RankMatch: Exploring the Better Consistency Regularization for Semi-supervised Semantic Segmentation

Huayu Mai[1*] Rui Sun[1*] Tianzhu Zhang[1,2†] Feng Wu[1]
[1]University of Science and Technology of China
[2]Deep Space Exploration Lab
{mai556, issunrui}@mail.ustc.edu.cn, {tzzhang, fengwu}@ustc.edu.cn

Abstract

The key lie in semi-supervised semantic segmentation is how to fully exploit substantial unlabeled data to improve the model's generalization performance by resorting to constructing effective supervision signals. Most methods tend to directly apply contrastive learning to seek additional supervision to complement independent regular pixel-wise consistency regularization. However, these methods tend not to be preferred ascribed to their complicated designs, heavy memory footprints and susceptibility to confirmation bias. In this paper, we analyze the bottlenecks exist in contrastive learning-based methods and offer a fresh perspective on inter-pixel correlations to construct more safe and effective supervision signals, which is in line with the nature of semantic segmentation. To this end, we develop a coherent RankMatch network, including the construction of representative agents to model inter-pixel correlation beyond regular individual pixel-wise consistency, and further unlock the potential of agents by modeling inter-agent relationships in pursuit of rank-aware correlation consistency. Extensive experimental results on multiple benchmarks, including mitochondria segmentation, demonstrate that RankMatch performs favorably against state-of-the-art methods. Particularly in the low-data regimes, RankMatch achieves significant improvements.

1. Introduction

Semantic segmentation, which aims to explain visual semantics at the pixel level, has achieved conspicuous achievements attributed to the recent advances in deep neural network [28] as a fundamental task in computer vision with widespread applications such as visual understanding [11], autonomous driving [12], *etc*. However, its data-driven nature makes it labor-intensive and time-consuming to gather massive pixel-level annotations as training data. To alleviate the data-hunger issue, considerable works [19, 44, 57] have turned their attention to semi-supervised semantic segmentation task. However, since only limited labeled data is accessible, how to fully exploit a large amount of unlabeled data to improve the model's generalization performance by resorting to constructing effective supervision signals is thus extremely challenging.

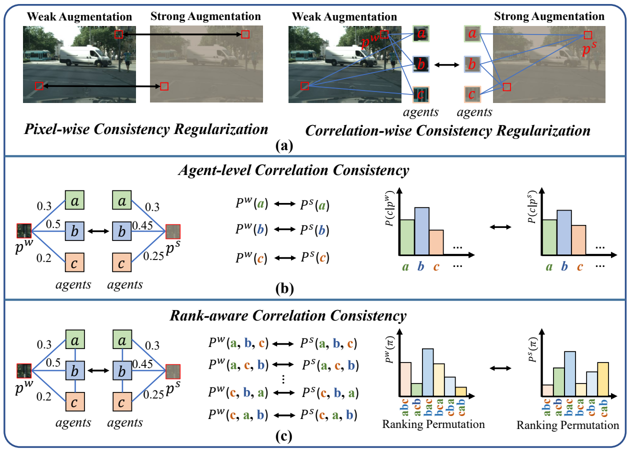
Weak Augmentation
Strong Augmentation
Weak Augmentation
Strong Augmentation
p^w
p^s
a
b
c
agents
agents
Pixel-wise Consistency Regularization
Correlation-wise Consistency Regularization
(a)
Agent-level Correlation Consistency
0.3
0.5
0.2
0.3
0.45
0.25
p^w
p^s
$P^w(a) \leftrightarrow P^s(a)$
$P^w(b) \leftrightarrow P^s(b)$
$P^w(c) \leftrightarrow P^s(c)$
$P(c|p^w)$
$P(c|p^s)$
a b c ...
(b)
Rank-aware Correlation Consistency
$P^w(a, b, c) \leftrightarrow P^s(a, b, c)$
$P^w(a, c, b) \leftrightarrow P^s(a, c, b)$
$P^w(c, b, a) \leftrightarrow P^s(c, b, a)$
$P^w(c, a, b) \leftrightarrow P^s(c, a, b)$
$P^w(\pi)$
$P^s(\pi)$
abc acb bac bca cba cab
Ranking Permutation
Ranking Permutation
(c)

Figure 1. Illustration of our motivation. (a) shows the differences between independent pixel-wise consistency regularization and correlation-wise consistency regularization. Evidently, taking the rich inter-pixel correlation into account can bring rich extra supervision. (b) shows the straightforward implementation of correlation-wise consistency regularization, *i.e*., treating each agent independently. (c) shows our core idea, rank-aware correlation consistency. We harness the inter-agent relationship by considering every possible agent rank permutation probability.

In previous literature, pseudo-labeling [1, 22] and consistency regularization [2, 21] have emerged as mainstream paradigms to leverage unlabeled data for semi-supervised semantic segmentation. Recently, these two paradigms are typically encapsulated into a teacher-student scheme [46] (where the teacher and student can be identical), that is,

[*]Equal contribution
[†]Corresponding author

the teacher network with weakly augmented perturbation view generates corresponding *pseudo-labels* to instruct the student network under the presence of strongly augmented perturbation view, using a form of *pixel-wise consistency regularization* (see Figure 1 (a)).

After an in-depth analysis of the teacher-student scheme, we argue that constructing extra supervision from substantial unlabeled samples matters in semi-supervised semantic segmentation, which is intuitively sensible from the definition of the task itself; that is, empowering regular consistency regularization to adapt to dense pixel-level prediction rather than suffering from the supervision signals of limited capacity derived at the individual pixel level (Figure 1 (a)). Inspired by the recent popularity of representation learning [8, 18], it naturally comes into mind to directly apply contrastive learning [33, 48, 50, 52] to semi-supervised semantic segmentation to establish an ample set of positive/negative samples in the representation space, aiming at seeking additional supervision to complement independent pixel-wise consistency regularization.

Despite their promising results, these methods tend not to be preferred ascribed to complicated designs and heavy memory footprint raised by the existence of ad hoc numerous positive/negative samples, inevitably compressing their capability and compromising the inherent simplicity of the teacher-student scheme. Plus, considering the absence of ground truth in unlabeled data, the determination of positive/negative samples is entirely conditioned on the model's biased predictions (erroneous pseudo-labels), leading to confirmation bias [14]. To make matters worse, the corollary of error accumulation is inevitably amplified by inbuilt low-data regimes of semi-supervised semantic segmentation, hindering the generalization ability of the model.

In this paper, we analyze the bottlenecks exist in contrastive learning-based methods for improving pixel-wise consistency regularization, and offer a fresh perspective on inter-pixel correlations to construct more safe and effective supervision signals for robust semi-supervised semantic segmentation. Intuitively, most methods neglect the fact that dense pixel prediction task carries rich inter-pixel information beyond basic individual pixel-wise consistency, shedding light on the possibility of closer collaboration between the inter-pixel correlation and the consistency regularization (*i.e.*, *correlation-wise consistency regularization*, right part of Figure 1 (a)) to comprehensively probe unlabeled data. The main idea is, we prepend the **agent-level correlation consistency** through a set of representative reference points (referred to as *agents*) to model the inter-pixel correlation (see Figure 1 (b)). For each pixel from the weakly augmented or corresponding strongly augmented view, we can obtain the agent-level correlation (*i.e.*, a likelihood vector) by comparing this pixel with a set of agents. In essence, the agent-level correlation reflects the consensus among representative agents with a broader receptive field, thus it encodes a higher-order consistency regularization to adapt to dense pixel-level prediction. However, it is nontrivial to attain the appropriate agents without any supervision signals for training. Intuitively, the agents should resonate favorably with diverse semantic cues from the original pixels with a wide range of semantic contrast descriptions. To this end, we devise an orthogonal selection strategy to pick the most representative agents from the feature map, preserving as much critical information as possible in the original pixels. In this way, benefiting from the richer description of the data distribution in agent-level correlation, we can achieve better exploitation of the unlabeled data.

Based on the above discussion, it is natural to integrate the resultant agent-level correlation into the teacher-student scheme and impose consistency constraint resorting to KL divergence, *etc*. (Figure 1 (b)). However, such a straightforward constraint treats each agent independently and heavily relies on strong i.i.d. assumption, hindering the potential for further optimization of the model. In fact, there exist specific relationships between agents that should also be considered in the agent-level correlation consistency regularization. For example, as shown in Figure 1, *agent a* and *agent b* are two pixels that reside in the same car while *agent c* situates from the road. Thus, *agent a* should hold a tighter relationship with *agent b* than *agent c*. To harness the inter-agent relationship modeling structure information for more effective supervision signals, instead of taking each agent independently, we carefully design the **rank-aware correlation consistency** to strive to further unlock the potential of agents by imposing the agent-level correlation rich in inter-agent relationship to be consistent between the teacher and student networks (*i.e.*, weak and strong augmented views, Figure 1 (c)). The core idea is that we take the agent ranking as a random event rather than a deterministic permutation. For instance, the correlation between different agents and a given pixel p^w varies, which can be regarded as the probabilities in ranking. The probability of being ranked first of the *agent a* is 0.3 while 0.5 of the *agent b*. From this perspective, the ranking permutation reflects the relationship of agents w.r.t. the pixel. In this way, for a given pixel, we consider every possible rank permutation of the agents (*e.g.*, *abc*, *cba*, *etc*.), and transform the agent-level correlation into the agent-ranking probability distribution. By constraining the consistency of the agent-ranking probability distribution between teacher and student networks, the model can be guided by more effective supervision signals. Ultimately, we term our final model as *RankMatch*.

In this work, our contributions can be summarized as follows: (1) We analyze the bottlenecks exist in contrastive learning-based methods for improving pixel-wise consistency regularization, and offer a fresh perspective

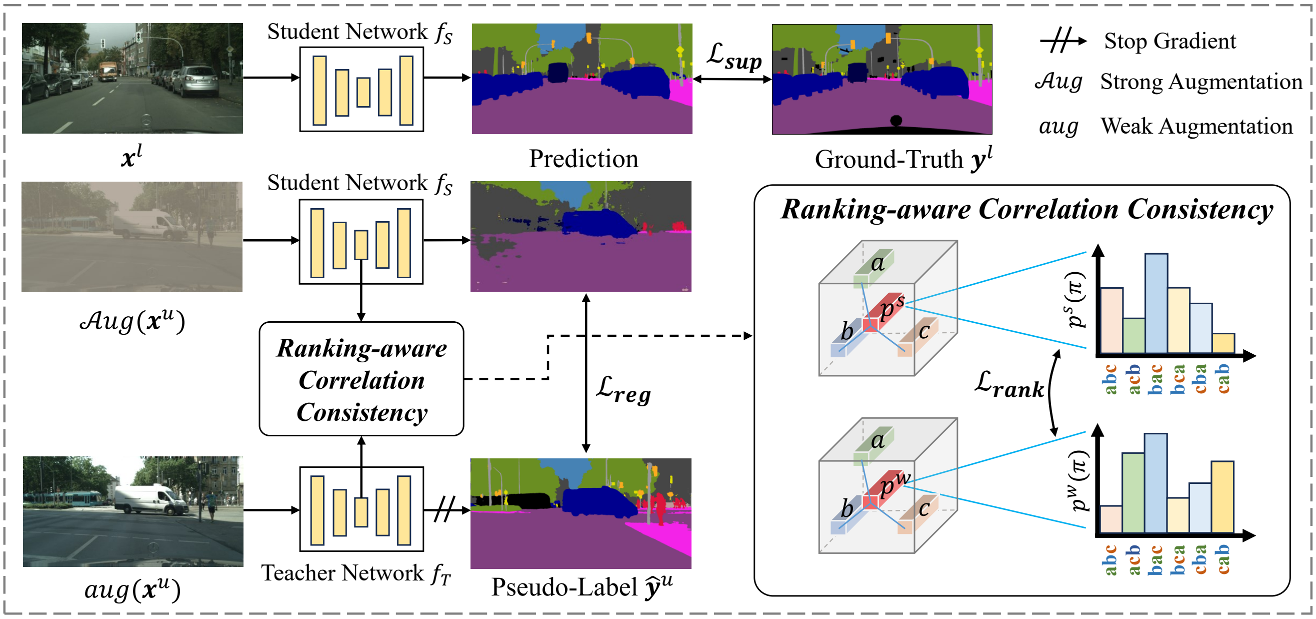

Figure 2. The framework of our RankMatch. The student network is guided by two sources of supervision, including the ground truth for the labeled data and the pseudo-labels generated by the teacher network for the unlabeled data. In previous consistency regularization methods, consistency is imposed at the pixel-level. While our work focuses on the rich correlations between pixels and imposes consistency constraint at the correlation-level. Furthermore, we design the ranking-aware correlation consistency for more effective supervision signals.

on inter-pixel correlations to construct more safe and effective supervision signals which is in line with the nature of semantic segmentation. (2) We develop a coherent RankMatch network, including the construction of representative agents to model inter-pixel correlation beyond regular individual pixel-wise consistency, and further unlock the potential of agents by modeling inter-agent relationships in pursuit of rank-aware correlation consistency. (3) Extensive experiments on three challenging benchmarks including mitochondria segmentation demonstrate that our RankMatch outperforms state-of-the-art semi-supervised semantic segmentation methods. Particularly in low-data regimes, RankMatch achieves significant improvements.

2. Related Work

Semi-supervised Learning. Semi-supervised learning [13, 37, 62] (SSL) is a well-studied topic and recent research can be summarized in two branches: Pseudo-labeling and consistency regularization. Pseudo-labeling [1, 5, 22, 59] methods involve training the model on unlabeled samples using pseudo-labels generated from the most up-to-date optimized model. On the other hand, consistency regularization-based [21, 46, 47, 55] methods leverage the smoothness assumption [32], encouraging the model to exhibit consistency when presented with the same example under different perturbations. Notably, recent SSL methods [3, 4, 15, 40, 56] have demonstrated the synergy between consistency regularization and Pseudo-labeling. One prominent example is FixMatch [40], which generates pseudo-labels from weakly augmented unlabeled images for strongly augmented versions of the same images. This concise yet powerful approach has gained widespread adoption in recent SSL studies.

Semi-supervised Semantic Segmentation. Benefits from the advances in deep neural network [30, 35, 41, 42, 45, 51, 53, 54] and various kinds of semi-supervised semantic segmentation (SSSS) algorithms [25, 27, 36, 44, 60, 61] have been proposed based on the mature combination of Pseudo-labeling and consistency regularization. Most of all, UniMatch [57] taking into account the nature of semantic segmentation tasks, incorporates suitable data augmentations into FixMatch, thus evolving into a concise yet powerful SSSS baseline. On top of these fundamental designs, motivated by representation learning, a series of works [33, 48, 50, 52] have incorporated contrastive learning into SSSS, tailoring it to the characteristics of the dense prediction task. In this paper, we offer a fresh perspective on inter-pixel correlations to construct more safe and effective supervision signals for robust semi-supervised semantic segmentation.

3. Method

In this section, we first formulate the semi-supervised semantic segmentation problem as preliminaries and introduce the core idea of the proposed RankMatch from the perspective of correlation. Then we describe the details of the construction of the agent-level correlation to mine more reliable information in the unlabeled data. Finally, rank-aware correlation consistency regularization is devised to harness the inter-agent relationship for more effective supervision signals. In Algorithm 1, we present the pseudo algorithm of our RankMatch to clearly summarize the method.

3.1. Preliminaries

Given a labeled set $\mathcal{D}^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N^l}$ and an unlabeled set $\mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$, where $N^u \gg N^l$, semi-supervised semantic segmentation aims to train a segmentation model with limited labeled data and vast unlabeled data. As shown in Figure 2, the popular teacher-student scheme consists of a teacher network f_T and a student network f_S. The student network is guided by two sources of supervision, including the ground truth for the labeled data and the pseudo-labels generated by the teacher network for the unlabeled data. The teacher network can either be identical to the student network or an exponentially moving average (EMA) version of the student network. In specific, for the labeled data, the supervised loss $\mathcal{L}_{sup}$ can be formulated as:

$$\mathcal{L}_{sup} = \frac{1}{N^l} \sum_{i=1}^{N^l} \frac{1}{HW} \sum_{j=1}^{HW} \ell_{ce} \left(\boldsymbol{y}_{ij}^l, f_S(\boldsymbol{x}_i^l)_j \right), \quad (1)$$

where H and W represent the height and width of the input image, ℓ_{ce} denotes the standard pixel-wise cross-entropy loss. For the unlabeled data, the teacher network takes the weak augmented view $aug(\boldsymbol{x}_i^u)$ as input and generates pseudo-labels $\hat{\boldsymbol{y}}_i^u$ for the student network as:

$$\hat{\boldsymbol{y}}_{ij}^u = \begin{cases} \arg\max f_T(aug(\boldsymbol{x}_i^u))_j, & c_{ij}^u > \gamma \\ \text{ignore\_index}, & \text{otherwise} \end{cases}, \quad (2)$$

where $c_{ij}^u = \max f_T(aug(\boldsymbol{x}_i^u))_j$ represents the confidence of the teacher prediction for j^{th} pixel and γ denotes the confidence threshold to exclude unreliable pseudo-labels from training. As result, we can obtain the consistency regularization loss $\mathcal{L}_{reg}$ as:

$$\mathcal{L}_{reg} = \frac{1}{N^u} \sum_{i=1}^{N^u} \frac{1}{HW} \sum_{j=1}^{HW} \ell_{ce} \left(\hat{\boldsymbol{y}}_{ij}^u, f_S(\mathcal{A}ug(\boldsymbol{x}_i^u))_j \right), \quad (3)$$

where $\mathcal{A}ug(\cdot)$ means the strong augmentation. By imposing consistency regularization, the model can learn reliable information from unlabeled data. The overall loss of the commonly used teacher-student scheme is $\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{reg}$.

Note that the above consistency regularization is operated at pixel-level, still stuck in the mindset of classification task. We contend that there exist substantial inter-pixel correlations within an image inherently, which should be taken into account in consistency regularization. What follows, we detail the process of modeling inter-pixel correlation.

3.2. Agent-level Correlation

To mine more reliable information in the unlabeled data, the idea arises naturally that we can impose consistency regularization at the correlation level, which is much richer than pixels. However, simply enforcing the correlations of all pixels (*i.e.*, the self-correlation matrix) to be consistent between teacher and student is not desirable. Lots of noise in the self-correlation matrix interferes with the optimization of the model, leading to a sub-optimal result.

In order to better model the inter-pixel correlation for consistency regularization, we construct the agent-level correlation by comparing each pixel with a set of representative reference points (referred to as agents). Intuitively, the agents should resonate favorably with diverse semantic cues from original pixels with a wide range of semantic contrast descriptions. For this purpose, we design an orthogonal selection strategy to pick the most representative agents from the image. Specifically, we obtain the feature map $\boldsymbol{F} \in \mathbb{R}^{C \times h \times w}$ for an unlabeled image $\boldsymbol{x}^u$ extracted by the feature extractor of the segmentation model. Then, we incrementally build a set of agents $\boldsymbol{A} = \{\boldsymbol{f}_i^a\}_{i=1}^N \in \mathbb{R}^{C \times N}$ sampled from the $\boldsymbol{F}$ such that a new agent is maximally orthogonal (*i.e.*, minimal cosine similarity) to the agents already selected, starting with a pixel feature at random, where N denotes the number of agents. This greedy strategy is dynamic, since it selects agents from the feature of the current image, preserving as much critical information as possible in the original pixels.

In this way, we can get the agent-level correlation $\boldsymbol{c} \in \mathbb{R}^{1 \times N}$ (omit the subscript i, j for convenience), *i.e.*, pixel-agent-level correlation for a given pixel feature $\boldsymbol{f}$ by

$$\boldsymbol{c} = softmax(\boldsymbol{f}\boldsymbol{A}^\mathsf{T}), \quad (4)$$

where T refers to the matrix transpose operation. Straightforwardly, we can impose the consistency regularization between the agent-level correlation $\boldsymbol{c}^w$ of teacher network and the $\boldsymbol{c}^s$ of student network resorting to KL divergence as:

$$\mathcal{L}_{corr} = \frac{1}{N^u} \sum_{i=1}^{N^u} \frac{1}{HW} \sum_{j=1}^{HW} \ell_{kl} \left(\boldsymbol{c}_{ij}^w, \boldsymbol{c}_{ij}^s \right). \quad (5)$$

However, such a naive constraint treats each agent independently hindering the potential for further model optimization. In the next, we introduce rank-aware correlation consistency regularization to model the specific relationships between agents.

3.3. Rank-aware Correlation Consistency

To harness the inter-agent relationship for more effective supervision signals, we carefully design the rank-aware consistency regularization. The core idea is that we take the agent ranking as a random event rather than a deterministic

Algorithm 1 Pseudo algorithms of RankMatch.

```
Inputs: Labeled Set $\mathcal{D}^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N^l}$, Unlabeled Set $\mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$ $(N^u \gg N^l)$
Define: Teacher Network $f_T$, Student Network $f_S$, Weak Augmentation $aug(\cdot)$, Strong Augmentation $\mathcal{A}ug(\cdot)$
Output: Student Network $f_S$
for each batch of $(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)$, $\boldsymbol{x}_i^u$ in $\mathcal{D}_l$, $\mathcal{D}_u$ do
    # Labeled Data:
    Calculate $\mathcal{L}_{sup}$ for $f_S$ by Equation (1)                          ▷ Supervised Loss
    # Unlabeled Data:
    Obtain pseudo-labels from $f_T$ by Equation (2)
    Calculate $\mathcal{L}_{reg}$ for $f_S$ by Equation (3)      ▷ Pixel-wise Consistency Regularization Loss
    Obtain agents for $f_T$ and $f_S$ respectively through orthogonal selection strategy
    Calculate the agent-level correlation by Equation (4)
    Transform the agent-level correlation into agent-ranking probability distribution by Equation (6)
    Calculate $\mathcal{L}_{rank}$ for $f_S$ by Equation (8)    ▷ Rank-aware Correlation Consistency Regularization Loss
    Gradient backward $\mathcal{L}_{sup} + \mathcal{L}_{reg} + \lambda\mathcal{L}_{rank}$                              ▷ Update Model
end for
```

permutation. That is to say, every permutation of the agents exists with some probability rather than only the permutation from largest to smallest exists. The probability of one permutation $\pi \in \mathcal{P}$ $(|\mathcal{P}| = N!)$ given $\boldsymbol{c}$ can be derived as:

$$P(\pi|\boldsymbol{c}) = \prod_{n=1}^{N} \frac{\boldsymbol{c}_{\pi(n)}}{\sum_{n'=n}^{N} \boldsymbol{c}_{\pi(n')}}, \tag{6}$$

where $\pi(n)$ denotes the n^{th} agent index of this permutation. For example, suppose we have three agents: *a*, *b* and *c*. One permutation of these three agents is $\pi = (a, b, c)$. Based on the agent-level correlation $\boldsymbol{c}$, we can derive the probability of permutation π:

$$P(\pi|\boldsymbol{c}) = \frac{\boldsymbol{c}(a)}{\boldsymbol{c}(a)+\boldsymbol{c}(b)+\boldsymbol{c}(c)} \cdot \frac{\boldsymbol{c}(b)}{\boldsymbol{c}(b)+\boldsymbol{c}(c)} \cdot \frac{\boldsymbol{c}(c)}{\boldsymbol{c}(c)}. \tag{7}$$

From this perspective, the ranking permutation reflects the relationship of agents. By calculating the probabilities for all $|\mathcal{P}|$ permutations, we transform the agent-level correlation $\boldsymbol{c}$ into agent-ranking probability distribution $P(\mathcal{P}|\boldsymbol{c}) \in \mathbb{R}^{1\times|\mathcal{P}|}$, which has modeled the inter-agent relationship. In fact, if we calculate the full permutations for all N agents, the computational overhead is indeed unacceptable. For computational efficiency, we focus on the permutations of the top-4 agents for each pixel, based on our observation that in every agent-level correlation, the top-4 agents have occupied almost all weight. Then, the rank-aware correlation consistency regularization can be obtained by:

$$\mathcal{L}_{rank} = \frac{1}{N^u} \sum_{i=1}^{N^u} \frac{1}{HW} \sum_{j=1}^{HW} \ell_{kl} \left(P(\mathcal{P}|\boldsymbol{c}_{ij}^w), P(\mathcal{P}|\boldsymbol{c}_{ij}^s) \right). \tag{8}$$

Finally, the overall loss objective of our RankMatch is derived as:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{reg} + \lambda\mathcal{L}_{rank}, \tag{9}$$

where the λ is the trade-off weight.

4. Experiments

4.1. Experimental Setup

Datasets: (1) PASCAL VOC 2012 [11] is an object-centric semantic segmentation dataset, containing 20 object classes in the foreground and a background class with 1,464 and 1,449 finely annotated images for training and validation, respectively. Many researches [9, 19] augment the original training set (*i.e.*, *classic*) with additional 9,118 coarsely annotated images in SBD [16] to get a *blender* training set. (2) Cityscapes [10] is an urban scene understanding dataset consisting of 2,975 images for training and 500 images for validation. The initial 30 semantic classes are re-mapped into 19 classes for the semantic segmentation task.

Implementation Details: For a fair comparison, we use ResNet-50/101 [17] pretrained on ImageNet [20] as the backbone and DeepLabv3+ [7] as the decoder. The crop size is set as 513 × 513 for PASCAL and 801 × 801 for Cityscapes, respectively. We adopt stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001 for PASCAL and 0.005 for Cityscapes. Polynomial Decay learning rate policy is applied throughout the whole training. The strong augmentation $\mathcal{A}ug(\cdot)$ contains random color jitter, grayscale and Gaussian blur. The weak augmentation $aug(\cdot)$ consists of random crop, resize and horizontal flip. The features used to construct the correlation consistency are extracted from the output of the ASPP module [6] and the channel number is 256. We set the number of agents $N = 128$ and trade-off weight $\lambda = 0.1$ for all experiments. The model is trained for 80 epochs on PASCAL and 240 epochs on Cityscapes with a batch size of 8, using 8× RTX 3090 GPUs (memory is 24G/GPU).

Table 1. Quantitative results of different SSL methods on Pascal *classic* set. We report mIoU (%) under various partition protocols and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Method	ResNet-50					ResNet-101				
	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)
Sup.-only	44.0	52.3	61.7	66.7	72.9	45.1	55.3	64.8	69.7	73.5
FixMatch$_{\text{[NeurIPS'20]}}$ [40]	60.1	67.3	71.4	73.7	76.9	63.9	73.0	75.5	77.8	79.2
iMAS$_{\text{[CVPR'23]}}$ [60]	–	–	–	–	–	68.8	75.3	79.1	80.2	82.0
AugSeg$_{\text{[CVPR'23]}}$ [61]	64.2	72.1	76.1	77.4	78.8	71.0	75.4	78.8	80.3	81.3
DGCL$_{\text{[CVPR'23]}}$ [50]	–	–	–	–	–	70.4	77.1	78.7	79.2	81.5
CSS$_{\text{[ICCV'23]}}$ [48]	68.0	71.9	74.9	77.6	–	–	–	–	–	–
LOGICDIAG$_{\text{[ICCV'23]}}$ [25]	–	–	–	–	–	73.2	76.6	77.9	79.3	–
NP-SemiSeg$_{\text{[ICML'23]}}$ [49]	65.7	72.3	75.7	77.4	–	–	–	–	–	–
DAW$_{\text{[NeurIPS'23]}}$ [44]	68.5	73.1	76.3	78.6	79.7	74.8	77.4	79.5	80.6	81.5
Switch$_{\text{[NeurIPS'23]}}$ [36]	70.7	74.5	76.4	77.6	78.1	–	–	–	–	–
UniMatch$_{\text{[CVPR'23]}}$ [57]	67.4	71.9	75.3	78.0	79.3	73.5	75.4	78.7	80.2	81.9
RankMatch (Ours)	**71.6**	**74.6**	**76.7**	**78.8**	**80.0**	**75.5**	**77.6**	**79.8**	**80.7**	**82.2**
$\Delta\uparrow$	**+27.6**	**+22.3**	**+15.0**	**+12.1**	**+7.1**	**+30.4**	**+22.3**	**+15.0**	**+11.0**	**+8.7**

Table 2. Quantitative results of different SSL methods on Pascal *blender* set. We report mIoU (%) under various partition protocols and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Method	ResNet-50			ResNet-101		
	1/16(662)	1/8(1323)	1/4(2646)	1/16(662)	1/8(1323)	1/4(2646)
Sup.-only	62.4	68.2	72.3	67.5	71.1	74.2
FixMatch$_{\text{[NeurIPS'20]}}$ [40]	70.6	73.9	75.1	74.3	76.3	76.9
ST++$_{\text{[CVPR'22]}}$ [58]	72.6	74.4	75.4	74.5	76.3	76.6
U^{2}PL$_{\text{[CVPR'22]}}$ [52]	–	–	–	77.2	79.0	79.3
AugSeg$_{\text{[CVPR'23]}}$ [61]	74.6	75.9	77.1	77.0	77.3	78.8
iMAS$_{\text{[CVPR'23]}}$ [60]	75.9	76.7	77.1	77.2	78.4	79.3
CFCG$_{\text{[ICCV'23]}}$ [23]	75.0	77.1	77.7	76.8	79.1	79.9
NP-SemiSeg$_{\text{[ICML'23]}}$ [49]	73.4	76.5	76.7	–	–	–
DAW$_{\text{[NeurIPS'23]}}$ [44]	76.2	77.6	77.4	78.5	78.9	79.6
UniMatch$_{\text{[CVPR'23]}}$ [57]	75.8	76.9	76.8	78.1	78.4	79.2
RankMatch (Ours)	**76.6**	**77.8**	**78.3**	**78.9**	**79.2**	**80.0**
$\Delta\uparrow$	**+14.2**	**+9.6**	**+6.0**	**+11.4**	**+8.1**	**+5.8**

4.2. Comparison with State-of-the-art Methods

For parameter efficiency, we adopt the popular consistency regularization framework UniMatch [57] as our baseline, that is the teacher and student networks are identical. We evaluate our method on both PASCAL (*classic* and *blender*) and Cityscapes datasets with both ResNet-50 and ResNet-101 backbone under diverse partition protocols, and make exhaustive comparisons with the state-of-the-art methods [23–25, 33, 36, 40, 44, 48–50, 52, 58, 60, 61]. The consistently dominant performance under all partition protocols with different backbones on all datasets proves the effectiveness of our RankMatch.

Results on PASCAL. Table 1 and Table 2 show the comparison of our method with the SOTA methods on PASCAL *classic* and *blender* set. Compared with the supervised-only (*Sup.-only*) model, our method achieves considerable performance improvements, suggesting that the information in unlabeled data is effectively utilized in our method. Moreover, we consistently observe substantial performance gains when compared to the baseline method, *i.e.*, UniMatch. Specifically, our approach achieves 71.6% and 75.5% under $1/16(92)$ partition on *classic* set with the backbone ResNet-50 and ResNet-101, boosting the baseline by 4.2% and 2.0%, respectively. These results underscore the powerful information mining capability of RankMatch under the extremely scarce labeled data setting.

Results on Cityscapes. Table 3 presents a comparative result of RankMatch against the SOTA methods on the Cityscapes dataset. Specifically, with the backbone ResNet-50, RankMatch outperforms the *Sup.-only* model by 12.1%, 7.5%, 6.1% and 2.9% under $1/16$, $1/8$, $1/4$ and $1/2$ partition protocols, respectively. Furthermore, when compared with the recent and competitive contrastive method ESL [33], our method maintains superior performance, *e.g.*, 2.0% performance lift under $1/16$ partition protocol with the ResNet-101 backbone, showing the superiority of our method over contrastive learning.

Qualitative Results. We compare the qualitative results of our method with different SOTA methods on the PASCAL dataset. As shown in Figure 3, RankMatch shows more

Table 3. Quantitative results of different SSL methods on Cityscapes. We report mIoU (%) under various partition protocols and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Method	ResNet-50				ResNet-101			
	1/16(186)	1/8(372)	1/4(744)	1/2(1488)	1/16(186)	1/8(372)	1/4(744)	1/2(1488)
Sup.-only	63.3	70.2	73.1	76.6	66.3	72.8	75.0	78.0
FixMatch[NeurIPS'20] [40]	72.6	75.7	76.8	78.2	74.2	76.2	77.2	78.4
AEL[NeurIPS'21] [19]	74.0	75.8	76.2	–	75.8	77.9	79.0	80.3
AugSeg[CVPR'23] [61]	73.7	76.4	78.7	79.3	75.2	77.8	79.5	80.4
iMAS[CVPR'23] [60]	74.3	77.4	78.1	79.3	–	–	–	–
ESL[ICCV'23] [33]	–	–	–	–	75.1	77.1	78.9	80.4
Co-Train[ICCV'23] [24]	–	76.3	77.1	–	75.0	77.3	78.7	–
NP-SemiSeg[ICML'23] [49]	73.0	77.1	78.8	78.7	–	–	–	–
Switch[NeurIPS'23] [36]	–	–	–	–	76.8	78.4	79.4	80.5
DAW[NeurIPS'23] [44]	75.2	77.5	79.1	79.5	76.6	78.4	79.8	80.6
UniMatch[CVPR'23] [57]	75.0	76.8	77.5	78.6	76.6	77.9	79.2	79.5
RankMatch (Ours)	**75.4**	**77.7**	**79.2**	**79.5**	**77.1**	**78.6**	**80.0**	**80.7**
$\Delta\uparrow$	+12.1	+7.5	+6.1	+2.9	+10.8	+5.8	+5.0	+2.7

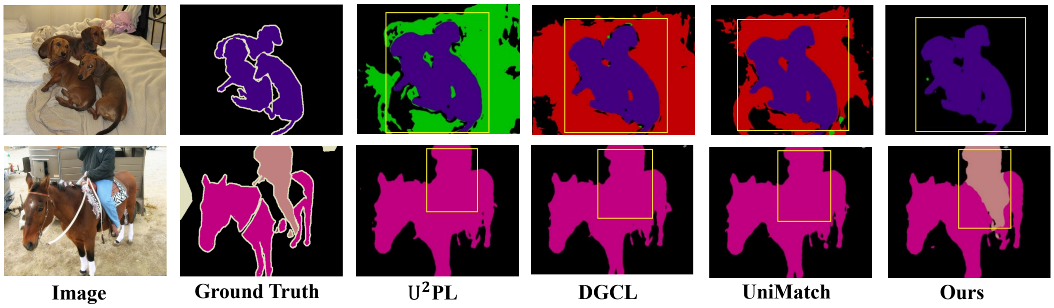

Figure 3. Qualitative comparison with different methods. Note that significant improvements are marked with yellow boxes.

Table 4. Ablation studies of different components.

Contrastive	*Correlation*	*Rank*	mIoU(92)	mIoU(1464)
			67.4	79.3
✓			68.6	79.6
	✓		70.3	79.5
	✓	✓	**71.6**	**80.0**

Table 7. Evaluation of the Agents number N.

Agents number N	mIoU
64	69.0
128	**71.6**
256	71.1
512	69.9

Table 8. Evaluation of the trade-off weight λ.

Trade-off weight λ	mIoU
0.05	70.8
0.1	**71.6**
0.2	71.2
0.5	70.9

Table 5. Ablation on different agent selection strategies.

Agent Selection	mIoU
All	69.8
Random	70.2
Top-N	70.6
Orthogonal	**71.6**

Table 6. Ablation on different correlation consistency.

Corr. Consis.	mIoU
L2	70.2
CE	70.1
KL	70.3
Rank-aware	**71.6**

powerful segmentation performance in fine-grained details (*e.g.*, the dogs on the bed and the man on horseback). With the help of rank-aware correlation consistency, RankMatch exhibits superior abilities in most scenarios.

4.3. Ablation Study and Analysis

To look deeper into our method, we perform a series of ablation studies on PASCAL *classic* set under $1/16(92)$ partition protocol with ResNet-50 to analyze our RankMatch.

Effectiveness of Components. In Table 4, we report the results of $1/16(92)$ and Full(1464) to clearly substantiate the effectiveness of our design. Note that, "*Contrastive*" denotes the reproduced results for U^{2}PL [52], a classic contrastive learning method in the semi-supervised semantic segmentation, based on our baseline (*i.e.*, UniMatch). The correlation-level consistency ("*Correlation*") without rank-aware ("*Rank*") means that treating each agent independently and straightforwardly imposing correlation-level regularization resorting to KL divergence, *i.e.*, $\mathcal{L}_{corr}$ in Equation (5). (1) Indeed, while contrastive learning can yield certain benefits for pixel-level consistency regularization baseline (1^{st} row *vs.* 2^{nd} row), it is still inferior to correlation-level consistency regularization (2^{nd} row *vs.* 3^{rd} & 4^{th} rows). (2) By comparing the results of the 3^{rd} and 1^{st} rows, a naive consideration of correlation-level con-

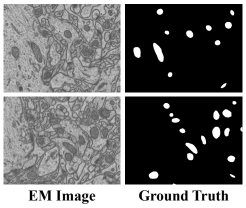

Figure 4. Visualization of the Lucchi dataset.

Method	*Spe.*	1/32(5)	1/16(10)	1/8(20)
Sup.-only		45.7	57.4	61.8
MT [46]	✗	71.8	72.4	75.4
CCT [38]	✗	84.7	85.4	85.8
CPS [9]	✗	84.5	84.6	85.8
DualRel [34]	✓	85.6	86.3	87.2
Ours	✗	**86.9**	**87.5**	**88.1**

Table 9. *Quant.* results of different SSL methods on Lucchi.

sistency reveals a significant performance improvement for pixel-level consistency regularization baseline. This observation indicates that the abundant correlations provide additional information gain for consistency regularization. (3) As 3^{rd} *vs.* 4^{th} row shows, harnessing the relationships between agents through the construction of the agent-ranking probability distribution can yield more effective supervision signals, manifesting in further performance improvements.

Effectiveness of Orthogonal Selection Strategy. In Table 5, we explore various strategies for agent selection, including "ALL" (considering all pixels in the feature map as agents), "Random" (randomly pick N pixels in the feature map as agents), "Top-N" (select top N pixels conditioned on the cumulative self-correlation matrix along the pixels dimension), and our proposed "Orthogonal". (1) Selecting all pixels as agents is not a desirable approach, as it inevitably introduces considerable noise among these pixels. This noise can adversely affect the quality of supervision signals, resulting in sub-optimal performance. (2) The strategy of "Orthogonal" achieves the best results, which is in line with our design purpose, that is, that representative agents can enjoy synergy with subsequent correlation-level consistency. We visualize the agent-pixel activation maps for those agents selected by "Orthogonal", as shown in Figure 5. It can be observed that the different agents activate different parts of the image, and resonate favorably with diverse semantic cues from the original pixels. These carefully selected agents retain as much critical information as possible in the original image, facilitating the subsequent construction of correlation consistency.

Effectiveness of Rank-aware Correlation Consistency. To investigate the effectiveness of rank-aware correlation consistency, we compare different modeling strategies for correlation consistency in Table 6. Among them, L2, CE, and KL belong to agent-independent correlation consistency, overlooking the inherent relationships between agents and resulting in sub-optimal performance. The proposed rank-aware correlation consistency achieve the best results, indicating that modeling the relationships between agents contributes to more effective supervision signals.

Hyperparameter Evaluations. (1) As shown in Table 7, it can be observed that the performance is optimal with $N = 128$. This result aligns with intuition, as too few agents can lead to information loss from the original image while too many can introduce noise into the training. Therefore, finding a balance for N is crucial. (2) λ controls the relative importance of the rank-aware correlation consistency loss, our model achieves much better performance when $\lambda = 0.1$ as shown in Table 8.

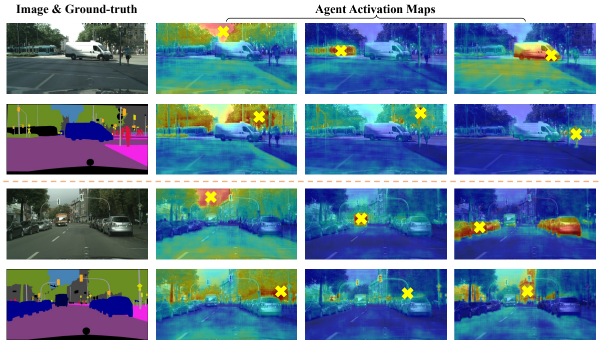

Figure 5. Visualization of the agent activation maps from our orthogonal selection for better illustration. The yellow cross denotes the position of agents in the original image.

Scalability for Other Scenarios. We extend our experimental evaluations on mitochondria segmentation [26, 27, 31, 39, 43] dataset Lucchi [29] to assess the scalability of our method. Figure 4 illustrates the images and ground truth of the Lucchi dataset, highlighting a common challenge in electron microscope images where instances are notably small and scattered. It underscores the need for more robust supervision during training within a semi-supervised framework. As depicted in Table 9, RankMatch exhibits superior performance compared to other competitive methods. Notably, our approach surpasses the specialized ("*spe.*") method DualRel [34] in the domain of electron microscopy images, underscoring the capability of our method to provide more rich and effective supervision.

5. Conclusion

In this paper, we offer a fresh perspective on inter-pixel correlations to construct more safe and effective supervision signals. To this end, We develop a coherent RankMatch network, including the construction of representative agents to model inter-pixel correlation beyond regular individual pixel-wise consistency, and further unlock the potential of agents by modeling inter-agent relationships in pursuit of rank-aware correlation consistency. Extensive experimental results on challenging benchmarks show the effectiveness.

6. Acknowledgments

This work was partially supported by the National Defense Basic Scientific Research Program of China (Grant JCKY2021601B013), National Nature Science Foundation of China (Grant 62021001), Youth Innovation Promotion Association CAS 2018166.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 1, 3

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 1

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3

[4] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*, 2021. 3

[5] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6912–6920, 2021. 3

[6] 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. 5

[7] 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. 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[9] 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. 5, 8

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1, 5

[12] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020. 1

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004. 3

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 2

[15] Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International Conference on Machine Learning*, pages 8082–8094. PMLR, 2022. 3

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 5

[17] 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. 5

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[19] 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. 1, 5, 7

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 3

[22] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 1, 3

[23] Shuo Li, Yue He, Weiming Zhang, Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, and Jingdong Wang. Cfcg: Semi-supervised semantic segmentation via cross-fusion and contour guidance supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16348–16358, 2023. 6

[24] Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*, 2023. 7

[25] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16197–16208, 2023. 3, 6

[26] Xiaoyu Liu, Bo Hu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Efficient biomedical instance segmentation via knowledge distillation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2022. 8

[27] Xiaoyu Liu, Wei Huang, Zhiwei Xiong, Shenglong Zhou, Yueyi Zhang, Xuejin Chen, Zheng-Jun Zha, and Feng Wu. Learning cross-representation affinity consistency for sparsely supervised biomedical instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21107–21117, 2023. 3, 8

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[29] Aurélien Lucchi, Kevin Smith, Radhakrishna Achanta, Graham Knott, and Pascal Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE transactions on medical imaging*, 31(2):474–486, 2011. 8

[30] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17927, 2023. 3

[31] Naisong Luo, Rui Sun, Yuwen Pan, Tianzhu Zhang, and Feng Wu. Electron microscopy images as set of fragments for mitochondrial segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 8

[32] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8896–8905, 2018. 3

[33] Jie Ma, Chuan Wang, Yang Liu, Liang Lin, and Guanbin Li. Enhanced soft label for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1185–1195, 2023. 2, 3, 6, 7

[34] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19617–19626, 2023. 8

[35] Huayu Mai, Rui Sun, Yuan Wang, Tianzhu Zhang, and Feng Wu. Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 3

[36] Jaemin Na, Jung-Woo Ha, Hyung Jin Chang, Dongyoon Han, and Wonjun Hwang. Switching temporary teachers for semi-supervised semantic segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 6, 7

[37] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018. 3

[38] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 8

[39] Yuwen Pan, Naisong Luo, Rui Sun, Meng Meng, Tianzhu Zhang, Zhiwei Xiong, and Yongdong Zhang. Adaptive template transformer for mitochondria segmentation in electron microscopy images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21474–21484, 2023. 8

[40] 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. 3, 6, 7

[41] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 3

[42] Rui Sun, Naisong Luo, Yuwen Pan, Huayu Mai, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Appearance prompt vision transformer for connectome reconstruction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1423–1431. International Joint Conferences on Artificial Intelligence Organization, 2023. Main Track. 3

[43] Rui Sun, Huayu Mai, Naisong Luo, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–533. Springer, 2023. 8

[44] Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: Exploring the better weighting function for semi-supervised semantic segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 3, 6, 7

[45] Rui Sun, Yuan Wang, Huayu Mai, Tianzhu Zhang, and Feng Wu. Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1218–1228, 2023. 3

[46] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1, 3, 8

[47] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022. 3

[48] 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. 2, 3, 6

[49] Jianfeng Wang, Daniela Massiceti, Xiaolin Hu, Vladimir Pavlovic, and Thomas Lukasiewicz. Np-semiseg: When neural processes meet semi-supervised semantic segmentation. 2023. 6, 7

[50] 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. 2, 3, 6

[51] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022. 3

[52] 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. 2, 3, 6, 7

[53] Yuan Wang, Naisong Luo, and Tianzhu Zhang. Focus on query: Adversarial mining transformer for few-shot segmentation. In *Advances in Neural Information Processing Systems*, 2023. 3

[54] Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: A buoys view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2023. 3

[55] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 3

[56] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021. 3

[57] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. *arXiv preprint arXiv:2208.09910*, 2022. 1, 3, 6, 7

[58] 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. 6

[59] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3

[60] Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23705–23714, 2023. 3, 6, 7

[61] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. 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*, pages 11350–11359, 2023. 3, 6, 7

[62] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005. 3