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

Chuanghao Ding\textsuperscript{1,2,5,*}, Jianrong Zhang\textsuperscript{1,3}, Henghui Ding\textsuperscript{4}, Hongwei Zhao\textsuperscript{1,3,†}, Zhihui Wang\textsuperscript{5}, Tengfei Xing\textsuperscript{5,†} and Runbo Hu\textsuperscript{5}

\textsuperscript{1}Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
\textsuperscript{2}College of Software, Jilin University, China
\textsuperscript{3}College of Computer Science and Technology, Jilin University, China
\textsuperscript{4}Nanyang Technological University, Singapore
\textsuperscript{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 Decoupled Semi-Supervised Semantic Segmentation (DeS\textsuperscript{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\textsuperscript{4} on two challenging benchmarks (PASCAL VOC 2012 and Cityscapes) and achieve remarkable improvements over the previous state-of-the-art methods.

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., 2022] strategies. These approaches typically employ the teacher-student
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 Decoupled Semi-Supervised Semantic Segmentation (DeSS) as an imbalanced semi-supervised semantic segmentation framework, as shown in Figure 1 (d). In the proposed DeSS, 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 Decoupled Semi-Supervised Semantic Segmentation (DeSS) 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. PC2Seg [Zhong et al., 2021] enforces label-space consistency regularization and feature contrastive property. U2PL [Wang et al., 2022] selects pixels based on their reliability and pushes away unreliable samples. RC2L [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;
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. Sec-
ond, 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 clas-
sification. 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, Decoupled
Semi-Supervised Semantic Segmentation (DeS²) framework.
Then, the detailed decoupled training is introduced in Sec-
tion 3.2. In Section 3.3, we develop the Multi-Entropy Sam-
pling strategy, which considers the number of inter-class sam-
ple s 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 unsu-
ervised training. Specifically, given a labeled image set
\( X_l = \{ (x_l^i, y_l^i) \}_{i=1}^{N_l} \) and \( x_l \in \mathbb{R}^{H \times W} \), \( y_l \in \mathbb{R}^{H \times W} \) 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 training procedures: supervised training and
unsupervised training. 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 $F^S_l$ and $F^T_l$ 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:

$$L^b_i = \frac{1}{N} \sum_{i=1}^{N_i} \mathcal{L}_{ce}(\arg\min\{\langle S(x^i_l), p_j \rangle \}_{j=1}^{C}),$$

$$L^{ub}_i = \frac{1}{N} \sum_{i=1}^{N_i} \mathcal{L}_{ce}(\arg\min\{\langle T(x^i_l), p_j \rangle \}_{j=1}^{C}),$$

where $\langle \ldots \rangle$ 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:

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

**Optimization goal.** To optimize our model, the total loss function consists of three components: a biased supervised loss $L^b_i$, an unbiased supervised loss $L^{ub}_i$, and an unsupervised loss $L_u$. Total loss can be written as:

$$\mathcal{L} = L^{ub}_i + \lambda_1 L^b_i + \lambda_2 L_u,$$

where $\lambda_1$ and $\lambda_2$ 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,

$$\theta_t = \tau_t \theta^t + (1 - \tau_t) \theta^s,$$

where $\theta^t$ and $\theta^s$ denote the model parameters of $E^T$ and student’s encoder, respectively, and $\tau_t \in (0, 1]$ is a constant to control the exponential moving.

We find that separating the training of the encoder and decoder with the classifier achieves better performance than only adjusting the classifier separately (experimental results are provided in Section 4). So DeS3, 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 inducitive bias of the feature distribution. Recent work [Zhou et al., 2021] 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 $\arg\min\text{ operation, as shown in Eq. (1)-(3), where cosine similarity is used as a distance measure: } \langle u, v \rangle = \frac{u^T v}{\|u\|\|v\|}.$

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

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

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:

$$W_i = \frac{(1 - \arg\max\{\sigma(z_{xi,p}\lambda)\})^2}{\sum((1 - \arg\max\{\sigma(z_{xi,p}\lambda)\})^2),}$$

where $\sum(\cdot)$ stands for sum operation, $\sigma$ denotes Softmax, and $z_{xi,p} = \arg\min\{\langle D^T(E^T(x_i^T)), p_j \rangle \}_{j=1}^{C}$.

Then we update Eq. (2) as:

$$L^{ub}_i = \frac{1}{N} \sum_{i=1}^{N_i} W_i \mathcal{L}_{ce}(z_{xi,p}, y_i).$$

Note that gradient updates only happen for the teacher’s balanced decoder via $L^{ub}_i$.

**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
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 in zones and employing the nor-

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].

<table>
<thead>
<tr>
<th>Method</th>
<th>VOC Train</th>
<th>VOC Aug</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/2(732)</td>
<td>1/4(366)</td>
</tr>
<tr>
<td>MT [Tarvainen and Valpola, 2017]</td>
<td>69.16</td>
<td>63.01</td>
</tr>
<tr>
<td>VAT [Miyato et al., 2018]</td>
<td>63.34</td>
<td>56.88</td>
</tr>
<tr>
<td>AdvSemSeg [Hung et al., 2018]</td>
<td>65.27</td>
<td>59.97</td>
</tr>
<tr>
<td>CCT [Ouali et al., 2020]</td>
<td>62.10</td>
<td>58.80</td>
</tr>
<tr>
<td>GCT [Ke et al., 2020]</td>
<td>70.67</td>
<td>64.71</td>
</tr>
<tr>
<td>CutMixSeg [French et al., 2020]</td>
<td>69.84</td>
<td>68.36</td>
</tr>
<tr>
<td>PseudoSeg [Zou et al., 2021]</td>
<td>72.41</td>
<td>69.14</td>
</tr>
<tr>
<td>CPS [Chen et al., 2021]</td>
<td>75.88</td>
<td>71.71</td>
</tr>
<tr>
<td>PC^2Seg [Zhong et al., 2021]</td>
<td>73.05</td>
<td>69.78</td>
</tr>
<tr>
<td>AEL [Hu et al., 2021]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RC^2L [Zhang et al., 2022]</td>
<td>77.06</td>
<td>72.24</td>
</tr>
<tr>
<td>U^2PL [Wang et al., 2022]</td>
<td>76.16</td>
<td>73.66</td>
</tr>
<tr>
<td>Supervised baseline</td>
<td>71.69</td>
<td>65.88</td>
</tr>
<tr>
<td>Ours</td>
<td>77.62</td>
<td>74.58</td>
</tr>
</tbody>
</table>

4 Experiment

4.1 Experimental Setup

Datasets. PASCAL VOC 2012 [Everingham et al., 2015] is the most widely used benchmark dataset in semi-supervised...
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 [Cords 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 $\lambda_1$ and $\lambda_2$ to 1 and 1.5. Empirically, we set both the EMA decay of $\tau_1$ and $\tau_2$ to 0.99. For the Multi-Entropy Sampling strategy, we set $k$ and $\alpha$ 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 $(1 - \frac{\text{iter}}{\text{max}\text{iter}})^{\text{power}}$ with $\text{power} = 0.9$.

### 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$^2$PL [Wang et al., 2022] on 1/8 partition protocol with 3.08%.

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

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.

We first investigate the impact of these three components. 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

### 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.

<table>
<thead>
<tr>
<th></th>
<th>D.T.</th>
<th>S.H.</th>
<th>P.R.</th>
<th>VOC Aug (1/4)</th>
<th>VOC Aug (1/8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>78.23</td>
<td>76.73</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.44</td>
<td>79.85</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.47</td>
<td>80.28</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>81.61</strong></td>
<td><strong>81.02</strong></td>
</tr>
</tbody>
</table>

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

![Graph showing class distribution comparison](image)

**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.

**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 (202002010373JC), and the Fundamental Research Funds for the Central Universities for Jilin University.

References


