Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision

Xiaokang Chen[1*] Yuhui Yuan[2] Gang Zeng[1] Jingdong Wang[2]
[1]Key Laboratory of Machine Perception (MOE), Peking University [2]Microsoft Research Asia

Abstract

In this paper, we study the semi-supervised semantic segmentation problem via exploring both labeled data and extra unlabeled data. We propose a novel consistency regularization approach, called cross pseudo supervision (CPS). Our approach imposes the consistency on two segmentation networks perturbed with different initialization for the same input image. The pseudo one-hot label map, output from one perturbed segmentation network, is used to supervise the other segmentation network with the standard cross-entropy loss, and vice versa. The CPS consistency has two roles: encourage high similarity between the predictions of two perturbed networks for the same input image, and expand training data by using the unlabeled data with pseudo labels. Experiment results show that our approach achieves the state-of-the-art semi-supervised segmentation performance on Cityscapes and PASCAL VOC 2012. Code is available at https://git.io/CPS.

1. Introduction

Image semantic segmentation is a fundamental recognition task in computer vision. The semantic segmentation training data requires pixel-level manual labeling, which is much more expensive compared to the other vision tasks, such as image classification and object detection. This makes semi-supervised segmentation an important problem to learn segmentation models by using the labeled data as well as the additional unlabeled data.

Consistency regularization is widely studied in semi-supervised semantic segmentation. It enforces the consistency of the predictions with various perturbations, e.g., input perturbation by augmenting input images [11, 19], feature perturbation [27], and network perturbation [18]. Self-training is also studied for semi-supervised segmentation [6, 43, 42, 9, 13, 25]. It incorporates pseudo segmentation maps on the unlabeled images obtained from the segmentation model trained on the labeled images to expand the training data, and retrains the segmentation model.

We present a novel and simple consistency regularization approach with network perturbation, called cross pseudo supervision. The proposed approach feeds the labeled and unlabeled images into two segmentation networks that share the same structure and are initialized differently. The outputs of the two networks on the labeled data are supervised separately by the corresponding ground-truth segmentation map. Our main point lies in the cross pseudo supervision that enforces the consistency between the two segmentation networks. Each segmentation network for an input image estimates a segmentation result, called pseudo segmentation map. The pseudo segmentation map is used as an additional signal to supervise the other segmentation network.

The benefits from the cross pseudo supervision scheme lie in two-fold. On the one hand, like previous consistency regularization, the proposed approach encourages that the predictions across differently initialized networks for the same input image are consistent and that the prediction decision boundary lies in low-density regions. On the other hand, during the later optimization stage the pseudo segmentation becomes stable and more accurate than the result from normal supervised training only on the labeled data. The pseudo labeled data behaves like expanding the training data, thus improving the segmentation network training quality.

Experimental results with various settings on two benchmarks, Cityscapes and PASCAL VOC 2012, show that the proposed cross pseudo supervision approach is superior to existing consistency schemes for semi-supervised segmentation. Our approach achieves the state-of-the-art semi-supervised segmentation performance on both benchmarks.

2. Related work

Semantic segmentation. Modern deep learning methods for semantic segmentation are mostly based on fully-convolutional network (FCN) [23]. The subsequent developments studies the models from three main aspects: resolution, context, and edge. The works on resolution enlargement include mediating the spatial loss caused in the classification network, e.g., using the encoder-decoder scheme [5] or dilated convolutions [36, 4], and maintaining high resolution, such as HRNet [34, 30].

The works on exploiting contexts include spatial context,

*This work was done when Xiaokang Chen was an intern at Microsoft Research, Beijing, P.R. China

e.g., PSPNet [41] and ASPP [4], object context [38, 37], and application of self-attention [33]. Improving the segmentation quality on the edge areas include Gated-SCNN [31], PointRend [20], and SegFix [39]. In this paper, we focus on how to use the unlabeled data, conduct experiments mainly using DeepLabv3+ and also report the results on HRNet.

Semi-supervised semantic segmentation. Manual pixel-level annotations for semantic segmentation is very time-consuming and costly. It is valuable to explore the available unlabeled images to help learn segmentation models.

Consistency regularization is widely studied for semi-supervised segmentation. It enforces the consistency of the predictions/intermediate features with various perturbations. Input perturbation methods [11, 19] augment the input images randomly and impose the consistency constraint between the predictions of augmented images, so that the decision function lies in the low-density region.

Feature perturbation presents a feature perturbation scheme by using multiple decoders and enforces the consistency between the outputs of the decoders [27]. The approach GCT [17] further performs network perturbation by using two segmentation networks with the same structure but initialized differently and enforces the consistency between the predictions of the perturbed networks. Our approach differs from GCT and enforces the consistency by using the pseudo segmentation maps with an additional benefit like expanding the training data.

Other than enforcing the consistency between various perturbations for one image, the GAN-based approach [25] enforce the consistency between the statistical features of the ground-truth segmentation maps for labeled data and the predicted segmentation maps on unlabeled data. The statistical features are extracted from a discriminator network that is learned to distinguish ground-truth segmentation and predicted segmentation.

Self-training, a.k.a., self-learning, self-labeling, or decision-directed learning, is initially developed for using unlabeled data in classification [15, 10, 1, 3, 22]. Recently it is applied for semi-supervised segmentation [6, 43, 42, 9, 13, 25, 14, 24]. It incorporates pseudo segmentation maps on unlabeled data obtained from the segmentation model previously trained on labeled data for retraining the segmentation model. The process can be iterated several times. Various schemes are introduced on how to decide the pseudo segmentation maps. For example, the GAN-based methods [13, 25, 29], use the discriminator learned for distinguishing the predictions and the ground-truth segmentation to select high-confident segmentation predictions on unlabeled images as pseudo segmentation.

PseudoSeg [44], concurrent to our work, also explores pseudo segmentation for semi-supervised segmentation. There are at least two differences from our approach. PseudoSeg follows the FixMatch scheme [28] via using the pseudo segmentation of a weakly-augmented image to supervise the segmentation of a strongly-augmented image based on a single segmentation network. Our approach adopts two same and independently-initialized segmentation networks with the same input image, and uses the pseudo segmentation maps of each network to supervise the other network. On the other hand, our approach performs back propagation on both the two segmentation networks, while PseudoSeg only performs back propagation for the strongly-augmented image.

Semi-supervised classification. Semi-supervised classification was widely studied in the first decade of this century [3]. Most solutions are based on the assumptions, such as smoothness, consistency, low-density, or clustered. Intuitively, neighboring data have a high probability of belonging to the same class, or the decision boundary should lie in low-density regions.

Deep learning methods impose the consistency over perturbed inputs or augmented images encouraging the model to produce the similar output/distributions for the perturbed inputs, such as temporal ensembling [21] and its extension mean teacher [32]. Dual student [18] makes modifications by jointly learning two classification networks that initialized differently with complex consistency on the predictions and different image augmentations.

Other development includes estimating labels for unlabeled data, e.g., MixMatch [2] combining the estimations on multiple augmentations, FixMatch [28] using pseudo labels on weak augmentation to supervise the labeling on strong augmentation.

3. Approach

Given a set $\mathcal{D}^l$ of N labeled images, and a set $\mathcal{D}^u$ of M unlabeled images, the semi-supervised semantic segmentation task aims to learn a segmentation network by exploring both the labeled and unlabeled images.

Cross pseudo supervision. The proposed approach consists of two parallel segmentation networks:

$$\mathsf{P}_1 = f(\mathsf{X}; \theta_1), \tag{1}$$
$$\mathsf{P}_2 = f(\mathsf{X}; \theta_2). \tag{2}$$

The two networks have the same structure and their weights, i.e., θ_1 and θ_2 , are initialized differently. The inputs X are with the same augmentation, and P_1 (P_2) is the segmentation confidence map, which is the network output after softmax normalization. The proposed approach is logically illustrated as below[1]:

$$\begin{aligned} \mathsf{X} \rightarrow \mathsf{X} \rightarrow f(\theta_1) \rightarrow \mathsf{P}_1 \rightarrow \mathsf{Y}_1 \\ \searrow f(\theta_2) \rightarrow \mathsf{P}_2 \rightarrow \mathsf{Y}_2. \end{aligned} \tag{3}$$

Here Y_1 (Y_2) is the predicted one-hot label map, called pseudo segmentation map. At each position i, the label vec-

[1] We use $f(\theta)$ to represent $f(\mathsf{X}; \theta)$ by dropping X for convenience.

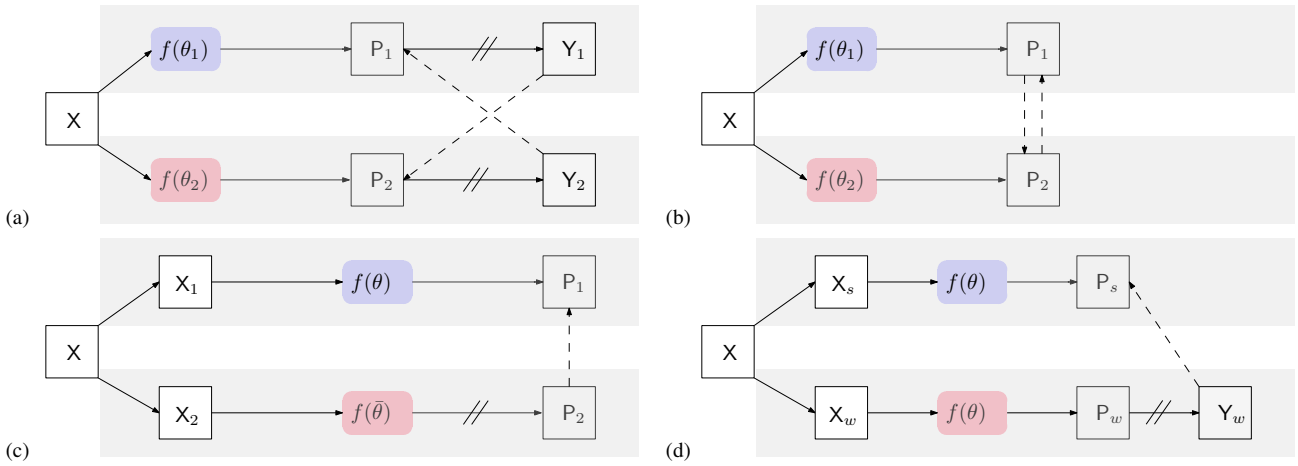

Figure 1: Illustrating the architectures for (a) our approach cross pseudo supervision, (b) cross confidence consistency (e.g., a component of GCT [17]), (c) mean teacher (used in CutMix-Seg [11]), and (d) PseudoSeg [44] structure (similar to FixMatch [28]). '$\rightarrow$' means forward operation and '$\dashrightarrow$' means loss supervision. '//' on '$\rightarrow$' means stop-gradient. More details are illustrated in the approach section.

tor $\mathbf{y}_{1i}$ ($\mathbf{y}_{2i}$) is a one-hot vector computed from the corresponding confidence vector $\mathbf{p}_{1i}$ ($\mathbf{p}_{2i}$). The complete version of our method is illustrated in Figure 1 (a) and we have not included the loss supervision in the above equations.

The training objective contains two losses: supervision loss $\mathcal{L}_s$ and cross pseudo supervision loss $\mathcal{L}_{cps}$. The supervision loss $\mathcal{L}_s$ is formulated using the standard pixel-wise cross-entropy loss on the labeled images over the two parallel segmentation networks:

$$\mathcal{L}_s = \frac{1}{|\mathcal{D}^l|}\sum_{\mathsf{X}\in\mathcal{D}^l}\frac{1}{W\times H}\sum_{i=0}^{W\times H}(\ell_{ce}(\mathbf{p}_{1i},\mathbf{y}^*_{1i}) + \ell_{ce}(\mathbf{p}_{2i},\mathbf{y}^*_{2i})), \quad (4)$$

where ℓ_{ce} is the cross-entropy loss function and $\mathbf{y}^*_{1i}(\mathbf{y}^*_{2i})$ is the ground truth. W and H represent the width and height of the input image.

The cross pseudo supervision loss is bidirectional: One is from $f(\theta_1)$ to $f(\theta_2)$. We use the pixel-wise one-hot label map Y_1 output from one network $f(\theta_1)$ to supervise the pixel-wise confidence map P_2 of the other network $f(\theta_2)$, and the other one is from $f(\theta_2)$ to $f(\theta_1)$. The cross pseudo supervision loss on the unlabeled data is written as

$$\mathcal{L}^u_{cps} = \frac{1}{|\mathcal{D}^u|}\sum_{\mathsf{X}\in\mathcal{D}^u}\frac{1}{W\times H}\sum_{i=0}^{W\times H}(\ell_{ce}(\mathbf{p}_{1i},\mathbf{y}_{2i}) + \ell_{ce}(\mathbf{p}_{2i},\mathbf{y}_{1i})). \quad (5)$$

We also define the cross pseudo supervision loss $\mathcal{L}^l_{cps}$ on the labeled data in the same way. The whole cross pseudo supervision loss is the combination of the losses on both the labeled and unlabeled data: $\mathcal{L}_{cps} = \mathcal{L}^l_{cps} + \mathcal{L}^u_{cps}$.

The whole training objective is written as:

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_{cps}, \quad (6)$$

where λ is the trade-off weight.

Incorporation with the CutMix augmentation. The CutMix augmentation scheme [40] is applied to the mean teacher framework for semi-supervised segmentation [11]. We also apply the CutMix augmentation in our approach. We input the CutMixed image into the two networks $f(\theta_1)$ and $f(\theta_2)$. We use the way similar to [11] to generate pseudo segmentation maps from the two networks: input two source images (that are used to generate the CutMix images) into each segmentation network and mix the two pseudo segmentation maps as the supervision of the other segmentation network.

4. Discussions

We discuss the relations of our method with several related works as following.

Cross probability consistency. An optional consistency across the two perturbed networks is cross probability consistency: the probability vectors (from pixel-wise confidence maps) should be similar (illustrated in Figure 1 (b)). The loss function is written as:

$$\mathcal{L}_{cpc} = \frac{1}{|\mathcal{D}|}\sum_{\mathsf{X}\in\mathcal{D}}\frac{1}{W\times H}\sum_{i=0}^{W\times H}(\ell_2(\mathbf{p}_{1i},\mathbf{p}_{2i}) + \ell_2(\mathbf{p}_{2i},\mathbf{p}_{1i})). \quad (7)$$

Here an example loss $\ell_2(\mathbf{p}_{1i},\mathbf{p}_{2i}) = \|\mathbf{p}_{1i}-\mathbf{p}_{2i}\|_2^2$ is used to impose the consistency. Other losses, such as KL-divergence, and the consistency over the intermediate features can also be used. We use $\mathcal{D}$ to represent the union of labeled set $\mathcal{D}^l$ and unlabeled set $\mathcal{D}^u$.

Similar to the feature/probability consistency, the proposed cross pseudo supervision consistency also expects the consistency between the two perturbed segmentation

networks. In particular, our approach in some sense augments the training data by exploring the unlabeled data with pseudo labels. The empirical results shown in Table 4 indicates that cross pseudo supervision outperforms cross probability consistency.

Mean teacher. Mean teacher [32] is initially developed for semi-supervised classification and recently applied for semi-supervised segmentation, e.g., in CutMix-Seg [11]. The unlabeled image with different augmentations is fed into two networks with the same structure: one is student $f(\theta)$, and the other one is mean teacher $f(\bar{\theta})$ with the parameter $\bar{\theta}$ being the moving average of the student network parameter θ:

$$\begin{aligned} \mathsf{X} \rightarrow \mathsf{X}_1 \rightarrow f(\theta) \rightarrow \mathsf{P}_1 \\ \searrow \mathsf{X}_2 \rightarrow f(\bar{\theta}) \nrightarrow \mathsf{P}_2. \end{aligned} \tag{8}$$

We use X_1 and X_2 to represent the differently augmented version of X. The consistency regularization aims to align the probability map P_1 of X_1 predicted by the student network to the probability map P_2 of X_2 predicted by the teacher network. During the training, we supervise P_1 with P_2 and apply no back propagation for the teacher network. We use $\nrightarrow$ to represent "no back propagation" in the following illustration. we have not included the loss supervision in the above equations and illustrate the complete version in Figure 1 (c). The results in Table 1 and Table 2 show that our approach is superior to the mean teacher approach.

Single-network pseudo supervision. We consider a downgraded version of our approach, single-network pseudo supervision, where the two networks are the same:

$$\begin{aligned} \mathsf{X} \rightarrow \mathsf{X} \rightarrow f(\theta) \rightarrow \mathsf{P} \nwarrow \\ \searrow f(\theta) \rightarrow \mathsf{P} \nrightarrow \mathsf{Y}. \end{aligned} \tag{9}$$

The structure is similar to Figure 1 (d), and the only difference is that the inputs to two streams are the same rather than one weak augmentation and one strong augmentation. We use $\nwarrow$ from Y to P to represent the loss supervision.

Empirical results show that single-network pseudo supervision performs poorly. The main reason is that supervision by the pseudo label from the same network tends to learn the network itself to better approximate the pseudo labels and thus the network might converge in the wrong direction. In contrast, supervision by the cross pseudo label from the other network, which differs from the pseudo label from the network itself due to network perturbation, is able to learn the network with some probability away from the wrong direction. In other words, the perturbation of pseudo label between two networks in some sense serves as a regularizer, free of over-fitting the wrong direction.

In addition, we study the single-network pseudo supervision in a way like [11] with the CutMix augmentation. We input two source images into a network $f(\theta)$ and mix the two pseudo segmentation maps as a pseudo segmentation map of the CutMixed image, which is used to supervise the output of the CutMixed image from the same network. Back propagation of the pseudo supervision is only done for the CutMixed image. The results show that our approach performs better (Table 6), implying that network perturbation is helpful though there is already perturbation from the way with the CutMix augmentation in [11].

PseudoSeg. PseudoSeg [44], similar to FixMatch [28], applies weakly-augmented image X_w to generate pseudo segmentation map, which is used to supervise the output of strongly-augmented image X_s from the same network with the same parameters. X_w and X_s are based on the same input image X. PseudoSeg only conducts back propagation on the path that processes the strongly-augmented image X_s (illustrated in Figure 1 (d)). It is logically formed as:

$$\begin{aligned} \mathsf{X} \rightarrow \mathsf{X}_s \rightarrow f(\theta) \rightarrow \mathsf{P}_s \nwarrow \\ \searrow \mathsf{X}_w \rightarrow f(\theta) \rightarrow \mathsf{P}_w \nrightarrow \mathsf{Y}_w. \end{aligned} \tag{10}$$

We use $\nwarrow$ from Y_w to P_s to represent the loss supervision. The above manner is similar to single-network pseudo supervision. The difference is that the pseudo segmentation map is from weak augmentation and it supervises the training over strong augmentation. We guess that besides the segmentation map based on weak augmentation is more accurate, the other reason is same as our approach: the pseudo segmentation map from weak augmentation also introduces extra perturbation to the pseudo supervision.

5. Experiments

5.1. Setup

Datasets. *PASCAL VOC* 2012 [8] is a standard object-centric semantic segmentation dataset, which consists of more than $13,000$ images with 20 object classes and 1 background class. The standard training, validation and test sets consist of $1,464$, $1,449$ and $1,456$ images respectively. We follow the previous work to use the augmented set [12] ($10,582$ images) as our full training set.

Cityscapes [7] is mainly designed for urban scene understanding. The official split has $2,975$ images for training, 500 for validation and $1,525$ for testing. Each image has a resolution of 2048×1024 and is fine-annotated with pixel-level labels of 19 semantic classes.

We follow the partition protocols of Guided Collaborative Training (GCT) [17] and divide the whole training set to two groups via randomly sub-sampling $1/2$, $1/4$, $1/8$ and $1/16$ of the whole set as the labeled set and regard the remaining images as the unlabeled set.

Evaluation. We evaluate the segmentation performance using mean Intersection-over-Union (mIoU) metric. For all partition protocols, we report results on the $1,456$ PASCAL VOC 2012 `val` set (or 500 Cityscapes `val` set) via only

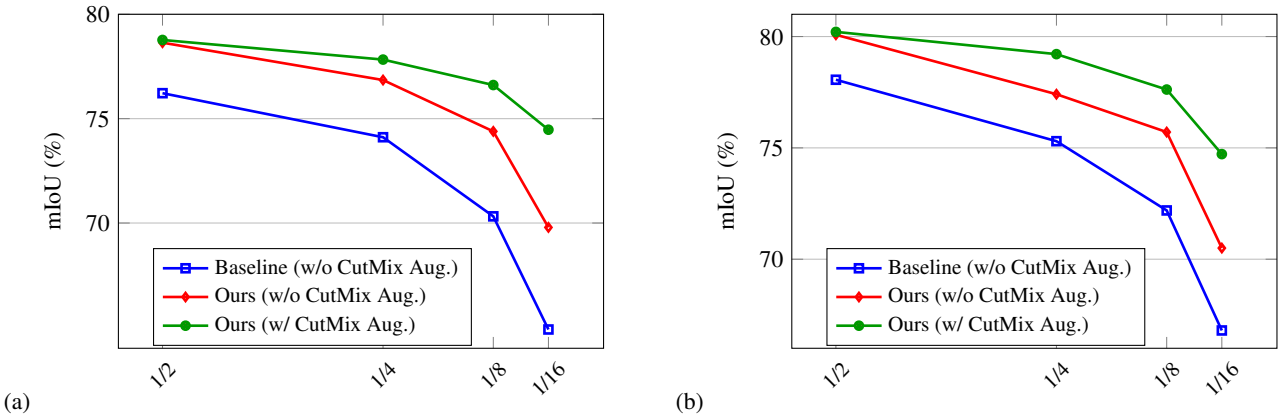

Figure 2: **Improvements over the supervised baseline** on the Cityscapes val set with (a) ResNet-50 and (b) ResNet-101.

single scale testing. We only use one network in our approach to generate the results for evaluation.

Implementation details. We implement our method based on PyTorch framework. We initialize the weights of two backbones in the two segmentation networks with the same weights pre-trained on ImageNet and the weights of two segmentation heads (of DeepLabv3+) randomly. We adopt mini-batch SGD with momentum to train our model with Sync-BN [16]. The momentum is fixed as 0.9 and the weight decay is set to 0.0005. We employ a poly learning rate policy where the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{0.9}$.

For the supervised baseline trained on the full training set, we use random horizontal flipping and multi-scale as data augmentation if not specified. We train PASCAL VOC 2012 for 60 epochs with base learning rate set to 0.01, and Cityscapes for 240 epochs with base learning rate set to 0.04. OHEM loss is used on Cityscapes.

5.2. Results

Improvements over baselines. We illustrate the improvements of our method compared with the supervised baseline under all partition protocols in Figure 2. All the methods are based on DeepLabv3+ with ResNet-50 or ResNet-101.

Figure 2 (a) shows our method consistently outperforms the supervised baseline on Cityscapes with ResNet-50. Specifically, the improvements of our method w/o CutMix augmentation over the baseline method w/o CutMix augmentation are 4.89%, 4.07%, 2.74%, and 2.42% under 1/16, 1/8, 1/4, and 1/2 partition protocols separately. Figure 2 (b) shows the gains of our method over the baseline method on Cityscapes with ResNet-101: 3.70%, 3.52%, 2.11%, and 2.02% under 1/16, 1/8, 1/4, and 1/2 partition protocols separately.

Figure 2 also shows the improvements brought by the CutMix augmentation. We can see that CutMix brings more gains under the 1/16 and 1/8 partitions than under the 1/4 and 1/2 partitions. For example, on Cityscapes with ResNet-101, the extra gains brought by CutMix augmentation are 4.22%, 1.91%, and 0.13% under 1/16, 1/8, and 1/2 partition protocols separately.

Comparison with SOTA. We compare our method with some recent semi-supervised segmentation methods including: Meat-Teacher (MT) [32], Cross-Consistency Training (CCT) [27], Guided Collaborative Training (GCT) [17], and CutMix-Seg [11] under different partition protocols. Specifically, we adopt the official open-sourced implementation of CutMix-Seg. For MT and GCT, we use implementations from [17]. We compare them using the same architecture and partition protocols for fairness.

PASCAL VOC 2012: Table 1 shows the comparison results on PASCAL VOC 2012. We can see that over all the partitions, with both ResNet-50 and ResNet-101, our method w/o CutMix augmentation consistently outperforms the other methods except CutMix-Seg that uses the strong CutMix augmentation [40].

Our approach w/ CutMix augmentation performs the best and sets new state-of-the-arts under all partition protocols. For example, our approach w/ CutMix augmentation outperforms the CutMix-Seg by 3.08% and 1.92% under 1/16 partition protocol with ResNet-50 and ResNet-101 separately. The results imply that our cross pseudo supervision scheme is superior to mean teacher scheme that is used in CutMix-Seg.

When comparing the results of our approach w/o and w/ CutMix augmentation, we have the following observation: the CutMix augmentation is more important for the scenario with fewer labeled data. For example, with ResNet-50, the gain 3.77% under the 1/16 partition is higher than 0.47% under the 1/8 partition.

Cityscapes: Table 2 illustrates the comparison results on the Cityscapes `val` set. We do not have the results for CutMix-Seg as the official CutMix-Seg implementation only supports single-GPU training and it is not feasible to run CutMix-Seg with DeepLabv3+ on Cityscapes due to the GPU memory limit. In comparison to other SOTA methods,

Table 1: **Comparison with state-of-the-arts** on the PASCAL VOC 2012 val set under different partition protocols. All the methods are based on DeepLabv3+.

Method	ResNet-50				ResNet-101			
	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
MT [32]	66.77	70.78	73.22	75.41	70.59	73.20	76.62	77.61
CCT [27]	65.22	70.87	73.43	74.75	67.94	73.00	76.17	77.56
CutMix-Seg [11]	68.90	70.70	72.46	74.49	72.56	72.69	74.25	75.89
GCT [17]	64.05	70.47	73.45	75.20	69.77	73.30	75.25	77.14
Ours (w/o CutMix Aug.)	68.21	73.20	74.24	75.91	72.18	75.83	77.55	78.64
Ours (w/ CutMix Aug.)	**71.98**	**73.67**	**74.90**	**76.15**	**74.48**	**76.44**	**77.68**	**78.64**

Table 2: **Comparison with state-of-the-arts** on the Cityscapes val set under different partition protocols. All the methods are based on DeepLabv3+.

Method	ResNet-50				ResNet-101			
	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
MT [32]	66.14	72.03	74.47	77.43	68.08	73.71	76.53	78.59
CCT [27]	66.35	72.46	75.68	76.78	69.64	74.48	76.35	78.29
GCT [17]	65.81	71.33	75.30	77.09	66.90	72.96	76.45	78.58
Ours (w/o CutMix Aug.)	69.79	74.39	76.85	78.64	70.50	75.71	77.41	80.08
Ours (w/ CutMix Aug.)	**74.47**	**76.61**	**77.83**	**78.77**	**74.72**	**77.62**	**79.21**	**80.21**

Table 3: **Comparison with state of the arts** on the Cityscapes val set under different partition protocols using HRNet-W48.

Method	Cityscapes			
	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
Base	66.90	72.79	75.23	78.09
Ours (w/o CutMix Aug.)	72.49	76.32	78.27	80.02
Ours (w/ CutMix Aug.)	**75.09**	**77.92**	**79.24**	**80.67**

our method achieves the best performance among all partition protocols with both ResNet-50 and ResNet-101 backbones. For example, our method w/ CutMix augmentation obtains 80.08% under the 1/2 partition with ResNet-101 backbone, which outperforms GCT by 1.50%. We report the additional results on HRNet in Table 3.

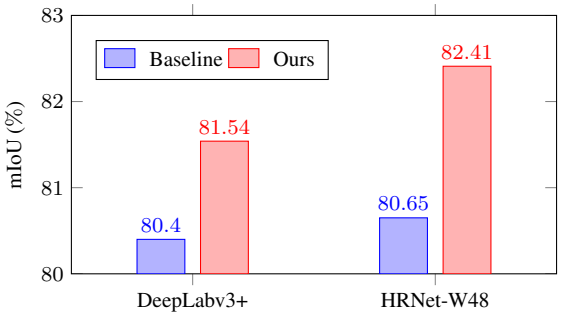

Figure 3: **Improving the fully-supervised baselines.** The baseline models (DeepLabv3+ with ResNet-101 and HRNet-W48) are trained using the full Cityscapes `train` set. Our approach uses $\sim 3,000$ images from Cityscapes `coarse` set as an additional unlabeled set for training. The superiority of our approach implies that our approach works well on the relatively large labeled data.

5.3. Improving Full- and Few-Supervision

Full-supervision. We verify our method using the full Cityscapes `train` set ($\sim 2,975$ images) and randomly sample $3,000$ images from the Cityscapes `coarse` set as the unlabeled set. For the unlabeled set, we do not use their coarsely annotated ground truth. Figure 3 illustrates the results on the Cityscapes `val` set with single-scale evaluation. We can see that even with a large amount of labeled data, our approach could still benefit from training with unlabeled data, and our approach also works well on the state-of-the-art segmentation network HRNet.

Few-supervision. We study the performance of our method on PASCAL VOC 2012 with very few supervision by following the same partition protocols adopted in PseudoSeg [44]. PseudoSeg randomly samples 1/2, 1/4, 1/8, and 1/16 of images in the standard training set (around 1.5k images) to construct the labeled set. The remaining images in the standard training set, together with the images in the augmented set [12] (around 9k images), are used as the unlabeled set.

We only report the results of our approach w/ CutMix augmentation as CutMix is important for few supervision. Results are listed in Table 5, where all methods use ResNet-101 as the backbone except CCT that uses ResNet-50. We can see that our approach performs the best and is supe-

Table 4: **Ablation study of different loss combinations** on PASCAL VOC 2012 and Cityscapes. The results are obtained under the 1/8 data partition protocol and the observations are consistent for other partition protocols. $\mathcal{L}_s$ represents the supervision loss on the labeled set. $\mathcal{L}^l_{cps}$ ($\mathcal{L}^u_{cps}$) represents the cross pseudo supervision loss on the labeled (unlabeled) set. $\mathcal{L}^l_{cpc}$ ($\mathcal{L}^u_{cpc}$) represents the cross probability consistency loss on the labeled (unlabeled) set. The overall performance with the cross pseudo supervision loss on both the labeled and unlabeled data is the best.

Losses					PASCAL VOC 2012		Cityscapes	
$\mathcal{L}_s$	$\mathcal{L}^l_{cps}$	$\mathcal{L}^u_{cps}$	$\mathcal{L}^l_{cpc}$	$\mathcal{L}^u_{cpc}$	ResNet-50	ResNet-101	ResNet-50	ResNet-101
✓					69.43	72.21	70.32	72.19
✓	✓				69.99	72.98	71.73	73.08
✓		✓			73.00	75.83	73.97	75.28
✓	✓	✓			**73.20**	**75.85**	**74.39**	**75.71**
✓			✓	✓	71.23	74.01	72.03	73.77

Table 5: **Comparison for few-supervision** on PASCAL VOC 2012. We follow the same partition protocols provided in PseudoSeg [44]. The results of all the other methods are from [44].

Method	#(labeled samples)			
	732	366	183	92
AdvSemSeg [13]	65.27	59.97	47.58	39.69
CCT [27]	62.10	58.80	47.60	33.10
MT [32]	69.16	63.01	55.81	48.70
GCT [17]	70.67	64.71	54.98	46.04
VAT [26]	63.34	56.88	49.35	36.92
CutMix-Seg [11]	69.84	68.36	63.20	55.58
PseudoSeg [44]	72.41	69.14	65.50	57.60
Ours (w/ CutMix Aug.)	**75.88**	**71.71**	**67.42**	**64.07**

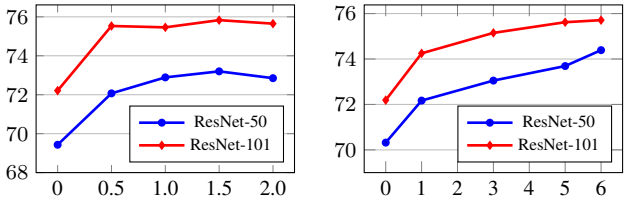

Figure 4: **Illustration on how the trade-off weight λ (x-axis) affects the mIoU score (y-axis)** on PASCAL VOC 2012 (left) and Cityscapes (right). All results are evaluated under the $1/8$ partition protocol.

rior to CutMix-Seg again on the few labeled case. Our approach is also better than PseudoSeg that uses a complicated scheme to compute the pseudo segmentation map. We believe that the reason comes from that our approach uses network perturbation and cross pseudo supervision while PseudoSeg uses a single network with input perturbation.

5.4. Empirical Study

Cross pseudo supervision. We investigate the influence of applying the proposed cross pseudo supervision loss to labeled set ($\mathcal{L}^l_{cps}$) or unlabeled set ($\mathcal{L}^u_{cps}$) in the Table 4. We can see that cross pseudo supervision loss on the unlabeled set brings more significant improvements than cross pseudo supervision loss on the labeled set in most cases. For example, with ResNet-50, cross pseudo supervision loss on the labeled set improves the performance of the baseline by 0.56% (1.41%) while cross pseudo supervision loss on the unlabeled set improves by 3.57% (4.07%) on PASCAL VOC 2012 (Cityscapes). The performance with cross pseudo supervision loss on both labeled set and unlabeled set is overall the best.

Table 6: **Comparison with single-network pseudo supervision** on PASCAL VOC 2012 `val`. SPS = single-network pseudo supervision. All methods are based on DeepLabv3+ are with ResNet-50. We can see that for both the two cases, w/ and w/o CutMix augmentation, our approach outperforms the single-network pseudo supervision.

Method	1/16	1/8	1/4	1/2
SPS (w/o CutMix Aug.)	59.54	69.05	72.55	75.17
Ours (w/o CutMix Aug.)	**68.21**	**73.20**	**74.24**	**75.91**
SPS (w/ CutMix Aug.)	65.62	71.27	73.70	74.87
Ours (w/ CutMix Aug.)	**71.98**	**73.67**	**74.90**	**76.15**

Comparison with cross probability consistency. We compare our method with the cross probability consistency on the last 2 rows of Table 4. We can see that our cross pseudo supervision outperforms the cross probability consistency on both benchmarks. For example, on Cityscapes, cross pseudo supervision outperforms cross probability consistency by 2.36% (1.94%) when applied to both labeled and unlabeled sets with ResNet-50 (ResNet-101).

The trade-off weight λ. We investigate the influence of different λ that is used to balance the supervision loss and cross pseudo supervision loss as shown in Equation 6. From Figure 4, we can see that $\lambda = 1.5$ performs best on PASCAL VOC 2012 and $\lambda = 6$ performs best on Cityscapes. We use $\lambda = 1.5$ and $\lambda = 6$ in our approach for all the experiments.

Single-network pseudo supervision vs. cross pseudo supervision. We compare the proposed approach with single-network pseudo supervision in Table 6. We can see that

Table 7: **Combination with self-training.** The CutMix augmentation is not used. We can see that the combination gets improves over both self-training and our approach.

Method	ResNet-50		ResNet-101	
	1/4	1/2	1/4	1/2
PASCAL VOC 2012				
Ours	74.24	75.91	77.55	78.64
Self-Training	74.47	75.97	76.63	78.15
Ours + Self-Training	**74.96**	**76.60**	**77.60**	**78.76**
Cityscapes				
Ours	76.85	78.64	77.41	80.08
Self-Training	75.88	77.64	77.55	79.46
Ours + Self-Training	**77.40**	**79.25**	**79.16**	**80.17**

our method outperforms the single-network pseudo supervision scheme either with CutMix augmentation or not. The single-network pseudo supervision with the CutMix augmentation is similar to the application of FixMatch [28] to semantic segmentation (as done in PseudoSeg). We think that this is one of the main reason that our approach is superior to PseudoSeg.

Combination/comparison with self-training. We empirically study the combination of our method and the conventional self-training [35]. Results on both benchmarks are summarized in Table 7. We can see that the combination of self-training and our approach outperforms both our method only and self-training only. The superiority implies that our approach is complementary to self-training.

As the self-training scheme consists of multiple stages (train over labeled set $\rightarrow$ predict pseudo labels for unlabeled set $\rightarrow$ retrain over labeled and unlabeled set with pseudo labels), it takes more training epochs than our approach. For a fairer comparison with self-training, we train our method for more epochs (denoted as Ours$^+$) to ensure our training epochs are also comparable with self-training. According to the results shown in Figure 5, we can see that ours$^+$ consistently outperforms self-training under various partition protocols. We guess that the reason lies in the consistency regularization in our approach.

5.5. Qualitative Results

Figure 6 visualizes some segmentation results on PASCAL VOC 2012. We can see the supervised baseline, shown in the Figure 6 column (c), mis-classifies many pixels due to limited labeled training samples. For example, in the 1-st row, the supervised only method (column (c)) mistakenly classifies many cow pixels as horse pixels while our method w/o CutMix augmentation (column (d)) fixes these errors. In the 2-nd row, both the supervised baseline and our method w/o CutMix augmentation, mislabel some dog pixels as horse pixels while our method w/ CutMix augmentation (column (e)) successfully corrects these errors.

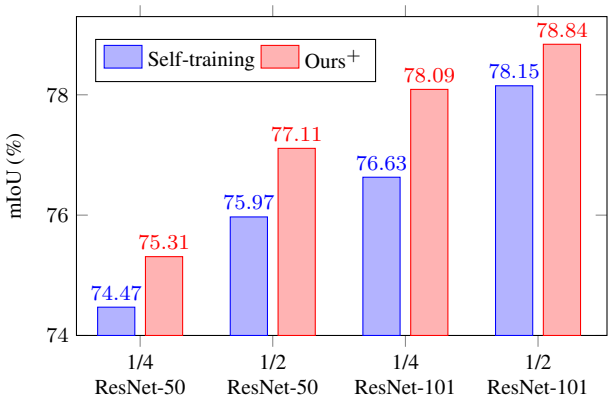

Figure 5: **Comparison with self-training** on PASCAL VOC 2012. The self-training approach is a two-stage approach which takes more training epochs. For a fair comparison, we train our approach with more training epochs (denoted by 'Ours$^+$') so that their epochs are comparable. The CutMix augmentation is not used.

Figure 6: **Example qualitative results from PASCAL VOC** 2012. (a) input, (b) ground truth, (c) supervised only, (d) ours (w/o CutMix Aug.), and (e) ours (w/ CutMix Aug.). All the approaches use DeepLabv3+ with ResNet-101 as the segmentation network.

6. Conclusion

We present a simple but effective semi-supervised segmentation approach, cross pseudo supervision. Our approach imposes the consistency between two networks with the same structure and different initialization, by using the one-hot pseudo segmentation map obtained from one network to supervise the other network. On the other hand, the unlabeled data with pseudo segmentation map, which is more accurate in the later training stage, serves as expanding the training data to improve the performance.

Acknowledgments: This work is supported by the National Key Research and Development Program of China (2017YFB1002601, 2016QY02D0304), National Natural Science Foundation of China (61375022, 61403005, 61632003), Beijing Advanced Innovation Center for Intelligent Robots and Systems (2018IRS11), and PEK-SenseTime Joint Laboratory of Machine Vision.

Appendix

A. More Implementation Details

Training details. The crop size for PASCAL VOC 2012 and Cityscapes are 512×512 and 800×800, respectively. For the multi-scale data augmentation, we randomly select scale from $\{0.5, 0.75, 1, 1.25, 1.5, 1.75\}$. For Cityscapes dataset, we use OHEM loss as the supervision loss ($\mathcal{L}_s$), and cross entropy loss as the cross pseudo supervision loss ($\mathcal{L}_{cps}$).

Training strategy. We use the similar training strategy as GCT [17] for semi-supervised segmentation. In the supervised baseline for all the partition protocols, we use the batch size 8. We ensure that the iteration number is the same as semi-supervised methods[2]. For semi-supervised methods, at each iteration, we sample additional 8 unlabeled samples. Our method and all the other semi-supervised methods in Table 1 and Table 2 of the main paper follow the same training strategy.

B. Network Perturbation

Our cross pseudo supervision approach (CPS) includes two perturbed segmentation networks, $f(\theta_1)$ and $f(\theta_2)$, which are of the same architecture and initialized differently. In the main paper, we pointed out that the pseudo segmentation results from the two networks are perturbed.

We empirically show the perturbation using the overlap ratio between them during training. The overlap ratio on the labeled set, the unlabeled set and the whole set are given in Figure 7. We can see that (1) the overlap ratio is small at the early training stage and (2) increases during the later training stage. The small overlap ratio at the early stage helps avoid the case the segmentation network converges towards a wrong direction. The large overlap ratio at the later stage implies that the pseudo segmentation results of the two segmentation networks are more accurate.

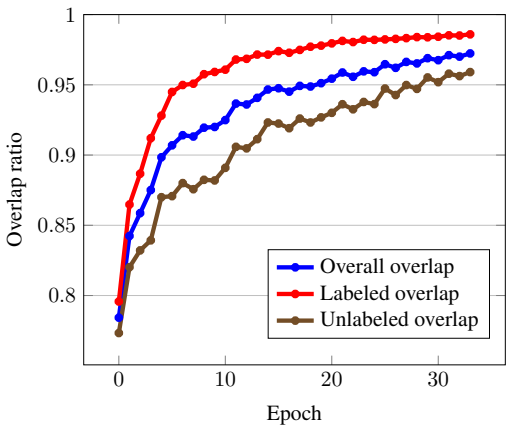

Figure 7: **Prediction overlap of the two networks on PASCAL VOC 2012 under the 1/8 partition**. We use DeepLabv3+ with ResNet-50 as the segmentation network. We only calculate the overlap ratio in the object region, and the pixels belong to the 'background' class are ignored.

References

[1] Ashok K. Agrawala. Learning with a probabilistic teacher. *IEEE Trans. Inf. Theory*, 16(4):373–379, 1970. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NIPS*, pages 5049–5059, 2019. 2

[3] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006. 2

[4] 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. *PAMI*, 2018. 1, 2

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

[6] Maxwell D Collins, Ekin D Cubuk, Hartwig Adam Barret Zoph, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*. Springer, 2020. 1, 2

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 4

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 4

[9] 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. *CoRR*, abs/2004.08514, 2020. 1, 2

[10] Stanley C. Fralick. Learning to recognize patterns without a teacher. *IEEE Trans. Inf. Theory*, 13(1):57–64, 1967. 2

[11] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *BMVC*, 2020. 1, 2, 3, 4, 5, 6, 7

[12] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 4, 6

[13] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-

[2] In GCT [17], the supervised baseline uses the batch size 16, and the number of iterations is much smaller than half of the number of iterations in the semi-supervised methods. Therefore, their supervised baseline results are worse than ours (we sure the same number of iterations and each iteration has the same number, 8, of labeled samples).

supervised semantic segmentation. In *BMVC*, 2018. 1, 2, 7

[14] Mostafa S Ibrahim, Arash Vahdat, Mani Ranjbar, and William G Macready. Semi-supervised semantic image segmentation with self-correcting networks. In *CVPR*, pages 12715–12725, 2020. 2

[15] H. J. Scudder III. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inf. Theory*, 11(3):363–371, 1965. 2

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[17] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, 2020. 2, 3, 4, 5, 6, 7, 9

[18] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *ICCV*, pages 6728–6736, 2019. 1, 2

[19] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. Structured consistency loss for semi-supervised semantic segmentation. *CoRR*, abs/2001.04647, 2020. 1, 2

[20] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9799–9808, 2020. 2

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ICLR*, 2017. 2

[22] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 2

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[24] Robert Mendel, Luis Antonio de Souza, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *ECCV*, pages 141–157. Springer, 2020. 2

[25] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *CoRR*, abs/1908.05724, 2019. 1, 2

[26] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 2018. 7

[27] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020. 1, 2, 5, 6, 7

[28] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *CoRR*, abs/2001.07685, 2020. 2, 3, 4, 8

[29] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, pages 5688–5696, 2017. 2

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 1

[31] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *ICCV*, pages 5229–5238, 2019. 2

[32] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017. 2, 4, 5, 6, 7

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[34] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 1

[35] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 8

[36] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016. 1

[37] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019. 2

[38] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv:1809.00916*, 2018. 2

[39] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *ECCV*, pages 489–506. Springer, 2020. 2

[40] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 3, 5

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[42] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R. Manmatha, Mu Li, and Alexander J. Smola. Improving semantic segmentation via self-training. *CoRR*, abs/2004.14960, 2020. 1, 2

[43] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training. *CoRR*, abs/2006.06882, 2020. 1, 2

[44] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 2, 3, 4, 6, 7