# CorrMatch: Label Propagation via Correlation Matching for Semi-Supervised Semantic Segmentation

Boyuan Sun[1] Yuqi Yang[1] Le Zhang[3] Ming-Ming Cheng[2,1] Qibin Hou[2,1*]
[1]VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian [3]SICE, UESTC
{boyuansun, yuqiyang2000}@mail.nankai.edu.cn, {lezhang}@uestc.edu.cn,
{cmm}@nankai.edu.cn, {andrewhoux}@gmail.com

## Abstract

*This paper presents a simple but performant semi-supervised semantic segmentation approach, called CorrMatch. Previous approaches mostly employ complicated training strategies to leverage unlabeled data but overlook the role of correlation maps in modeling the relationships between pairs of locations. We observe that the correlation maps not only enable clustering pixels of the same category easily but also contain good shape information, which previous works have omitted. Motivated by these, we aim to improve the use efficiency of unlabeled data by designing two novel label propagation strategies. First, we propose to conduct pixel propagation by modeling the pairwise similarities of pixels to spread the high-confidence pixels and dig out more. Then, we perform region propagation to enhance the pseudo labels with accurate class-agnostic masks extracted from the correlation maps. CorrMatch achieves great performance on popular segmentation benchmarks. Taking the DeepLabV3+ with ResNet-101 backbone as our segmentation model, we receive a 76%+ mIoU score on the Pascal VOC 2012 dataset with only 92 annotated images. Code is available at https://github.com/BBBBchan/CorrMatch.*

## 1. Introduction

With the development of deep learning techniques, especially convolutional neural networks (CNNs) [12,14,21,55,60,69], many significant semantic segmentation methods [5,15,18,28,65,68,71] have achieved remarkable results. However, methods based on deep learning often require large-scale pixel-wise annotated datasets with a massive amount of labeled images. Compared to the image classification and object detection tasks [8,38], the accurate annotations for segmentation datasets are very expensive and time-consuming.

Recently, many researchers have sought to address the above challenge by reducing the demand for large-scale

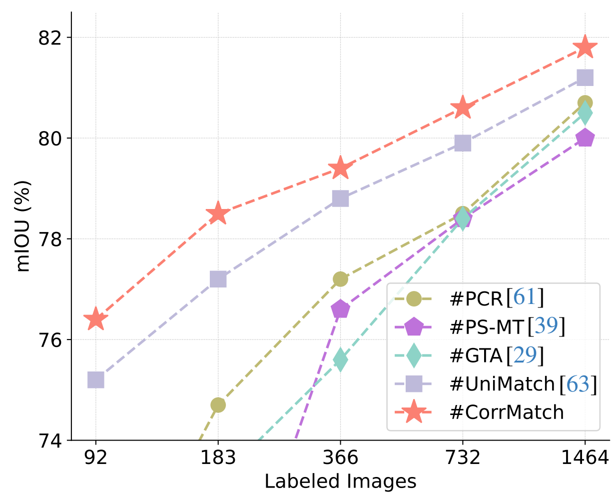



Figure 1. Comparison with state-of-the-art methods on the Pascal VOC dataset. Our CorrMatch outperforms all others for all splits.

accurately annotated data in the semantic segmentation task by presenting weakly-supervised [26,27,53,56], semi-supervised [11,22,23,41], or even unsupervised segmentation methods [13,19,24,50]. Among these schemes, semi-supervised semantic segmentation only requires a small amount of labeled data accompanied by a large amount of unlabeled data for training, which approaches real-world scenarios more and hence attracts the favor of more and more researchers from academia and industry.

In the literature of semi-supervised semantic segmentation, most works adopt the Mean Teacher architecture [23,29,39,61] or self-training strategy [31,64,66] to enable consistency regularization. As shown in Tab. 1, these methods often require extra networks or training stages, complicating the training process. Although the recent UniMatch [63] has shown that a single-stage pipeline is sufficient, it still demands multiple strong augmentation data streams. Unlike them, our CorrMatch is a simpler framework with no need for multiple networks, training stages, or strong augmentation data streams.

Furthermore, in previous works [39,61,64], the most pop-

*Corresponding author.

Table 1. Differences between our CorrMatch and some representative approaches. SDA denotes strong data augmentation.

| Method | Multiple networks | Multi-train stages | Multiple SDA streams | Pairwise similarity |
|---|---|---|---|---|
| PS-MT [39] | ✓ | ✘ | ✘ | ✘ |
| ST++ [64] | ✘ | ✓ | ✘ | ✘ |
| ELN [34] | ✓ | ✓ | ✘ | ✘ |
| UniMatch [63] | ✘ | ✘ | ✓ | ✘ |
| **CorrMatch** | ✘ | ✘ | ✘ | ✓ |

ular way to leverage unlabeled data is setting a fixed threshold to screen reliable pixels as pseudo labels. However, those methods often struggle to efficiently utilize unlabeled data due to the trade-off between pseudo-label proportion and accuracy via threshold adjustments. Beyond that, motivated by the fact that the correlations between pixels can reflect the pairwise similarities, which indicates semantically similar pixels exhibit higher similarity on the correlation map, we reconsider the challenge of accurately assigning pseudo labels to unlabeled data from a label propagation perspective.

First, considering the correlation maps embed the global pairwise similarities, we propose the pixel propagation strategy. With correlation maps constructed from extracted features, the pixel propagation strategy spreads them into predictions, which enriches predictions with global similarities information and fosters semantic consistency. Meanwhile, with the observation that every row of a correlation map is equipped with local shape information, a series of binary maps that capture the objects' shapes can be acquired. Thus, coupled with the most salient predicted class within the intersection of the shapes and high-confidence regions, we propose the region propagation strategy to enhance pseudo labels by accurately assigning class labels to these shapes. By considering the union of shapes and high-confidence regions as the new ones, the high-confidence regions can be expanded, consequently improving the use efficiency of unlabeled data. As shown in Fig. 1, our CorrMatch outperforms all previous approaches.

Our main contributions can be summarized as follows:

- We demonstrate the two advantages of correlation maps in improving the use efficiency of unlabeled data.
- We design a simple but performant semi-supervised semantic segmentation framework containing two novel label propagation strategies.
- Our CorrMatch achieves new state-of-the-art performance on the Pascal VOC 2012 and Cityscapes datasets without any computation burden during inference.

## 2. Related Work

### 2.1. Semi-Supervised Learning

Semi-supervised learning [44, 76] is proposed to settle a paradigm that how to construct models using both labeled and unlabeled data and has been studied long before the deep learning era [2, 3, 30]. And certainly, semi-supervised learning has gained more attention with advancements in deep learning and computer vision [4, 16, 37, 58, 59, 77].

Since Bachman *et al.* [1] proposed a consistency regularization-based method, many approaches, such as Π-Model [36, 43], Mean Teacher [48] and Dual Student [33] have migrated it into the semi-supervised learning field. Recently, FixMatch [46] provides a simple weak-to-strong consistency regularization framework and serves as many other relevant methods' baseline [17, 47, 49, 63]. However, many follow-up works [51, 62, 67] have pointed out that simply setting a manually fixed threshold may lead to inferior performance and slow convergence speed. Among them, FreeMatch [51] provides a dynamic threshold scheme connected with the model's learning process. However, these strategies designed for classification are not suitable for segmentation as multiple categories often exist in each image.

### 2.2. Semi-Supervised Semantic Segmentation

As semi-supervised learning has achieved surprising results in the image classification [36, 37, 46, 48], many works adopt the same setting for semantic segmentation [22, 41, 57].

One type of methods [11, 23, 39, 52, 61, 70, 72, 75] adopt the Mean Teacher architecture. $U^2$PL [52] attempts to use unreliable predictions via contrastive learning better. PS-MT [39] builds a stricter teacher with the VAT [40] technique. ELN [34] uses an error localization network to mitigate the performance degradation caused by confirmation bias due to invalid pseudo labels. All of these methods demand multi-networks for training. Meanwhile, another type of method, self-training based methods [9, 31, 64, 66], often require multiple training stages. Among them, ST++ [64] proposes a three-stage paradigm with strong augmentation. SimpleBase [66] uses separated batch normalization [25] for images with different augmentation. $PC^2$Seg [74] uses feature-space contrastive learning besides consistency training. Recently, UniMatch [63] adopted a single-stage framework based on FixMatch [46] via multiple strong augmentation branches. Unlike all the above, CorrMatch explores how to take advantage of correlation maps better to improve the use efficiency of unlabeled data via label propagation.

## 3. CorrMatch

The goal of semi-supervised semantic segmentation is to train a semantic segmentation network $\mathcal{F}$ with a small labeled image set and a large unlabeled image set. We present a single-stage framework CorrMatch, which leverages pairwise correlations to achieve two label propagation strategies.

### 3.1. Preliminaries

CorrMatch is built upon a simple framework [63] with weak-to-strong consistency regularization. A standard cross-

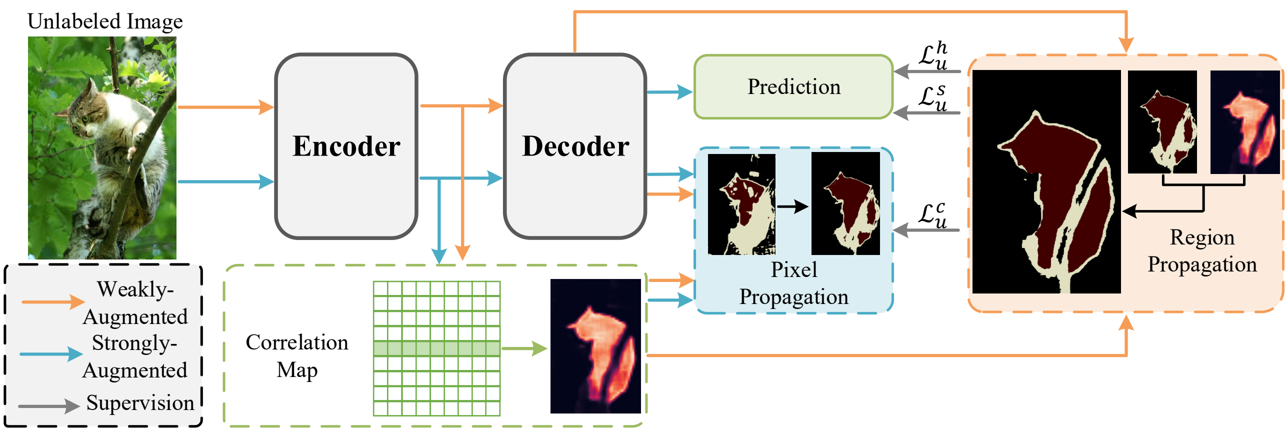



Figure 2. Illustration of our CorrMatch pipeline for unlabeled images. We build it upon the DeepLabv3+ framework [5]. Besides consistency regularization, CorrMatch adopts two label propagation strategies with correlation matching.

entropy loss is applied for labeled images $\{x_i^l\}$ and their corresponding labels $\{y_i^l\}$. And unlabeled images $\{x_i^u\}$ are mainly leveraged by enforcing prediction consistency. For an unlabeled image, $x_i^w$ and $x_i^s$ represent its augmented version with weak and strong augmentation, respectively. The consistency regularization treats the prediction of $x_i^w$ as the pseudo label for $x_i^s$. We demonstrate the pipeline of unlabeled images in Fig. 2.

Given a mini-batch of $N$ unlabeled images, we encourage the outputs to be consistent for both weakly and strongly augmented inputs with hard supervision:

$$\mathcal{L}_u^h = \frac{1}{N}\sum_i^N \ell_c(\mathcal{F}(x_i^s), \mathcal{F}(x_i^w)) \odot \mathcal{M}_i, \quad (1)$$

where $\ell_c$ is the pixel-wise cross-entropy loss function and $\odot$ is the element-wise multiplication. $\mathcal{M}_i$ is a binary map indicating the positions with high confidence predictions in $\mathcal{F}(x_i^w)$, which can be written as:

$$\mathcal{M}_i = \mathbb{1}(\max(\hat{\mathcal{F}}(x_i^w)) > \tau), \quad (2)$$

where $\hat{\mathcal{F}}(x_i^w) \in \mathbb{R}^{K\times HW}$ is the logits output produced by the semantic segmentation network $\mathcal{F}$ and $K$ is the class number. $\tau$ is a threshold used to screen high-confidence predicted pixels as the pseudo label.

However, $\mathcal{L}_u^h$ only treats $\mathcal{F}(x_i^w)$ as the hard pseudo label for $\mathcal{F}(x_i^s)$ and thus ignores additional information stored in logits $\hat{\mathcal{F}}(x_i^w)$. Taking this into account, we further consider the consistency between the logits of the weakly and strongly augmented images in high-confidence regions. In Eqn. (3), we give the formula of $\mathcal{L}_u^s$ for soft supervision.

$$\mathcal{L}_u^s = \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\hat{\mathcal{F}}(x_i^s), \hat{\mathcal{F}}(x_i^w)) \odot \mathcal{M}_i, \quad (3)$$

where $\mathrm{KL}(\cdot)$ is Kullback-Leibler Divergence loss function. We view the above framework as our baseline.

### 3.2. Pixel Propagation

As discussed in Sec. 1, pseudo labels obtained through threshold-based selection overlook the semantic similarity between pixels, constraining the utilization of unlabeled data. In this section, we propose the pixel propagation strategy to enhance the model's overall awareness of pairwise similarities and consequently improve the utilization of unlabeled data, which involves two steps: (1) calculating correlation maps and (2) spreading correlation maps into predictions.

We first extract features $w_1$ and $w_2 \in \mathbb{R}^{D\times HW}$ through linear layers after the encoder of the network, where $D$ is the channel dimension and $HW$ is the number of feature vectors. These extracted features enable correlation matching to quantify the degree of pairwise similarity. Thus, we compute the correlation map $\mathcal{C}$ by performing a matrix multiplication between all pairs of feature vectors:

$$\mathcal{C} = \mathrm{Softmax}(w_1^\top \cdot w_2)/\sqrt{D}, \quad (4)$$

where $^\top$ denotes the matrix transpose operation. The correlation map $\mathcal{C} \in \mathbb{R}^{HW\times HW}$ is a 2D matrix and is activated by a $\mathrm{Softmax}$ function to yield pairwise similarities. $\mathcal{C}$ enables accurate delineation of the corresponding regions belonging to the same object as shown in Fig. 2 and inspires us to propagate it into pseudo labels using correlation matching. More visualizations can be found in Fig. 3.

To enhance the model's awareness of pairwise similarity, we spread the correlation map $\mathcal{C}$ into model logits outputs $\hat{\mathcal{F}}(x_i^u)$ to attain another representation of the prediction $\mathbf{z}_i^u \in \mathbb{R}^{K\times HW}$ via label propagation:

$$\mathbf{z}_i^u = f_1(\hat{\mathcal{F}}(x_i^u)) \cdot \mathcal{C}, \quad (5)$$

where $f_1(\cdot)$ is a bilinear interpolation for shape matching. The resulting $\mathbf{z}_i^u$ emphasizes the pairwise similarities of the same object through the correlation map.

Therefore, a correlation loss $\mathcal{L}_u^c$ can be calculated between $\mathbf{z}_i^u$ and the high-confidence pseudo labels as the supervision, which can be written as follows:

$$\mathcal{L}_u^c = \frac{1}{|N|}\sum_{i=1}^{N}(\ell_c(\mathbf{z}_i^u, \mathcal{F}(x_i^w))) \odot \mathcal{M}_i. \quad (6)$$

For the labeled images $\{x_i^l\}$, we also compute the cross-entropy loss between $\mathbf{z}_i^l$ and $y_i^l$ as the supervised correlation loss $\mathcal{L}_s^c$, where $\mathbf{z}_i^l$ can be attained using Eqn. (5). So far, given a weakly augmented unlabeled image $x_i^w$, its correlation map $\mathcal{C}_i^w$ can effectively model pairwise similarities.

### 3.3. Region Propagation

During experiments, we also observe that every row $\mathbf{c}$ in $\mathcal{C}_i^w$ denotes the similarity between individual feature vectors and all vectors within the entire feature map, which implicitly encapsulates shape information. With this observation, we propose the region propagation strategy to enhance pseudo labels with these shapes information. Specifically, we first normalize $\mathbf{c}$ and turn it into a binary map $\hat{\mathbf{c}}$:

$$\hat{\mathbf{c}} = f_2(\mathbb{1}(\frac{\mathbf{c} - \min(\mathbf{c})}{\max(\mathbf{c}) - \min(\mathbf{c})} > 0.5)), \quad (7)$$

where $f_2(\cdot)$ is a shape-matching function to align the shapes of $\hat{\mathbf{c}}$ and $\mathcal{F}(x_i^w)$. As shown in Fig. 3, the shape information $\hat{\mathbf{c}} \in \mathbb{R}^{H\times W}$ explicitly embeds class agnostic shape information. For every $\hat{\mathbf{c}}$, we can calculate the overlap ratio $r_1$ between $\hat{\mathbf{c}}$ and the high-confidence regions $\mathcal{M}_i$. When $\hat{\mathbf{c}}$ has a large overlap with $\mathcal{M}_i$, (i.e., $r_1 > \tau$), we are able to use $\hat{\mathbf{c}}$ to adjust the pseudo label $\mathcal{F}(x_i^w)$.

Given the current pseudo labels $\mathcal{F}(x_i^w)$, we can calculate the quantity of each unique class $l \in L$ within high-confidence shape $(\mathcal{F}(x_i^w) \odot \mathcal{M}_i \odot \hat{\mathbf{c}})$ by a function $G(l)$ and locate the most significant class $k^*$ with the following equation:

$$k^* = \text{argmax}_{l\in L} G(l), \quad (8)$$

$$G(l) = \sum_{HW} \mathbb{1}[(\mathcal{F}(x_i^w) \odot \mathcal{M}_i \odot \hat{\mathbf{c}}) = l], \quad (9)$$

where $L$ is the set of all unique classes that present in predictions $\mathcal{F}(x_i^w)$. With the most significant class $k^*$, we can calculate its proportion $r_2$ within the high-confidence shape.

When $k^*$ highly coincides with the high-confidence shape, (i.e., $r_2 > \tau$), we can propagate the specific class $k^*$ into the enhanced pseudo label $\mathcal{F}(x_i^w)$ and expanded high-confidence regions $\mathcal{M}_i$ by matching the certain shape $\hat{\mathbf{c}}$.

$$\mathcal{F}(x_i^w) = \begin{cases} k^*, & \hat{\mathbf{c}} = 1 \\ \mathcal{F}(x_i^w), & \hat{\mathbf{c}} = 0 \end{cases}, \mathcal{M}_i = \mathcal{M}_i \cup \hat{\mathbf{c}} \quad (10)$$

However, considering the intricate computations required for each specific shape within the correlation map and the frequent occurrence of similar semantic information in adjacent regions, resulting in similar shapes in the correlation map, it becomes evident that involving every row of the correlation map in pseudo labels optimization is redundant. Hence, we employed a random sampling approach within the correlation map to expedite label propagation. As shown in Fig. 3, region propagation significantly expands high-confidence regions with shape information and the most salient class.


Figure 3. Illustration of our proposed propagation strategies. White areas are ignored regions due to low confidence. Combining the shape information with the most salient class, CorrMatch can significantly enhance pseudo labels and expand high-confidence regions.

It is also worth mentioning that the correlation map construction process and label propagation only participate in the training process and hence do not bring any additional computational burdens during the inference process.

### 3.4. More Details

**Dynamic threshold.** As mentioned in FreeMatch [51], using a fixed threshold $\tau$ that is too strict or too loose is detrimental to model convergence. At the same time, we observe that the most suitable thresholds are different for different experimental settings (Fig. 5d). Thus, We provide a dynamic threshold strategy that is related to the training process.

Given the threshold $\tau$ a relatively small value (0.85) as initialization, the strategy of updating $\tau$ depends on the logits $\hat{\mathcal{F}}(x_i^w)$. We use the exponential moving average (EMA) [42] to iteratively update $\tau$. Each increment is defined as:

$$\Delta\tau = \frac{1}{|L|}\sum_{l\in L}\max[\mathbb{1}(\mathcal{F}(x_i^w) = l) \odot \overset{c}{\max}(\hat{\mathcal{F}}(x_i^w))], \quad (11)$$

where $\overset{c}{\max}(\cdot)$ denotes taking the maximum value along the channel dimension. This operation aims to take the maximum confidence of all predicted classes in $\hat{\mathcal{F}}(x_i^w)$ and use their average as the increment for each iteration. We found that such a simple threshold updating strategy works well. We will further show in Sec. 4.3 that $\tau$ is insensitive to initialization. The corresponding pseudo code is provided in the supplementary materials.

Table 2. Comparisons of CorrMatch with the state-of-the-art approaches on the Pascal VOC 2012 val set in terms of mIoU (%). All methods are trained on the classic setting, *i.e*., the labeled images are selected from the original VOC train set, which consists of 1,464 images.

| Method | Training Size | 1/16 (92) | 1/8 (183) | 1/4 (366) | 1/2 (732) | Full (1464) |
|---|---|---|---|---|---|---|
| ST++ [64] | 321 × 321 | 65.2 | 71.0 | 74.6 | 77.3 | 79.1 |
| UniMatch [63] | 321 × 321 | 75.2 | 77.2 | 78.8 | 79.9 | 81.2 |
| Mean Teacher [48] | 513 × 513 | 51.7 | 58.9 | 63.9 | 69.5 | 71.0 |
| CutMix-Seg [11] | 513 × 513 | 52.2 | 63.5 | 69.5 | 73.7 | 76.5 |
| PseudoSeg [78] | 513 × 513 | 57.6 | 65.5 | 69.1 | 72.4 | 73.2 |
| CPS [6] | 513 × 513 | 64.1 | 67.4 | 71.7 | 75.9 | - |
| PC $^2$ Seg [74] | 513 × 513 | 57.0 | 66.3 | 69.8 | 73.1 | 74.2 |
| U$^2$ PL [52] | 513 × 513 | 68.0 | 69.2 | 73.7 | 76.2 | 79.5 |
| PS-MT [39] | 513 × 513 | 65.8 | 69.6 | 76.6 | 78.4 | 80.0 |
| GTA [29] | 513 × 513 | 70.0 | 73.2 | 75.6 | 78.4 | 80.5 |
| PCR [61] | 513 × 513 | 70.1 | 74.7 | 77.2 | 78.5 | 80.7 |
| RC$^2$L [70] | 513 × 513 | 65.3 | 68.9 | 72.2 | 77.1 | 79.3 |
| CCVC [54] | 513 × 513 | 70.2 | 74.4 | 77.4 | 79.1 | 80.5 |
| **CorrMatch** | 321 × 321 | **76.4** | **78.5** | **79.4** | **80.6** | **81.8** |

**Loss function.** The overall objective function $\mathcal{L}$ is a combination of supervised loss $\mathcal{L}_s$ and unsupervised loss $\mathcal{L}_u$: $\mathcal{L} = \frac{1}{2}(\mathcal{L}_s + \mathcal{L}_u)$. Like most methods, we use the cross-entropy loss function $\mathcal{L}_s^h$ as the basic supervision of labeled data $\mathcal{D}^l$. Therefore, the supervised loss $\mathcal{L}_s$ is defined as the combination of $\mathcal{L}_s^h$ and supervised correlation loss $\mathcal{L}_s^c$: $\mathcal{L}_s = \frac{1}{2}(\mathcal{L}_s^h + \mathcal{L}_s^c)$. As for unsupervised loss $\mathcal{L}_u$ on unlabeled data $\mathcal{D}^u$, it can be expressed as follows:

$$\mathcal{L}_u = \lambda_1 \mathcal{L}_u^h + \lambda_2 \mathcal{L}_u^s + \lambda_3 \mathcal{L}_u^c, \tag{12}$$

where $\mathcal{L}_u^h, \mathcal{L}_u^s$ and $\mathcal{L}_u^c$ denote the unsupervised hard loss, soft loss, and correlation loss. And $[\lambda_1, \lambda_2, \lambda_3]$ are set to $[0.5, 0.25, 0.25]$ by default.

# 4. Experiments

## 4.1. Experiment Setup

**Datasets.** We report results on the Pascal VOC 2012 and Cityscapes datasets. Pascal VOC 2012 is a semantic segmentation benchmark with 21 classes, consisting of 1,464 high-quality annotated images for training and 1,449 images for evaluation originally [10]. We also conduct experiments on the aug Pascal VOC 2012 dataset, which contains more coarsely annotated images from the Segmentation Boundary Dataset (SBD) [20], resulting in 10,582 training images in total. Cityscapes is an urban scene understanding dataset, including 2,975 training and 500 validation images with fine annotations [7]. It contains 19 classes of urban scenes, and all images have the resolution of 1024×2048.

**Implementation details.** Following most previous semi-supervised semantic segmentation methods, we use DeepLabV3+ [5] with ResNet-101 [21] pre-trained on ImageNet [8] as the backbone. For the training on the Pascal VOC 2012 dataset, we use stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001, weight decay of 1e$-$4, crop size of 321×321 or 513×513, batch size of 16, and training epochs of 80. For the Cityscapes dataset, following UniMatch [63], we use stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.005, weight decay of 1e$-$4, crop size of 801 × 801, batch size of 16, and training epochs of 240 with 4 × A40 GPUs.

As for evaluation metrics, we report the mean Intersection-over-Union (mIoU) with original images following previous papers [6, 11, 39] for the Pascal VOC 2012 dataset. For Cityscapes, same as previous methods [6,52,63], we apply slide window evaluation with a fixed crop in a sliding window manner and then calculate mIoU on these cropped images. All the results are measured on the standard validation set based on single-scale inference.

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

**Results on classic Pascal VOC 2012**. We show the performance of our method with other state-of-the-art methods on the classic Pascal VOC 2012 Dataset in Tab. 2. Our experiments are conducted on various splits of the original train set following the data partition in CPS [6]. On the full split, our method gets 81.8% mIoU. Also, CorrMatch achieves consistent performance gains compared to existing state-of-art approaches. Particularly, CorrMatch outperforms UniMatch by 1.2%, 1.3%, 0.6%, 0.7% and 0.6% on each split.

**Results on aug Pascal VOC 2012**. In Tab. 3, we show our performance and compare it with existing methods on the aug Pascal VOC 2012 Dataset. It is clear that our results are consistently much better than the existing best ones. We conduct experiments on 1/16, 1/8, and 1/4 splits, respectively. Under the 321×321 training size (top-left of Tab. 3), compared to the supervised baseline, CorrMatch gets +12.0%, +7.4%, and +5.5% improvements. In addition, our approach

Table 3. Comparisons of state-of-the-art methods on the Pascal VOC 2012 val set with mIoU (%) metric. All methods are trained on the aug setting, i.e., the labeled images are selected from the aug VOC train set, which consists of 10, 582 images. † means using U$^2$PL [52]'s splits.

| Method | Train size | 1/16 (662) | 1/8 (1323) | 1/4 (2646) |
|---|---|---|---|---|
| Supervised | 321 × 321 | 65.6 | 70.4 | 72.8 |
| ST++ [64] | 321 × 321 | 74.5 | 76.3 | 76.6 |
| CAC [35] | 321 × 321 | 72.4 | 74.6 | 76.3 |
| UniMatch [63] | 321 × 321 | 76.5 | 77.0 | 77.2 |
| **CorrMatch** | 321 × 321 | **77.6** | **77.8** | **78.3** |
| U2PL† [52] | 513 × 513 | 77.2 | 79.0 | 79.3 |
| GTA† [29] | 513 × 513 | 77.8 | 80.4 | 80.5 |
| PCR† [61] | 513 × 513 | 78.6 | 80.7 | 80.7 |
| CCVC† [61] | 513 × 513 | 76.8 | 79.4 | 79.6 |
| AugSeg† [73] | 513 × 513 | 79.3 | 81.5 | 80.5 |
| **CorrMatch**† | 513 × 513 | **81.3** | **81.9** | **80.9** |

| Method | Train size | 1/16 (662) | 1/8 (1323) | 1/4 (2646) |
|---|---|---|---|---|
| CutMix-Seg [11] | 513 × 513 | 71.7 | 75.5 | 77.3 |
| CCT [41] | 513 × 513 | 71.9 | 73.7 | 76.5 |
| GCT [32] | 513 × 513 | 70.9 | 73.3 | 76.7 |
| CPS [6] | 513 × 513 | 74.5 | 76.4 | 77.7 |
| AEL [23] | 513 × 513 | 77.2 | 77.6 | 78.1 |
| FST [9] | 513 × 513 | 73.9 | 76.1 | 78.1 |
| ELN [34] | 513 × 513 | - | 75.1 | 76.6 |
| U$^2$PL [52] | 513 × 513 | 74.4 | 77.6 | 78.7 |
| PS-MT [39] | 513 × 513 | 75.5 | 78.2 | 78.7 |
| AugSeg [73] | 513 × 513 | 77.0 | 77.3 | 78.8 |
| **CorrMatch** | 513 × 513 | **78.4** | **79.3** | **79.6** |

outperforms UniMatch by 1.1%, 0.8%, and 1.1% on each split. As for the 513×513 training size (right of Tab. 3), CorrMatch also consistently outperforms current state-of-the-art methods. For instance, we get 79.3% mIoU on the 1/8 split with a gain of around 2% compared to AugSeg [73].

We also report the results using the same splits as in U$^2$PL [52] with 513×513 training size (bottom-left of Tab. 3), which contain more well-annotated labels and have higher expectations of results. Compared to the best method AugSeg [73], our method gains 2.0% improvement on the 1/16 split. Furthermore, same to other methods, we observe that, as the split size increases from 1/8 to 1/4, the performance decreases under this setting. This is because in the 1/8 split, almost all of the accurately labeled images are included, and most of the images added to the larger split are coarsely labeled, which results in no improvement in performance.

**Results on Cityscapes**. In Tab. 4, we compare the performance of CorrMatch with state-of-the-art methods on the Cityscapes dataset. We follow sliding window evaluation and online hard example mining (OHEM) loss [45] techniques, which have been widely applied in previous SOTA works [6, 23, 39, 52, 61, 63]. It can be clearly seen that our method can consistently outperform other methods under all splits. Compared to UniMatch [63], our CorrMatch achieves +0.7%, +0.6%, +0.2%, and +0.9% on 1/16, 1/8, 1/4, 1/2 splits, respectively.

## 4.3. Ablations Studies

In this part, we conduct a series of ablations studies to verify the designs of proposed strategies in CorrMatch. We report the results of the DeepLabV3+ network using ResNet-101 as the encoder on the original Pascal VOC 2012 dataset with training size 321 × 321.

**Effectiveness of components.** We first conduct ablation studies on different components of our CorrMatch to demonstrate their effectiveness in Tab. 5. With the hard unsupervised loss and dynamic threshold, we get 73.6% on the 92 split and 80.0% on the 1464 split. Adding soft loss $\mathcal{L}_u^s$ as the basic framework brings 0.8% and 0.5% improvements. With the help of label propagation, we achieve another 2.0% and 1.3% improvements. These results demonstrate the effectiveness of each of our components individually. Also, replacing $\mathcal{L}_u^h$ with $\mathcal{L}_u^s$ results in a performance decrease, which illustrates the importance of $\mathcal{L}_u^h$. Finally, the complete CorrMatch achieves 76.4% and 81.8% mIoU, which is +2.8% and +1.8% compared to the baselines.

Table 4. Comparing results of state-of-the-art algorithms on the Cityscapes val set. All the experiments are conducted with ResNet-101 as the backbone.

| Method | 1/16 (186) | 1/8 (372) | 1/4 (744) | 1/2 (1488) |
|---|---|---|---|---|
| Supervised | 65.7 | 72.5 | 74.4 | 77.8 |
| CCT [41] | 69.3 | 74.1 | 76.0 | 78.1 |
| CPS [6] | 69.8 | 74.3 | 74.6 | 76.8 |
| AEL [23] | 74.5 | 75.5 | 77.5 | 79.0 |
| U$^2$PL [52] | 70.3 | 74.4 | 76.5 | 79.1 |
| PS-MT [39] | - | 76.9 | 77.6 | 79.1 |
| UniMatch [63] | 76.6 | 77.9 | 79.2 | 79.5 |
| PCR [61] | 73.4 | 76.3 | 78.4 | 79.1 |
| **CorrMatch** | **77.3** | **78.5** | **79.4** | **80.4** |

We also conduct experiments with the fixed threshold (0.95). It can be observed that compared to the fixed baselines (73.1% and 79.9%), changing it into a dynamic manner only brings +0.5% and +0.1%. Meanwhile, after adding all components, the corresponding improvements can be lifted to +0.9% and +1.0%. This proves our threshold strategy cooperates well with our label propagation strategy.

**Impact of label propagation strategies.** In Tab. 6, we conduct the ablation study of our label propagation strategies. Our pixel propagation strategy, which constructs the cor-

Table 5. Ablation study on the effectiveness of different components, including threshold $\tau$ (Dyna. denotes our dynamic strategy), hard loss $\mathcal{L}_u^h$, soft loss $\mathcal{L}_u^s$, label propagation $\mathcal{P}$.

| $\tau$ | $\mathcal{L}_u^h$ | $\mathcal{L}_u^s$ | $\mathcal{P}$ | 92 | 1464 |
|---|---|---|---|---|---|
| Dyna. | ✓ | | | 73.6 | 80.0 |
| Dyna. | | ✓ | | 73.1 | 79.6 |
| Dyna. | ✓ | ✓ | | 74.4 | 80.5 |
| Dyna. | ✓ | | ✓ | 74.6 | 80.6 |
| Dyna. | ✓ | ✓ | ✓ | **76.4** | **81.8** |
| Fixed | ✓ | | | 73.1 | 79.9 |
| Fixed | ✓ | ✓ | | 73.3 | 79.9 |
| Fixed | ✓ | | ✓ | 74.3 | 80.1 |
| Fixed | ✓ | ✓ | ✓ | 75.5 | 80.8 |

Table 6. Ablation study on the label propagation strategies.

| Method | 92 | 366 | 1464 |
|---|---|---|---|
| w/o Propagation | 74.4 | 78.5 | 80.5 |
| w/ Pixel Propagation | 75.8 | 78.9 | 81.3 |
| w/ Pixel & Region Propagation | **76.4** | **79.4** | **81.8** |

relation maps and spreads them into predictions as a new representation with the supervision of correlation loss $\mathcal{L}^c$, brings 1.4%, 0.4%, and 0.8% improvements. Furthermore, equipped with our region propagation strategy, more detailed local shape information is mined and thus enhanced pseudo labels are obtained. This strategy further improves 0.6%, 0.5%, and 0.5% on 92, 366, and 1464 splits, respectively.

**Where to extract features.** In the default setting, we choose to extract features from the backbone, which makes the proposed strategies more convenient to be transplanted to other segmentation networks. Actually, given a specific network structure, the position of feature extraction can be flexible. Here, we consider the impact of different feature extraction positions on performance. In Tab. 7, we demonstrate the performance of extracting features after different positions for the Deeplabv3+ decoder under different splits. The results show that using the backbone features consistently outperforms other alternatives.

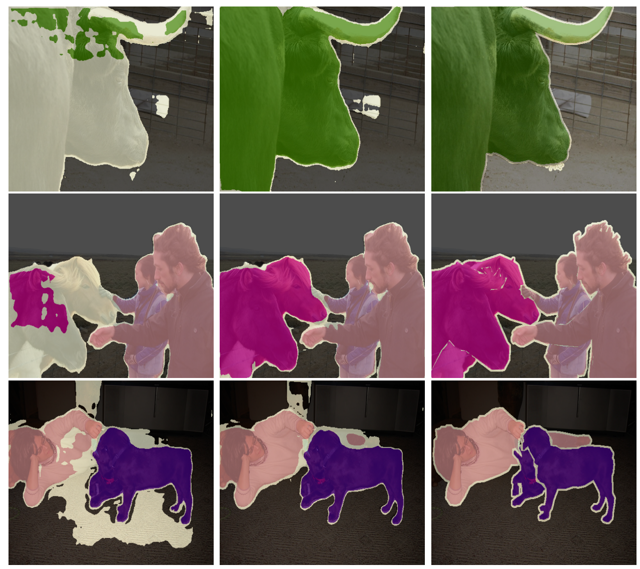
**Different sampling strategies.** Since using all shapes within the correlation map to enhance pseudo labels would incur a substantial computational burden, it is imperative to sample a subset of shapes from it. Here we conduct experiments about sampling methods and quantities in Tab. 8. We conduct experiments on random sampling $\mathcal{R}$ and uniform sampling $\mathcal{U}$ methods, with 16, 32, 64, 128, and 256 sampling numbers on the 1464 split. The results show random sampling continuously outperforms uniform sampling. Among these, random sampling with 128 sample numbers yields the best performance, with marginal differences compared to the 256-sample strategy. Thus, we choose to randomly sample 128 shapes from the correlation map as a trade-off between computational efficiency and performance.

Table 7. Ablation study on feature extraction positions. We take features after each specific module of DeepLabV3+ to build correlation maps and adopt label propagation strategies.

| Position | Backbone | ASPP | Fusion | Classifier |
|---|---|---|---|---|
| 732 | **80.4** | 79.5 | 79.1 | 79.5 |
| 1464 | **81.8** | 80.6 | 80.1 | 80.8 |

Table 8. Ablation study on the different sampling methods. $\mathcal{R}$ denotes random sampling; $\mathcal{U}$ denotes uniform sampling.

| Numbers | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|
| $\mathcal{R}$ | 81.1 | 81.2 | 81.4 | **81.8** | 81.7 |
| $\mathcal{U}$ | 81.0 | 81.1 | 81.2 | 81.4 | 81.0 |



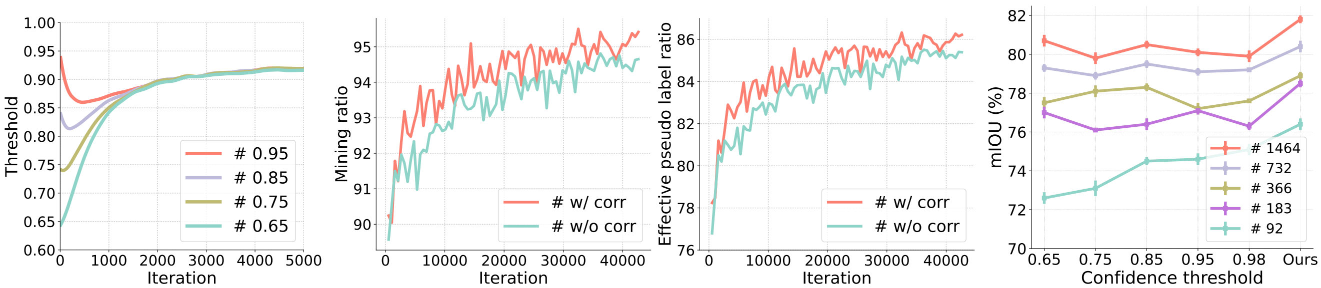
(a) w/o propagation (b) w/ propagation (c) GT

Figure 4. Qualitative results on the Pascal VOC 2012 dataset. (a) Pseudo labels without label propagation; (b) Pseudo labels with CorrMatch; (c) Ground truth. White areas in (a) and (b) are ignored regions due to low confidence.

**Different initial values for CorrMatch.** Since our EMA-based threshold updating strategy needs an initial value for $\tau$, we discuss the impact of different initialization values for $\tau$ in Fig. 5a. The conclusion is that our threshold strategy is insensitive to different initialization values. Even with different threshold initialization values, all the thresholds tend to approach a similar value very quickly (around 1500 iterations) in the early stage of training (around 40000 iterations in total) under all experiment settings.

## 4.4. Correlation Helps Mining Reliable Regions

**Statistics.** Ideally, all correctly predicted points should be regarded as pseudo labels for the unlabeled data. To demonstrate the ability of correlation matching to help label propa-




(a) Different threshold initialization (b) Mining ratio (c) Effective pseudo label ratio (d) Different fixed thresholds

Figure 5. Some statistics on label propagation and the threshold strategy. For (a), (b), and (c), experiments are conducted on the 1464 split.

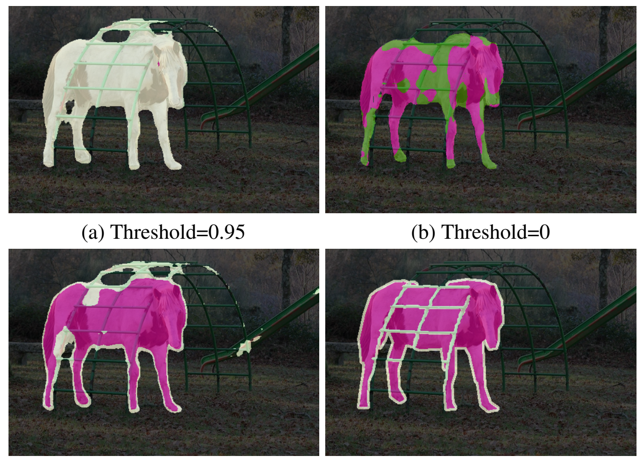
gation, we count the mining ratio and effective pseudo label ratio in Fig. 5b and Fig. 5c. The mining ratio is the proportion of selected high-confidence pixels among all correctly predicted pixels. The effective pseudo label ratio is the proportion of accurately predicted pseudo labels to the whole image, which can reflect effective pseudo label numbers. It can be clearly seen that with the proposed label propagation strategies, the mining ratio and effective pseudo label ratio are significantly higher than those without them, which illustrates that the utilization of unlabeled data has improved effectively. This further indicates our strategies can improve the overall quality of pseudo labels by leveraging similarity and shape information from correlation maps.

**Qualitative analysis.** In Fig. 4, we give some visualization results to further demonstrate the effectiveness of our label propagation strategies. Comparing Fig. 4b and Fig. 4a, it is obvious that with the support of label propagation, the number of pixels and completeness of the high-confidence regions are significantly better than those without it. This means that our method can effectively expand high-confidence regions and populate these regions with the correct categories. We will provide more detailed qualitative results in the supplementary materials.

## 5. Discussions on Label Propagation Strategy

Traditionally, semi-supervised semantic segmentation methods mostly rely on adjusting thresholds to expand high-confidence regions [52, 63]. However, selecting the most suitable threshold could be a challenging task. For instance, our observations illustrated in Fig. 5d, indicate that the optimal threshold can vary significantly. Fig. 6a and Fig. 6b further demonstrate that a too-strict threshold restricts the unlabeled data utilization, while a lenient threshold results in fragmented incorrect pixel predictions.

Different from the scheme of directly adjusting the threshold, label propagation does not merely expand the high-confidence regions; it assigns accurate predictions to pseudo labels by utilizing accurate shapes within the correlation map, which helps maintain more consistent semantic structures within high-confidence regions and thus mitigates the discontinuity issue. In Fig. 6c and the last column of Fig. 5d, we show the pseudo label and performance of CorrMatch. This indicates that our CorrMatch consistently obtains more accurate and complete pseudo labels and achieves the highest results on all splits.



(a) Threshold=0.95 (b) Threshold=0

(c) Label propagation (d) GT

Figure 6. Comparisons of pseudo labels with different strategies.

## 6. Conclusions

We present CorrMatch that can utilize label propagation with correlation matching to discover more accurate high-confidence regions for semi-supervised semantic segmentation. The key contributions of our CorrMatch are reconsidering the use of correlation maps and designing two label propagation strategies to enrich the pseudo label. Equipped with these strategies, CorrMatch significantly expands the high-confidence regions and thus can utilize unlabeled data more efficiently. Experiments show the superiority of our CorrMatch over other methods.

**Acknowledgments.** This research was supported by NSFC (NO. 62225604, No. 62276145), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63223049), CAST through Young Elite Scientist Sponsorship Program (No. YESS20210377). Computations were supported by the Supercomputing Center of Nankai University (NKSC).

## References


[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *NeurIPS*, 27, 2014. 2

[2] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006. 2

[3] Kristin Bennett and Ayhan Demiriz. Semi-supervised support vector machines. *NeurIPS*, 11, 1998. 2

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 32, 2019. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 1, 3, 5

[6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021. 5, 6

[7] 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 *CVPR*, pages 3213–3223, 2016. 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1, 5

[9] Ye Du, Yujun Shen, Haochen Wang, Jingjing Fei, Wei Li, Liwei Wu, Rui Zhao, Zehua Fu, and Qingjie Liu. Learning from future: A novel self-training framework for semantic segmentation. *arXiv preprint arXiv:2209.06993*, 2022. 2, 6

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–308, 2009. 5

[11] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *Brit. Mach. Vis. Conf.*, 2020. 1, 2, 5, 6

[12] Shanghua Gao, Zhong-Yu Li, Qi Han, Ming-Ming Cheng, and Liang Wang. Rf-next: Efficient receptive field search for convolutional neural networks. *IEEE TPAMI*, pages 1–19, 2022. 1

[13] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE TPAMI*, pages 1–20, 2022. 1

[14] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(2):652–662, 2021. 1

[15] Lixue Gong, Yiqun Zhang, Yunke Zhang, Yin Yang, and Weiwei Xu. Erroneous pixel prediction for semantic image segmentation. *Computational Visual Media*, 8:165–175, 2022. 1

[16] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. 2

[17] Sascha Grollmisch and Estefanía Cano. Improving semi-supervised learning for audio classification with fixmatch. *Electronics*, 10(15):1807, 2021. 2

[18] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022. 1

[19] Robert Harb and Patrick Knöbelreiter. Infoseg: Unsupervised semantic image segmentation with mutual information maximization. In *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings*, pages 18–32. Springer, 2022. 1

[20] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, pages 991–998. IEEE, 2011. 5

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 5

[22] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. *NeurIPS*, 28, 2015. 1, 2

[23] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *NeurIPS*, 34:22106–22118, 2021. 1, 2, 6

[24] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, pages 7334–7344, 2019. 1

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. pmlr, 2015. 2

[26] Peng-Tao Jiang, Ling-Hao Han, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Online attention accumulation for weakly supervised semantic segmentation. *IEEE TPAMI*, 44(10):7062–7077, 2022. 1

[27] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *CVPR*, 2022. 1

[28] Rui Jiang, Ruixiang Zhu, Hu Su, Yinlin Li, Yuan Xie, and Wei Zou. Deep learning-based moving object segmentation: Recent progress and research prospects. *Machine Intelligence Research*, 20(3):335–369, 2023. 1

[29] Ying Jin, Jiaqi Wang, and Dahua Lin. Semi-supervised semantic segmentation via gentle teaching assistant. In *NeurIPS*, 2022. 1, 5, 6

[30] Thorsten Joachims et al. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999. 2

[31] Rihuan Ke, Angelica I Aviles-Rivero, Saurabh Pandey, Saikumar Reddy, and Carola-Bibiane Schönlieb. A three-stage self-training framework for semi-supervised semantic segmentation. *IEEE Trans. Image Process.*, 31:1805–1815, 2022. 1, 2

[32] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 6

[33] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *ICCV*, pages 6728–6736, 2019. 2

[34] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *CVPR*, pages 9957–9967, 2022. 2, 6

[35] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, pages 1205–1214, 2021. 6

[36] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2

[37] 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*, volume 3, page 896, 2013. 2

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[39] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *CVPR*, pages 4258–4267, 2022. 1, 2, 5, 6

[40] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, 41(8):1979–1993, 2018. 2

[41] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020. 1, 2, 6

[42] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992. 4

[43] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *NeurIPS*, 29, 2016. 2

[44] Matthias Seeger. Learning with labeled and unlabeled data, 2000. 2

[45] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016. 6

[46] 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. *NeurIPS*, 33:596–608, 2020. 2

[47] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 2

[48] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 2, 5

[49] Pratima Upretee and Bishesh Khanal. Fixmatchseg: Fixing fixmatch for semi-supervised semantic segmentation. *arXiv preprint arXiv:2208.00400*, 2022. 2

[50] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, pages 10052–10062, 2021. 1

[51] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022. 2, 4

[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 *CVPR*, pages 4248–4257, 2022. 2, 5, 6, 8

[53] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, pages 12275–12284, 2020. 1

[54] Zicheng Wang, Zhen Zhao, Luping Zhou, Dong Xu, Xiaoxia Xing, and Xiangyu Kong. Conflict-based cross-view consistency for semi-supervised semantic segmentation. *arXiv preprint arXiv:2303.01276*, 2023. 5

[55] Xiu-Shen Wei, Yu-Yan Xu, Chen-Lin Zhang, Gui-Song Xia, and Yu-Xin Peng. Cat: a coarse-to-fine attention tree for semantic change detection. *Visual Intelligence*, 1(1):3, 2023. 1

[56] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, pages 7268–7277, 2018. 1

[57] Hui Xiao, Dong Li, Hao Xu, Shuibo Fu, Diqun Yan, Kangkang Song, and Chengbin Peng. Semi-supervised semantic segmentation with cross teacher training. *Neurocomputing*, 508:36–46, 2022. 2

[58] Jin Xie, San-Yang Liu, and Jia-Xi Chen. A framework for distributed semi-supervised learning using single-layer feedforward networks. *Machine Intelligence Research*, 19(1):63–74, 2022. 2

[59] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 2

[60] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 1

[61] Hai-Ming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. *arXiv preprint arXiv:2210.04388*, 2022. 1, 2, 5, 6

[62] 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 *ICML*, pages 11525–11536. PMLR, 2021. 2

[63] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, 2023. 1, 2, 5, 6, 8

[64] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *CVPR*, pages 4268–4277, 2022. 1, 2, 5, 6

[65] Bowen Yin, Xuying Zhang, Zhongyu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. *arXiv preprint arXiv:2309.09668*, 2023. 1

[66] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *ICCV*, pages 8229–8238, 2021. 1, 2

[67] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NeurIPS*, 34:18408–18419, 2021. 2

[68] Dong Zhang, Liyan Zhang, and Jinhui Tang. Augmented fcn: rethinking context modeling for semantic segmentation. *Science China Information Sciences*, 66(4):142105, 2023. 1

[69] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *CVPR*, pages 2736–2746, 2022. 1

[70] Jianrong Zhang, Tianyi Wu, Chuanghao Ding, Hongwei Zhao, and Guodong Guo. Region-level contrastive and consistency learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:2204.13314*, 2022. 2, 5

[71] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 1

[72] Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *CVPR*, 2023. 2

[73] 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 *CVPR*, 2023. 6

[74] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *ICCV*, pages 7273–7282, 2021. 2, 5

[75] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *ICCV*, pages 7036–7045, 2021. 2

[76] Xiaojin Jerry Zhu. Semi-supervised learning literature survey, 2005. 2

[77] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *NeurIPS*, 33:3833–3845, 2020. 2

[78] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 5