Consistency and Separation Regularization for Contrastive Learning in Semi-supervised Semantic Segmentation

***Abstract*—The practice of annotation for semantic segmentation is conducted in pixel-level, prompting researchers to explore semi-supervised approaches. Pseudo-labeling is one major approach to explore semantic information from unlabeled images. However, existing pseudo-labeling approaches produce pseudo-labels with either substantial errors or considerable noise. Contrastive learning methods have been proposed to learn and obtain rich semantic information in the training process. Nonetheless, during our experiments, we found that these methods have overlooked issues related to representation inconsistency and an excessive similarity in inter-class features, both of which significantly influence the performance of their models. Based on these observations, we present a contrastive learning module designed to maintain consistency among intra-class image features, while ensuring sufficient separation between inter-class image features. This method is straightforward and does not require any elaborate techniques, making it easy to deploy on any existing segmentation network and semi-supervised framework without introducing additional computational costs or memory burdens. We also have proposed a novel loss function, designed to reduce noisy pseudo-labels and to collaborate with our contrastive learning module. To evaluate the performance of our proposed method, we conduct a comprehensive set of experiments and ablation studies on the Pascal VOC 2012 and Cityscapes datasets using our consistency and separation regularization for contrastive learning (CSC) approach. The results demonstrate that our method achieves state-of-the-art (SOTA) performance in various semi-supervised settings.**

***Index Terms*—Semi-Supervised Semantic Segmentation, Con-**

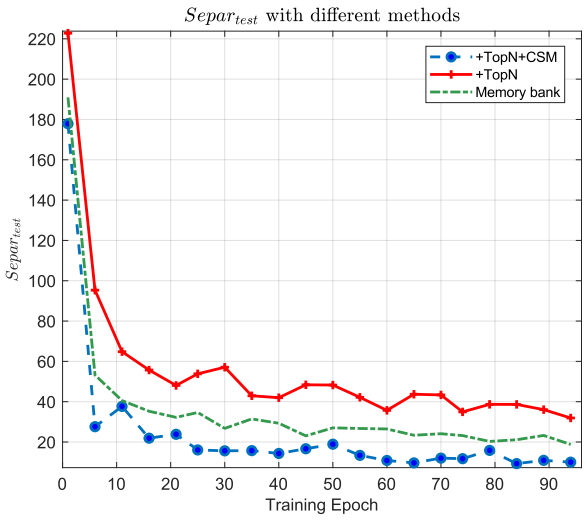

Fig. 1: The measure of separation between inter-class image features learned by the semantic segmentation model using different method. Red line denotes contrastive learning with our restriction $TopN$ on positive samples in loss function. Green line denotes memory bank application. Blue line denotes contrastive learning using our Consistency and Separation Module (CSM) on memory bank.

I. Introduction

SEMANTIC segmentation is a fundamental task in computer vision, which has been widely applied in various visual tasks [1]–[5]. Despite the significant progress made in this field, the current high-performance segmentation methods mostly rely on full supervision. Full supervision requires large amounts of annotated data at the pixel level, which is costly to obtain. To address the problem of insufficient data annotation, semi-supervised semantic segmentation has been proposed.

Recent research on semi-supervised semantic segmentation has shown that pseudo-labeling [6]–[8] is one of the most promising approaches to explore semantic information from unlabeled images. Pseudo-labeling utilizes predictions made by a model trained on labeled data as pseudo-labels to further train the model on unlabeled data. Nonetheless, pseudo-labeling methods have not yet reached parity with fully supervised semantic segmentation models. This disparity [9], [10] can be attributed to the reliance of pseudo-labeling methods on the cross-entropy loss function, which lacks the ability to distinguish between inter-class and intra-class features.

To mitigate this challenge, researchers have introduced contrastive learning techniques, as detailed in [11]–[14], which aim to distinguish features corresponding to distinct classes. Contrastive learning imposes consistency among pixel-level features (or region-level features) within the same class while concurrently promoting separation between pixel features associated with different classes, as discussed in [15]–[19]. Consequently, to create embedding spaces conducive to class differentiation, certain techniques have proposed fully supervised pixel-wise contrastive learning techniques, as exemplified in [4]. However, when applied to the semi-supervised domain, some methods simply utilize pseudo-labels as the guidance for correct feature representations and thus use these feature vectors to compute pixel-wise contrastive loss [15], [17], [19]. Although these methods aim to establish similarity between positive pixel pairs and dissimilarity between negative pixel pairs, the outcome is not as expected. This discrepancy can be attributed to the lack of separation among inter-class features and consistency among intra-class features. Firstly, prior research [20]–[23] pointed out the significance of effective feature sampling strategies as the diversity of training samples is crucial for contrastive learning. Instead

of selecting all the pixels in an image to learn, a sampling strategy only sorts out valuable data samples that complement contrastive learning. Unfortunately, existing strategies [22], [23] fail to select features with adequate class separation, as visually depicted in Fig. I. Furthermore, as training progresses, there is a notable decline in the similarity between features originating from the same image, as shown in Fig. II-A. Training with such inconsistent features leads to continuously evolving feature mappings for the same class, impeding the model's ability to achieve stable training outcomes. As a result, in the context of semi-supervised semantic segmentation, we present a novel solution called the Contrastive Segmentation Consistency (**CSC**) module, designed to enhance the accuracy of pseudo-labels through the incorporation of consistency and separation regularization principles.

Additionally, there is substantial noise among pseudo-labels reported in previous works [16], [24], which hampers the effectiveness of the methods. To mitigate this impact of noise in pseudo-labels, many methods have been devised. One prevalent approach establishes a specific threshold based on the confidence scores of model predictions. Other methods take into account the problem of class imbalance [16], [25], [26], and thus assign lower threshold values to tail classes, vice versa. However, it is crucial to note that these methods cannot be directly adapted for contrastive loss. Therefore, we present a novel pixel-level contrastive learning loss adding restrictions on the positive sample pairs. In this way, the model can focus on learning only the most accurate feature of positive samples while targeting on difficult negative samples. To sum up, our approach addresses these aforementioned critical challenges in the following ways:

1. Intra-Class Feature Consistency: The **CSC** module preserves intra-class feature consistency during model iterations, promoting stable and consistent learning. This mitigates the issue of significant feature fluctuations over time, allowing the model to be trained more effectively.

2. Inter-Class Feature Separation: Our method emphasizes the separation between features corresponding to different classes (inter-class features). This emphasis on feature separation enhances the model's ability to distinguish between classes, resulting in improved segmentation performance.

3. Pixel-Level Contrastive Learning Loss: We introduce a novel pixel-level contrastive learning loss function tailored to combat the substantial noise encountered in the semi-supervised setting. This loss function is specifically designed to align contrastive learning with the challenges posed by pseudo-labels, making it a more effective solution.

Our experimental evaluation shows that the proposed **CSC** module optimizes the performance of pseudo-labeling methods by making critical adjustments to both feature dimensions and loss function. Our module thereby enhances the overall quality of semi-supervised semantic segmentation outcomes.

II. Related Work

A. Semi-supervised Learning

Semi-supervised learning (SSL) aims at leveraging a few labeled data and a large amount of unlabeled data together to train a computer vision model. Existing methods mainly focus on pseudo-labeling. Pseudo-labels based method assigns pseudo-labels to unlabeled data during training. Then, they retrain the network and acquire higher quality pseudo-labels iteratively. For the purpose of getting better pseudo-labels, threshold-based methods [6] are proposed to filter pseudo-labels. Some methods like [7] further propose using dynamic thresholding based on confidence, entropy, and other measures to address the issue of class-imbalance. Since the aforementioned methods are prone to confirmation bias, [8] employs the feature map generated via memory-smoothed pseudo-labeling for contrastive learning. [8] proposes a graph contrast learning method to improve the pseudo-labels, which jointly learns class probabilities and low-dimensional embeddings of the training data to achieve mutual improvement. Similarly, the proposed method employs a contrastive module to construct the feature, aiming to minimize confirmation bias.

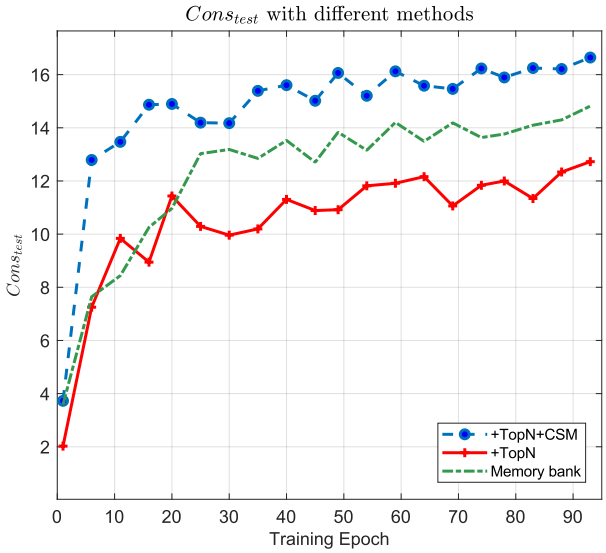

Fig. 2: The measure of consistency between inter-class image features learned by the semantic segmentation model using different method. Redline denotes contrastive learning with our restriction $TopN$ on positive samples in loss function. Green line denotes memory bank application. Blue line denotes contrastive learning using our Consistency and Separation Module (CSM) on memory bank.

B. Semi-supervised Semantic Segmentation

Early studies in the domain of semi-supervised semantic segmentation utilize adversarial generative models for the generation of high quality pseudo-labels [27]–[29]. Recently, SSL principles such as consistency regularization and pseudo-labeling have expanded from image classification to semantic segmentation. In the line of consistency regularization,

[30]–[36] focus on perturbation strategies that employ different augemented views; [37]–[39] introduce feature-level perturbations to maintain consistency in model predictions. Among these works, [39] explores the potential of perturbation in larger scale and utilizes a dual-stream perturbation method.

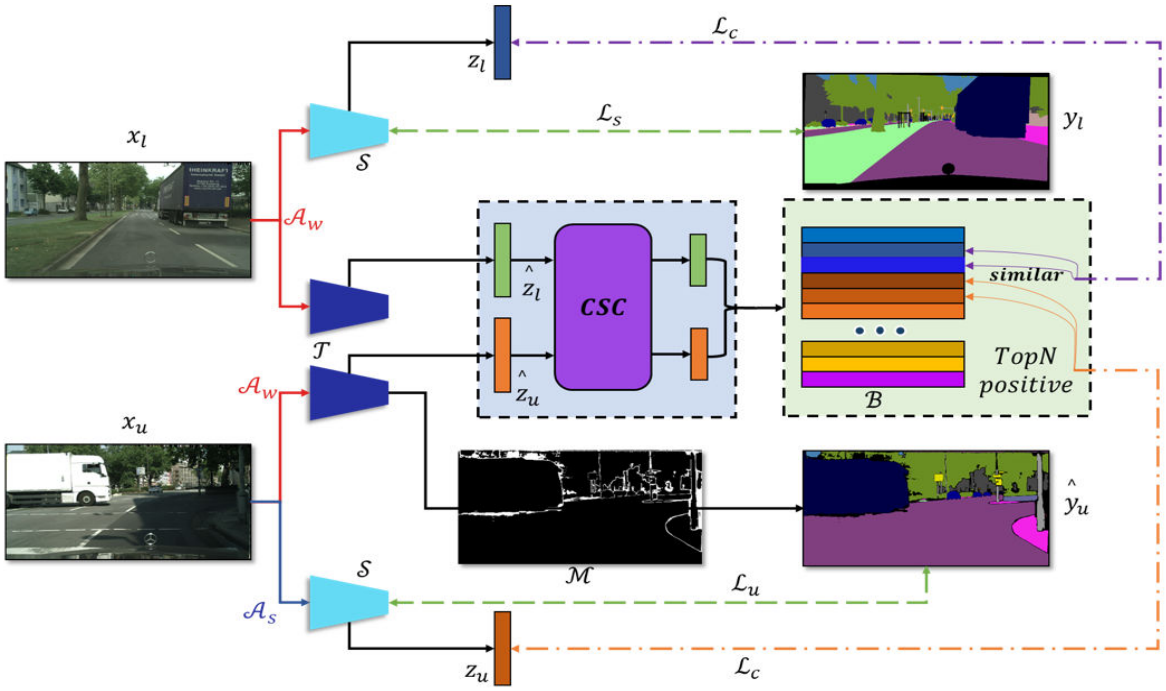

Fig. 3: A complete pipeline for our **CSC** approach. In this scheme, labeled data x_l is weakly augmented by $\mathcal{A}_w$. For unlabeled data x_u, it goes through both weak augmentation $\mathcal{A}_w$ and strong augmentation $\mathcal{A}_s$. For teacher model $\mathcal{T}$, it is trained on labeled data but generates pseudo-label $\hat{y}_i^u$. Feature vectors generated by the teacher model will go through our CSC module and store in memory bank B. For student model $\mathcal{S}$, it is train on both labeled and unlabeled data, supervised by both labels and pseudo-labels. In addition, the feature representations z_l and z_u generated by the student model will be supervised through a contrastive loss including our $TopN$ positive method.

[38], [40], [41] encourage consistency between the predictions from different networks. The core idea of pseudo-labeling is high confidence scores and model retraining [18], [24], [25], [35]. Basic approaches, such as [6], [25], [38], [39], [42], employ various criteria like confidence, margin, and entropy to effectively minimize noise by identifying and filtering out false labels. Furthermore, [40], [43]–[48] introduce auxiliary modules to correct pseudo-labels. Although shown to be effective, training an auxiliary network poses challenges [40], [44]. Certain modules may not integrate well with other methods [45]. Recently, contrastive learning is used in semi-supervised segmentation to adjust feature alignment and mitigate confirmation bias [15]–[17], [19], [49], [50]. Similar to these works, our method also employs a contrastive learning module to enhance separation among inter-class features. While their works primarily focus on the impact of region or pixel features on constructing the embedding space, our work emphasizes on how to avoid noise and confirmation biases in semi-supervised segmentation and how to effectively separate inter-class features.

C. Supervised Contrastive Learning

While supervised contrastive learning [51] can be directly applied to downstream tasks, self-supervised contrastive learning [11], [52], [53] is commonly applied to pre-training models. Nevertheless, both of them learn representations in a discriminative fashion. [4] introduces a pixel-wise contrastive loss for semantic segmentation. In the task of semantic segmentation, positive samples consist of pixels belonging to the same class as the given pixel, while negative samples encompass pixels from different classes. However, it is impractical to sample all pixels for high-resolution images which would cost huge memory and slow training time. [5] simply discard negative samples from different images. While [4] considers the pixels that the model predicts incorrectly as hard examples and further proposes semi-hard example sampling strategy. In [15], pseudo-labels are simply employed instead of discarding them. [50] utilizes a memory bank to exclusively store high-quality pixel-level features obtained from labeled data. It is worth noting that the aforementioned methods either have not fully utilized unlabeled data or have introduced significant amounts of noise, leading to a decrease in model performance. Based on these methods, we further explore how to construct a discriminative embedding space to eliminate noise. In addition, an effective sampling method is proposed to better exploit unlabeled data.

III. METHOD

A. Overview

Given a labeled dataset $\mathcal{D}_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ and an unlabeled dataset $\mathcal{D}_u = \{(x_i^u)\}_{i=1}^{N_u}$ where $N_l \ll N_u$. Our task aims to learn a segmentation network by leveraging both the labeled and unlabeled data. The proposed framework is illustrated in Fig. 3. Similar to other approaches [19], [24], [47], [50], our **CSC** framework also contains a student model $\mathcal{S}$ and a teacher model $\mathcal{T}$. Besides, we employ various augmentation methods

in different branches as illustrated in Fig. 3. In addition, the model incorporates a projection head p, placed after the encoder f, while maintaining the original structure of the decoder d. In this work, the overall optimization target can be formulated as:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u + \lambda_c \mathcal{L}_c \tag{1}$$

where $\mathcal{L}_s, \mathcal{L}_u, \mathcal{L}_c$ represent the supervised loss, unsupervised loss and contrastive loss respectively. λ_u, λ_c are weights of the unsupervised loss and contrastive loss respectively. Given a labeled image, a typical supervised cross-entropy loss applied at per-pixel locations is:

$$\mathcal{L}_s = \frac{1}{|\mathcal{D}_l|} \sum_{(x_i^l, y_i^l) \in \mathcal{D}_l} l^{ce}(d_{\mathcal{S}}(f_{\mathcal{S}}(\mathcal{A}_w(x_i^l))), y_i^l) \tag{2}$$

where l^{ce} represents the cross-entropy (CE) loss, and $f_{\mathcal{S}}, d_{\mathcal{S}}$ represent the encoder and the decoder of the student model. $\mathcal{A}_w$ is the weak augmentation method. Correspondingly, the strong augmentation method is denoted as $\mathcal{A}_s$. For unlabeled images, the unsupervised loss $\mathcal{L}_u$ can be computed as:

$$\mathcal{L}_u = \frac{1}{|\mathcal{D}_u|} \sum_{x_i^u \in \mathcal{D}_u} l^{ce}(d_{\mathcal{S}}(f_{\mathcal{S}}(\mathcal{A}_s(x_i^u))), \hat{y}_i^u) \tag{3}$$

where $\hat{y}_i^u$ is the pseudo-label of x_i^u, it can be obtained by:

$$\hat{y}_i^u(j,k) = \begin{cases} \underset{c}{\text{argmax}}\ prob(c,j,k) & \text{if } \mathcal{M}_{jk} = 1 \\ \text{ignore} & \text{else} \end{cases} \tag{4a}$$

$$prob = softmax(d_{\mathcal{T}}(f_{\mathcal{T}}(\mathcal{A}_w(x)))) \tag{4b}$$

where $prob \in \mathbb{R}^{C \times H \times W}$ represents the probability of each class ($c \in C$) at a given position ($j \in H, k \in W$) on the image, and $\mathcal{M}$ represents a binary mask that utilizes different criteria to filter unreliable pixels. $f_{\mathcal{T}}, d_{\mathcal{T}}$ represent the encoder and the decoder of the teacher model. Most of the existing methods directly use the model predictions to filter unreliable pixels, which inevitably carry a lot of noise. We simply utilize the methods in [25] to obtain our binary mask $\mathcal{M}$. As a result, our contrastive learning module can further eliminate the confirmation bias and noise during training.

B. Pixel-wise Contrastive Learning

To further explore the intra-class structural information and enhance feature discrimination ability, we leverages contrastive learning to pull embeddings from the same class closer together than embeddings from different classes. Consider an augmented image x either from the labeled set or unlabeled set. It is mapped into a representation vector $f(x) \in \mathbb{R}^{D_f \times hw}$ by encoder network f where h, w is respectively the $\frac{1}{s}$ times of H, W. s means the output stride of the encoder f. It is further mapped into a normalized feature map $z = p(f(x)), z \in \mathbb{R}^{D \times hw}$ by projection head. (Consistent with the work of [4], [24], $D = 256$ in all our experiments) Let z_i signify the i-th pixel's normalized feature vector. In supervised contrastive learning, the positive set $\mathcal{P}_i$ is determined to be formed by the pixel feature vectors with the same category while the negative set $\mathcal{N}_i$ is formed by the pixel embedding with the other category correspondingly. When it comes to semi-supervision, the definition is almost identical except that pseudo-labels are used for supervision. Hence the pixel-wise contrastive loss $\mathcal{L}_p$ is formulated as:

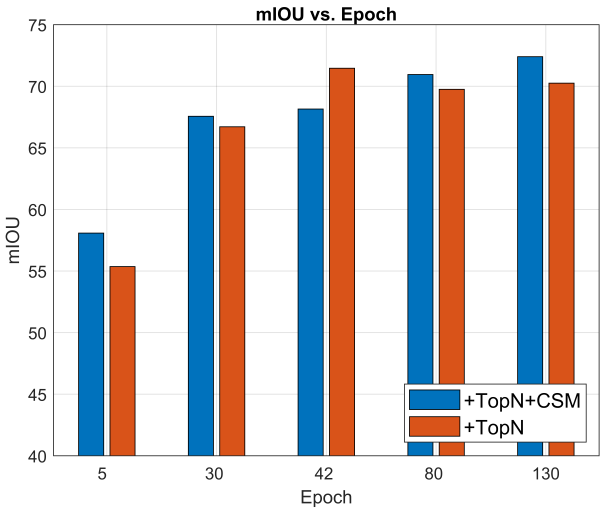

Fig. 4: A concrete contrast between inter-class image features learned by the semantic segmentation model using different method. Orangered column denotes contrastive learning with our restriction $TopN$ on positive samples in loss function. Blue column denotes contrastive learning using our CSM on memory bank.

$$\mathcal{L}_p = -\frac{1}{hw} \sum_{i}^{hw} \sum_{z_j^+ \in \mathcal{P}_i} \log \frac{\mathcal{E}(z_i \cdot z_j^+)}{\mathcal{E}(z_i \cdot z_j^+) + \sum_{z_k^- \in \mathcal{N}_i} \mathcal{E}(z_i \cdot z_k^-)} \tag{5a}$$

$$\mathcal{E}(z_i, z_j) = \exp(\frac{z_i \cdot z_j}{\tau}) \tag{5b}$$

for which τ is the temperature coefficient to control the softness of logit distribution, and ($\cdot$) represents the dot product of normalized feature vectors or the cosine similarity between feature vectors.

As mentioned before, the pixel-wise loss requests all pixels in a mini-batch, thus requiring memory and time costs. Besides, it only considers the intra-image structural information while ignoring the demand of data diversity. As a result, we transition to using the sampling strategy (semi-hard sampling) following [4] and a category-wise memory bank to store image features. Hence the positive set $\mathcal{P}$ and negative set $\mathcal{N}$ are supposed to be selected from the memory bank $\mathcal{B}$ and the mini-batch, and the query feature vector z_i is sampled from limited pixels in the mini-batch rather than all pixels. In general, the memory bank $\mathcal{B} \in \mathbb{R}^{C \times D \times L}$ stores feature vectors from all categories and the length L of $\mathcal{B}$ fulfills the condition: $L \gg hw$.

C. Consistency and Separation Module

Most existing pixel-wise contrastive learning methods rely on using pseudo-labels as the source of supervised information

for unlabeled images. However, in semi-supervision, the model introduces a considerable amount of noise. This high level of noise hinders the application of contrastive learning. As discussed in [24], disregarding false negative samples can result in incorrect separation within the contrastive loss. In this study, we further explore the impact of both false negative samples and false positive samples. From a sampling perspective, even if an original negative sample is wrongly predicted, it is still more likely to be a negative sample. On the contrary, an originally positive sample is more prone to being misclassified as negative. As a result, with the original NCE loss function used in self-supervised learning, there will be excessive positive pairs for a query feature. This may even cause feature misalignment between dissimilar pixels. Therefore, we optimize the contrastive loss function by adding such restriction on positive samples:

$$\mathcal{L}_c = -\frac{1}{n}\sum_{i}^{n}\sum_{z_j^{+'}} \log \frac{\mathcal{E}(z_i, z_j^{+'})}{\mathcal{E}(z_i, z_j^{+'}) + \sum\limits_{z_k^- \in \mathcal{N}_i} \mathcal{E}(z_i, z_k^-)} \tag{6a}$$

$$\mathcal{P}_i^{TopN} = Top(\{z_i \cdot z_j^+ \,|\, z_j^+ \in \mathcal{P}_i, j = 1, ..., |\mathcal{P}_i|\}, N) \tag{6b}$$

$$z_j^{+'} \in \mathcal{P}_i^{TopN} \tag{6c}$$

where n represents the length of the query feature set $\{z\}_i^n$, $Top(A, b) = \{a_i \,|\, a_i \in A,$ sorted in descending order, for $i = 1$ to $b\}$. By selecting the positive pair with the maximum of cosine similarity, we not only focus more on the 'correct' positive sample by filtering out 'false' positive samples easily, but also save on computation and cost.

With further research, we find that the direction of intra-class feature vector changes so drastically that contrastive learning cannot be conducted to train the model in a stable fashion. On the other hand, memory bank is always used in place of large-batch which costs huge memory. But rapidly changing feature makes it difficult to use memory bank to simulate a large-batch. Therefore, we want to make the features more continuous and consistent. Firstly, we define a symbol $Cons$ to explicitly measure the extent of consistency for intra-class features. It is formulated as:

$$Cons(T) = \frac{1}{\mathbb{N}} \sum_{x \in \{x,y\}^{\mathbb{N}}}^{\mathbb{N}} \cos(z^{x,t}, z^{x,t+T}) \tag{7}$$

where $\cos(\ ,\ )$ is the function calculating cosine similarity. x is a given image. $\{x, y\}^{\mathbb{N}}$ is the mini-batch set and $\mathbb{N}$ is the number of images in a mini-batch. y for a labeled image represents label while for an unlabeled image, it represents pseudo-label. $z^{x,t}, z^{x,t+T}$ represent the corresponding features for x at time $t, t+T$ respectively. By calculating the average cosine similarity in pixel level, $Cons$ can accurately reflect the amount of changes in the feature vectors of a given image. Based on this measure, we further fine-tune the target feature in each mini-batch to reduce $Cons(T)$ in order to achieve better representation consistency. The formula for this update method is given by:

$$\bar{z}^c = Norm(\frac{1}{\mathbb{N}}\frac{1}{n_x^c} \sum_{x \in \{x,y\}^{\mathbb{N}}} \sum_{i, y_i = c}^{n_x^c} z_i^x) \tag{8a}$$

$$u_i^c = Norm(\alpha^c \times u_i^c + (1 - \alpha^c) \times \bar{z}^c, u_i^c \in \mathcal{B}(c)) \tag{8b}$$

$$\alpha^c = \frac{L_c}{L_c + \sum_{x \in \{x,y\}^{\mathbb{N}}} n_x^c} \tag{8c}$$

where $\bar{z}^c$ is the average value of features in category c. n_x^c denotes the number of samples in category c for a given image x. $\sum_{i,y_i=c}^{n_x^c}$ indicates the summation over the features x_i whose corresponding class label y_i is equal to c. $\mathcal{B}(c)$ stores the features in category c, and u_i^c denotes the feature vector in memory bank $\mathcal{B}$. α^c is a trade-off variable that balances the impact of mini-batch on memory bank of category c. $Norm()$ is normalization function for feature vectors. This update method enables the memory bank to better imitate the image features learned by model in the current mini-batch. As a result, the model can incrementally learn image features through contrastive learning. Since training model with such deficiency is one potential reason for unsatisfying pseudo-label generation, our method directly targets on this problem and thereby gives full scope to contrastive learning.

In contrastive learning, one concrete measure of its performance is the separation between inter-class features. As a result, we define a symbol $Separ$ to explore how separate image features for each class are.

$$Separ = \sum_{c_i, c_j \in C} \bar{u}^{c_i} \cdot \bar{u}^{c_j} + V \tag{9a}$$

$$\bar{u}^c = Norm(\frac{1}{L_c} \sum_{u_k^c \in \mathcal{B}(c)} u_k^c) \tag{9b}$$

where $\bar{u}^c$ denotes the average value of the feature in memory bank $\mathcal{B}(c)$, V is a constant to keep $Separ$ a positive value.

In the experiment shown by Fig. I, the amount of $Separ$ for $TopN$ and memory bank has always been changing and maintained in a high range of values. As a result, we develop a regularization method to ensure the model to always learn sufficient amount of separation between inter-class features. Specifically, the model records the best separation $Separ_b$ during the entire training process. Every time the memory bank is going to have an update, the model checks whether $Separ$ is within a specific range x of $Separ_b$. If not, this update will be canceled and the model continues to learn the next few iterations. Therefore, the image features learned by the model are always more separate and experimental results also proved this.

Overall, the pseudocode of our **CSC** module is illustrated in Algorithm 1.

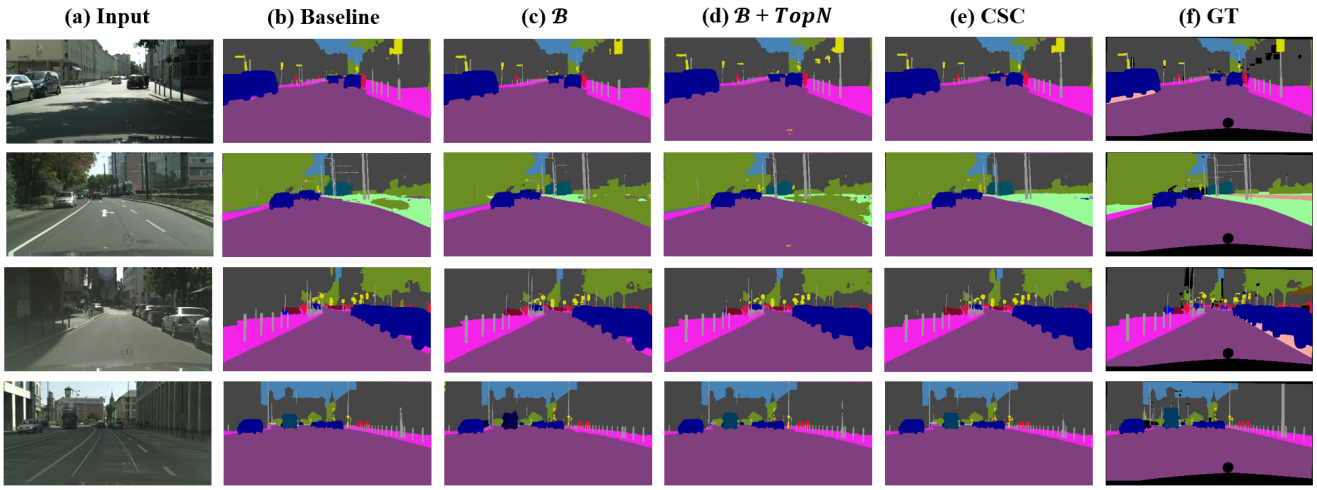

Fig. 5: Visual comparison between different components and baseline on $1/8$ of labeled data. For each row from left to right: (a) input image, (b) semi-supervised baseline, (c) memory bank application, (d) positive sample restriction in loss function, (e) our complete **CSC** approach, (f) ground truth.

Algorithm 1 CSC Pseudocode

SET the student model $\mathcal{S}$, teacher model $\mathcal{T}$
SET the encoder f, decoder d, projection head p
SET the weak augmentation $\mathcal{A}_w$, strong augmentation $\mathcal{A}_s$
SET the semi-hard sampling method $Sampling$, memory bank at time t $\mathcal{B}_t$
SET pixel-wise loss function $\mathcal{L}_p$, our loss function $\mathcal{L}_c$
SET consistence operation f_{cons}, dequeue and enqueue operation f_{que}, Separation symbol $Separ$
SET the labeled data x_l, labels y_l, unlabeled data x_u, pseudo-labels y_u in a mini-batch.

$feats_l, \hat{y}_l = p_{\mathcal{S}}(f_{\mathcal{S}}(x_l)), d_{\mathcal{S}}(f_{\mathcal{S}}(x_l))$
$feats_u, \hat{y}_u = p_{\mathcal{S}}(f_{\mathcal{S}}(x_u)), d_{\mathcal{S}}(f_{\mathcal{S}}(x_u))$
for $i = 1$ **to** c **do**
  if len($\mathcal{B}_t$(i))==0 **then**
    $L = \mathcal{L}_p$
    break
  else
    $L = \mathcal{L}_c$
  end if
end for

$\hat{feats}_l = Sampling(feats_l, \hat{y}_l, y_l)$
$\hat{feats}_u = Sampling(feats_u, \hat{y}_u, y_u)$

Update $\mathcal{S}$ to minimize $L(\hat{feats}_l)$, $L(\hat{feats}_u)$
EMA Update $\mathcal{T}$ with $\mathcal{S}$

$\mathcal{B}_{tmp} = f_{cons}(\mathcal{B}_t, \hat{feats}_l, \hat{feats}_u)$
$\mathcal{B}_{t+1} = \mathcal{B}_{tmp}\, if\ Separ(\mathcal{B}_{tmp}) < Separ(\mathcal{B}_t)\, else\, \mathcal{B}_t$
$\mathcal{B}_{tmp} = f_{que}(\mathcal{B}_{t+1}, feats_l, feats_u)$
$\mathcal{B}_{t+2} = \mathcal{B}_{tmp}\, if\ Separ(\mathcal{B}_{tmp}) < Separ(\mathcal{B}_{t+1})\, else\, \mathcal{B}_{t+1}$

Return the trained neural network model $\mathcal{S}$, $\mathcal{T}$

IV. Experiments

We conduct our experiments on two commonly used dataset: Cityscapes and PASCAL VOC 2012. The mean intersection-over-union (mIoU) scores of our method on different benchmarks are reported in the following sections.

A. Datasets

The Cityscapes dataset [54] is a comprehensive autonomous driving dataset comprising a diverse collection of stereo video sequences captured in real-world urban environments across 50 different cities. It includes meticulously annotated pixel-level information for 5,000 frames, alongside a larger set of 20,000 frames with weak annotations for 19 distinct semantic categories. Each image in this dataset maintains a fixed resolution of 2048 pixels in width and 1024 pixels in height. The PASCAL VOC 2012 dataset [55], initially designed for visual object class recognition, features 20 object classes of interest along with a background class. Its standard dataset partitions for training, validation, and testing encompass 1,464, 1,449, and 1,556 images, respectively. In our experiment, we choose the blender setting to select labeled data. In this case, it selects among the entire pool of 10,582 images. For each dataset, we compare our method **CSC** with other methods under $1/2, 1/4, 1/8$, and $1/16$ partition protocols.

B. Implementation Details

1) Network Structure: For effective comparison with previous studies, we use ResNet-101 [56] pre-trained on ImageNet [57] as the backbone and DeepLabv3+ [58] as the decoder in our experiment. Next, we proceed to train both the teacher and student models using cross-entropy loss, utilizing both strong and weak data augmentation techniques on supervised data. The student model is trained on both labeled and unlabeled target data, while the teacher model is exclusively trained

on labeled data. Following [24], Both the segmentation head and the projection head consist of two Convolution-Batch Normalization-ReLU (Conv-BN-ReLU) blocks. These blocks maintain the feature map resolution, with the initial block reducing the number of channels by half. The segmentation head functions as a pixel-level classifier, transforming the 512-dimensional features generated by the Atrous Spatial Pyramid Pooling (ASPP) module into C classes, where C represents the number of semantic classes.

$\mathcal{B}$	$\mathcal{L}_p$	1/8(372)
		71.68%
✓		73.03%
	✓	72.53%

TABLE I: Ablation study on the essence of memory bank, including the semi-supervised baseline, category-wise memory bank $\mathcal{B}$ and pixel-wise contrastive loss $\mathcal{L}_p$ on $1/8$ labeled data. The mean IoU is reported on this Cityscapes benchmark. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

$\mathcal{B}$	$\mathcal{L}_p$	1/8(1323)
		77.15%
✓		77.68%
	✓	77.42%

TABLE II: Ablation study on the essence of memory bank, including the semi-supervised baseline, category-wise memory bank $\mathcal{B}$ and pixel-wise contrastive loss $\mathcal{L}_p$ on $1/8$ labeled data. The mean IoU is reported on this PASCAL VOC 2012 benchmark. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

2) Optimization: For cityscapes, we use stochastic gradient descent (SGD) optimizer with initial learning rate 0.00125, weight decay 0.0005, crop size 640×640, batch size 2 and training epochs 200. For PASCAL VOC 2012, we also use stochastic gradient descent (SGD) optimizer with initial learning rate 0.00025, weight decay 0.0001, crop size 480×480, batch size 4 and training epochs 80. In the meantime, we employ the polynomial policy to gradually reduce the learning rate throughout the training process: $lr = lr_{init} \cdot (1 - \frac{iter}{totaliter})^{0.9}$

C. Ablation Study of *CSC*

1) Essence of Memory Bank: The fundamental reason behind incorporating memory bank into our method is to elevate the significance of data diversity to generalize more accurate image features in semi-supervised semantic segmentation. Therefore, by applying memory bank to our default setting (semi-supervised baseline with only $\mathcal{L}_l$ and $\mathcal{L}_u$), model performance can immediately increase by 1.35%, as shown by TABLE I. In addition, as our method targets on addressing the problem of learning effective image features, this switch to memory bank can give full scope to our method. As shown by ablation study result (TABLE III) for each component, our method has an overall 3.96% increase in mIoU score. Unlike the conventional memory bank, ours maintains the dictionary as a queue of data samples. This not only allows us to reuse the embedding from the immediate preceding mini-batches but also abandons the oldest mini-batch in each update. This change is particularly important in a semantic segmentation scenario since inconsistent features will hinder the encoder's ability to learn better features. Similar experimental pattern can also be seen on PASCAL VOC 2012 dataset in TABLE II

$\mathcal{B}$	$TopN$	CSM	1/8(372)
			71.68%
✓			73.03%
✓	✓		74.16%
✓	✓	✓	**75.64%**

TABLE III: Ablation study on the contribution of each component, including the semi-supervised baseline, category-wise memory bank $\mathcal{B}$, restriction on positive samples $TopN$ and our Consistency and Separation Module CSM on $1/8$ labeled data. The mean IoU is reported on this Cityscapes benchmark. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

$\mathcal{B}$	$TopN$	CSM	1/8(1323)
			77.15%
✓			77.68%
✓	✓		78.47%
✓	✓	✓	**78.68%**

TABLE IV: Ablation study on the contribution of each component, including the semi-supervised baseline, category-wise memory bank $\mathcal{B}$, restriction on positive samples $TopN$ and our Consistency and Separation Module CSM on $1/8$ labeled data. The mean IoU is reported on this PASCAL VOC 2012 benchmark. † means we reproduce the approach under our limited configuration. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

2) Contribution of Each Component: We conduct an ablation study on Cityscapes to verify the effectiveness of the proposed Consistency and Separation Module(CSM) and adjustment on loss function. As shown in TABLE III, when the NCE loss function is adjusted to choose the most similar positive pair, the performance of our model can be improved at about 1.13%. When our regularization on consistency and separation is established, there is an additional boost of 1.48%. By adding our method, the mIoU can be increased from 73.03% to 75.64% for Cityscapes when utilizing $1/8$ of the labeled data. Similarly, for PASCAL VOC 2012(TABLE IV), the mIoU can be elevated from 77.68% to 78.47%. This result confirms that adopting our CSM approach enhances inter-class feature distinguishability and intra-class feature similarity in contrastive learning. In addition, as shown by the third and fourth rows in TABLE III and TABLE IV, it becomes evident that the new contrastive loss significantly improves the ultimate performance.

3) Impact of Hyperparameters: In our approach, there are three kinds of hyperparameters: $TopN$, λ_c, λ_u, memory bank size and its update frequency. The selection of hyperparameters is paramount in fine-tuning the model's behavior and performance. To simplify the construction of our loss function, we set the weights for both λ_c and λ_u to 1 following previous work [24], [50]. By eliminating unnecessary weights, we not only preserve but also verify the significance of contrastive loss, where our major contribution lies in. $TopN$ is an another configuration of our loss function, we want to evaluate its

Method	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)
SupOnly	65.74%	71.55%	75.80%	77.13%
MT [59]	70.51%	71.53%	73.02%	76.58%
CCT [37]	71.86%	73.68%	76.51%	77.40%
GCT [40]	70.90%	73.29%	76.66%	77.98%
U2PL† [24]	74.73%	77.32%	77.89%	78.20%
CSC(w/CutMix)	**76.16%**	**78.68%**	**78.17%**	**78.44%**

TABLE V: Comparison with state-of-the-arts on the Pascal VOC 2012 *val* set under different partition protocols. † means we reproduce the approach under our limited configuration. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

significance on the performance of contrastive learning. We thereby conduct a systematic exploration of these hyperparameters to provide a comprehensive understanding of their impact. Empirical results for different memory bank size and update frequency settings on Cityscapes with only 1/8 training data are shown in TABLE VII. From these results, it becomes evident that across a broad spectrum of memory bank size and its update frequency configurations, our **CSC** method consistently demonstrates high mIoU performance in Cityscapes. On one hand, we notice that different memory bank size S doesn't make a huge difference in model performance. On the other hand, memory bank update frequency U_f delivers the best performance when setting to 10. Therefore, in our experiment, we choose to set S to 2000 and U_f to 10. For $TopN$, we have tested $1, 3, 5$ and 10 to investigate how radical the loss function abandons unreliable positive samples from the memory bank. As shown in TABLE VI, model performance decreases as $TopN$ increases. As a result, we choose 1 as our final value for $TopN$.

$TopN$	1	3	5	10
1/8(1323)	**78.47%**	78.31%	77.51%	77.57%

TABLE VI: Ablation study on the impact of different hyperparameters $TopN$ on $1/8$ labeled data. The mean IoU is reported on this PASCAL VOC 2012 benchmark. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

S	U_f	1/8(1323)
1000	20	77.83%
2000	10	**78.47%**
2000	20	77.72%
2000	30	78.24%
3000	20	77.60%

TABLE VII: Ablation study on the impact of different hyperparameters S and U_f on $1/8$ labeled data. The mean IoU is reported on this PASCAL VOC 2012 benchmark. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

D. Comparison with State-of-the-Arts

We compare our approach with recent state-of-the-art models and the semi-supervised baseline. Experimental results on Pascal VOC 2012 dataset are listed in TABLE V. We test our results under $1/16, 1/8, 1/4, 1/2$ partition protocols in the dataset to verify the generalization capability of our **CSC** approach. As shown in TABLE V, **CSC** outperforms the sup-only approach by $+10.43, +7.13, +2.37$ and $+1.31$ with $1/16, 1/8, 1/4, 1/2$ amount of labeled data available. To compare directly with the SOTA model, we reproduce **U2PL** [24]. As shown in the last two rows of TABLE V, **CSC** outperforms **U2PL** [24] impressively when lower amount of labeled data is available. This pattern can also be seen in the performance of other existing methods. The results indicate that our contrastive learning method has been excessively powerful in a semi-supervision setting. Our method has effectively utilized the limited labeled data to learn rich semantic information through a consistent memory bank. Due to memory bank's ability to restore large amounts of feature vectors, our model can have access to the entire dataset throughout the training process. Note that our approach has also targeted on positive sample restriction, the segmentation model can thereby focus more on negative samples instead of positive samples. As a result, the feature representations learned through contrastive learning are more suitable in the task of segmentation.

Method	1/16(186)	1/8(372)
SupOnly	62.96%	69.81%
MT [59]	68.05%	73.56%
CCT [37]	69.32%	74.12%
CPS [41]	69.78%	74.31%
U2PL† [24]	67.12%	73.68%
CSC(w/CutMix)	**69.13%**	**75.64%**

TABLE VIII: Comparison with state-of-the-arts on the Cityscapes *val* set under different partition protocols. † means we reproduce the approach under our limited configuration. (DeepLabv3+ and ResNet-101 ImageNet pre-trained)

On Cityscapes dataset, experimental results are illustrated in TABLE VIII. Although our **CSC** method has a small amount of deficiency compared to recent existing methods under $1/16$ of labeled data, we have achieved huge success when only $1/8$ of labeled data is available.

V. Conclusion

We investigate three problems of contrastive learning in the context of semi-supervised semantic segmentation. Through visualized experimental results of these problems, we proposed a novel contrastive loss function to mitigate the challenge of substantial noise among pseudo-labels. In addition, we also present a **CSC** module designed for learning better feature representations. The module fundamentally improves the quality of pseudo-labels and thus help the model to achieve SOTA performance on both Cityscapes and Pascal VOC 2012 datasets.

REFERENCES

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:3429309

[2] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3141–3149, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52180375

[3] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:140309863

[4] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. V. Gool, "Exploring cross-image pixel contrast for semantic segmentation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7283–7293, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231719378

[5] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, "Contrastive learning for label efficient semantic segmentation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 603–10 613, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229152360

[6] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *ArXiv*, vol. abs/2001.07685, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210839228

[7] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Neural Information Processing Systems*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239016453

[8] J. Li, C. Xiong, and S. C. Hoi, "Comatch: Semi-supervised learning with contrastive graph regularization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9475–9484.

[9] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[10] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5982–5991.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *ArXiv*, vol. abs/2002.05709, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211096730

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:207930212

[13] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *ArXiv*, vol. abs/2006.09882, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219721240

[14] J.-B. Grill, F. Strub, F. Altch'e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *ArXiv*, vol. abs/2006.07733, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219687798

[15] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," *ArXiv*, vol. abs/2104.04465, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233204603

[16] Y. Zhou, H. Xu, W. Zhang, B.-B. Gao, and P.-A. Heng, "C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7016–7025, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244114902

[17] J. Zhang, T. Wu, C.-Y. Ding, H. Zhao, and G. Guo, "Region-level contrastive and consistency learning for semi-supervised semantic segmentation," *ArXiv*, vol. abs/2204.13314, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248427159

[18] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia, "Semi-supervised semantic segmentation with directional context-aware consistency," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1205–1214, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233305151

[19] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7253–7262, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237259945

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[21] M. Kaya and H. Ş. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[22] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," *arXiv preprint arXiv:2010.04592*, 2020.

[23] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 798–21 809, 2020.

[24] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4238–4247, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247315180

[25] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," in *Neural Information Processing Systems*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238582727

[26] Y. Wang, H. Chen, Q. Heng, W. Hou, M. Savvides, T. Shinozaki, B. Raj, Z. Wu, and J. Wang, "Freematch: Self-adaptive thresholding for semi-supervised learning," *ArXiv*, vol. abs/2205.07246, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248811614

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1033682

[28] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *British Machine Vision Conference*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3456798

[29] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 1369–1379, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201058657

[30] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *ArXiv*, vol. abs/1708.04552, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:23714201

[31] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. J. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:152282661

[32] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1368–1377, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220545989

[33] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. D. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv: Computer Vision and Pattern Recognition*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:218487557

[34] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, "A simple baseline for semi-supervised semantic segmentation with strong data augmentation*," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8209–8218, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233241099

[35] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *ArXiv*, vol. abs/2010.09713, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:224705241

[36] S. Chen, X. Jia, J. He, Y. Shi, and J. Liu, "Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation," *2021 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR), pp. 11 013–11 022, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232147231

[37] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 671–12 681, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214605688

[38] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4248–4257, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244709639

[39] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," *ArXiv*, vol. abs/2208.09910, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251719486

[40] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. H. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," *ArXiv*, vol. abs/2008.05258, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221103770

[41] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235293837

[42] Z. Feng, Q. Zhou, G. Cheng, X. Tan, J. Shi, and L. Ma, "Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum," *ArXiv*, vol. abs/2004.08514, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:215827846

[43] D. Kwon and S. Kwak, "Semi-supervised semantic segmentation with error localization network," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9947–9957, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247958032

[44] R. Mendel, L. A. D. Souza, D. Rauber, J. P. Papa, and C. Palm, "Semi-supervised segmentation based on error-correcting supervision," in *European Conference on Computer Vision*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222178391

[45] L. Wu, L. Fang, X. He, M. He, J. Ma, and Z. Zhong, "Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 8827–8844, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255667855

[46] P. Hu, S. Sclaroff, and K. Saenko, "Leveraging geometric structure for label-efficient semi-supervised scene segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 6320–6330, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252714800

[47] D. Guan, J. Huang, A. Xiao, and S. Lu, "Unbiased subclass regularization for semi-supervised semantic segmentation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9958–9968, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247594520

[48] Y. Jin, J. Wang, and D. Lin, "Semi-supervised semantic segmentation via gentle teaching assistant," *ArXiv*, vol. abs/2301.07340, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255998490

[49] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 656–11 665, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249980756

[50] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8199–8208, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233423665

[51] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *ArXiv*, vol. abs/2004.11362, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216080787

[52] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *ArXiv*, vol. abs/2003.04297, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:212633993

[53] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4591284

[54] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:502946

[55] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:4246903

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84 – 90, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:195908774

[58] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3638670

[59] A. Tarvainen and H. Valpola, "Weight-averaged consistency targets improve semi-supervised deep learning results," *ArXiv*, vol. abs/1703.01780, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:2759724