C^3-SemiSeg: Contrastive Semi-supervised Segmentation via Cross-set Learning and Dynamic Class-balancing

Yanning Zhou
CUHK

Hang Xu*
Huawei Noah's Ark Lab

Wei Zhang
Huawei Noah's Ark Lab

Bin Gao
Huawei Noah's Ark Lab

Pheng-Ann Heng
CUHK

Abstract

The semi-supervised semantic segmentation methods utilize the unlabeled data to increase the feature discriminative ability to alleviate the burden of the annotated data. However, the dominant consistency learning diagram is limited by a) the misalignment between features from labeled and unlabeled data; b) treating each image and region separately without considering crucial semantic dependencies among classes. In this work, we introduce a novel C^3-SemiSeg to improve consistency-based semi-supervised learning by exploiting better feature alignment under perturbations and enhancing the capability of discriminative feature cross images. Specifically, we first introduce a cross-set region-level data augmentation strategy to reduce the feature discrepancy between labeled data and unlabeled data. Cross-set pixel-wise contrastive learning is further integrated into the pipeline to facilitate feature representation ability. To stabilize training from the noisy label, we propose a dynamic confidence region selection strategy to focus on the high confidence region for loss calculation. We validate the proposed approach on Cityscapes and BDD100K dataset, which significantly outperforms other state-of-the-art semi-supervised semantic segmentation methods.

1. Introduction

Semantic segmentation is a fundamental and challenging problem in the computer vision community and has been studied for the long term. It aims to generate high-resolution pixel-wise categories prediction given an image and can be applied to many applications such as autonomous driving [39, 51, 9] and medical image analysis [36, 53]. Most of the methods enjoy the merit of Convolutional Neural Networks (CNNs) and improve it by designing specific architectures as well as training strategies [26, 36, 5, 43]. However, these data-driven methods depend on the large scale and the high quality of the annotated dataset, which becomes a burden to apply in the real world. Regarding limited annotations, the network cannot discern the various appearance within the category and is easily over-fitted to the restricted samples, which results in error prediction in some confusing categories.

Semi-supervised learning aims to utilize datasets that have labels for only a fraction of their samples [18] by learning representation from both labeled and unlabeled data. The trained network usually has better generalization ability on unseen data than that trained with fully supervised setting. Adding consistency regularization is a common way in semi-supervised learning [23, 38, 13]. It encourages the network to generate similar predictions for the same unlabeled image with different augmentation by calculating the difference between outputs as the loss function. Nevertheless, previous methods only facilitate the intra-image feature consistency inside the unlabeled dataset. Although both labeled data and unlabeled data are sampled $i.i.d.$ from the same data distribution, it is observed the empirical distribution of labeled data often deviates from the true samples distribution [44], which further leads to the misalignment in the feature space [27] and even hurts the performance [31]. Therefore, reducing the feature misalignment and enhancing feature discriminative ability is crucial for semi-supervised pixel-level recognition.

In this work, we introduce a novel C^3-SemiSeg to alleviate those constraints of consistency-based semi-supervised methods by exploiting better feature alignment under perturbations, and enhancing discriminative of the inter-class features cross images from both labeled and unlabeled set.

Specifically, we adopt the mean teacher networks [38]

*Corresponding Author: xbjxh@live.com

into our framework, where each model contains a shared CNN encoder followed by a segmentation head and a projection head in parallel during training. To fully enjoy the merit of consistency regularisation, we propose the asymmetric data augmentation strategy with the cross-set region-level data mixing that feeds the strong augmented data into the student to match the prediction of weak augmented data from the teacher network. The data mixing method can further narrow the feature misalignment between labeled and unlabeled data throughout cross-set fusion.

Meanwhile, a pixel-wise contrastive loss is added on both labeled and unlabeled features to simultaneously promote the embedding to be close to that from the same category while being far away from different categories. The intuition is to enforce the feature compactness within the class and increase the discriminative across classes, similar to [45], but the target scope is different. [45] only conducts contrastive learning within labeled data, while our method leverages both labeled and unlabeled data. Therefore, our method could not only enhance feature discriminative, but also reduce feature misalignment between two sets, and extend the hard negative sampling space. Furthermore, to reduce the negative effect brought by noisy predictions, we proposed the Dynamic Confident Region Selection (DCRS) to preserve class-balanced samples adaptively with high confidence for network optimization at each step.

We conduct experiments on Cityscapes and BDD100K datasets with different proportions of labeled data to demonstrate the effectiveness of our proposed approach in different situations. It even closes the performance gap between $1/4$ labeled data and fully labeled data by 79%. Our contributions can be summarised as follow:

- We propose a novel C^3-SemiSeg framework to improve conventional consistency-based semi-supervised learning by the asymmetric data augmentation with the cross-set region-level data mixing to narrow the feature misalignment between labeled and unlabeled data.

- A pixel-wise contrastive learning loss function is proposed to enhance inter-class feature discrepancy and inter-class feature compactness across the dataset, with the Dynamic Confident Region Selection module to further prevent the misleading from noisy predictions.

- Extensive experiments on two autonomous driving dataset, named Cityscapes and BDD100K demonstrates C^3-SemiSeg outperforms other state-of-the-art methods significantly with all the labeled data ratio.

2. Related Works

Semi-supervised learning. Semi-supervised learning is considered a promising way to reduce the need for expensive annotations. Most methods are designed according to one or several following considerations: (1) Consistency regularisation. Some methods [23, 38] assume that by giving input with different perturbations, the predictions should be consistent. Different data augmentation are added to the same unlabeled data, equipped with the loss function to encourage the predictions to be close to each other. (2) Pseudo-labeling. These methods [24, 42] gives pseudo-label to unlabeled data through the network pre-trained from labeled data. Then, they retrain the network and refine the pseudo-label iteratively. (3) Entropy regularisation. It encourages the network to be confident in the decision making by minimizing the entropy [14].

Semi-supervised Semantic Segmentation. Early works [30, 20] introduce GAN-based framework and adversarial training to encourage the predictions from labeled and unlabeled data to be indistinguishable.

Recently, people investigates the self-training strategy on this task [57, 58, 59, 4, 12, 28]. These methods focus on calibrating the pseudo-label from reliable predictions. Zou *et al.* [57] jointly performed network learning and pseudo-label estimation with the class-specific threshold for class-balanced self-training. [59] further incorporated two types of confidence regularisation to encourage the network output's smoothness. Based on [57], Mei *et al.* [28] proposed an exponential moving average method to generate the threshold of each instance for unsupervised domain adaptation. Compared with the previous method, it is more flexible to adjust the class-aware threshold dynamically. Hence, we adapt [28] into our method. Recently, [59] introduced a calibrated fusion strategy by combining the self-attention Grad-CAM maps with predictions. Nevertheless, to acquire reasonable Grad-CAM results on the dataset where most images have the same image-level label, more aggressive geometric data augmentation is needed.

The self-training method's main disadvantage is that it requires a well-trained teacher model, while this prerequisite is not always held, especially when data is extremely limited. Compared with it, the consistency training method [13, 34, 32] performs better in low-data regime. The main factor to their success is the data augmentation strategy, where either input augmentation [13, 32] or feature perturbation are studied [34]. [13] demonstrated the effectiveness of CutOut [11] and CutMix [54] in this task. [32] used the predicted labels to synthesise new images by mixing half of the semantic class region between two images. [19] further mixed images according to depth information from a depth estimation model trained by extra sequential data. Instead of mixing unlabeled data only [13, 32], we argue that (1) conducting data mixing [54] between labeled and unlabeled data, and (2) using weakly-augmented data for teacher and strong-augmented data for the student is the best choice for semi-supervised semantic segmentation.

Contrastive Learning for Dense Prediction. Recently,

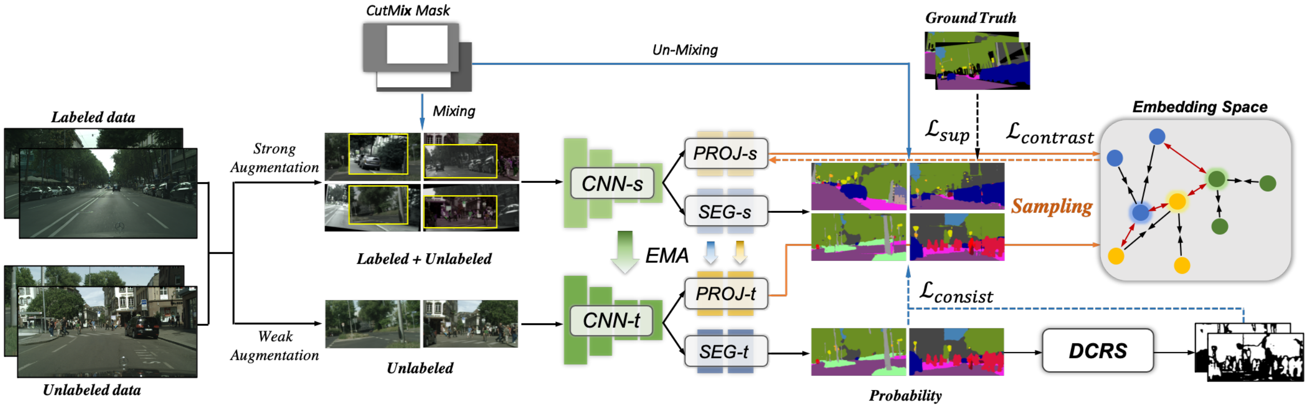

Figure 1. Overview of the proposed C^3-SemiSeg framework. The approach consists of two networks with the same architecture to carry out semi-supervised semantic segmentation. Each network contains a shared CNN feature extractor (CNN-t/ CNN-s) followed by a projection head (PROJ-t/ PROJ-s) and a segmentation head (SEG-t/ SEG-s). We perform the region-level data mixing across strong augmented data from $\mathcal{D}_U$ and $\mathcal{D}_L$, and encourage the consistency predictions between the teacher and the student. Meanwhile, the unit-normalized embedding features are sampled according to its segmentation result for pixel-wise contrastive learning to encourage intra-class compactness and inter-class discriminative. In each forward pass, Dynamic Confident Region Selection (DCRS)strategy is proposed to update the class-balanced confident threshold adaptively and then select the high confident region for loss calculation.

contrastive learning methods raise researchers' attention due to the success on representation learning and other applications [15, 33, 7, 8, 16]. The core idea of these approaches is to pull the embedding of positive samples and push negative samples' embedding in the projection space. Here we focus on the most related literature for dense prediction tasks and refer readers to [25, 21] for other details. Methods of self-supervised pre-training for dense prediction [47, 48, 49, 3] focus on defining novel positive/negative pairs and designing specific learning framework. Recently, Wang *et al.* [46] proposed a pixel-wise metric learning paradigm by exploring labeled pixels' structure using contrastive learning and shows promising results on supervised semantic segmentation. Zhao *et al.* [56] designed a contrastive learning-based training strategy for a semi-supervised setting. However, [56] still needs a contrastive pre-training step before adding unlabeled data. In contrast, we aim to encourage the class-specific embedding to be discriminative and the predictions of augmented inputs to be consistent together throughout end-to-end training to enjoy the merit of the complementary information from both labeled and unlabeled data.

3. Method

3.1. Overview of C^3-SemiSeg.

Following the setting of semi-supervised semantic segmentation, we are provided with a small set of labeled data with pixel-level annotation $\mathcal{D}_L$ and a large set of unlabeled data $\mathcal{D}_U$. Let $\mathcal{B}_L$ and $\mathcal{B}_U$ denotes the labeled and unlabeled data in each batch, a standard pixel-wise cross-entropy loss is applied on the segmentation head for labeled data, similar to previous semi-supervised methods [13]:

$$\mathcal{L}_{sup} = -\frac{1}{|\mathcal{B}_L|}\sum_{x\in\mathcal{B}_L}\frac{1}{M_L}\sum_{i=1}^{M_L} y_i^T \log(f\,(\theta; A\circ x_i)), \quad (1)$$

where θ is the learnable weight of the encoder and the segmentation head (CNN-s and SEG-s in Figure 1), M_L denotes the number of valid pixels in one image, $y_i \in \mathbb{R}^c$ is the one-hot vector label, and $A(\cdot)$ represents the weak augmentation function applied on the labeled image.

For unlabeled data, an unsupervised consistency loss term is applied to encourages consistent predictions in response to one image with different perturbations. To construct prediction pairs, we adopt the mean teacher framework [38]. An exponential moving average weight of the student network is used to update the teacher: $\hat{\theta}_t = \alpha\hat{\theta}_{t-1} + (1-\alpha)\theta_t$, where α is the hyper-parameter that controls the update ratio. Hence, it tends to produce a more accurate model [35]. Let $f(\hat{\theta};\cdot)$ denotes the combination of the encoder and the segmentation head from the teacher network, the unsupervised consistency loss term is formed:

$$\mathcal{L}_{consist} = -\frac{1}{|\mathcal{B}_U|}\sum_{x\in\mathcal{B}_U}\frac{1}{M_U}\sum_{i=1}^{M_U} f\left(\hat{\theta}; A\circ x_i\right)^T \log\left(f\left(\theta; \hat{A}\circ x_i\right)\right), \quad (2)$$

where $\hat{A}(\cdot)$ is the strong augmentation function for the unlabeled data (details in Section 3.2), M_U represents the size of confidence region (details in Section 3.4). Equation 2 can be considered as the cross-entropy loss function that utilises the teacher's prediction as the soft target label.

Additionally, a pixel-wise contrastive loss is calculated on the pixel embedding from the projection head. Let f_i denotes the unit-normalized features for pixel i, $\mathcal{P}_i$ and $\mathcal{N}_i$ denotes the corresponding positive set and negative set. The pixel-wise contrastive loss is formed:

$$\mathcal{L}_{contrast} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{P}_i|}$$
$$\sum_{j\in\mathcal{P}_i}\log\frac{exp\left(f_i^T\cdot f_j/\tau\right)}{exp\left(f_i^T\cdot f_j/\tau\right)+\sum_{k\in\mathcal{N}_i}exp\left(f_i^T\cdot f_k/\tau\right)}, \tag{3}$$

where τ represents the temperature.

3.2. Cross-Set Data Augmentation

Previous semi-supervised methods normally carried data augmentation within the unlabeled set [2, 13, 59]. But when there exists distribution mismatching between labeled and unlabeled set, the feature misalignment would hurt the performance. To reduce this effect and fully enjoy the merit of consistency learning, we propose (1) conducting region-level data mixing across data from D_U and D_L, and (2) using asymmetric data augmentation for two networks.

Region-level data mixing. Data mixing is the augmentation technique by combing two images in pixel-level [55, 40] or region level [54], which encourages the network to attend on less discriminative parts and therefore utilises a broader variety of features. Previous methods suggested applying CutMix to unlabeled data [13]. Though it enriches the diversity of unlabeled samples, it does not deduce the feature misalignment between $\mathcal{D}_L$ and $\mathcal{D}_U$. Especially when there is a large ratio gap between $\mathcal{D}_L$ and $\mathcal{D}_U$, it is possible to exist a large distribution shift between two sets. Recently, [50, 41] illustrates the effectiveness of MixUp for domain mix-up strategy in unsupervised domain adaptation. Therefore, we propose to conducted cross-set data mixing by combining $\mathcal{B}_L$ and $\mathcal{B}_U$ data together for CutMix. Specifically, given two images x_a,$x_b \in \mathcal{B}$, where $\mathcal{B} = \mathcal{B}_U \cup \mathcal{B}_L$, the region-level mixing process is:

$$x_{mix} = m \odot x_a + (1-m) \odot x_b, \tag{4}$$

where m is a binary mask initialized as one with a random rectangle of pixels as zero. $\mathcal{L}_{sup}$ and $\mathcal{L}_{con}$ is applied to the labeled and unlabeled region of x_{mix}.

Asymmetric data augmentation. In terms of classification, previous methods [23, 38, 2] utilised the same weak augmentation for both the teacher and the student's inputs. Recent methods [37, 1] shows better results by applying substantial augmentation for the student while weak augmentation for the teacher. Here, instead of using the same intensity of augmentations for both networks [13], we propose to apply weak augmentation for the teacher to acquire more precious predictions as supervised signals. Then, the strong augmentation with proposed region-level data mixing and RandAugment [10] is applied to the input of the student network.

3.3. Pixel-wise Contrast Learning

Although the consistency regularisation encourages the invariant predictions given an image with small perturbations, it does not consider the cross-image structure information. To further enhance the feature discriminative ability, we propose to take the merit of contrastive learning that pulls the pixel-wise features from the same category and pushes features away from different categories across images. Specifically, features from the shared encoder are feed into the projection head and map into the embedding space. Let f_i denotes the i-th pixel's unit-normalized embedding. Its corresponding positive set $\mathcal{P}_i$ is the pixel embedding with the same category, while the negative set $\mathcal{N}_i$ is that with different categories. To reduce the feature misalignment between labeled and unlabeled set, $\mathcal{P}_i$ and $\mathcal{N}_i$ are constructed by combining features from two sets in the same batch together. By learning from comparison, it facilitates features to be closer to those that belonged to the same class and be discrepant to those from different classes.

The proposed pixel-wise contrastive loss requires the category information to sample positive and negative sets. To enable the usage of features from unlabeled data, we assume that the teacher's predictions are correct in most areas and use its prediction to assign positive and negative pairs. To reduce the negative effect from noisy predictions, we propose the dynamic confidence region selection strategy (Section 3.4) to filter out the uncertain region. Therefore, Equation 3 can be re-written as:

$$\mathcal{L}_{contrast} = -\frac{1}{\hat{M}_L+\hat{M}_U}\sum_{i=1}^{\hat{M}_L+\hat{M}_U}\frac{1}{|\mathcal{P}_i|}\sum_{j\in\mathcal{P}_i}$$
$$\log\frac{exp\left(f_i^T\cdot\hat{f}_j/\tau\right)}{exp\left(f_i^T\cdot\hat{f}_j/\tau\right)+\sum_{k\in\mathcal{N}_i}exp\left(f_i^T\cdot\hat{f}_k/\tau\right)}, \tag{5}$$

where $\hat{f}_j$ and $\hat{f}_k$ are positive and negative embeddings from the teacher network, $\hat{M}_L$ and $\hat{M}_U$ denotes the summation of M_L and M_U in one batch, respectively. We also utilise the Segmentation-Aware Hard Anchor Sampling in [46] to let segmentation results help find informative hard samples. Specifically, for the data from $\mathcal{B}_L$, hard samples are points from the wrong prediction area, and for the data from $\mathcal{B}_U$, it is defined as the inconsistent outputs between two networks.

3.4. Dynamic Confident Region Selection

It is natural for the predicted probability to exist errors. These noisy soft labels would harm the learning process significantly. Therefore, designing a sample selection strategy to filter the noisy label is desirable. [37] proposed to ignore the region with the confidence less than the threshold. [57] and [58] further proposed the class-balanced threshold that estimates an individual threshold for each class to prevent the class domination in the pseudo label. However, a fixed threshold is not suitable for the continuously updated teacher network in our framework. Hence, we proposed to utilise a dynamic class-balanced threshold for region selection. Specifically, for each forward pass, we sort the predictions with a reservation ratio s for the c-th class to find the class-specific threshold $\delta_{t,c}$ for current batch. Yet DCRS does not directly use the current threshold to preserve regions. Instead, it uses the overall threshold which is updated online by averaging the consecutive thresholds from different forward pass via EMA:

$$\hat{\delta}_{t,c} = \beta\hat{\delta}_{t-1,c} + (1-\beta)\delta_{t,c}, \tag{6}$$

where β controls the update ratio, $\hat{\delta}_{t,c}$ denotes the final confidence threshold for the c-th class at t-step, which is smoother by the past threshold information. After updating $\hat{\delta}_{t,c}$, it is used to mask the teacher's predictions in this batch. Regions with confidence less than the threshold are ignored in consistency regularisation and contrastive learning.

Our DCRS is similar to the IAS [28], with two differences: (i) We do not give a tighter sample ratio for the harder classes and (ii) we utilise DCRS during training instead of IAS during the pseudo-labels generation.

3.5. Overall Loss Function

The proposed semi-supervised semantic segmentation can be trained in an end-to-end fashion. The total loss is

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_1\mathcal{L}_{consist} + \lambda_2\mathcal{L}_{contrast}, \tag{7}$$

where λ_1 and λ_2 are hyper-parameters to balance each term's intensity. Note that the projection head will be removed after training. Therefore, it does not add any computation cost at inference time.

4. Experiments

4.1. Experimental Setup

We conduct experiments and report the mean intersection-over-union (mIOU) score on two commonly used dataset, namely Cityscapes and BDD100K.

Cityscapes [9]: This is an autonomous driving dataset captured from 50 cities in the real world. It contains high-quality pixel-level annotations for 19 semantic categories with a fixed resolution of 2048×1024. The training and validation splits contain 2975 and 500, respectively. Following previous works [13, 32, 12], we down-sample the images to 1024×512. We randomly sample $\frac{1}{30}$, $\frac{1}{8}$, and $\frac{1}{4}$ of training data as $\mathcal{D}_L$, and remain others in $\mathcal{D}_U$. Besides, we also evaluate our framework on the full supervised condition, where both $\mathcal{D}_U$ and $\mathcal{D}_L$ contain all samples. Same as [32, 12], we conduct our method over 3 runs.

BDD100K [52]: This is another large-scale autonomous driving dataset. For semantic segmentation task, it has the same label space like that in Cityscapes [9]. The training and validation splits contain 7000 and 1000 images, respectively. Previously, no semi-supervised semantic segmentation method conducted experiments on this dataset. We choose to estimate the performance with the same data ratio as that on Cityscapes ($\frac{1}{30}$, $\frac{1}{8}$, and $\frac{1}{4}$).

Network Structure. Same as previous methods [13, 32], we utilize DeepLab V2 [5], which contains Atrous Spatial Pyramid Pooling (ASPP) module to extract multi-scale representations based on ImageNet pre-trained ResNet-101 [17] in our experiments. Specifically, the ResNet-101 denotes CNN in Figure 1. ASPP and the following classifier are considered as the segmentation head (SEG). The projection head is implemented as a two-layer feed-forward network with a non-linear function between layers: `Conv-ReLU-Conv`, to map the 2048-d features from the backbone into 256-d embedding space.

Implementation Details. Same data augmentation is used in experiments on two datasets. Notably, images from $\mathcal{D}_L$ are only applied weak augmentation, including randomly horizontal flip and randomly rotation within 10°. On the other hand, images from $\mathcal{D}_U$ are conducted weak augmentation and strong augmentation for the teacher and the student. In terms of strong augmentation, it contains the geometric operation mentioned before, colour jittering and the colour transformation from RandAugment [10] (details in the supplementary material). For pixel-wise contrastive loss, we select twenty samples per class in each image to build a positive set and a negative set. Half of them are chosen from the wrong prediction region, which is recommended by [46]. The temperature τ is set to 0.15 in all experiments. For Cityscapes, images from $\mathcal{D}_L$ and $\mathcal{D}_U$ are randomly cropped into 256×512 as inputs. To prevent the model from confusing by initial noisy predictions, it conducts purely supervised learning at the first ten epochs. Then the teacher net is initialised by the weight from the student network and updated by the exponential moving average of the weight from the student in each step with $\alpha = 0.99$. For $L_{consist}$, a sigmoid ramp-up function [23] is used to adapt the intensity at the beginning $\lambda_1 = 50e^{-5(1-curr\_iter/4000)^2}$. We use the Adam [22] optimization algorithm with a learning rate of 0.00012, and adopt the polynomial annealing policy [6] to schedule the learning rate, which is multiplied by $\left(1 - \frac{curr\_iter}{total\_iter}\right)^{0.9}$ at each iteration. The network is trained with a batch size of 32 on 4 GPUs for 25000 iterations. We set the ratio of labeled and unlabeled data in each batch to $1 : 1$. For BDD100K, data is cropped into 512×512 as the input. We add ex-

Labeled samples	**1/30 (100)**		**1/8 (372)**		**1/4 (744)**		**Full (2975)**	
Baseline	-		55.5		59.9		66.4	
Adversarial [20]	-		58.8	(+3.3)	62.3	(+2.4)	-	
Baseline	-		56.2		60.2		66.0	
s4GAN [30]	-		59.3	(+3.1)	61.9	(+1.7)	65.8	(-0.2)
Baseline	-		55.96±0.86		60.54±0.85		-	
ECS [29]	-		60.26±0.84	(+4.30)	63.77±0.65	(+3.23)	-	
Baseline	44.41±1.11		55.25±0.66		60.57±1.13		67.53±0.35	
French *et al.* [13]	51.20±2.29	(+6.79)	60.34±1.24	(+3.30)	63.87±0.71	(+3.30)	67.68±0.37	(+0.15)
Baseline	45.5		56.7		61.1		66.9	
DST-CBC [12]	48.7	(+3.2)	60.5	(+3.8)	64.4	(+3.3)	-	
Baseline	43.84±0.71		54.84±1.14		60.08±0.62		66.19±0.11	
ClassMix [32]	54.07±1.61	(+10.23)	61.35±0.62	(+6.51)	63.63±0.33	(+3.55)	-	
Baseline	44.83±0.38		55.10±0.66		60.20±0.53		66.87±0.06	
Ours (C^3-SemiSeg)	**55.17**±0.86	(**+10.88**)	**63.23**±0.45	(**+8.13**)	**65.50**±1.08	(**+5.30**)	**69.53**±0.21	(**+2.06**)

Table 1. Performance (mIoU) on Cityscapes validation set under different proportions of labeled samples, presented as mean ± std-dev computed from 3 runs. Our proposed C^3-SemiSeg outperforms other methods at each labeled ratio.

Method	1/30 (233)		1/8 (875)		1/4 (1750)	
Baseline	40.4		47.7		52.6	
Ours (C^3-SemiSeg)	**49.1**	+8.7	**52.2**	+4.5	**55.2**	+2.6

Table 2. Performance (mIoU) on BDD100K validation set.

Method	1/30 (100)		1/8 (372)		1/4 (744)	
Baseline	45.0		55.7		60.8	
Ours w/o $\mathcal{L}_{consist}$	48.8	+3.8	60.2	+4.5	64.4	+3.6
Ours w/o $\mathcal{L}_{contrast}$	54.4	+9.4	63.3	+7.6	65.8	+5.0
Ours (C^3-SemiSeg)	**55.0**	**+10.0**	**63.7**	**+8.0**	**66.4**	**+5.6**

Table 3. Performance analysis over different loss components on Cityscapes validation set.

tra random scaling of $0.75, 1.0, 1.25$ into the augmentation strategy mentioned before. We use a batch size of 16 on 4 GPUs to train the network for 20000 iterations. For other hyper-parameters, we set s, β, λ_2 as $0.8, 0.9$, and 0.1 for both datasets, respectively. All the experiments are conducted on Tesla V100 GPUs.

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

Cityscapes. In Table 1 we present our results of mean Intersection over Union (mIoU) on the Cityscapes validation dataset under different proportions of labeled samples. We also show the corresponding baseline at the top of each method, which denotes the purely supervised learning results trained by the the same labeled data. Note that all the methods use the DeepLab V2 [5] for a fair comparison.

Our proposed method not only achieves the highest performance (55.17%, 63.23%, and 65.50%), but also the largest gains ($+10.88\%$, $+8.13\%$, and $+5.30\%$) for the case of $1/30$, $1/8$, and $1/4$ labeled data. When the ratio of labeled data becomes higher (e.g., $1/4$), the performance improvement brought by other semi-supervised learning approaches becomes smaller, especially [13], and [32]. Compared to 3.27% from [32], the gain (5.64%) from our method is significant larger. Moreover, when we adapt our approach to the fully supervised setting by assigning all data, our method can still beat the baseline with 2.06% improvement. The factors that contribute to this include the gain from contrastive learning, which can also be proved in Table 3. When only contrastive learning is conducted (row 2), the network gets consistent improvements in different labeled data ratios. Besides, the self-training based method [12] has limited performance improvement on this task. It may because when labeled data is limited, the noisy pseudo-label confuses the network during the iterative learning process.

BDD100K. To further prove the generalization ability of our method, we conduct experiments on BDD100K, which contains more complicated scenarios with various weather conditions. Table 2 shows the experiment results of mean Intersection over Union (mIoU) on the validation dataset with different proportions of labeled samples. Compared with $40.4\%, 47.7\%, 52.6\%$ mIoU from the supervised baseline, our method get 8.7%, 4.5% and 2.6% performance gains on $1/30, 1/8, 1/4$ labeled data ratio, respectively.

4.3. Ablation Studies

Next, we analyse the contribution of each component in our framework. Without further mentioned, all the experiments are conducted on the Cityscapes dataset using the same training strategy with the first seed from [13].

D	S	1/30 (100)		1/8 (372)		1/4 (744)	
N.A.		54.4		63.3		65.7	
$\mathcal{D}_L$	✓	53.5	-0.9	61.9	-1.4	65.8	+0.1
$\mathcal{D}_U$	✓	54.6	+0.2	63.5	+0.2	**66.4**	**+0.8**
$\mathcal{D}_L + \mathcal{D}_U$		54.6	+0.2	**64.2**	**+0.9**	65.8	+0.1
$\mathcal{D}_L + \mathcal{D}_U$	✓	**55.0**	**+0.6**	63.7	+0.4	**66.4**	**+0.8**

Table 4. Performance analysis over different contrastive learning strategy. N.A.: Without $\mathcal{L}_{contrast}$, D: Source of features used in $\mathcal{L}_{contrast}$, S: Segmentation-aware sampling.

Table 3 shows the performance analysis over different loss components. Compared with supervised baseline, applying contrastive learning improves the performance by 3.8%, 4.5% and 3.6% for $1/30, 1/8, 1/4$ proportions of labeled data. Meanwhile, adding consistency regularization results in an improvement of mIoU from 45.0% to 54.4%, 55.7% to 63.3%, and 60.8% to 65.8%. Combining them leads to the best results, which demonstrates that the complementary information from contrastive learning and consistency regularisation can assist the network to have a better discriminative ability of feature representations.

Effectiveness of each component in contrastive learning.

We first select features from different sources ($\mathcal{D}_L$ and $\mathcal{D}_U$) to apply the proposed pixel-wise contrastive loss. Table 4 shows that applying it only on $\mathcal{D}_L$ does not give improvements (row 2). We believe it is because the network over-fits to the limited number of annotated samples easily. On the other hand, adding contrastive learning on $\mathcal{D}_U$ shows benefits (row 3) on each labeled data ratio, and applying it across $\mathcal{D}_U$ and $\mathcal{D}_L$ is the best choice (row 6), which demonstrates the essential to enhance the intra-class feature compactness and reduce the feature misalignment across labeled and unlabeled data.

Furthermore, we evaluate the effectiveness of the sampling strategy. As suggested in [46], we sample twenty feature points in each class per image. Half of them are hard samples, while others are randomly sampled. Specifically, for the data from $\mathcal{D}_L$, hard samples are points from the wrong prediction area, and for the data from $\mathcal{D}_U$, the wrong prediction is defined as the inconsistent outputs between two networks. Meanwhile, we train another network by randomly sampling twenty feature points in each class for comparison. As shown in Table 4 row 5, the segmentation-aware sampling strategy gives slightly improvements.

Teacher	Student	1/30 (100)		1/8 (372)		1/4 (744)	
Supervised		45.0		55.7		60.8	
S+R	S+R	43.2	-1.8	54.8	-0.9	57.1	-3.7
W	W	50.6	+5.6	59.3	+3.6	62.0	+1.2
W	S	53.9	+8.9	63.2	+7.5	64.8	+4.0
W	S+R	**54.4**	**+9.4**	**63.3**	**+7.6**	**65.8**	**+5.0**

Table 5. Performance analysis over different augmentation strategies. S: Strong augmentation, W: Weak augmentation, R: RandAugment [10].

Mixing	1/30 (100)		1/8 (372)		1/4 (744)	
N.A.	51.1		60.7		63.3	
Intra-set	53.3	+2.2	62.4	+1.7	65.2	+1.9
Cross-set	**54.4**	**+3.3**	**63.3**	**+2.6**	**65.8**	**+2.5**

Table 6. Performance analysis over different data mixing strategy. Intra-set: Perform independent data mixing over $\mathcal{D}_U$ and $\mathcal{D}_L$, Cross-set: Perform data mixing across $\mathcal{D}_U$ and $\mathcal{D}_L$.

Effectiveness of the augmentation strategy for consistency regularization. Consistency regularisation aims to encourage the consistent prediction of the given image with small perturbations. Therefore, it is crucial to design the augmentation strategy carefully. We first conduct experiments of adding the different intensity of augmentation to the teacher and the student. Details for the definition of augmentation can be seen in Supplementary Material.

As can be seen in Table 5, directly applying strong augmentation with RandAugment [10] to both networks leads to the performance drop. It may cause by the significantly increasing of wrong predictions from the teacher network, misleading the network's optimisation direction. When the weak augmentation is added on both networks, it brings 5.6%, 3.6% and 1.2% improvements under $1/30, 1/8, 1/4$ proportions of labeled data. Nevertheless, it does not fully take the potential benefits from consistency regularisation compared with networks equipped with proposed asymmetry data augmentation, which yields 8.9%, 7.5% and 4.0% gains. Furthermore, adding RandAugment [10] let it get extra 0.5%, 0.1% and 1.0% improvements.

We also evaluate the effectiveness of cross-set region-level data mixing. Table 6 shows that applying intra-set region-level data mixing has 2.2%, 1.7% and 1.9% performance gains. In addition, extend the data mixing across both labeled and unlabeled data further improves $1.1\%, 0.9\%$ and 0.6%, which illustrates the proposed cross-set data mixing is a more powerful tool to enhance features alignment for dense prediction tasks.

Effectiveness of the Dynamic Confident Region Selection. To narrow the negative effect brought by the teacher's wrong predictions, we present the DCRS module added after the teacher output. One important hyper-parameter for

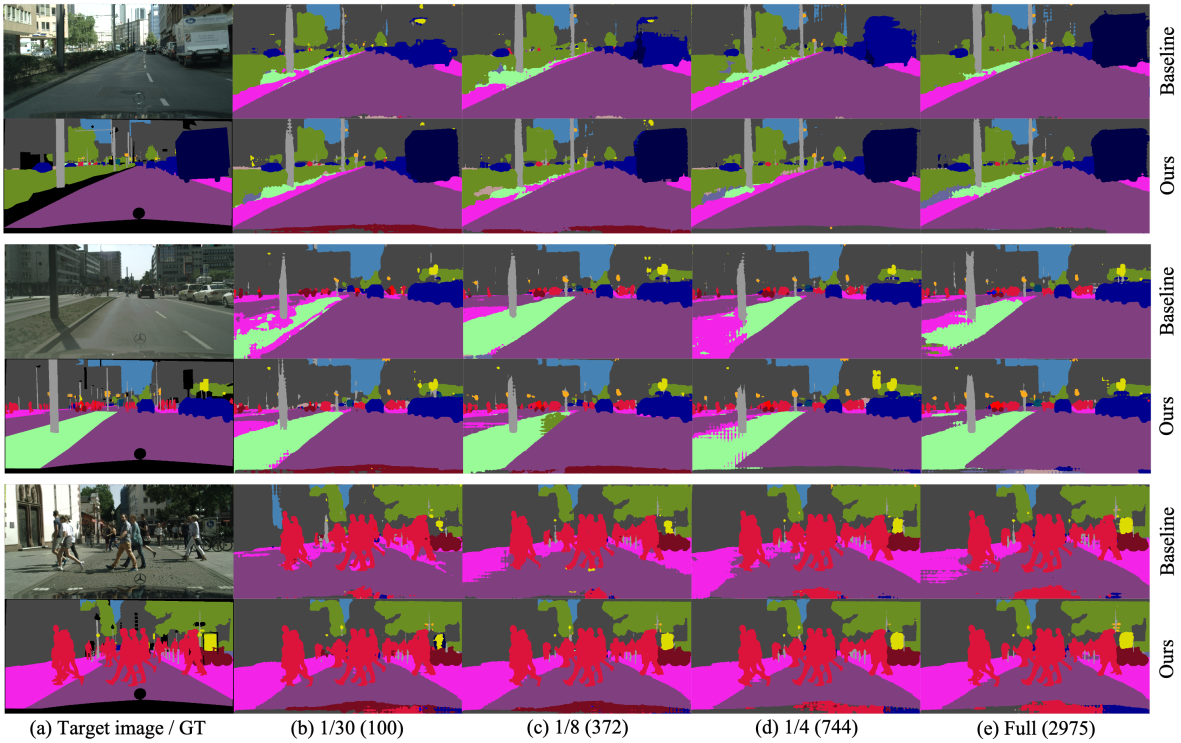

Figure 2. Qualitative results of our method and baseline method on different proportions of labeled images on Cityscapes validation dataset. (a) target images and corresponding ground truth (GT), (b)-(e) segmentation results of different proportions of labeled images.

the success of the DCRS is s, which controls the ratio of preserved region to calculate the loss functions.

Table 7 shows the network performance of mIoU on Cityscapes validation set using $1/8$ proportion of labeled data, in the case of different s. Note that $s = 1.0$ means all the samples are preserved, or in other words, DCRS is removed from the framework. As we can see, setting s too small brings negative effects to the network. This might because it only preserves easy samples with high confidence, which leads to the lost of informative regions. When we increase s to 0.8 to include more regions for loss calculation, the network reaches the highest mIoU of 63.3%. Furthermore, assigning s too large is also problematic and leads to the performance decrease since the wrong prediction always comes from the low confidence region.

s	0.4	0.6	0.8	0.9	1.0
mIoU	61.9	62.9	**63.3**	63.1	62.8

Table 7. Performance analysis over different s in DCRS. Calculated on $1/8$ proportion of labeled data (372).

Quantitative Evaluation. In Figure 2, we further display some qualitative segmentation results of our method and baseline method on various proportions of labeled images. Overall, our method achieves more complete segmentation results than the baseline model in the same split of labeled images, especially for the region with complexity texture and that needs long range feature consistency.

5. Conclusion

We propose a novel end-to-end learning framework for semi-supervised semantic segmentation. The asymmetric data augmentation with cross-set data mixing strategy to enjoys the merit of consistency regularisation. Furthermore, to extend the intra-class feature compactness and the inter-class discriminative ability across all images, we introduce the pixel-wise contrastive learning. DCRS is added to eliminate the negative effects from noisy predictions during loss calculation. Our experiments on two commonly use autonomous driving datasets demonstrate that the proposed framework can fully take advantage of labeled and unlabeled data and achieve superior performance.

Acknowledgments

This work was supported by Hong Kong Research Grants Council with Project No. CUHK 14201620. The authors are also grateful to the Huawei Noah's Ark Lab for funding the collaboration between Huawei and CUHK.

References

[1] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *ICLR*, 2020.

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

[3] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *NeurIPS*, 2020.

[4] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*, 2020.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[10] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshop*, 2020.

[11] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[12] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020.

[13] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020.

[14] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *CAP*, 2005.

[15] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(2), 2012.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[18] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[19] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. *arXiv preprint arXiv:2012.10782*, 2020.

[20] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018.

[21] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[23] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

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

[25] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 2020.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[27] Christoph Mayer, Matthieu Paul, and Radu Timofte. Adversarial feature distribution alignment for semi-supervised learning. *Computer Vision and Image Understanding*, 2021.

[28] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020.

[29] Robert Mendel, Luis Antonio De Souza, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. 2020.

[30] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[31] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018.

[32] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. *arXiv preprint arXiv:2007.07936*, 2020.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[34] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020.

[35] Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 1992.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

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

[39] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*.

[40] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *CVPR*, 2018.

[41] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021.

[42] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowl. Inf. Syst.*, 42(2):245–284, 2015.

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[44] Qin Wang, Wen Li, and Luc Van Gool. Semi-supervised learning by augmented distribution alignment. In *ICCV*, 2019.

[45] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021.

[46] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021.

[47] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *arXiv preprint arXiv:2011.09157*, 2020.

[48] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. *arXiv preprint arXiv:2102.04803*, 2021.

[49] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *arXiv preprint arXiv:2011.10043*, 2020.

[50] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020.

[51] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[52] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.

[53] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*, 2017.

[54] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[55] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[56] Xiangyun Zhao, Raviteja Vemulapalli, Philip Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label-efficient semantic segmentation. *arXiv preprint arXiv:2012.06985*, 2020.

[57] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018.

[58] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019.

[59] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *ICLR*, 2021.