ScaleMatch: Multi-scale Consistency Enhancement for Semi-supervised Semantic Segmentation

Liang Lv , Lefei Zhang*

National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
{lianglyu, zhanglefei}@whu.edu.cn

Abstract

Semi-supervised learning improves semantic segmentation performance by leveraging unlabeled data, thereby significantly reducing labeling costs. Previous semi-supervised semantic segmentation (S4) methods explored perturbations at the image level but neglected to adequately utilize multi-scale information. When labeled information is insufficient, the scale variation between different objects makes learning instances with extreme scales even more difficult. To address this issue, we propose ScaleMatch, which aims to learn scale-invariant features by obtaining a mixed dual-scale pseudo-label and scale consistency learning. Specifically, the cross-scale interaction fusion (CIF) module enforces interactive information across different scaled-views, allowing for more reliable pseudo-label generation. More importantly, ScaleMatch introduces variable scale branches to utilize scale-invariant supervision. It consists of image-level scale variation consistency (ISVC) and feature-level scale variation consistency (FSVC). Consequently, our ScaleMatch enhances the model's generalization under scale variation, outperforming existing state-of-the-art methods on both the Pascal VOC and Cityscapes datasets under various partition protocols.

Code — https://github.com/lvliang6879/ScaleMatch

Introduction

Semantic segmentation (SS) is a dense prediction task aiming to assign pixel-level classification labels in an image. It is a fundamental task in computer vision, and plays an important role in various fields, including autonomous driving, medical image analysis, and remote sensing image perception (Minaee et al. 2021; Shamshad et al. 2023; Zhang and Zhang 2022). Tremendous success has been achieved by deep learning algorithms in SS, due to the support of large-scale, accurately pixel-annotated datasets. However, obtaining pixel-wise annotation, which is time-consuming and labor-intensive, is extremely challenging compared to image classification or object detection.

To reduce dependence on labor-intensive manual labeling, semi-supervised semantic segmentation (S4) methods

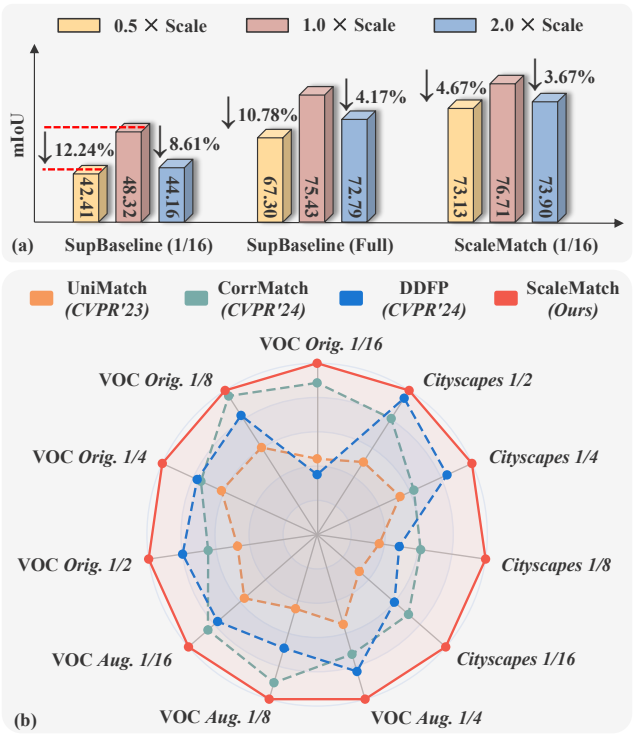

Figure 1: (a) Comparison of mIoU variations on the Pascal VOC validation set between models trained with full supervision using 1/16 labeled data and all labeled data, as well as models trained with the ScaleMatch method using 1/16 labeled data. The input scaling factors are 0.5, 1.0, and 2.0. (b) Comparison with state-of-the-art methods on the Pascal VOC and Cityscapes datasets.

(Yang et al. 2023; Chen et al. 2021; Wang et al. 2022) are well-studied to enhance the segmentation performance using limited labeled data and a large amount of unlabeled data. Early S4 research encompassed the transition from GAN-based methods (Souly, Spampinato, and Shah 2017; Mittal, Tatarchenko, and Brox 2019) to the widespread adoption of consistency regularization (Sohn et al. 2020; Zhang et al. 2021) frameworks. Existing S4 methods further consider new data augmentation (Fang et al. 2023; Zhao et al. 2023b)

*Corresponding author.

techniques or mutual learning between multiple models (Li et al. 2023b; Na et al. 2024) to improve the performance.

Although introducing a large amount of unlabeled data (Lu et al. 2020) helps improve the model's performance, these methods are not specifically designed for the SS tasks, resulting in limited performance gains. Compared to image classification (He et al. 2022), segmenting objects in the SS tasks involves multi-scale variations across different images. Current fully supervised segmentation models can handle this issue by learning on large-scale labeled data. However, for the S4 model, learning multi-scale information becomes much more challenging as the labeled data is scarce. As shown in Figure 1 (a), when we scale or enlarge the input image twice, the model's performance drops significantly under 1/16 labeled data. In contrast, models trained with full supervision exhibit a slight performance degradation and demonstrate greater robustness to scale variations. This phenomenon indicates that when annotated data is insufficient, *S4 models are sensitive to the scale variation*.

To address the aforementioned issues, we incorporate multi-resolution inputs into the weak-to-strong consistency learning framework to enhance pixel-level semantic consistency of the same image targets at different scales. With this spirit, one intuitive solution is introducing both a high-resolution and a low-resolution branch in S4, and this simple strategy achieves significant performance promotion. However, incorporating multi-scale inputs significantly increases computational and memory burden. Thus, to fully exploit multi-scale information without substantially increasing computational burden, we propose a novel and elegant S4 framework, ***ScaleMatch***.

ScaleMatch devises a cross-scale interaction fusion (CIF) module to generate high-quality pseudo-labels. The CIF module combines the Vision-RWKV (Duan et al. 2024) to fully mine information from different scales to generate spatial activation maps for the various scale predictions. By adaptively fusing pseudo-labels from different scales, CIF produces more accurate pseudo-labels, further guiding model training. To further exploit the scale consistency, we develop an image-feature level consistency to facilitate model learning. Specifically, ***ScaleMatch*** introduces variable-scale branches at both the image and feature levels. The outputs of the branches are supervised by pseudo-labels to ensure image-level scale variation consistency (ISVC) and feature-level scale variation consistency (FSVC). This approach allows the model to learn rich scale information while mitigating the GPU memory overhead caused by multiple resolution inputs.

Our contributions can be summarized as follows:

- We reveal that when labeled data is insufficient, segmentation models struggle to learn target scale variations in images.
- We propose a simple and effective S4 framework, ***ScaleMatch***, to address the challenge of learning scale variations.
- ***ScaleMatch*** leverages the CIF module to adaptively fuse pseudo-labels from different scale views, improving the quality of the pseudo-labels. Additionally, ***ScaleMatch*** includes ISVC and FSVC to enhance scale-invariance learning.
- Extensive experiments on two widely recognized benchmarks, PASCAL VOC and Cityscapes, demonstrate that the proposed ***ScaleMatch*** significantly outperforms existing methods (as illustrated in Figure 1 (b)).

Related Work

Semi-Supervised Image Semantic Segmentation (S4) aims to enhance the performance of SS models by leveraging limited labeled data and a large amount of unlabeled data. Current research typically focuses on areas such as data augmentation, co-training, and improving the quality of pseudo labels. For instance, CPS (Chen et al. 2021) introduces two differently initialized models that exchange pseudo labels to achieve cross-supervision. ST++ (Yang et al. 2022) gradually and selectively generates pseudo labels to ensure high-quality self-training (Wei et al. 2022). U2PL (Wang et al. 2022) extracts negative samples from unreliable predictions and contrasts them with positive samples to further mine information from pseudo labels. UniMatch (Yang et al. 2023) explores consistency between perturbed features and original features at the feature level by introducing random channel dropout. Additionally, CorrMatch (Sun et al. 2024a) leverages correlation maps to achieve enhanced label propagation, yielding significant improvements. AllSpark (Wang et al. 2024a) employs a channel-wise cross-attention mechanism to better utilize the features of labeled data. Despite the progress, existing methods neglect the challenges faced by segmentation models when learning multi-scale information in scenarios with limited labeled data. Our goal is to explore and mitigate this issue in S4 methods.

Multi-scale Training techniques have been developed to address the significant challenges posed by scale variations in vision-related tasks. In semi-supervised object detection (SSOD), objects with extreme scale tend to have low confidence, which makes them miss the pseudo labels for most SSOD methods due to strict filtering conditions. Some methods try to add consistent regularization on varying scales. SED (Guo et al. 2022) enhances model robustness by distilling predictions between a standard view and a reduced-scale view. Pseco (Li et al. 2022) uses a down-sampled perspective and modifies the feature pyramid layer, enabling the reuse of identical-scale pseudo boxes as those used in a standard view. These methods have inspired us to introduce scale-vary consistency into the S4, learning representations at multiple scales while reducing the computational cost of multi-scale training and achieving excellent results.

Methodology

Preliminaries and Overview

In S4, the training data includes both labeled and unlabeled sets. The labeled set is $\{x_i^l, y_i^l\}_{i=1}^{B_l}$, where $x_i^l \in \mathbb{R}^{H\times W\times 3}$ is an input image and $y_i \in \mathbb{R}^{H\times W\times K}$ is the pixel-wise label with K classes. The unlabeled set is $\{x_j^u\}_{j=1}^{B_u}$. According

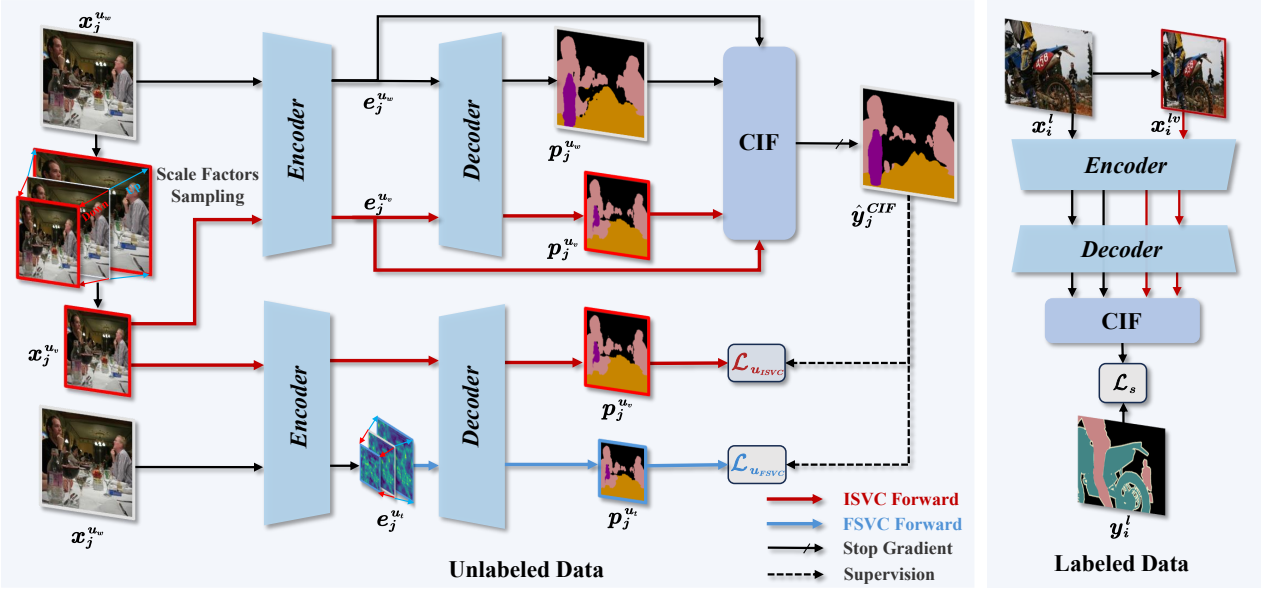

Figure 2: An overview of the proposed ScaleMatch. Besides two scale-variantion consistency regularization (ISVC and FSVC), ScaleMatch adopts cross-scale interaction fusion (CIF) strategies for feature enhancement. CIF combines the outputs from different scale inputs and adaptively fuses the pseudo-labels from these scales, resulting in more accurate pseudo-labels to guide model training. ISVC and FSVC enforce that objects in the same image maintain consistent pixel-level semantic representations under different scale inputs.

to the FixMatch (Sohn et al. 2020) framework, each unlabeled image x_j^u first undergoes weak augmentation $\mathcal{A}_w(x_j^u)$ to obtain the weakly-augmented image $x_j^{u_w}$. This weakly-augmented image is then further processed with strong augmentation $\mathcal{A}_s(x_j^{u_w})$ to obtain the strongly-augmented image $x_j^{u_s}$. The loss function for semi-supervised segmentation is:

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_{u_s} \tag{1}$$

where $\mathcal{L}_s$ and $\mathcal{L}_{u_s}$ denote supervised and unsupervised cross-entropy losses, respectively, and λ is coefficient. The supervised loss $\mathcal{L}_s$ and unsupervised loss $\mathcal{L}_{u_s}$ are respectivelu given by:

$$\mathcal{L}_s = \frac{1}{B_l} \sum_{i=1}^{B_l} \mathcal{L}_{ce}(y_i^l, p_i) \tag{2}$$

$$\mathcal{L}_{u_s} = \frac{1}{B_u} \sum_{j=1}^{B_u} \mathbb{1}(\max(p_j^{u_w}) \geq \tau) \mathcal{L}_{ce}(\hat{y}_j, p_j^{u_s}) \tag{3}$$

Where τ is the confidence threshold, $p_j^{u_w} = f(x_j^{u_w})$ and $p_j^{u_s} = f(x_j^{u_s})$ are the predicted probability maps, and $\hat{y}_j = \arg\max(p_j^{u_w})$ is the pseudo-label. The segmentation model $f(\cdot)$ includes an encoder $g(\cdot)$ and a decoder $h(\cdot)$.

Many mainstream S4 methods are based on the aforementioned framework. However, existing methods neglect the challenge of learning from extreme-scale variations in semi-supervised scenarios. In this paper, we explore multi-scale consistency in S4 and leverage scale information to obtain reliable pseudo-labels. As shown in Figure 2, the proposed ***ScaleMatch*** consists of three parts: A CIF module, ISVC and FSVC. We detail these parts sequentially.

Cross-scale Interaction Fusion

In SS tasks, multi-scale reasoning significantly enhances segmentation results. By introducing scale-variant branches and obtaining predictions from different resolutions, we can obtain high-quality pseudo-label. Different from conventional methods that average predictions across scales, this paper aims to exploit assigning a soft weight to each pixel, giving higher weight to more accurate predictions for precise pseudo-labels.

To this aim, we propose the Cross-scale Interaction Fusion (CIF) module. During training, this module uses feature maps from different resolutions to learn spatial activation maps for each prediction. As shown in Figure 3, the CIF module includes two key components: VRWKV-based Cross-scale Feature Interaction (VCFI) and the Scale-Aware Gate (SAG).

VCFI. Specifically, we first set up a group of scale factors $V = [v_1, v_2, \ldots, v_n]$. During each iteration of training, we randomly sample from V:

$$v = \text{Sample}(V) \tag{4}$$

Once obtaining the scale factor v_s, we interpolate a weakly-augmented image $x_j^{u_w} \in \mathbb{R}^{H \times W \times 3}$ with v:

$$x_j^{u_v} = \text{Interpolate}(x_j^{u_w}, [H^v, W^v]) \tag{5}$$

where Interpolate uses bilinear interpolation, with H^v and W^v representing the height and width scaled by a factor of v, $x_j^{u_v} \in \mathbb{R}^{H^v \times W^v \times 3}$ denotes the scaled image.

Next, the weakly augmented image $x_j^{u_w}$ and the scaled image $x_j^{u_v}$ are fed into the encoder $g(\cdot)$ to extract their

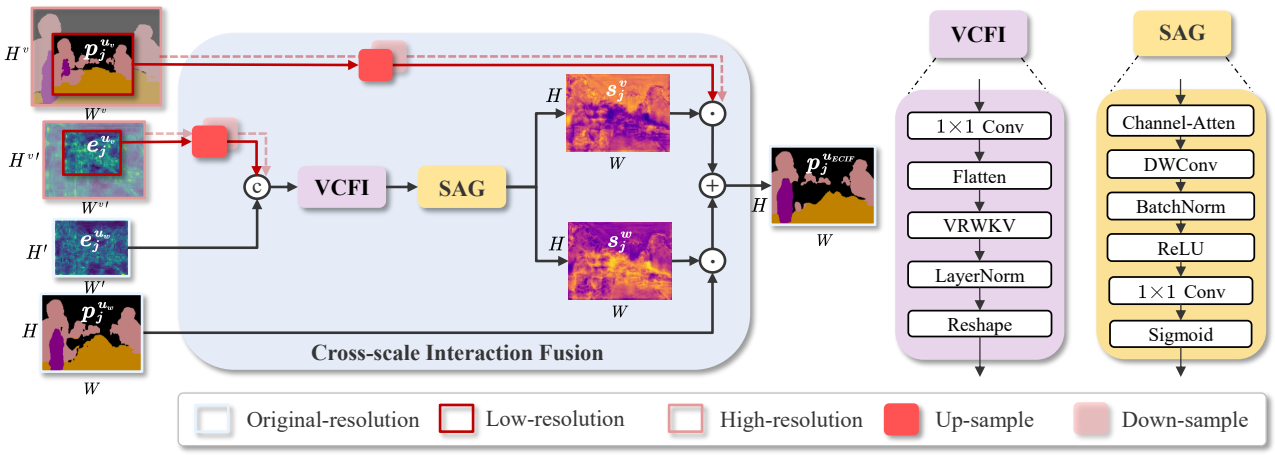

Figure 3: The details of the Cross-scale Interaction Fusion (CIF) module. VCFI: VRWKV-based Cross-scale Feature Interaction. SAG: Scale-Aware Gate.

respective feature maps $e_j^{u_w} \in \mathbb{R}^{H' \times W' \times 3}$ and $e_j^{u_v} \in \mathbb{R}^{H^{v'} \times W^{v'} \times 3}$.

Then, $e_j^{u_w}$ and $e_j^{u_v}$ are fed into the decoder $h(\cdot)$ and generate the prediction probability maps $p_j^{u_w} \in \mathbb{R}^{H \times W \times K}$ and $p_j^{u_v} \in \mathbb{R}^{H^v \times W^v \times K}$.

Meanwhile, to facilitate interaction between the feature maps at different scales, we interpolate $e_j^{u_v}$ to match the dimensions of $e_j^{u_w}$:

$$e_j^{u_{vw}} = \text{Interpolate}(e_j^{u_v}, [H', W']) \tag{6}$$

Since the resolution of x_j^v varies, we need to choose either upsampling or downsampling interpolation based on different scale factors, as indicated by the red squares in Figure 3. Next, we concatenate the two feature maps:

$$e_j^{u_{cat}} = \text{Concat}(e_j^{u_w}, e_j^{u_{vw}}) \tag{7}$$

The concatenated feature map $e_j^{u_{cat}}$ is then input into the VCFI module for feature interaction computation:

$$e_j^{u_F} = \text{Flatten}\left(\text{Conv}_{1\times1}\left(e_j^{u_{cat}}\right)\right) \tag{8}$$

$$e_j^{u_I} = \text{Reshape}\left(\text{LN}\left(\text{VRWKV}(e_j^{u_F})\right)\right) \tag{9}$$

where the $\text{Conv}_{1\times1}$ layer compresses the channel dimensions, Flatten unrolls the features for computation, $e_j^{u_F}$ represents the flattened features, VRWKV refers to Vision-RWKV (Yuan et al. 2024) layers works for capturing global information with linear complexity. $e_j^{u_I}$ represents the interacted features, which are reshaped back to the original input feature map's dimensions after LayerNorm (LN).

SAG. Next, we interpolate $e_j^{u_I}$ to the original resolution of $x_j^{u_w}$. The scaled $e_j^{u_{Iw}}$ is then input into the SAG module to obtain the spatial activation map, which can be expressed as:

$$e_j^{u_{Iw}} = \text{DWConv}(\text{CA}(\text{Interpolate}(e_j^{u_I}, [H, W]))), \tag{10}$$

where DWConv and Channel-Attention (CA) (Chen, Zhang, and Zhang 2023) are utilized to facilitate information exchange between channels and assign greater importance to significant channels:

$$s_j^w = \sigma(\text{Conv}_{1\times1}(\text{RELU}(\text{BN}(e_j^{u_{Iw}})))) \tag{11}$$

Here, the spatial activation map s_j^w corresponding to $x_j^{u_w}$ is generated using the $\text{Conv}_{1\times1}$. This map is mapped to the range $(0, 1)$ using the sigmoid function σ as a gate function. For the spatial activation map of s_j^v, we have $s_j^v = 1 - s_j^w$.

Based on the attention weight in Eq. 11, we can obtain the final prediction probability maps $p_j^{u_w}$ and $p_j^{u_{vw}}$:

$$p_j^{u_{CIF}} = s_j^w \odot p_j^{u_w} + s_j^v \odot p_j^{u_{vw}} \tag{12}$$

$$\hat{y}_j^{CIF} = \arg\max(p_j^{u_{CIF}}) \tag{13}$$

where $p_j^{u_{vw}} \in \mathbb{R}^{H \times W \times K}$ is the result of interpolating $p_j^{u_v}$. $\odot$ denotes the element-wise product. Finally, the fused pseudo-label $\hat{y}_j^{CIF}$ is used as a supervisory signal to guide the model's learning on unlabeled data. It is worth noting that we also use CIF in the supervised training of labeled data, where real labels are used to supervise and update the parameters of CIF, following a calculation process identical to the one described above.

Scale-variation Consistency Learning

To further address the challenge of learning target scale variations in SS models due to limited labeled data, ***ScaleMatch*** enforces pixel-level semantic consistency in segmentation predictions across multi-scale views of the same image, referring to scale-variation consistency learning. To avoid the memory overhead associated with merging multiple resolution inputs, we introduce scale variation branches. These branches ensure multi-scale invariance by continuously varying their input or feature resolution. To enrich the representation of the same target at different scales, we employ different levels of scale variation consistency as regularization, including **ISVC** and **FSVC**.

Image-level Scale Variation Consistency (ISVC). We use the pseudo-labels $\hat{y}_j^{CIF}$ from the CIF module to supervise the output $p_j^{u_{vw}}$, calculating the image-level scale vari-

ation consistency loss, which can be formulated as:

$$\mathcal{L}_{u_{ISCV}} = \frac{1}{B_u}\sum_{j=1}^{B_u}\mathbb{1}(\max(p_j^{u_w}) \geq \tau)\mathcal{L}_{ce}(\hat{y}_j^{CIF}, p_j^{u_{vw}}) \quad (14)$$

Feature-level Scale Variation Consistency (FSVC). To build a broader perturbation space, UniMatch (Yang et al. 2023) leverages perturbations to the features of weakly augmented images, achieving impressive results. Based on this finding, we further introduce scale variations to the features of weakly augmented images to enhance scale variation consistency learning. We retain the dropout perturbations for features as used in UniMatch, and we predefine a set of scale variation factors $T = [t_1, t_2, \ldots, t_n]$ for the feature map and sample a factor t. We then interpolate the feature map according to the factor t:

$$e_j^{u_t} = \text{Interpolate}(e_j^{u_w}, [H^t, W^t]) \quad (15)$$

The scaled feature map $e_j^{u_t}$ is fed into the network decoder $h(\cdot)$, producing the corresponding output $p_j^{u_t}$. The output is then aligned back to the original resolution through interpolation, resulting in $p_j^{u_{tw}}$, in order to calculate the feature-level scale variation consistency loss:

$$\mathcal{L}_{u_{FSCV}} = \frac{1}{B_u}\sum_{j=1}^{B_u}\mathbb{1}(\max(p_j^{u_w}) \geq \tau)\mathcal{L}_{ce}(\hat{y}_j^{CIF}, p_j^{u_{tw}}) \quad (16)$$

Therefore, the overall objective function is the sum of the loss functions from Eq. 1, Eq. 14, and Eq. 16:

$$\mathcal{L} = \mathcal{L}_s + \lambda_1\mathcal{L}_{u_s} + \lambda_2\mathcal{L}_{u_{ISCV}} + \lambda_3\mathcal{L}_{u_{FSCV}} \quad (17)$$

where λ_1, λ_2, and λ_3 are the weights of each loss.

Experiments

Datasets. We conduct experiments on two widely-used datasets, Pascal VOC 2012 and Cityscapes.

- **Pascal VOC 2012** is a SS benchmark (Everingham et al. 2015), consisting of 1,464 high-quality annotated images for training and 1,449 images for evaluation. Additionally, we perform experiments on the augmented Pascal VOC 2012 dataset, which includes coarsely annotated images from the SBD (Hariharan et al. 2011), totaling 10,582 training images.
- **Cityscapes** is designed for semantic analysis of urban street scenes and includes 2,975 high-resolution images for training and 500 images for validation, primarily covering 19 categories within urban environments. Following prior research (Yang et al. 2023), we evaluate both datasets using various label partitions.

Implementation Details. Consistent with previous research (Sun et al. 2024a), we employ DeepLabV3+ (Chen et al. 2018) as our network architecture and use a ResNet-101 pretrained on ImageNet as the backbone. For the Pascal dataset, we use an SGD optimizer with an initial learning rate of 0.001, weight decay of 1e-4, and crop sizes of either 321×321 or 513×513, with batch sizes of 16 and 8, respectively, over a total of 80 training epochs, and a confidence threshold τ of 0.95. For the Cityscapes dataset, we also use an Adamw optimizer with an initial learning rate of

PASCAL *Original*	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full(1464)
SupBaseline | 513×513	48.32	56.20	66.65	71.32	75.43
SemiCVT [CVPR'23]	68.56	71.26	74.99	78.54	80.32
FPL [CVPR'23]	69.30	71.72	75.73	78.95	-
CCVC [CVPR'23]	70.20	74.40	77.40	79.10	80.50
AugSeg [CVPR'23]	71.09	75.45	78.80	80.33	81.36
iMAS [CVPR'23]	68.80	75.30	79.10	80.20	82.00
DGCL [CVPR'23]	70.47	77.14	78.73	79.23	81.55
DAW [NeurIPS'23]	74.80	77.40	79.50	80.60	81.50
ESL [ICCV'23]	70.97	74.06	78.14	79.53	81.77
LogicDiag [ICCV'23]	73.25	76.66	77.93	79.39	-
AllSpark [CVPR'24]	76.07	78.41	79.77	80.75	82.12
RankMatch [CVPR'24]	75.50	77.60	79.80	80.70	82.20
DDFP [CVPR'24]	74.95	78.01	79.51	81.21	81.96
ScaleMatch	**76.71**	**78.64**	**80.53**	**82.02**	**83.01**
ScaleMatch†	**77.80**	**79.72**	**81.46**	**83.15**	**84.43**
SupBaseline | 321×321	45.10	55.30	64.80	69.70	73.50
ST++ [CVPR'22]	65.20	71.00	74.60	77.30	79.10
UniMatch [CVPR'23]	75.20	77.19	78.80	79.90	81.20
Diverse Co-T. [ICCV'23]	75.40	76.80	**79.60**	80.40	81.60
CorrMatch [CVPR'24]	**76.40**	78.50	79.40	80.60	**81.80**
ScaleMatch	76.13	**78.56**	79.55	**80.72**	**81.80**
ScaleMatch†	**77.07**	**79.77**	**80.32**	**81.93**	**83.04**

Table 1: Comparison with SOTAs on Pascal *original* dataset. The † indicates using multi-scale inference with the CIF, while SupBaseline represents supervised training using only labeled data.

PASCAL *Augmented*	1/16 (662)	1/8 (1323)	1/4 (2646)
SupBaseline | 513×513	67.24	70.57	73.85
CCVC [CVPR'23]	77.20	78.40	79.00
AugSeg [CVPR'23]	77.01	78.20	78.82
UniMatch [CVPR'23]	78.10	78.40	79.20
iMAS [CVPR'23]	77.20	78.40	79.30
ESL [ICCV'23]	76.36	78.57	79.02
DLG [ICCV'23]	77.75	79.31	79.14
RankMatch [CVPR'24]	**78.90**	79.20	80.00
DDFP [CVPR'24]	78.32	78.88	79.83
CorrMatch [CVPR'24]	78.40	79.30	79.60
ScaleMatch	78.56	**79.50**	**80.19**
ScaleMatch†	**79.41**	**80.28**	**80.72**
SupBaseline$^{\diamondsuit}$ | 513×513	70.60	75.00	76.50
U^{2}PL$^{\diamondsuit}$ [CVPR'22]	77.21	79.01	79.30
SemiCVT$^{\diamondsuit}$ [CVPR'23]	79.20	79.95	80.20
AugSeg$^{\diamondsuit}$ [CVPR'23]	79.30	81.50	80.50
UniMatch$^{\diamondsuit}$ [CVPR'23]	80.94	81.92	80.41
LogicDiag$^{\diamondsuit}$ [ICCV'23]	79.65	80.24	80.62
Dual Teacher$^{\diamondsuit}$ [NeurIPS'23]	80.10	81.50	80.50
AllSpark$^{\diamondsuit}$ [CVPR'24]	80.67	82.04	80.92
CorrMatch$^{\diamondsuit}$ [CVPR'24]	81.30	81.90	80.90
ScaleMatch$^{\diamondsuit}$	**81.49**	**82.75**	**81.14**
ScaleMatch$^{\diamondsuit}$†	**82.69**	**83.51**	**82.12**

Table 2: Comparison with SOTAs on Pascal *Augumented* dataset. The $\diamondsuit$ represents using the same split as U^{2}PL.

Cityscape	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
SupBaseline | 801×801	67.24	70.57	73.85	76.21
SemiCVT [CVPR'23]	72.19	75.41	77.17	79.55
CCVC [CVPR'23]	77.20	78.40	79.00	-
AugSeg [CVPR'23]	77.01	78.20	78.82	80.43
UniMatch [CVPR'23]	76.60	77.90	79.20	79.50
iMAS [CVPR'23]	77.20	78.40	79.30	-
DGCL [CVPR'23]	73.18	77.29	78.48	80.71
CSS [ICCV'23]	74.02	76.93	77.94	79.62
ESL [ICCV'23]	76.36	78.57	79.02	79.98
DLG [ICCV'23]	77.75	79.31	79.14	79.54
Diverse Co-T. [ICCV'23]	75.70	77.40	78.50	-
CFCG [ICCV'23]	77.28	79.09	80.07	80.59
DAW [NeurIPS'23]	76.60	78.40	79.80	80.60
Dual Teacher [NeurIPS'23]	76.80	78.40	79.50	80.50
RankMatch [CVPR'24]	77.10	78.60	80.00	80.70
DDFP [CVPR'24]	77.10	78.19	79.88	80.82
CorrMatch [CVPR'24]	77.30	78.50	79.40	80.40
ScaleMatch	**77.83**	**79.44**	**80.24**	**80.98**
ScaleMatch†	**78.70**	**80.42**	**81.47**	**82.01**

Table 3: Comparison with SOTAs on Cityscape dataset.

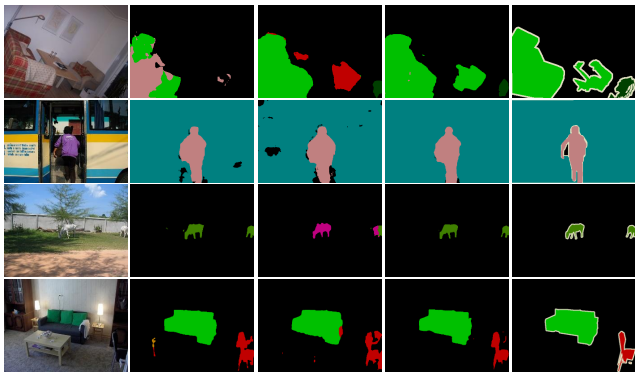

Figure 4: Visualization of the segmentation results on the Pascal validation set, in comparison with UniMatch and AllSpark.

0.00005, weight decay of 1e-2, a crop size of 801×801, and a batch size of 4, over a total of 240 training epochs with a confidence threshold τ of 0. The scale factors V are [0.25, 0.5, 1.5, 2.0], T are [0.75, 1.0, 1.25], and the weights for each loss function $\lambda_1, \lambda_2, \lambda_3$ are set to 0.25, 0.25, and 0.5, respectively.

In all experiments, we implement our proposed method using the PyTorch framework and perform computations on four NVIDIA RTX 4090 GPUs (24GB VRAM each). All segmentation performance evaluations are based on the mean Intersection over Union (mIoU).

Comparison with State-of-the-Art Methods. In this part, we demonstrate the outstanding performance of ***ScaleMatch*** with comparison of SOTAs on both datasets under different partition protocols including ST++ (Yang et al. 2022), U^{2}PL (Wang et al. 2022), SemiCVT (Huang et al. 2023), FPL (Qiao et al. 2023), CCVC (Wang et al. 2023c), AugSeg (Zhao et al. 2023b), DGCL (Wang et al. 2023b), UniMatch (Yang et al. 2023), ESL (Ma et al. 2023), LogicDiag (Liang et al. 2023), iMAS (Zhao et al. 2023a), DLG (Li et al. 2023a), Diverse Co-T. (Li et al. 2023c), Dual Teacher (Na et al. 2024), CSS (Wang et al. 2023a), CFCG (Li et al. 2023b), DAW (Sun et al. 2024b), CorrMatch(Sun et al. 2024a), DDFP (Wang et al. 2024b), RankMatch (Mai et al. 2024), and Allspark (Wang et al. 2024a). Furthermore, we also provide visual verification of the proposed method on the Pascal VOC validation set, as shown in Figure 4.

Pascal VOC 2012 dataset. We present the performance of our method alongside other state-of-the-art methods on the Pascal VOC 2012 original and augmented datasets in Table 1 and Table 2. Compared to the SupBaseline method, which trains a model without SSL, ***ScaleMatch*** achieves higher performance with only a few labeled images. Additionally, ***ScaleMatch*** demonstrates state-of-the-art performance across most experimental splits at both 513 and 321 resolutions. Specifically, on the **Pascal** *original* dataset, the proposed method outperforms existing SOTA methods by 0.64%, 0.23%, 0.73%, 0.81%, and 0.81% on each split, respectively. To further leverage the powerful multi-scale feature fusion capabilities provided by the CIF module, we conduct CIF-based multi-scale inference (combining 0.5× and 2.0× scales) for each experimental setting, denoted by †. The results show that integrating multi-scale fusion during inference further improves the model's segmentation performance. The visualization results in Figure 4 demonstrate that through scale-variation consistency learning, ***ScaleMatch*** exhibits better segmentation performance on objects of different scales compared to existing methods.

Cityscapes dataset. Table 3 compares our method with others on the Cityscapes validation set. Despite the complex street scenes, our method performs remarkably well across different partitions. ***ScaleMatch*** significantly outperforms existing SOTA methods by 0.53%, 0.13%, 0.17%, and 0.16% under the same partition protocols.

Ablation Studies

We conduct ablation studies to verify the proposed strategies in ***ScaleMatch***, reporting results for DeepLabV3+ with ResNet-101 on the Pascal VOC original dataset (513 × 513 training size).

Component Analysis. ***ScaleMatch*** comprises three components: ISVC (Image-level Scale Variation Consistency), FSVC (Feature-level Scale Variation Consistency), and CIF (Efficient Cross-scale Interaction Fusion). We evaluate their effectiveness in Table 4 through detailed ablation studies conducted at 1/4 and 1/2 splits. The ISVC component brings notable performance improvements, increasing scores by 2.94% and 3.17% at 1/4 and 1/2 splits compared to the baseline. Adding FSVC on top of ISVC further enhances performance by 1.01% and 0.51%. Finally, incorporating the CIF module achieves the highest performance, with mIoU reaching 80.53% and 82.02% at 1/4 and 1/2 scales, respectively. These results highlight the complementary nature of these components.

Baseline	ISVC	FSVC	CIF	1/4 (366)	1/2 (732)
✓	-	-	-	75.78	77.89
✓	✓	-	-	78.72	81.06
✓	✓	✓	-	79.73	81.57
✓	✓	✓	✓	**80.53**	**82.02**

Table 4: Effectiveness of the proposed components on Pascal *original* dataset. We leverage the FixMatch (Sohn et al. 2020) as our baseline.

Scale Factors Group	ISVC		MRI	
	mIoU	Memory	mIoU	Memory
Baseline	77.98	5.19	77.98	5.19
0.5	79.12	7.70	79.12	7.70
2.0	78.93	15.47	78.93	15.47
0.5, 2.0	80.86	15.41	80.83	17.50
0.25, 0.5, 1.5, 2.0	**81.06**	15.50	**81.08**	20.83
$[0.5, 2.0](random)$	79.66	15.58	-	-

Table 5: Comparison of mIoU and memory (GB) usage between ISVC and multi-resolution input (MRI) under different scale factor groups.

Method	Fusion Mode	Warm-up	1/2 (732)
w/o Multi-scale Fusion	-	-	81.57
w/ Multi-scale Fusion	Average Fusion	-	81.68
	CIF	-	81.83
	CIF	epoch = 10	82.02
	CIF	epoch = 20	81.95

Table 6: Ablation study of different multi-scale fusion modes.

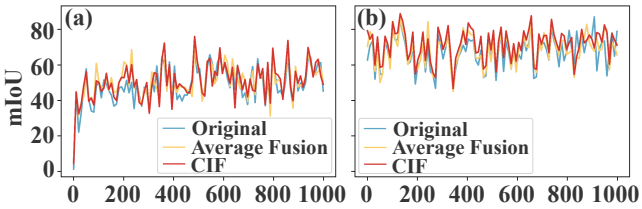

Figure 5: Comparison of pseudo-label mIoU at different training stages for a batch of data, (a) epoch=1, (b) epoch=10.

Different Scale Factors. To validate the impact of different scale factors, we conduct an ablation study with results shown in Table 5. Experimental results demonstrate that using ISVC to introduce single-scale branches (0.5× or 2.0×) yields performance improvement, but the gains are limited. For example, adding 0.5× to the baseline results in a 1.14% increase, while adding 2.0× results in a 0.95% increase. However, when large (2.0×) and small (0.5×) scales are introduced simultaneously, the performance promotion is more significant 2.88%. Further increasing the scale factors and using a combination of (0.25×, 0.5×, 1.5×, 2.0×), mIoU further improves to 81.06%. When expanding the scale factors by randomly sampling from the range [0.5, 2.0], we observed a decrease in mIoU to 79.66%. The potential reason is that excessive scale variation leads to learning difficulties for the model, thereby degrading performance. Therefore, using a limited combination of scale factors and simultaneously introducing both large and small scales can effectively enhance model performance. Additionally, we compare our method with multi-resolution input, which involves training with images of various resolutions directly. Both methods achieve high accuracy, however, multi-resolution input significantly increases memory consumption. In contrast, the memory usage of ISVC depends solely on the largest scale factor, enabling more efficient multi-scale training.

Effectiveness of CIF. We conduct a detailed quantitative analysis and visual comparison of the CIF module. As shown in Figure 5, we calculate the mIoU of pseudo-labels generated by each batch during training to observe the changes in the accuracy of the pseudo-labels. It is observed that in the early stages of training, specifically at epoch 1, the results using average fusion and CIF are often worse than the original pseudo-labels. This is because the model's performance is relatively poor at the beginning, even the fusion of predictions at different scales is hard to improve performance. After some training time (by the 10th epoch), the mIoU curves for average fusion and CIF are generally higher than those for the original pseudo-labels. Therefore, we implement a warm-up strategy for CIF in the training process. As shown in Table 6, the proposed CIF module achieves an mIoU of 81.83%. When warm-up is applied, starting from the 10th epoch, segmentation performance reaches 82.02%. However, extending the warm-up period to 20 epochs causes the performance to slightly decrease to 81.95%. This suggests starting CIF too early may introduce noise into the pseudo-labels, while starting too late does not fully leverage the higher-quality pseudo-labels. Thus, we choose to begin the warm-up at epoch 10 to make better use of the more accurate pseudo-labels.

Conclusion

In this paper, we discover that a major reason for the poor performance of current S4 methods is the difficulty segmentation models face in effectively learning multi-scale information when labeled data is limited. To address this issue, we propose a novel S4 framework called ***ScaleMatch***. ***ScaleMatch*** introduces an CIF module that adaptively fuses predictions at different scales, improving the quality of pseudo-labels and more effectively guiding model training. By incorporating scale variation branches, our framework reduces the computational burden of multi-resolution inputs while leveraging scale variation consistency at both the image and feature levels. This enhances the segmentation model's ability to learn multi-scale information and improves its robustness to scale variations. Extensive experimental results show that our method outperforms the current state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62122060. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

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

Chen, S.; Zhang, L.; and Zhang, L. 2023. Msdformer: Multiscale deformable transformer for hyperspectral image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2613–2622.

Duan, Y.; Wang, W.; Chen, Z.; Zhu, X.; Lu, L.; Lu, T.; Qiao, Y.; Li, H.; Dai, J.; and Wang, W. 2024. Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures. *arXiv preprint arXiv:2403.02308*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Fang, Y.; Zhu, F.; Cheng, B.; Liu, L.; Zhao, Y.; and Wei, Y. 2023. Locating noise is halfway denoising for semi-supervised segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16612–16622.

Guo, Q.; Mu, Y.; Chen, J.; Wang, T.; Yu, Y.; and Luo, P. 2022. Scale-Equivalent Distillation for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14522–14531.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 991–998.

He, R.; Han, Z.; Lu, X.; and Yin, Y. 2022. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14585–14594.

Huang, H.; Xie, S.; Lin, L.; Tong, R.; Chen, Y.-W.; Li, Y.; Wang, H.; Huang, Y.; and Zheng, Y. 2023. SemiCVT: Semi-supervised convolutional vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11340–11349.

Li, G.; Li, X.; Wang, Y.; Wu, Y.; Liang, D.; and Zhang, S. 2022. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *European Conference on Computer Vision*, 457–472.

Li, P.; Purkait, P.; Ajanthan, T.; Abdolshah, M.; Garg, R.; Husain, H.; Xu, C.; Gould, S.; Ouyang, W.; and Van Den Hengel, A. 2023a. Semi-supervised semantic segmentation under label noise via diverse learning groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1229–1238.

Li, S.; He, Y.; Zhang, W.; Zhang, W.; Tan, X.; Han, J.; Ding, E.; and Wang, J. 2023b. CFCG: Semi-Supervised Semantic Segmentation via Cross-Fusion and Contour Guidance Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16348–16358.

Li, Y.; Wang, X.; Yang, L.; Feng, L.; Zhang, W.; and Gao, Y. 2023c. Diverse Cotraining Makes Strong Semi-Supervised Segmentor. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16009–16021.

Liang, C.; Wang, W.; Miao, J.; and Yang, Y. 2023. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16197–16208.

Lu, X.; Wang, W.; Shen, J.; Tai, Y.-W.; Crandall, D. J.; and Hoi, S. C. 2020. Learning video object segmentation from unlabeled videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8960–8970.

Ma, J.; Wang, C.; Liu, Y.; Lin, L.; and Li, G. 2023. Enhanced soft label for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1185–1195.

Mai, H.; Sun, R.; Zhang, T.; and Wu, F. 2024. RankMatch: Exploring the Better Consistency Regularization for Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3391–3401.

Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; and Terzopoulos, D. 2021. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7): 3523–3542.

Mittal, S.; Tatarchenko, M.; and Brox, T. 2019. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4): 1369–1379.

Na, J.; Ha, J.-W.; Chang, H. J.; Han, D.; and Hwang, W. 2024. Switching temporary teachers for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*.

Qiao, P.; Wei, Z.; Wang, Y.; Wang, Z.; Song, G.; Xu, F.; Ji, X.; Liu, C.; and Chen, J. 2023. Fuzzy positive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15465–15474.

Shamshad, F.; Khan, S.; Zamir, S. W.; Khan, M. H.; Hayat, M.; Khan, F. S.; and Fu, H. 2023. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88: 102802.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020.

Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Souly, N.; Spampinato, C.; and Shah, M. 2017. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*, 5688–5696.

Sun, B.; Yang, Y.; Zhang, L.; Cheng, M.-M.; and Hou, Q. 2024a. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3097–3107.

Sun, R.; Mai, H.; Zhang, T.; and Wu, F. 2024b. DAW: exploring the better weighting function for semi-supervised semantic segmentation. *Advanc es in Neural Information Processing Systems*, 36.

Wang, C.; Xie, H.; Yuan, Y.; Fu, C.; and Yue, X. 2023a. Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 931–942.

Wang, H.; Zhang, Q.; Li, Y.; and Li, X. 2024a. AllSpark: Reborn Labeled Features from Unlabeled in Transformer for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3627–3636.

Wang, X.; Bai, H.; Yu, L.; Zhao, Y.; and Xiao, J. 2024b. 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*, 3303–3312.

Wang, X.; Zhang, B.; Yu, L.; and Xiao, J. 2023b. Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3114–3123.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4248–4257.

Wang, Z.; Zhao, Z.; Xing, X.; Xu, D.; Kong, X.; and Zhou, L. 2023c. Conflict-based cross-view consistency for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19585–19595.

Wei, Q.; Sun, H.; Lu, X.; and Yin, Y. 2022. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *European Conference on Computer Vision*, 516–532.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7236–7246.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4268–4277.

Yuan, H.; Li, X.; Qi, L.; Zhang, T.; Yang, M.-H.; Yan, S.; and Loy, C. C. 2024. Mamba or rwkv: Exploring high-quality and high-efficiency segment anything model. *arXiv preprint arXiv:2406.19369*.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419.

Zhang, L.; and Zhang, L. 2022. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2): 270–294.

Zhao, Z.; Long, S.; Pi, J.; Wang, J.; and Zhou, L. 2023a. Instance-specific and Model-adaptive Supervision for Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23705–23714.

Zhao, Z.; Yang, L.; Long, S.; Pi, J.; Zhou, L.; and Wang, J. 2023b. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11350–11359.