Locating Noise is Halfway Denoising for Semi-Supervised Segmentation

Yan Fang[1,2] *, Feng Zhu[3], Bowen Cheng[4] †, Luoqi Liu[5], Yao Zhao[1,2], Yunchao Wei[1,2] ✉

[1] Institute of Information Science, Beijing Jiaotong University
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology
[3] University of Technology Sydney [4] University of Illinois Urbana-Champaign
[5] MT Lab, Meitu Inc

Abstract

We investigate semi-supervised semantic segmentation with self-training, where a teacher model generates pseudo masks to exploit the benefits of a large amount of unlabeled images. We notice that the noisy label from the generated pseudo masks is the major obstacle to achieving good performance. Previous works all treat the noise in pixel level and ignore the contextual information of the noise. This work shows that locating the patch-wise noisy region is a better way to deal with noise. To be specific, our method, named Uncertainty-aware Patch CutMix (UPC), first estimates the uncertainty of per-pixel prediction for pseudo masks of unlabeled images. Then UPC splits the uncertainty map into patches and calculates patch-wise uncertainty. UPC selects top-k most uncertain patches to generate the uncertain regions. Finally, uncertain regions are replaced with reliable ones from labeled images. We conduct extensive experiments using standard semi-supervised settings on Pascal VOC and Cityscapes. Experiment results show that UPC can significantly boost the performance of the state-of-the-art methods. In addition, we further demonstrate that our UPC is robust to out-of-distribution unlabeled images, e.g.*, MSCOCO.*

1. Introduction

Past decades have witnessed significant progress in semantic segmentation thanks to the rapid growth of deep learning [20, 6, 9, 8, 30] and large-scale datasets [29, 10, 13] with accurate pixel-level annotations. However, annotating a massive number of training images is expensive and time-consuming, which limits the ability to scale to various segmentation tasks or scenarios. To reduce the annotation cost, semi-supervised semantic segmentation algorithms [33, 31, 15] emerged by exploiting a small number of labeled data and a large amount of unlabeled data.

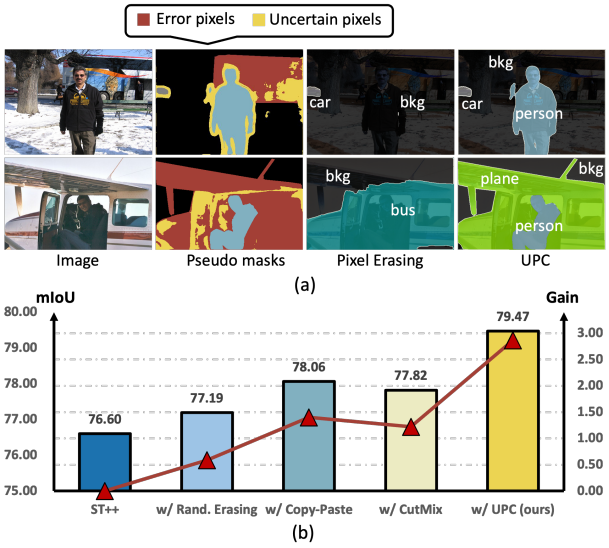

Figure 1. (a) Illustrations of the difference between the Pixel Erasing method and our UPC. The Pixel Erasing method directly removes highly uncertain pixels in pseudo masks. (b) Comparison with other popular data augmentation methods, based on ST++.

State-of-the-art semi-supervised semantic segmentation can be categorized into two groups: consistency regularization and entropy minimization. Particularly, consistency regularization-based approaches [15, 33, 7] enforce similar predictions of the same image under different transformations, and a representative work is the Mean-Teacher [38]. However, consistency regularization usually has complex training pipelines to achieve good performance, *e.g.*, contrastive learning [27] or class-balanced strategies [21, 3, 14]. In contrast, entropy minimization [17] often follows a self-training [14] pipeline by retraining the model with pseudo masks on unlabeled images, which leads to a simpler

*Work done during internships at MT Lab, Meitu Inc.
†Work done while at University of Illinois Urbana-Champaign.
✉Corresponding author.

pipeline. Nevertheless, models with self-training tend to over-fit noisy pseudo masks, leading to a degradation in performance compared to consistency regularization. Thus, how to reduce the impact of noise becomes the key to improving the performance of self-training methods.

Previous works try a variety of ways to moderate the effect of noise in pseudo labels. [15, 23, 33] employ a strict threshold to filter out low-confidence pixels. [21] proposes dynamic re-weighting to increase the contribution of reliable pixels. [39] stores unreliable pixels into a category-wise memory bank to serve as negative samples. Although these approaches have achieved significant progress, they all ignore a critical point: *contextual information of the noisy pixel*. As shown in Fig. 1 (a), the contextual information includes two folds: (1) Noisy pixels do not appear alone. The pixels around the noisy pixel tend to be noise even if they have high confidence. (2) noisy pixels always lie on the object edge, and ignoring them will corrupt the learning of object shape and structure. Thus, we ask the question: *Can we locate the uncertain region instead of pixels?*

To address the question, we introduce a straightforward data augmentation technique called Uncertainty-aware Patch CutMix (UPC). UPC leverages uncertainty as a guide and identifies noisy regions in a patch-wise manner rather than a pixel-wise approach. Specifically, we first train a segmentation model with labeled images, and then use this model to create pseudo masks and uncertainty maps on unlabeled images. We then feed these uncertainty maps into UPC to pinpoint the uncertain regions. UPC begins by dividing the uncertainty map into patches of size $N \times N$. The uncertainty of each patch is calculated by summing the uncertainty of all pixels within it. UPC then selects the top-k unreliable patches as uncertain regions. Once identified, UPC replaces these uncertain regions using Patch CutMix with reliable regions from another labeled image. To ensure the denoising function is both robust and stable, UPC employs a redundant augmentation strategy, which generates multiple transferred pairs for each unlabeled image using different values of k and labeled images. There are several notable advantages to our UPC method. Firstly, the patch-wise identification of uncertain regions preserves contextual information and facilitates subsequent CutMix operations. Secondly, UPC is model-structure and training-procedure agnostic, making it easy to combine with other semi-supervised methods. Thirdly, UPC is simple and straightforward to implement, requiring minimal effort to incorporate into existing workflows.

We conduct extensive experiments on Pascal VOC [13] and Cityscapes [10] datasets using standard settings to evaluate the effectiveness of our UPC method. Our UPC yields outstanding performance on both datasets when combined with previous competitive approaches. Specifically, on Pascal VOC and Cityscapes, our UPC improves the competitive method U^{2}PL by 3.33% and 5.01%, respectively. Additionally, our UPC outperforms other data augmentation methods on the competitive method ST++, as shown in Fig. 1 (b). Furthermore, the spirit of semi-supervised learning is to fully utilize knowledge from unlabeled images, regardless of their distribution. However, most previous practices use unlabeled images with similar distributions as labeled images, which is impractical and contradicts the semi-supervised learning spirit. In this study, we examine the ability of our UPC by taking MSCOCO [29] as out-of-distribution (o.o.d) unlabeled data and Pascal VOC as labeled data. The results demonstrate excellent generalization capabilities of our UPC, regardless of the distribution of unlabeled images.

Overall, our contributions can be concluded as below:

1. We propose the Uncertainty-aware Patch CutMix (UPC) method from an accurate denoising perspective. Through patch-wise locating mechanism, our method effectively reduces the impact of noisy pseudo masks in semi-supervised semantic segmentation.

2. Without introducing any extra hyper-parameters or retraining stages, our method boosts semi-supervised semantic segmentation by a large margin. Importantly, our method is a plug-and-play method that can be easily combined with other semi-supervised works.

2. Related Work

2.1. Semi-supervised Learning

Consistency regularization [15, 33, 34] and entropy minimization [17, 28] are two primary approaches to semi-supervised learning. Consistency regularization [1] aims to enforce the model to generate stable and consistent predictions under various perturbations, which can be categorized into input perturbations and network perturbations. Recent works have explored data augmentation as input perturbation [15] or used different initialized networks as network perturbation [33]. On the other hand, entropy minimization originally introduced by Grandvalet and Bengio [17], has gained popularity through self-training schemes [4, 5, 49, 14, 22]. This approach adopts an explicit bootstrapping technique, where unlabeled data is assigned pseudo-labels for iterative retraining. Recent studies have further advanced the self-training paradigm by incorporating additional retraining stages [4] and employing data augmentation strategies [42, 41], building upon the fundamental insight of entropy minimization. MixMatch [2] is different from prior works by harvesting the advantages of both methods and proposing a hybrid framework to exploit

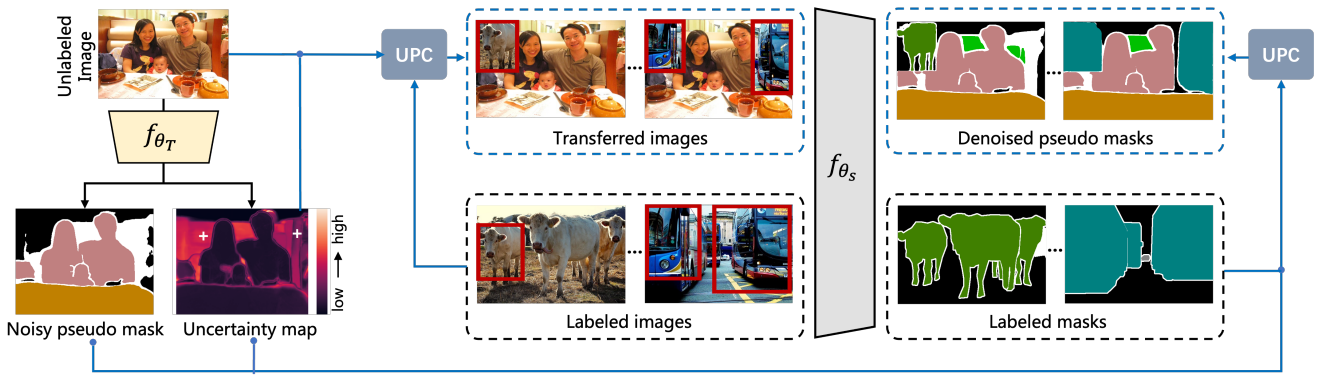

Figure 2. A brief overview of the UPC pipeline. We use f_{θ_T} and f_{θ_S} to represent the teacher model and student model respectively. First, UPC estimates the uncertainty of pseudo masks, producing uncertainty maps. In this illustration, the light-colored areas in the uncertainty map always correspond to noisy regions. Using this uncertainty guidance, UPC replaces these noisy regions with substitutes from different labeled pairs in a patch-wise manner, resulting in multiple denoised results for training.

unlabeled data from weak and strong augmentation. FixMatch [35] inherits the spirit of MixMatch [2] with simplified mechanisms and has a significant impact on subsequent works. FlexMatch [44], utilizes curriculum learning to filter low-confidence labels with class-wise thresholds.

2.2. Semi-supervised Semantic Segmentation

In semi-supervised semantic segmentation, preliminary works [37, 31, 22] utilize GANs [16] to generate supervision signals for unlabeled data but face difficulties due to mode collapse. Inspired by consistency regularization in semi-supervised learning, subsequent methods use simpler mechanisms to enforce similar predictions under multiple perturbed embeddings [33], contextual crops [27], and dual initialized models [24]. PseudoSeg [52] extends the weak-to-strong consistency of FixMatch [35] to segmentation and adds a module to refine pseudo masks. Recent works introduce more supervision, such as contrastive learning [39]. It is worth noting that recent studies have directed their attention to enhancing the category capacity of models through incremental learning [45, 46] and open-vocabulary [18] techniques. These approaches have the potential to complement semi-supervised methods. Despite dedicated mechanisms, we propose that simple data augmentation is effective enough to improve semi-supervised segmentation.

2.3. Self Training

The technique of self-training via pseudo labeling, initially proposed approximately a decade ago [28], has gained renewed interest across various domains. It has recently garnered attention in fully-supervised image recognition [5, 50], semi-supervised learning [4, 42, 36, 40], and domain adaptation [26, 51]. This classical method has experienced a resurgence and is now being explored in diverse research areas. Especially, it has been revisited in several semi-supervised tasks, including image classification, object detection [36], and semantic segmentation [4, 42, 41]. Among them, the most related ones are Naive Student [4], ST++ [41]. Nevertheless, our work is fundamentally different from these works in that we propose an effective denoising method on pseudo masks, which is extremely beneficial to semi-supervised semantic segmentation.

2.4. Strong Data Augmentation

Strong data augmentation (SDA) [11, 43, 32, 12] is widely used as one input perturbation in different tasks. Previous works [42, 15] have proved SDA is effective to semi-supervised semantic segmentation. The former one [42] uses different data augmentations together on unlabeled images. And the latter one [15] tries several commonly used data augmentation and proposes a semi-supervised baseline using augmentation [43] as a perturbation. Differently, SDA used in our method is different from the above two works, we design data augmentation from a pseudo mask denoising perspective other than input perturbation. We use the proposed UPC in the plainest self-training semi-supervised baseline, producing excellent performance.

3. Method

We first introduce the formulation of semi-supervised semantic segmentation task and an overview of our proposed Uncertainty-aware Patch CutMix (UPC) in Sec.3.1. The details of UPC will be introduced in Sec.3.2.

3.1. Overview

Semi-supervised semantic segmentation utilizes a training set consisting of labeled set $\mathcal{D}^l = \{(x_i^l, y_i^l) \mid i = 1, ..., L\}$ and unlabeled set $\mathcal{D}^u = \{(x_j^u) \mid j = 1, ..., U\}$, where $U \gg L$ in ideal cases. Labeled images x_i^l are annotated with human-annotations y_i^l for C classes as supervi-

sion, while unlabeled images x_j^u use their pseudo masks $\hat{y}_j^u$ generated by the teacher model f_{θ_T} as supervision. The teacher model f_{θ_T} and student model f_{θ_S} are two models used in self-training methods, one for generating pseudo labels and the other for final usage. To simplify the expression, we use $P = (x, y)$ to represent an image with its mask, and superscript l, u to indicate the labeled or unlabeled pair. In this representation, a mini-batch used in training can be denoted as $B = (B_l, B_u)$, where $B_l = \{P_0^l, P_1^l, ..., P_{|B_l|}^l\}$ and $B_u = \{P_0^u, P_1^u, ..., P_{|B_u|}^u\}$.

Fig. 2 illustrates the whole pipeline of our proposed method. First, labeled images with their annotations are fed to the teacher model f_{θ_T}. The trained f_{θ_T} then generates predictions on unlabeled images. These predictions are processed into the pseudo masks and corresponding uncertainty maps. The calculation of uncertainty map $\mathcal{H}$ is formulated as following:

$$\mathcal{H}_{i,j} = -\sum_{c=1}^{C} p_{i,j}^c \log p_{i,j}^c, \tag{1}$$

where $p_{i,j}^c$ is the classification probability of pixel (i, j) on the class c, and C is the number of classes. Using the uncertainty map, our UPC mixes the unlabeled images and labeled images into several transferred images, which contain less noise and richer context. The student model f_{θ_S} is trained on those transferred images and labeled images with the optimization target as follows:

$$\mathcal{L} = \mathcal{L}_l + \lambda \mathcal{L}_u, \tag{2}$$

where $\mathcal{L}_l$ and $\mathcal{L}_u$ denote loss on labeled images and loss on unlabeled images respectively. Additionally, λ is the trade-off between $\mathcal{L}_l$ and $\mathcal{L}_u$. Both $\mathcal{L}_l$ and $\mathcal{L}_u$ can be implemented based on cross-entropy loss.

3.2. Uncertainty-aware Patch CutMix

The simple self-training is easy to degrade on semi-supervised semantic segmentation because noise in pseudo labels tend to accumulate and considerably affect model performance, summarized as overfitting and coupling problems. To effectively reduce the noise in pseudo masks, the key problem is how to accurately find the noise and replace them in a proper way.

3.2.1 Patch Split

In previous works [21, 39], using the uncertainty to locate the noise has been proven effective. However, existing works all deal with the noise at the pixel level and ignore the fact that the noisy pixel has rich context information. The context information has two aspects: (1) noisy pixels appear in group. The pixels around the noisy pixel are more likely to be misclassified compared with other pixels. (2) noisy pixels often lie on the object boundary, and have no regular shape. Ignoring the noisy pixels will reshape the object during training, hindering the learning of object shape and structure. To make full use of the context information, we propose Patch Split to locate uncertain regions instead of uncertain pixels.

In Fig. 3, the UPC method locates the uncertain regions in a patch-wise manner. The uncertainty map is split into $N \times N$ patches, and the uncertainty of each patch G is computed by summing the uncertainty of all the pixels in it:

$$\mathcal{H}_G = \sum_{(i,j) \in G} \mathcal{H}_{i,j}. \tag{3}$$

The k patches with the top-k highest uncertainty are selected as uncertain regions G_k, and a binary mask M_k is generated to represent these regions. Formally, M_k can be calculated as follows:

$$M_k(x, y) = \begin{cases} 1, & (x, y) \in G_k \\ 0, & (x, y) \notin G_k. \end{cases} \tag{4}$$

This method enables UPC to locate uncertain regions more accurately and provides better guidance for the patch-wise CutMix operation.

We make statistical analysis on the patch uncertainty and noise area. The statistical result proves that regions with high uncertainty always correspond to the noise area of pseudo masks. The details are provided in the supplementary material.

3.2.2 Patch CutMix

After acquiring the uncertain regions, we have two ways to deal with them: erasing or replacing them with reliable regions. Obviously, replacing uncertain regions with reliable ones is a better way, which achieves reducing noise and introducing reliable supervision signals simultaneously.

For the replacing technique, we choose the CutMix [43], which is widely used in image recognition and semi-supervised semantic segmentation tasks [15, 7]. The most obvious feature of CutMix is using extra images to perform the fusion transformation. Here we introduce CutMix in the semi-supervised task as follows:

$$f(z^u, z^l, M) = z^l \times M + z^u \times (1 - M), \tag{5}$$

where z can be images or masks with the same input size (H, W). z^l represents labeled data, z^u represents unlabeled data, and M indicates a binary matrix with the same size (H, W). CutMix randomly generates one bounding box with its binary mask M according to the pre-set scale and aspect ratio. Then the binary mask M and its supplementary $1 - M$ are used to combine two images in linear combination manner.

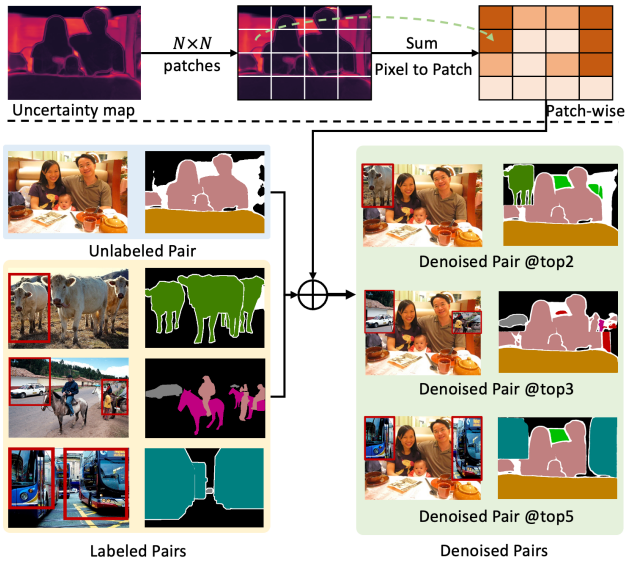

Figure 3. Illustration of detailed data processing procedure of UPC. UPC fuses pixel-wise uncertainty into patches and uses labeled pairs to replace patches in unlabeled pairs. In the uncertainty map, the darker color represents the more reliable region with low uncertainty. The region surrounded by the red box is the crop of labeled images and their corresponding annotation. Further, UPC generates three different denoised pairs with top-2, top-3, and top-5 uncertainty patches.

As shown in Fig.3, UPC performs the above-mentioned CutMix in a patch-wise manner. Given an unlabeled pair $P_j^u = (x_j^u, \hat{y}_j^u)$, a labeled pair $P_i^l = (x_i^l, y_i^l)$ and a binary mask M_k, UPC conducts transformation as follows: For the unlabeled image x_j^u with size (H, W), UPC first splits it into $N \times N$ patches and generate the guidance binary mask M_k in the same way as Patch Split. Then UPC merges P_i^l and P_j^u using the binary mask M_k to produce a transformed pair P_j^t, formulated as,

$$P_j^t(k) = (x_j, \hat{y}_j) = (f(x_j^u, x_i^l, M_k), f(\hat{y}_j^u, y_i^l, M_k)). \quad (6)$$

After the aforementioned transformation, mini-batches are comprised of labeled training pairs and denoised unlabeled training pairs. These denoised unlabeled pairs and labeled pairs form a new mini-batch $B_{new} = (B_l, B_t), B_t = \{P_0^t, P_1^t, ..., P_{|B_t|}^t\}$ that is used as the training input and supervision for the student model f_{θ_S}. The loss functions $\mathcal{L}_l$ and $\mathcal{L}_u$ are calculated on the labeled batch B_l and denoised unlabeled pairs B_t, respectively. Based on Patch Split and Patch CutMix, the noisy regions are replaced with reliable ones, thereby enabling robust semi-supervised retraining.

3.2.3 Redundant Augmentation Strategy

Although Patch CutMix can effectively locate and replace uncertain regions, it is unsuitable to set the same hyper-parameter ($i.e.$ k) for all unlabeled pairs. Moreover, the pseudo masks and uncertainty maps produced by the teacher model f_{θ_T} can not be completely accurate, introducing some unreliable factors into training. To enhance the robustness of model training and fully utilize unlabeled images, we propose the Redundant Augmentation Strategy (RAS) to generate multiple denoised pairs with different top-k values and labeled pairs. Specifically, let x_j^u be the unlabled pair, k_1, k_2, and k_3 be different top-k values, $P_{i_1}^l$, $P_{i_2}^l$, and $P_{i_3}^l$ indicate different labeled pairs. The transformed pairs $P_j^t(k_1)$, $P_j^t(k_2)$, and $P_j^t(k_3)$ can be formulated as,

$$\begin{aligned} P_j^t(k_1) &= (f(x_j^u, x_{i_1}^l, M_{k_1}), f(\hat{y}_j^u, y_{i_1}^l, M_{k_1})), \\ P_j^t(k_2) &= (f(x_j^u, x_{i_2}^l, M_{k_2}), f(\hat{y}_j^u, y_{i_2}^l, M_{k_2})), \\ P_j^t(k_3) &= (f(x_j^u, x_{i_3}^l, M_{k_3}), f(\hat{y}_j^u, y_{i_3}^l, M_{k_3})). \end{aligned} \quad (7)$$

As shown in Fig. 3, UPC produces three transferred samples by selecting different top-k uncertainty regions and reliable contexts, further improving the effect of denoising.

4. Experiment

4.1. Setup

Dataset. The Pascal VOC 2012 [13] is composed of 1464 images for training and 1449 images for validation originally. Following previous works, we use SBD [19] as the augmented set with 9118 additional training images, total 10582 training images. Due to coarse annotations of SBD, PseudoSeg [52] takes only the standard $1,464$ images as the whole labeled set, while other methods take all $10,582$ images as candidate labeled data. Following U^{2}PL [39], we evaluate our method on both the *classic* set ($1,464$ candidate labeled images) and the *blender* set ($10,582$ candidate labeled images). It needs to mention that all protocols of *classic* set use remaining images from a total 10582 images as unlabeled images. The Cityscapes [10] contains 2975 images with fine-grained masks for training and 500 images for validation. For each dataset, we compare with each competitors under commonly used semi-supervised scenarios, which refer to $1/2$, $1/4$, $1/8$, and $1/16$ partition protocols. Additionally, the "Full" setting used in *classic* set refers to using all 1464 labeled images from *classic* set with the remaining training images as unlabeled images.

Except for commonly used semi-supervised dataset setting, we use MSCOCO [29] as a large-scale unlabeled dataset complementary to Pascal VOC training set. To be specific, we use all 10582 images from Pascal VOC training set as labeled images, and images from the MSCOCO training set as unlabeled images.

Network Architecture. We use ResNet-101 [20] pretrained on ImageNet as the backbone and the deeplab

Method	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
MT[NIPS2017] [38]	51.72	58.93	63.86	69.51	70.96
CutMix-Seg[BMCV2017] [15]	52.16	63.47	69.46	73.73	76.54
PseudoSeg[ICLR2021] [52]	57.60	65.50	69.14	72.41	73.23
PC^2Seg[ICCV2021] [47]	57.00	66.28	69.78	73.05	74.15
CPS[CVPR2021] [7]	64.10	67.40	71.70	75.90	-
w/ UPC	**69.50** ↑5.40	**72.35** ↑4.95	**74.94** ↑3.24	**77.46** ↑1.56	-
ST++[CVPR2022] [41]	65.20	71.00	74.60	77.30	79.10
w/ UPC	**69.19** ↑3.99	**74.16** ↑3.16	**76.72** ↑2.12	**79.08** ↑1.78	**80.65** ↑1.55
U^2PL[CVPR2022] [39]	67.98	69.15	73.66	76.16	79.49
w/ UPC	**71.31** ↑3.33	**73.53** ↑4.37	**76.07** ↑2.41	**77.96** ↑1.80	**80.22** ↑0.73
Mask2Former[CVPR2022] [8]	64.85	67.53	71.39	74.23	78.71
w/ UPC	**69.25** ↑4.40	**75.52** ↑7.99	**78.91** ↑7.52	**81.06** ↑6.83	**83.31** ↑4.60

Table 1. Comparison with state-of-the-art methods on *classic* **PASCAL VOC 2012** `val` set under different partition protocols. The labeled images are selected from the original VOC `train` set, which consists of $1,464$ samples in total. The fractions denote the percentage of labeled data used for training, followed by the actual number of images. The results with UPC are emphasized with **bold** font, and the red upper arrow means improvement.

Method	1/16(662)	1/8(1323)	1/4(2646)	1/2 (5291)
MT[NIPS2017] [38]	70.51	71.53	73.02	76.58
CutMix-Seg[BMCV2019] [15]	71.66	75.51	77.33	78.21
CCT[ECCV2020] [23]	71.86	73.68	76.51	77.40
GCT[CVPR2020] [33]	70.90	73.29	76.66	77.98
AEL[NIPS2021] [21]	77.20	77.57	78.06	80.29
CPS[CVPR2021] [7]	74.48	76.44	77.68	78.64
w/ UPC	**76.61** ↑2.13	**77.80** ↑1.36	**78.53** ↑0.85	**79.35** ↑0.71
ST++[CVPR2022] [41]	74.50	76.30	76.60	-
w/ UPC	**77.65** ↑3.15	**78.62** ↑2.32	**79.47** ↑2.87	-
U^2PL[CVPR2022] [39]	77.21	79.01	79.30	80.50
w/ UPC	**78.53** ↑1.32	**79.92** ↑0.81	**80.36** ↑0.94	**81.05** ↑0.55
Mask2Former[CVPR2022] [8]	73.26	74.51	77.32	78.62
w/ UPC	**78.35** ↑5.09	**80.57** ↑6.06	**82.07** ↑4.75	**82.53** ↑3.91

Table 2. Comparison with state-of-the-art methods on *blender* **PASCAL VOC 2012** `val` set under different partition protocols. All labeled images are selected from the augmented VOC `train` set, which consists of $10,582$ samples in total.

v3+ [6] as the decoder, same as previous works. Main results reported in our paper are implemented based on previous representative semi-supervised segmentation works, CPS [7], ST++ [41] and U^2PL [39]. Our UPC serves as data augmentation method to rectify and enhance pseudo masks during training, **without changing their original architecture and training procedures.** Besides, we also apply UPC on Mask2Former [8] which has a strong ability to deal with the global context, to prove its generability.

Implementation Details. For the training on the *blender* and *classic* PASCAL VOC 2012 dataset on CPS, ST++ and U^2PL, we add UPC into data and pseudo label processing without changing network architecture and training hyper-parameters, just keeping the same as the original. When using Mask2Former, we use AdamW optimizer [25] with initial learning rate 0.0001, weight decay as 0.0001, crop size as 512×512, batch size as 16. We train Mask2Former with UPC in the self-training paradigm. We trained labeled data and the combination set for 30000 and 60000 steps respectively. Moreover, the semi-supervised training schedule is 120000 steps when using MSCOCO as unlabeled images.

For the training on the Cityscapes dataset, we keep the same setting but apply UPC on our candidate baselines. But for Mask2Former, we use AdamW optimizer with ini-

Method	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
SupOnly	65.74	72.53	74.43	77.83
ST++	-	72.70	73.80	-
w/ UPC	-	**75.36**↑2.66	**77.10**↑3.30	-
U^2PL	70.30	74.37	76.47	79.05
w/ UPC	**75.31**↑5.01	**77.35**↑2.98	**79.03**↑2.56	**79.62**↑0.57

Table 3. Comparison with state-of-the-art methods on **Cityscapes** `val` set under different partition protocols. All labeled images are selected from the Cityscapes `train` set, which consists of $2,975$ samples in total."SupOnly" stands for supervised training without using any unlabeled data.

tial learning rate 0.0001, weight decay as 0.0005, crop size as 1024×1024, batch size as 16 and training steps as 90000, slightly different from [8]. On both Pascal VOC and Cityscapes, UPC splits the unlabeled images into 4×4 patches and use $k = 2, 3, 5$ in RAS.

4.2. Comparison with state-of-the-art methods

We compare our method with recent semi-supervised semantic segmentation methods as the following, Mean Teacher (MT) [38], CCT [33], GCT [23], PseudoSeg [52], CutMix-Seg [15], CPS [7], PC^2Seg [47], AEL [21], U^2PL [39] and ST++ [41]. Specifically, we validate our UPC on CPS, ST++ and U^2PL and compare with their original performance, all using deeplab v3+ with resnet-101 backbone. When applying UPC on Mask2Former, we compare UPC with the original supervised Mask2Former to validate its effectiveness. Compared with different baseline methods, we provide detailed results on Pascal VOC 2012 and Cityscapes datasets.

Results on PASCAL VOC 2012 Dataset. We exhibit the comparision results between our method and the state-of-the-art methods on *classic* PASCAL VOC 2012 Dataset as Table 1. On three different competitive methods, our UPC improves them by a large margin on all protocols. Surprisingly, UPC improves CPS, ST++ and U^2PL by the most 5.40%, 3.99% and 4.37% respectively. Besides, the additional results on Mask2Former further show its great potential, improving at least 4.40% under all protocols.

Table 2 shows the comparison results on *blender* Pascal VOC datasets. Our proposed UPC improves four candidate baselines under all protocols obviously. Especially, our UPC improves ST++ by at least 2.32% under all protocols. The results further prove the effectiveness and generality of our UPC under various semi-supervised settings.

Results on Cityscapes Dataset. Table 3 shows the comparison results on the cityscapes. We apply our UPC on two recently published works, ST++ and U^2PL. On these two baselines, our method achieves consistent performance gains over all partition protocols. The improvement demonstrates the effectiveness and generalization of UPC in a more challenging dataset containing complicated image contexts.

Method	SupOnly	+MSCOCO	
		w/o label	w / label
PseudoSeg	76.96	78.20↑1.24	79.28↑2.32
UPC		**81.43**↑4.47	**82.41**↑5.45
UPC†	81.72	**84.93**↑3.21	**85.21**↑3.49

Table 4. Semi-supervised results using all Pascal VOC training set as labeled images while MSCOCO training set as unlabeled images. "SupOnly" stands for supervised training without using any unlabeled data. w/ label represents using image-level label for unlabeled images. "†" means implementation on Mask2Former.

Method	1/8 (1323)	1/4 (2646)
w/o	76.30	76.60
w/ Random Erasing	77.06	77.19
w/ Copy-Paste	77.73	78.06
w/ CutMix	77.55	77.82
w/ UPC (ours)	**78.62**↑2.32	**79.47**↑2.87

Table 5. The effectiveness of different denoising methods. Random Erasing,Copy-Paste and Cutmix are three competitors.

CutMix	Patch CutMix	RAS	mIoU
			76.60
✓			77.82
	✓		78.73
	✓	✓	**79.47**↑2.87

Table 6. Ablation study on different components of our UPC, including Patch CutMix and RAS.

Results using o.o.d unlabeled data. Table 4 provides comparison results with o.o.d unlabeled MSCOCO. Under this setting, we use the whole Pascal VOC training set as labeled images and the whole MSCOCO training setting as unlabeled images. Due to the lack of works trying this setting, we can only compare our UPC with PseudoSeg. We establish the simplest baseline on deeplab v3+ and Mask2Former respectively. On deeplab v3+ decoder, our method surpasses PseudoSeg by a large margin whether using image-level labels for unlabeled images or not. And UPC achieves 84.93% without image-level labels on Mask2Former. Those results demonstrate that UPC has robust generalization on o.o.d unlabeled images.

4.3. Ablations

To prove our core insight, *i.e.*, UPC is simple but effective to promote semi-supervised semantic segmentation as data augmentation, we conduct experiments about analyzing the effect of similar data augmentations, the effect of

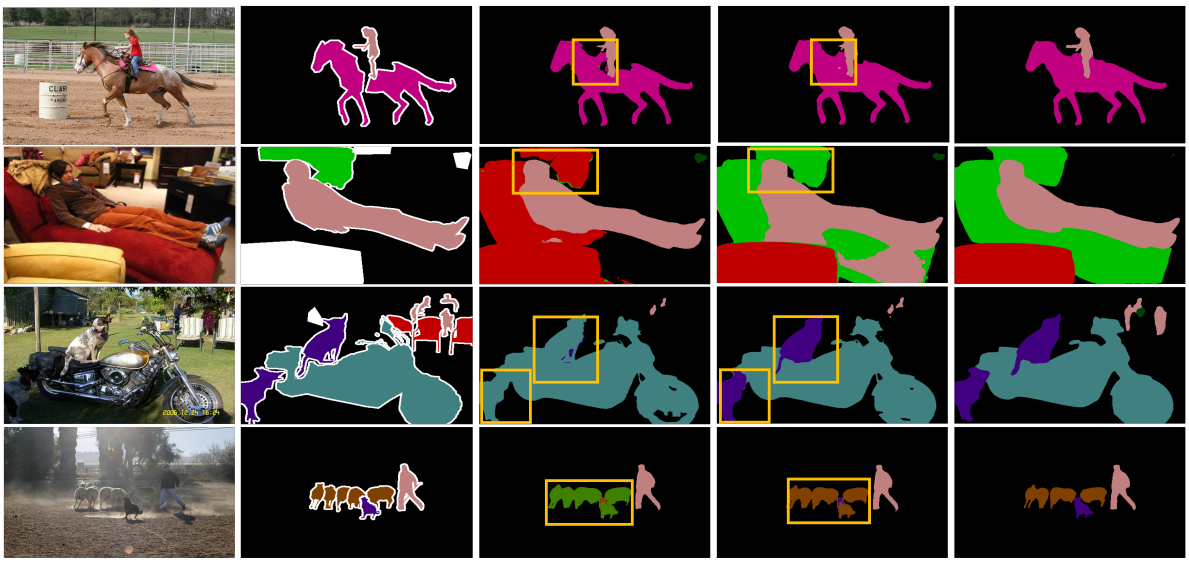

Figure 4. Qualitative results on Pascal VOC `val` set. col.(b) are human-annotated masks. The models of col.(c) and col.(d) are trained on 1464 labeled images. col.(c) is our supervised-only baseline, and col.(d) is semi-supervised baseline corresponding to "Full" protocol on *classic* Pascal VOC dataset. Additionally, we also provide results of our UPC trained on o.o.d MSCOCO, shown as col.(e) "UPC (w/ o.o.d)". Areas surrounded by orange boxes represent the improvement of UPC.

k	1	2	3	5
Random	76.66	76.90	77.53	77.85
UPC	77.41	77.95	78.47	78.73
Oracle	**77.71**	**78.45**	**78.82**	**79.24**

Table 7. Ablation study on the uncertainty guidance and k value. k refers to replacing k patches. Random means random selecting k patches and Oracle means using ground truth as guidance.

N \ k	1	2	3	5	7	9	11
3	77.73	78.12	78.24	**78.42**	78.04	-	-
4	77.41	77.95	78.47	**78.73**	77.81	77.35	76.83
5	76.96	77.37	77.86	78.02	78.61	**78.76**	78.13

Table 8. Ablation study on the effect of parameters N and k.

our proposed Patch CutMix and RAS. The following ablations are conducted under "$1/4$ (2646)" protocol on *blender* Pascal VOC dataset with ST++. The other settings keep unchanged if without claim.

Effectiveness of Denoising Methods. Table 5 shows the effectiveness and limitations of some random denoising methods on ST++. We compare multiple methods including Random Erasing [48], CopyPaste [36], and CutMix [43]. Random Erasing randomly removes a region on an unlabeled image and its pseudo mask. Meanwhile, CutMix and CopyPaste prefer to erase one region and replace it with one from labeled images. Note that baseline work ST++ already introduced noise-aware loss in their work, but using CutMix directly on ST++ cannot bring obvious improvement. Compared with competitors, UPC achieves highest improvement due to its accurate denoising function in a patch-wise way.

Effectiveness of Patch CutMix and RAS. Table 6 shows results comparison between results with CutMix and Patch CutMix as row 1 and row 2. Compared to simple CutMix, Patch CutMix brings more improvement. It is obvious that uncertain region guidance helps UPC a lot to realize accurate denoising. Moreover, we ablate our proposed strategy RAS which generates more than one transferred image. In Table 6, it can be witnessed that RAS brings further improvement over single Patch CutMix.

Effectiveness of Uncertainty Guidance and Ablation Study on Hyper-parameters. Table 7 answers the question. One is the benefit of uncertainty guidance is not clear, another is what k value makes Patch CutMix the best. For the first question, we design two control groups, Random and Oracle, selecting substitute patches by random and IoU with ground truth respectively. Uncertainty surpasses the Random with large margin and has a small gap to the Oracle under different k value. For another question, when $k = 5$ setting, the UPC (Uncertainty) performs the best. Finally, we use multiple k value $2, 3, 5$ in RAS to further improve UPC, resulting in higher performance.

Ablations on N and k of Patch CutMix. We further ablate the influence of parameters N and k on UPC. Table 8 presents the results obtained by varying the values of k and N in the "Patch CutMix" approach. Specifically, we explore the impact of k by considering values ranging from

1 to 11, and the effect of N by varying it between 3 and 5. The results show larger k does not result in a sustained improvement. We find that k should be increased with a larger N and UPC will not get sustained improvement with a larger k and N. Based on our experiments, the default setting of $N = 4$ and $k = 5$ achieves nearly the best performance. Furthermore, Table 8 shows the top-2, top-3, and top-5 perform well under $N = 4$ settings.

4.4. Qualitative visualization

Fig. 4 shows the results of different methods on Pascal VOC `val` set. Compared with the supervised-only method, UPC generates better predictions with correct class prediction results and smooth boundaries. Besides the "Full" protocol setting, the results from the model trained using o.o.d unlabeled MSCOCO also demonstrate these features.

5. Conclusion

In this paper, we propose a simple yet effective data augmentation mechanism, named Uncertainty-aware Patch CutMix (UPC), to advance semi-supervised semantic segmentation. For unlabeled images, we predict the uncertainty map by segmentation model. To accurately locate the uncertain region and maintain context information, we use Patch Split to estimate the regional uncertainty. Those regions with high uncertainty scores are then replaced with certain ones cut out from labeled data. With UPC, the training process of semi-supervised segmentation algorithm becomes more robust to the noise from pseudo masks, obtaining competitive results over other methods under various partition protocols accordingly. More importantly, our UPC shows a strong generalization ability to the unlabeled out-of-distribution data. We hope our UPC can serve as a solid training strategy and help ease future research on semi-supervised semantic segmentation.

Acknowledgement. This work is sponsored by the National Key R&D Program of China (No.2021ZD0112100), National NSF of China (No.U1936212, No.62120106009).

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 2, 3

[3] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920, 2021. 1

[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 *European Conference on Computer Vision*, pages 695–714. Springer, 2020. 2, 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 3

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 6

[7] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 1, 4, 6, 7

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1, 6, 7

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 5

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 3

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 2, 5

[14] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. *arXiv preprint arXiv:2004.08514*, 1(2):5, 2020. 1, 2

[15] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 1, 2, 3, 4, 6, 7

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and

Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[17] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004. 1, 2

[18] Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, and Yunchao Wei. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[19] 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. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5

[21] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. 1, 2, 4, 6, 7

[22] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 2, 3

[23] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. pages 429–445, 2020. 2, 6, 7

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6728–6736, 2019. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020. 3

[27] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. pages 1205–1214, 2021. 1, 3

[28] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2, 3

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 5

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[31] 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*, 43(4):1369–1379, 2019. 1, 3

[32] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021. 3

[33] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1, 2, 3, 6, 7

[34] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 2

[35] 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. *Advances in neural information processing systems*, 33:596–608, 2020. 3

[36] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 3, 8

[37] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*, pages 5688–5696, 2017. 3

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1, 6, 7

[39] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 2, 3, 4, 5, 6, 7

[40] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 3

[41] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022. 2, 3, 6, 7

[42] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic seg-

mentation with strong data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8229–8238, 2021. 2, 3

[43] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3, 4, 8

[44] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3

[45] Zekang Zhang, Guangyu Gao, Zhiyuan Fang, Jianbo Jiao, and Yunchao Wei. Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. *Advances in Neural Information Processing Systems*, 35:24340–24353, 2022. 3

[46] Zekang Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Coinseg: Contrast inter- and intra- class representations for incremental segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[47] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 6, 7

[48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 8

[49] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R Manmatha, Mu Li, and Alexander Smola. Improving semantic segmentation via self-training. *arXiv preprint arXiv:2004.14960*, 2020. 2

[50] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020. 3

[51] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 3

[52] 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. 3, 5, 6, 7