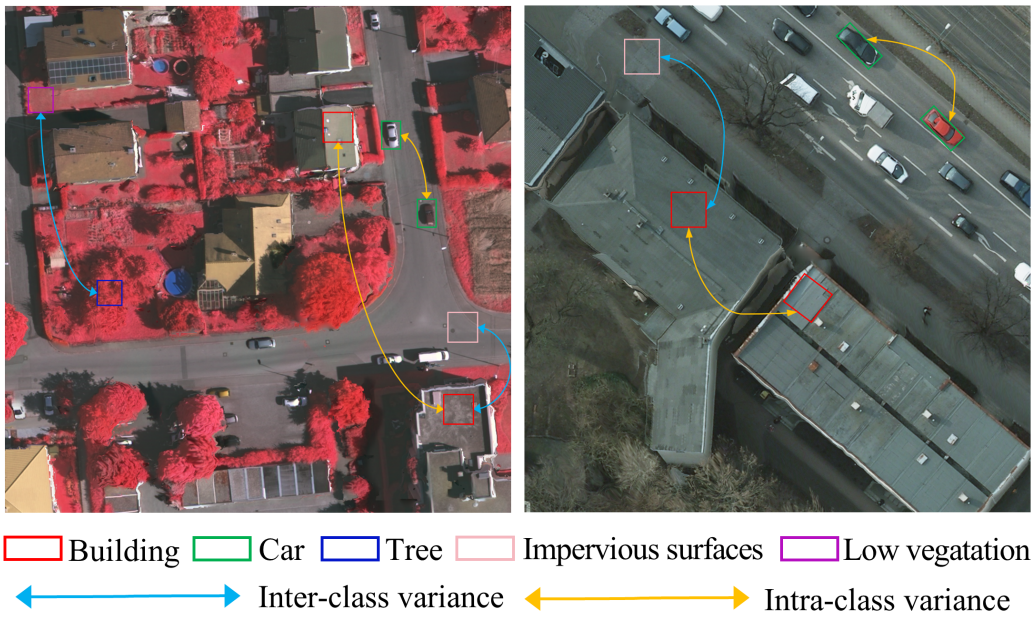



**Figure 1.** Illustration of the large intra-class variance and small inter-class variance in remote sensing (RS) images. Yellow arrows denote intra-class variance, while blue arrows indicate inter-class variance.

To overcome these limitations, relational context aggregation methods employ attention mechanisms to adaptively model long-range dependencies, allowing contextual cues to be selected based on semantic relevance rather than spatial proximity. According to the type of relationships, these methods can be categorized into pixel-level and class-level relational modeling methods.

Pixel-level relational modeling explicitly computes pairwise dependencies among all pixels to perform dynamic feature aggregation. For example, NonLocal [17] constructs a pairwise affinity matrix to capture global dependencies, while DANet [21] introduces spatial and channel attention mechanisms to refine pixel representations. Transformer-based segmentation frameworks [22–25] generalize this idea by applying self-attention to image tokens or patches, enabling powerful global contextual reasoning. Related efforts have also explored Transformer-inspired models or Transformer frameworks, for cross-view feature matching [26,27], providing valuable methodological insights for multimodal semantic segmentation [28,29]. However, the quadratic computational and memory complexity of full attention limits their scalability to high-resolution RS imagery. Although sparse attention variants [30,31] and efficient Transformer architectures [32–34] alleviate computational burdens through structured or hierarchical attention mechanisms, they still operate at the pixel–pixel or token–token level, failing to explicitly model class-level semantic structures. Consequently, in complex RS scenes with subtle inter-class differences, the absence of such class-aware guidance may lead the model to aggregate context from visually similar yet semantically irrelevant regions, ultimately degrading segmentation performance.

In contrast, class-level context modeling aims to suppress interference from irrelevant or confusing regions and enhance feature discriminability by explicitly capturing the relationships between pixels and class-level semantic representations. Representative approaches such as ACFNet [35] and OCRNet [18] generate class centers (or object region representations) either by averaging the embeddings of pixels assigned to each predicted class or by aggregating all pixel features using coarse segmentation probability maps. While effective in natural image segmentation, these strategies encounter notable limitations when applied to high-resolution RS imagery, which typically features complex backgrounds, low

inter-class variability, and large intra-class variability. Under these conditions, existing class-level context modeling approaches encounter two fundamental limitations.

***Inaccuracy and weakly discriminative class prototypes:*** Under these challenging characteristics of RS imagery , class prototypes tend to become highly sensitive to noise, prediction errors, and ambiguous regions. Simple averaging (Figure 2a) assigns equal importance to all pixels and disregards variations in pixel confidence and representativeness, often leading to biased class prototypes. Although probability-weighted aggregation (Figure 2b) partially alleviates this issue, the soft aggregation of all pixels inevitably allows non-target pixels to contribute to class prototypes, potentially contaminating the learned representations. Moreover, most existing methods lack explicit inter-class separation constraints, leading to insufficiently discriminative class prototypes, an issue particularly pronounced in RS imagery with inherently low inter-class separability.

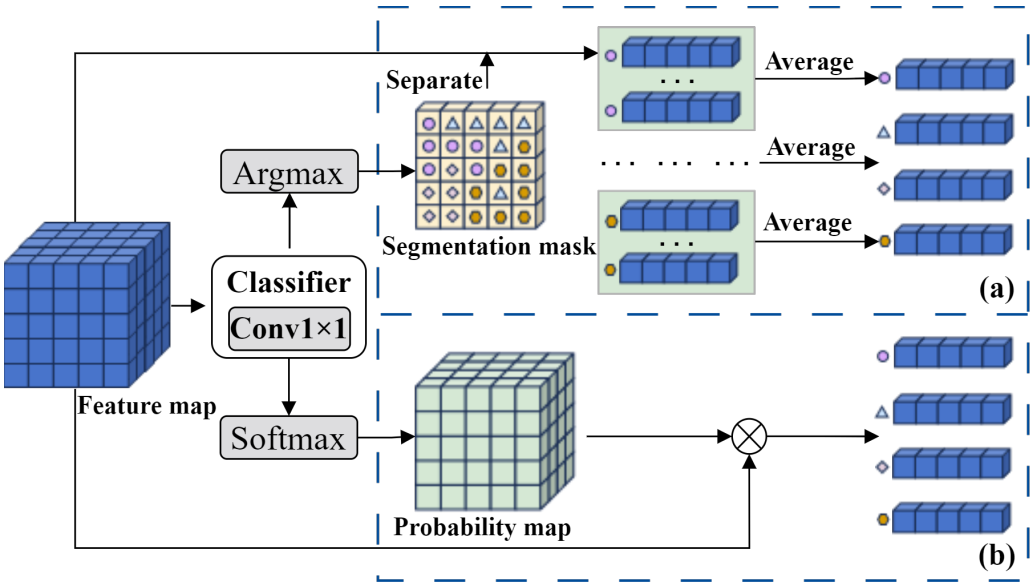



**Figure 2.** Illustration of two representative class center computation approaches: (**a**) simple averaging of pixel features assigned to each predicted class using a coarse segmentation mask; (**b**) probability-weighted aggregation of all pixel features using coarse segmentation probability maps. Different symbols (e.g., circles, triangles) indicate pixels belonging to different classes.

***Insufficient optimization of hard samples:*** These characteristics of RS imagery also lead to highly uneven pixel-wise segmentation difficulty. Boundary pixels, occluded regions, and pixels belonging to visually similar categories are significantly harder to classify than those in homogeneous regions. Despite this inherent imbalance, class-level context modeling frameworks are usually trained using the standard cross-entropy (CE) loss, which implicitly assumes uniform segmentation difficulty across pixels. Consequently, abundant easy pixels dominate optimization process, while hard samples receive inadequate supervision, leading to biased feature learning and reduced robustness in complex RS scenes.

To address these challenges, we propose CRECA-Net, a class representation-enhanced class-aware segmentation framework that simultaneously improves class prototypes and strengthens hard sample learning. Concretely, CRECA-Net integrates three key components: a class prototype refinement (CPR) module, which produces more reliable and discriminative prototypes by applying pixel selection, confidence-aware contribution weighting, and inter-class prototype separation regularization; a set of class-level context aggregation (CLCA) modules, which explicitly model pixel-to-class prototype correlations via cross-attention and progressively inject class-aware context into decoder features; and a

difficulty-aware (DA) loss that dynamically estimates pixel-wise segmentation difficulty and redistributes the intra-image loss to emphasize hard samples.

Our main contributions are summarized as follows:

- We present CRECA-Net, a class-aware segmentation framework that simultaneously enhances class representation learning and difficulty-adaptive optimization, effectively alleviating segmentation difficulty arising from low inter-class separability in high-resolution RS imagery.
- We propose a CPR module that constructs reliable and discriminative class prototypes through pixel selection, confidence-aware contribution weighting, and inter-class prototype separation regularization, thereby mitigating prototype bias and facilitating more effective class-level context modeling in subsequent CLCA modules.
- We introduce DA loss that dynamically estimates pixel-wise difficulty and adjusts the intra-image loss distribution to strengthen hard samples learning, resulting in more stable optimization and enhanced robustness. Extensive experiments on two benchmark datasets demonstrate that CRECA-Net achieves consistent performance improvements with minimal computational overhead.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the architecture of CRECA-Net and details its key components. Section 4 describes the datasets, evaluation metrics, and experimental setup. Section 5 reports the comparative and ablation studies. Finally, Sections 6 and 7 present the discussion of experimental results and the conclusion with future research directions, respectively.

## 2. Related Work

### *2.1. Semantic Segmentation*

Semantic segmentation extends image classification to the pixel level by assigning a semantic label to each pixel, thereby enabling fine-grained scene understanding. The Fully Convolutional Network (FCN) [36] pioneered this field by replacing fully connected layers with convolutional ones to achieve end-to-end dense prediction, laying the foundation for modern segmentation networks. UNet [37] and its variants [38,39] employ encoder–decoder architectures with skip connections to restore spatial details lost during downsampling. However, due to their reliance on local convolutional operations, these networks still suffer from a limited receptive field and inadequate global context modeling. Consequently, later studies have primarily focused on spatial or relational context aggregation to enhance global feature and improve segmentation performance.

### *2.2. Spatial Context Aggregation Methods*

Spatial context aggregation methods expand the receptive field by capturing contextual information from multiple spatial scales using dilated convolutions [40,41] or pooling operations. For example, PSPNet [15] employs a pyramid pooling module that performs parallel pooling with different kernel sizes to aggregate features across various spatial regions, enhancing multi-scale representations. Similarly, the DeepLab series [16,19,20] introduces the atrous spatial pyramid pooling (ASPP) module, which enlarges the receptive field via dilated convolutions with varying dilation rates while preserving spatial resolution. DenseASPP [42] further strengthens this framework by using dense connectivity for richer multi-scale feature extraction.

Several studies have further adapted these ideas to RS imagery. GCDNet [43] introduces a global context dependencies attention module to establish long-range relationships among multi-scale features extracted via dilated convolutions. AFNet [44] employs scale-feature and scale-layer attention mechanisms to adaptively enhance and fuse multi-scale features based on object sizes, thereby improving the segmentation performance for objects

with large scale variations. Although effective for multi-scale context extraction, these methods rely on fixed geometric receptive fields determined by kernel shapes or dilation rates. As a result, the aggregated context is prone to including target and non-target regions, producing unreliable semantic cues and exacerbating category confusion, an issue that becomes especially severe in RS images featuring cluttered backgrounds and subtle inter-class differences.

### 2.3. Relational Context Aggregation Methods

Relational context aggregation methods leverage attention mechanisms to adaptively capture long-range dependencies and can be broadly categorized into pixel-level and class-level approaches according to the type of relationships they model.

***Pixel-level Relational Context Aggregation:*** Early pixel-level approaches capture dependencies among all pixels to dynamically aggregate contextual information. For instance, NonLocal [17] models pairwise relationships across all spatial positions through non-local operations, enabling each pixel feature to access global context, while DANet [21] further introduces dual-attention to jointly capture spatial and channel dependencies. In recent years, Transformer architectures, represented by SegFormer [23], Segmenter [22], and Mask2Former [24], have further generalized pixel-level attention by applying self-attention to image tokens or patches, enabling powerful global reasoning. However, these dense attention operations require constructing full affinity matrices, resulting in quadratic computational and memory complexity.

To mitigate this limitation, numerous sparse and efficient attention mechanisms have been proposed. ANN [30] selects representative key-value pairs via pyramid sampling, CCNet [31] restricts attention computation to pixels along the same row and column, and EMANet [45] estimates attention maps through an expectation–maximization algorithm to reduce computational cost. In parallel, efficient Transformer architectures such as Swin Transformer [32], EfficientViT [33], and Biformer [34] employ window-based attention, lightweight ReLU-based attention mechanisms and bi-level routing strategies to further improve efficiency.

Given the large spatial dimensions and high computational demands of high-resolution RS imagery, numerous studies have developed efficient relational aggregation mechanisms to model long-range dependencies. MANet [46] and MAResU-Net [47] propose kernel and linear attention mechanisms to reduce the quadratic complexity of dot-product attention. ABCNet [48] leverages linear attention to construct global context paths, whereas UNetFormer [49] adopts a dual-attention design to jointly capture global and local dependencies. Similarly, SLCNet [50] explicitly supervises attention maps to guide long-range correlation modeling, and CGGLNet [51] integrates class-guided attention with adaptive multi-scale local features enhancement. Furthermore, MSGCNet [52] employs efficient cross-attention and window-based multihead self-attention for multiscale feature interaction and long-range dependency modeling, while AMMUNet [53] enhances global correlation modeling through multiscale attention map merging and granular multi-head self-attention. DLNet [54] further employs macro-level and micro-level self-attention in conjunction with cross-attention mechanisms to facilitate multi-scale feature extraction and fusion. Transformer-based models such as BANet [55] and DC-Swin [25] further integrate efficient transformer variants, including ResT [56] and Swin Transformer [32], to better capture long-range dependencies with reduced computational overhead.

Despite their effectiveness in extending receptive fields and enhancing global contextual reasoning, pixel-level and token-level attention mechanisms predominantly rely on appearance-based similarity to compute affinities. This reliance can lead to context being aggregated from irrelevant background regions or visually similar yet semantically distinct

categories, especially in high-resolution RS imagery where large intra-class variability and low inter-class separability make appearance-based correlations inherently unreliable. The lack of explicit class-level semantic guidance thus limits the reliability and discriminative power of pixel-level relational aggregation, motivating the development of class-aware context modeling paradigms.

***Class-level Relational Context Aggregation:*** Class-level modeling aims to overcome the limitations of pixel-level relational aggregation by explicitly modeling relationships between pixels and class representations (i.e., class centers or class prototypes). These prototypes are generally computed either by averaging the feature embeddings of pixels within each predicted class or by aggregating all pixel features through probability-based weighting. Representative methods include ACFNet [35], which aggregates contextual information directly through class centers and coarse segmentation results, and OCRNet [18], which enhances pixel features by computing their similarity to object region representations. ISNet [57] further integrates image-level and semantic-level context to refine pixel embeddings, while MCIBI [58] maintains a feature memory bank of dataset-level class prototypes mined across images.

Class-level relational modeling has also been extensively explored in the RS domain. HMANet [59] combines class-enhanced, class-channel, and region-shuffle attention to improve aerial image segmentation. CCANet [60] introduces a class-constrained, coarse-to-fine attention module that leverages category information as explicit constraints to capture long-range context. LOGCAN++ [61] applies affine transformations to adaptively extract local class representations serving as intermediate bridges between pixels and global class centers, mitigating intra-class variance.

Despite their effectiveness in introducing semantic priors and mitigating certain limitations of pixel-level/token-level attention, existing class-level relational modeling methods still face challenges when applied to high-resolution RS imagery, which typically exhibits complex backgrounds, substantial intra-class variability, and small inter-class differences. These factors can cause current prototype estimation strategies to produce unreliable and weakly discriminative class representations. Moreover, these properties result in highly uneven pixel-wise segmentation difficulty across regions, causing hard samples to receive insufficient supervision under conventional CE loss. These issues collectively compromise the robustness and overall effectiveness of existing class-level approaches when deployed in complex RS environments.

To address these challenges, we propose CRECA-Net, a class-aware network designed to integrate class prototype refinement and difficulty-aware learning. By constructing more reliable and discriminative class prototypes while dynamically emphasizing hard samples during training, CRECA-Net achieves more robust and accurate segmentation performance on high-resolution RS imagery.

## 3. Proposed Method

### *3.1. Overall Architecture*

The overall architecture of the proposed CRECA-Net is illustrated in Figure 3. The network follows a classical encoder-decoder structure with skip connections to balance global context modeling and spatial detail preservation. We employ a ResNet-50 [62] backbone, pre-trained on ImageNet [63], as the encoder for hierarchical feature extraction. Given an input RS image $I$, the encoder generates multi-scale feature maps $F_1$, $F_2$, $F_3$, and $F_4$ with spatial resolutions of 1/4, 1/8, 1/16, and 1/32 of the input size, respectively. In the decoder, all feature maps are first projected into a unified channel dimension ($d = 128$) via a $1 \times 1$ convolution. Next, we introduce a CPR module that operates on the deepest feature map $F_4$ to generate refined class prototypes $C$. These prototypes then guide top-down

feature refinement through four class-level context aggregation modules ($\mathrm{CLCA}_1 - \mathrm{CLCA}_4$). Specifically, $\mathrm{CLCA}_4$ takes $(F_4, C)$ as input and outputs an enhanced representation $F_4^{o'}$. This representation is fused with the higher-resolution feature map $F_3$ and processed by $\mathrm{CLCA}_3$ to obtain $F_3^{o'}$. The same refinement procedure is subsequently applied through $\mathrm{CLCA}_2$ and $\mathrm{CLCA}_1$, yielding progressively enriched features $F_2^{o'}$ and $F_1^{o'}$ that incorporate both semantic cues and spatial details. Finally, all enhanced feature maps $(F_i^{o'})_{i=1}^4$ are upsampled to a uniform spatial resolution and aggregated through element-wise summation, followed by a final $1 \times 1$ convolution to produce the segmentation prediction. The entire network is trained end-to-end using a joint objective that combines DA loss and inter-class prototype separation loss, enabling more effective learning from hard samples while encouraging discriminative prototype formation.

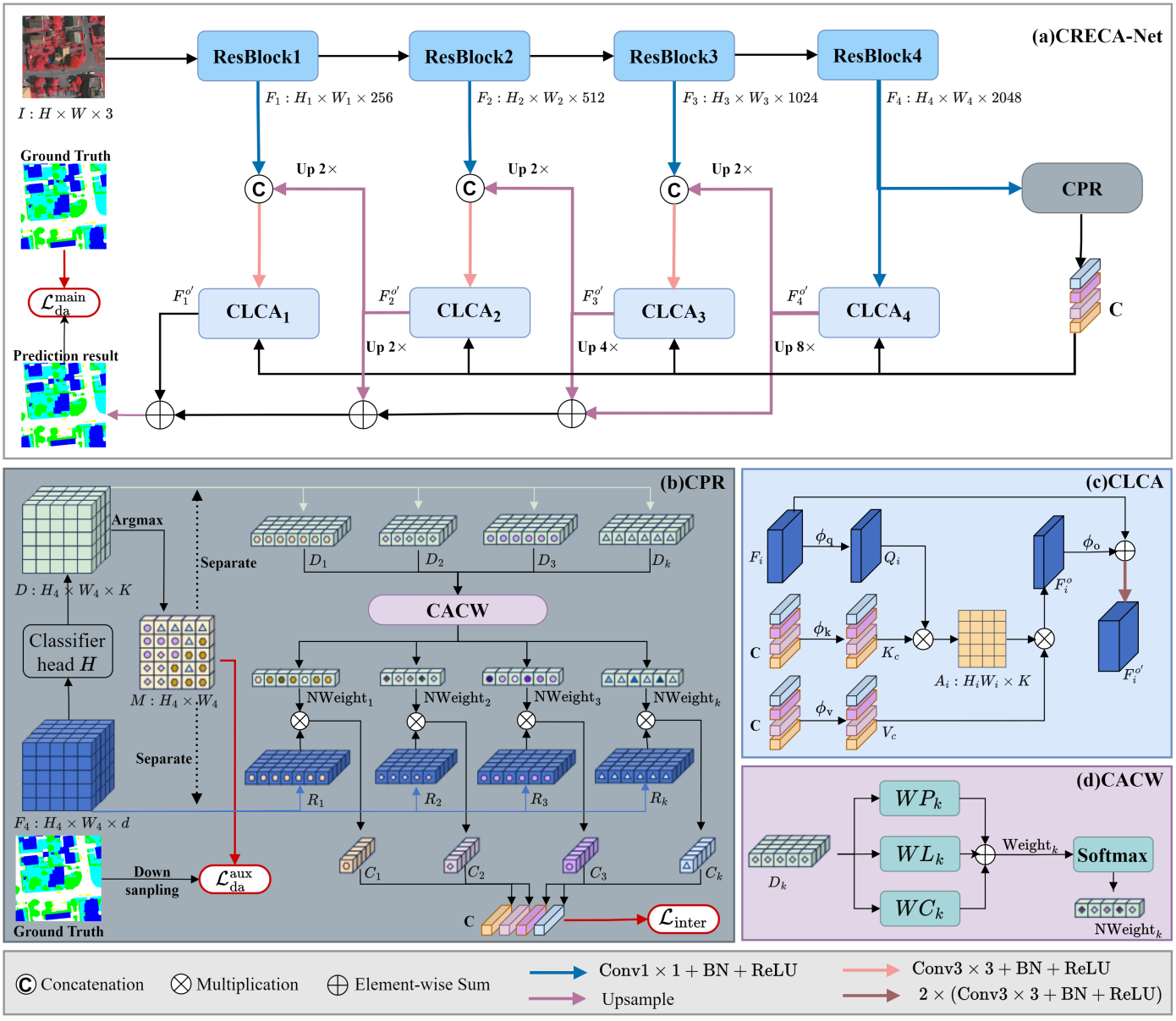



**Figure 3.** Overall architecture of CRECA-Net, with a ResNet-50 backbone as the encoder for hierarchical feature extraction, and a decoder comprising a class prototype refinement (CPR) module for generating discriminative class centers, and class-level context aggregation (CLCA) modules for modeling class-aware contextual relationships and reducing inter-class confusion.

### *3.2. Class Prototype Refinement Module*

High-resolution RS images often contain visually similar objects from different categories, such as trees vs. low vegetation, or buildings vs. impervious surfaces, resulting in low inter-class variance. Complex backgrounds further aggravate this issue, making it challenging for existing class prototype estimation methods to derive accurate and well-separated class prototypes. Since reliable prototypes are crucial for constructing a discriminative embedding space and guiding class-aware context reasoning, we introduce the CPR module to refine class representations from three complementary aspects: pixel selection, confidence-aware contribution weighting, and inter-class prototype separation. These strategies collectively improve the accuracy, representativeness, and discriminability of class prototypes, thus enabling subsequent CLCA modules to perform more precise class-aware context aggregation.

#### 3.2.1. Pixel Selection

Pixel selection improves the reliability of class prototype estimation by restricting the pixels used for prototype computation. Instead of estimating class prototypes through probability-weighted averaging over all pixels, as commonly adopted in prior work, we retain only those pixels predicted as class $k$. This selective mechanism reduces cross-class interference and ensures that each prototype is computed from semantically consistent feature regions.

As shown in Figure 3b, given the deepest feature map $F_4 \in \mathbb{R}^{H_4 \times W_4 \times d}$, where $H_4$ and $W_4$ denotes its spatial height and width, respectively, a classification head $H$ implemented using a $1 \times 1$ convolution projects $F_4$ into the class space, producing the pre-classification logits:

$$D = H(F_4), \quad D \in \mathbb{R}^{H_4 \times W_4 \times K} \tag{1}$$

where $K$ denotes the number of semantic categories. A coarse segmentation mask $M \in \mathbb{R}^{H_4 \times W_4}$ is then obtained by assigning each pixel to the class with the highest logit response along the category dimension. This mask is subsequently supervised using the proposed DA loss, as described in Section 3.4:

$$M(i,j) = \underset{c \in \{1,\ldots,K\}}{\arg\max}\, D(i,j,c) \tag{2}$$

where $M(i,j)$ denotes the predicted class label for pixel $(i,j)$. Based on $M$, the feature map $F_4$ is partitioned into class-specific regions $R_k$, each containing the pixel features predicted as class $k$:

$$R_k = \{F_4(i,j,:) \mid M(i,j) = k\}, \quad R_k \in \mathbb{R}^{N_k \times d} \tag{3}$$

where $N_k$ is the number of pixels assigned to class $k$. Similarly, the pre-classification logits $D$ are grouped into class-specific subsets:

$$D_k = \{D(i,j,:) \mid M(i,j) = k\}, \quad D_k \in \mathbb{R}^{N_k \times K} \tag{4}$$

This pixel grouping strategy ensures that each class prototype is computed from semantically coherent and prediction-consistent regions. Consequently, the estimated prototypes better capture the intrinsic characteristics of each category and are more robust to interference from visually similar but semantically distinct classes.

#### 3.2.2. Confidence-Aware Contribution Weighting

Although the pixel selection module confines the candidate set for class $k$ to the pixels predicted as belonging to class $k$, these pixels still vary in confidence and representativeness.

Treating them equally may bias prototype estimation. To address this, we propose a confidence-aware contribution weighting (CACW) strategy, which assigns each pixel a contribution score computed from three complementary confidence indicators: prediction probability, logit margin, and prediction certainty. These indicators are defined as follows.

***Prediction Probability Confidence:*** The predicted probability for the assigned class provides a direct estimate of model confidence for each pixel. Pixels with higher predicted probabilities are considered more reliable and therefore exert a stronger influence on the class prototype.

Given the pre-classification logits $D_k \in \mathbb{R}^{N_k \times K}$ (as defined above), the class probability distribution is obtained by applying the softmax operation along the category dimension:

$$P_k(i,c) = \frac{\exp(D_k(i,c))}{\sum_{c'=1}^{K} \exp(D_k(i,c'))}, \quad P_k \in \mathbb{R}^{N_k \times K} \tag{5}$$

Here, $P_k(i,c)$ denotes the predicted probability that the $i$-th pixel is classified into category $c$ (for $c = 1, \ldots, K$). The probability-based confidence vector for class $k$ is computed as:

$$\mathrm{WP}_k = \left[P_k(i,k)\right]_{i=1}^{N_k} \tag{6}$$

This indicator naturally down-weights low-confidence pixels while emphasizing more reliable pixels.

***Logit Margin Confidence:*** To complement the absolute confidence provided by probability values, we further incorporate the margin between the largest and second-largest logits. A larger margin indicates clearer separation between the top competing classes, reflecting stronger representativeness of the assigned class.

Let $\mathrm{Top}_i(x)$ denote the $i$-th largest element of a vector $x$. Given the pre-classification logits $D_k \in \mathbb{R}^{N_k \times K}$, the logit margin score for each pixel is computed as:

$$\mathrm{WL}_k = \left[\mathrm{Top}_1(D_k(i,:)) - \mathrm{Top}_2(D_k(i,:))\right]_{i=1}^{N_k} \tag{7}$$

Here, $D_k(i,:) \in \mathbb{R}^K$ denotes the logit vector of the $i$-th pixel.

This indicator complements the probability-based confidence by capturing relative certainty among competing classes, encouraging contributions from pixels with higher discriminativeness.

***Prediction Certainty Confidence:*** In addition to probability and logit margin, entropy provides a measure of predictive uncertainty. Prior studies on uncertainty estimation [64,65] have shown that greater entropy indicates greater prediction uncertainty and a higher likelihood of misclassification. Accordingly, pixels with lower entropy should contribute more to the prototype.

For each pixel $i$, the entropy is computed from its predicted probability distribution $P_k(i,:)$ as:

$$\mathrm{Ent}_k(i) = -\sum_{c=1}^{K} P_k(i,c) \log P_k(i,c) \tag{8}$$

where $i = 1, \ldots, N_k$ denotes the index of the pixels assigned to class $k$. To convert entropy into a confidence-like weight, we first normalize the entropy values by the maximum possible entropy and then invert them:

$$\mathrm{WC}_k = \left[1 - \frac{\mathrm{Ent}_k(i)}{\log K}\right]_{i=1}^{N_k} \tag{9}$$

where $\log K$ corresponds to the maximum entropy of a uniform $K$-class distribution. Pixels with lower entropy, indicating higher certainty, receive larger weights. This indicator

explicitly captures the overall uncertainty in predictions and provides a complementary perspective to both the probability and logit margin confidence measures.

***Confidence Fusion and Normalization:*** After obtaining the three confidence indicators, they are integrated to derive a contribution score for each pixel. Specifically, for pixels predicted as class $k$, the aggregated confidence is computed as the sum of the three confidence indicators:

$$\text{Weight}_k = \text{WP}_k + \text{WL}_k + \text{WC}_k \tag{10}$$

Here, $\text{Weight}_k \in \mathbb{R}^{N_k \times 1}$ represents the contribution scores of the $N_k$ pixels predicted as class $k$. To ensure comparability among pixels within the same class, these aggregated scores are normalized via a softmax function:

$$\text{NWeight}_k(i) = \frac{\exp(\text{Weight}_k(i))}{\sum_{i'=1}^{N_k} \exp(\text{Weight}_k(i'))}, \quad i = 1, 2, \ldots, N_k \tag{11}$$

The normalized scores $\text{NWeight}_k(i)$ quantify each pixel's relative importance in shaping the final class prototype.

***Weighted Prototype Computation:*** Based on the normalized contribution weights, the class prototype for class $k$ is computed as a weighted mean of its selected pixel features:

$$C_k = \sum_{i=1}^{N_k} \text{NWeight}_k(i) \cdot R_k(i, :) \tag{12}$$

Here, $R_k(i, :)$ denotes the feature vector of the $i$-th pixel in $R_k$. Repeating this process for all $K$ categories yields the full prototype matrix:

$$C = [C_1, C_2, \ldots, C_K]^\top \in \mathbb{R}^{K \times d} \tag{13}$$

where each row corresponds to a refined class representation. The complete computation procedure for class prototype generation is summarized in Algorithm 1.

**Algorithm 1** Class Prototype Generation

**Require:** Feature map $F_4 \in \mathbb{R}^{H_4 \times W_4 \times d}$,
Classifier Head $H(\cdot)$,
Number of categories $K$
**Ensure:** Refined class prototypes $C = [C_1, C_2, \ldots, C_K] \in \mathbb{R}^{K \times d}$

**Feature Projection:** Obtain the pre-classification logits $D$ according to Equation (1)
**Mask Generation:** Generate the coarse segmentation mask $M$ as defined in Equation (2)
**for** each class $k = 1$ to $K$ **do**
  Extract the feature subset $R_k$ and logit subset $D_k$ following Equations (3) and (4)
  Compute the probability distribution $P_k$ according to Equation (5)
  **for** each pixel $i = 1$ to $N_k$ **do**
    Compute the confidence indicators: prediction probability confidence $WP_k^{(i)}$, logit margin confidence $WL_k^{(i)}$, and prediction certainty confidence $WC_k^{(i)}$ as defined in Equations (6), (7), and (9)
    Compute aggregated confidence weight $\text{Weight}_k^{(i)}$ using Equation (10)
  **end for**
  Compute Normalize confidence weight $\text{NWeight}_k(i)$ according to Equation (11)
  Compute the refined class prototype $C_k$ with Equation (12)
**end for**
**Return:** Refined class prototypes $C$

This weighting scheme ensures that highly confident and semantically representative pixels contribute more strongly, while uncertain or noisy samples contribute less. As a result, the generated prototypes are more robust and discriminative, providing a reliable basis for subsequent class-aware context modeling.

#### 3.2.3. Inter-Class Prototype Separation Loss

Although the CPR module improves the reliability of individual class prototypes, ensuring sufficient separability among different categories remains essential for robust semantic segmentation, particularly in RS imagery, where semantically distinct classes often exhibit highly similar appearances (e.g., tree vs. low vegetation, building vs. impervious surface). However, existing class-level context aggregation methods typically use the derived class centers directly for context modeling without enforcing explicit separation, which may lead to overly clustered prototypes and weakened class discriminability.

To mitigate this issue, we introduce an inter-class prototype separation loss, which explicitly encourages larger angular margins between class centers by penalizing excessive similarity. This constraint prevents feature-space collapse and promotes the formation of more discriminative and well-separated class embeddings.

***Definition of Inter-class Similarity Penalty:*** To explicitly penalize overly close class centers, we define a penalty term based on the cosine similarity between class centers:

$$P_{\text{inter}}(C_p, C_q) = \max(0, \text{Sim}(C_p, C_q) - \beta) \tag{14}$$

where $C_p, C_q \in \mathbb{R}^d$ denote the refined prototypes for classes $p$ and $q$, respectively. The cosine similarity is defined as:

$$\text{Sim}(C_p, C_q) = \frac{C_p \cdot C_q}{\|C_p\|_2 \cdot \|C_q\|_2} \tag{15}$$

The threshold $\beta \in [0, 1]$ sets the maximum allowable similarity between two class centers. Only prototype pairs whose similarity exceeds $\beta$ incur a penalty, thereby pushing overly close prototypes farther apart.

***Inter-class Prototype Separation Loss:*** The loss evaluates pairwise similarities among all class prototypes and penalizes those pairs whose similarity exceeds the threshold $\beta$:

$$\mathcal{L}_{\text{inter}} = \frac{1}{K} \sum_{p=1}^{K} \sum_{\substack{q=1 \\ q \neq p}}^{K} P_{\text{inter}}(C_p, C_q) \tag{16}$$

Minimizing $\mathcal{L}_{\text{inter}}$ effectively enlarges the angular margins among class centers, enhances inter-class separability, and produces clearer semantic boundaries particularly in visually ambiguous regions. This loss operates solely on learned prototypes, requires no additional supervision, and introduces negligible computational overhead, making it practical for end-to-end training.

It is noteworthy that the class centers are computed exclusively from the deepest feature map $F_4$, which possesses a larger receptive field and captures richer semantic information [15]. This enables more accurate coarse segmentation masks and more reliable prototype initialization. The coarse masks are then refined under ground-truth supervision.

Unlike previous works [18,61] that compute prototypes by aggregating all pixel features via probability-based weighting, our method forms each class representation using only the pixels predicted as belonging to that class, preventing cross-class interference. Furthermore, each pixel's contribution is adaptively modulated by three complementary confidence indicators: **(1) absolute confidence measured by the predicted probability;**

**(2) relative confidence quantified by the logit margin; and (3) overall prediction certainty indicated by entropy.** By combining these indicators, high-confidence and semantically representative pixels receive larger contribution weights, producing prototypes that are robust and discriminative. Furthermore, the proposed inter-class prototype separation loss enforces angular margins among prototypes, effectively alleviating the prevalent problem of low inter-class variance in high-resolution RS imagery.

### *3.3. Class-Level Context Aggregation Module*

Self-attention-based pixel-level relational modeling has been extensively employed in RS image segmentation owing to its strong capability in capturing long-range dependencies. However, computing pairwise similarities among all pixels incurs a quadratic complexity of $O(n^2d)$, where $n$ and $d$ denote the number of pixels and feature dimension, respectively. For high-resolution RS images, this leads to prohibitive computational and memory costs. Additionally, aggregating information from all pixels may incorporate features from visually similar but semantically irrelevant regions, especially in scenes containing complex backgrounds or visually similar land-cover categories. These factors make pixel-to-pixel affinities unreliable and ultimately degrade segmentation accuracy.

To address these limitations, we introduce the CLCA module, which replaces dense pixel-to-pixel interactions with pixel-to-class prototype correlations. Unlike standard self-attention that captures dependencies only at the pixel level without semantic guidance, CLCA leverages class prototypes to guide attention and aggregates semantic information across all pixels predicted as belonging to the same class. This design reduces the computational complexity of pixel-level self-attention from $O(n^2d)$ to $O(nKd)$ (where $K$ denotes the number of semantic classes), while injecting explicit semantic guidance into the attention mechanism. Consequently, CLCA allows the network to capture long-range contextual cues while suppressing activations from irrelevant or confusing regions, thus enhancing the discriminability of pixel representations. Compared with generic cross-attention variants, which compute attention between two feature maps without explicit class-level guidance, CLCA explicitly incorporates class-level prototypes to guide semantic aggregation, which is not easily achievable with standard self-attention or cross-attention alone.

As illustrated in Figure 3a, the CLCA module operates at each decoder stage and takes two inputs **(1)** the stage-wise feature map $F_i \in \mathbb{R}^{H_i \times W_i \times d}$ and **(2)** the refined class prototypes $C \in \mathbb{R}^{K \times d}$ generated by the CPR module. Following a cross-attention paradigm, $F_i$ is projected into a query matrix $Q_i = \phi_q(F_i) \in \mathbb{R}^{(H_iW_i) \times d}$, while the class prototypes are mapped to a key matrix $K_c = \phi_k(C) \in \mathbb{R}^{K \times d}$ and a value matrix $V_c = \phi_v(C) \in \mathbb{R}^{K \times d}$. Here, $\phi_q$, $\phi_k$, and $\phi_v$ represent learnable $1 \times 1$ convolutions followed by batch normalization and ReLU activation. The pixel-to-class attention map $A_i \in \mathbb{R}^{(H_iW_i) \times K}$ is computed using a scaled dot-product followed by a row-wise softmax:

$$A_i = \text{Softmax}\left(\frac{Q_i K_c^\top}{\sqrt{d}}\right) \tag{17}$$

Based on $A_i$, the aggregated feature $F_i^o$ is obtained by:

$$F_i^o = A_i V_c \in \mathbb{R}^{(H_iW_i) \times d} \tag{18}$$

The resulting feature is first reshaped to match the spatial dimensions of the original feature $F_i$, then passed through an output mapping $\phi_o$, and fused with the original feature map $F_i$ via channel concatenation (denoted as $\oplus$). The fused feature is subsequently refined using two $3 \times 3$ convolution layers with batch normalization and ReLU activation ($\text{Conv}_{3\times3}$) to produce the refined aggregated feature $F_i^{o'}$:

$$F_i^{o'} = \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{3\times3}(F_i \oplus \phi_{\mathrm{o}}(F_i^o))) \tag{19}$$

The above procedure defines a single CLCA module, which outputs $F_i^{o'}$ for stage $i$.

The CLCA module is applied in a top-down manner to progressively refine decoder features. For decoder stages $i = 1, 2, 3$, the current-stage feature is first updated by fusing it with the upsampled refined aggregated feature from the deeper stage:

$$F_i = \mathrm{Conv}_{3\times3}\big(\mathrm{Upsample}(F_{i+1}^{o'}) \oplus F_i\big) \tag{20}$$

where $\mathrm{Upsample}(\cdot)$ denotes $2\times$ bilinear upsampling. The updated feature $F_i$ is then fed into the CLCA module at the current stage to generate the corresponding refined aggregated output:

$$F_i^{o'} = \mathrm{CLCA_i}(F_i, C) \tag{21}$$

This hierarchical refinement strategy progressively strengthens semantic consistency while preserving spatial detail. By repeatedly incorporating class-aware global cues from deep layers into fine-grained spatial structures of shallower layers, CLCA improves segmentation coherence and boundary localization.

### *3.4. Loss Function*

***Difficulty-aware Loss:*** Semantic segmentation in high-resolution RS imagery presents substantial variability in pixel-wise classification difficulty due to complex backgrounds, large intra-class diversity, and small inter-class differences. Pixels located at object boundaries, in occluded areas, or belonging to visually similar categories are inherently harder to classify than those in homogeneous regions. However, the standard CE loss assigns equal importance to all pixels, implicitly assuming uniform difficulty. This mismatch leads to a training imbalance: **(1)** Early in training, gradients are dominated by abundant easy samples, impeding effective learning of hard samples; **(2)** Later in training, although hard samples generate larger errors, their small quantity limits their impact on optimization. Consequently, hard regions receive insufficient supervision, ultimately degrading segmentation performance.

These limitations arise from the static nature of the standard CE loss, which cannot adaptively adjust sample importance as training evolves. To address this issue, we propose a DA loss that incorporates a dynamic difficulty estimation mechanism and an adaptive loss scheduling strategy, which stabilizes early-stage optimization while progressively emphasizing challenging samples as training proceeds.

***Dynamic Difficulty Estimation Mechanism:*** To estimate pixel-wise classification difficulty, we adopt a normalized, confidence-driven weighting scheme inspired by focal loss [66]. The difficulty weight for the $i$-th pixel is defined as:

$$W_i = \frac{(1-p_i)^\gamma}{\sum_{i'=1}^{N}(1-p_{i'})^\gamma} \tag{22}$$

where $p_i \in [0, 1]$ denotes the predicted probability of the ground-truth class, $N$ is the total number of pixels, and $\gamma \geq 0$ is a focusing parameter controlling the emphasis on hard samples. The resulting difficulty-weighted loss is expressed as:

$$\mathcal{L}_{\mathrm{weight}} = \sum_{i=1}^{N} W_i \mathcal{L}_{\mathrm{ce}}^i = -\sum_{i=1}^{N} W_i \log(p_i) \tag{23}$$

The normalization term $\sum_{i=1}^{N} W_i = 1$ preserves gradient stability by keeping the loss magnitude comparable to the standard CE loss. This difficulty-aware redistribution of pixel-

wise loss within an image assigns greater weights to lower confidence pixels, encouraging the model to focus on hard regions.

***Adaptive Loss Scheduling Mechanism:*** Because model predictions are unreliable at early training stages, directly applying difficulty weighting may negatively impact optimization stability. To address this, we introduce an adaptive loss scheduling mechanism based on an annealing function that gradually shifts the optimization focus from easy samples to hard samples. The overall DA loss is defined as:

$$\mathcal{L}_{\text{da}} = \big(1 - \lambda(t)\big)\mathcal{L}_{\text{ce}} + \lambda(t)\mathcal{L}_{\text{weight}} \tag{24}$$

where $\mathcal{L}_{\text{ce}}$ denotes the standard CE loss, and $\lambda(t) \in [0, 1]$ is a monotonically increasing annealing function of training step $t$. As $\lambda(t)$ increases, the loss smoothly transitions from uniform CE supervision to difficulty-aware weighting, which stabilizes early-stage optimization while progressively increasing the model's focus on hard regions. Representative annealing strategies are listed in Table 1, where $T_{\text{step}}$ denotes the annealing step count marking the end of the annealing period; its selection is discussed in Section 5.3.3.

**Table 1.** Different forms of annealing functions.

| Annealing Function | Hyperparameter | Mathematical Expression |
|---|---|---|
| Linear | $T_{\text{step}}$ | $\lambda(t) = \dfrac{t}{T_{\text{step}}}$ |
| Polynomial | $T_{\text{step}}$, decay_factor | $\lambda(t) = \left(\dfrac{t}{T_{\text{step}}}\right)^{\text{decay_factor}}$ |
| Cosine | $T_{\text{step}}$ | $\lambda(t) = 0.5 \cdot \left(1 - \cos\left(\dfrac{t}{T_{\text{step}}} \cdot \pi\right)\right)$ |

***Overall loss:*** The total training loss combines the DA loss with the inter-class prototype separation loss to jointly optimize pixel-level learning and inter-class discrimination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{da}}^{\text{main}} + \lambda_{\text{aux}}\mathcal{L}_{\text{da}}^{\text{aux}} + \mathcal{L}_{\text{inter}} \tag{25}$$

where $\mathcal{L}_{\text{da}}^{\text{main}}$ and $\mathcal{L}_{\text{da}}^{\text{aux}}$ represent the DA loss applied to the final prediction and the auxiliary coarse mask $M$, respectively, and $\mathcal{L}_{\text{inter}}$ is the inter-class prototype separation loss introduced in Section 3.2.3. The coefficient $\lambda_{\text{aux}}$ controls the relative contribution of the auxiliary branch.

## 4. Experiment Settings

### *4.1. Datasets*

The performance of the proposed CRECA-Net was evaluated on two widely used benchmark datasets for RS image segmentation: the ISPRS Vaihingen [67] and ISPRS Potsdam [67] datasets.

***ISPRS Potsdam Dataset:*** The Potsdam dataset consists of 38 orthophoto tiles, each sized at 6000 × 6000 pixels with a ground sampling distance (GSD) of 5 cm. Each tile provides four multispectral bands, near-infrared (NIR), red (R), green (G), and blue (B), along with corresponding digital surface model (DSM) and normalized DSM (NDSM) data. In our experiments, only the RGB bands were used, excluding DSM and NDSM data. The dataset contains dense annotations for six land-cover categories: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. Following established protocols, 23 tiles with IDs 2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_8, 7_9, 7_11, and 7_12 were allocated for training (note that tile 7_10 was excluded due to annotation errors), while the remaining 14 tiles

were reserved for testing. Each tile was further divided into fixed-size, non-overlapping patches of 1024 × 1024 pixels for training and evaluation.

***ISPRS Vaihingen Dataset:*** The Vaihingen dataset comprises 33 orthophoto tiles with an average resolution of 2494 × 2494 pixels and a GSD of 9 cm. Each tile includes three spectral bands, near-infrared (NIR), red (R), and green (G), together with corresponding DSM and NDSM data. The land-cover categories are the same six classes as defined for the Potsdam dataset. Following prior work, 15 tiles (IDs: 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 32, 34, and 37) were used for training, and the remaining 17 tiles were used for testing. All tiles were cropped into 1024 × 1024 patches using a sliding window with a 512-pixel stride, ensuring uniform patch dimensions across the dataset.

### *4.2. Evaluation Metrics*

To comprehensively evaluate the performance of CRECA-Net and ensure fair comparisons with existing methods, three widely recognized evaluation metrics were used: Overall Accuracy (OA), mean Intersection over Union (mIoU), and F1 score (F1).

***Overall Accuracy (OA):*** This metric measures the proportion of correctly classified pixels among all pixels and is defined as:

$$\mathrm{OA} = \frac{\sum_{k=1}^{K} \mathrm{TP}_k}{\sum_{k=1}^{K} (\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{TN}_k + \mathrm{FN}_k)} \tag{26}$$

where $K$ denotes the total number of categories, and $\mathrm{TP}_k$, $\mathrm{FP}_k$, $\mathrm{TN}_k$, and $\mathrm{FN}_k$ represent the true positives, false positives, true negatives, and false negatives for the $k$-th class, respectively.

***Mean Intersection over Union (mIoU):*** This metric measures the average overlap between the predicted and ground-truth regions across all categories, defined as the ratio of their intersection to their union:

$$\mathrm{mIoU} = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{27}$$

***F1 Score (F1):*** The F1 score for class $k$ is the harmonic mean of precision and recall. The overall F1 score (mean F1) is computed as the mean of $F1_k$ over all classes.

$$\mathrm{F1_k} = 2 \times \frac{\mathrm{Precision_k} \times \mathrm{Recall_k}}{\mathrm{Precision_k} + \mathrm{Recall_k}} \tag{28}$$

For class $k$, $\mathrm{Precision_k}$ denotes the proportion of true positives among all predicted positives, and $\mathrm{Recall_k}$ is the ratio of true positives to the total number of actual positives, defined as:

$$\mathrm{Precision_k} = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k} \tag{29}$$

$$\mathrm{Recall_k} = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{30}$$

### *4.3. Implementation Details*

To ensure consistent and fair comparisons, all experiments were implemented using the PyTorch (version 2.0.0) framework and conducted on a single NVIDIA RTX A40 GPU (NVIDIA Corporation, Santa Clara, CA, USA). We employed the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01 and a weight decay of 0.0001. A polynomial decay policy [41] was adopted to dynamically adjust the learning rate according to the schedule $(1 - \frac{\mathrm{iter}}{\mathrm{max_iter}})^{power}$, where the *power* was set to 0.9. The mini-batch size was 8, with 80 epochs on the Potsdam dataset and 150 epochs on the Vaihingen dataset. The

introduced hyperparameters, including the similarity threshold $\beta$, focusing parameter $\gamma$, and auxiliary loss coefficient $\lambda_{aux}$, were empirically determined through ablation studies and set to 0.125, 1.0, and 0.8, respectively, and kept fixed across all experiments. To enhance generalization, several data augmentation techniques were applied during training, including random cropping (crop size of 512 × 512), random scaling (scale factors of 0.5, 0.75, 1.0, 1.25, 1.5), random horizontal/vertical flipping, random rotation, and random Gaussian blur. During testing, multi-scale inference, flipping, and rotation were applied as test-time augmentation to improve robustness.

## 5. Experiment Results and Analysis

This section presents a comparative evaluation of CRECA-Net against fifteen representative semantic segmentation models. The compared methods span various architectural paradigms, including **(1)** classical convolutional neural network-based frameworks FCN [36]; **(2)** spatial context aggregation models PSPNet [15] and DeepLabV3+ [16]; **(3)** pixel-level relational modeling methods, such as DANet [21], SegFormer [23], MAResU-Net [47], BANet [55], DC-Swin [25], UNetFormer [49], AMMUNet [53], SLCNet [50], MSGC-Net [52], and CGGLNet [51]; and **(4)** class-level context modeling approaches comprising OCRNet [18] and LOGCAN++ [61]. All baseline implementations follow the optimal configurations reported in their original publications.

### *5.1. Comparison with State-of-the-Art Methods on the Potsdam Dataset*

#### 5.1.1. Quantitative Analysis

To evaluate the effectiveness of CRECA-Net, we conduct a quantitative comparison with state-of-the-art methods on the Potsdam dataset. As shown in Table 2, CRECA-Net achieves the highest OA (92.00%), MeanF1 (93.18%), and mIoU (87.44%), consistently outperforming all competing approaches. Specifically, compared with spatial context aggregation models, CRECA-Net yields substantial mIoU improvements of 10.77% and 8.85% over PSPNet and DeepLabV3+, respectively. It also surpasses recent pixel-level relational modeling methods, achieving mIoU gains of 1.16%, 1.30% and 0.41% over UNetFormer, MSGCNet, and CGGLNet, respectively. Furthermore, compared with LOGCAN++, a state-of-the-art class-level context modeling framework specifically designed for RS image segmentation, CRECA-Net achieves a notable 1.08% improvement in mIoU. This consistent superiority can be primarily attributed to the proposed CPR module, which produces more accurate and discriminative class centers. These refined class centers enable more effective class-level context modeling, thereby improving feature representations and segmentation performance. Overall, these results highlight the critical role of high-quality class prototypes in enhancing class-level context modeling for RS image segmentation.

#### 5.1.2. Computational Complexity Analysis

To quantitatively assess computational complexity, we compare CRECA-Net with representative baselines with respect to model parameters (Params), computational cost (FLOPs), and inference speed (FPS). All evaluations were conducted on a single NVIDIA RTX A40 GPU with an input resolution of 1024 × 1024, and the results are summarized in Table 2. Among models using the same backbone, CRECA-Net achieves superior segmentation performance while maintaining comparable or lower computational cost. For instance, it outperforms LogCAN++ by 1.08% in mIoU with similar FLOPs and Params, and surpasses CGGLNet by 0.41% in mIoU while using 38.36G fewer FLOPs and 5.39M fewer Params. This efficiency can be attributed to two design aspects: (1) the CPR module and the DA loss introduce no learnable parameters; and (2) the CLCA module performs context aggregation at the class level, which requires significantly fewer parameters

than spatial or pixel-level relational modeling approaches. These findings indicate that CRECA-Net achieves a favorable trade-off between segmentation accuracy and computational cost, underscoring the advantage of class-level context modeling in high-resolution RS segmentation.

**Table 2.** Quantitative comparison with state-of-the-art methods on the ISPRS Potsdam dataset. Per-class segmentation performance is reported using IoU scores, with the best value highlighted in bold. Params, FLOPs, and FPS denote the number of model parameters, computational cost, and inference speed, respectively.

| Method | Backbone | Imp. Surf | Building | Lowveg | Tree | Car | OA | Mean F1 | mIoU | FLOPs (G) | Params (M) | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FCN [36] | ResNet50 | 80.08 | 86.78 | 68.80 | 72.16 | 68.42 | 83.31 | 85.69 | 75.25 | 28.77 | 49.49 | 53.28 |
| PSPNet [15] | ResNet50-D8 | 81.02 | 87.27 | 68.71 | 71.47 | 74.90 | 83.54 | 86.64 | 76.67 | 178.64 | 48.97 | 13.93 |
| DeepLabV3+ [16] | ResNet50-D8 | 83.53 | 89.44 | 70.89 | 75.14 | 73.93 | 85.15 | 87.84 | 78.59 | 176.51 | 43.58 | 14.73 |
| DANet [21] | ResNet50-D8 | 81.22 | 87.44 | 69.19 | 73.25 | 72.59 | 83.85 | 86.68 | 76.73 | 211.04 | 49.82 | 12.37 |
| OCRNet [18] | ResNet50-D8 | 82.00 | 87.54 | 69.51 | 73.63 | 71.57 | 84.03 | 86.74 | 76.85 | 153.07 | 36.51 | 19.58 |
| SegFormer [23] | MiT-B1 | 82.30 | 89.17 | 70.45 | 74.58 | 71.81 | 84.70 | 87.25 | 77.66 | 15.42 | 13.68 | 31.25 |
| BANet [55] | ResT-Lite | 86.95 | 91.51 | 76.29 | 79.31 | 90.16 | 90.42 | 91.69 | 84.84 | 15.15 | 12.73 | 10.13 |
| MAResU-Net [47] | ResNet34 | 86.96 | 92.23 | 76.28 | 79.53 | 91.34 | 90.52 | 91.92 | 85.27 | 28.76 | 26.28 | 45.19 |
| DC-Swin [25] | Swin-S | 88.20 | 92.78 | 77.91 | 79.00 | 88.94 | 91.10 | 92.00 | 85.37 | 72.18 | 66.95 | 15.16 |
| UNetFormer [49] | ResNet18 | 88.62 | 93.63 | 77.85 | 79.27 | 92.02 | 91.35 | 92.50 | 86.28 | 11.76 | 11.73 | 85.51 |
| AMMUNet [53] | ResNet50 | 83.72 | 90.74 | 72.02 | 75.92 | 73.06 | 85.58 | 88.15 | 79.09 | 52.51 | 34.71 | 19.06 |
| SLCNet [50] | ResNet50 | 85.34 | 90.98 | 76.07 | 79.02 | 91.76 | 89.81 | 91.55 | 84.63 | 315.47 | 187.15 | 6.18 |
| MSGCNet [52] | ResNet34 | 88.25 | 92.86 | 77.56 | 79.44 | 92.61 | 91.11 | 92.42 | 86.14 | 29.11 | 27.61 | 43.27 |
| LOGCAN++ [61] | ResNet50 | 88.77 | 93.60 | 77.27 | 79.89 | 92.28 | 91.34 | 92.55 | 86.36 | 50.18 | 31.03 | 16.30 |
| CGGLNet [51] | ResNet50 | 88.92 | 94.04 | 78.42 | 80.40 | 93.35 | 91.71 | 92.93 | 87.03 | 87.94 | 36.89 | 11.09 |
| Ours | ResNet50 | **89.57** | **94.25** | **79.00** | **81.07** | 93.30 | **92.00** | **93.18** | **87.44** | 49.48 | 31.00 | 29.10 |

#### 5.1.3. Qualitative Analysis

To qualitatively evaluate segmentation performance, Figure 4 compares the visual results of CRECA-Net with several representative methods on typical samples from the Potsdam dataset. As illustrated, our model produces clearer and more precise segmentation maps, especially in complex or visually ambiguous regions. For instance, in the first example, CRECA-Net accurately delineates buildings from adjacent impervious surfaces while reducing misclassification in surrounding low vegetation areas. In the second case, our model effectively distinguishes intertwined low vegetation and tree regions, demonstrating stronger class discriminability. The third image further highlights CRECA-Net's robustness in fine-grained object segmentation. Collectively, these results indicate that CRECA-Net delivers higher recognition accuracy and improved spatial coherence compared to competing approaches.

Furthermore, to gain deeper insight into the feature learning capability of different models, we employ t-SNE [68] to visualize the high-level features extracted from the final layers of UNetFormer, LOGCAN++, and CRECA-Net on the Potsdam test set, as depicted in Figure 5. The visualization shows that in UNetFormer and LOGCAN++, the feature embeddings of the "building" and "impervious surface" classes remain relatively close, indicating limited inter-class separability. In contrast, CRECA-Net forms a distinctly separated cluster for the "building" class, which is distant from the "impervious surface" cluster. This pronounced separation demonstrates CRECA-Net's ability to enhance feature discriminability and refine inter-class boundaries in the embedding space.

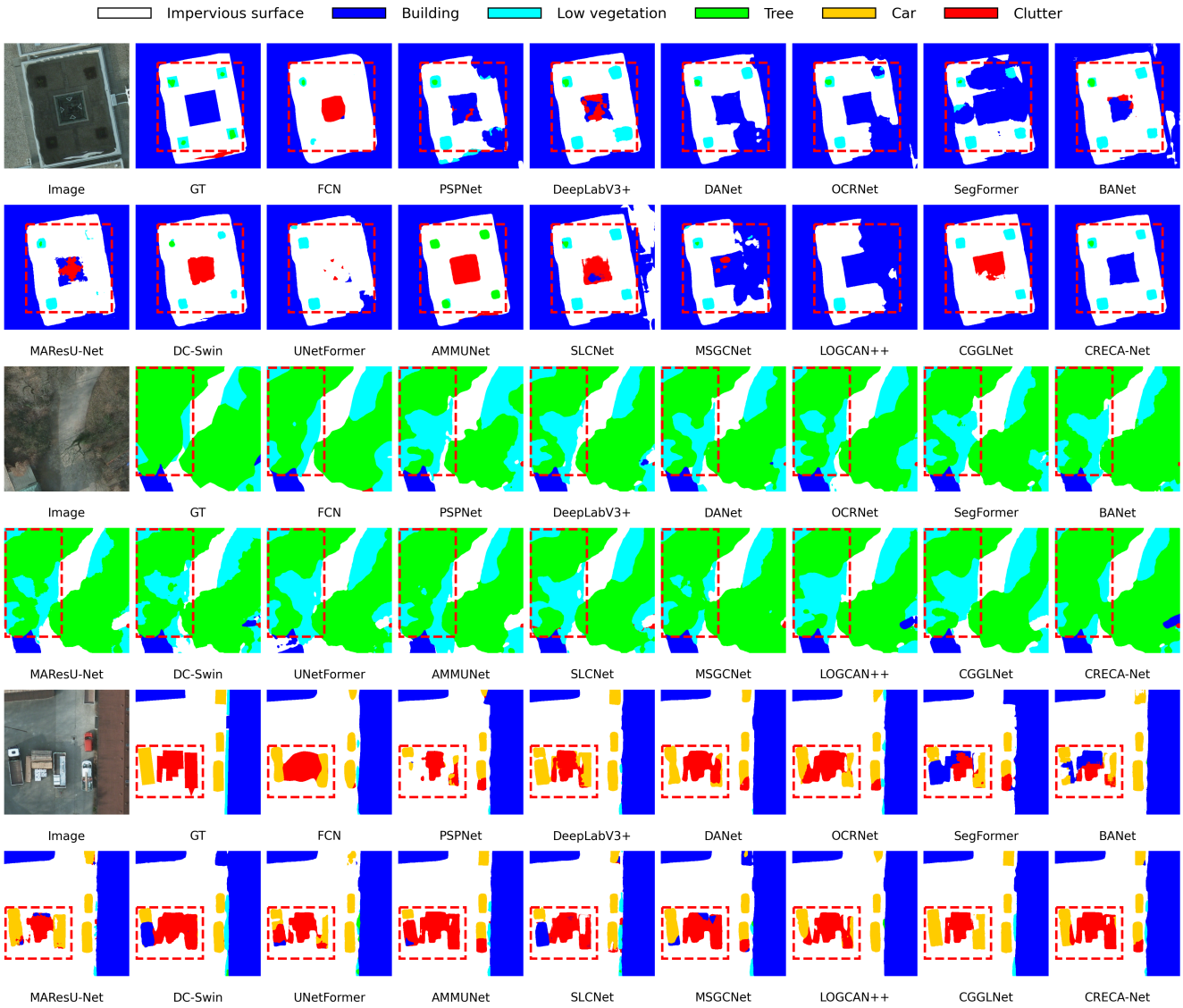



**Figure 4.** Qualitative comparison of different models on the Potsdam test set. Red dashed boxes indicate areas of focus.

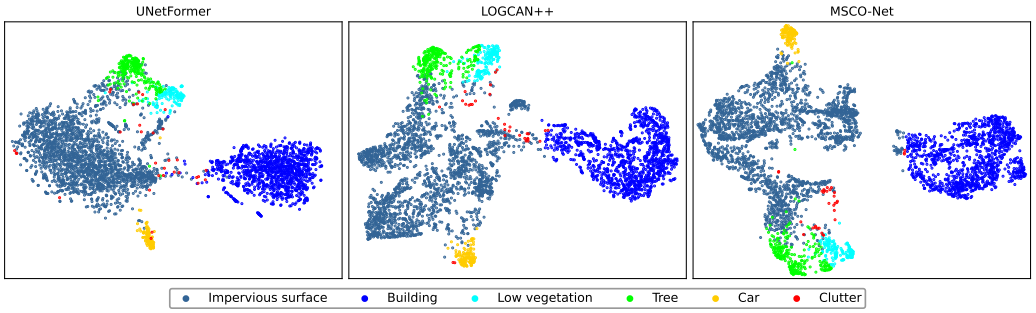



**Figure 5.** t-SNE visualizations of the last-layer feature embeddings extracted from UNetFormer, LOGCAN++, and the proposed CRECA-Net on the Potsdam test.

### *5.2. Comparison with State-of-the-Art Methods on the Vaihingen Dataset*

#### 5.2.1. Quantitative Analysis

To further evaluate the generalization capability of CRECA-Net, we conduct a quantitative comparison on the Vaihingen dataset, as presented in Table 3. CRECA-Net consistently outperforms all baseline methods across all metrics, further validating its effectiveness

and generalization capability. In terms of mIoU, CRECA-Net achieves improvements of 8.29% and 8.20% over the classical spatial context aggregation models PSPNet and DeepLabV3+, respectively. Compared with pixel-level relational modeling methods, it obtains mIoU gains of 1.00%, 1.01%, and 0.88% over UNetFormer, SLCNet, and CGGLNet, respectively. Furthermore, CRECA-Net surpasses LOGCAN++, a state-of-the-art class-level context modeling method for RS image segmentation, by 0.38% in mIoU. These improvements primarily stem from the proposed CPR module, which refines class centers and facilitates more reliable class-level context aggregation. Collectively, these results underscore the superior performance of CRECA-Net across different context modeling mechanisms and demonstrate its robustness and adaptability in diverse RS scenarios.

**Table 3.** Quantitative comparison with state-of-the-art methods on the ISPRS Vaihingen dataset. Per-class segmentation performance is reported using IoU scores, and the best value is highlighted in bold.

| Method | Backbone | Imp. Surf | Building | Lowveg | Tree | Car | OA | Mean F1 | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| FCN [36] | ResNet50 | 88.38 | 85.77 | 66.61 | 74.78 | 47.50 | 92.87 | 83.22 | 72.61 |
| PSPNet [15] | ResNet50-D8 | 89.69 | 86.83 | 67.23 | 74.87 | 63.22 | 93.52 | 86.21 | 76.37 |
| DeepLabV3+ [16] | ResNet50-D8 | 89.28 | 86.99 | 66.89 | 75.23 | 63.89 | 93.43 | 86.27 | 76.46 |
| DANet [21] | ResNet50-D8 | 89.43 | 86.48 | 66.69 | 74.60 | 65.10 | 93.31 | 86.30 | 76.46 |
| OCRNet [18] | ResNet50-D8 | 89.21 | 86.14 | 66.83 | 75.41 | 62.64 | 93.30 | 86.00 | 76.05 |
| SegFormer [23] | MiT-B1 | 89.45 | 85.64 | 66.38 | 74.82 | 60.24 | 93.16 | 85.46 | 75.31 |
| BANet [55] | ResT-Lite | 92.96 | 90.59 | 71.28 | 80.70 | 73.55 | 92.58 | 89.75 | 81.82 |
| MAResU-Net [47] | ResNet34 | 93.51 | 91.03 | 73.14 | 81.83 | 74.46 | 93.07 | 90.36 | 82.79 |
| DC-Swin [25] | Swin-S | 93.84 | 92.03 | 73.45 | 81.91 | 70.39 | 93.31 | 90.01 | 82.33 |
| UNetFormer [49] | ResNet18 | 93.58 | 91.34 | 73.48 | 81.98 | 77.90 | 93.20 | 90.91 | 83.66 |
| AMMUNet [53] | ResNet50 | 89.69 | 86.37 | 65.73 | 75.10 | 61.21 | 93.24 | 85.66 | 75.62 |
| SLCNet [50] | ResNet50 | 93.64 | 90.70 | 73.49 | 82.16 | 78.26 | 93.15 | 90.91 | 83.65 |
| MSGCNet [52] | ResNet34 | 93.22 | 90.57 | 73.06 | 82.22 | 73.12 | 92.96 | 90.14 | 82.44 |
| LOGCAN++ [61] | ResNet50 | 93.95 | 91.81 | 74.59 | 82.57 | 78.48 | 93.53 | 91.29 | 84.28 |
| CGGLNet [51] | ResNet50 | 93.56 | 91.59 | 73.09 | 81.41 | 79.25 | 93.15 | 90.98 | 83.78 |
| Ours | ResNet50 | **94.08** | **92.27** | **74.85** | **82.81** | **79.27** | **93.65** | **91.52** | **84.66** |

#### 5.2.2. Qualitative Analysis

The visual segmentation performance of CRECA-Net is further compared with several representative models on the Vaihingen test set, as illustrated in Figure 6. Visual inspection of challenging samples shows that CRECA-Net delivers more accurate and spatially coherent segmentation results. Specifically, in the first sample, our method successfully detects all vehicles while significantly reducing both missed and false detections. In the second case, it yields more complete building extraction with smoother and more continuous boundaries. Remarkably, in the third example, where shadow occlusion often causes trees to be misclassified as low vegetation, CRECA-Net effectively identifies shadowed trees, thus mitigating confusion between these visually similar categories. These visual results collectively demonstrate that CRECA-Net excels at preserving structural integrity and alleviating inter-class ambiguity in complex RS scenarios.

Likewise, we use t-SNE to visualize the high-level feature distributions of different models on the Vaihingen test set, as shown in Figure 7. Consistent with observations on the Potsdam dataset, the feature embeddings of UNetFormer and LOGCAN++ exhibit relatively overlapping clusters for the “building”,“tree”, and “low vegetation” classes. In contrast, CRECA-Net produces a clearly isolated cluster for the “building” category, which is well separated from the “tree” and “low vegetation” clusters. These observations indicate that CRECA-Net possesses strong feature discriminability and demonstrates robust across-dataset generalization.

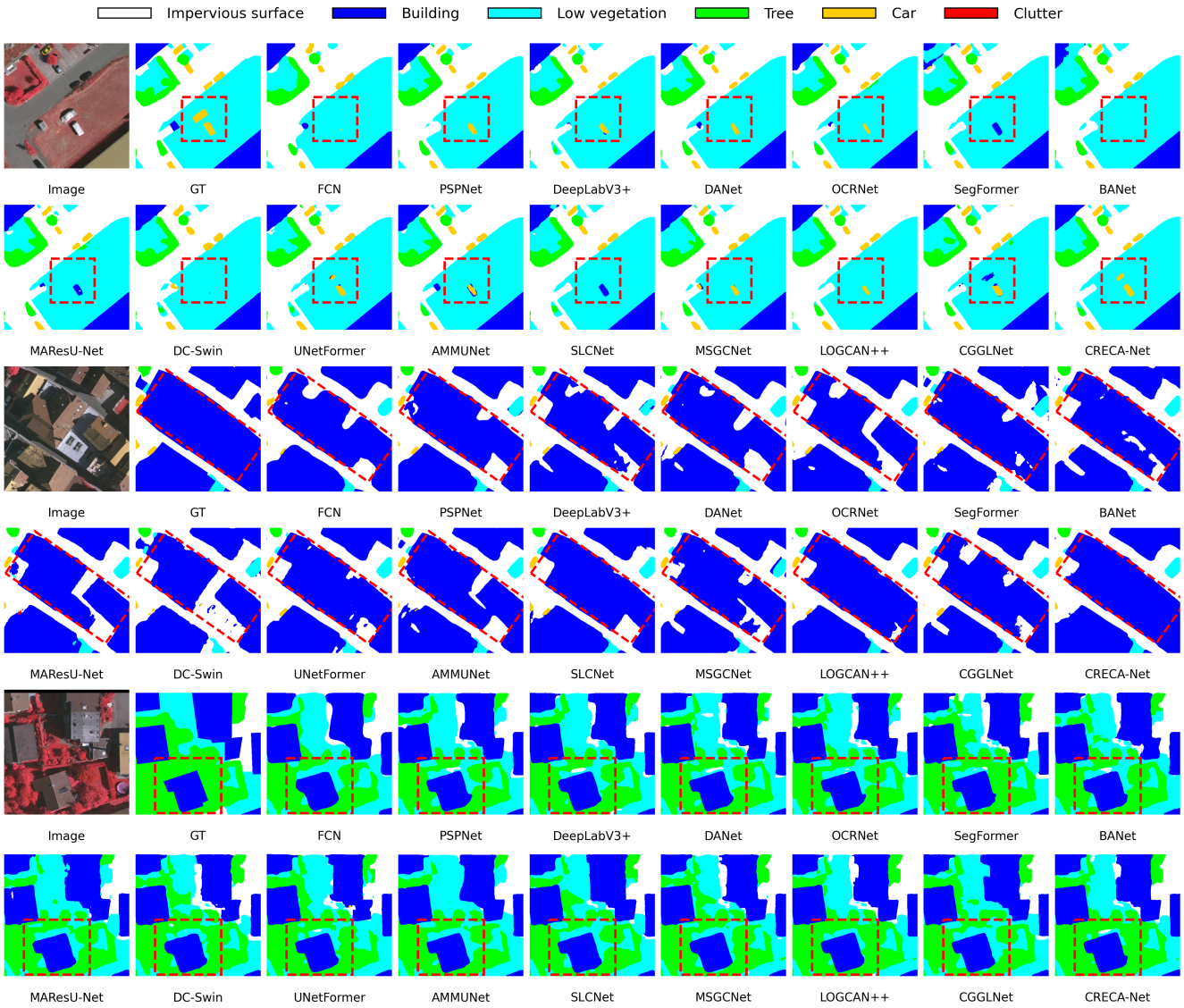



**Figure 6.** Qualitative comparison of different models on the Vaihingen test set. Red dashed boxes indicate areas of focus.

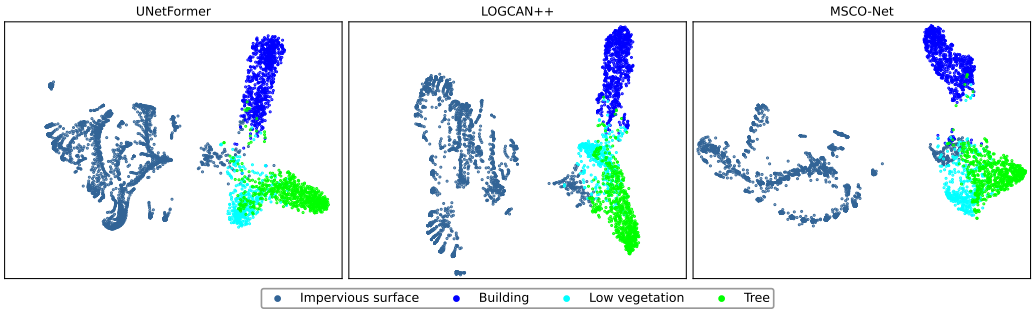



**Figure 7.** t-SNE visualizations of the last-layer features embeddings extracted from UNetFormer, LOGCAN++, and the proposed CRECA-Net on the Vaihingen test set.

### *5.3. Ablation Studies and Analysis*

We conducted a series of ablation experiments on the Potsdam dataset to determine the optimal model architecture and to assess the influence of key components and hyperparameter settings. The detailed experimental results and analyses are presented below.

#### 5.3.1. Ablation of the Model Structure

To systematically analyze the contribution of each major component in CRECA-Net, we conduct a comprehensive ablation study, with experimental configurations and results summarized in Table 4. Specifically, six architectural variants were constructed by gradually simplifying the full model: **(1) CRECA-Net:** The complete model integrating both the CPR and CLCA modules, trained with the proposed DA loss. **(2) $\mathcal{L}_{da} \rightarrow \mathcal{L}_{ce}$:** A variant that replaces the DA loss with the standard CE loss, while retaining both the CPR and CLCA modules. **(3) $\mathcal{L}_{da} \rightarrow \mathcal{L}_{ce}$ & CPR $\rightarrow$ WCC:** Derived from variant (2) by substituting the CPR module with the weighted class center (WCC) scheme, where class centers are computed via probability-based weighting aggregation (see Figure 2b). **(4) $\mathcal{L}_{da} \rightarrow \mathcal{L}_{ce}$ & CPR $\rightarrow$ ACC:** A variant of (2) in which the CPR module is replaced with the average class center (ACC) scheme, where class centers are obtained by averaging all pixel features within each predicted class (see Figure 2a). **(5) -CPR-CLCA:** A configuration that removes both CPR and CLCA modules from CRECA-Net. **(6) -CPR-CLCA & $\mathcal{L}_{da} \rightarrow \mathcal{L}_{ce}$:** The most simplified version, obtained by further replacing the DA loss with CE loss in variant (5), leaving only the ResNet-50 backbone with a basic decoder.

**Table 4.** Ablation study evaluating the contributions of major components in CRECA-Net on the Potsdam dataset. "$\rightarrow$" indicates component or loss replacement; "-" denotes module removal; "&" represents the combination of multiple modifications; "✓" indicates the presence of the component, while "✗" denotes its absence.

| Method | $\mathcal{L}_{ce}$ | $\mathcal{L}_{da}$ | Imp. Surf | Building | Lowveg | Tree | Car | mIoU |
|---|---|---|---|---|---|---|---|---|
| (1) CRECA-Net | ✗ | ✓ | 89.57 | 94.25 | 79.00 | 81.07 | 93.30 | 87.44 |
| (2) $\mathcal{L}_{\text{da}} \rightarrow \mathcal{L}_{\text{ce}}$ | ✓ | ✗ | 89.66 | 94.08 | 78.89 | 80.49 | 92.63 | 87.15 |
| (3) $\mathcal{L}_{\text{da}} \rightarrow \mathcal{L}_{\text{ce}}$ & CPR$\rightarrow$WCC | ✓ | ✗ | 88.61 | 93.64 | 77.81 | 80.09 | 91.92 | 86.41 |
| (4) $\mathcal{L}_{\text{da}} \rightarrow \mathcal{L}_{\text{ce}}$ & CPR$\rightarrow$ACC | ✓ | ✗ | 88.10 | 93.29 | 77.47 | 79.92 | 91.58 | 86.07 |
| (5) -CPR-CLCA | ✗ | ✓ | 86.98 | 92.19 | 76.33 | 79.41 | 91.23 | 85.23 |
| (6) -CPR-CLCA & $\mathcal{L}_{\text{da}} \rightarrow \mathcal{L}_{\text{ce}}$ | ✓ | ✗ | 86.09 | 90.89 | 76.52 | 77.91 | 91.09 | 84.50 |

The ablation results yield several key insights. First, variant (2), which preserves both the CPR and CLCA modules, achieves a 2.65% mIoU improvement over the most simplified variant (6), confirming the effectiveness of the proposed CPR and CLCA modules. Second, introducing the proposed DA loss into this configuration further enhances model performance, enabling CRECA-Net to achieve the highest mIoU of 87.44%, indicating that the DA mechanism effectively facilitates learning from hard samples. Finally, when the CPR module is replaced with either WCC or ACC, mIoU drops by 0.74% and 1.08%, respectively, highlighting the superiority of the proposed CPR mechanism in generating accurate and discriminative class centers. Overall, these ablation results confirm that the CPR module, CLCA module, and DA loss jointly contribute to the performance gains of CRECA-Net.

#### 5.3.2. Ablation of the CPR Module

To further investigate the internal design of the CPR module, we conducted a dedicated ablation study, with results summarized in Table 5. Using the complete CPR configuration as the baseline, we systematically evaluated two aspects: (1) the contribution of each component, by removing individual components from the CACW submodule, namely, $\text{WP}_k$, $\text{WL}_k$, $\text{WC}_k$, as well as the inter-class prototype separation loss $\mathcal{L}_{inter}$; (2) the effectiveness of each confidence metric independently by constructing variants that retain only one of $\text{WP}_k$, $\text{WL}_k$, or $\text{WC}_k$.

The results show that removing any component or relying on a single confidence metric consistently degrades segmentation performance. This observation suggests that all components contribute positively to the refinement of discriminative class centers,

and that their complementary effects are crucial for improving downstream segmentation performance. Among all component removal variants, removing $\mathcal{L}_{inter}$ results in the largest performance drop. This can be attributed to the high inter-class similarity commonly observed in RS images: without the inter-class prototype separation constraint imposed by $\mathcal{L}_{inter}$, class centers tend to collapse toward each other (i.e., become less separated), resulting in reduced inter-class separability and the overall discriminative power of the learned feature space. Overall, these results demonstrate that both the multi-metric confidence weighting strategy and the inter-class prototype separation loss are indispensable for constructing high-quality class centers, thereby validating the design of the proposed CPR module.

**Table 5.** Ablation study evaluating the internal design of the CPR module on the Potsdam dataset, with the first row serves as the baseline. "✓" indicates the presence of the component, while "✗" denotes its absence.

| $WP_k$ | $WL_k$ | $WC_k$ | $\mathcal{L}_{inter}$ | Imp. Surf | Building | Lowveg | Tree | Car | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ | 89.57 | 94.25 | 79.00 | 81.07 | 93.30 | 87.44 |
| ✗ | ✓ | ✓ | ✓ | 89.35 | 94.07 | 78.94 | 80.83 | 92.66 | 87.17 |
| ✓ | ✗ | ✓ | ✓ | 89.50 | 94.20 | 78.98 | 80.58 | 92.54 | 87.16 |
| ✓ | ✓ | ✗ | ✓ | 89.20 | 93.88 | 78.88 | 81.07 | 92.92 | 87.19 |
| ✓ | ✓ | ✓ | ✗ | 88.99 | 93.80 | 78.75 | 81.05 | 92.77 | 87.07 |
| ✓ | ✗ | ✗ | ✓ | 89.18 | 93.87 | 77.48 | 80.64 | 93.09 | 86.85 |
| ✗ | ✓ | ✗ | ✓ | 89.04 | 93.93 | 78.36 | 81.07 | 92.56 | 86.99 |
| ✗ | ✗ | ✓ | ✓ | 89.16 | 93.97 | 78.09 | 80.68 | 92.78 | 86.94 |

***Choice of the Similarity Threshold*** $\beta$***:*** As a key hyperparameter within the inter-class prototype separation constraint of the CPR module, the threshold $\beta$ controls the maximum allowable cosine similarity between class centers and thus directly influences the strength of the inter-class separation constraint. A smaller $\beta$ enforces a stricter angular margin by lowering the threshold that triggers the penalty between class centers. To select an appropriate value, we performed a grid search over the range $\beta \in [0.05, 0.15]$, and the corresponding results are reported in Table 6. As $\beta$ increases from 0.05 to 0.125, the mIoU gradually improves and reaches its peak value of 87.44% at $\beta = 0.125$, indicating that moderate relaxation of the separation constraint helps avoid over-penalizing semantically related classes. However, further increasing $\beta$ causes a performance decline, suggesting that an overly weak separation constraint reduces inter-class discriminability. Overall, these results demonstrate a clear trade-off between enforcing class separation and preserving meaningful semantic relationships, with $\beta = 0.125$ striking an effective balance between the two.

**Table 6.** Ablation study on the effect of the similarity threshold $\beta$ on the Potsdam dataset.

| $\beta$ | 0.05 | 0.75 | 0.1 | 0.125 | 0.15 |
|---|---|---|---|---|---|
| mIoU | 87.11 | 87.33 | 87.39 | 87.44 | 87.20 |

### 5.3.3. Analysis of Difficulty-Aware Loss

To assess the effectiveness of the proposed DA loss and investigate the influence of its key hyperparameters, we conducted a series of controlled ablation studies on the Potsdam dataset.

***Effect of Difficulty-aware Loss:*** As shown in rows (1) and (5) of Table 4, replacing the standard CE loss with the DA loss for configurations (2) and (6) leads to mIoU improvements of 0.29% and 0.73%, respectively, without introducing any additional computational

overhead. These improvements demonstrate that the DA loss effectively mitigates the optimization bias toward easy samples inherent in standard CE loss by adaptively emphasizing hard samples, particularly those in boundary regions and visually ambiguous areas, leading to more effective and stable training.

***Visualization and Analysis of Difficulty-Aware Loss Weight*** $W_i$***:*** To further verify whether the proposed DA loss indeed focuses on hard-to-segment regions, we visualize the pixel-wise DA loss weight maps generated during training, as shown in Figure 8. The DA mechanism assigns higher weights to pixels with higher classification difficulty, thereby guiding the network to focus on challenging pixels. As observed in the first row, higher weights are concentrated along object boundaries, where segmentation ambiguity is inherently high. In the second row, regions containing visually similar categories, such as trees vs. low vegetation and cars vs. impervious surfaces, are also assigned elevated weights, reflecting increased classification uncertainty. Additionally, in the third row, areas affected by shadow occlusion receive higher weights, indicating the model's sensitivity to complex visual conditions. Collectively, these visual results provide clear evidence that the DA mechanism successfully captures pixel-wise classification difficulty arising from boundary ambiguity, visual similarity, and occlusion, thereby guiding the model to focus on challenging regions and improving overall segmentation performance.

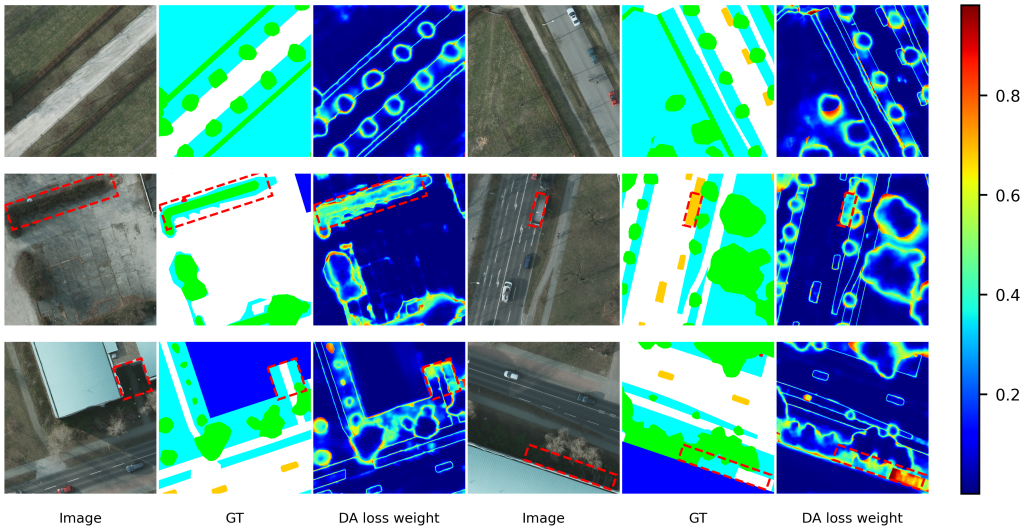



**Figure 8.** Visualization of DA loss weight maps. From left to right: raw image, ground truth, and the corresponding DA loss weight map generated during training. Warmer colors indicate higher difficulty-aware weights. Red dashed boxes highlight challenging regions, including visually similar categories and shadow-occluded areas.

***Choice of the Focusing Parameter*** $\gamma$***:*** The parameter $\gamma$ controls the degree of emphasis placed on hard samples in the DA loss. Following the principle of focal loss [66,69], we evaluated $\gamma$ within the range [0.5, 2.0], as shown in Table 7. The model's performance gradually improves as $\gamma$ increases from 0.5 to 1.0, reaching the highest mIoU of 87.44% at $\gamma = 1.0$. However, further increasing $\gamma$ to 2.0 does not yield additional performance gains, likely because excessively large $\gamma$ over-penalizes low-confidence predictions and inadvertently amplifies noisy or mislabeled pixels [70]. These observations indicate that a moderate focusing strength achieves an optimal balance between emphasizing hard samples and suppressing noise.

**Table 7.** Ablation study on the effect of the focusing parameter $\gamma$ in the DA loss on the Potsdam dataset.

| $\gamma$ | **0.5** | **1.0** | **1.5** | **2.0** |
|---|---|---|---|---|
| mIoU | 87.29 | 87.44 | 87.14 | 87.08 |

***Effect of Normalization:*** The dynamic difficulty estimation mechanism introduces image-level normalization of pixel-wise difficulty weights, serving two purposes: (1) re-distributing the loss within each image to emphasize hard regions; and (2) maintaining consistency in the overall loss magnitude, thereby preventing images with a high density of difficult pixels from dominating the optimization.

To evaluate the effectiveness of this image-level normalization strategy, we compared three loss formulations, as shown in Table 8: row (1) reports the proposed DA loss with image-level normalization, row (2) corresponds to the naïve focal loss with un-normalized weights $w_i = (1 - p_i)^\gamma$, and row (3) presents the foreground-aware loss proposed in FarSeg++ [69], which normalizes the naïve focal loss by scaling it with the ratio between the focal loss and the CE loss.

Without normalization, the naïve focal loss biases training toward images containing many hard samples, leading to a 0.42% decrease in mIoU. In contrast, our image-level normalization balances the per-image contribution to the loss, increasing mIoU to 87.28% and surpassing the FarSeg++ normalization by an additional 0.34%. These results indicate that image-level normalization effectively stabilizes per-image contributions during training, allowing the model to learn more evenly from hard regions of each image and preventing overfitting to image subsets dominated by difficult samples. Overall, this confirms the critical role of image-level normalization in enabling robust and stable training of the proposed DA loss.

***Choice of the Annealing Function:*** The annealing function enables a smooth transition from the standard CE loss to the DA loss, thereby mitigating training instability caused by unreliable difficulty estimation during the early training phase. To evaluate its impact, we compare multiple annealing strategies, as reported in rows (4)–(6) of Table 8. All annealing strategies outperform the direct application of the DA loss, confirming that a gradual transition from CE to DA loss is critical for stable training. Among them, cosine annealing achieves the best performance, reaching an mIoU of 87.44%, due to its gentle slope at the beginning and end of training, which avoids abrupt changes in gradient contributions and facilitates stable convergence. These findings demonstrate that selecting an appropriate annealing schedule is essential for stable and effective training with the proposed DA loss.

**Table 8.** Ablation study of the effect of normalization and annealing function in the DA loss on the Potsdam dataset.

| **Method** | **Normalization** | **Annealing Function** | **mIoU** |
|---|---|---|---|
| (1) Standard cross-entropy loss | ✗ | ✗ | 87.15 |
| (2) Loss weight $w_i = (1 - p_i)^\gamma$ [66] | ✗ | ✗ | 86.73 |
| (3) F-A Optimization [69] | ✓ | ✓ | 86.94 |
| (4) Difficult estimation | ✓ | ✗ | 87.28 |
| (5) Difficult estimation + Linear Annealing | ✓ | Linear | 87.35 |
| (6) Difficult estimation + Polynomial Annealing | ✓ | Polynomial | 87.40 |
| (7) Difficult estimation + Cosine Annealing | ✓ | Cosine | 87.44 |

***Choice of Annealing Step $T_{step}$:*** The DA loss introduces a hyperparameter $T_{\text{step}}$ to control the duration of the annealing process. Since the reliability of difficulty estimation depends on the model's predictions quality, setting $T_{\text{step}}$ too small may lead to unreliable difficulty estimates, while an overly large value may reduce the number of training iterations devoted to optimizing hard samples. To determine a stable and principled value, we reference the training dynamics of the CE-trained CRECA-Net (where the DA loss is replaced with standard CE loss). Specifically, $T_{\text{best}}$ is defined as the training step at which the CE-trained model achieves its best validation performance, and $T_{\text{step}}$ is set as a fraction of $T_{\text{best}}$. Accordingly, we evaluate four settings: $T_{\text{step}} \in \{T_{\text{best}}, 0.9T_{\text{best}}, 0.8T_{\text{best}}, 0.7T_{\text{best}}\}$, as summarized in Table 9. Results on both the Potsdam and Vaihingen datasets indicate that $T_{\text{step}} = 0.8T_{\text{best}}$ yields the best performance, while the model's performance remains relatively stable across neighboring values. These findings suggest that the DA loss is not overly sensitive to the precise choice of $T_{\text{step}}$, and a reliable setting can be derived from CE-based training behavior without introducing additional hyperparameter complexity.

**Table 9.** Ablation study of the effect of the annealing step $T_{\text{step}}$ in the DA loss on the Potsdam and Vaihingen datasets.

| $T_{\text{step}}$ | $T_{\text{best}}$ | $0.9T_{\text{best}}$ | $0.8T_{\text{best}}$ | $0.7T_{\text{best}}$ |
|---|---|---|---|---|
| Potsdam(mIoU) | 87.27 | 87.28 | 87.44 | 87.35 |
| Vaihingen(mIoU) | 84.45 | 84.57 | 84.66 | 84.48 |

***Choice of Loss Coefficient $\lambda_{\text{aux}}$:*** We further analyze the influence of the auxiliary loss coefficient $\lambda_{\text{aux}}$, which regulates the contribution of the auxiliary segmentation branch, by fixing all other loss terms and varying $\lambda_{\text{aux}}$ within 0.7, 0.8, 0.9, 1.0. As shown in Table 10, increasing $\lambda_{\text{aux}}$ from 0.7 to 0.8 steadily improves performance, suggesting that moderate auxiliary supervision enhances optimization. However, excessively large values tend to hinder the optimization of the main prediction branch, resulting in a slight degradation of accuracy. Based on these observations, $\lambda_{\text{aux}} = 0.8$ is selected as the default setting, achieving a balanced trade-off between the auxiliary supervision and the main segmentation objective.

**Table 10.** Ablation study on the effect of the auxiliary loss coefficient $\lambda_{\text{aux}}$ on the Potsdam dataset.

| $\beta$ | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|
| mIoU | 87.29 | 87.44 | 87.38 | 87.36 |

## 6. Discussion

The consistent performance improvements of CRECA-Net on the ISPRS Potsdam and Vaihingen datasets primarily resulted from the design of the CPR module, CLCA module, and the DA loss. The CPR module enhances the reliability and discriminability of class prototypes through three key components: pixel selection, CWCA, and the inter-class prototype separation loss. Specifically, pixel selection ensured that class prototypes were computed from semantically coherent and prediction-consistent regions, CWCA allowed high-confidence pixels to contribute more to prototype estimation, while the inter-class prototype separation loss explicitly promoted separability between different category prototypes. The CLCA module employed cross-attention to model pixel-to-class prototype correlations, thereby capturing long-range contextual dependencies under semantic prototype guidance and reducing prediction ambiguity in complex scenes. In addition, the DA loss introduced a dynamic difficulty estimation mechanism with an adaptive loss scheduling strategy, which adaptively adjusted pixel-wise loss weights within

each image, enabling the model to gradually shift its learning focus from easy samples to more challenging ones, such as boundary regions and visually confusing categories, while ensuring stable training.

These results indicated that jointly improving class representation quality and training dynamics provided an effective strategy for addressing the challenges of high-resolution RS image segmentation, particularly in scenes characterized by large intra-class variation and complex backgrounds. Despite these improvements, accurately delineating fine-grained boundary details and distinguishing subtle transitions between visually similar categories remained challenging, particularly in complex scenes where boundary ambiguity and category confusion were more pronounced.

## 7. Conclusions

In this work, we investigated the key challenges inherent in high-resolution RS image segmentation, including complex backgrounds, large intra-class variance, and small inter-class variance, all of which commonly limit the performance of conventional segmentation models. To address these challenges, we proposed CRECA-Net, a class-aware segmentation network that integrates a CPR module, CLCA modules, and a DA loss. Specifically, the CPR module constructed more reliable and discriminative class representations, the CLCA module aggregated long-range contextual information under semantic prototype guidance, and the DA loss enabled the model to focus more effectively on hard samples, such as boundary regions and visually similar categories. Extensive experiments on two benchmark datasets demonstrated that CRECA-Net consistently outperformed several state-of-the-art methods across multiple evaluation metrics, achieving superior segmentation accuracy and strong generalization in diverse RS scenarios. Future work will explore integrating contrastive learning to further enhance intra-class compactness and inter-class separability at the representation level. Additionally, edge-aware loss functions and boundary-refinement mechanisms will be incorporated to explicitly alleviate structural ambiguity and improve boundary delineation in complex RS scenes.

**Author Contributions:** Conceptualization, R.L.; methodology, R.L.; software, L.Y.; validation, R.L.; resources, B.C.; data curation, L.Y. and S.Z.; writing—original draft preparation, R.L.; writing—review editing, R.L.; visualization, R.L.; supervision, L.Y. and S.Z.; funding acquisition, B.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Scientific Research Foundation of Chongqing University of Technology (Grant No. 2025ZDZ012), the Industry-University-Research Collaboration Innovation Fund for Chinese Higher Education Institutions, Ministry of Education (Grant No. 2025ZX00), the Major Project of Chongqing Technology Innovation and Application Development (Grant No. CSTB2024TIAD-STX0026), the National Science Foundation of China (Grant No. 61961040, U1903215 and 61771089), and the Sichuan Provincial Key Research and Development Program (Grant No. 2021YFQ0011).

**Data Availability Statement:** Vaihingen and Potsdam datasets are available at: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/default.aspx (accessed on 2 January 2025).

**Acknowledgments:** The authors would like to express their sincere gratitude to Dolla Mihretu Samuel for his valuable assistance in refining the English expression and improving the overall clarity and readability of the manuscript during the revision process.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References


1. Lian, R.; Wang, W.; Mustafa, N.; Huang, L. Road extraction methods in high-resolution remote sensing images: A comprehensive review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5489–5507. [CrossRef]
2. Wang, C.; Xu, R.; Xu, S.; Meng, W.; Wang, R.; Zhang, J.; Zhang, X. Toward Accurate and Efficient Road Extraction by Leveraging the Characteristics of Road Shapes. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4404616. [CrossRef]
3. Zhang, P.; Li, J.; Wang, C.; Niu, Y. SAM2MS: An Efficient Framework for HRSI Road Extraction Powered by SAM2. *Remote Sens.* **2025**, *17*, 3181. [CrossRef]
4. Wei, S.; Zeng, X.; Zhang, H.; Zhou, Z.; Shi, J.; Zhang, X. LFG-Net: Low-Level Feature Guided Network for Precise Ship Instance Segmentation in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5231017. [CrossRef]
5. Azimi, S.M.; Fischer, P.; Körner, M.; Reinartz, P. Aerial LaneNet: Lane-Marking Semantic Segmentation in Aerial Imagery Using Wavelet-Enhanced Cost-Sensitive Symmetric Fully Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2920–2938. [CrossRef]
6. Zhang, K.; Ming, D.; Du, S.; Xu, L.; Ling, X.; Zeng, B.; Lv, X. Distance Weight-Graph Attention Model-Based High-Resolution Remote Sensing Urban Functional Zone Identification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5608518. [CrossRef]
7. Zhao, Y.; Sun, G.; Zhang, L.; Zhang, A.; Jia, X.; Han, Z. MSRF-Net: Multiscale Receptive Field Network for Building Detection From Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5515714. [CrossRef]
8. Chicchon, M.; Colosi, F.; Malinverni, E.S.; León Trujillo, F.J. Urban Sprawl Monitoring by VHR Images Using Active Contour Loss and Improved U-Net with Mix Transformer Encoders. *Remote Sensing* **2025**, *17*, 1593. [CrossRef]
9. Li, R.; Zheng, S.; Duan, C.; Wang, L.; Zhang, C. Land cover classification from remote sensing images based on multi-scale fully convolutional network. *Geo-Spat. Inf. Sci.* **2022**, *25*, 278–294. [CrossRef]
10. Ma, A.; Zheng, C.; Wang, J.; Zhong, Y. Domain Adaptive Land-Cover Classification via Local Consistency and Global Diversity. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5606317. [CrossRef]
11. Liu, L.; Tong, Z.; Cai, Z.; Wu, H.; Zhang, R.; Le Bris, A.; Olteanu-Raimond, A.M. HierU-Net: A Hierarchical Semantic Segmentation Method for Land Cover Mapping. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4404614. [CrossRef]
12. Saleh, T.; Holail, S.; Zahran, M.; Xiao, X.; Xia, G.S. LiST-Net: Enhanced Flood Mapping with Lightweight SAR Transformer Network and Dimension-Wise Attention. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5211817. [CrossRef]
13. Yadav, R.; Nascetti, A.; Ban, Y. Attentive Dual Stream Siamese U-Net for Flood Detection on Multi-Temporal Sentinel-1 Data. In Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 5222–5225. [CrossRef]
14. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. 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, Seattle, WA, USA, 13–19 June 2020; pp. 4096–4105.
15. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
16. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
17. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
18. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 173–190.
19. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
20. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
21. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
22. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7262–7272.
23. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
24. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.

25. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
26. Guan, F.; Zhao, N.; Wang, H.; Fang, Z.; Zhang, J.; Yu, Y.; Jiang, L.; Huang, H. Dual-branch transformer framework with gradient-aware weighting feature alignment for robust cross-view geo-localization. *Inf. Fusion* **2025**, *127*, 103808. [CrossRef]
27. Guan, F.; Zhao, N.; Fang, Z.; Jiang, L.; Zhang, J.; Yu, Y.; Huang, H. Multi-level representation learning via ConvNeXt-based network for unaligned cross-view matching. *Geo-Spat. Inf. Sci.* **2025**, *28*, 2344–2357. [CrossRef]
28. Ye, Z.; Li, Y.; Li, Z.; Liu, H.; Zhang, Y.; Li, W. Attention-Multi-Scale Network for Semantic Segmentation of Multi-Modal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5610315. [CrossRef]
29. Zhao, Y.; Qiu, L.; Yang, Z.; Chen, Y.; Zhang, Y. MGF-GCN: Multimodal interaction Mamba-aided graph convolutional fusion network for semantic segmentation of remote sensing images. *Inf. Fusion* **2025**, *122*, 103150. [CrossRef]
30. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 593–602.
31. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
32. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
33. Cai, H.; Li, J.; Hu, M.; Gan, C.; Han, S. Efficientvit: Multi-scale linear attention for high-resolution dense prediction. *arXiv* **2022**, arXiv:2205.14756.
34. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
35. Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; Ding, E. Acfnet: Attentional class feature network for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6798–6807.
36. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597. [CrossRef]
38. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv*, **2018**, arXiv:1804.03999. [CrossRef]
39. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
40. Yu, F. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
41. Chen, L.C. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
42. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-22 June 2018; pp. 3684–3692.
43. Cui, J.; Liu, J.; Wang, J.; Ni, Y. Global context dependencies aware network for efficient semantic segmentation of fine-resolution remoted sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2505205. [CrossRef]
44. Liu, R.; Mi, L.; Chen, Z. AFNet: Adaptive fusion network for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7871–7886. [CrossRef]
45. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9167–9176.
46. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [CrossRef]
47. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8009205. [CrossRef]
48. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
49. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
50. Yu, D.; Ji, S. Long-range correlation supervision for land-cover classification from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4409814. [CrossRef]

51. Ni, Y.; Liu, J.; Chi, W.; Wang, X.; Li, D. CGGLNet: Semantic Segmentation Network for Remote Sensing Images Based on Category-Guided Global-Local Feature Interaction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5615617. [CrossRef]
52. Zeng, Q.; Zhou, J.; Tao, J.; Chen, L.; Niu, X.; Zhang, Y. Multiscale global context network for semantic segmentation of high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622913. [CrossRef]
53. Yang, Y.; Zheng, S.; Wang, X.; Ao, W.; Liu, Z. AMMUNet: Multiscale Attention Map Merging for Remote Sensing Image Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 3506718. [CrossRef]
54. Meng, W.; Shan, L.; Ma, S.; Liu, D.; Hu, B. Dlnet: A dual-level network with self-and cross-attention for high-resolution remote sensing segmentation. *Remote Sens.* **2025**, *17*, 1119. [CrossRef]
55. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
56. Zhang, Q.; Yang, Y.B. Rest: An efficient transformer for visual recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15475–15485.
57. Jin, Z.; Liu, B.; Chu, Q.; Yu, N. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7189–7198.
58. Jin, Z.; Gong, T.; Yu, D.; Chu, Q.; Wang, J.; Wang, C.; Shao, J. Mining contextual information beyond image for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7231–7241.
59. Chen, Y.; Dong, Q.; Wang, X.; Zhang, Q.; Kang, M.; Jiang, W.; Wang, M.; Xu, L.; Zhang, C. Hybrid attention fusion embedded in transformer for remote sensing image semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* **2024**, *17*, 4421–4435. [CrossRef]
60. Deng, G.; Wu, Z.; Wang, C.; Xu, M.; Zhong, Y. CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4401120. [CrossRef]
61. Ma, X.; Lian, R.; Wu, Z.; Guo, H.; Ma, M.; Wu, S.; Du, Z.; Song, S.; Zhang, W. LOGCAN++: Adaptive Local-global class-aware network for semantic segmentation of remote sensing imagery. *arXiv* **2024**, arXiv:2406.16502. [CrossRef]
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
63. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
64. Huang, Y.H.; Proesmans, M.; Georgoulis, S.; Van Gool, L. Uncertainty based model selection for fast semantic segmentation. In Proceedings of the 2019 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 27–31 May 2019; pp. 1–6.
65. Liu, Z.; Zhang, Z.; Khaki, S.; Yang, S.; Tang, H.; Xu, C.; Keutzer, K.; Han, S. Sparse refinement for efficient high-resolution semantic segmentation. In *Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: Cham, Switzerland, 2025; pp. 108–127.
66. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
67. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Bnitez, S.; Breitkopf, U. International Society for Photogrammetry and Remote Sensing, 2d Semantic Labeling Contest. **2020**, *29*. Available online: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/Default.aspx (accessed on 1 October 2025).
68. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
69. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. FarSeg++: Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13715–13729. [CrossRef] [PubMed]
70. Li, B.; Liu, Y.; Wang, X. Gradient harmonized single-stage detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, 27 January–1 February 2019; Volume 33, pp. 8577–8584.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.