Decoupling with Entropy-based Equalization for Semi-Supervised Semantic Segmentation

Chuanghao Ding[1,2,5,*] , **Jianrong Zhang**[1,3] , **Henghui Ding**[4] , **Hongwei Zhao**[1,3,†] ,
Zhihui Wang[5] , **Tengfei Xing**[5,†] and **Runbo Hu**[5]
[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[2] College of Software, Jilin University, China
[3] College of Computer Science and Technology, Jilin University, China
[4]Nanyang Technological University, Singapore
[5]Didi Chuxing, China

Abstract

Semi-supervised semantic segmentation methods are the main solution to alleviate the problem of high annotation consumption in semantic segmentation. However, the class imbalance problem makes the model favor the head classes with sufficient training samples, resulting in poor performance of the tail classes. To address this issue, we propose a **D**ecoupled **S**emi-**S**upervised **S**emantic **S**egmentation (**DeS**[4]) framework based on the teacher-student model. Specifically, we first propose a decoupling training strategy to split the training of the encoder and segmentation decoder, aiming at a balanced decoder. Then, a non-learnable prototype-based segmentation head is proposed to regularize the category representation distribution consistency and perform a better connection between the teacher model and the student model. Furthermore, a Multi-Entropy Sampling (MES) strategy is proposed to collect pixel representation for updating the shared prototype to get a class-unbiased head. We conduct extensive experiments of the proposed **DeS**[4] on two challenging benchmarks (PASCAL VOC 2012 and Cityscapes) and achieve remarkable improvements over the previous state-of-the-art methods.

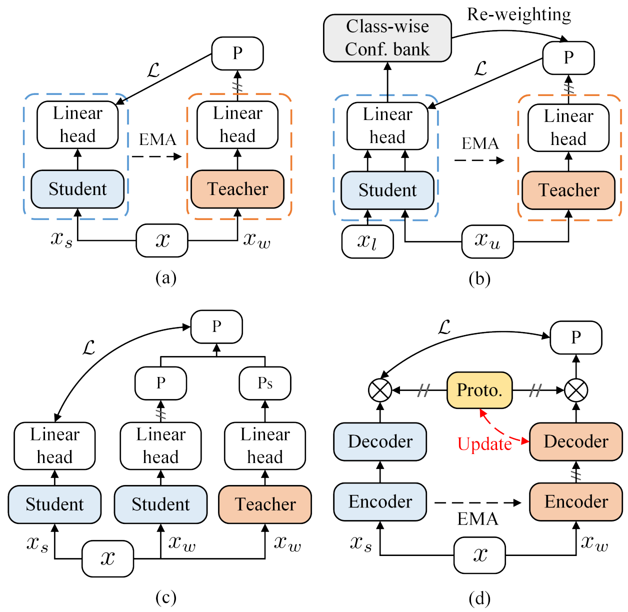


Figure 1: Comparison with existing semi-supervised semantic segmentation methods for class-imbalance learning. (a) Common Teacher-Student framework [French *et al.*, 2020], (b) Combination of re-sampling and re-weighting [Hu *et al.*, 2021], (c) Additional unbiased subclass regularization networks [Guan *et al.*, 2022], (d) Decoupling balance training network (Ours). " // " on "→" means stop-gradient. P: pseudo labels. Proto.: share semantic prototype.

1 Introduction

Semantic segmentation is one of the most fundamental tasks in the computer vision field, it can be applied in many applications like autonomous vehicles and movie editing. In recent years, remarkable progress has been made in semantic segmentation based on Deep Neural Networks [He *et al.*, 2016; Chen *et al.*, 2018a] as well as large-scale well-annotated segmentation datasets [Everingham *et al.*, 2015; Cordts *et al.*, 2016]. Existing fully-supervised deep-learning-based segmentation methods are data-hungry and require large-scale datasets for training. It is however very time-consuming and labor-intensive to obtain segmentation datasets because they are dense annotations of pixel-wise masks. To alleviate this high annotation consumption issue, semi-supervised semantic segmentation has been widely concerned [French *et al.*, 2020; Zou *et al.*, 2021; Chen *et al.*, 2021], it offers the potential of leveraging limited annotations and a large set of unlabeled images.

Many semi-supervised segmentation efforts aim at applying consistency regularization [French *et al.*, 2020; Chen *et al.*, 2021; Zhang *et al.*, 2022] and self-training [Bachman *et al.*, 2019; Chen *et al.*, 2020; Fan *et al.*, 2022a] strategies. These approaches typically employ the teacher-student

[*]Interns at Didi Chuxing
[†]Corresponding author

paradigm [French *et al.*, 2020] and supervise the student model by the pseudo label generated by the teacher model, as shown in Figure 1 (a). However, since models are trained using imbalanced data, most methods are limited by the pixel-wise classification accuracy of the semantic segmentation, which leads to the degradation of tailed categories learning. Recently, a few works attempt to alleviate the imbalance problem in semi-supervised semantic segmentation [Guan *et al.*, 2022; Hu *et al.*, 2021]. For example, Distribution Alignment and Random Sampling (DARS) [He *et al.*, 2021] and UCC [Fan *et al.*, 2022a] explore the mismatch problem between the true distribution and pseudo-labeled distribution, and propose a progressive data augmentation strategy and Dynamic Cross-Set Copy-Paste (DCSCP), respectively. AEL [Hu *et al.*, 2021] tackles the biased training problem with re-sampling and re-weighting, as shown in Figure 1 (b). It proposes two adaptive-based data augmentation methods and a sampling strategy for the confidence bank. Differently, USRN [Guan *et al.*, 2022] presents a class-balance subclass framework with clustered subclasses, which is illustrated in Figure 1 (c). However, existing methods learn the encoder and decoder jointly and such a learning fashion ignores the impact of the long-tailed problem on different components.

In this work, inspired by a recent successful imbalanced semi-supervised classification algorithm [Fan *et al.*, 2022b], we propose **D**ecoupled **S**emi-**S**upervised **S**emantic **S**egmentation (**DeS**^{4}) as an imbalanced semi-supervised semantic segmentation framework, as shown in Figure 1 (d). In the proposed **DeS**^{4}, we decouple the encoder and pixel-level representation (from the decoder) for long-tail semantic segmentation. Specifically, the training of the encoder and segmentation decoder are decoupled without gradient propagation, and we aim to get a robust encoder and unbiased segmentation decoder. Under the teacher-student pattern, we connect the student model and teacher model via a shared segmentation head for exchanging unbiased information between the two models, which is based on non-learnable prototypes rather than relying only on pseudo-label supervision. Besides, we propose a Multi-Entropy Sampling (MES) strategy to update the unbiased prototype non-parametrically. The entropy level of the category-wise representation distribution is divided into several zones, and balance subsampling is conducted for each zone of entropy level. The proposed MES strategy greatly improves the diversity of the category-wise representation while maintaining the balance property. Furthermore, we utilize the sampled category representations to update the prototype via exponential moving average (EMA). Then, the pixel representations find the nearest prototype of the same category with metric learning for classification.

We outperform other methods on two widely used datasets: PASCAL VOC 2012 [Everingham *et al.*, 2015] and Cityscapes [Cordts *et al.*, 2016]. For example, our method achieves 81.61% and 80.64% on the VOC `Aug` dataset under 1/2 and 1/4 partitions, which shows an improvement of 2.31% and 1.63% over the previous state-of-the-art methods.

In summary, this paper makes the following contributions:

- We propose **D**ecoupled **S**emi-**S**upervised **S**emantic **S**egmentation (**DeS**^{4}) as an imbalanced semi-supervised semantic segmentation framework, in which we separate the training of the encoder and decoder.
- We propose non-learnable prototypes as a shared and balanced segmentation head, which links the teacher model and the student model better. Meanwhile, a Multi-Entropy sampling strategy is proposed for updating prototypes in a balanced manner.
- We outperform existing state-of-the-art semi-supervised semantic segmentation methods on two public datasets consistently.

2 Related Works

Semantic segmentation. Fully Convolutional Network [Long *et al.*, 2015] learns dense features effectively in an end-to-end fashion. Since it was a pioneering work, several enhancements were proposed based on FCN from various aspects, e.g. enhancing the receptive field [Chen *et al.*, 2018a], incorporating multi-scale contextual features [Chen *et al.*, 2016; Zhao *et al.*, 2017; Ding *et al.*, 2018], and investigating attention operations [Fu *et al.*, 2019; Ding *et al.*, 2019]. Besides, significant improvements in semantic segmentation in recent years have been made by stronger backbone architectures, such as ResNet [He *et al.*, 2016] in CNN-based methods, and ViT [Dosovitskiy *et al.*, 2020] in Transformer-based methods. Currently, many efforts have been made for exploring long-range dependency with Transformers in the segmentation head [Xie *et al.*, 2021; Cheng *et al.*, 2021; Zheng *et al.*, 2021; Ding *et al.*, 2021], which showed remarkable results.

Semi-supervised semantic segmentation. Semi-supervised semantic segmentation methods pay attention to training by combining labeled images with unlabeled images, which reduces the time-consuming of manual annotation. Previous approaches investigate the generative adversarial networks (GANs) [Hung *et al.*, 2018] for unlabeled data via discriminating pseudo labels. Recently, several works motivated by the remarkable progress in semi-supervised learning based on consistency regularization [Chen *et al.*, 2021; Wang *et al.*, 2022; Zhang *et al.*, 2022] and self-training [Lee, 2013; Fan *et al.*, 2022a]. For example, GCT [Ke *et al.*, 2020] enforces consistency between two models with different initializations but the same architecture. PseudoSeg [Zou *et al.*, 2021] introduces Grad-CAM for better quality pseudo-labels. CPS [Chen *et al.*, 2021] proposes dual parallel models and performs cross-model supervision for the training of semantic segmentation networks. Furthermore, many works benefited from learning pixel-level representations with unsupervised contrastive learning. PC^{2}Seg [Zhong *et al.*, 2021] enforces label-space consistency regularization and feature contrastive property. U^{2}PL [Wang *et al.*, 2022] selects pixels based on their reliability and pushes away unreliable samples. RC^{2}L [Zhang *et al.*, 2022] encourages region-level consistency and contrastive properties to solve the false-negative problem and simplify the contrast learning training process. Besides, many efforts [Hu *et al.*, 2021;

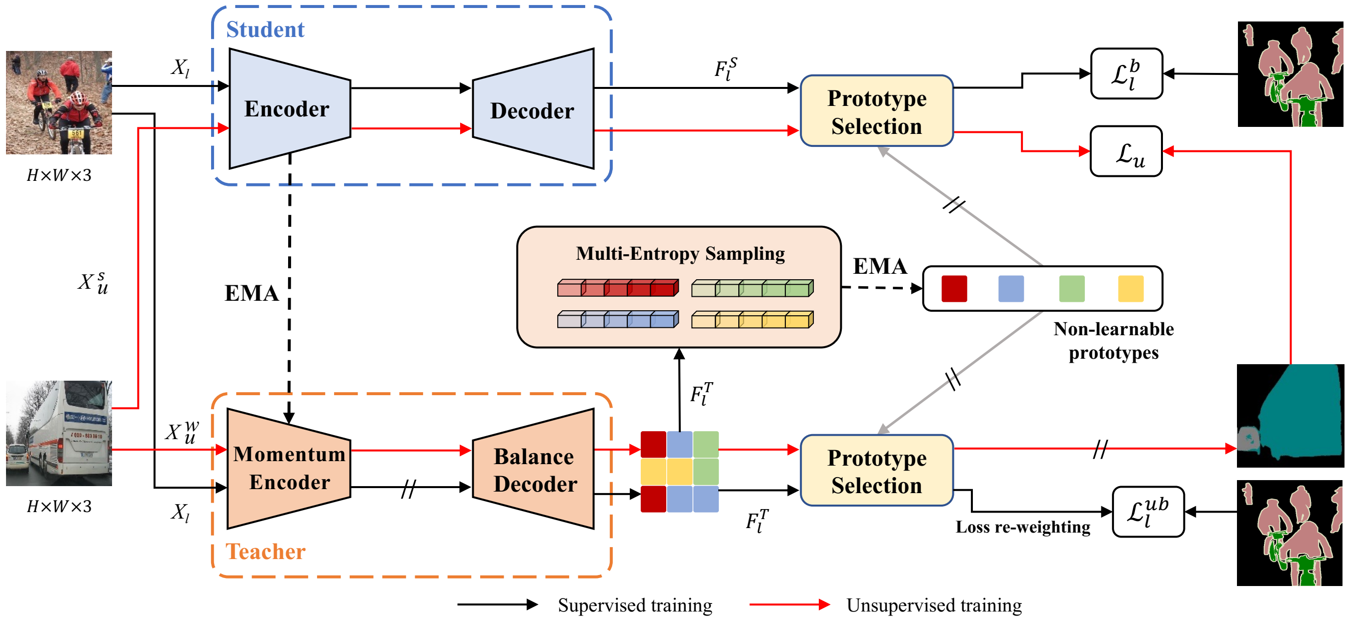


Figure 2: Illustrate the architecture of our framework. H and W are the height and width of the input image. X_l, X_u^w, and X_u^s denote the labeled image set, unlabeled image set with weak augmentation, and unlabeled image set with strong augmentation, respectively. " // " on "→" means no gradient update happen.

Guan *et al.*, 2022] have been devoted to overcoming the pixel class imbalance issue. AEL [Hu *et al.*, 2021] proposes adaptive data augmentation methods and sampling strategy, USRN [Guan *et al.*, 2022] trains an unbiased subclass classifier to regularize imbalanced pseudo-labels and designs a gate module based on the entropy. UCC [Fan *et al.*, 2022a] proposes Dynamic Cross-Set Copy-Paste (DCSCP) strategy to address the misalignment and class imbalance problem.

In this work, we first propose a decoupled training strategy for semantic segmentation in a semi-supervised fashion. It decouples the training of the encoder and the decoder. Second, different from [Guan *et al.*, 2022] which performs K-Means clustering and uses prototypes as additional class centers, we raise a non-learnable prototype-based classifier for both the teacher model and the student model, and we also propose a novel balance sampling strategy for the prototype updating.

Class-imbalance learning. Class-imbalance learning is a fundamental problem that has been widely studied. Many works attempt to tackle the class imbalance problem via loss function re-weighting. For example, Focal loss [Lin *et al.*, 2017] adjusts the loss weight of each sample to suit different class labels for training data, resulting in much more noise from the dataset. There are also some works that obtain re-sampled data with a balanced number of training samples via random linear interpolation [Chawla *et al.*, 2002], multi-stage training [Yin *et al.*, 2019], *etc*. Besides, SPE [Liu *et al.*, 2020] proposes a self-paced ensemble strategy with re-sampling to balance the dataset effectively. Recent efforts demonstrate that decoupling the representation and classifier [Kang *et al.*, 2019; Tang *et al.*, 2020] can be beneficial for long-tailed classification. Inspired by these approaches, CoSSL [Fan *et al.*, 2022b] proposes a Co-Learning framework for imbalanced semi-supervised classification.

Different from SPE [Liu *et al.*, 2020], we employ a multi-entropy sampling on various categories rather than binary classification. Meanwhile, CoSSL [Fan *et al.*, 2022b] links the teacher model and student model via pseudo-label only, we propose a shared non-learnable prototype as a bridge to transfer class-unbiased information to the student and unify the category-wise embedding space for both the teacher model and student model.

3 Method

We first describe our framework in Section 3.1, **D**ecoupled **S**emi-**S**upervised **S**emantic **S**egmentation (**DeS**^{4}) framework. Then, the detailed decoupled training is introduced in Section 3.2. In Section 3.3, we develop the Multi-Entropy Sampling strategy, which considers the number of inter-class samples and the entropy of intra-class samples at the same time.

3.1 Overview

The overall architecture is illustrated in Figure 2. Our method has two training procedures: supervised training and unsupervised training. Specifically, given a labeled image set $X_l = \{(x_l^i, y_l^i)\}_{i=1}^{N_l}$ and $x_l \in \mathbb{R}^{H\times W\times 3}$, $y_l \in \mathbb{R}^{H\times W}$ for supervised training where H and W represent the height and the width of the image, N_l is the size of labeled dataset. We employ the commonly used teacher-student framework, which has two models with the same architecture. Denoting S as the student model, and the pixel-level feature map can be computed as $F_l^S = S(X_l)$. On the other hand, we also feed X_l into the teacher model T. We get the feature map F_l^T from the teacher model following $F_l^T = T(X_l)$.

We propose to use a shared prototype-based classifier $P = [p_1, p_2, ..., p_C] \in \mathbb{R}^{C\times d}$ for predicting the probability (for more details, see Section 3.2), where C is the total number of classes of the dataset, and d is the dimension of prototypes. Each pixel feature in F_l^S and F_l^T is projected in the same category as the nearest prototype class centers, and then minimizes the cross-entropy loss. The supervised loss can be formulated as:

$$\mathcal{L}_l^b = \frac{1}{N_l}\sum_{i=1}^{N_l}\mathcal{L}_{ce}(\arg\min\{\langle S(x_l^i), p_j\rangle\}_{j=1}^C, y_l^i), \quad (1)$$

$$\mathcal{L}_l^{ub} = \frac{1}{N_l}\sum_{i=1}^{N_l}\mathcal{L}_{ce}(\arg\min\{\langle T(x_l^i), p_j\rangle\}_{j=1}^C,\ y_l^i), \quad (2)$$

where $\langle \dots \rangle$ denotes the distance measure, $\mathcal{L}_{ce}$ denotes the cross-entropy loss.

For unlabeled image set $X_u = \{x_u^i\}_{i=1}^{N_u}$, where N_u is the number of the unlabeled images, we perform weakly augmentation (*e.g.* random flip, random crop, *etc.*) and strong augmentation (consists of all weak augmentation approaches and CutMix) to obtain X_u^w and X_u^s, respectively. As shown in Figure 2, pseudo labels $Y_u = \{y_u^i\}_{i=1}^{N_u}$ are generated by the teacher model to supervise the training of the student. The loss of the unsupervised branch can be written as:

$$\mathcal{L}_u = \frac{1}{N_u}\sum_{i=1}^{N_u}\mathcal{L}_{ce}(\arg\min\{\langle S(x_u^i)), p_j\rangle\}_{j=1}^C,\ y_u^i). \quad (3)$$

Optimization goal. To optimize our model, the total loss function consists of three components: a biased supervised loss $\mathcal{L}_l^b$, an unbiased supervised loss $\mathcal{L}_l^{ub}$, and an unsupervised loss $\mathcal{L}_u$. Total loss can be written as:

$$\mathcal{L} = \mathcal{L}_l^{ub} \ + \ \lambda_1\mathcal{L}_l^b \ + \ \lambda_2\mathcal{L}_u, \quad (4)$$

where λ_1 and λ_2 are hyper-parameters to balance losses.

3.2 Decoupled Semi-Supervised Semantic Segmentation

Decoupling for long-tailed classification is proposed in [Kang *et al.*, 2019], which demonstrates that only adjusting a classifier is possible to get a good performance. We apply the decoupling to semi-supervised semantic segmentation for the first time, and propose a shared classifier based on non-learnable prototypes to better connect the teacher and student models.

Supervised training procedure. We represent the encoder and decoder of the teacher model in terms of E^T and D^T. Different from previous approaches [Hu *et al.*, 2021; Guan *et al.*, 2022; Zhang *et al.*, 2022] , only the model weights of E^T are exponential moving average (EMA) updated by the weights of the student model's encoder,

$$\theta^t = \tau_1\theta^t + (1-\tau_1)\theta^s, \quad (5)$$

where θ^t and θ^s donate the model parameters of E^T and student's encoder, respectively, and $\tau_1 \in [0, 1]$ is a constant to control the exponential moving.

We find that separating the training of the encoder and the decoder with the classifier achieves better performance than only adjusting the classifier separately (experimental results are provided in Section 4). So **DeS**^{4}, with a goal of learning a class-unbiased segmentation decoder and classifier for the teacher model.

First, for the classifier, the learnable prototype is equivalent to a linear classifier, which is hard to maintain the balanced property with gradient propagation and ignores the inductive bias of the feature distribution. Recent work [Zhou *et al.*, 2022] explores a non-learnable prototype-based method for semantic segmentation. We propose shared non-learnable prototypes as class centers, which can represent the feature space of each class $c \in \{1, ..., C\}$. We propose a novel pixel-level feature sampling strategy to update prototypes in a balanced manner (for more details, see Section 3.3). More specifically, given a pixel latent feature f, classify through prototypes is to find the nearest element in P with `argmin` operation, as shown in Eq. (1)-(3), where cosine similarity is used as a distance measure: $\langle u, v\rangle = \frac{u^{\mathrm{T}}v}{\|u\|\|v\|}$.

Denoting the probability distribution of pixel latent feature as: $p(c|f) = \frac{\exp(S_f^c)}{\sum_{c'=1}^C \exp(S_f^{c'})}$, where S_f^c is defined as the similarity between f and closest prototype of category c. We optimize the log-likelihood of the distribution:

$$\mathcal{L}_{ce} = \mathbb{E}_{c\in C}[-\log p(c|f)]. \quad (6)$$

As to the training of the unbiased decoder, we introduce the recently successful pixel-level loss re-weighting. For i-th image, the loss weight can be computed as:

$$\mathcal{W}_i = \frac{(1-\arg\max(\sigma(z_{x_l^i,p})))^2}{sum((1-\arg\max(\sigma(z_{x_l^i,p})))^2)}, \quad (7)$$

where $sum(\cdot)$ stands for sum operation, σ denotes `Softmax`, and $z_{x_l^i,p} = \arg\min\{\langle D^T(E^T(x_l^i)), p_j\rangle\}_{j=1}^C$. Then we update Eq. (2) as:

$$\mathcal{L}_l^{ub} = \frac{1}{N_l}\sum_{i=1}^{N_l}\mathcal{W}_i\mathcal{L}_{ce}(z_{x_l^i,p}, y_l^i). \quad (8)$$

Note that gradient updates only happen for the teacher's balanced decoder via $\mathcal{L}_l^{ub}$.

Pseudo label supervision. Previous approaches simply generate pseudo-labels as a signal for information interaction through the teacher model. However, it is not enough for relying only on pseudo labels, due to two reasons: 1). Ignoring the different feature spaces between the teacher and student. 2). Pseudo labels cannot pass unbiased information to the student model. We tackle this question through the shared and balanced prototype-based classifier in both supervised and unsupervised training. As normal, the momentum encoder and class-unbiased decoder extract pixel-level representations from unlabeled images. Then, pseudo labels are generated via prototype-based metric learning updated by Multi-Entropy Sampling (Section 3.3) against data imbalance. Furthermore, the student model also gets predictions via the same prototype with balance property and guarantees identical feature space. This enhances the information

Method	VOC Train					VOC Aug			
	1/2(732)	1/4(366)	1/8(183)	1/16(92)	1.4k(1464)	1/2(5291)	1/4(2646)	1/8(1323)	1/16(662)
MT [Tarvainen and Valpola, 2017]	69.16	63.01	55.81	48.70	-	77.61	76.62	73.20	70.59
VAT [Miyato *et al.*, 2018]	63.34	56.88	49.35	36.92	-	-	-	-	-
AdvSemSeg [Hung *et al.*, 2018]	65.27	59.97	47.58	39.69	68.40	-	-	-	-
CCT [Ouali *et al.*, 2020]	62.10	58.80	47.60	33.10	69.40	77.56	76.17	73.00	67.94
GCT [Ke *et al.*, 2020]	70.67	64.71	54.98	46.04	-	77.14	75.25	73.30	69.77
CutMixSeg [French *et al.*, 2020]	69.84	68.36	63.20	55.58	-	75.89	74.25	72.69	72.56
PseudoSeg [Zou *et al.*, 2021]	72.41	69.14	65.50	57.60	73.23	-	-	-	-
CPS [Chen *et al.*, 2021]	75.88	71.71	67.42	64.07	-	78.64	77.68	76.44	74.48
PC^{2}Seg [Zhong *et al.*, 2021]	73.05	69.78	66.28	57.00	74.15	-	-	-	-
AEL [Hu *et al.*, 2021]	-	-	-	-	-	80.29	78.06	77.57	77.20
RC^{2}L [Zhang *et al.*, 2022]	77.06	72.24	68.87	65.33	79.33	80.43	79.71	77.49	75.56
U^{2}PL [Wang *et al.*, 2022]	76.16	73.66	69.15	67.98	79.49	80.50	79.30	79.01	77.21
Supervised baseline	71.69	65.88	54.92	45.77	72.50	77.13	75.80	71.55	67.87
Ours	**77.62**	**74.58**	**72.23**	**68.02**	**80.86**	**82.11**	**81.61**	**81.02**	**77.28**

Table 1: Comparison with the state-of-the-art methods on VOC 2012 `Val` set. The supervised baseline is trained only with labeled images. We follow [Ouali *et al.*, 2020; Ke *et al.*, 2020; Zhang *et al.*, 2022] to use the encoder pretrained on COCO [Lin *et al.*, 2014] for 1/8 and 1/16 VOC `Train`. For the rest, we use the backbone pretrained on ImageNet [Deng *et al.*, 2009].

exchange with no gradient, resulting in a more balanced student model and further leading to a stronger and more robust momentum encoder.

3.3 Multi-Entropy Sampling

Previous strategies mainly sample pixel features randomly or rely on the confidence of feature softmax probability distribution. The former is more likely to choose high-confidence samples, and the latter prefers to explore low-confidence samples, leading to the noise-raised problem due to the different learning difficulties of each category. Inspired by SPE [Liu *et al.*, 2020], we propose a Multi-Entropy Sampling strategy that selects features by considering both the entropy balance of intra-class samples and the quantitative balance of inter-class samples at the same time.

Precisely, we first account for the number of pixels in each category in the whole labeled image set, obtaining the ratio $R = [r_1, r_2, ..., r_C]$ among categories and employing the normalization based on the category c^l ($r_{c^l} = 1$) with the lowest number. Then, we calculate the amount of per category samples $N^b = [n_1^b, n_2^b, ..., n_C^b]$ in a batch. In the training step of a batch, the number of samples per category can be calculated as $\hat{N}^b = R \cdot N^b$ to encourage inter-class quantity balance property.

As normal, for intra-class balance, the entropy can be formulated as:

$$\text{Entropy}(z) = -\sum_{i=1}^{C} z_i \log(z_i + \epsilon), \tag{9}$$

where z stands for the predicted probability distribution, and ϵ is a constant which is set to $1e-10$. Differently, it is not reasonable to focus totally on low-confidence or high-confidence (low entropy or high entropy) samples for prototype-based classifier updating. Furthermore, we propose a multi-entropy-based method for each category rather than selecting randomly. Given a set of pixel feature $F_l^{\hat{c}} = [f_1^{\hat{c}}, f_2^{\hat{c}}, ..., f_{\hat{n}_{\hat{c}}^b}^{\hat{c}}]$, and the corresponding entropy predictions $E^{\hat{c}} = [e_1^{\hat{c}}, e_2^{\hat{c}}, ..., e_{\hat{n}_{\hat{c}}^b}^{\hat{c}}]$ of the category $\hat{c}$ from the teacher model. We split these samples into k zones according to their entropy as the Multi-Entropy, which can be formulated as:

$$Z_i = \{[E_{\min}^{\hat{c}} + (i-1) \times L,\ \ E_{\min}^{\hat{c}} + i \times L)\}_{i=1}^{k}. \tag{10}$$

Denoting that $E_{\min}^{\hat{c}}$ and $E_{\max}^{\hat{c}}$ are the min and max entropy in $E^{\hat{c}}$. k is a hyper-parameter to control the number of zones. $[\cdot)$ represents the interval of left closed and right open. L denotes the length of the zone: $\frac{E_{\max}^{\hat{c}} - E_{\min}^{\hat{c}}}{k}$.

We calculate the entropy-based sampling ratio of each zone as follows:

$$Z_i^{ent} = \frac{1}{\alpha + \sum_{j \in Z_i} e_j^{\hat{c}} / |Z_i|}. \tag{11}$$

Here, α is a constant to biasing sampling more towards difficult samples. Finally, we conduct under-sampling for $F_l^{\hat{c}}$ in i-th zone with the number of $N_{\hat{c}}^i = \frac{Z_i^{ent}}{\sum_{i \in k} Z_i^{ent}} \times \hat{n}_{\hat{c}}^b$, which aims to have balanced entropy in zones. We define the under-sampling set as $F_{sample}^{\hat{c}} = [f_1^{\hat{c}}, f_2^{\hat{c}}, ..., f_{\sum_{i \in k} N_{\hat{c}}^i}^{\hat{c}}]$. Note that if there are insufficient samples in a zone, we sample from the neighboring zones to ensure balanced entropy as much as possible.

With sampled pixel feature set available, prototypes can be updated via EMA. The process can be formulated as follow:

$$p_j = \tau_2 p_j + (1 - \tau_2) GAP(F_{sample}^{\hat{c}}), \tag{12}$$

where GAP indicates the global average pooling, $\tau_2 \in [0, 1]$ is a hyper-parameter to control the exponential moving. Note that no gradient propagation happens at updating non-learnable prototypes.

4 Experiment

4.1 Experimental Setup

Datasets. PASCAL VOC 2012 [Everingham *et al.*, 2015] is the most widely used benchmark dataset in semi-supervised

Method	1/4(744)	1/8(372)
CutMixSeg [French *et al.*, 2020]	68.33	65.82
CMB [Alonso *et al.*, 2021]	65.9	64.4
CPS [Chen *et al.*, 2021]	74.58	74.31
PseudoSeg [Zou *et al.*, 2021]	72.36	69.81
PC^2Seg [Zhong *et al.*, 2021]	75.15	72.29
AEL [Hu *et al.*, 2021]	77.48	75.55
USRN [Guan *et al.*, 2022]	-	75.0
RC^2L [Zhang *et al.*, 2022]	76.47	74.04
U^2PL [Wang *et al.*, 2022]	76.47	74.37
Supervised baseline	74.43	72.53
Ours	**77.87**	**75.74**

Table 2: Comparison with the state-of-the-art methods on Cityscapes `Val` set. We report the results on 1/4 and 1/8 Cityscapes dataset with the model pretrained on the COCO dataset.

semantic segmentation with a background category and 20 foreground categories. The original dataset consists of 1464 images for training and 1449 images for evaluation. Several works combine the coarse annotated images with the original train set (VOC `Train`) to get the augmented dataset (VOC `Aug`) for training. Following common practice, we evaluate our proposed model in both two settings. Cityscapes [Cordts *et al.*, 2016] is a high-resolution urban scene dataset with a total of 19 classes. We follow previous works [Zhang *et al.*, 2022] to select 1/4 and 1/8 training images as labeled data.

Evaluation metrics. We report mean Intersection-over-Union (mIoU) as the evaluation metric. All the experimental results are evaluated on either the VOC `Val` set or the Cityscapes `Val` set, and ablation studies are conducted on the 1/4 and 1/8 VOC `Aug` dataset.

Implementation details. We use DeepLab v3+ [Chen *et al.*, 2018b] as the semantic segmentation network with the ResNet101 backbone. All experiments are trained on 8 NVIDIA RTX A6000 GPUs with a batch size of 16, and we use stochastic gradient descent (SGD) to optimize the model, and set balance weights λ_1 and λ_2 to 1 and 1.5. Empirically, we set both the EMA decay of τ_1 and τ_2 to 0.99. For the Multi-Entropy Sampling strategy, we set k and α to 5 and 0.1, respectively. For both PASCAL VOC `Train` and `Aug` datasets, the initial learning rate is set to 0.001, and the weight decay is 0.0001. We follow previous settings [Wang *et al.*, 2022] to train our model for 80 epochs with the crop size of 513 $\times$ 513. For the Cityscapes dataset, the initial learning rate is 0.01, weight decay is 0.0006, and the crop size is 769 $\times$ 769. Furthermore, we employ a poly learning rate policy that the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with $power = 0.9$.

4.2 Comparison to the State-of-the-Arts

We compare **DeS**^{4} to state-of-the-art methods (*e.g.* [Hu *et al.*, 2021; Wang *et al.*, 2022; Zhang *et al.*, 2022], *etc.*) on VOC `Val` set and Cityscapes `Val` set.

Results on PASCAL VOC 2012. We show the comparison results in Table 1. For VOC `Train` set, our proposed **DeS**^{4} outperforms existing state-of-the-art method, for example, we achieve the improvements of 0.56% and 0.92% with partition protocols of 1/2 and 1/4, and significantly outperform U^2PL [Wang *et al.*, 2022] on 1/8 partition protocol with 3.08%. As to VOC `Aug` set, we obtain the improvements of 1.61%, 2.31%, 2.01%, and 0.07% under 1/2, 1/4, 1/8, and 1/16 partitions compared with U^2PL, respectively. We also conduct the experiment under a 1.4k/9k split, where all the original VOC 2012 dataset is used as labeled data and the augmented dataset is used as unlabeled data. The performance of **DeS**^{4} is 1.37% and 1.53% higher than U^2PL and RC^2L [Zhang *et al.*, 2022], respectively.

Results on Cityscapes. Experimental results on Cityscapes `val` set are shown in Table 2. Our method improves the supervised baseline by 3.44% and 3.21% under 1/4 and 1/8 partitions. We also outperform recent state-of-the-art approaches. In particular, **DeS**^{4} outperforms AEL [Hu *et al.*, 2021], RC^2L [Zhang *et al.*, 2022], and U^2PL, with improvements of 0.39%, 1.4%, and 1.4% under 1/4 partition, and 0.19%, 1.7%, and 1.37% under 1/8 partition.

D.T.	S.H.	P.R.	VOC `Aug` (1/4)	VOC `Aug` (1/8)
			78.23	76.73
✓			80.44	79.85
✓	✓		80.47	80.28
✓	✓	✓	**81.61**	**81.02**

Table 3: Impact of various components, where D.T., S.H., and P.R. stand for 'Decoupled Training', 'Shared segmentation Head', and 'Pixel Re-weighting', respectively.

BSS	VOC `Aug` (1/4)	VOC `Aug` (1/8)
Quantity balance	81.10	80.50
Confidence balance	81.17	80.66
ME balance	**81.61**	**81.02**

Table 4: Study on sampling strategies for updating prototypes, where 'BSS' denotes the balance sampling strategy.

4.3 Ablation Study

Investigating each component. We first investigate the impact of three vital components of **DeS**^{4}. The results are provided in Table 3. The first line is the baseline, which combines MeanTeacher [Tarvainen and Valpola, 2017] and CutMixSeg [French *et al.*, 2020] based on the prototype classifier. It can be seen that adopting the decoupled training strategy improves the baseline method by 2.21% and 3.12%. We find that the shared prototype-based segmentation head leads to slightly better performance under 1/4 VOC `Aug`, but achieves an improvement of 0.43% under 1/8 partition. Pixel re-weighting improves "Baseline + Decoupled Training + Shared segmentation head" significantly, obtaining the improvement of 1.14% and 0.74% under 1/4 and 1/8 partition protocols, respectively. The studies demonstrated the effectiveness of these three components.

Impact of balanced sampling strategy. In this subsection, we study the effect of our proposed Multi-Entropy Sampling. The quantity balance denotes that we randomly sample the same number of pixel representations for each category, and

Method	back.	aero.	bicy.	bird	boat*	bott.*	bus	car	cat	chair	cow*	table	dog	horse	motor	pers.	plant*	sheep*	sofa	train	tv*
Supervised	89.9	73.6	33.8	75.1	42.0	54.4	80.0	75.8	78.9	24.7	50.2	43.1	72.6	50.2	68.2	77.2	34.9	64.8	30.6	67.6	55.1
DARS [He *et al.*, 2021]	91.3	82.6	37.4	81.9	50.5	58.6	88.5	82.9	82.8	25.5	56.3	49.1	75.3	64.6	73.6	79.7	42.2	64.0	37.1	73.4	57.9
USRN [Guan *et al.*, 2022]	91.9	84.1	36.1	84.9	52.8	66.4	87.9	81.8	86.4	26.5	75.2	58.6	83.0	73.3	74.7	80.2	40.7	76.2	42.0	78.5	59.8
Ours	96.1	87.9	40.3	82.0	68.2	54.5	93.2	85.1	83.5	34.5	87.6	53.3	78.1	85.3	84.0	85.5	48.2	82.9	49.2	86.3	66.8

Table 5: Quantitative comparisons of **DeS**^{4} with other class-imbalance learning methods under VOC `Aug` 1/32 partition protocol. Red and Blue indicate the best and the second-best result. The class name marked in "*" is the tailed-class.

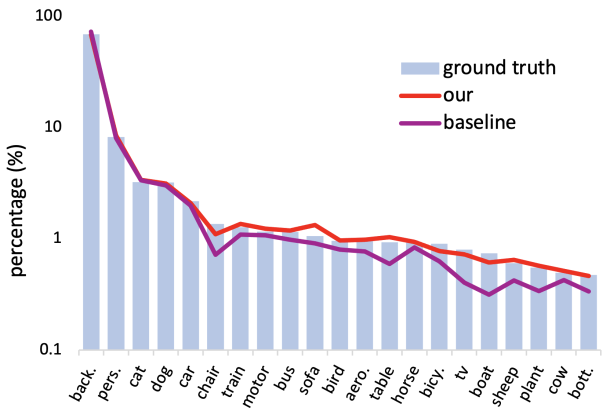


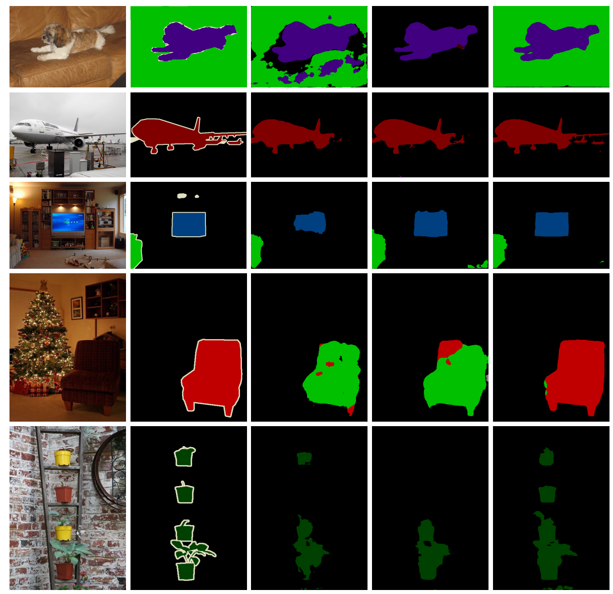
Figure 3: Comparison of the class distribution among ground truth, semi-supervised baseline, and our method on 1/2 VOC `Aug` unlabeled dataset.

the confidence balance is based on the quantity balance while class-wise representation sampling is driven by the softmax distribution. As shown in Table 4, it can be seen that MES achieves the best performance on both 1/4 VOC `Aug` and 1/8 VOC `Aug` sets. This suggests that the MES can generate a more general class clustering center which leads to better performance.

Analysis of the class imbalance problem. We present experimental results on the class imbalance problem. The class distribution on 1/2 VOC `Aug` unlabeled dataset is shown in Figure 3. To be fair, we also provide a per-class comparison with USRN [Guan *et al.*, 2022] and DARS [He *et al.*, 2021] in Table 5 under VOC `Aug` 1/32 partition protocol. Our method outperforms the previous state-of-the-art on tailed-classes.

4.4 Visualization

Figure 4 shows visual results on PASCAL VOC 2012 `Val` set [Everingham *et al.*, 2015], and the model is trained on 1.4k/9k split. We present more visual results associated with tail classes to demonstrate the superiority of our approach. One can see that our **DeS**^{4} corrects more wrong predictions compared to the supervised baseline and the semi-supervised baseline. For example, some pixels are mistakenly classified in the 4th row of (c) and (d). Both the supervised baseline and the semi-supervised baseline have the mislabeling issue in the 1st row and 5th row. Besides, our method has better segmentation boundaries for foreground objects, which are shown in the 2nd and 3rd rows.



(a) (b) (c) (d) (e)

Figure 4: Visual results on PASCAL VOC 2012 `Val` set. (a) original image, (b) ground truth, (c) supervised baseline, (d) semi-supervised baseline, (e) ours.

5 Conclusion

We developed a Decoupled Semi-Supervised Semantic Segmentation (**DeS**^{4}) framework. We proposed to decouple the training of the encoder and decoder to achieve a balanced segmentation decoder of the teacher model. Then, we proposed a shared non-learnable prototype-based classifier to connect and unify the category-wise embedding space of the teacher model and student model. Furthermore, the Multi-Entropy Sampling strategy is presented to update the shared prototype non-parametrically for a class-unbiased classifier of the teacher model. Experimental results demonstrated that our method achieved better performance than previous state-of-the-art methods.

Contribution Statement

Chuanghao Ding, Jianrong Zhang, and Henghui Ding designed the method and wrote the paper (equal contribution). Hongwei Zhao, Zhihui Wang, and Tengfei Xing analyzed experimental results and provided valuable comments. Hongwei Zhao and Runbo Hu are the project leaders that supervised this work.

Acknowledgments

This research was funded by the Provincial Science and Technology Innovation Special Fund Project of Jilin Province (20190302026GX), the Natural Science Foundation of Jilin Province (20200201037JC), and the Fundamental Research Funds for the Central Universities for Jilin University.

References

[Alonso *et al.*, 2021] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 8219–8228, 2021.

[Bachman *et al.*, 2019] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *arXiv preprint arXiv:1906.00910*, 2019.

[Chawla *et al.*, 2002] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. In *Journal of artificial intelligence research*, pages 321–357, 2002.

[Chen *et al.*, 2016] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3640–3649, 2016.

[Chen *et al.*, 2018a] 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. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 834–848, 2018.

[Chen *et al.*, 2018b] 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.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *arXiv preprint arXiv:2002.05709*, 2020.

[Chen *et al.*, 2021] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, 2021.

[Cheng *et al.*, 2021] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 17864–17875, 2021.

[Cordts *et al.*, 2016] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[Ding *et al.*, 2018] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2393–2402, 2018.

[Ding *et al.*, 2019] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8885–8894, 2019.

[Ding *et al.*, 2021] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 16321–16330, 2021.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *arXiv preprint arXiv:2010.11929*, 2020.

[Everingham *et al.*, 2015] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. In *International Journal of Computer Vision (IJCV)*, pages 98–136, 2015.

[Fan *et al.*, 2022a] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9947–9956, 2022.

[Fan *et al.*, 2022b] Yue Fan, Dengxin Dai, Anna Kukleva, and Bernt Schiele. Cossl: Co-learning of representation and classifier for imbalanced semi-supervised learning. In *cvpr*, pages 14574–14584, 2022.

[French *et al.*, 2020] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[Fu *et al.*, 2019] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[Guan *et al.*, 2022] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Unbiased subclass regularization for semi-supervised semantic segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9968–9978, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[He *et al.*, 2021] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6930–6940, 2021.

[Hu *et al.*, 2021] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 22106–22118, 2021.

[Hung *et al.*, 2018] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[Kang *et al.*, 2019] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *arXiv preprint arXiv:1910.09217*, 2019.

[Ke *et al.*, 2020] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 429–445, 2020.

[Lee, 2013] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning Workshops (ICMLW)*, page 896, 2013.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, James Hays Serge Belongie, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[Liu *et al.*, 2020] Zhining Liu, Wei Cao, Zhifeng Gao, Jiang Bian, Hechang Chen, Yi Chang, and Tie-Yan Liu. Self-paced ensemble for highly imbalanced massive data classification. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 841–852, 2020.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[Miyato *et al.*, 2018] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1979–1993, 2018.

[Ouali *et al.*, 2020] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12674–12684, 2020.

[Tang *et al.*, 2020] K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1513–1524, 2020.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1195–1204, 2017.

[Wang *et al.*, 2022] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4248–4257, 2022.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12077–12090, 2021.

[Yin *et al.*, 2019] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5704–5713, 2019.

[Zhang *et al.*, 2022] Jianrong Zhang, Tianyi Wu, Chuanghao Ding, Hongwei Zhao, and Guodong Guo. Region-level contrastive and consistency learning for semi-supervised semantic segmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1622–1628, 2022.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[Zheng *et al.*, 2021] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021.

[Zhong *et al.*, 2021] 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 International Conference on Computer Vision (ICCV)*, pages 7273–7282, 2021.

[Zhou *et al.*, 2022] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2582–2593, 2022.

[Zou *et al.*, 2021] 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 *International Conference on Learning Representations (ICLR)*, 2021.