RegionMatch: Pixel-Region Collaboration for Semi-Supervised Semantic Segmentation in Remote Sensing Images

Xiaoqian Zhu[1] , **Xiangrong Zhang**1 , **Tianyang Zhang**[1] , **Chaowei Fang**[1] ,
Xu Tang[1] and **Licheng Jiao**[1]

[1]School of Artificial Intelligence, Xidian University.
zxq_enjoy@126.com, xrzhang@mail.xidian.edu.cn, tianyangzhang@stu.xidian.edu.cn

Abstract

Semi-supervised semantic segmentation (S4) has shown significant promise in reducing the burden of labor-intensive data annotation. However, existing methods mainly rely on pixel-level information, neglecting the strong region consistency inherent in remote sensing images (RSIs), which limits their effectiveness in handling the complex and diverse backgrounds of RSIs. To address this, we propose RegionMatch, a novel approach that leverages unlabeled data from a fresh object-level perspective, which is more tailored to the nature of semantic segmentation. We design the Pixel-Region Synergy Pseudo-Labeling strategy, which explicitly injects object-level contextual information into the S4 pipeline and promotes knowledge collaboration between pixel and region perspectives for generating high-quality pseudo-labels. In addition, we propose the Region Structure-Aware Correlation Consistency, which models object-level relationships by establishing inter-region correlations across images and pixel correlations within regions, providing more effective supervision signals for unlabeled data. Experimental results demonstrate that RegionMatch outperforms state-of-the-art methods on multiple authoritative remote sensing datasets, highlighting its superiority in the RSIs.

1 Introduction

Semantic segmentation in remote sensing images (RSIs) is a fundamental task in remote sensing data interpretation, aiming to classify each pixel into distinct categories. It is widely applied in land cover monitoring, urban planning, and disaster assessment. Currently, deep learning-based models have dominated this field. However, these models require labor-intensive pixel-level annotations to achieve satisfactory performance. To address this challenge, semi-supervised semantic segmentation (S4) [Wang *et al.*, 2022a; Ding *et al.*, 2023] uses a small labeled dataset combined with a large amount of unlabeled data, reducing the dependence on annotation data.

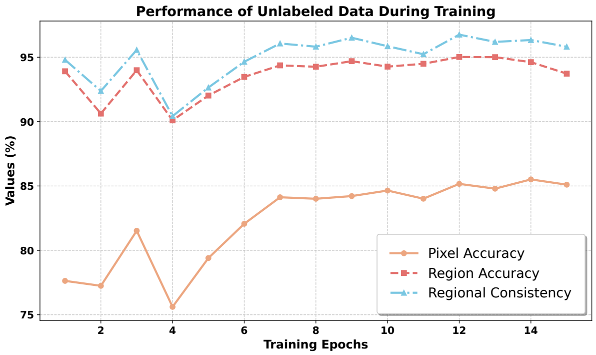


Figure 1: Explanation of pixel accuracy, region accuracy, and regional consistency in pseudo-labels from DeeplabV3+ model on the 1/2 labeled protocol of the Vaihingen dataset. Current methods only use pixel-level pseudo-labels, ignoring valuable object-level information. Region labels are derived from pixel predictions, with the true category defined as the ground truth of the dominant pixel. Based on this, the region accuracy is calculated. Regional consistency measures the proportion of pixels in a region belonging to the dominant category.

The key to S4 lies in the effective utilization of unlabeled data. Currently, the combination of pseudo-labeling and consistency regularization techniques has become the dominant approach [Sohn *et al.*, 2020; Yang *et al.*, 2023a]. These methods select reliable pixels predicted from one augmented version of unlabeled data and use them as pseudo-labels to train another. Most of the improved methods [Wang *et al.*, 2022b; Qiao *et al.*, 2023; Wang *et al.*, 2024] still utilize unlabeled data only from an individual pixel perspective. Recent studies revisit the potential of semi-supervised learning from the perspective of semantic segmentation tasks, focusing on more effective strategies for exploiting unlabeled data. For instance, RankMatch [Mai *et al.*, 2024] designs consistency regularization methods from the perspective of inter-pixel correlations, CorrMatch [Sun *et al.*, 2024a] utilizes pairwise similarity maps between pixels to propagate labels, and MPMC [Howlader *et al.*, 2025] mitigates pseudo-labeling noise by utilizing multi-label classification predictions for image patches. However, these approaches are limited to pixel-level information when leveraging unlabeled data. This limitation is particularly pronounced in RSIs, where models are more vulnerable to noise due to complex backgrounds and diverse landscapes.

Semantic segmentation tasks inherently depend on understanding object-level relationships, where a pixel's label is often determined by the object region to which it belongs. RSIs cover wide ground areas, with ground objects exhibiting strong coherence, particularly in areas such as residential zones, forests, and agricultural lands. We observed that semantic segmentation models exhibit strong object region recognition ability in RSIs. As illustrated in Fig. 1, classification accuracy for object regions significantly outperforms pixel-level accuracy, indicating that region-level information is more stable and reliable than pixel-level's. Furthermore, the predicted regions exhibit high consistency, as evidenced by the dominance of a single category within these regions. Current S4 methods fail to incorporate valuable object-level information, limiting their performance in RSIs.

In this paper, we analyze the bottleneck of existing methods based on pixel-level information and propose a novel approach, RegionMatch, to leverage unlabeled data from a fresh object-level perspective for semi-supervised semantic segmentation of RSIs. Specifically, we introduce an object region classifier that trains the model to identify object regions composed of pixels, explicitly injecting object-level contextual information into the S4 pipeline. We then design the Pixel-Region Synergy Pseudo-Labeling strategy (PRSPL) to integrate both pixel-level and object-level information, generating high-quality pseudo-labels for model training and promoting knowledge collaboration across the two perspectives. To further unleash the potential of unlabeled data, we propose the Region-Structure Aware Correlation Consistency (RSCC). RSCC explicitly models object-level relationships by establishing multi-level correlations both within and between pixels and regions, providing stronger supervision signals for unlabeled data that are more in line with the nature of semantic segmentation. RSCC consists of two key components: Cross-Image Region-Region Correlation Consistency (CIR²R) and Pixel-Region Correlation Consistency (PRCC). CIR²R establishes inter-region correlation across images to capture object-level consistency, while PRCC strengthens intra-region pixel consistency, thus improving the model's understanding of object-level relationships.

The contributions of this paper can be summarized as follows:

(1) We propose RegionMatch, a novel S4 method for RSIs, which utilizes unlabeled data from a fresh object-level perspective that is more tailored to the nature of semantic segmentation.

(2) We designed the Pixel-Region Synergy Pseudo-Labeling strategy, which encourages knowledge cooperation between pixel and object-region perspectives to generate high-quality pseudo-labels for model training.

(3) We propose Region-Structure Aware Correlation Consistency which models object-level relationships to provide more effective supervisory signals for unlabeled data, thus inspiring the potential of unlabeled data further.

2 Related Work

2.1 Semi-Supervised Semantic Segmentation

The primary challenge in S4 is effectively leveraging unlabeled data to enhance model generalization. Key strategies include adversarial method [Ke *et al.*, 2020; Zhang *et al.*, 2020], consistency regularization [Chen *et al.*, 2021; Liu *et al.*, 2022], pseudo-labeling [Yang *et al.*, 2022; Teh *et al.*, 2022], and contrastive learning [Zhou *et al.*, 2021; Wang *et al.*, 2023b; Wang *et al.*, 2023a]. Among these, the combination of pseudo-labeling and consistency regularization [Sohn *et al.*, 2020; Yang *et al.*, 2023a; Chen *et al.*, 2021] has emerged as the dominant approach. These methods generate pseudo-labels from a segmentation model's predictions on weakly augmented unlabeled data, which are then used to train strongly augmented versions. While previous approaches [Wang *et al.*, 2022b; Qiao *et al.*, 2023; Sun *et al.*, 2024b; Na *et al.*, 2024; Yang *et al.*, 2025; Wang *et al.*, 2024] utilize unlabeled data from the perspective of individual pixels, recent approaches construct supervised signals for unlabeled data from the perspective of inter-pixel relationships and multi-label classification of image patches. For example, RankMatch [Mai *et al.*, 2024] models pixel correlation using representative proxies to optimize rank-aware consistency, while CorrMatch [Sun *et al.*, 2024a] employs pairwise pixel similarity for label propagation. Another approach [Howlader *et al.*, 2025] introduces a multi-label classifier with multi-scale patches to reduce pseudo-label noise. However, these methods utilize unlabeled data only from a pixel-level perspective, ignoring the fact that the semantic segmentation task is more dependent on the understanding of object-level relationships, thus limiting their performance.

Relatively few studies have focused on semi-supervised semantic segmentation for RSIs. Current research primarily concentrates on developing more effective data perturbation strategies [Lu *et al.*, 2023; Lv and Zhang, 2024; Bai *et al.*, 2024; Liu *et al.*, 2024; Luo *et al.*, 2024] to improve the effect of consistency regularization, and constructing representation learning methods for unlabeled data [Xin *et al.*, 2024; Luo *et al.*, 2024; Yang *et al.*, 2023b]. These methods rely solely on individual pixel information when utilizing unlabeled data, overlooking the strong region consistency inherent in RSIs, which limits the performance of S4.

2.2 Context for Semantic Segmentation

Pixel classification is heavily influenced by its context, typically defined by spatially related locations, such as the surrounding neighborhood. Many semantic segmentation models enhance feature representations by incorporating contextual information. Early works [Zhao *et al.*, 2017; Zhao *et al.*, 2018; Chen *et al.*, 2017] used convolution and pooling kernels of varying sizes to capture multi-scale context. However, these fixed-region aggregation methods are non-specific, and information collected from semantically irrelevant regions may introduce noise. Subsequent research has focused on more effective aggregation of contextual information from semantically relevant regions. For instance, [Ding *et al.*, 2019] restricts context regions by learning semantic masks that match object shape and scale. [Yu *et al.*, 2020] addresses

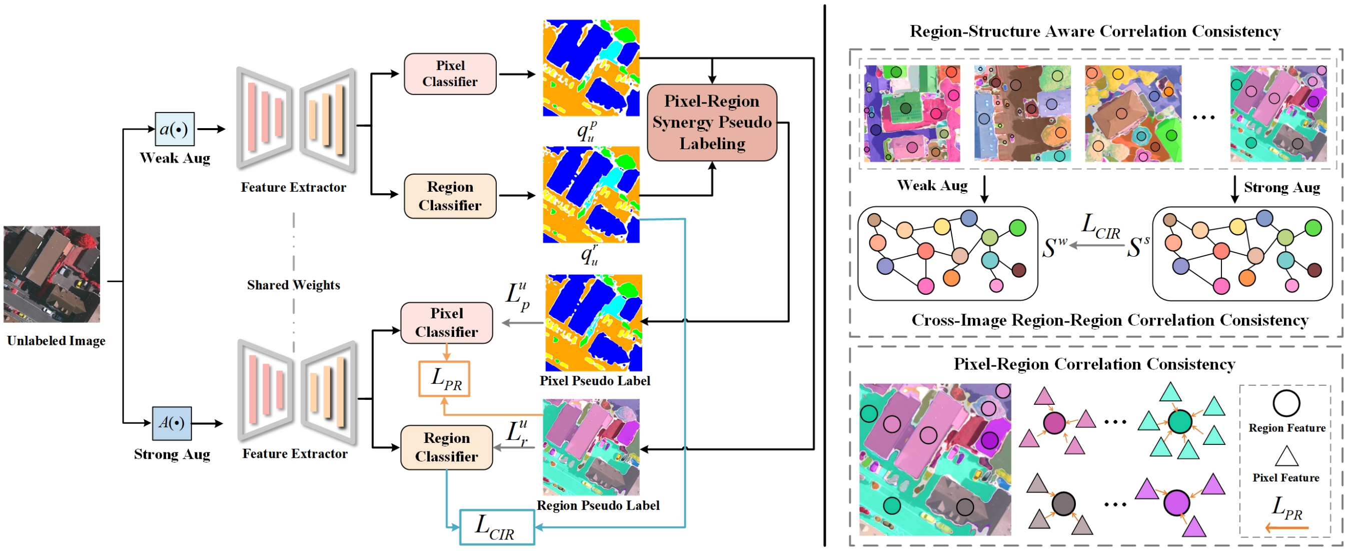


Figure 2: Illustration of our RegionMatch pipeline for unlabeled images. Object-level contextual information is explicitly introduced to the S4 pipeline through an object-region classification task. We propose Pixel-Region Synergy Pseudo-Labeling and Region-Structure Aware Correlation Consistency to promote knowledge collaboration between the pixel and object region perspectives, while modeling object-level relationships to provide stronger supervision signals for unlabeled data.

intra- and inter-class context dependencies through affinity loss supervision. [Shen *et al.*, 2020] enhances pixel representation by modeling contextual interactions between regions. [Hu *et al.*, 2021a] proposed region-aware contrast learning to improve pixel features. Unlike these fully-supervised approaches, this paper aims to explore object-level information from unlabeled data as supervisory signals to maximize its utilization.

3 Method

S4 aims to train a semantic segmentation model using a small labeled dataset and a large unlabeled dataset. We propose RegionMatch, leveraging unlabeled data from a novel object-level perspective.

3.1 Preliminaries

RegionMatch is built upon the classic pseudo-labeling and weak-to-strong consistency regularization framework [Sohn *et al.*, 2020]. Given a labeled set $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{N^l}$ and an unlabeled set $\mathcal{D}^u = \{x_i^u\}_{i=1}^{N^u}$, where $N^u \gg N^l$. Pixel-wise cross-entropy loss L_p is applied to both labeled and unlabeled data:

$$L_p = L_p^l + L_p^u \tag{1}$$

For labeled data, the loss L_p^l is calculated as follows:

$$L_p^l = \frac{1}{N^l} \sum_{l=1}^{N^l} l_{ce}(y^l, p^l) \tag{2}$$

where l_{ce} represents the standard cross-entropy loss function, $p^l = f_p(g(x^l))$ is the predicted probability distribution of the pixel classifier $f_p(\cdot)$, and $g(\cdot)$ is the feature extractor of the segmentation model. For unlabeled images, weak augmentation $\alpha(\cdot)$ and strong augmentation $A(\cdot)$ are applied. The prediction from the weakly augmented version is used to supervise the strongly augmented version. The unlabeled loss L_p^u is computed as follows:

$$L_p^u = \frac{1}{N^u} \sum_{u=1}^{N^u} \mathbb{I}(q^u > \tau) \cdot l_{ce}(\hat{y}^u, p^u) \tag{3}$$

where $p^u = f_p(g(A(x^u)))$ represents the predicted probability distribution of the pixel classifier, $q_u = f_p(g(\alpha(x^u)))$, $\hat{y}^u = \arg\max(q^u)$, and τ is a fixed confidence threshold.

3.2 Overview

Fig. 2 shows the pipeline of RegionMatch for unlabeled data. We generate region pseudo label using the model's predictions on a weakly augmented version of the image, dividing the image into multiple semantically consistent regions consisting of adjacent pixels that represent complete objects or parts of objects. An object region classifier is then introduced to classify these regions, explicitly incorporating object-level contextual information into the pipeline. We then propose the Pixel-Region Synergy Pseudo-Labeling strategy (PRSPL), which merges information from both pixel- and region-level perspectives to generate high-quality pseudo-labels for training. Additionally, we propose Region-Structure Aware Correlation Consistency (RSCC) to model object-level relationships in images to build supervisory signals for unlabeled data. RSCC consists of Cross-Image Region-Region Correlation Consistency (CIR²R) and Pixel-Region Correlation Consistency (PRCC). The overall loss function of RegionMatch consists of the pixel classification loss L_p, the region classification loss L_r, the CIR²R loss L_{CIR}, and the PRCC loss L_{PR}:

$$L_{\text{all}} = L_p + L_r + \lambda_1 L_{\text{CIR}} + \lambda_2 L_{\text{PR}} \tag{4}$$

where λ_1 and λ_2 are the corresponding weight coefficients.

3.3 Object Region Classification

In the method described in Section 3.1, the supervised signal for unlabeled data relies only on pixel-level information, which makes pixel classification susceptible to noise in RSIs with complex and variable backgrounds. RSIs typically have extensive ground coverage and strong ground object coherence. We find that the model provides more stable and reliable pseudo-labels at the object level compared to the pixel level. Therefore, we introduce explicit supervision for object-region classification to inject object-level contextual information into the S4 pipeline, improving the model's understanding of object-level relationships of the image.

For each image, both pixel-level and region-level labels are provided. The pixel label denotes the semantic category of each pixel, while the region label specifies the object region to which the pixel belongs. Region labels $r_l \in \{1, \dots, N_l\}^{H \times W}$ for labeled image x_l and $r_u \in \{1, \dots, N_u\}^{H \times W}$ for unlabeled image x_u are both obtained using a connected component algorithm. Labeled image uses the pixel-level ground truth y_l, and unlabeled image uses the pixel-level pseudo-label $\hat{y}_u$. For further details, see the appendix. We introduce a region classifier f_r to classify the regions in the image. For each region i, the feature extractor outputs feature maps $F \in \mathbb{R}^{H \times W \times d}$, which are aggregated using mask average pooling within the region and passed through a projection layer $h(\cdot)$ to obtain the corresponding region feature R_i^l. The region classification loss for the labeled image x_l is defined as:

$$L_r^l = \frac{1}{N_r^l} \sum_{i=1}^{N_r^l} l_{ce}(y_i^r, f_r(R_i^l)) \tag{5}$$

where y_i^r denotes the ground truth class label of region i.

To minimize interference from noisy regions in the unlabeled data, we apply a dual thresholding mechanism with τ_{low} and τ_{high} to compute the region classification loss for unlabeled data. Specifically, we calculate the average confidence score $\bar{c}_i^u$ of each region i as the quality assessment of the region. Noisy regions are filtered out using the lower threshold:

$$\mathcal{I} = \{i \mid \bar{c}_i^u \geq \tau_{\text{low}}\} \tag{6}$$

For regions with $\tau_{\text{low}} \leq \bar{c}_i^u < \tau_{\text{high}}$, we generate soft labels $\tilde{q}_i^u$ by averaging the pseudo-label probability within the region. For regions with $\bar{c}_i^u > \tau_{\text{high}}$, hard labels are used: $\hat{y}_i^r = \arg\max(\tilde{q}_i^u)$. The region classification loss for the unlabeled image x_u is computed as:

$$L_r^u = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \begin{cases} l_{ce}(\hat{y}_i^r, f_r(R_i^u)) & \text{if } \bar{c}_i^u \geq \tau_{\text{high}}, \\ l_{ce}(\tilde{q}_i^u, f_r(R_i^u)) & \text{if } \tau_{\text{low}} \leq \bar{c}_i^u < \tau_{\text{high}}. \end{cases} \tag{7}$$

The total region classification loss is defined as:

$$L_r = \frac{1}{N_l} \sum_{l=1}^{N_l} L_r^l + \lambda \frac{1}{N_u} \sum_{u=1}^{N_u} L_r^u \tag{8}$$

where $\lambda = 0.5$ is a balancing factor that controls the weight of the unlabeled loss relative to the labeled loss.

3.4 Pixel-Region Synergy Pseudo-Labeling

The pixel classifier captures features at the individual pixel level, while the region classifier catches class attributes at the object level. A natural idea is to promote cooperation between the two perspectives to improve the model's understanding of unlabeled data. To this end, we propose the Pixel-Region Synergy Pseudo-Labeling strategy (PRSPL) strategy, which combines the knowledge of both perspectives to generate high-quality pseudo-labels for training models. If the two classifiers reach consensus on the category prediction of a pixel, we consider the pseudo-label of this pixel to be reliable. We therefore select pixels with consistent predictions from both perspectives to train the model, which in turn promotes knowledge cooperation between the two perspectives.

For the image x_u, we use the region classifier f_r to predict pixel-wise probability distributions q_u^r under the region-level view. For the pixel-level view, the probability distribution is denoted as q_u^p. We calculate the entropy for both the region-level and pixel-level views, defined as H_u^r and H_u^p, respectively:

$$H_u^r = -\sum_{c=1}^{C} q_u^r(c) \log q_u^r(c) \tag{9}$$

$$H_u^p = -\sum_{c=1}^{C} q_u^p(c) \log q_u^p(c) \tag{10}$$

where C is the number of classes. To avoid interference from noisy regions, we compute reliable pseudo labels in the low entropy regions of both views. In low-entropy regions, the predictions from both classifiers are considered more confident. The low-entropy region i for the region-level perspective $\mathcal{I}^r$ and the pixel-level perspective $\mathcal{I}^p$ are defined as:

$$\mathcal{I}^r = \{i \mid H_u^r(i) \leq T_\alpha^r\} \tag{11}$$

$$\mathcal{I}^p = \{i \mid H_u^p(i) \leq T_\alpha^p\} \tag{12}$$

The threshold values for entropy are calculated as:

$$T_\alpha^r = \text{Quantile}(\{H_u^r(i) \mid i \in \{1, \dots, H \times W\}\}, 1 - \alpha) \tag{13}$$

$$T_\alpha^p = \text{Quantile}(\{H_i^p(i) \mid i \in \{1, \dots, H \times W\}\}, 1 - \alpha) \tag{14}$$

Here, $\text{Quantile}(\cdot, \alpha)$ represents the calculation of the α-th quantile. We use a dynamic quantile α_t calculated as:

$$\alpha_t = \alpha_0 \cdot \left(1 - \frac{t}{\text{total epoch}}\right) \tag{15}$$

where t is the current epoch number during training.

In the intersection of the low-entropy regions from both views, if a pixel has consistent classification predictions from both the region-level and pixel-level views, we consider the pixel's pseudo-label to be reliable. Therefore, the pixel classification loss for unlabeled data in our method can be expressed as:

$$L_p^u = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{1}{HW} \sum_{i=1}^{HW} l_{ce}(\hat{y}^u(i), p^u(i)) \odot M_i \tag{16}$$

where:

$$M_i = \begin{cases} 1, & \text{if } i \in \mathcal{I}^r \cap \mathcal{I}^p \text{ and } \arg\max q_u^r(i) = \arg\max q_u^p(i), \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

Here, $\odot$ denotes element-wise multiplication.

3.5 Region-Structure Aware Correlation Consistency

In semantic segmentation tasks, pixel class labels typically depend on the object regions to which they belong, making it essential to understand object-level relationships. When constructing supervisory signals for unlabeled data, these object-level relationships should be fully considered. To this end, we propose Region-Structure Aware Correlation Consistency (RSCC), which provides strong supervisory signals that align with the nature of semantic segmentation tasks by explicitly modeling object-level relationships. RSCC achieves this through two key components: Cross-Image Region-Region Correlation Consistency (CIR²R) and Pixel-Region Correlation Consistency (PRCC). By incorporating object-level relationships into the S4 pipeline, RSCC enhances the model's ability to understand complex scenes in RSIs.

Cross-Image Region-Region Correlation Consistency

CIR²R aims to compute feature correlations across image regions, explicitly model object-level semantic relationships and capture patterns shared between ground objects. By establishing object-level semantic consistency across augmented images, this approach captures richer semantic descriptions and strengthens the supervisory signal of unlabeled data. Specifically, for a batch of unlabeled images $\{x^u\}_{u=1}^n$, we apply weak augmentation to obtain the augmented image set $\{x_u^w\}_{u=1}^n$. The segmentation model extracts region features $\{R_i^w\}_{i=1}^{N_b}$ for the weakly augmented images, where R_i^w represents the feature of the i-th region. We calculate the relationship matrix $S^w \in \mathbb{R}^{N_b \times N_b}$ for all regions in the weakly augmented images. Each element of S^w is computed as:

$$S_{ij}^w = \frac{(R_i^w)^\top R_j^w}{\|R_i^w\| \cdot \|R_j^w\|} \tag{18}$$

where $\|\cdot\|$ denotes the L_2 norm, and S_{ij}^w represents the semantic similarity between region i and region j in the weakly augmented images. Similarly, we apply strong augmentation to the same batch of unlabeled images, obtaining the augmented image set $\{x_u^s\}_{i=1}^n$. The segmentation model extracts region features $\{R_i^s\}_{i=1}^{N_s}$, and the relationship matrix $S^s \in \mathbb{R}^{N_b \times N_b}$ is calculated. Finally, the loss function of this part is computed as the MSE between the similarity matrices of the weakly and strongly augmented images over all batches:

$$L_{\text{CIR}} = \frac{1}{B}\sum_{b=1}^{B} \|S_b^w - S_b^s\|_2^2 \tag{19}$$

where B denotes the total number of batches in the unlabeled dataset, S_b^w and S_b^s are the similarity matrices of weakly and strongly augmented images in batch b, respectively, and $\|\cdot\|_2^2$ represents the squared L_2 norm (i.e., MSE).

Pixel-Region Correlation Consistency

Unlike CIR²R, which focuses on the correlations between image regions, PRCC emphasizes the semantic consistency among pixels within a region. This enables the model to capture the complete shape of ground objects and more superior object-level features, thus improving the understanding of object-level semantic relationships. To enhance pixel consistency within a region, a straightforward approach is to minimize the feature distance between any two pixels in the region. However, this approach requires computing pairwise distances between all pixels, leading to substantial computational overhead, especially for large areas. To address this, we use regional features as a proxy for pixel features, indirectly promoting pixel consistency by aligning regional and pixel features. This strategy significantly reduces the computational burden and improves the efficiency of model training.

For each region i and pixel j within that region, the similarity between pixel feature F_j and the region feature R_i is computed as:

$$E_{ij} = e^{-\|R_i - F_j\|_2} \tag{20}$$

where $\|\cdot\|_2$ is the Euclidean norm. The goal is to maximize the similarity between each pixel and its corresponding region feature. The loss function is defined as:

$$L_{\text{PR}} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{j\in\text{region } i} \log E_{ij} \tag{21}$$

4 Experiments

4.1 Experiment Setup

Dataset. We evaluated the performance of the proposed RegionMatch method using three widely used remote sensing (RS) semantic segmentation datasets: Vaihingen [Rottensteiner *et al.*, 2012], Potsdam [Rottensteiner *et al.*, 2012], and LoveDA [Wang *et al.*, 2021]. The Vaihingen dataset consists of 33 images, with sizes ranging from 1996×1995 pixels to 3816×2550 pixels. For our experiments, 16 images are used for training, and 17 images are used for testing. The Potsdam dataset contains 38 images, each with a size of 6000×6000 pixels. Of these, 24 images are used for training and 14 for testing. The validation categories for both the Vaihingen and Potsdam datasets include impervious surfaces, buildings, low vegetation, trees, and cars. The LoveDA dataset comprises 1024×1024 images from urban and rural areas, with categories including background, building, road, water, barren land, forest, and agriculture. The dataset contains 2522 training images and 1669 testing images. For all datasets, we crop the images into 512×512 pixel sections.

Evaluation Metric. The segmentation performance is evaluated using the mean Intersection over Union (mIoU) on the union of the sets.

Implementation Details. All experiments were conducted on an NVIDIA 4090 GPU. The DeeplabV3+ [Chen *et al.*, 2018] segmentation model with a ResNet50 [He *et al.*, 2016] backbone was employed. Stochastic Gradient Descent (SGD) was used as the optimizer. The base learning rate was set to 0.002, weight decay was set to 0.0001, and momentum was set to 0.9. The total number of training epochs was set to 80. Data augmentation strategies follow the FixMatch [Sohn *et al.*, 2020]. τ_{low} and τ_{high} are set to 0.85 and 0.95, respectively. λ_1 and λ_2 are set to 2 and 1, respectively. α_0 is set to 20%.

4.2 Comparison with State-of-the-art Methods

Results on Vaihingen. Table 1 presents the experimental results of different semi-supervised semantic segmentation methods on the Vaihingen dataset. Under the labeled data

Table 1: Comparisons of state-of-the-art methods on the Vaihingen with mIoU (%) metric.

Method	1/8	1/4	1/2
Supervised	61.52	65.45	68.62
WSCL [Lu *et al.*, 2023]	67.51	68.59	69.22
AEL [Hu *et al.*, 2021b]	67.81	69.80	71.64
U2PL [Wang *et al.*, 2022b]	65.87	67.82	69.14
FixMatch [Sohn *et al.*, 2020]	66.69	68.73	69.41
UniMatch [Yang *et al.*, 2023a]	66.01	69.38	71.44
CorrMatch [Sun *et al.*, 2024a]	66.36	68.29	69.35
MPMC [Howlader *et al.*, 2025]	66.44	68.56	70.02
DWL[Huang *et al.*, 2024]	68.58	70.29	71.72
Ours	**69.25**	**71.20**	**72.68**

protocols of 1/8, 1/4, and 1/2, our method achieves mIoU scores of 69.25%, 71.20%, and 72.68%, respectively, significantly outperforming existing semi-supervised methods. Compared to supervised learning using only labeled data, our method improves by 7.38%, 5.75%, and 4.06% under the 1/8, 1/4, and 1/2 labeled data settings. These results demonstrate that our approach effectively leverages the advantages of semi-supervised learning, significantly enhancing the model's generalization ability.

Table 2: Comparisons of state-of-the-art methods on the Potsdam with mIoU (%) metric.

Method	1/8	1/4	1/2
Supervised	74.78	76.71	78.15
WSCL [Lu *et al.*, 2023]	76.74	77.38	78.53
AEL [Hu *et al.*, 2021b]	78.09	79.06	79.13
U2PL [Wang *et al.*, 2022b]	77.39	77.87	78.39
FixMatch [Sohn *et al.*, 2020]	77.85	78.23	78.93
UniMatch [Yang *et al.*, 2023a]	79.08	79.30	79.44
CorrMatch [Sun *et al.*, 2024a]	78.01	78.68	79.11
MPMC [Howlader *et al.*, 2025]	78.31	78.90	79.23
DWL [Huang *et al.*, 2024]	78.30	78.56	78.91
Ours	**80.24**	**80.68**	**81.17**

Results on Potsdam. The results on the Potsdam dataset are shown in Table 2, where we compare our method with state-of-the-art semi-supervised semantic segmentation approaches. Under the 1/8, 1/4, and 1/2 labeled data protocols, our method achieves mIoU scores of 80.24%, 80.68%, and 81.17%, respectively. These results significantly outperform methods that rely on pixel-level relationships [Sun *et al.*, 2024a] or contextual information from image patches [Howlader *et al.*, 2025], indicating that our approach more effectively unleashes the potential of unlabeled data from an object-level perspective. Furthermore, compared to supervised learning, our method improves mIoU by 5.46%, 3.97%, and 3.02% under the 1/8, 1/4, and 1/2 labeled data protocols, respectively, providing further evidence of its efficacy.

Results on Loveda. Table 3 presents a comparison of various methods on the Loveda dataset. Our approach consistently outperforms existing state-of-the-art methods across different annotated data protocols. Specifically, under the 1/16 labeled data protocol, our method achieves 52.55% mIoU, representing 5.23% improvement over supervised learning. Under the 1/8 protocol, our method attains 52.96% mIoU, surpassing supervised learning by 4.45%. At the 1/4 and 1/2 protocols, our method continues to lead, achieving 53.37% and 53.96% mIoU, respectively, exceeding the supervised approach by 3.33% and 3.71%. These results highlight our method's superior ability to leverage unlabeled data.

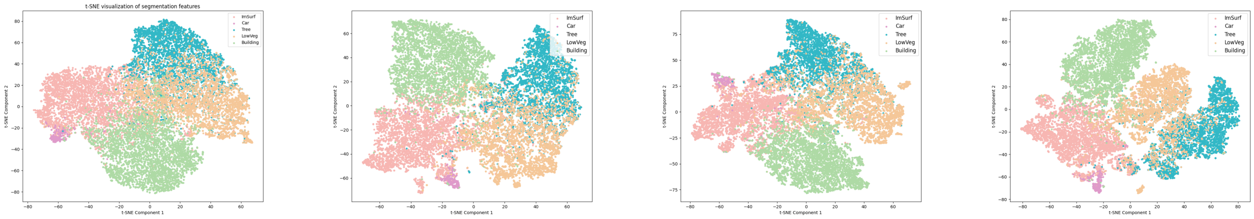
Table 3: Comparisons of state-of-the-art methods on the Loveda with mIoU (%) metric.

Method	1/16	1/8	1/4	1/2
Supervised	47.32	48.51	50.04	50.25
WSCL [Lu *et al.*, 2023]	49.68	50.23	51.74	52.40
AEL [Hu *et al.*, 2021b]	47.68	49.24	50.80	51.76
U2PL [Wang *et al.*, 2022b]	48.36	49.97	50.89	51.83
FixMatch [Sohn *et al.*, 2020]	50.36	50.42	51.37	52.66
UniMatch [Yang *et al.*, 2023a]	51.22	51.42	52.08	52.75
CorrMatch [Sun *et al.*, 2024a]	48.94	50.41	52.11	52.97
MPMC [Howlader *et al.*, 2025]	49.05	50.73	51.76	52.52
DWL[Huang *et al.*, 2024]	51.68	51.86	52.92	53.30
Ours	**52.55**	**52.96**	**53.37**	**53.96**

4.3 Ablations Studies

Effectiveness of Components. Table 4 illustrates the impact of various components on performance using the Vaihingen dataset. As shown, the proposed components yield significant improvements across all protocols. By collaboratively generating pseudo-labels from both pixel-level and region-level perspectives during model training, PRSPL facilitates knowledge sharing and mitigates the negative effects of erroneous pseudo-labels. Fig. 4 shows the influence of PRSPL on pseudo-label accuracy under the 1/8 labeled data protocol. The results confirm that PRSPL significantly improves pseudo-label accuracy, validating its effectiveness. Fig. 3 illustrates the T-SNE feature visualization on the Vaihingen dataset under the 1/8 protocol for different components. The results indicate that incorporating PRSPL allows the model to capture region-level contextual information, which is beneficial for pixel classification and enhances class separability. RSCC further strengthens the model's understanding of object-level semantic relationships by establishing multi-level relationships both within and between pixels and regions, aiding the model in learning more compact intra-class feature representations. Finally, the combination of PRSPL and RSCC enables the model to achieve both improved inter-class separation and more compact intra-class representations.

Impact of Region Classification. Table 5 demonstrates the impact of object region classification on the Vaihingen dataset. It can be observed that incorporating unlabeled data yields greater benefits for object region classification compared to labeled data. Due to the absence of labels, unlabeled data is more challenging to classify, as it relies on useful contextual information. By introducing the region classification task, contextual information of object regions can be supplemented, which aids in pixel classification and subsequently



(a) Baseline (b) Baseline + PRSPL (c) Baseline + RSCC (d) Ours

Figure 3: Visualization of t-SNE features of different components on the Vaihingen dataset under 1/8 labeled data protocol.

Table 4: Ablation studies of different components.

PRSPL	CIR^{2}R	PRCC	1/8	1/4	1/2
			66.69	68.73	69.41
✓			67.98	69.90	71.47
✓	✓		68.74	70.65	72.02
✓		✓	68.56	70.57	71.92
✓	✓	✓	69.25	71.20	72.68

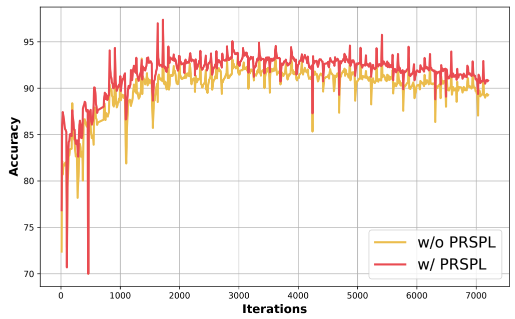


Figure 4: The accuracy of pseudo-labels under the 1/8 labeled protocol on the Vaihingen dataset.

improves model performance.

Table 5: Data setting in region classification.

Labeled	Unlabeled	1/8	1/4	1/2
✓		68.17	70.35	71.98
	✓	68.53	70.74	72.07
✓	✓	69.25	71.20	72.68

Threshold Strategy in PRSPL. In PRSPL, we adopt a dynamic entropy threshold strategy. Table 6 presents the impact of dynamic thresholds on the Vaihingen dataset. It can be observed that the dynamic threshold plays a crucial role in model performance. Since deep learning models first memorize clean labels and then gradually memorize noisy labels, a lower threshold can be set initially to provide more training data. A higher threshold is applied in later stages to prevent the model from fitting noisy data.

Table 6: Threshold strategy in PRSPL.

Strategy	1/8	1/4	1/2
Fixed	68.38	70.74	72.15
Dynamic	69.25	71.20	72.68

Table 7: Labeling strategy in region classification.

Strategy	1/8	1/4	1/2
Hard	68.98	70.84	72.37
Soft	68.76	70.69	72.11
Hard+Soft	69.25	71.20	72.68

Labeling Strategy in Region Classification. We employed a combined labeling strategy of both hard and soft labels for region classification. Table 7 illustrates the impact of this strategy on the Vaihingen dataset. When using the τ_{low} threshold, neither the hard labels nor the soft labels alone achieved optimal performance. The best performance was attained when both soft and hard labels were used in combination.

5 Conclusion

This paper presents a novel semi-supervised semantic segmentation method, RegionMatch, effectively leveraging unlabeled data from a fresh object-level perspective. We propose the Pixel-Region Synergy Pseudo-Labeling strategy to explicitly inject contextual information in the S4 pipeline and encourage knowledge cooperation from pixel and region perspectives. In addition, our proposed Region-Structure Aware Correlation Consistency (RSCC) models object-level relationships by establishing the multi-level correlation between regions and pixels, providing stronger supervision signals for unlabeled data that fit the semantic segmentation task. Extensive experiments on multiple authoritative remote sensing datasets demonstrate that our method effectively improves the potential of unlabeled data.

References

[Bai *et al.*, 2024] Lubin Bai, Haoyu Wang, Xiuyuan Zhang, Wei Qin, Bo Liu, and Shihong Du. Ap-semi: Improving the semi-supervised semantic segmentation for vhr images through adaptive data augmentation and prototypical sample guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

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

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

[Chen *et al.*, 2021] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021.

[Ding *et al.*, 2019] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019.

[Ding *et al.*, 2023] Chuanghao Ding, Jianrong Zhang, Henghui Ding, Hongwei Zhao, Zhihui Wang, Tengfei Xing, and Runbo Hu. Decoupling with entropy-based equalization for semi-supervised semantic segmentation. In *IJCAI*, pages 663–671, 2023.

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

[Howlader *et al.*, 2025] Prantik Howlader, Srijan Das, Hieu Le, and Dimitris Samaras. Beyond pixels: Semi-supervised semantic segmentation with a multi-scale patch-based multi-label classifier. In *European Conference on Computer Vision*, pages 342–360. Springer, 2025.

[Hu *et al.*, 2021a] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16291–16301, 2021.

[Hu *et al.*, 2021b] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021.

[Huang *et al.*, 2024] Wei Huang, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. Decouple and weight semi-supervised semantic segmentation of remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212:13–26, 2024.

[Ke *et al.*, 2020] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020.

[Liu *et al.*, 2022] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4258–4267, 2022.

[Liu *et al.*, 2024] Ruizhong Liu, Tingzhang Luo, Shaoguang Huang, Yuwei Wu, Zhen Jiang, and Hongyan Zhang. Crossmatch: Cross-view matching for semi-supervised remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Lu *et al.*, 2023] Xiaoqiang Lu, Licheng Jiao, Lingling Li, Fang Liu, Xu Liu, Shuyuan Yang, Zhixi Feng, and Puhua Chen. Weak-to-strong consistency learning for semisupervised image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[Luo *et al.*, 2024] Yuan Luo, Bin Sun, Shutao Li, and Yulong Hu. Hierarchical augmentation and region-aware contrastive learning for semi-supervised semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Lv and Zhang, 2024] Liang Lv and Lefei Zhang. Advancing data-efficient exploitation for semi-supervised remote sensing images semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Mai *et al.*, 2024] Huayu Mai, Rui Sun, Tianzhu Zhang, and Feng Wu. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3391–3401, 2024.

[Na *et al.*, 2024] Jaemin Na, Jung-Woo Ha, Hyung Jin Chang, Dongyoon Han, and Wonjun Hwang. Switching temporary teachers for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

[Qiao *et al.*, 2023] Pengchong Qiao, Zhidan Wei, Yu Wang, Zhennan Wang, Guoli Song, Fan Xu, Xiangyang Ji, Chang Liu, and Jie Chen. Fuzzy positive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2023.

[Rottensteiner *et al.*, 2012] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; I-3*, 1(1):293–298, 2012.

[Shen *et al.*, 2020] Dingguo Shen, Yuanfeng Ji, Ping Li, Yi Wang, and Di Lin. Ranet: Region attention network for semantic segmentation. *Advances in Neural Information Processing Systems*, 33:13927–13938, 2020.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[Sun *et al.*, 2024a] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3097–3107, 2024.

[Sun *et al.*, 2024b] Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: exploring the better weighting function for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

[Teh *et al.*, 2022] Eu Wern Teh, Terrance DeVries, Brendan Duke, Ruowei Jiang, Parham Aarabi, and Graham W Taylor. The gist and rist of iterative self-training for semi-supervised segmentation. In *2022 19th Conference on Robots and Vision (CRV)*, pages 58–66. IEEE, 2022.

[Wang *et al.*, 2021] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[Wang *et al.*, 2022a] Xiaoyang Wang, Jimin Xiao, Bingfeng Zhang, and Limin Yu. Card: Semi-supervised semantic segmentation via class-agnostic relation based denoising. In *IJCAI*, pages 1451–1457, 2022.

[Wang *et al.*, 2022b] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022.

[Wang *et al.*, 2023a] Changqi Wang, Haoyu Xie, Yuhui Yuan, Chong Fu, and Xiangyu Yue. Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 931–942, 2023.

[Wang *et al.*, 2023b] Xiaoyang Wang, Bingfeng Zhang, Limin Yu, and Jimin Xiao. Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3114–3123, 2023.

[Wang *et al.*, 2024] Xiaoyang Wang, Huihui Bai, Limin Yu, Yao Zhao, and Jimin Xiao. Towards the uncharted: Density-descending feature perturbation for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3303–3312, 2024.

[Xin *et al.*, 2024] Yi Xin, Zide Fan, Xiyu Qi, Yidan Zhang, and Xinming Li. Confidence-weighted dual-teacher networks with biased contrastive learning for semi-supervised semantic segmentation in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Yang *et al.*, 2022] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4268–4277, 2022.

[Yang *et al.*, 2023a] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023.

[Yang *et al.*, 2023b] Zhujun Yang, Zhiyuan Yan, Wenhui Diao, Qiang Zhang, Yuzhuo Kang, Junxi Li, Xinming Li, and Xian Sun. Label propagation and contrastive regularization for semi-supervised semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[Yang *et al.*, 2025] Lihe Yang, Zhen Zhao, and Hengshuang Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Yu *et al.*, 2020] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12416–12425, 2020.

[Zhang *et al.*, 2020] Jia Zhang, Zhixin Li, Canlong Zhang, and Huifang Ma. Robust adversarial learning for semi-supervised semantic segmentation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 728–732. IEEE, 2020.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[Zhao *et al.*, 2018] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018.

[Zhou *et al.*, 2021] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7036–7045, 2021.