Adversarial Dense Contrastive Learning for Semi-Supervised Semantic Segmentation

Ying Wang, Ziwei Xuan, *Member, IEEE*, Chiuman Ho, and Guo-Jun Qi, *Fellow, IEEE*

***Abstract*— Semi-supervised dense prediction tasks, such as semantic segmentation, can be greatly improved through the use of contrastive learning. However, this approach presents two key challenges: selecting informative negative samples from a highly redundant pool and implementing effective data augmentation. To address these challenges, we present an adversarial contrastive learning method specifically for semi-supervised semantic segmentation. Direct learning of adversarial negatives is adopted to retain discriminative information from the past, leading to higher learning efficiency. Our approach also leverages an advanced data augmentation strategy called AdverseMix, which combines information from under-performing classes to generate more diverse and challenging samples. Additionally, we use auxiliary labels and classifiers to prevent over-adversarial negatives from affecting the learning process. Our experiments on the Pascal VOC and Cityscapes datasets demonstrate that our method outperforms the state-of-the-art by a significant margin, even when using a small fraction of labeled data.**

***Index Terms*— Semi-supervised learning, semantic segmentation, adversarial contrastive learning, data augmentation.**

I. Introduction

As ONE of the fundamental computer vision tasks, semantic segmentation plays a key role in a broad range of applications such as autonomous driving, computational photography, augmented reality and medical imaging. It involves pixel-wise classification of an image into a set of target categories. Deep learning-based models have shown notable advancements in semantic segmentation [1], [2], [3], [4], [5]. However, training a strong model for segmentation relies on large amount of finely-annotated data. Getting pixel-level labels is very costly and has become a major bottleneck in many applications. Our focus is on semi-supervised scenario where the training data contains a small set of labeled ones and a large set of unlabeled ones. We aim at fully exploiting the benefit of both labeled and unlabeled data in a unified framework. Various techniques have been presented for semi-supervised learning (SSL) [6], [7], [8], [9], [10]. Generally, traditional semantic segmentation in semi-supervised settings tends to utilize spatially nearby information while ignoring the global context. To mitigate this issue, it is natural to resort to methods, which are efficient in extracting class-level semantics from a global context, to enhance the model.

In recent years contrastive learning [11], [12], [13], [14], [15] has received tremendous attention due to its prominent performance in self-supervised learning. The basic idea is to utilize InfoNCE [16] to maximize the mutual information between different views. It learns good feature representations by pulling together similar sample pairs while pushing apart dissimilar ones. Hence the downstream tasks benefit from the more discriminative representations. This characteristic of contrastive learning complements the class-level information to traditional semi-supervised semantic segmentation and enhances the performance of segmentation models [17]. However, combining semi-supervised semantic segmentation with contrastive learning also faces the challenges from constructing *hard negative samples* and designing *proper data augmentation*.

Negative samples are critical ingredients in contrastive learning and it is desirable to contrast the queries with hard negative samples from different classes. Some researchers deal with this challenge by adopting a large batch [12], which nevertheless suffers from the large computational overhead. Alternatively, others maintain a memory bank or queue of negative samples [11], [18] and update a small portion in each iteration. In particular, adversarial contrastive learning [19], [20], [21], [22], [23] stands out not only in image classification but also in several dense prediction tasks. Instead of updating a small portion of negatives in each iteration, AdCo [20] proposed to directly learn the negative adversaries through min-max optimization. Compared with other methods, it enables continuous tracking of the rapid changes in the feature representations without need for a large batch size and avoids disposal of rich information in the past by the queue-based method.

As an extended work of AdCo [20], this paper explores the benefits of adversarial contrastive learning for dense prediction in semi-supervised settings. We construct hard negative samples by solving a minimax problem w.r.t. to the InfoNCE based loss. On one side, the optimization of the segmentation model learns to minimize the contastive loss; on the other side, the updates for the whole memory bank of negative samples play against the segmentation model so as to mine hard negative samples. As a consequence, the segmentation model can efficiently discriminate queries against adversarially

Manuscript received 28 May 2022; revised 6 March 2023 and 6 June 2023; accepted 11 July 2023. Date of publication 1 August 2023; date of current version 7 August 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ajmal S. Mian. *(Ying Wang and Ziwei Xuan contributed equally to this work.) (Corresponding author: Ying Wang.)*

Ying Wang and Ziwei Xuan are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: yingwang0022@gmail.com; xuan64@tamu.edu).

Chiuman Ho and Guo-Jun Qi are with the OPPO U.S. Research Center, Palo Alto, CA 94303 USA (e-mail: chiuman@oppo.com; guojunq@gmail.com).

Digital Object Identifier 10.1109/TIP.2023.3299196

hard negatives which are learned by the memory bank. However, direct application of adversarial contrastive learning to dense prediction in semi-supervised setting can raise new concerns. There is a chance that adversarial contrastive learning pushes negatives across class boundaries towards the positive anchors due to over-adversarial gradient updates [24]. For dense prediction, the false negative problem can be more severe with many fine-grained instances such as pixels, patches and objects. The false negatives are likely to undermine the model and harm its performance in the absence of supervised labels. Though it is natural to think of using the segmentation head to track false negatives in the memory bank, it poses a new risk: the head usually relies on neighboring pixels for prediction, but negatives in the memory bank are not necessarily related to each other. This can cause biased detection of over-adversarial negatives. To address this false-negative challenge in such adversarial contrastive learning framework, we utilize an auxiliary classifier to generate 'auxiliary labels' on top of a class-wise memory bank storing negative keys. We propose to timely update auxiliary labels and identify 'pseudo'-false negatives when their labels flip to the same category of the corresponding queries, and further replace those 'pseudo'-false negatives with reliable ones from the current mini-batch. The false negatives can be effectively alleviated by supervision from auxiliary labels. In addition, we adopt an informative sampling strategy for memory bank initialization which leads to more efficient adversarial training. The involved negative keys are actively sampled from reliable representations. To avoid under/over-confident samples and better leverage convincing ones, queries and positive keys are sampled from thresholded representations based on the confidence of associated logits.

The design of *proper data augmentation* is another crucial component in contrastive learning. A proper data augmentation is expected to generate positive counterparts by creating non-essential variations but reserving essential features of the input w.r.t. the downstream task [12], [25]. Specifically, proper data augmentation is desired to force the learned representations to focus on features related to semantic segmentation under the circumstance of large sample redundancy. In literature, many advanced augmentations have been presented for image classification, such as [26] and [27], while approaches for semantic segmentation are under-explored. Moreover, uniform sampling, which is adopted by many traditional works in semi-supervised learning, is prone to causing sample insufficiency on under-performing classes (i.e. classes where the segmentation model has compromised performance), especially when the dataset is imbalanced over classes. Considering that samples from under-performing classes are natural components of hard negatives contrasting queries, such insufficiency can further undermine the efficiency of adversarial contrastive learning for a segmentation model. To this end, we design an advanced data augmentation method named AdverseMix for semantic segmentation in semi-supervised settings. As a data mixing strategy, AdverseMix emphasizes the importance of sampling instances from under-performing classes. Built on class-wise 'Copy-and-Paste' operations which retain boundary information of instances, AdverseMix assigns higher probability to select samples from under-performing classes by timely tracking class accuracy metric. This keeps the contours of instances from under-performing classes being intact and preserves the boundary pixels which are usually hard for models to segment correctly. AdverseMix also leverages the inter-class correlation and actively chooses images containing more semantically related classes to mix. The composite images entail novel and challenging samples, which encourage generalization to under-performing patterns and thus improve the efficiency of adversarial contrastive learning.

The key contributions of our work include the following:

- We propose an adversarial contrastive learning method for semi-supervised dense prediction. An auxiliary classifier is designed to instantly track and exclude potential false negatives which is critical for dense prediction.
- We present an advanced data mixing method termed AdverseMix which fully exploits information from minor classes and stitches more diverse yet challenging samples. The composite images consist of more semantically closed patterns which provide incentives to contrastive learning.
- We informatively sample negative keys of high quality to initialize the memory bank, as well as to replace the excluded 'pseudo'-false negatives. This not only alleviates biased issue in adversarial contrastive learning, but also promotes efficiency in dense prediction which is typically characterized by high sample-redundancy.
- Extensive experiments have shown significant gains of our method over state-of-art approaches for a wide range of labeled data fractions, without incurring extra computation at inference.

II. Related Work

A. Semi-Supervised Learning

For semi-supervised learning, various algorithms have been proposed to exploit the knowledge from unlabeled data, such as VAT [28], Π model [29], mean teacher [30] and dual student [31], etc. Dual Student [31] generates perturbed outputs for the same input via two networks with different initializations and enforces consistency training. Temporal Model [29] enforces consistency between outputs and associated self-ensembles. Mean Teacher [30] yields the target samples via exponential moving average. MixMatch [32] performs entropy minimization on unlabeled data while retains consistency with the help of MixUp, and this method is further extended by ReMixMatch [33] using distribution alignment and augmentation anchoring. In FixMatch [34], consistency regularization is maintained by a pair of weak-strong data augmentation and sample confidence is realized by a simple threshold.

Semi-Supervised Semantic Segmentation: Training models for dense prediction is usually bottlenecked by the demand for sufficient informative pixels, and SSL on semantic segmentation (e.g. [35], [36], [37], [38], [39], [40]) can significantly alleviate this with marginally compromised performance. To exploit the unlabeled data, adversarial learning and consistency training are leveraged for semi-supervised segmentation. AdvSemiSeg [39] utilizes a discriminator to provide additional supervision to unlabeled samples. CCT [35] proposed to perturb outputs of decoders for consistency regularization, while DCC [38] enforces context-aware consistency to mitigate contextual bias. Data mixing methods are used jointly with mean-teacher framework in [40] and [41] showing impressive gains. Self-training based approach [36] selectively re-trains the model by picking more reliable data in priority. PseudoSeg [42] proposes a unique redesign of the pseudo-labeling strategy, generating well-calibrated structured pseudo labels for training with unlabeled data. ECS [37] introduces a secondary network, termed the "corrector," to identify and rectify incorrect predictions made by the primary segmentation network. On top of consistency preservation, our method adopts contrastive learning to enhance the representation compactness at the class level.

B. Contrastive Learning

Contrastive learning has shown its strength in shrinking the gap between supervised and unsupervised learning (e.g., [11], [12], [14], [25], [38], [43], [44]). SimCLR [12] proposed to use data augmentation and nonlinear projection head to learn representations in a contrastive manner. SwAV [14] simultaneously clusters the data while enforcing consistency between cluster assignments produced for different views. Particularly, negative samples play a critical role in contrastive learning. MoCo [11] regularly updates the memory bank with negatives from every mini-batch. A Hard negative mixing approach was presented in [45] to make contrasting more efficient. Instead of direct sampling of negatives, AdCo [20] proposed an adversarial contrastive learning algorithm to learn harder negatives. VLT [46] enhances the robustness of segmentation under the framework of vision-language transformer via masked contrastive learning. Further, for dense prediction tasks, DenseCL [47] and PixPro [48] proposed dense self-supervised learning methods by optimizing a contrastive loss, while SemiContrast [49] and Reco [17] implemented contrastive learning on semi-supervised semantic segmentation with a class-wise memory bank. PC^{2}Seg [50] incorporates pixel-level l_2 loss and pixel contrastive loss to enhance label-space consistency. Different from previous works on contrastive aided semi-supervised learning, we focus on hard negatives and proper data augmentation as two key challenges therein. We further propose adversarial training and design a proper data augmentation to address the two challenges respectively.

In addition, we notice that some non-contrastive methods (e.g. [13], [51]) have recently gained popularity among researchers. BYOL [13] designed two networks to learn from each other and achieved impressive performance without negative samples. SimSiam [51] achieved similar performance with asymmetric structure and a stop-gradient operation. Our work aims to enhance semi-supervised semantic segmentation using contrastive learning, and we defer the integration of non-contrastive methods to future research.

C. Data Augmentation

Data augmentation has been a de facto standard as a preprocessing for computer vision tasks (e.g. [12], [25], [52]). It mitigates the issue of limited data, prevents the model from overfitting and improves the model's generalization capability. Geometric and color-space transformation based augmentation methods are commonly used, such as random cropping, resizing, horizontal flipping and blurring, etc. Many recent works explored the benefit of data mixing based methods. Mixup [27] conducted convex combination of pairs of images and labels; Cutmix [26] replaced pixels in a rectangular area of an image with one from another; Classmix [41] did copy and paste of the whole objects of an image to another one which preserves the object boundaries. Based on 'Copy-and-Paste' operations, our proposed data mixing strategy actively generates new yet challenging samples which force the model to improve on under-performing classes.

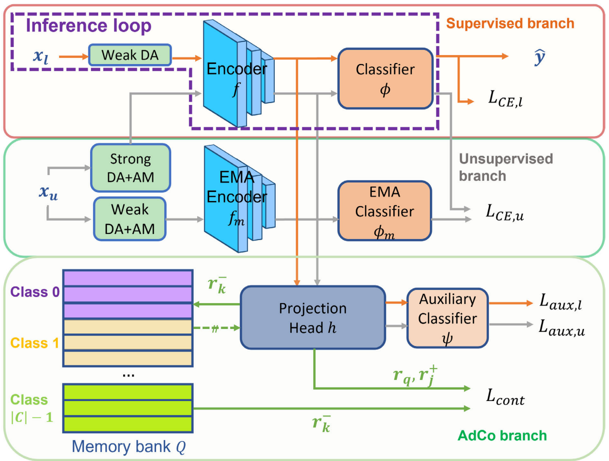


Fig. 1. Our semi-supervised semantic segmentation framework. AM: AdverseMix; DA: data augmentation; AdCo: adversarial contrastive learning. Note that our scheme does not increase computational complexity in the inference loop.

III. The Proposed Method

Suppose that the dataset is composed of a subset of fully annotated data and a subset of unlabeled ones. We denote the labeled subset as $D_l = \{(x_{l,i}, y_{l,i})\}_{i=0}^{N_l}$ where $x_{l,i}$ and $y_{l,i}$ are the i-th image and pixel-level annotation, respectively. Likewise, we denote the unlabeled subset as $D_u = \{(x_{u,j})\}_{j=0}^{N_u}$, where $x_{u,j}$ is an unlabeled image. In practice, $N_l \ll N_u$. Our goal is to construct a framework for semantic segmentation based on D_l and D_u, to take full advantage of both labeled and unlabeled data.

A. Preliminaries

Sohn et al. [34] leveraged consistency and confidence to propose a simple framework for semi-supervised learning which can be adapted to semantic segmentation task.

1) Supervised Branch: For a training batch from labeled data $\{(x_{l,i}), y_i\}_{i=1}^{B}$, weak data augmentation is applied to each data sample $x_{l,i}$. The features are extracted through an encoder $f(\cdot)$ and pixel-wise predictions $\hat{y}_{l,i}$ are generated by a classifier $\phi(\cdot)$. The pixel-wise cross-entropy loss $L_{ce,l}$ is computed based on the predictions and the ground-truth segmentation. As marked in Fig. 1, we denote this part by supervised branch in our model.

2) Unsupervised Branch: As proposed in [34], consistency is kept by the combination of teacher-student model and strong-weak data augmentation on unlabeled data, and confidence is forced by introducing a hard-threshold on model predictions. Within a teacher-student framework, the encoder and the classifier in the supervised branch constitute the student model, and the teacher model is constructed as the exponential-moving-average (EMA) [30] of the student model. For a training batch from unlabeled data $\{(x_{u,i})\}_{i=1}^{B}$, strong augmentations are applied to each data sample $x_{u,i}$ and the student model takes the resultant augmented sample as the input; correspondingly, the teacher model takes the weakly-augmented data sample as the input. Then the teacher model's predictions are made from the EMA encoder $f_m(\cdot)$ and the associated classifier $\phi_m(\cdot)$. A hard-threshold [34] is applied to the model predictions and the resultant one-hot pseudo-labels are used to supervise the student model by training with a cross-entropy loss $L_{ce,u}$. As plotted in Fig. 1, we denote this part by unsupervised branch in our model.

3) Disadvantages: Although solely applying the above framework in a semi-supervised dense prediction setting can enforce model invariance to the non-essential augmentation and enhance consistency, the learned embeddings may fail to improve on contrastive property. For a semantic segmentation model, pixel-wise predictions are possible to be biased towards the class of their neighbors, in which case rare objects and pixels on the object boundary are easy to be mislabeled. Meanwhile, the increase in inter-class separability over the entire dataset can relax over-reliance on neighboring pixels under aforementioned situation and regularize the model such that the learned representations are semantically compact. Hence, it is beneficial to introduce contrastive learning to enforce higher discriminativeness in the intermediate feature space. It is also noticed that natural datasets are imbalanced over classes, and models are prone to being compromised on those tail classes. Sampling in the uniform way can lead to sample insufficiency for those under-performing classes. Without sufficient samples from under-performing classes, the learned features may not well represent corresponding semantics.

B. Contrastive Learning

1) Contrastive Loss: To enhance the discriminativeness of feature embeddings, contrastive learning can be adopted on both supervised- and unsupervised-learning part [53] to force the representations of similar sample pairs to stay close to each other while forcing dissimilar ones to be far apart. Here we use the InfoNCE loss [16] as the contrastive loss L_{cont} for dense prediction as:

$$L_{cont} = \frac{1}{|C|}\sum_{c\in C}\frac{1}{|R_c^q|}\sum_{r_i\in R_q^c}\frac{1}{|R_c^+|}\sum_{r_j^+\sim R_c^+} -\log \\ \times \frac{\exp\left(\mathrm{sim}(r_i, r_j^+)/\tau\right)}{\exp\left(\mathrm{sim}(r_i, r_j^+)/\tau\right) + \sum\limits_{r_k^-\sim R_c^-}\exp\left(\mathrm{sim}(r_i, r_k^-)/\tau\right)}. \tag{1}$$

where R_c^q is the set of pixel level queries/anchors r_i from class c, R_c^+ is the set of positive keys r_j^+, R_c^- is the set of negative keys r_k^-, $\mathrm{sim}(\cdot)$ denotes the cosine similarity, C is the set of all classes in a dataset, $|\cdot|$ is the set cardinality, and τ is the temperature.

In each iteration, the gradients of L_{cont} are computed and propagated back w.r.t. the queries. We note that the gradients for vectors engaged in cosine similarity are as follows,

$$\begin{aligned}\frac{\partial \mathrm{sim}(u, v)}{\partial v} &= \frac{\partial}{\partial v}\frac{u\cdot v}{||u||||v||} = \frac{u}{||u||||v||} - \frac{\mathrm{sim}(u, v)v}{||v||^2}\\ &= \frac{1}{||v||}(\bar{u} - \mathrm{sim}(u, v)\bar{v}),\end{aligned} \tag{2}$$

where $||\cdot||$ is the L2 norm and $\bar{\boldsymbol{r}}$ denotes the L2-normalization of r. Eq. 2 can be geometrically interpreted as the normalized new information vector provided by $\bar{v}$ to $\bar{u}$, which is perpendicular to the direction of u.

Following the above derivative, the gradients over the queries can be derived as

$$\begin{aligned}\frac{\partial L_{cont}}{\partial r_i} &= \frac{1}{\tau}\frac{1}{||r_i||}\frac{1}{|R_c^+|}\sum_{r_j^+\sim R_c^+}\\ &\times\Bigg[-(1 - p(r_j^+|r_i))\left(\bar{r}_j^+ - \mathrm{sim}(r_i, r_j^+)\bar{r}_i\right)\\ &\quad + \sum_{r_k^-\sim R_c^-} p(r_k^-|r_i)\left(\bar{r}_k^- - \mathrm{sim}(r_i, r_k^-)\bar{r}_i\right)\Bigg],\end{aligned} \tag{3}$$

where $p(r_j^+|r_i)$ and $p(r_k^-|r_i)$ represent the probability of the sample r_j^+ and r_k^- being positive to r_i, respectively, that is,

$$\begin{aligned}p(r_j^+|r_i) &:= \frac{\exp\left(\mathrm{sim}(r_i, r_j^+)/\tau\right)}{\exp\left(\mathrm{sim}(r_i, r_j^+)/\tau\right) + \sum\limits_{r_k^-\sim R_c^-}\exp\left(\mathrm{sim}(r_i, r_k^-)/\tau\right)}\\ p(r_k^-|r_i) &:= \frac{\exp\left(\mathrm{sim}(r_i, r_k^-)/\tau\right)}{\exp\left(\mathrm{sim}(r_i, r_j^+)/\tau\right) + \sum\limits_{r_k^-\sim R_c^-}\exp\left(\mathrm{sim}(r_i, r_k^-)/\tau\right)}.\end{aligned} \tag{4}$$

And queries are updated with a learning rate η_θ:

$$r_i \leftarrow r_i - \eta_\theta\frac{\partial L_{cont}}{\partial r_i}.$$

Eq. 3 suggests that queries will be updated towards the weighted combination of the positive keys, while being pushed away from the weighted combination of negatives. The more likely a positive sample r_j^+ is true with a larger $p(r_j^+|r_i)$, the less the query should be pulled towards it as the query is already sufficiently close to the sample. On the contrary, the more likely a negative sample r_k^- is false with a larger $p(r_k^-|r_i)$, the query is closer to the sample, and thus should be more strongly pushed away from it.

2) Structure: Following the structures in [12] for contrastive learning, on top of the supervised branch and unsupervised branch, features are extracted by a projection head $h(\cdot)$ from the output of $f(\cdot)$ for both labeled and unlabeled data, i.e. $r \in \{h(f(x_{l,i})), h(f(x_{u,j}))\}$. As in [43], a memory bank Q can be constructed to store negative samples and is updated by a small portion in each iteration, while queries and positive keys are sampled from the current minibatch. The contrastive loss is then computed through discriminating negative keys from their positive counterparts.

3) Weakness: However, under the circumstance of dense prediction, there are two key ordeals for the success of contrastive learning: **(i)** mining hard negative samples and **(ii)** designing proper data augmentation.

Hard negatives are those samples with the corresponding class different from anchor class but have representations close to anchor representation in the feature space. Mining hard negatives benefits contrastive learning by improving discrimination on pixels from their counterparts whose categories are easy to be mislabeled. There are two main methods to mine hard negative samples: using negatives in the current minibatch (e.g. [12]) and maintaining a memory bank (e.g. [11], as described before). Nevertheless, as pointed in the work of AdCo [20], the former approach abandons rich information embedded in negative samples in past batches and usually a large batch size is required, which increases the memory usage; the latter approach has a small portion of negatives updated in each iteration, as a consequence the most critical negative samples may not be well covered by the memory bank and the learned features are under-representative. This issue is exacerbated in dense prediction due to the imbalance over classes in natural datasets. On one hand, insufficient samples from rarer objects, which are usually from tail class, may bias the semantic segmentation towards the class of dominant neighboring pixels; on the other hand, without actively drawing most representative ones, sample redundancy from head class may impair the efficiency of the algorithm.

Meanwhile, proper data augmentation is crucial for contrastive learning based training. It benefits the learning of representative and generalizable embeddings by introducing non-essential variations while retaining essential semantic features. Hence, the resultant model can focus on the features critical for dense prediction. This calls for the design of a specific data augmentation scheme for semantic segmentation.

Overall, the lack of hard negative samples and proper data augmentations could hinder efficient representation learning for dense prediction. To solve the issue of hard negative samples, we propose **contrastive learning with adversarial training**. We also put forward a new data mixing based augmentation scheme, termed **AdverseMix**, specialized for semantic segmentation task.

C. Adversarial Training

1) Adversarial Contrastive Learning: As an extended work of AdCo [20] which was designed for unsupervised learning on image classification, we adopt adversarial training on dense feature maps to directly learn hard negatives, providing additional discriminative information to learn more expressive dense representations that can distinguish positive queries from these adversarial negatives. This part is illustrated in Fig. 2.

First, the memory bank is initialized by employing negative keys, which are sampled from pixel-level representations of images within the training set. These representations are produced by the projection head, denoted as $h(\cdot)$. Once the memory bank is fully loaded, the negative keys in the memory bank are directly updated through an end-to-end learning process. As illustrated in [20], this is equivalent to solving a minimax problem such that both the encoder parameters θ_f and memory bank Q reach an equilibrium w.r.t. the loss L:

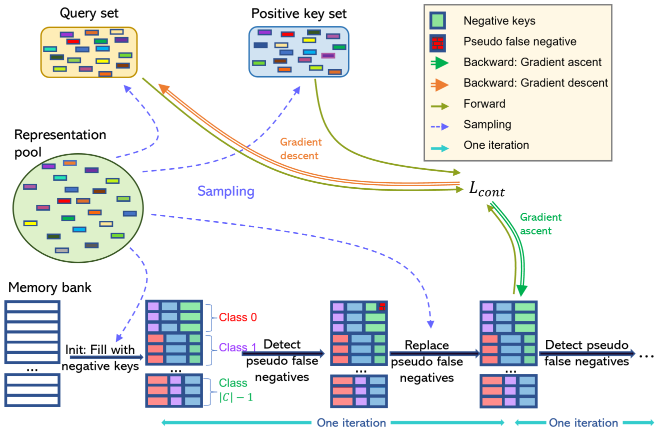


Fig. 2. Illustration of adversarial contrastive learning with sampling and false negative replacement.

$$\theta_f^*, Q^* = \arg\min_{\theta_f}\max_{Q} L(\theta_f, Q)$$

Then, on top of updating encoder weights through the gradients of contrastive loss w.r.t. the queries as in Eq. 3, we calculate the gradients of negative samples in the memory bank using $p(r_k^-|r_i)$ (cf. Eq. 4),

$$\frac{\partial L_{cont}}{\partial r_k^-} = \frac{1}{\tau}\frac{1}{|R_c^+|}\sum_{r_j^+\in R_c^+}\frac{p(r_k^-|r_i)}{||r_k^-||}\left(\bar{r_i} - \mathrm{sim}(r_i, r_k^-)\bar{r}_k^-\right)$$

and following the minimax rule, the negative keys get updated given a learning rate $\eta_{\mathcal{N}}$:

$$r_k^- \leftarrow r_k^- + \eta_{\mathcal{N}}\frac{\partial L_{cont}}{\partial r_k^-}.$$

This implies that the memory bank is updated towards the weighted combination of the queries, where the weights are proportional to the probability of the negatives being false. This update process makes it more difficult for the model to accurately distinguish the queries from learned negative samples, forcing the model to improve its discrimination capability.

Concern of False Negatives: In spite of its benefits, adversarial contrastive learning could introduce an additional risk of learning false negatives. As interpreted above, hard negatives are updated towards the anchor along the direction perpendicular to the negative keys. Over-adversaries could happen if the update pushes the negative too close to the query and the resultant vector becomes positive to the query in reality, which is viewed as a false negative. Those false negatives participate in contrastive learning in the next iteration and could mislead the model as a consequence. This issue is exacerbated in semi-supervised dense prediction since in semantic segmentation, there are abundant pixels from correlated classes (semantically coupled pairs such as bikes and persons, spatially close pairs such as terrain and vegetation) which usually have proximate representations in feature space. There is a higher chance for negatives from those classes to be pushed too close to their counterparts by aggressive updates, and hence, the encoder is compromised. To mitigate this issue, we build an auxiliary classifier for detection and propose to replace those detected false negatives.

2) False Negatives and Auxiliary Classifier: The reason for building the auxiliary classifier is threefold. First, during adversarial training, some negatives may become over-adversarial as they get too close to positive queries. These over-adversarial negatives would become false negatives as they cross the class boundaries. This problem could be exacerbated particularly in dense prediction tasks as it is easier to have more false negatives at pixel level.

Second, the main classifier ϕ works on the feature map produced by the encoder, where pixel predictions often rely on convolutional filters with receptive field covering the neighboring pixels. However, the learned negatives in the memory bank are independent of each other, i.e., it is not necessary to suppose they are from the same real image. Thus it is impossible to directly apply the pixel-level classifier ϕ to predict their (pseudo-)labels.

Third, for the dense prediction, we need such an auxiliary classifier to predict auxiliary-labels to detect and replace the false negatives in the memory bank.

To address these issues, we propose a simple yet effective auxiliary classifier ψ consisting of 1×1 convolutional networks without involving neighboring pixels. During adversarial training, the auxiliary classifier keeps track of the auxiliary-labels of negative adversaries and detects r_k^- as a 'pseudo'-false negative w.r.t. its anchor r_i if $\arg\max \psi(r_k^-) = \arg\max \psi(r_i)$. We keep the initial classes of such 'pseudo'-false negatives, sample a new set from these classes with high confidence scores over the current mini-batch, and replace these 'pseudo'-false ones with reliable ones. In this way, the 'pseudo'-false negative replacement also keeps the same number of negatives for each query class in the memory bank.

For training the auxiliary classifier, we select most uncertain samples measured by the difference between top-2 predicted class probabilities, similar to [54]. The auxiliary classifier is trained on both labeled and unlabeled features with a cross-entropy loss. The associated losses are denoted by $L_{aux,l}$ and $L_{aux,u}$ respectively, and overall auxiliary loss is computed as $L_{aux} = L_{aux,l} + L_{aux,u}$.

3) Informative Sampling: As aforementioned about the concerns in semi-supervised framework (cf. III-A), uniform sampling may lead to learning insufficiency on under-performing classes. Meanwhile, hard negatives can be mined not only during adversarial training, but also at memory bank initialization, which requires to sample informative negatives. Hence, we actively sample pixel-level representations at three separate stages: query and positive key sampling, memory bank initialization, and 'pseudo'-false negatives replacement.

Queries and Positives Sampling. At each iteration, we sample queries and positive keys involved in contrastive loss. To remove data redundancy, we utilize a pair of low and high thresholds for selecting highly confident samples and removing non-informative over-confident ones, respectively.

Memory Bank Initialization. In order to reflect semantic relation in the memory bank initialization, we assign the number of sampled negatives from class c_j w.r.t. query class $c_i|_{i \neq j}$ proportionally to the inter-class similarity between (c_i, c_j). The inter-class similarity is calculated based on class-level representation prototypes. Once the memory bank is fully loaded with negatives of high confidence, each sampled query from class c_i will be contrasted with a group of these negatives.

'Pseudo'-False Negatives Replacement. To maintain the same relationship between semantic classes, we replace 'pseudo'-false negatives by new reliable ones from their initial classes. As illustrated in Fig. 2, this step is executed instantly after gradient updates and false negative detection, with new negatives filled in at the same positions of replaced ones.

The overall algorithm for contrastive learning with adversarial training is summarized in Algorithm 1 and demonstrated in Fig. 2. We denote this part by 'AdCo branch' in Fig. 1. We also remark that, during inference or evaluation stage, the data will only be processed within the inference loop in Fig. 1.

Algorithm 1 Contrastive Learning With Adversarial Training

```
input Class-wise memory bank $Q := [q_{c,i}]$, main classifier $\phi$,
    auxiliary classifier $\psi$, encoder $f$, projection head $h$
Initialize $Q$: (see details in Sec. III-C3)
    Sample neg. keys of high confidence to fill $Q$
for each mini-batch do
    for each class $c$ do                                    } False
        for each neg. key $q_{c,i}$ in class $c$ do          } negative
            Update auxiliary-label for $q_{c,i}$             } replace-
            if $\arg\max \psi(q_{c,i}) == c$ then:           } ment
                Sample a new neg. to replace $q_{c,i}$       }
    Sample queries $\{r_i\}$ and positives $\{r_j^+\}$ to calculate $L_{cont}$
    Optimize $f$ by $\min L_{ce,l} + \lambda_u L_{ce,u} + \lambda_c L_{cont} + \lambda_a L_{aux}$;
    optimize $\phi$ by $\min L_{ce,l} + \lambda_u L_{ce,u}$.
    Update $Q$ w.r.t. $L_{cont}$ (cf. Sec. III-C1)
    Optimize $\psi$ and $h$ by $\min L_{aux}$.
```

D. AdverseMix

For dense prediction tasks, the problem of long-tailed classes and their relations in feature space are under-explored in many mixing-based augmentation approaches. As illustrated in Fig. 3, we propose a novel paradigm of data augmentation method termed AdverseMix by selecting hard images with under-performing classes to mix with those that tend to be confused with the former in the feature space. In other words, the AdverseMix will generate mixed images containing masked regions of both *under-performing classes* and their *easy-to-confuse counterparts*. This will push the classifier to distinguish the under-performing classes from their most-confusing counterparts, thereby focusing on the most adversarially hard cases in training the model.

Specifically, we denote the set of label classes contained in a labeled image $x_{l,i}$ by $\mathcal{A}_{l,i}$ and the set of pseudo-label classes

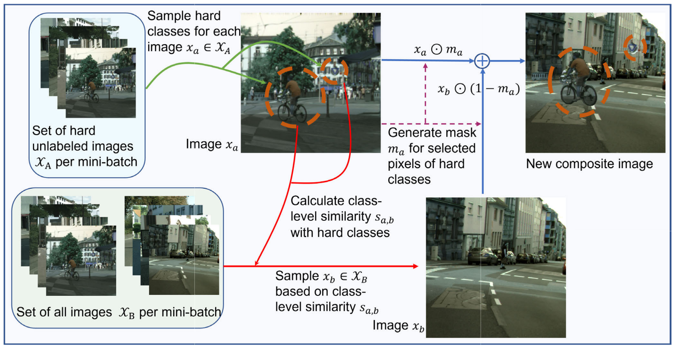

Sample hard classes for each image $x_a \in \mathcal{X}_A$
Set of hard unlabeled images $\mathcal{X}_A$ per mini-batch
Image x_a
$x_a \odot m_a$
$x_b \odot (1 - m_a)$
Generate mask m_a for selected pixels of hard classes
New composite image
Calculate class-level similarity $s_{a,b}$ with hard classes
Set of all images $\mathcal{X}_B$ per mini-batch
Sample $x_b \in \mathcal{X}_B$ based on class-level similarity $s_{a,b}$
Image x_b

Fig. 3. The diagram of AdverseMix. Note that the selected hard classes – rider, bicycle, and traffic sign – from image x_a have a high correlation to instances in image x_b.

contained in an unlabeled image $x_{u,i}$ by $\mathcal{A}_{u,i}$, with $|\cdot|$ the set cardinality. The most recent IoU value of class c evaluated in the training set is denoted by IoU_c.

First, we select foreground images with under-performing classes from unlabeled data. Based on the IoU value, we define a score to measure the difficulty level for each unlabeled image $x_{u,i}$, i.e. $\alpha_i := \frac{1}{|\mathcal{A}_{u,i}|} \sum_{c \in \mathcal{A}_{u,i}} \frac{1}{\text{IoU}_c}$. We select hardest images with top-n_{cp} α-values and denote them by $\{x_{u_h}\}$. We generate another view x'_{u_h} of x_{u_h} to augment the selected hard images from under-performing classes. By excluding n_{cp} easiest images by x_{u_e}, we construct the set of hard unlabeled images over the current mini-batch by $\mathcal{X}_A = \{x_u\} \cup \{x'_{u_h}\} \backslash \{x_{u_e}\}$ as the hard foreground image set. For each image x_a from $\mathcal{X}_A$, half of its classes are sampled based on their difficulty level $\frac{1}{\text{IoU}_c}$, $c \in \mathcal{A}_a$, yielding the set of hard classes $\mathcal{A}_{a_s}$, and their pixel mask m_a.

Second, we mix these hard foreground images with the images that contain the easy-to-confuse classes in the feature space. We view the set of all images in the current mini-batch as the background image set and denote it by $\mathcal{X}_B = \mathcal{X}_A \cup \{x_{u_e}\} \cup \{x_l\}$. For each pair of images $(x_a, x_b) \in \mathcal{X}_A \times \mathcal{X}_B|_{x_a \neq x_b}$, we evaluate their class-level similarity $s_{a,b}$ by averaging the cosine similarities between the set of classes $\mathcal{A}_{a_s}$ and $\mathcal{A}_b$ from x_a and x_b.

$$s_{a,b} = \frac{1}{|\mathcal{A}_{a_s}||\mathcal{A}_b|} \sum_{c_i \in \mathcal{A}_{a_s}, c_j \in \mathcal{A}_b} \text{sim}(c_i, c_j), \tag{5}$$

where $\mathcal{A}_b$ is the set of classes from x_b. In practice, we use class-level representation prototype of each class for this calculation. Then a background image x_b is sampled proportionally to the scores, and the mixed image is generated by 'copy-and-paste' operation: $x_{\text{new}} = x_a \odot m_a + x_b \odot (1 - m_a)$. A diagram of AdverseMix is illustrated in Fig. 3. It shows that AdverseMix selects pixels from harder classes, e.g. the bicycle, the rider and traffic signs in image x_a, and the paired one x_b has classes most relevant to those selected classes, e.g., the street background in this example. The mixed images provide new yet challenging contents which can be leveraged by contrastive learning. Thus the model's generalization capability is further improved.

In addition, we remark that during 'copy-and-paste', there is a chance that pixels from x_a cover all pixels from certain class in x_b. This may cause the resultant composite images to be less close in semantics if the covered class in x_b is the key class correlated to selected classes in x_a. Hence, we modify $\mathcal{A}_b$ in Eq. 5 to $\mathcal{A}_{b_s} := \{c_j | c_j \in y_b \odot (1 - m_a)\}$ given the mask m_a associated with selected hard classes from x_a. To this end, $\mathcal{A}_{b_s}$ excludes x_b's classes whose associated instances are fully covered by m_a. Furthermore, we notice that the majority of images in datasets like Pascal VOC [55], contain single or very few objects, while other datasets like Cityscapes [56], consist of images comprising a larger number of instances. For the former type of datasets, pasted pixels from x_a sometimes greatly change the shape of the contextual object in image x_b due to significant overlapping. The resultant composite image may bias the learning process towards unnatural patterns. Less strong performance on Pascal VOC than on Cityscapes has been reported by ClassMix [41]. We simply shrink the selected objects from x_a and copy them to one of the four quadrants in x_b where least non-background pixels from x_b will be covered. In this way, the new image will alleviate the excessive information loss from x_b.

Algorithm 2 AdverseMix

```
input (in current mini-batch) unlabeled images $\{x_u\}$, pseudo labels $\{\tilde{y}_u\}$, the set of pseudo-label classes $\mathcal{A}_{u,i}$ for $x_{u,i}$; labeled images $\{x_l\}$, ground-truth labels $\{y_l\}$; per class $\{\text{IoU}_c\}$; class-level similarity $\{\text{sim}(c_i, c_j)\}$ for each class pair $(c_i, c_j)$
For each $x_{u,i} \in \{x_u\}$, calculate difficulty score $\alpha_i = \frac{1}{|\mathcal{A}_{u,i}|} \sum_{c \in \mathcal{A}_{u,i}} \frac{1}{\text{IoU}_c}$
Select $n_{cp}$ hardest images $\{x_{u_h}\}$ and $n_{cp}$ easiest images $\{x_{u_e}\}$ according to $\{\alpha_i\}$
For each $x_{u_h}$, generate its augmented view $x'_{u_h}$
Construct the set of harder unlabeled images $\mathcal{X}_A = \{x_u\} \cup \{x'_{u_h}\} \backslash \{x_{u_e}\}$, and the set of associated pseudo labels $\mathcal{Y}_A = \{\tilde{y}_u\} \cup \{\tilde{y}'_{u_h}\} \backslash \{\tilde{y}_{u_e}\}$; construct the set of all images $\mathcal{X}_B = \mathcal{X}_A \cup \{x_{u_e}\} \cup \{x_l\}$, and the set of associated ground-truth/pseudo labels $\mathcal{Y}_B = \mathcal{Y}_A \cup \{\tilde{y}_{u_e}\} \cup \{y_l\}$
for each $x_a \in \mathcal{X}_A$ do
    Sample hard classes of $x_a$, and generate mask $m_a$.
    Calculate pairing score $\{s_{a,b}\}$ for each $x_b \in \mathcal{X}_B$
    Sample $x_b \in \mathcal{X}_B$ w.r.t. pairing score $\{s_{a,b}\}$ ▷ sampled $x_b$ is more semantically related to $x_a$
    $x_{\text{new}} = x_a \odot m_a + x_b \odot (1 - m_a)$.
    $y_{\text{new}} = \tilde{y}_a \odot m_a + y_b \odot (1 - m_a)$ ▷$y_b \in \mathcal{Y}_B$ is pixel-wise pseudo-label (or ground-truth label) w.r.t. $x_b$
```

The algorithm of AdverseMix is summarized in Algorithm 2. We present an ablation study of the effect of the number of augmented views n_{cp} on the performance in the experiment.

IV. Experimental Results

A. Implementation Details

1) Datasets: We conduct experiments on Pascal VOC 2012 [55] and Cityscapes [56] datasets. Pascal VOC contains 21 classes. Its training set comprises 10582 images, with 1464 finely annotated and the rest coarsely annotated.

TABLE I

SEMANTIC SEGMENTATION mIoU FOR PASCAL VOC *val* DATASET. "[]": NUMBER OF LABELED IMAGES; "**v2**": DEEPLABV2; "**v3+**": DEEPLABV3+; †: USING MULTI-SCALE AND HORIZONTAL FLIPPING DURING INFERENCE

Scheme	Method	Backbone	0.57%[60]	1/16[662]	1/8[1323]	1/4[2645]	Full[10582]
Sup only	v3+	R-50	40.49	63.90	68.30	71.20	76.30
ECS [37]	v3+	R-50	–	–	70.2	72.6	76.3
SemiContrast [49]	v3+	R-50	–	–	71.8	–	75.9
DCC [38]	v3+	R-50	–	70.1	72.4	74.0	76.5
ST† [36]	v3+	R-50	–	71.6	73.3	75.0	–
ST++ † [36]	v3+	R-50	–	72.6	74.4	75.4	–
Ours	v3+	R-50	64.19	72.41	74.03	75.90	77.24
Ours†	v3+	R-50	<u>65.80</u>	<u>74.14</u>	<u>75.61</u>	<u>77.10</u>	<u>79.16</u>
ClassMix [41]	v2	R-101	–	–	71.00	–	74.13
SemiContrast [49]	v2	R-101	–	–	71.6	–	74.1
ReCo [17]	v2	R-101	–	–	71.00	–	74.36
GCT [61]	v2	R-101	–	67.19	72.14	73.62	75.73
Sup only	v3+	R-101	44.00	66.40	71.0	73.31	77.79
ReCo [17]	v3+	R-101	53.31	–	74.62	–	77.75
DCC [38]	v3+	R-101	–	72.4	74.6	76.3	78.2
ST† [36]	v3+	R-101	–	72.9	75.7	76.4	–
ST++ † [36]	v3+	R-101	–	74.5	76.3	76.6	–
Ours	v3+	R-101	63.47	73.75	76.91	77.72	79.62
Ours†	v3+	R-101	<u>65.34</u>	<u>74.99</u>	<u>78.14</u>	<u>79.13</u>	<u>81.14</u>

It has 1449/456 images for validation and testing. Cityscapes consists of 19 classes of finely annotated images with high resolution. It contains 2975/500/1525 training/validation/testing images, respectively.

2) Network Structure: We adopt DeepLabv3+ [57] as our baseline model with ResNet-50/101 [58] pretrained on ImageNet [59] as the backbone. The encoder and projection head both have 256 output channels. The model is evaluated on *val* set with single-scale as well as multi-scale inference with horizontal flipping [60]. We adopt the mean Intersection-over-Union (mIoU) as the evaluation metric.

At unsupervised branch, we implement random resizing, cropping and horizontal flipping for weak data augmentation; on top of that, Gaussian blurring is additionally used for strong augmentation. At supervised branch, we utilize random resizing, cropping, horizontal flipping and color jittering as the default augmentation. Images are randomly cropped to size 320×320 for Pascal VOC and size 720×720 for Cityscapes. The batch size is 8 for Cityscapes and 16 for Pascal VOC. We run experiments on 2 Nvidia Tesla V100 cards, except for Cityscapes with ResNet-101 backbone where 2 Nvidia A100 cards are used. Synchronous BatchNorm is adopted for multi-GPU training. We set initial learning rate η_0 as 0.02 for Pascal VOC and 0.05 for Cityscapes. The learning rate of the backbone is 10 times smaller. We adopt a polynomial learning rate schedule with a decay factor 0.9. We use SGD optimizer with momentum 0.9, adopt weight decay as 0.0001 and set the number of training epochs as 300. We first train the encoder f for the first 5 epochs by optimizing cross-entropy loss $L_{ce,l}$ alone (supervised branch in Fig. 1), then train both supervised and unsupervised branches jointly for the next 10 epochs. After that, we train all branches jointly to the end. We use adversarial training with 'pseudo'-false negative replacement and AdverseMix with $n_{cp} = 1$ by default.

B. Main Results

We extensively compare our method with recent approaches and underline the best result in tables. Table I compares different semi-supervised methods for various rates of labeled data from Pascal VOC: 0.57%, 1/16, 1/8 and 1/4 and full set. We also compare with supervised-only baseline. Both single-scale and multi-scale inference results are provided. For works adopting memory bank based contrastive learning, our method outperforms SemiContrast [49] and Reco [17] across all partition rates, indicating the effectiveness of adversarial training of negatives and AdverseMix. Compared with Reco, our method with ResNet-101 backbone achieves 2.29% gain at 1/8 split. Particularly, our method achieves impressively large gain at smaller partition rates (10.1% gain at 0.57% split). Our method also surpasses ST++ [36], a self-training based method, by 2.5% mIoU at 1/4 split benchmark with ResNet-101 backbone. Interestingly, we observe that by using only 1/4 labeled data, our method can achieve comparable performance with supervised training on the whole dataset. If having access to the whole fully annotated data, our method can achieve additional gains with contrastive learning, which is consistent with [53].

Comparison is also made on different splits of labeled images for Cityscapes: 1/30, 1/16, 1/8, 1/4 and full set. In Table II, all methods are based on ResNet-50 backbone. We observe that our method outperforms DCC [38] by 5.9% and PC^{2}Seg [50] by 3.5% at 1/8 split benchmark, respectively. In addition, for both datasets, our method consistently achieves large gains over different partition rates. Especially, at the regime of extreme low number of labeled images, the proposed method promotes the supervised-only baseline with a larger margin. With 1/16 labeled data on Pascal VOC, our method with ResNet-50 backbone boosts mIoU by 8.5% (63.9% vs 72.4%) compared with supervised-only baseline. Similarly on Cityscapes, with 1/30 labeled data, our method provides a significant gain of 14% (55.3% vs 69.3%) over supervised-only baseline. In Table III we deliver the comparison results on Cityscapes using ResNet-101 backbone. We note some recent published works (e.g., [62], [63]) on this topic. We don't include them since they utilized extra deep stem network structure, OHEM auxiliary loss, and larger crop

TABLE II

Semantic Segmentation mIoU for Cityscapes *val* Dataset. "**v3+0**": DeepLabv3+; †: Using Multi-Scale and Horizontal Flipping During Inference; ‡: With Data Selection

Scheme	Method	Backbone	1/30[100]	1/16[186]	1/8[372]	1/4[744]	Full[2975]
Sup only	v3+	R-50	55.25	65.36	68.06	72.86	78.42
ECS [37]	v3+	R-50	–	–	67.4	70.7	74.8
SemiContrast [49]	v3+	R-50	64.9	–	70.0	71.6	74.2
DCC [38]	v3+	R-50	–	–	69.7	72.7	77.5
PC^{2}Seg [50]	v3+	R-50	60.37	–	72.11	73.80	75.39
ST† [36]	v3+	R-50	60.9	–	71.6	73.4	–
ST++† [36]	v3+	R-50	61.4	–	72.7	73.8	–
Ours	v3+	R-50	69.32	74.00	75.64	76.02	78.98
Ours†	v3+	R-50	70.44	75.41	77.07	77.84	80.16

TABLE III

Semantic Segmentation mIoU (%) for Cityscapes *val* Dataset. "X-65": Xception-65 [57]; "R-101": ResNet-101 [58]; "**v2**": DeepLabv2 [65]; "**v3**": DeepLabv3 [60]; "**v3+**": DeepLabv3+ [57]; †: Using Multi-Scale and Horizontal Flipping During Inference; ‡: With Data Selection

Scheme	Method	Backbone	1/30[100]	1/16[186]	1/8[372]	1/4[744]	Full[2975]
AdvSemSeg [39]	v2	R-101	–	–	58.80	62.30	66.40
S4GAN [66]	v2	R-101	–	–	59.30	61.90	65.80
CutMix [40]	v2	R-101	51.20	–	60.34	63.87	67.68
ClassMix [41]	v2	R-101	54.07	–	61.35	63.63	66.19
DMT [67]	v2	R-101	54.81	–	63.03	–	68.16
SemiContrast [49]	v2	R-101	59.4	–	64.4	–	67.3
C^3-SemiSeg [68]	v2	R-101	55.17	–	63.23	65.50	69.53
ReCo [17]	v2	R-101	56.53	–	64.94	67.53	68.60
DSBN-based [64]	v2	R-101	–	–	67.6	69.3	70.1
Sup only	v3+	R-101	59.39	67.45	72.12	73.92	78.77
DepthMix [69]	v3	R-101	58.40	–	66.66	68.43	71.16
DepthMix‡ [69]	v3	R-101	62.09	–	68.01	69.38	–
CutMix [40]	v3+	R-101	55.71	–	65.82	68.33	–
ECS [37]	v3+	R-101	–	–	67.38	70.70	74.76
ReCo [17]	v3+	R-101	60.28	–	66.44	68.50	71.45
PseudoSeg [42]	v3+	R-101	60.96	–	69.81	72.36	–
PC^{2}Seg [50]	v3+	R-101	62.89	–	72.29	75.15	75.99
DSBN-based [64]	v3+	X-65	–	–	74.1	77.8	78.7
Ours	v3+	R-101	71.05	75.82	77.20	78.76	80.22
Ours†	v3+	R-101	72.44	77.54	78.32	80.27	81.65

size of images on Cityscapes dataset. At low partition rate regime, our method has a gain over PC^{2}Seg [50] by 7.7% for 1/30 split, which shows resembling performance on aforementioned dataset. When data partition rate is 1/8, we achieve a gain of 2.9% mIoU over BSBN-based method [64] which used Xception-65 [57] as backbone. Our method even improves by 0.5% over supervised baseline when the full dataset is given.

TABLE IV

Ablation of Each Component of the Proposed Method for Semi-Supervised Semantic Segmentation. "NDM": No Data Mixing; "CM": ClassMix; "AM": AdverseMix; "AdCo": Adversarial Contrastive Learning; "City": Cityscapes; "VOC": Pascal VOC

(a) Performance on Cityscapes 1/16 split with backbone ResNet-50 and Pascal VOC 1/16 split with backbone ResNet-101.

Data	Sup	Semi (NDM)	Semi (CM)	Semi (AM + AdCo)
City	65.36	71.27	72.10	74.00
VOC	66.40	70.10	70.65	73.75

(b) Study on adversarial contrastive learning and AdverseMix modules. Results are obtained on 1/4 Pascal VOC with backbone ResNet-101 and 1/30 Cityscapes with backbone ResNet-50.

Data	Sup	Semi(AM)	Semi(AdCo)	Semi(AM+AdCo)
VOC	73.31	76.59 (+3.28)	76.67 (+3.36)	77.72 (+4.41)
City	55.25	67.71 (+12.46)	68.38 (+13.13)	69.32 (+14.07)

C. Ablation Study

1) Impact of Adversarial Contrastive Learning and AdverseMix: In this part we illustrate the effectiveness of the proposed AdCo branch and AdverseMix. As showcased in Table IV(a), we compare our proposed method with three baselines. The first one only optimizes the supervised loss $L_{ce,l}$ (denoted by *Sup*); the second one optimizes $L_{ce,l} + \lambda_u L_{ce,u}$ except that there is no any data mixing method at unsupervised branch (denoted by *Semi (NDM)*); the third one adopts ClassMix [41] as data mixing on top of the second baseline (denoted by *Semi (CM)*). As presented in Table IV(a), our method outperforms the *Semi (NDM)* baseline by 2.73% and 3.65% on 1/16 Cityscapes and 1/16 Pascal VOC respectively; the gains are 1.90% and 3.10% over the *Semi (CM)* baseline. This demonstrates the performance improvement brought by the modules of AdCo branch and AdverseMix jointly.

Furthermore, we check the individual contribution by the AdCo branch and AdverseMix separately. As shown in Table IV(b), besides the full proposed method *Semi(AM+ AdCo)*, we also consider removing either the AdCo branch (denoted by *Semi (AM)*) or the AdverseMix data mixing

TABLE V

ABLATION STUDY ON SOME HYPERPARAMETER SELECTION

(a) Study on the number of negative keys from memory bank assigned to per query. Results are based on Pascal VOC val *dataset with split 1/4, with backbone ResNet-101.*

# of neg	64	256	512
mIoU	76.84	77.24 (+0.40)	<u>77.72</u> (+0.88)

(b) Ablation study on false adversarial negatives with tracking by auxiliary classifier (aux-cls) and without.

	w/o aux-cls		w/ aux-cls
# total neg.	detection rate	flipped neg. rate	flipped neg. rate
155648	98.20%	3.38%	0.07%

(c) Study on the number of crops in AdverseMix. Results are based on Pascal VOC val *1/4 split with backbone ResNet-101.*

# of crops	0	1	4
mIoU	77.07	<u>77.72</u> (+0.65)	76.82 (-0.25)

module (denoted by *Semi (Adv)*). The experiment is on 1/4 Pascal VOC and 1/30 Cityscapes. The comparison between *Semi(AM)* and *Semi(AM+AdCo)* shows that the implementation of adversarial contrastive learning branch on top of an SSL setting with AdverseMix (*Semi(AM)*) provides an mIoU gain by 1.13% on 1/4 Pascal VOC and 1.61% on 1/8 Cityscapes, respectively. Similarly, the comparison between *Semi(AdCo)* and *Semi(AM+AdCo)* presents that, for 1/4 Pascal VOC and 1/8 Cityscapes, 1.05% and 0.94% improvement on mIoU for can be achieved by implementing AdverseMix on top of a data-mixing-free SSL setting with adversarial contrastive learning branch. The joint use of AdverseMix and the AdCo branch enhances the supervised-only baseline with a gain of 4.41% on 1/4 Pascal VOC, and at the low split rate regime for Cityscapes, the two modules together provide a gain of 14.07% over supervised-only baseline. This indicates that, both adversarial contrastive learning and AdverseMix effectively benefit model training, and the combination of the two results in graceful synergy between each other, which also works better than the nontrivial *Semi (CM)* baseline. Since in [20], adversarial contrastive learning has already been shown to outperform pure memory bank based methods such as MoCo [11] and MoCo v2 [43], we omit further comparison between our method and other memory bank based methods in this work.

2) Memory Bank Size: Table V(a) provides a comparison of different number of negatives from the memory bank assigned to per query in the AdCo branch. The experiment is on Pascal VOC of 1/4 split rate and with ResNet-101 backbone. As the table indicates, increasing the number of negatives steadily improves performance, which is consistent with previous research on self-supervised learning [11], [20]. Due to resource limits, we use memory bank size 512 by default.

3) Auxiliary Classifier: In Table V(b), we show the effectiveness of auxiliary classifiers in tracking ‘pseudo’-false negative keys, by comparing our method to the baseline without kicking-out ‘pseudo’-false negatives. We aim to calculate a detection rate which measures the percentage of false negatives captured by auxiliary classifier being true false negatives. In practice, the experiment is on Cityscapes 1/8 data partition, where a memory bank with a total of 155,648 negatives is employed. In each current mini-batch, we collect the ground-truth label for each representation output by the project head. For the baseline method which does not take care of over-adversaries, we use the trained auxiliary classifier to detect potential ‘pseudo’-false negatives. We estimate the ground-truth label for each detected ‘pseudo’-false negative by searching its *nearest neighbor among representations* from the current mini-batch and using the label of the nearest neighbor as the estimated ground-truth label. We collect verified ‘pseudo’-false negatives as those determined to have the same class as their positive counterparts by nearest neighbor estimation. Thus the detection rate is calculated as the ratio of verified ‘pseudo’-false negatives over all tracked ‘pseudo’-false negatives. The high detection rate shows that auxiliary classifier can provide high accuracy on distinguishing ‘pseudo-false ones for learned negatives in the memory bank. We also compare the rate of detected ‘pseudo’-false negatives (denoted by *flipped neg. rate* in Table V(b)) for the case with and without auxiliary classifier after gradient updates on adversarial negatives during training. As Table V(b) indicates, with ‘pseudo’-false negative replacement, the over-adversaries are mitigated by a large portion.

4) Number of Crops in AdverseMix: As mentioned in Sec. III-D we generate another view(s) of the hardest image(s) to bootstrap the challenging patterns in the composite images. The impact of the number of crops n_{cp} is presented in Table V(c). The comparison between one-crop case and no-crop case indicates that augmenting another view of hardest image (cf. Sec. III-D) is beneficial for learning underperforming patterns, while the comparison between one-crop case and four-crop case implies that too many crops may compromise the model performance. We hypothesize that with excessive augmented crops, more images are deemed as ‘easier’ ones which are factually nontrivial. Hence, features from ‘easier’ images are not engaged in the new composite images and the model cannot efficiently leverage information from those images. The model performance is thus undermined. By default, we set n_{cp} as 1. For the semi-supervised baseline, the total number of passes through the EMA encoder and classifier is equivalent to the batch size. AdverseMix with $n_{cp} = 1$ introduces only one extra pass, which slightly increases the computational load.

D. Visualization Results

1) Predicted Labels: In Fig. 4, we plot the predicted segmentation masks of our method and compare them with DCC [38] and the supervised-only baseline. It is observed that our method has better prediction precision than others, particularly on instance pairs such as juxtaposed car and truck, adjacent bicycles and poles. We hypothesize that these spatially and semantically correlated patterns are particularly bootstrapped by AdverseMix; adversarial training also generates challenging negative pairs and forces the model to enhance performance on such adversarially hard patterns through contrastive learning. Similar results are observed on the segmentation predictions for Pascal VOC 1/8 data partition in Fig. 5. The comparison between our method, DCC [38] and

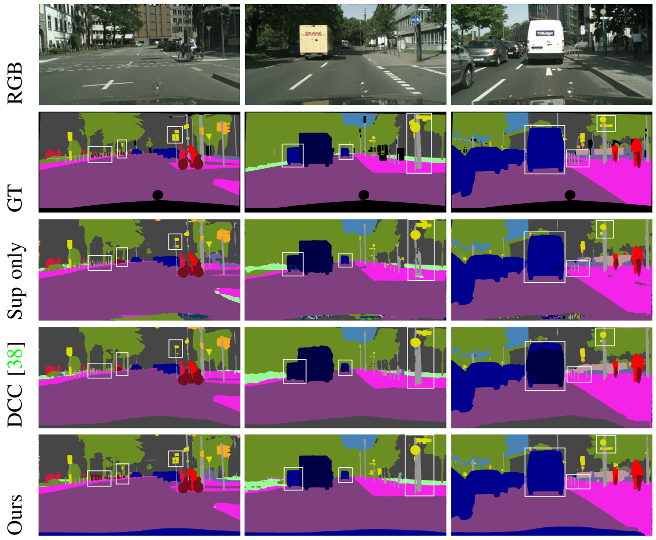


Fig. 4. Cityscapes segmentation examples of different methods. Our method generates better predictions at boxed area, particularly for adjacent bicycles and poles (left boxed area in the first column), juxtaposed car and truck (left boxed area in the middle column), car of unusual shape (left boxed area in the right column).

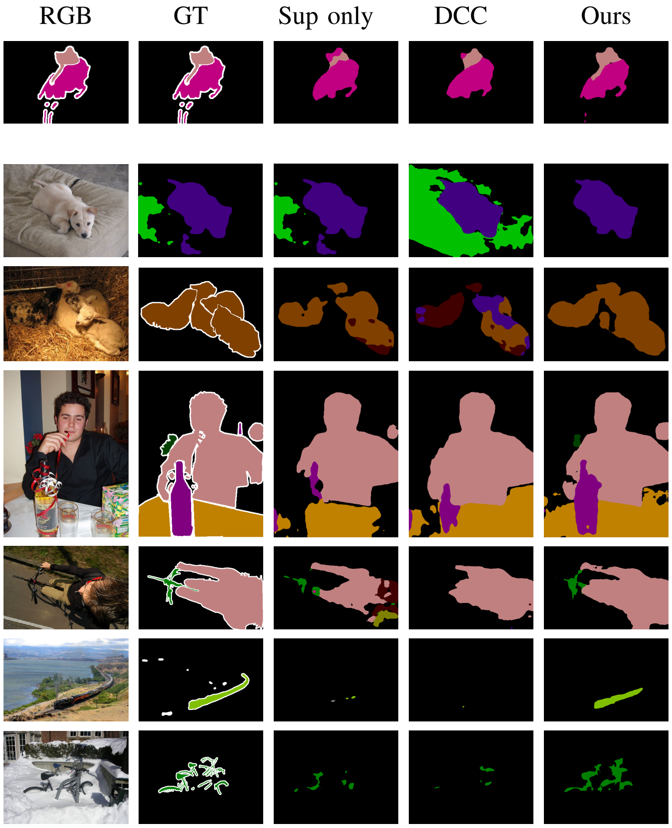
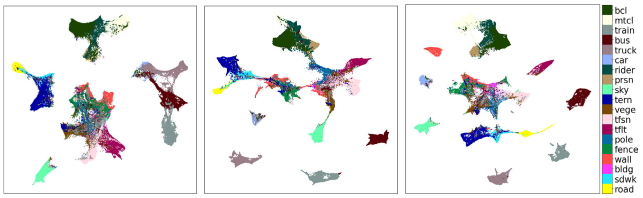


Fig. 5. Pascal VOC segmentation examples of different methods. Our method is compared with supervised-only baseline and DCC [38].

the supervised-only baseline reveals that our method works better on various complicated situations, including composite patterns (row 1: equestrian and horse), cluster of instances (row 3: a flock of sheep), scene shot from an atypical angle (row 5: rider and bicycle), and object of unusual shape (row 6: train rounding the bend).



(a) supervised-only (b) Semi (CM) (c) Our method

Fig. 6. Visualization of features (i.e. the input to the classifier) using t-SNE [70] on Cityscapes with 1/8 split. a) is trained with supervised-only baseline; b) is produced with aforementioned semi-supervised baseline with ClassMix as data mixing strategy; c) is our method with adversarial contrastive learning and AdverseMix. We observe more distinct separation on semantic classes by our method.

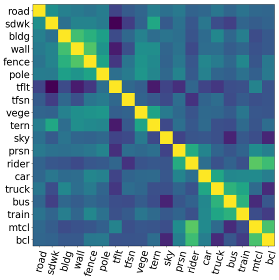


Fig. 7. Correlation heatmap on class-wise representation prototypes for Cityscapes 1/8 partition by our method.

2) Embedding Space: In Fig. 6, we plot the t-SNE visualization of features at the input of the main classifier on Cityscapes 1/8 split, and compare it with aforementioned ClassMix [41] based method. We observe that the features by ClassMix has a higher degree of separation than supervised-only baseline. Compared with ClassMix, our method learns an increased number of completely separated clusters, and generates more compactly-concentrated features. In addition, the corresponding inter-class correlation between class-level prototypes is presented in Fig. 7. The comparison between Fig. 6c and Fig. 7 indicates that the entangled group of semantic classes in t-SNE plot is likely to include those of high correlations in Fig. 7 (class pairs in brighter color), e.g. rider, person, bicycle and motorcycle (semantically correlated) as well as pole, building, wall and fence (spatially adjacent). We hypothesize that because adversarial contrastive learning and AdverseMix work in synergy to generate challenging patterns and force the model to improve on under-performing semantic classes, we can even observe graceful separation on some hard class pairs, such as car vs truck, bus vs train. However, there still remain some classes not well distinguished, e.g. wall vs building, and we speculate that such patterns are beyond the capability of the model at 1/8 data partition.

V. Conclusion

In this work we propose a joint design of adversarial training and AdverseMix for application of contrastive learning to semi-supervised semantic segmentation. For the purpose of mining hard negative samples, we impose direct learning of negative adversaries in contrastive learning. Considering the false negative issue which is critical for dense prediction, we employ an auxiliary classifier to instantly identify the over-adversarial negatives, which are timely tracked and replaced by new reliable samples. We also present an informative sampling approach for memory bank initialization and 'pseudo'-false negative replacement, leading to more efficient adversarial training. To address the challenge of proper data augmentation for dense prediction and enable efficient extraction of informative features from under-performing classes, we propose AdverseMix that generates more diverse yet challenging samples. Extensive experiments demonstrate the strength of our scheme for a wide range of labeled fractions.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[4] J. Donahue, P. Krähenbuhl, and T. Darrell, "Adversarial feature learning," 2016, *arXiv:1605.09782*.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[7] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with GANs: Manifold invariance with improved inference," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[8] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.

[9] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 367, 2005, pp. 281–296.

[10] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3239–3250.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[13] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.

[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.

[15] X. Chen and K. He, "Exploring simple Siamese representation learning," 2020, *arXiv:2011.10566*.

[16] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[17] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," 2021, *arXiv:2104.04465*.

[18] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[19] Z. Jiang, T. Chen, T. Chen, and Z. Wang, "Robust pre-training by adversarial contrastive learning," in *Proc. NIPS*, 2020, pp. 16199–16210.

[20] Q. Hu, X. Wang, W. Hu, and G.-J. Qi, "AdCo: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1074–1083.

[21] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "VideoMoCo: Contrastive video representation learning with temporally adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11200–11209.

[22] M. Kim, J. Tack, and S. J. Hwang, "Adversarial self-supervised contrastive learning," 2020, *arXiv:2006.07589*.

[23] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra, "Can contrastive learning avoid shortcut solutions?" 2021, *arXiv:2106.11230*.

[24] A. Joey Bose, H. Ling, and Y. Cao, "Adversarial contrastive estimation," 2018, *arXiv:1805.03642*.

[25] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" 2020, *arXiv:2005.10243*.

[26] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[28] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[29] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*.

[30] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017, *arXiv:1703.01780*.

[31] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6727–6735.

[32] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," 2019, *arXiv:1905.02249*.

[33] D. Berthelot et al., "ReMixMatch: Semi-supervised learning with distribution alignment and augmentation anchoring," 2019, *arXiv:1911.09785*.

[34] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," 2020, *arXiv:2001.07685*.

[35] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12671–12681.

[36] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make self-training work better for semi-supervised semantic segmentation," 2021, *arXiv:2106.05095*.

[37] R. Mendel, L. A. de Souza, D. Rauber, J. P. Papa, and C. Palm, "Semi-supervised segmentation based on error-correcting supervision," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 141–157.

[38] X. Lai et al., "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1205–1214.

[39] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018, *arXiv:1802.07934*.

[40] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," 2019, *arXiv:1906.01916*.

[41] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1368–1377.

[42] Y. Zou et al., "PseudoSeg: Designing pseudo labels for semantic segmentation," 2020, *arXiv:2010.09713*.

[43] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[44] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 776–794.

[45] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," 2020, *arXiv:2010.01028*.

[46] H. Ding, C. Liu, S. Wang, and X. Jiang, "VLT: Vision-language transformer and query generation for referring segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7900–7916, Jun. 2023.

[47] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3023–3032.

[48] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16679–16688.

[49] 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, *arXiv:2104.13415*.

[50] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7253–7262.

[51] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[52] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3008–3017.

[53] P. Khosla et al., "Supervised contrastive learning," 2020, *arXiv:2004.11362*.

[54] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9796–9805.

[55] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[56] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[57] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[60] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[61] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 429–445.

[62] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[63] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," 2021, *arXiv:2110.05474*.

[64] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, "A simple baseline for semi-supervised semantic segmentation with strong data augmentation," 2021, *arXiv:2104.07256*.

[65] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," 2016, *arXiv:1606.00915*.

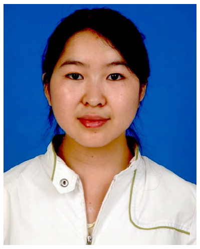
[66] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2021.

[67] Z. Feng et al., "DMT: Dynamic mutual training for semi-supervised learning," 2020, *arXiv:2004.08514*.

[68] Y. Zhou, H. Xu, W. Zhang, B. Gao, and P.-A. Heng, "C3-SemiSeg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7016–7025.

[69] L. Hoyer, D. Dai, Y. Chen, A. Köring, S. Saha, and L. Van Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," 2020, *arXiv:2012.10782*.

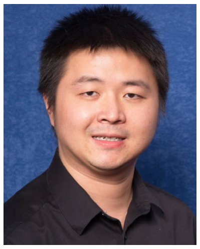
[70] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.



Ying Wang received the Ph.D. degree in electrical and computer engineering from Texas A&M University in 2017. Before that, she was a Research Member with the Chinese Academy of Sciences, from 2010 to 2012. From August 2017 to January 2021, she was a Senior Machine Learning Engineer with the Qualcomm AI Research Center. In 2021, she was a Senior Staff Machine Learning Researcher with the OPPO U.S. Research Center. Since 2022, she has been with Amazon, as a Senior Applied Scientist. She has published her works in mainstream journals and conferences including, *JSAC*, CVPR, ECCV, NeurIPS, and ISIT. She has over 30 patents in the field of 5G wireless communications and machine learning. Her research interests include machine learning for computer vision, auto ML, semi-supervised learning, and 5G wireless communications. She acted as a TPC Member of IEEE INFOCOM in 2018 and 2019. She has been a reviewer for a series of journals and conferences, such as IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and CVPR.

Ziwei Xuan (Member, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Texas A&M University, TX, USA. His current research interests include computer vision with deep learning and machine learning aided wireless communications.

Chiuman Ho received the Ph.D. degree from the University of Pittsburgh. He has completed the postdoctoral research with UC Berkeley. He is the Senior Director of AI with the OPPO U.S. Research Center. Before leaving the academia, he was a Research Assistant Professor with Michigan State University. He is highly interested in AI research and its commercialization. He also helps to develop the AI strategies for OPPO and leads a team to apply AI to improve user experience. He works on deep learning, generative adversarial networks, and reinforcement learning. He is broadly interested in applying these techniques to computer vision and natural language processing. He also works to accelerate and compress deep learning models.



Guo-Jun Qi (Fellow, IEEE) is the Chief Scientist, and has been leading and overseeing the International Research and Development Team for Multiple Artificial Intelligent Services, Huawei Cloud, since August 2018. He has been a Faculty Member with the Department of Computer Science and the Director of the MAchine Perception and LEarning (MAPLE) Laboratory, University of Central Florida, since August 2014. Prior to that, he was also a Research Staff Member with the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. He has published more than 100 papers in a broad range of venues in pattern recognition, machine learning, and computer vision. His research interests include machine learning and knowledge discovery from multi-modal data sources to build smart and reliable information and decision-making systems. He has served/serves as the General Co-Chair for ICME 2021; the Technical Program Co-Chair for ACM Multimedia 2020, ICIMCS 2018, and MMM 2016; and the area chair (a senior program committee member) for multiple academic conferences. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition* (PR), and *ACM Transactions on Knowledge Discovery from Data* (T-KDD).