Self-Decoupling and Ensemble Distillation for Efficient Segmentation

Yuang Liu, Wei Zhang*, Jun Wang*

School of Computer Science and Technology, East China Normal University
{frankliu624, zhangwei.thu2011, wongjun}@gmail.com

Abstract

Knowledge distillation (KD) is a promising teacher-student learning paradigm that transfers information from a cumbersome teacher to a student network. To avoid the training cost of a large teacher network, the recent studies propose to distill knowledge from the student itself, called Self-KD. However, due to the limitations of the performance and capacity of the student, the soft-labels or features distilled by the student barely provide reliable guidance. Moreover, most of the Self-KD algorithms are specific to classification tasks based on soft-labels, and not suitable for semantic segmentation. To alleviate these contradictions, we revisit the label and feature distillation problem in segmentation, and propose Self-Decoupling and Ensemble Distillation for Efficient Segmentation (SDES). Specifically, we design a decoupled prediction ensemble distillation (DPED) algorithm that generates reliable soft-labels with multiple expert decoders, and a decoupled feature ensemble distillation (DFED) mechanism to utilize more important channel-wise feature maps for encoder learning. The extensive experiments on three public segmentation datasets demonstrate the superiority of our approach and the efficacy of each component in the framework through the ablation study.

Introduction

Semantic segmentation (SS) has been a longstanding challenge in computer vision, which is the foundation of numerous advanced intelligent applications, such as image editing (Zhu et al. 2020), scene understanding (Zheng et al. 2021) and automatic pilot (Levinson et al. 2011; Huang et al. 2018). As a dense prediction task, it aims to assign a semantic label for each pixel in an image. Thanks to the renaissance of deep learning, the approaches based on fully convolutional networks (FCN) (Long, Shelhamer, and Darrell 2015) have been the mainstream of semantic segmentation and achieved remarkable performance. In addition to utilizing deeper backbone networks (Simonyan and Zisserman 2015; He et al. 2016; Huang et al. 2017; Xie et al. 2017), most of the works focus on preserving the global information with context aggregation, and introduce some efficient modules, *e.g.*, pyramid pooling module (Zhao et al. 2017), dilated convolution (Chen et al. 2017a), self-attention (Fu et al. 2019). However, these complex and strong architectures usually have heavy parameters and require overwhelmed computation, which limit the deployment of segmentation networks on resource-constrained mobile devices. In contrast, designing efficient segmentation networks has attracted increasing attentions due to their balance between performance and cost, *e.g.*, ENet (Paszke et al. 2016), ERFNet (Romera et al. 2017), ESNet (Lyu et al. 2019). Compared with hand-crafted lightweight model design, knowledge distillation (KD) is a more popular and general solution to model compression and stimulating the potential small models.

Currently, most of the KD methods (Zagoruyko and Komodakis 2017; Peng et al. 2019; Liu et al. 2019; Shu et al. 2021; Liu, Zhang, and Wang 2022) are with teacher-student learning architectures (abbreviated as T-S KD), as shown in Figure 1(a), in which the compact student network is trained with the supervision of the soft-labels or intermediate features from the well-trained teacher. T-S KD terribly suffers from a two-stage training scheme and tremendous computation and memory cost of the cumbersome teacher. And this issue is further amplified in complex tasks like SS. To alleviate this issue, some researchers (Zhang et al. 2019; Yuan et al. 2020; Kim et al. 2021; Zhang et al. 2021a; Hou et al. 2019; Ji et al. 2021) investigate to distill information from the student network itself, which is called Self-KD, as shown in Figure 1(b). However, most of the Self-KD methods are specific to classification models and depend on image-level soft-labels or cross-layer features. The soft-labels, constructed manually (Szegedy et al. 2016; Yuan et al. 2020) or obtained from the multi-exit architecture (Zhang et al. 2019, 2021a), often contain uncertain information and cannot substitute the output by strong teachers. In fact, the label-based Self-KD methods are usually regarded as a specific label smoothing regularization (LSR) in some researches (Mobahi, Farajtabar, and Bartlett 2020; Kim et al. 2021; Zhang and Sabuncu 2020). But this mechanism is not suitable for SS, which is a pixel-level classification task and depends on context information instead of independent image labels. Zhang *et al*. (Zhang et al. 2021a) extended the multi-exit Self-KD architecture to SS networks but bring limited improvements. In contrast to label-based Self-KD, some works (Hou et al. 2019; Ji et al. 2021) fo-

*Corresponding authors.

cus on refining or distilling feature knowledge cross layers, which only brings tiny improvement due to contradictory constraints between the front and rear layers.

Therefore, in this paper, to address the issues of involuted T-S KD and insufficient Self-KD in SS, we revisit the distillation problem and propose a novel self-decoupled and ensemble distillation framework tailed for efficient segmentation (SDES). The core idea of our method is to construct reliable soft-labels and enhanced features to provide rich information for the decoder and backbone of the student, respectively. These two kinds of knowledge both are distilled from the student itself, which can avoid complicated teacher-student learning and tickle label smoothing. In particular, we introduce two expert modules to decouple and integrate information, following the decoder and backbone respectively, as shown in Figure 1(c). On the one hand, a decoupled prediction ensemble distillation (DPED) algorithm is developed to generate reliable soft-labels from multiple expert decoders with disentangled class-wise information and reduced learning complexity. The expert decoders can work as strong teachers in terms of the decoupled labels. On the other hand, we introduce a channel-wise decoupled feature ensemble distillation (DFED) mechanism to enhance the global attention for self-regulation of the backbone. The feature maps with richer information are selected from each layer and aggregated as an expert-level attention map to supervise the whole backbone. With these insights, we implement our SDES framework and conduct extensive experiments on three public segmentation datasets to evaluate the effectiveness of the framework. In fact, it combines the advantages of both previous T-S KD and Self-KD, balancing model performance and training efficiency.

In a nutshell, our main contributions are as follows:

- We investigate the class-disentangled segmentation problem and propose a decoupled prediction ensemble distillation (DPED) algorithm for reliable soft-labels generation.
- A decoupled feature ensemble distillation (DFED) scheme is developed for attention enhancement with channel-wise selection at multiple layers. To the best of our knowledge, this is the first research effort to exploit decoupled predictions and features for Self-KD on SS.
- Without the assistance of cumbersome teacher networks, our SDES framework can make the compact student SS networks achieve very competitive performance in one-stage distillation, as compared to both previous T-S KD and Self-KD methods.

Related Work

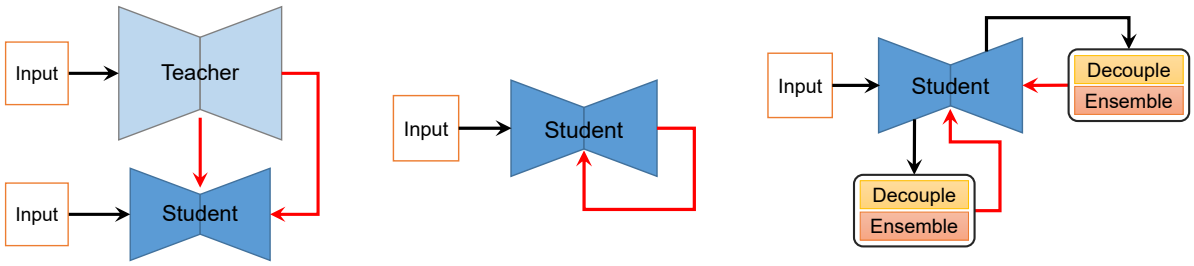
Semantic Segmentation (SS)

Recent works in semantic segmentation are mainly based on fully convolutional networks (FCN) (Long, Shelhamer, and Darrell 2015) and gain improvements through receptive field expansion and context aggregation. Badrinarayanan *et al*. (Badrinarayanan, Kendall, and Cipolla 2017) extended the VGG (Simonyan and Zisserman 2015) model to an encoder-decoder architecture to refine the latent features extracted from the backbone. PSPNet (Zhao et al. 2017) exploits the capability of global context information by context aggregation through pyramid pooling module. The DeepLab series (Chen et al. 2015, 2017a,b) introduce dilated convolutions and multi-scale features to enlarge receptive field. Moreover, some works (Yu et al. 2018; Fu et al. 2019; Huang et al. 2019) focus on obtaining full-image contextual information through attention mechanisms. Recently, some researchers are starting to explore Transformer-based segmentation networks (Xie et al. 2021; Strudel et al. 2021), which can mine global dependency information by self-attention.

Meanwhile, designing highly efficient segmentation networks is a promising direction for practical deployment. In addition to adopt compact classification backbone (*e.g*., MobileNet (Sandler et al. 2018), ShuffleNet (Ma et al. 2018)), some researchers dedicate to designing real-time computation architectures, such as ENet (Paszke et al. 2016), ERFNet (Romera et al. 2017), ESNet (Lyu et al. 2019), and so on. Beyond the above works, we focus on boosting the performance of compact segmentation networks via knowledge distillation approach.

Knowledge Distillation (KD)

Since KD was proposed as a teacher-student learning architecture by Hinton *et al*. (Hinton, Vinyals, and Dean 2015), it has been a popular model compression way in visual classification (Park et al. 2019; Xie et al. 2020; Zhang et al. 2020) and other tasks (Li et al. 2022; Porrello, Bergamini, and Calderara 2020; Weinzaepfel et al. 2020). According to whether a teacher network is participant, most of the KD methods can be roughly divided into two categories (Wang and Yoon 2020; Gou et al. 2021): T-S KD and Self-KD. Obviously, the T-S KD methods (Zagoruyko and Komodakis 2017; Peng et al. 2019; Chen et al. 2020; Liu, Zhang, and Wang 2022) aim to transfer information from the well-trained teacher to the compact student, which is the mainstream. The main idea of Self-KD is to distill soft-labels (Zhang et al. 2019; Yuan et al. 2020; Mobahi, Farajtabar, and Bartlett 2020; Zhang et al. 2022; Kim et al. 2021; Zhang et al. 2021a) or features (Hou et al. 2019; Ji et al. 2021; Li 2022) from the student itself, without a teacher network. Some studies (Zhang et al. 2019, 2021a) introduce a multi-exit architecture to boost the student itself, while the studies (Zhang et al. 2022; Kim et al. 2021) propose to refine the soft-targets for the student. In addition, the works (Hou et al. 2019; Ji et al. 2021) aim to guide the intermediate layers of the student with attention mechanism or refined features. However, most of the Self-KD methods are tailored for classification tasks (Zhang et al. 2019; Yuan et al. 2020; Mobahi, Farajtabar, and Bartlett 2020; Ji et al. 2021) and not suitable for SS (Zhang et al. 2021a). Meanwhile, there are tiny KD methods for segmentation (Liu et al. 2019; Shu et al. 2021; Liu, Zhang, and Wang 2022) that all rely on high-cost teacher networks. Zhang *et al*. (Zhang et al. 2021a) extended the multi-exit architecture to SS, which is the first effort of the Self-KD in SS task. LSR (Szegedy et al. 2016) regularizes model training by replacing the one-hot labels with smoothed ones, which can provide a smoothing distribution to make the model avoid over-confidence. It can be



(a) T-S KD (b) Self-KD (c) Ours

Figure 1: Overview of different kinds of knowledge distillation methods.

regarded as a teacher-free / self-boosting method. LSR is often associated with KD in some theoretical researches (Yuan et al. 2020; Müller, Kornblith, and Hinton 2019; Mobahi, Farajtabar, and Bartlett 2020; Yun et al. 2020; Shen et al. 2021; Lukasik et al. 2020). Our method is teacher-free but can guide the student reliably with the two expert modules like T-S KD.

Methodology

Preliminary

Semantic segmentation is mainly formulated as a pixel-wise dense classification problem that aims to assign an individual semantic label from C classes to each pixel in an image. The segmentation network is usually decomposed of a backbone/encoder and a decoder, which can be denoted as E and D, respectively. With an RGB image $X \in \mathbb{R}^{H\times W\times 3}$ as input, the corresponding prediction map output by the segmentation network is $P = D \circ E(X)$. H and W denote the height and width of both the input image and prediction map. The loss function $\mathcal{L}_{\text{SS}}$ of the segmentation task is formulated as pixel-wise cross-entropy with the ground-truth label map Y:

$$\mathcal{L}_{\text{SS}} = -\frac{1}{H\times W}\sum_{i}^{H\times W}\sum_{j}^{C} Y_{i,j}\log\sigma(P_i)_j\,, \qquad (1)$$

in which $Y_{i,j}$ represents the j-th one-hot value of the i-th pixel, P_i is the categorical logits of the i-th pixel. $\sigma(\cdot)$ denotes the softmax function, and the j-th class probability of the i-th pixel can be calculated by $\sigma(P_i)_j = \frac{\exp(P_{i,j})}{\sum_{c=1}^{C}\exp(P_{i,c})}$.

Figure 2 depicts the proposed SEDS framework, consisting of the DPED and DFED modules that work as experts for soft-label and feature distillation, respectively. We will elaborate them in the following.

Decoupled Prediction Ensemble Distillation (DPED)

Motivation. To obtain an excellent teacher network, current T-S KD methods (Zagoruyko and Komodakis 2017; Liu et al. 2019; Shu et al. 2021) mainly pre-train a stronger backbone with more parameters and computation. In general classification, we empirically find that if a classification network takes charge fewer or a part of categories, it can easily achieve better performance. In view of the multi-class and pixel-annotated label maps for SS, we derive that an entire semantic label map can be decomposed of more subset-class even one-class maps. This can debias the class-wise dependency and imbalanced category distribution. With these insights, we argue that multiple perfect sub-area experts can be trained to guide the student network with reliable soft-labels.

We construct N expert decoders $\{D_1^T, D_2^T, \cdots, D_N^T\}$ following the weight-shared encoder E and in parallel with the student decoder D. As for the architectures of the expert decoders, they can be very lightweight with tiny parameters as discussed in experiment part. Different from D, the expert decoders are trained to classify a subset of the fully semantic categories, which can reduce the entanglement among classes in one image. Intuitively, allocating fewer categories (*e.g.*, one category) to each expert decoder could ensemble more accurate soft-labels. However, this needs more decoders and increases the computation and memory usage in training. Moreover, over-mighty experts could cause big gap with the student model which is not conducive to the soft-label distillation. To this end, we empirically apply "more to less" subset split strategy to each expert decoder. In particular, we divide the entire label set $\Psi = \{1, 2, \cdots, C\}$ with C semantic categories into N subsets, $\Psi = \Psi_1 \cup \Psi_2 \cup \cdots \cup \Psi_N, N \ll C$. And there is no intersection between any two subsets.

Additionally, to alleviate the class imbalance problem in each semantic map, we group the categories with a similar number of pixels into a subset, which is motivated by the long-tailed classification methods (Li, Wang, and Wu 2021; Zhang et al. 2021b). We rank and split the N subsets according to pixel-wise class cardinality (more to less), which indicates that these subsets become less imbalanced than the original.

To allocate the ground truth label map with subset categories for each expert decoder D_n^T, we calculate a mask matrix $M^n \in \mathbb{R}^{H\times W}$ about the whole label map Y. The i-th element in M^n is defined as

$$M_i^n = \begin{cases} 1\,, & \text{if } \texttt{argmax}(Y_i) \in \Psi_n \\ 0\,, & \text{otherwise} \end{cases}\,, i \in \{1, 2, \cdots, H\times W\}\,. \qquad (2)$$

Then the new subset label map to train the decoder D_n^T can be obtained by taking the dot product of the corresponding mask M^n and the original label map Y: $Y^n = M^n \odot Y$.

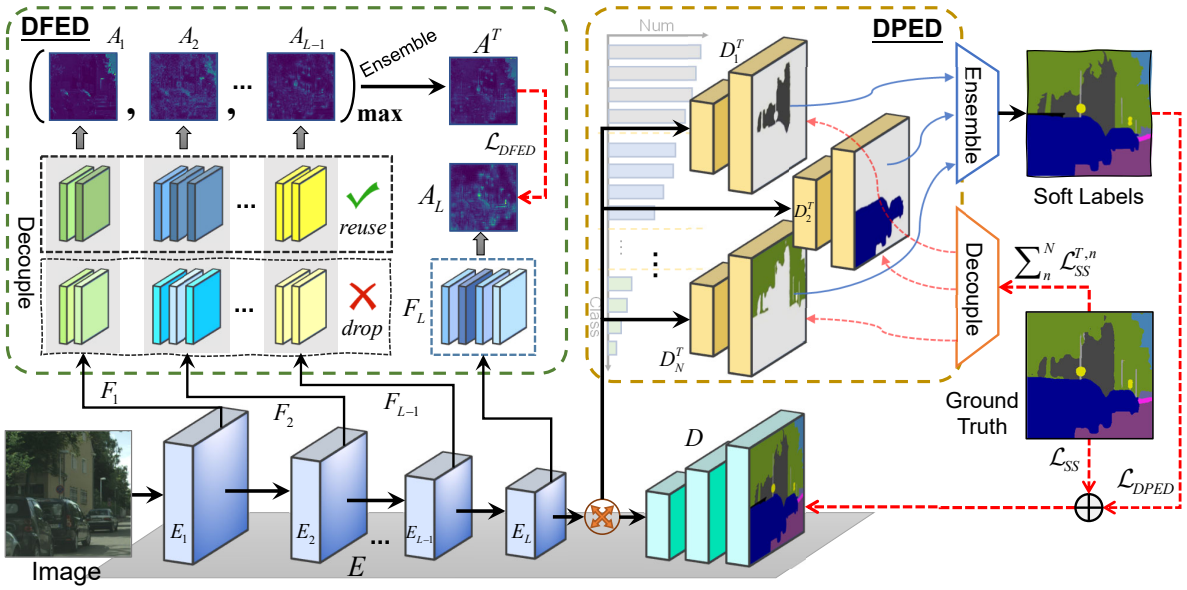


Figure 2: The proposed SDES framework. To better depict the prediction decoupling mechanism, we only show three typical semantic classes in the expert decoders and overlook the "more to less" subset splits in experiments. In DFED, the shades of color reflect the importance of channels.

The segmentation loss of the expert decoder D_n^T is to train each pixel in the sub-area with the label Y^n using cross-entropy. We can combine the losses of all the expert decoders as

$$\begin{aligned}\mathcal{L}_{\text{SS}}^{T} &= \frac{1}{N}\sum_{n}^{N}\mathcal{L}_{\text{SS}}^{T,n} \\ &= -\frac{1}{N}\sum_{n}^{N}\frac{1}{|M^n|}\sum_{i,M_i^n\neq 0}^{H\times W}\sum_{j}^{C} Y_{i,j}^{n}\log\sigma(P_i^{T,n})_j\,,\end{aligned} \tag{3}$$

where $P_i^{T,n}$ is the predicted logits of the i-th pixel from expert decoder D_n^T, and $P^{T,n} = D_n^T \circ E(X)$. $|M^n|$ is the number of non-zero elements in the mask M^n. It's worth noting that the parameters of all expert decoders are updated synchronously with the student decoder D. We assume that the student encoder E trained with the segmentation loss $\mathcal{L}_{SS}$ is sufficient to extract rich features. Hence, it's not necessary to be updated by the back-propagation from each expert decoder. And this can speed up the training process.

The ensemble of expert decoders can output reliable soft-labels for prediction distillation. We collect all the disjoint subset prediction maps $\{P^{T,1}, P^{T,2}, \cdots, P^{T,N}\}$ from the expert decoders and integrate them into an ensemble prediction map P^T with the corresponding subset masks. This can be formulated as

$$P^T = \sum_{n}^{N}\left(M^n \odot P^{T,n}\right) + \left(\mathbf{1} - \sum_{n}^{N} M^n\right)\odot P\,, \tag{4}$$

in which $\odot$ is the dot product operation. Generally, there are some ignored categories in the segmentation label sets, which are not considered in the mask calculation but can contribute to an incomplete expert prediction map. Thus, we introduce the student prediction map P to fill in the gap.

With the ensemble prediction map P^T, the decoupled prediction ensemble distillation (DPED) loss is formulated as a Kullback-Leibler (KL) divergence between the student soft-label map and the ensemble soft-label map, similar to the vanilla KD (Hinton, Vinyals, and Dean 2015):

$$\begin{aligned}\mathcal{L}_{\text{DPED}} &= \frac{1}{H\times W}\sum_{i}^{H\times W}\text{KL}\left(\sigma(\frac{P_i}{\tau})\|\sigma(\frac{P_i^T}{\tau})\right) \\ &= \frac{1}{H\times W}\sum_{i}^{H\times W}\sum_{j}^{C}\sigma\left(P_i/\tau\right)_j\log\frac{\sigma\left(P_i/\tau\right)_j}{\sigma\left(P_i^T/\tau\right)_j}\,,\end{aligned} \tag{5}$$

where $\sigma(P_i/\tau)$ and $\sigma(P_i^T/\tau)$ calculate the i-th soft-label of the student and ensemble expert, respectively. $\tau > 0$ is the distillation temperature.

Decoupled Feature Ensemble Distillation (DFED)

Motivation. In addition to the prediction expert, we can also construct a feature expert module to guide the student encoder at feature level. Analogy to the interminable distillation in T-S KD, the key is to obtain an expert feature map with richer information than the student extraction. With these insights, we innovatively propose to decouple the feature maps from multiple layers of the student itself, and then select the more important parts for aggregation and distillation.

Model filter pruning (Li et al. 2017; He, Zhang, and Sun 2017; Lin et al. 2020) is a model compression way that aims to remove some unimportant convolutional filters according to certain criteria. We take inspiration from this and introduce a channel-wise feature selection mechanism to reuse

the meaningful features and drop the tickle part, ensuring to provide the student with trustworthy information. Different from filter pruning, the measure and selection operations in our scheme are applied on the feature maps extracted by each layer, instead of the convolutional filters.

We denote each layer of the student encoder E as $E_\ell, \ell \in \{1, 2, \cdots, L\}$, *i.e.*, $E(X) = E_L \circ E_{L-1} \circ \cdots \circ E_1(X)$. Thus, the feature map extracted by the ℓ-th layer is represented by $F_\ell = E_\ell \circ E_{\ell-1} \circ E_1(X)$. Following the study (Li et al. 2017), we introduce a sum of absolute values raised to the power of p to measure the importance of the k-th channel in the feature map F_ℓ,

$$w_{\ell,k} = \sum_{i}^{H_\ell \times W_\ell} |F_{\ell,k,i}|^p \,, \tag{6}$$

where H_ℓ, W_ℓ are the height and width of the feature map F_ℓ. $|\cdot|^p$ is the p-power of the activation scalar, and we set $p = 1$. $F_{\ell,k,i}$ represents the i-th scalar at the k-th channel of the F_ℓ.

The channels with larger weights are selected for reuse, and the others are dropped. The reuse set of features is defined by

$$\mathcal{F}_\ell^{reuse} = \{F_{\ell,k} | w_{\ell,k} \geqslant h_\gamma(w_\ell)\} \,. \tag{7}$$

Here, $h_\gamma(\cdot)$ is a rank and selection function to get the threshold of the top γ weights. We set $\gamma = 0.5$ by default. The other channels at the ℓ-th layer belong to the drop set $\mathcal{F}_\ell^{drop}$,

$$\mathcal{F}_\ell^{reuse} \cup \mathcal{F}_\ell^{drop} = \mathcal{F}_\ell \,, \mathcal{F}_\ell^{reuse} \cap \mathcal{F}_\ell^{drop} = \emptyset \,. \tag{8}$$

We concatenate the reused channels in $\mathcal{F}_\ell^{reuse}$ into a new feature map $F_\ell^{reuse} \in \mathbb{R}^{\lfloor C_\ell \times \gamma \rfloor \times H_\ell \times W_\ell}$, in which $\lfloor \cdot \rfloor$ is the floor operation. To unify the shapes of different reused feature maps from different layers, we downsample all the front $(L-1)$ feature maps to the same size as the last one, *i.e.*, $H_L \times W_L$. The whole last-layer feature map F_L without pruning works as the student-side features in distillation. The attention maps of the previous layers and the last layer are calculated by

$$A_\ell = \frac{\hat{F}_\ell^{reuse}}{\left\| \hat{F}_\ell^{reuse} \right\|_2} \,, A_L = \frac{F_L}{\|F_L\|_2} \,. \tag{9}$$

Note that $\hat{F}_\ell^{reuse}$ is the downsampling version of F_ℓ^{reuse}, and $A^T, A_L \in \mathbb{R}^{H_L \times W_L}$.

To aggregate the previous attention maps as an ensemble version A^T, we compute the element-wise maximum by

$$A_i^T = \max(A_{1,i}, A_{2,i}, \cdots, A_{L-1,i}) \,, \tag{10}$$

where $i \in \{1, 2, \cdots, H_L \times W_L\}$. The decoupled feature ensemble distillation loss function can be simply formulated as an L_2 distance between the expert and student-side attention map:

$$\mathcal{L}_{\text{DFED}} = \left\| A_L - A^T \right\|_2 \,. \tag{11}$$

Overall Framework

We summarize our decoupled prediction and feature ensemble distillation together to train the student network. The cross-entropy segmentation losses for the student and auxiliary decoders are also employed. As such, the total loss of the SDES framework is formulated as

$$\mathcal{L}_{\text{total}} = (\mathcal{L}_{\text{SS}} + \mathcal{L}_{\text{SS}}^T) + \alpha \mathcal{L}_{\text{DPED}} + \beta \mathcal{L}_{\text{DFED}} \,. \tag{12}$$

Here, α and β are the weight coefficients to balance different components. The auxiliary decoders will be discarded in inference and not increase parameters or computation.

Experiments

Datasets

Pascal VOC 2012 (Everingham and Winn 2011) is a visual object segmentation dataset that that consists of 21 classes (20 foreground object classes and an extra background class). We extend it with additional annotation provided by the previous study (Hariharan et al. 2011), resulting in 10582/1449/1456 images for train/val/test.

Cityscapes (Cordts et al. 2016) contains 5000 high-resolution images (2975 fine annotation images for training, 500 for validation, and 1525 for testing) for urban scene parsing. It covers more than 30 classes, but only 19 classes are adopted for evaluation.

CamVid (Brostow et al. 2008) is an automotive dataset, containing 367/101/233 images for train/val/test, each with 720×960 pixels. Methods are evaluated on the most frequent 11 classes.

Implementation Details

Network architectures. For T-S KD experiments, we employ the segmentation architecture DeepLabV3 (Chen et al. 2017b) with ResNet-50 (He et al. 2016) backbone as the strong teacher network, without specific instruction. As for the student networks in T-S KD and Self-KD, we use a broad range of compact network architectures, including DeepLabV3 (Chen et al. 2017b) with lightweight backbones (*e.g.*, ResNet-18 (He et al. 2016), MobileNetV2 (Sandler et al. 2018), and EfficientNet-B0 (Tan and Le 2019)) and some real-time segmentation networks (*e.g.*, ENet (Paszke et al. 2016), ERFNet (Romera et al. 2017) and ESNet (Lyu et al. 2019)).

Training setup. Our approach is implemented by PyTorch with two NVIDIA 2080Ti GPUs. Following standard data augmentation, we apply random horizontal flipping and random cropping with size of 512×512 during training. The student networks are optimized by mini-batch stochastic gradient descent (SGD) with the momentum (0.9) and the weight decay (0.0001). We set the initial learning rate as 0.01 and use "poly" learning rate decay where the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iters})^{0.9}$ after each iteration. The number of the total training iterations is 30K/45K/5K for VOC/Cityscapes/CamVid with a batch size of 8. The hyperparameter τ, α and β are set to 10, 0.5 and 100 by default, respectively.

Method	Student	LSR	Tf-KD	SAD	SR-SS	KL*	AT*	CWD*	**Ours**
Teacher-free	✓	✓	✓	✓	✓	✗	✗	✗	✓
MobileNetV2	71.12	72.08	72.23	71.92	72.23	71.60	71.45	71.66	**72.62**
ResNet-50	76.75	77.24	77.46	76.98	77.20	77.13	77.02	77.71	**78.00**

Table 1: Performance of different distillation methods on the VOC dataset in terms of DeepLabV3. We adopt another pre-trained student network as the teacher in T-S KD methods which are marked with "*". "Teacher-free" indicates whether or not using pre-trained teacher networks.

Network	Method	Teacher-free (TF)	mIoU	PA	Params	FLOPs
DeepLabV3 (ResNet-50)	Teacher	-	74.85	95.53	39.64 M	327.97 G
DeepLabV3 (ResNet-18)	Student	✓	69.19	94.79	15.90 M	134.08 G
	Ens-T	-	76.26	96.64		
	LSR (Szegedy et al. 2016)	✓	68.94	94.71		
	Tf-KD (Yuan et al. 2020)	✓	70.35	94.96		
	SAD (Hou et al. 2019)	✓	70.10	94.98		
	SR-SS (Zhang et al. 2021a)	✓	70.46	95.01		
	Tf-FD (Li 2022)	✓	70.83	95.04		
	KL (Hinton, Vinyals, and Dean 2015)	✗	70.30	95.00		
	AT (Zagoruyko and Komodakis 2017)	✗	71.02	94.95		
	CWD (Shu et al. 2021)	✗	71.53	95.12		
	Ours(w/o DFED)	✓	71.81	95.12		
	Ours(w/o DPED)	✓	71.12	95.06		
	Ours	✓	**72.15**	**95.20**		
ERFNet	Student	✓	67.68	-	2.07 M	26.86 G
	Ours	✓	**69.24**	-		
ESNet	Student	✓	67.05	-	1.66 M	24.35 G
	Ours	✓	**69.07**	-		
ENet	Student	✓	60.14	-	0.36 M	4.35 G
	Ours	✓	**61.90**	-		

Table 2: Performance of different distillation methods on the Cityscapes dataset. We tag the self-ensemble teacher as "Ens-T". FLOPs is measured based on the fixed size of 512×1024.

Evaluation metrics. We employ mean Intersection over Union (mIoU) and Pixel Accuracy (PA) to measure the segmentation performance. Floating point operations (FLOPs) and parameters (Params) are utilized to measure the computation and storage cost of the models.

Comparison with State-Of-The-Art

Results on VOC. We evaluate the proposed distillation method on Pascal VOC 2012 with two student networks: DeepLabV3 with MobileNetV2 and ResNet-50, respectively. We set $N = 5$ and divide the background class into a group while the other 20 foreground classes into 4 groups. For a fair comparison, we introduce another pre-trained student network, instead of stronger networks, as the teacher in T-S KD, which is marked with "*" in the tables. The mIoU performance of the students is presented in the Table 1. Our method can outperform both the state-of-the-art Self-KD and T-S KD approaches and improve about 1.5% mIoU on both compact and large networks.

Results on Cityscapes. Table 2 demonstrates the performance of different distillation methods on the Cityscapes in terms of four compact or real-time segmentation networks. We set $N = 4$ and employ the expert decoders with simple architectures, *i.e.*, two fully convolutional layers. The parameters and FLOPs of the ensemble teacher for ResNet-18 are 15.94 M and 134.34 G, respectively. Each decoder has only about 0.01 M parameters and brings less than 0.1 G FLOPs in the training stage. Without heavy parameters and pre-training, the self-ensemble teacher can achieve the similar performance to a cumbersome teacher. We can see that the four Self-KD methods bring limited improvement, *i.e.*, only about 1% mIoU, and the two intermediate T-S KD methods (AT and CWD) boost the student by more than 2% mIoU. The proposed method achieves competitive performance and even outperforms the T-S KD methods. Comparing to T-S KD methods, ours does not need the teacher that

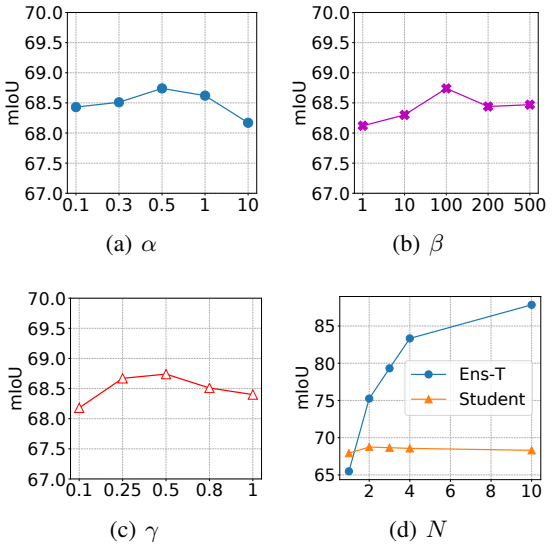


Figure 3: Impacts of (a) the weight coefficient α, (b) β, (c) the feature selection ratio γ, and (d) the number of expert decoders N. We adjust DeepLabV3 with EfficientNet-B0 on CamVid.

has 39.64 M parameters and requests extra 327.97 G FLOPs in distillation. Moreover, only with DPED or DFED, the proposed SEDS method can improve the student by 1.9 $\sim$ 2.6% mIoU. We conduct extensive experiments on three real-time segmentation networks, *i.e.*, ERFNet, ESNet and ENet, and the student is improved by nearly 2% mIoU.

Results on CamVid. In Table 3, we compare the proposed SDES against the state-of-the-art distillation methods on the CamVid dataset in terms of DeepLabV3 with two different lightweight backbones. We adopt two expert decoders the same as the student decoder D. There are only 11 semantic classes, so we set the distillation temperature $\tau = 5$. For MobileNetV2, our SEDS method can improve the student network by nearly 1% mIoU, while the other Self-KD methods that are specific to classification even degrade the performance. Compared to complicated T-S KD methods, our method achieves 67.74% mIoU on EfficientNet-B0. It's worth noting that even the massive ResNet-50 segmentation network with 39.64M parameters only has 68.88% mIoU on the CamVid dataset.

MobileNetV2			EfficientNet-B0		
Params: 12.65 M			Params: 7.31 M		
Method	TF	mIoU	Method	TF	mIoU
Student	✓	66.74	Student	✓	67.56
LSR	✓	64.45	Ens-T	-	75.25
Tf-KD	✓	65.75	KL*	✗	68.05
SAD	✓	66.45	AT*	✗	67.71
SR-SS	✓	66.85	CWD*	✗	68.55
Ours	✓	**67.71**	**Ours**	✓	**68.74**

Table 3: Performance of different distillation methods on the CamVid dataset in terms of DeepLabV3.

$\mathcal{L}_{\text{SS}}$	$\mathcal{L}_{\text{DPED}}$	$\mathcal{L}_{\text{DFED}}$	mIoU
✓			67.56
✓	✓		68.13
✓		✓	68.37
✓	✓	✓	68.74
✓	Rand	✓	68.19

Table 4: Ablation study of self-distillation loss terms on CamVid.

Ablation Study

Impact of hyperparameters. Figure 3 depicts the impacts of four hyperparameters in our framework. The α and β balance the two self-distillation loss functions. As shown in Figure 3(a)(b), we adjust $\alpha \in [0.1, 10], \beta \in [1, 500]$ and find $\alpha = 0.5, \beta = 100$ are the better choice. As for the channel-wise feature selection ratio $\gamma \in [0, 1]$, we find that too large or too small γ is not suitable for segmentation, and $0.25 \sim 0.5$ is the better choice. Finally, we investigate the impact of the number of expert decoders N and the corresponding results are shown in Figure 3(d). As can be seen, the self-ensemble teacher can easily achieve remarkable performance with more than two experts. But the big gap between the self-ensemble teacher and the student impedes the distillation when introducing too many experts. In general, $N \in [2, 5]$ is enough.

Effectiveness of loss terms. As shown in Table 4, we examine the contribution of each self-distillation loss, based on EfficientNet-B0. The decoupled prediction ensemble distillation loss $\mathcal{L}_{\text{DPED}}$ improves the baseline by 0.57%, while the decoupled feature ensemble distillation $\mathcal{L}_{\text{DFED}}$ brings the 0.81% mIoU gain. Applying these two distillation losses can lead to 1.18% mIoU gain over $\mathcal{L}_{\text{SS}}$. Additionally, to verify the efficacy of the "more to less" subset split strategy, we apply a random rank of categories, and it reduces the mIoU from 68.74% to 68.19%.

Conclusion

This paper presents a specialized self-knowledge distillation architecture for semantic segmentation. The DPED module provides reliable soft-label information for the student in each sub-area. And the DFED mechanism enhances the student backbone by aggregating the important feature maps from the precedent layers. These two distillation components address the issues in current T-S KD and Self-KD methods. We confirm the large performance improvements quantitatively, and verify the efficacy of each component with various ablation studies. In the future, we will further investigate this issue and extend our approach to other dense prediction tasks (Deng et al. 2019) and Transformer-based SS networks (Strudel et al. 2021; Xie et al. 2021).

Acknowledgments

This work was supported in part by KLATASDS-MOE and the Fundamental Research Funds for the Central Universities.

References

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12): 2481–2495.

Brostow, G. J.; Shotton, J.; Fauqueur, J.; and Cipolla, R. 2008. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 44–57.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4): 834–848.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017b. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, Z.; Zheng, X.; Shen, H.; Zeng, Z.; Zhou, Y.; and Zhao, R. 2020. Improving Knowledge Distillation via Category Structure. In *ECCV*, 205–219.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 3213–3223.

Deng, J.; Pan, Y.; Yao, T.; Zhou, W.; Li, H.; and Mei, T. 2019. Relation distillation networks for video object detection. In *ICCV*, 7023–7032.

Everingham, M.; and Winn, J. 2011. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 8.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *CVPR*, 3146–3154.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *IJCV*, 129(6): 1789–1819.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *ICCV*, 991–998.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, Y.; Zhang, X.; and Sun, J. 2017. Channel pruning for accelerating very deep neural networks. In *ICCV*, 1389–1397.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hou, Y.; Ma, Z.; Liu, C.; and Loy, C. C. 2019. Learning lightweight lane detection cnns by self attention distillation. In *ICCV*, 1013–1021.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708.

Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; and Yang, R. 2018. The apolloscape dataset for autonomous driving. In *CVPRW*, 954–960.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. CCNet: Criss-cross attention for semantic segmentation. In *ICCV*, 603–612.

Ji, M.; Shin, S.; Hwang, S.; Park, G.; and Moon, I.-C. 2021. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *CVPR*, 10664–10673.

Kim, K.; Ji, B.; Yoon, D.; and Hwang, S. 2021. Self-knowledge distillation with progressive refinement of targets. In *ICCV*, 6567–6576.

Levinson, J.; Askeland, J.; Becker, J.; Dolson, J.; Held, D.; Kammel, S.; Kolter, J. Z.; Langer, D.; Pink, O.; Pratt, V.; et al. 2011. Towards fully autonomous driving: Systems and algorithms. In *IEEE IVS*, 163–168.

Li, G.; Li, X.; Wang, Y.; Zhang, S.; Wu, Y.; and Liang, D. 2022. Knowledge Distillation for Object Detection via Rank Mimicking and Prediction-guided Feature Imitation. In *AAAI*.

Li, H.; Kadav, A.; Durdanovic, I.; Samet, H.; and Graf, H. P. 2017. Pruning filters for efficient convnets. In *ICLR*.

Li, L. 2022. Self-Regulated Feature Learning via Teacher-free Feature Distillation. In *ECCV*, 347–363.

Li, T.; Wang, L.; and Wu, G. 2021. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 630–639.

Lin, M.; Ji, R.; Wang, Y.; Zhang, Y.; Zhang, B.; Tian, Y.; and Shao, L. 2020. HRank: Filter Pruning using High-Rank Feature Map. In *CVPR*, 1529–1538.

Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; and Wang, J. 2019. Structured knowledge distillation for semantic segmentation. In *CVPR*, 2604–2613.

Liu, Y.; Zhang, W.; and Wang, J. 2022. Multi-Knowledge Aggregation and Transfer for Semantic Segmentation. In *AAAI*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *ICML*, 6448–6458. PMLR.

Lyu, H.; Fu, H.; Hu, X.; and Liu, L. 2019. ESNet: Edge-based segmentation network for real-time semantic segmentation in traffic scenes. In *ICIP*, 1855–1859.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 116–131.

Mobahi, H.; Farajtabar, M.; and Bartlett, P. 2020. Self-distillation amplifies regularization in hilbert space. *NeurIPS*, 33: 3351–3361.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In *NeurIPS*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *CVPR*, 3967–3976.

Paszke, A.; Chaurasia, A.; Kim, S.; and Culurciello, E. 2016. ENet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*.

Peng, B.; Jin, X.; Liu, J.; Li, D.; Wu, Y.; Liu, Y.; Zhou, S.; and Zhang, Z. 2019. Correlation congruence for knowledge distillation. In *ICCV*, 5007–5016.

Porrello, A.; Bergamini, L.; and Calderara, S. 2020. Robust Re-Identification by Multiple Views Knowledge Distillation. In *ECCV*, 93–110.

Romera, E.; Alvarez, J. M.; Bergasa, L. M.; and Arroyo, R. 2017. ERFNet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE TITS*, 19(1): 263–272.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 4510–4520.

Shen, Z.; Liu, Z.; Xu, D.; Chen, Z.; Cheng, K.-T.; and Savvides, M. 2021. Is label smoothing truly incompatible with knowledge distillation: An empirical study. In *ICLR*.

Shu, C.; Liu, Y.; Gao, J.; Yan, Z.; and Shen, C. 2021. Channel-Wise Knowledge Distillation for Dense Prediction. In *ICCV*, 5311–5320.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *CVPR*, 7262–7272.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Wang, L.; and Yoon, K.-J. 2020. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *arXiv preprint arXiv:2004.05937*.

Weinzaepfel, P.; Brégier, R.; Combaluzier, H.; Leroy, V.; and Rogez, G. 2020. DOPE: Distillation Of Part Experts for whole-body 3D pose estimation in the wild. In *ECCV*, 380–397.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *CVPR*, 10687–10698.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*, 1492–1500.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 1857–1866.

Yuan, L.; Tay, F. E.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, 3903–3911.

Yun, S.; Park, J.; Lee, K.; and Shin, J. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *CVPR*, 13876–13885.

Zagoruyko, S.; and Komodakis, N. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2021a. Self-Regulation for Semantic Segmentation. In *ICCV*, 6953–6963.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, 3713–3722.

Zhang, P.; Kang, Z.; Yang, T.; Zhang, X.; Zheng, N.; and Sun, J. 2022. LGD: Label-guided Self-distillation for Object Detection. In *AAAI*.

Zhang, S.; Chen, C.; Hu, X.; and Peng, S. 2021b. Balanced knowledge distillation for long-tailed learning. *arXiv preprint arXiv:2104.10510*.

Zhang, Y.; Lan, Z.; Dai, Y.; Zeng, F.; Bai, Y.; Chang, J.; and Wei, Y. 2020. Prime-Aware Adaptive Distillation. In *ECCV*, 658–674.

Zhang, Z.; and Sabuncu, M. 2020. Self-distillation as instance-specific label smoothing. *NeurIPS*, 33: 2184–2195.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.

Zheng, W.; Yan, L.; Wang, F.-Y.; and Gou, C. 2021. Progressive Knowledge-Embedded Unified Perceptual Parsing for Scene Understanding. In *CVPR*, 1633–1642.

Zhu, J.; Shen, Y.; Zhao, D.; and Zhou, B. 2020. In-domain gan inversion for real image editing. In *ECCV*, 592–608.