Intra-class Feature Variation Distillation for Semantic Segmentation

Yukang Wang[2], Wei Zhou[2], Tao Jiang[2], Xiang Bai[2], and Yongchao Xu[Γ1]

[1] School of Computer Science, Wuhan University, Wuhan, China
yongchao.xu@whu.edu.cn
[2] School of EiC, Huazhong University of Science and Technology, Wuhan, China
{wangyk,weizhou,taojiang,xbai}@hust.edu.cn

Abstract. Current state-of-the-art semantic segmentation methods usually require high computational resources for accurate segmentation. One promising way to achieve a good trade-off between segmentation accuracy and efficiency is knowledge distillation. In this paper, different from previous methods performing knowledge distillation for densely pairwise relations, we propose a novel intra-class feature variation distillation (IFVD) to transfer the intra-class feature variation (IFV) of the cumbersome model (teacher) to the compact model (student). Concretely, we compute the feature center (regarded as the prototype) of each class and characterize the IFV with the set of similarity between the feature on each pixel and its corresponding class-wise prototype. The teacher model usually learns more robust intra-class feature representation than the student model, making them have different IFV. Transferring such IFV from teacher to student could make the student mimic the teacher better in terms of feature distribution, and thus improve the segmentation accuracy. We evaluate the proposed approach on three widely adopted benchmarks: Cityscapes, CamVid and Pascal VOC 2012, consistently improving state-of-the-art methods. The code is available at https://github.com/YukangWang/IFVD.

Keywords: Semantic segmentation, knowledge distillation, intra-class feature variation.

1 Introduction

Semantic segmentation is a fundamental topic in computer vision, which aims to assign each pixel in the input image with a unique category label. The recent surge of work based on fully convolutional networks [25] (FCNs) has lead to vast performance improvements for semantic segmentation algorithms. However, seeking for high segmentation accuracy often comes at a cost of more runtime. Most state-of-the-art semantic segmentation frameworks [46,10,41,43,13] usually require high computational resources, which limits their use in many real-world applications such as autonomous driving, virtual reality, and robots. To tackle this problem, some real-time architectures for semantic segmentation have been proposed, *e.g.*, ENet [28], ESPNet [26], ICNet [45] and BiSeNet [40].

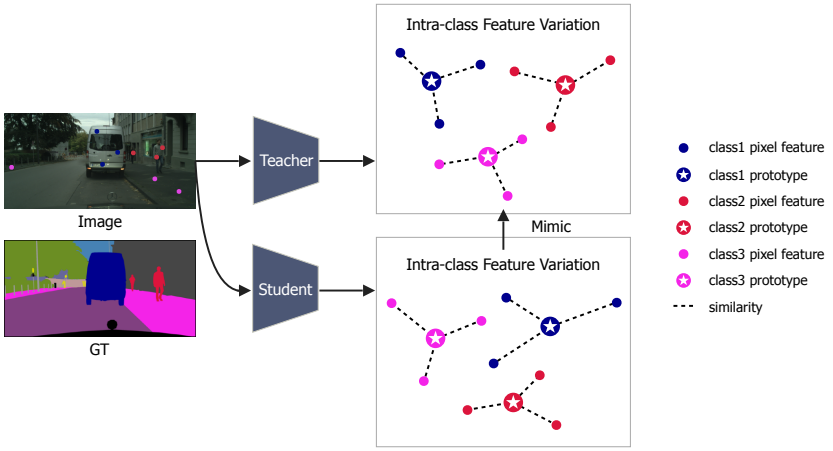

Fig. 1. The teacher model and the student model are endowed with different intra-class feature variation, which can be characterized as the set of similarity (dashed lines) between the feature on each pixel and its corresponding class-wise prototype. Higher similarity means lower variation. Our motivation is to transfer such variation of the teacher model to the student model, which makes the student model mimic the teacher model better, and thus improves the accuracy of the student model.

Model compression is a popular way to achieve high efficiency. In general, existing methods can be roughly divided into three categories: quantization [31,37], pruning [14,2,15,36] and knowledge distillation [19,33,44]. The quantization-based methods represent the parameters of filter kernels and weighting matrices using fewer bits. The pruning-based approaches aim to trim the network by removing redundant connections between neurons of adjacent layers. The notion of knowledge distillation is first proposed in [7] and then popularized by Hinton *et al.* [19]. The key idea is to utilize the soft probabilities of a cumbersome model (teacher) to supervise the training of a compact model (student). Later, in [44,21], the authors suggest transferring attention maps of the teacher model to the student model. In [35,30,27], the authors also attempt to transfer pairwise or triple-wise relations. Prior works using knowledge distillation are mostly devoted to classification tasks and achieve impressive results.

Some recent works [18,24] adopt knowledge distillation for semantic segmentation. Similar to the classification task, a straightforward scheme is to align individual pixel-wise outputs. This forces the student model to mimic the teacher model in terms of output probability maps. Different from the classification task, semantic segmentation has a structured output. The long-range dependencies are crucial for semantic segmentation, and the teacher model and student model usually capture different long-range contextual information due to their differences in receptive fields. Motivated by this, in [18,24], the authors propose to transfer the densely pairwise relations computed in the feature space. Moreover, in [24], the authors also align the outputs in a holistic manner via adversarial learning.

These knowledge distillation strategies are proved to be effective for semantic segmentation.

In this paper, we also leverage knowledge distillation for semantic segmentation. Different from previous works that transfer knowledge on densely pairwise relations, we propose a novel notion of intra-class feature variation distillation (IFVD). More specifically, the teacher model is able to produce a more robust intra-class feature representation than the student model, making them have different degrees of variation. Based on this property, we propose to transfer such variation of the teacher model to the student model (see Figure 1). For that, we first compute the feature center of each class, regarded as the class-wise prototype, which represents each class with one prototypical feature vector. We then characterize the intra-class feature variation (IFV) with the set of similarity between the feature on each pixel and its corresponding class-wise prototype and make the student model mimic such IFV of the teacher model, improving the segmentation accuracy of the student model. Extensive experiments demonstrate that the proposed IFVD consistently achieves noticeable improvements on the student model.

The main contributions of this paper are two-fold: 1) We propose a novel notion of intra-class feature variation distillation for semantic segmentation. More specifically, we force the student model to mimic the set of similarity between the feature on each pixel and its corresponding class-wise prototype, alleviating the difference of feature distributions between the student model and the teacher model. This helps to improve the segmentation accuracy of the student model. To the best of our knowledge, this is the first application of the intra-class feature variation concept to knowledge distillation for semantic segmentation. 2) The proposed intra-class feature variation distillation consistently improves upon existing methods using knowledge distillation for semantic segmentation, further boosting the state-of-the-art results of the compact model on three popular benchmark datasets.

The reminder of this paper is organized as follows. We shortly review some related works in Section 2 and clarify the differences with our approach. We then detail the proposed method, aptly named IFVD in Section 3, followed by extensive experiments in Section 4. Lastly, we conclude and give some perspectives on the future work in Section 5.

2 Related Work

We shortly review some related works on semantic segmentation and vision tasks leveraging knowledge distillation for boosting the accuracy while maintaining the efficiency of the compact model.

2.1 Semantic segmentation

Semantic segmentation is one of the most fundamental topics in computer vision. The recent rapid development of deep neural networks has had a tremendous

impact on semantic segmentation. Following the pioneer work [25] that adopts fully convolution network for semantic segmentation, many efforts have been made to boost the segmentation performance by exploiting the multi-scale context. For instance, Chen *et al.* [8] and Yu *et al.* [42] utilize dilated convolution to enlarge the receptive field and preserve the spatial size of the feature map. Chen *et al.* further develop DeeplabV3+ [10] with an encoder-decoder structure to recover the spatial information. PSPNet [46] apply the pyramid pooling to aggregate contextual information. Recently, some methods resort to the attention mechanism to guide the network learning and alleviate inconsistency in segmentation. For example, Yu *et al.* [41] adopt channel attention to select the features. OCNet [43] focuses on the context aggregation by spatial attention. In [13], the authors consider the combination of spatial and channel attention. These state-of-the-art methods aim to boost the segmentation performance at the cost of high computational resources.

Highly efficient semantic segmentation has been recently studied to address the above issue. ENet [28] explores spatial decomposition of convolutional kernels and achieves similar accuracy to SegNet [4] with 79x less parameters. ESPNet [26] designs an efficient spatial pyramid module that decomposes the standard convolution into point-wise convolution followed by spatial pyramid to reduce computational cost. In [45], the authors propose ICNet, an image cascade network based on the compressed PSPNet for real-time semantic segmentation. Yu *et al.* [40] introduce BiSeNet contains a spatial path and a context path to raise efficiency.

2.2 Vision tasks using knowledge distillation

Knowledge distillation has been widely studied in recent years. The concept is popularized by Hinton *et al.* in [19], which represents the process of training a student model with the objective of matching the soft probabilities of a teacher model. Similar ideas can also be found in [5,7,3]. With knowledge distillation, the student model performs well in terms of accuracy while maintaining efficiency. Various knowledge distillation schemes have been proposed recently. Romero *et al.* [33] utilize additional linear projection layers to minimize the discrepancy of high-level features. Zagoruyko *et al.* [44] and Huang *et al.* [21] transfer the attention map of the teacher model to the student model. Yim *et al.* [39] consider the flow knowledge between layers. Peng *et al.* [30] introduce correlation congruence for knowledge distillation to transfer not only the instance-level information but also the correlation between instances. Xu *et al.* [38] apply knowledge distillation based on conditional adversarial networks.

Prior works are mostly devoted to image classification. With growing interests in this topic, knowledge distillation approaches are proposed in other vision tasks, including semantic segmentation. He *et al.* [18] adapt the knowledge distillation with an additional auto-encoder and also transfer the densely pairwise affinity maps to the student model. Liu *et al.* [24] propose structured knowledge distillation (SKD), which also transfers pairwise relations and forces the outputs

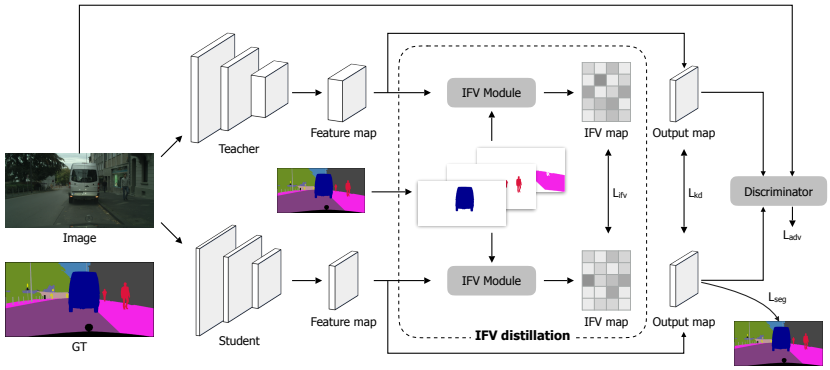

Fig. 2. Pipeline of the proposed intra-class feature variation distillation (IFVD). We introduce an IFV module to obtain the intra-class feature variation (IFV) maps. Knowledge transfer is then applied to the IFV maps of the teacher model and the student model. The original knowledge distillation loss (the KL divergence on outputs of teacher and student models) and adversarial learning are also included in our pipeline to further align the student model to the teacher model in the output space.

of the student model to mimic the teacher model from a holistic view via adversarial learning. The self-attention distillation (SAD) is introduced in [20] to explore attention maps derived from high-level features as the distillation target for shallow layers.

Most of the existing knowledge distillation approaches for semantic segmentation rely on transferring pairwise relations. However, our proposed IFVD solves the problem from a different aspect, which focuses on the intra-class feature variation (IFV). We propose to characterize the IFV with the set of similarity between the feature on each pixel and its corresponding class-wise prototype. The class-wise prototype is a prototypical representation for each class. This kind of similarity indicates how compact the intra-class feature distribution is. On the other hand, the stronger teacher model usually provides a more robust intra-class feature representation than the student model. Different feature distributions make the difference in semantic segmentation. The proposed IFVD forces the student model to explicitly mimic the intra-class feature variation, alleviating the difference in feature distribution between the student model and the teacher model. This is beneficial for improving the segmentation accuracy of the student model.

3 Method

3.1 Overview

Semantic segmentation densely classifies each pixel into a class category. Though many efforts have been made to maintain the intra-class consistency, intra-class

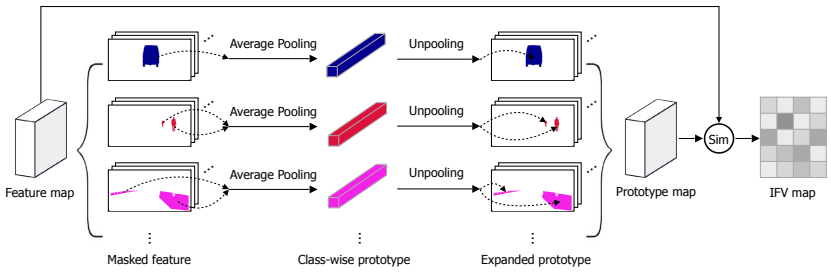

Fig. 3. Illustration of the proposed IFV module for computing the IFV map of a model. The masked feature for each class is generated from the feature map and the down-sampled label map (of the same size as the underlying feature map). Then the prototype of each class is obtained by masked average pooling and expanded to form the prototype map. Finally, we characterize the intra-class feature variation by computing the pixel-wise cosine similarity along channel dimension between the feature map and the prototype map.

variation in feature space still exists. Indeed, it is almost impossible for current CNN models to learn exactly the same feature for those pixels within the same category. Equipped with different feature extractors, the cumbersome model (teacher) and the compact model (student) have different degrees of intra-class feature variation. On the other hand, feature representation learning plays an important role in semantic segmentation. Different feature distributions lead to different segmentation results. The intra-class feature variation (IFV) is closely related to the feature distribution. Transferring such knowledge from teacher to student could make the student mimic the teacher better in terms of feature distribution, and thus improve the performance of the student model. Therefore, we propose to perform the knowledge distillation on the intra-class feature variation. The overall pipeline of the proposed method, dubbed intra-class feature variation distillation (IFVD), is depicted in Figure 2. In Section 3.2, we introduce the intra-class feature variation map to characterize the IFV of a model. We then detail the intra-class feature variation distillation in Section 3.3.

3.2 Intra-class feature variation map

We characterize the intra-class feature variation of a model using the map of feature similarity between each pixel and its corresponding class-wise prototype. Such intra-class feature variation (IFV) maps can be easily obtained in two steps. First, we compute the prototype for each class c by averaging the features on all pixels having the same class label c. Then we perform the cosine similarity function between the feature of each pixel and its corresponding class-wise prototype. Formally, the IFV map M is computed as follows:

$$M(p) = sim(f(p), \frac{1}{|\mathcal{S}_p|} \sum_{q \in \mathcal{S}_p} f(q)), \tag{1}$$

where $f(p)$ denotes the feature on pixel p, $\mathcal{S}_p$ is the set of pixels having the same label as pixel p, $|\mathcal{S}_p|$ stands for the size of the set $\mathcal{S}_p$, and sim is a similarity function. Specifically, we adopt the $Cosine$ similarity for all experiments in this paper.

As shown in Figure 3, to embed the proposed IFVD in existing deep neural networks, we propose an IFV module including the above steps. Concretely, we first down-sample the label map with the nearest interpolation to match the spatial size of the feature map. Then we select the region of the same label for each class and apply the average pooling on the masked feature along the spatial dimension. In this way, the prototype of each class is obtained. We then expand each class-wise prototype by unpooling operation on the masked region. In consequence, a prototype map with the same size as the input feature map is produced, in which each position stores the corresponding class-wise prototypical feature vector. Finally, the IFV map M is obtained by computing pixel-wise cosine similarity along channel dimension between the feature map and the prototype map.

3.3 Intra-class feature variation distillation

The intra-class feature variation (IFV) of a model can be well characterized by the IFV map described in the previous section. As described in the beginning of Section 3.1, the cumbersome model (teacher) and the compact model (student) usually have different intra-class feature variation. Moreover, we also found that there is still a bias of IFV after using existing knowledge distillation strategies. Therefore, we propose the intra-class feature variation distillation (IFVD), which aims to make the student model mimic better the teacher model.

A straightforward idea to achieve this goal is to minimize the distance between intra-class feature variation maps of the teacher model and the student model. Specifically, we employ the conventional Mean Squared (L2) loss as below:

$$L_{ifv} = \frac{1}{N} \sum_{p \in \Omega} (M_s(p) - M_t(p))^2, \tag{2}$$

where N is the number of pixels, Ω denotes the image domain, M_t and M_s represent the corresponding intra-class feature variation map (computed by Equation (1)) of the teacher model and the student model, respectively.

The loss function in Equation (2) makes the student model to mimic the intra-class feature variation of the teacher model. The original knowledge distillation loss and adversarial learning are also included in our pipeline to make the student model not only mimic the feature distribution but also the output score map of the teacher model.

The original KD loss is a conventional and widely adopted objective for many vision tasks. It adds a strong congruent constraint on predictions. Formally, we minimize the Kullback-Leibler (KL) divergence between the output score maps

S of the teacher model and the student model as follows:

$$L_{kd} = \frac{1}{N} \sum_{p \in \Omega} \sum_{i=1}^{C} S_s^i(p) \; log \; \frac{S_s^i(p)}{S_t^i(p)}, \tag{3}$$

where C denotes the total number of classes, $S_s^i(p)$ and $S_t^i(p)$ denote the probability of i-th class on pixel p produced by the student model and the teacher model, respectively.

Adversarial learning for knowledge distillation can be first found in [38]. Liu *et al.* [24] shares the similar idea for semantic segmentation, named holistic distillation. We also leverage the adversarial learning performed in the output space. More specifically, we first train a discriminator to distinguish whether an input is from the teacher model or the student model, by assessing how well the raw image and the segmentation map match. Then the segmentation network is trained to fool the discriminator. Formally, the loss for training discriminator L_d and adversarial item L_{adv} can be formulated as follows:

$$L_d = \mathbb{E}_{z_s \sim p_s(z_s)}[D(z_s|I)] - \mathbb{E}_{z_t \sim p_t(z_t)}[D(z_t|I)], \tag{4}$$

$$L_{adv} = \mathbb{E}_{z_s \sim p_s(z_s)}[D(z_s|I)], \tag{5}$$

where $\mathbb{E}[\cdot]$ represents the expectation operator. $D(\cdot)$ is an embedding network as the discriminator. I and z are the input image and the corresponding segmentation map.

For the proposed intra-class feature variation distillation (IFVD), the whole training objective is composed of a conventional cross-entropy loss L_{seg} for semantic segmentation and three loss items for knowledge distillation:

$$L = L_{seg} + \lambda_1 L_{kd} - \lambda_2 L_{adv} + \lambda_3 L_{ifv}, \tag{6}$$

where λ_1, λ_2, λ_3 are set to 10, 0.1 and 50, respectively.

During training, we alternatively optimize the discriminator D with L_d in Equation (4) and the segmentation network with L in Equation (6).

4 Experiments

To validate the effectiveness of the proposed IFVD, we conduct experiments on three common segmentation benchmark datasets: Cityscapes [11], CamVid [6] and Pascal VOC 2012 [12].

4.1 Datasets and evaluation metrics

Cityscapes [11] is a challenging benchmark collected for urban scene parsing. The dataset contains 5000 finely annotated images divided into 2975, 500 and

1525 images for training, validation and testing, respectively. It provides 30 common classes and 19 of them are used for evaluation. Similar to [24], we do not use the coarsely labeled data.

CamVid [6] is an automotive dataset extracted from high resolution video frames. It is split into 367 images for training and 233 images for testing. 11 classes are utilized for evaluation. The 12th class represents the unlabeled data that we ignore during training.

Pascal VOC 2012 [12] is a segmentation benchmark containing 20 foreground object categories and one background class. Following prior works [8,46], we use the augmented data with extra annotations provided by [16] resulting in 10582, 1449 and 1456 images for training, validation and testing, respectively. We use the *train* split for training and report performance on the *val* split.

Evaluation metrics. In all experiments, we adopt the commonly used mean Intersection-over-Union (mIoU) to measure segmentation accuracy. All models are tested under a single-scale setting. For a more robust and fair comparison, we report the average results of multiple models from the final epoch. The number of parameters is obtained by summing the number of elements for every parameter group in PyTorch [29] and "FLOPs" are calculated with the PyTorch version implementation [1] on a fixed input size (512×1024).

4.2 Implementation details

Network architectures. For a fair comparison, we experiment on the same cumbersome and compact networks as [24]. More specifically, we adopt the segmentation architecture PSPNet [46] with ResNet101 [17] backbone as the teacher model for all experiments. The student model also utilizes PSPNet [46] as the segmentation architecture but with different backbone networks. For the backbone of student model, we conduct experiments on ResNet18 [17] and ResNet18 (0.5), the width-halved version of ResNet18, respectively. We further replace the student backbone with EfficientNet-B0 [34] and EfficientNet-B1 [34] to validate the effectiveness of the proposed IFVD when the teacher model and the student model are of different architectural types.

Training details. For all our experiments, we first pretrain the teacher model following the training process of [46] and then keep the parameters frozen during the distillation progress. For the training process of the student, we use SGD as the optimizer with the "poly" learning rate policy where the learning rate equals to the base one multiplying $base\_lr * (1 - \frac{iter}{total\_iter})^{power}$. The base learning rate $base\_lr$ is initialized to 0.01 and the power is set to 0.9. We employ a batch size of 8 and 40000 iterations without specification. For the data augmentation, we only apply random flipping and random scaling in the range of $[0.5, 2]$. We choose image crop size as 512×512 for the limited GPU memory. The implementation is based on the PyTorch [29] platform. All our distillation experiments are carried out on a workstation with an Intel Xeon 16-core CPU (3.5GHz), 64GB RAM, and a single NIVIDIA Titan Xp GPU card of 12GB memory.

Table 1. Ablation study on Cityscapes.

Method			*val* mIoU (%)
T: ResNet101			78.56
S: ResNet18			69.10
$Loss_{kd}$	$Loss_{adv}$	$Loss_{ifv}$	*val* mIoU (%)
✓			70.51 (+1.41)
✓	✓		72.47 (+3.37)
✓	✓	✓	74.54 (+5.44)

4.3 Ablation study

As introduced in Section 3.3, the proposed IFVD contains three loss items for knowledge distillation, $Loss_{kd}$, $Loss_{adv}$ and $Loss_{ifv}$. Therefore, we study their contributions, respectively, on Cityscapes. Specifically, we first train the teacher model with ResNet101 backbone and then perform the knowledge distillation on the student model with ResNet18 backbone. As shown in Table 1, the original KD loss improves the student model without distillation by 1.41%. The gain increases to 3.37% when adversarial learning is also adopted. Further aligning the intra-class feature variation (IFV) boosts the improvement to 5.44%. The gap between student and teacher is reduced from 9.46% to 4.02%. These results demonstrate the effectiveness of the proposed IFVD, which is also complementary to other existing methods.

4.4 Results

Cityscapes. We first evaluate the proposed IFVD on the Cityscapes dataset [11]. Since the method in [24] does not provide experimental results with EfficientNet, we have implemented [24] with EfficientNet using the code released by [24]. The quantitative results are listed in Table 2. IFVD improves the student model built on ResNet18 without distillation by 5.44% on *val* set and 5.14% on *test* set. We also apply the proposed distillation scheme on ResNet18 (0.5), which is a width-halved version of the original ResNet18 and not pretrained on ImageNet. The proposed IFVD leads to an improvement of 7.95% on *val* set and 9.58% on *test* set. When the teacher model and the student model are of different architectural types, similar consistent improvements can also be obtained. Specifically, with the student model built on EfficientNet-B0, IFVD achieves a 6.36% and 4.46% mIoU boosting over the baseline model, on *val* set and *test* set, respectively. The gains shift to 6.10% and 4.51% when EfficientNet-B1 is adopted as the student model. Compared with SKD [24] relying on transferring pairwise relations, the proposed IFVD achieves consistent improvements, ranging from 0.72% to 3.37% on all involved student networks. These results demonstrate the effectiveness of the proposed IFVD. Some qualitative comparisons, *e.g.*, with ResNet18 as the student backbone, are illustrated in Figure 4.

Table 2. Quantitative results on Cityscapes. * means the results we reproduced using the released code of [24], which does not provide experimental results with EfficientNet.

Method	*val* mIoU(%)	*test* mIoU(%)	Params (M)	FLOPs (G)
Some related semantic segmentation methods				
ENet [28]	-	58.3	0.3580	3.612
ESPNet [26]	-	60.3	0.3635	4.422
ERFNet [32]	-	68.0	2.067	25.60
ICNet [45]	-	69.5	26.5	28.3
FCN [25]	-	65.3	134.5	333.9
RefineNet [23]	-	73.6	118.1	525.7
OCNet [43]	-	80.1	62.58	548.5
PSPNet [46]	-	78.4	70.43	574.9
Comparison with different distillation schemes				
T: ResNet101	78.56	76.78	70.43	574.9
S: ResNet18	69.10	67.60	13.07	125.8
+ SKD [24]	72.70	71.40		
+ IFVD (ours)	74.54	72.74		
S: ResNet18 (0.5)	55.40	54.10	3.27	31.53
+ SKD [24]	61.60	60.50		
+ IFVD (ours)	63.35	63.68		
S: EfficientNet-B0	58.37	58.06	4.19	7.967
+ SKD* [24]	62.90	61.80		
+ IFVD (ours)	64.73	62.52		
S: EfficientNet-B1	60.40	59.91	6.70	9.896
+ SKD* [24]	63.13	62.59		
+ IFVD (ours)	66.50	64.42		

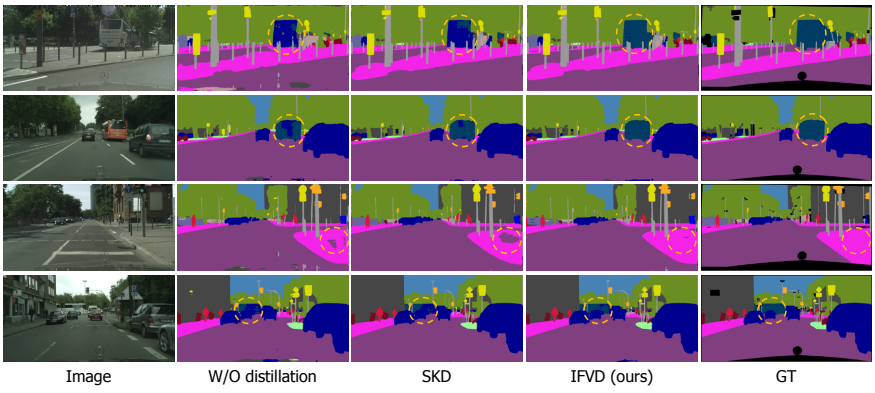

Fig. 4. Some qualitative comparisons on the Cityscapes *val* split.

Table 3. The segmentation performance on CamVid *test* set. * means the results we reproduced using the released code of [24], which does not provide experimental results with EfficientNet.

Method	*test* mIoU (%)	Params (M)
Some related semantic segmentation methods		
FCN [25]	57.0	134.5
ENet [28]	51.3	0.3580
ESPNet [26]	57.8	0.3635
FC-DenseNet56 [22]	58.9	1.550
SegNet [4]	55.6	29.46
ICNet [45]	67.1	26.5
BiSeNet-ResNet18 [40]	68.7	49.0
Comparison with different distillation schemes		
T: ResNet101	77.52	70.43
S: ResNet18	70.3	13.07
+ SKD [24]	71.0	
+ IFVD (ours)	71.8	
S: EfficientNet-B0	61.9	4.19
+ SKD* [24]	63.9	
+ IFVD (ours)	64.4	

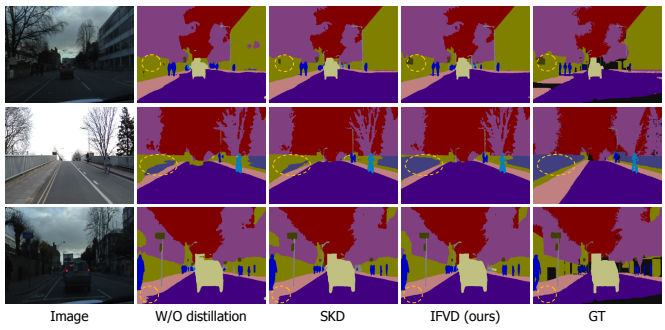

Fig. 5. Some qualitative results on the CamVid *test* split.

CamVid. We then evaluate the proposed IFVD on CamVid dataset [6]. The quantitative results are listed in Table 3. The proposed IFVD improves the model without distillation by 1.5% while slightly improving SKD [24] by 0.8%. Moreover, the gains shift to 2.5% and 0.5% when employing the student model built on EfficientNet-B0. Some qualitative comparisons based on ResNet18 are depicted in Figure 5.

Pascal VOC. We also conduct experiments on PASCAL VOC dataset [12] to further verify the distillation ability of the proposed IFVD on visual object segmentation. As depicted in Table 4, IFVD improves the baseline model by 3.27% while outperforming SKD [24] by 1.00%. We then evaluate our method with

Table 4. The performance on Pascal VOC 2012 *val* set. * means the results we reproduced using the released implementation of [24], which does not conduct experiments on this dataset.

Method	*val* mIoU (%)
Some related semantic segmentation methods	
CRF-RNN [47]	72.90
DeepLab-LargeFOV [8]	75.54
DeepLabV3 [9]	78.51
Comparison with different distillation schemes	
T: ResNet101	77.82
S: ResNet18	70.78
+ SKD* [24]	73.05
+ IFVD (ours)	74.05
S: EfficientNet-B0	69.28
+ SKD* [24]	70.24
+ IFVD (ours)	71.07

Fig. 6. Visual improvements on the Pascal VOC 2012 *val* split.

the student model built on EfficientNet-B0. The proposed IFVD surpasses the baseline model by 1.79% while improving SKD by 0.83%. Visualization results when employing ResNet18 as the student backbone are given in Figure 6.

4.5 Discussion

Experimental results prove that the proposed IFVD can consistently boost the accuracy of the student model. Besides, we also analyze the discrepancy between the IFV of teacher and student models before and after distillation on Cityscapes. As depicted in Figure 7, we observe that both the student models without distillation and with the state-of-the-art SKD [24] have a relatively high average

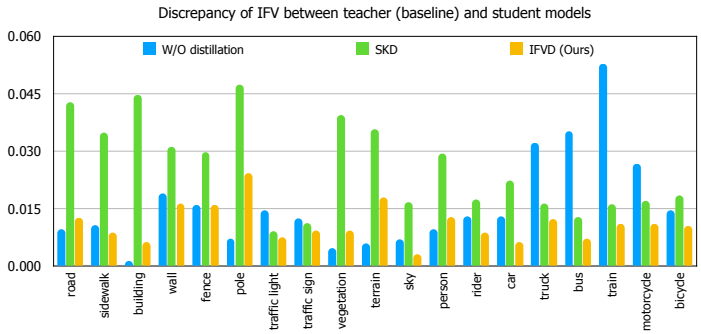

Fig. 7. Discrepancy between the intra-class feature variation of teacher and student models. We first obtain the IFV maps of teacher and student models, and then compute the average discrepancy between them with L1 distance for each class on Cityscapes.

bias to the teacher model. After applying the proposed IFVD, the average discrepancy is significantly decreased, implying that IFVD can make the student better mimic the teacher in terms of feature distribution and thus improve the performance. Finally, one may wonder what would happen if we use the global prototype computed on the whole training dataset. We have conducted such an experiment and the mIoU slightly reduced to 73.86% (-0.68%) on Cityscapes. This is probably due to the high intra-class variability in training data.

5 Conclusion

We propose a novel intra-class feature variation distillation (IFVD) for semantic segmentation. Different from existing methods that perform knowledge distillation on pairwise relations, we attempt to alleviate the difference in feature distribution of the teacher model and student model. This is achieved by transferring the set of similarity between the feature on each pixel and its corresponding class-wise prototype. We conduct extensive experiments on three popular benchmark datasets, and consistently improve the model without distillation by a large margin. Comparison with the state-of-the-art knowledge distillation method for semantic segmentation also demonstrates the effectiveness of the proposed IFVD. In the future, we would like to explore the inter-class feature separability in addition to intra-class feature variation for knowledge distillation. We also plan to explore such spirit in other tasks than semantic segmentation.

Acknowledgement

This work was supported in part by the Major Project for New Generation of AI under Grant no. 2018AAA0100400, NSFC 61703171, and NSF of Hubei Province of China under Grant 2018CFB199. Dr. Yongchao Xu was supported by the Young Elite Scientists Sponsorship Program by CAST.

References

1. https://github.com/warmspringwinds/pytorch-segmentation-detection/blob/master/pytorch_segmentation_detection/utils/flops_benchmark.py 9
2. Alvarez, J.M., Salzmann, M.: Learning the number of neurons in deep networks. In: Proc. of NIPS. pp. 2270–2278 (2016) 2
3. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Proc. of NIPS. pp. 2654–2662 (2014) 4
4. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. **39**(12), 2481–2495 (2017) 4, 12
5. Breiman, L., Shang, N.: Born again trees. University of California, Berkeley, Berkeley, CA, Technical Report **1**, 2 (1996) 4
6. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Proc. of ECCV. pp. 44–57 (2008) 8, 9, 12
7. Bucilu, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proc. of SIGKDD. pp. 535–541 (2006) 2, 4
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence. **40**(4), 834–848 (2018) 4, 9, 13
9. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 13
10. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. of ECCV. pp. 801–818 (2018) 1, 4
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of CVPR. pp. 3213–3223 (2016) 8, 10
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010) 8, 9, 12
13. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proc. of CVPR. pp. 3146–3154 (2019) 1, 4
14. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: Proc. of ICLR (2016) 2
15. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. In: Proc. of NIPS. pp. 1135–1143 (2015) 2
16. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proc. of ICCV. pp. 991–998 (2011) 9
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of CVPR. pp. 770–778 (2016) 9
18. He, T., Shen, C., Tian, Z., Gong, D., Sun, C., Yan, Y.: Knowledge adaptation for efficient semantic segmentation. In: Proc. of CVPR. pp. 578–587 (2019) 2, 4
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Proc. of NIPS Workshop (2014) 2, 4
20. Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning lightweight lane detection cnns by self attention distillation. In: Proc. of ICCV. pp. 1013–1021 (2019) 5

21. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017) 2, 4
22. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proc. of CVPR. pp. 11–19 (2017) 12
23. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proc. of CVPR. pp. 1925–1934 (2017) 11
24. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proc. of CVPR. pp. 2604–2613 (2019) 2, 4, 8, 9, 10, 11, 12, 13
25. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. of CVPR. pp. 3431–3440 (2015) 1, 4, 11, 12
26. Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H.: Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proc. of ECCV. pp. 552–568 (2018) 1, 4, 11, 12
27. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proc. of CVPR. pp. 3967–3976 (2019) 2
28. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016) 1, 4, 11, 12
29. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: Proc. of NIPS Workshop (2017) 9
30. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proc. of ICCV. pp. 5007–5016 (2019) 2, 4
31. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks. In: Proc. of ECCV. pp. 525–542 (2016) 2
32. Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems **19**(1), 263–272 (2017) 11
33. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: Proc. of ICLR (2015) 2, 4
34. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proc. of ICML. pp. 6105–6114 (2019) 9
35. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proc. of ICCV. pp. 1365–1374 (2019) 2
36. Wen, W., Wu, C., Wang, Y., Chen, Y., Li, H.: Learning structured sparsity in deep neural networks. In: Proc. of NIPS. pp. 2074–2082 (2016) 2
37. Wu, J., Leng, C., Wang, Y., Hu, Q., Cheng, J.: Quantized convolutional neural networks for mobile devices. In: Proc. of CVPR. pp. 4820–4828 (2016) 2
38. Xu, Z., Hsu, Y.C., Huang, J.: Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. In: Proc. of ICLR Workshop (2018) 4, 8
39. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proc. of CVPR. pp. 4133–4141 (2017) 4

40. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proc. of ECCV. pp. 325–341 (2018) 1, 4, 12
41. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: Proc. of CVPR. pp. 1857–1866 (2018) 1, 4
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: Proc. of ICLR (2016) 4
43. Yuan, Y., Wang, J.: Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018) 1, 4, 11
44. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proc. of ICLR (2017) 2, 4
45. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proc. of ECCV. pp. 405–420 (2018) 1, 4, 11, 12
46. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proc. of CVPR. pp. 2881–2890 (2017) 1, 4, 9, 11
47. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proc. of CVPR. pp. 1529–1537 (2015) 13