Adaptive Perspective Distillation for Semantic Segmentation

Zhuotao Tian*, Pengguang Chen*, Xin Lai, Li Jiang, Shu Liu, Hengshuang Zhao, Bei Yu, Ming-Chang Yang, Jiaya Jia, *Fellow, IEEE*

Abstract—Strong semantic segmentation models require large backbones to achieve promising performance, making it hard to adapt to real applications where effective real-time algorithms are needed. Knowledge distillation tackles this issue by letting the smaller model (student) produce similar pixel-wise predictions to that of a larger model (teacher). However, the classifier, which can be deemed as the perspective by which models perceive the encoded features for yielding observations (i.e., predictions), is shared by all training samples, fitting a universal feature distribution. Since good generalization to the entire distribution may bring the inferior specification to individual samples with a certain capacity, the shared universal perspective often overlooks details existing in each sample, causing degradation of knowledge distillation. In this paper, we propose Adaptive Perspective Distillation (APD) that creates an adaptive local perspective for each individual training sample. It extracts detailed contextual information from each training sample specifically, mining more details from the teacher and thus achieving better knowledge distillation results on the student. APD has no structural constraints to both teacher and student models, thus generalizing well to different semantic segmentation models. Extensive experiments on Cityscapes, ADE20K, and PASCAL-Context manifest the effectiveness of our proposed APD. Besides, APD can yield favorable performance gain to the models in both object detection and instance segmentation without bells and whistles.

Index Terms—Scene Understanding, Semantic Segmentation, Knowledge Distillation.

1 Introduction

Deep learning has significantly boosted the performance of semantic segmentation. Powerful segmentation models [2], [47] require strong feature extractors [8], [30], [35] to reach high performance. While real-time algorithms are more preferred in practice. Designing efficient segmentation models [46], [18], [38] is thus important.

Compared to hand-crafted efficient model design, knowledge distillation (KD) [10] is a more general technique for achieving high efficiency since KD can be applied to any existing models without structural constraints. Specifically, "knowledge" is distilled from a large model (teacher) to a smaller one (student) by minimizing the Kullback-Leibler divergence (KLD) between student output and soft target yielded by the teacher.

KD has been shown effective in classification [10], [24], [29], [36], while in segmentation, models are required to maintain the encoded features in certain resolutions and accomplish pixel-wise labeling by up-sampling to the original size. Contextual information is essential in segmentation because models cannot make predictions merely based on the RGB value of every single pixel. Design for contextual information enrichment (i.e., global pooling [16], pyramid pooling [47], dilated convolution [3] and attention [32]) can significantly improve the baselines. Previous methods [17], [33] propose distillation schemes to extract and transfer structured information on features, while it is notable that one important factor "perspective" in semantic segmentation is seldom studied.

Z. Tian, P. Chen, X. Lai, L. Jiang, B. Yu, M. Yang and J. Jia are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong.
S. Liu is with Smartmore.
H. Zhao is with University of Oxford.
Z. Tian and P. Chen contribute equally.

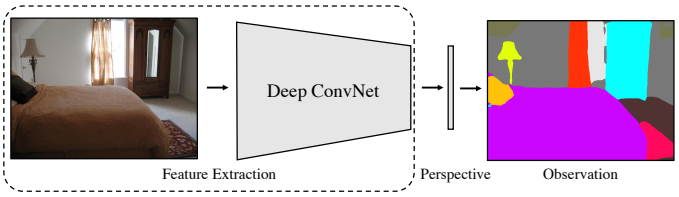


Fig. 1: Deep semantic segmentation framework is abstracted as the process that the final pixel-wise observation (prediction) is obtained from the perspective (classifier) based on encoded features produced by the deep neural networks.

Perspective works by representing the light that passes from a scene through a plane to the viewer's eye. In fact, *deep models perceive the encoded semantic features and make final predictions from the essential "perspective"*. We can consider the final classifier as a form of perspective for a model. Put differently, the inference of a segmentation model can be deemed as a process that *the perspective (classifier) projects the encoded high-level semantic information to yield observations (predictions) for the viewer*, as illustrated in Fig. 1. Compared to the student, the teacher usually has a better perspective because of the large feature encoder that can produce high-quality features to learn a good perspective, providing more accurate observations (predictions) used as soft targets in normal KD loss [10].

During KD, the teacher's feature encoder and perspective are fixed. Both of them generally fit the universal distribution given that they have been sufficiently trained on the entire training set. The fixed "universal perspective" of teacher achieves high-quality evaluation results by generalizing to all testing samples. However, the soft targets exploited with such a good generalization might not be the optimal choice for transferring knowledge from the teacher

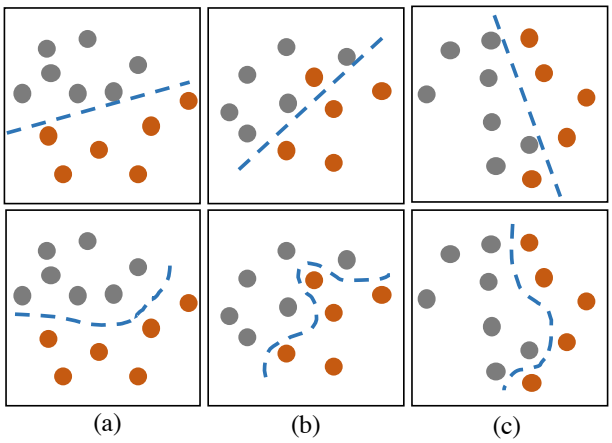


Fig. 2: Abstracted illustrations of the fixed universal perspective (top) and adaptive perspective (bottom) represented by blue dashed lines. They yield observations on features of two classes (gray and orange dots). Top examples show that generally correct observations can be obtained by the decision boundaries drawn by the fixed universal perspective, and the lack of specification to individual samples causes a few mistakes. However, in our proposed method, models learn to form adaptive perspectives that are more accurate decision boundaries as demonstrated by the bottom examples and the green dashed line in Fig. 3. The adaptive perspective reveals additional detailed co-occurring semantic cues for each individual sample, thus it might accomplish the knowledge distillation better than the fixed universal one.

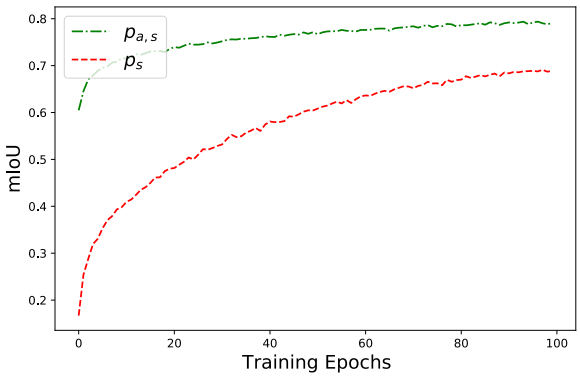


Fig. 3: Training mIoU curves of the auxiliary prediction $\boldsymbol{p}_{a,s}$ and the main prediction $\boldsymbol{p}_s$ of the student model on PASCAL-Context. $\boldsymbol{p}_{a,s}$ and $\boldsymbol{p}_s$ are obtained from the adaptive perspective and fixed universal perspective respectively. The auxiliary prediction $\boldsymbol{p}_{a,s}$ achieves much higher mIoU on the training set because $\boldsymbol{p}_{a,s}$ is generated by the adaptive perspective $\mathcal{A}_s$ that is with high specification to each image, mining more details for knowledge distillation and forming better decision boundaries as depicted by the bottom examples in Fig. 2. The comparison on the validation set is presented in Fig. 5.

to student, because, with a certain capacity, high generalization might cause poor specification that can reveal more useful information of the encoded features for decent knowledge distillation. To maintain good specification, the feature maps of different training samples should be projected by different perspectives to yield predictions, because even the same object may occur with varying co-occurrence information in different training samples, and a fixed universal perspective might not be able to well handle all the individual cases.

To address this key issue, we propose a new knowledge distillation method based on the concept of perspective for semantic segmentation. Our method enables models to form the adaptive perspective for every input image, i.e., different images are processed by different perspectives, based on their contextual contents. As illustrated in Figures 2 and 3, the adaptive perspective is generated for each image and it can better describe the encoded feature distribution, which reveals more contextual details that are conducive to knowledge distillation. As teacher always learns a better universal perspective, we also align the adaptive perspectives of teacher and student. It makes the student learn to form better adaptive perspectives under the teacher's guidance. Besides, the auxiliary observations (predictions) are obtained from the adaptive perspectives of the teacher and student. They are then used for distillation from the adaptive perspectives, further boosting performance.

We name our method Adaptive Perspective Distillation (APD) since it offers an adaptive perspective to reveal more contextual cues for semantic segmentation. Our method is effective in boosting different models on various benchmark datasets, achieving advanced performance compared with state-of-the-art algorithms. Note only two light-weight projectors are introduced for knowledge distillation, and, after training, they are simply discarded without causing any structural modification to the original model during evaluation, manifesting the substantial practical merit. In summary, our contribution is threefold.

- Different from the common practice in KD, we exam individual images and generate adaptive perspectives and observations to improve knowledge distillation.
- The proposed APD is model-agnostic and achieves great success by significantly improving different semantic segmentation models on popular datasets without structural constraints.
- Our method is also effective for knowledge distillation on the tasks of object detection and instance segmentation, further demonstrating the generalization ability.

2 Related Work

Semantic segmentation. Semantic segmentation is a fundamental and challenging task that requires accurate pixel-wise predictions for each image. FCN [27] is the first to adopt the convolution layers instead of the fully-connected layer to accomplish the semantic segmentation task. Encoder-decoder is developed [20], [1], [25] to let the encoded latent features refined by the decoder in steps. Dilated convolution [2], [40] enlarges the receptive field that is important for per-pixel predictions based on the contextual information. Pooling is another way for providing more contextual cues, such as global pooling [16], pyramid pooling [2], [47], [37], and strip pooling [11]. Note attention mechanism further boosts the performance by leveraging the long-range relationship across features [42], [48], [44], [6], [12], [14], [7], [41].

Recently, in order to perform pixel-wise semantic segmentation in real-time on mobile devices, efficient segmentation models are developed [22], [18], [46], [39]. E-Net [22] incorporates early

down-sampling, filter factorization, and pooling in parallel with strided convolution to reduce the computation overhead without compromising accuracy. ESPNet [18] builds the efficient spatial pyramid (ESP) module with factorized convolutions to accelerate the model. ICNet [46] leverages the multi-resolution branches with label guidance to accomplish real-time inference effectively. BiSeNet [39] proposes the spatial- and context-path to obtain sufficient contextual cues efficiently.

Knowledge distillation. Knowledge distillation was proposed by Hinton in [10]. It supervises a compact model by a larger pre-trained teacher in classification. The teacher provides soft labels, which contain useful "dark knowledge" for the student. The student could learn better results from the soft labels. Later, FitNet [24] distills knowledge from the features instead of the final prediction, which opened a new door in knowledge distillation. Following work [43], [21], [9] studied how to extract useful information from the features to better transfer to the student.

The study of knowledge distillation in semantic segmentation tasks commences in recent years. SKD [17] extracts structured information from the features. It also leverages a GAN network on top of the prediction of teacher and student to distill the holistic knowledge. After that, IFVD [33] extracts the intra-class feature variation on the features. They replace the transformation in SKD with an IFV transformation. The study of knowledge distillation in semantic segmentation is still far from satisfactory.

We analyze the knowledge distillation problem from a new view and propose the Adaptive Perspective Distillation that achieves advanced performance on different baselines and datasets.

3 Method

In this section, we start with a brief introduction of knowledge distillation in Sec. 3.1 followed by the introduction of our proposed method in Sec. 3.2.

3.1 Knowledge Distillation

Large models always achieve better performance than the small ones because of the large capacity. As suggested by Hinton *et al.* [10], knowledge of a large model (teacher) can be transferred to the smaller one (students) via soft labels that are more informative than the one-hot hard labels. This process is called knowledge distillation (KD). By mimicking the soft labels predicted by the teacher, the student gradually obtains the "dark knowledge" contained in the teacher model, such as correlation between different entities, which is conducive to the representation learning and cannot be expressed by the hard labels.

Liu *et al.* [17] apply KD to semantic segmentation where the Kullback-Leibler divergence (KLD) is calculated in a pixel-wise manner. Formally, the knowledge distillation loss $\mathcal{L}_{kd}$ is the average KLD of all pixels as

$$\mathcal{L}_{kd} = \frac{1}{H \times W} \sum_{x=1}^{H \times W} KLD(\boldsymbol{p}_s^x || \boldsymbol{p}_t^x), \tag{1}$$

where $\boldsymbol{p}_t^x$ and $\boldsymbol{p}_s^x$ represent the class probabilities of pixel x predicted by teacher and student models respectively.

It is worth noting that, normally, the teacher model is fixed during training to provide consistent soft targets p_t^i to student, and $\mathcal{L}_{kd}$ is used as an auxiliary loss that is optimized together with the main loss $\mathcal{L}_{ce}$ produced by p_s^i and one-hot hard labels. Therefore, the overall training objective $\mathcal{L}$ is

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{kd} \mathcal{L}_{kd}, \tag{2}$$

where λ_{kd} is set to 10 following [17], [33].

3.2 Adaptive Perspective Distillation

Overview. All semantic segmentation models can be decomposed into two components: 1) feature generator $\mathcal{G}$ and 2) classifier $\mathcal{C}$. Both $\mathcal{G}$ and $\mathcal{C}$ are fixed in the teacher model during distillation. Teacher's classifier $\mathcal{C}_t$ takes the features $\boldsymbol{f}_t$ extracted from $\mathcal{G}_t$ and produces soft targets for $\mathcal{L}_{kd}$. However, $\mathcal{C}_t$ fits the entire training set, and thus it provides a fixed universal perspective for mining knowledge from each feature map extracted by $\mathcal{G}_t$ of the teacher.

To further investigate the "dark knowledge" inside the teacher, we take a closer look at each training sample by forming individual adaptive perspectives $\mathcal{A}_t$ that are composed of semantic anchors (i.e., representative vectors for individual semantic classes) obtained from the encoded features $\boldsymbol{f}_t$, which serves as another auxiliary task providing local perspectives for distilling knowledge. Auxiliary observations $\boldsymbol{p}_{a,t}$ are then generated by adaptive perspectives $\mathcal{A}_t$ and encoded features $\boldsymbol{f}_t$ for transferring the knowledge from teacher to student. The student feature generator $\mathcal{G}_s$ is required to mimic $\mathcal{G}_t$ to yield similar adaptive perspectives $\mathcal{A}_s$, as well as the auxiliary observations $\boldsymbol{p}_{a,s}$ obtained from $\mathcal{A}_s$. Since both the adaptive perspective and auxiliary observations are generated specifically for each training sample, they provide more informative cues for KD. Our method is abstracted in Fig. 4.

Adaptive perspective. In the following, we introduce the way to generate adaptive perspectives to better distill the knowledge between the teacher and student models. First, two light-weight projectors, i.e., two 2-layer MLPs with an intermediate ReLU activation layer, are used to produce the adapted features for constructing new perspectives with the same channel numbers, making our method model-agnostic because the teacher and student models usually have different output channels. We can formalize this procedure as

$$\boldsymbol{f}_{a,t} = \mathcal{P}_t(\boldsymbol{f}_t), \quad \boldsymbol{f}_{a,s} = \mathcal{P}_s(\boldsymbol{f}_s). \tag{3}$$

Masked average pooling (MAP) is then applied to $\boldsymbol{f}_{a,t}$ and $\boldsymbol{f}_{a,s}$ to generate the C-dimensional semantic anchors $\mathcal{A}_t^i$ and $\mathcal{A}_s^i \in \mathcal{R}^{[1\times C]} (i \in \{1, ..., N\})$ as shown in Eq. (4), where $\boldsymbol{M}_i \in \mathcal{R}^{[H\times W\times 1]}$ is the binary mask obtained from the ground truth label, indicating whether the features belong to class c_i, and x denotes the feature position. N represents the number of classes contained in the current image, and different images may have different values of N. For simplicity, we only discuss the case with one single image.

$$\mathcal{A}_t^i = \frac{\sum_{x=1}^{HW} \boldsymbol{f}_{a,t}^x \cdot \boldsymbol{M}_i^x}{\sum_{x=1}^{HW} \boldsymbol{M}_i^x}, \quad \mathcal{A}_s^i = \frac{\sum_{x=1}^{HW} \boldsymbol{f}_{a,s}^x \cdot \boldsymbol{M}_i^x}{\sum_{x=1}^{HW} \boldsymbol{M}_i^x}. \tag{4}$$

These semantic anchors are then put together to form an adaptive perspective for each image. With the information provided by the ground-truth labels, the adaptive perspective can better describe the encoded semantic distribution. Thus, though it cannot be used for the final prediction due to the use of the ground-truth label, it is suitable to distill knowledge between the student and teacher with deeper insight, i.e., how the model interprets the encoded features for different images. Note it is normal to add

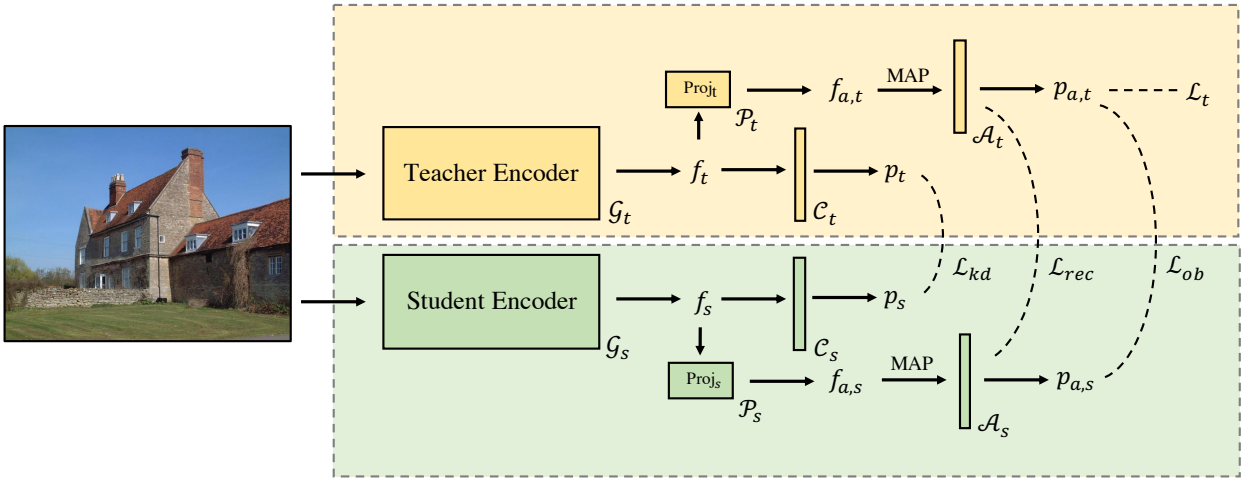


Fig. 4: Illustration of our method. The input image is first processed by teacher and student encoders ($\mathcal{G}_t$ and $\mathcal{G}_s$) respectively to get the encoded feature maps $\boldsymbol{f}_t$ and $\boldsymbol{f}_s$. To accomplish normal KD, $\mathcal{L}_{kd}$ [10] is applied to the predictions obtained from the main classifiers $\boldsymbol{C}_t$ and $\boldsymbol{C}_s$, offering a global perspective. $\boldsymbol{f}_t$ and $\boldsymbol{f}_s$ are also transformed by projectors ($\mathcal{P}_t$ and $\mathcal{P}_s$) to form adaptive classifiers $\mathcal{A}_t$ and $\mathcal{A}_s$, serving as local views that reveal useful details. The distillation from the adaptive perspectives is accomplished by the proposed $\mathcal{L}_{rec}$ and $\mathcal{L}_{ob}$ that rectifies adaptive classifiers and aligns auxiliary predictions ($\boldsymbol{p}_{a,t}$ and $\boldsymbol{p}_{a,s}$) respectively. $\mathcal{L}_t$ only updates teacher's projector $\mathcal{P}_t$. We note that the gradients yielded by $\mathcal{L}_{kd}$, $\mathcal{L}_{rec}$ and $\mathcal{L}_{ob}$ will not be back-propagated to $\boldsymbol{p}_t$, $\mathcal{A}_t$ and $\boldsymbol{p}_{a,t}$. The normal cross entropy loss $\mathcal{L}_{ce}$ applied to $\boldsymbol{p}_s$ is omitted in this figure for simplicity.

extra modules during distillation in literature. The proposed two projectors are not used during inference, so the model efficiency is not adversely affected.

After we get the adaptive perspectives, additional explicit observations can be obtained by calculating the cosine similarity between the adapted features ($\boldsymbol{f}_{a,t}$ and $\boldsymbol{f}_{a,s}$) and adaptive perspectives ($\mathcal{A}_t$ and $\mathcal{A}_s$) as

$$\boldsymbol{p}_{a,t}^{x,i} = \frac{\exp(\cos(\boldsymbol{f}_{a,t}^{x}, \mathcal{A}_t^i)/\tau)}{\sum_{j=1}^{N} \exp(\cos(\boldsymbol{f}_{a,t}^{x}, \mathcal{A}_t^j)/\tau)}, \tag{5}$$

$$\boldsymbol{p}_{a,s}^{x,i} = \frac{\exp(\cos(\boldsymbol{f}_{a,s}^{x}, \mathcal{A}_s^i)/\tau)}{\sum_{j=1}^{N} \exp(\cos(\boldsymbol{f}_{a,s}^{x}, \mathcal{A}_s^j)/\tau)}. \tag{6}$$

Learning objective for teacher's adaptive perspective. Teacher's projector $\mathcal{P}_t$ is randomly initialized by the default setting of PyTorch, thus it will collapse with meaningless interpretation without optimization. To ensure that $\mathcal{P}_t$ can provide representative perspectives $\mathcal{A}_t \in \mathcal{R}^{[N\times C]}$ that reveal more contextual details for each image, an explicit regularization is indispensable – features belonging to class c_i should get closer to $\mathcal{A}_t^i$ and are far from the semantic anchors of the other co-occurring categories. Therefore, we introduce the learning objective for teacher's projector $\mathcal{P}_t$ as

$$\mathcal{L}_t = \frac{1}{H\times W}\sum_{x=1}^{H\times W} -\log \frac{\exp(\cos(\boldsymbol{f}_{a,t}^{x}, \mathcal{A}_t^{c(x)})/\tau)}{\sum_{i=1}^{N} \exp(\cos(\boldsymbol{f}_{a,t}^{x}, \mathcal{A}_t^{i})/\tau)} \tag{7}$$

where $c(x)$ indicates the class that $\boldsymbol{f}_{a,t}^{x}$ belongs to, and τ is set as 0.1 for cosine similarity. We note that the teacher model is fixed during KD, and $\mathcal{L}_t$ only optimizes the teacher's projector $\mathcal{P}_t$.

Learning objective for the student. Misaligned perspectives may result in different observations. Therefore, student's feature generator $\mathcal{G}_s$ and projector $\mathcal{P}_s$ are first required to mimic teacher by producing similar perspectives. To realize this objective, we apply $\mathcal{L}_{rec}$ to accomplish the rectification on the adaptive perspectives of teacher and student. $\mathcal{L}_{rec}$ directly encourages the similarity between $\mathcal{A}_t$ and $\mathcal{A}_s$ as

$$\mathcal{L}_{rec} = 1 - \frac{1}{N}\sum_{i=1}^{N} \cos(\mathcal{A}_s^i, \mathcal{A}_t^i). \tag{8}$$

Furthermore, the observation obtained from the student's perspective also needs to imitate the teacher's observation, which can be achieved by minimizing KLD between their observations $\boldsymbol{p}_{a,t}$ and $\boldsymbol{p}_{a,s}$ as

$$\mathcal{L}_{ob} = \frac{1}{H\times W}\sum_{x=1}^{H\times W} KLD(\boldsymbol{p}_{a,s}^{x} || \boldsymbol{p}_{a,t}^{x}). \tag{9}$$

The overall Adaptive Perspective Distillation objective for student extends the loss in Eq. (2) with $\mathcal{L}_{ob}$ and $\mathcal{L}_{rec}$ providing extra informative cues for distillation as

$$\mathcal{L}_s = \mathcal{L}_{ce} + \lambda_{kd}(\mathcal{L}_{kd} + \mathcal{L}_{ob}) + \lambda_{rec}\mathcal{L}_{rec}, \tag{10}$$

where λ_{kd} for $\mathcal{L}_{kd}$ is set to 10, the same as those in SKD and IFVD for fair comparison. As for $\mathcal{L}_{ob}$ that minimizes the Kullback-Leibler divergence from the adaptive observations, its loss weight is empirically set to λ_{kd}. The weighting factor λ_{rec} is set to 10. τ for scaling the cosine similarity is 0.1 in $\mathcal{L}_{ob}$ and $\mathcal{L}_{rec}$. The sensitivity analysis of λ_{rec} and τ is given in Sec. 4.4. They both work well on all datasets with different backbones without further tuning.

Optimization. $\mathcal{L}_t$ only optimizes the teacher's projector $\mathcal{P}_t$ because the gradients yielded by $\mathcal{L}_{kd}$, $\mathcal{L}_{rec}$ and $\mathcal{L}_{ob}$ will not be back-propagated to $\boldsymbol{p}_t$, $\mathcal{A}_t$ and $\boldsymbol{p}_{a,t}$, as shown in Fig. 4. On the other hand, $\mathcal{L}_s$ optimizes the entire student model, i.e., feature generator $\mathcal{G}_s$ and classifier $\mathcal{C}_s$, as well as the projector $\mathcal{P}_s$. Therefore $\mathcal{L}_t$ and $\mathcal{L}_s$ optimize independently in each training batch.

4 EXPERIMENTS

4.1 Dataset Description

Cityscapes [5] focuses on semantic understanding of urban street scenes. It contains 5000 finely annotated images. Specifically, 2975, 500 and 1525 images for training, validation and testing respectively. 19 classes are required in prediction for evaluation.

ADE20K [49] is a rather challenging dataset that spans diverse annotations of scenes, objects, parts of objects, and in some cases even parts of parts. ADE20K contains up to 150 classes and diverse scenes for semantic segmentation. 20000, 2000 and 3000 images are used for training, validation and testing.

PASCAL-Context [19] extends the original PASCAL VOC semantic segmentation task with more detailed annotations for the whole scene. 4998 and 5105 images are used for training and validation, and 9637 images are used for testing. We evaluate all models on 60 categories (59 + background), following the practice of MMSegmentation [4].

COCO [15] is the most popular and challenging dataset for object detection and instance segmentation. In this paper, we use "COCO" to represent COCO 2017 dataset. It contains more than 200,000 images and 80 object categories for train, validation, and test sets. We use the COCO 2017 *train* set for training and report the validation results on the COCO 2017 *val* set. The results are reported in COCO-style mAP.

4.2 Implementation Details

We adopt three popular scene parsing benchmark datasets (Cityscapes [5], ADE20K [49] and PASCAL-Context [19]) in experiments. Models are trained and evaluated on the training and validation sets of these datasets respectively by default.

Both projectors $\mathcal{P}_t$ and $\mathcal{P}_s$ are composed of two 1×1 convolutional layers (denoted as $d_{in} \times d_{out}$) with an intermediate ReLU activation layer, while the difference lies in the input & output dimensions of the convolutional layers. Let d_t and d_s represent the dimensions of features $\boldsymbol{f}_t$ and $\boldsymbol{f}_s$ yielded by teacher and student feature generators $\mathcal{G}_t$ and $\mathcal{G}_s$, respectively. Usually, because the teacher is with a larger capacity and $d_t \geq d_s$, teacher's projector $\mathcal{P}_t$ is required to compress the dimension of $\boldsymbol{f}_t$ from d_t to d_s, matching that of the student feature $\boldsymbol{f}_s$. Therefore, the structure of $\mathcal{P}_t$ is as: $[d_t \times d_s \rightarrow \text{ReLU} \rightarrow d_s \times d_s]$, and the structure of $\mathcal{P}_s$ is: $[d_s \times d_s \rightarrow \text{ReLU} \rightarrow d_s \times d_s]$. Then, the projected features are l-2 normalized for calculating the cosine similarity.

The semantic segmentation models are built upon Semseg [45]. Student models are trained following the default configuration of PSPNet [47] except for the initial learning rate and batch size because PSPNet uses 8 GPUs by default while we use 4 GPUs for training. Specific epoch numbers, initial learning rates and training patch sizes used for different datasets are summarized in Table 1. SGD is used for optimization. Weight decay and momentum are set to 0.0001 and 0.9 respectively. The "poly" learning rate decay [2] is used by multiplying the initial learning rate with $(1 - \frac{current\_iter}{max\_iter})^{power}$, where $power$ is set to 0.9. All models are optimized without OHEM. As for the teacher, since the feature generator and classifier are fixed during training, only the projector $\mathcal{P}_t$ requires gradients. $\mathcal{P}_t$ is optimized by Adam optimizer with initial learning rate 1e-5 and beta (0.9, 0.99), which generalize well on all datasets without additional tuning.

TABLE 1: Training configurations on different datasets. Epoch: Training epoch number. BS: Batch size. InitLR: Initial training learning rate. PS: Patch size for training.

Dataset	Epoch	BS	InitLR	PS
Cityscapes [5]	200	8	5e-3	713
ADE20K [49]	100	8	5e-3	473
PASCAL-Context [19]	100	12	7.5e-4	473

Data augmentation includes mirroring, re-scaling from 0.5 and 2.0, and random rotation from -10 to 10 degrees. Finally, image patches are cropped from the original images as training samples. We output the prediction without additional post-processing (e.g., fully connected conditional random field (CRF) [13] and multi-scale testing). All experiments are conducted on PyTorch with four NVIDIA GTX 2080Ti GPUs, and results are obtained without altering the original labels. We will make our code publicly available for reproducing all experimental results in this paper.

4.3 Comparison with State-of-the-art

In this section, we show quantitative and qualitative comparison with state-of-the-art methods SKD [17] and IFVD [33]. For a fair comparison, we reproduce these two methods in the same training and testing setting with our method based on their official code.

Statistical comparisons & analysis. As shown in Table 2, we make comparison between the teacher PSPNet-R101 and student models on different backbones, i.e., ResNet-18 [8], MobileNet-V2 [26] and EfficientNet [28]. Since our method enables models to form new local perspectives that mine extra useful information, the proposed Adaptive Perspective Distillation achieves better performance compared to other methods when different student backbones are adopted.

We note SKD and IFVD only distill knowledge from an unchanged global view with a fixed classifier of the teacher. It is via $\mathcal{L}_{kd}$ [10] without new perspectives, causing limited knowledge that can be transferred. Contrarily, the proposed method mines extra cues for distillation by creating a new perspective for every single image specifically, and thus our method consistently yields significant performance gain to all student models. Besides, in Sec. 4.4, we show that our proposed APD is complementary to SKD and IFVD.

The efficiency comparison is illustrated in Table 3 with the test mIoU results on Cityscapes. We also conduct experiments with PSPNet on ADE20K and PASCAL-Context to show the superiority of our method on different datasets. Results are shown in Table 4.

Cross-model distillation. To further manifest the generalization ability of the prosed method, we conduct experiments across different models, i.e., PSPNet $\rightarrow$ DeepLab-V3 and DeepLab-V3 $\rightarrow$ PSPNet. The cross-model distillation Results are shown in Table 5. It can be observed that IFVD and SKD may adversely affect the performance for cross-model distillation as sometimes they may cause performance degradation compared to the results of KD proposed by Hinton *et al.* [10]. On the contrary, the proposed method still consistently brings decent performance gain in the practical cross-model setting.

Comparison with validation curves. Qualitative comparison with validation curves is presented in Fig. 5. We note that these validation results are obtained from the center regions cropped

TABLE 2: Performance comparison with state-of-the-art methods on Cityscapes with PSPNet [47] and DeepLab-V3 [2]. RN, MN2 and EN represent ResNet [8], MobileNet-V2 [26] and EfficientNet [28] respectively.

Methods	Backbone	PSPNet	DeepLab-V3
Teacher	RN-101	78.15	78.47
Student-I	RN-18	74.15	74.47
+ KD	RN-18	74.81	73.67
+ SKD	RN-18	74.56	74.03
+ IFVD	RN-18	74.10	74.99
+ Ours	RN-18	**75.68**	**75.45**
Student-III	MN2-1.0	71.34	71.40
+ KD	MN2-1.0	71.91	71.94
+ SKD	MN2-1.0	72.40	71.34
+ IFVD	MN2-1.0	72.94	70.79
+ Ours	MN2-1.0	**73.66**	**74.47**
Student-IV	EN-B0	72.30	71.54
+ KD	EN-B0	73.32	72.55
+ SKD	EN-B0	73.45	69.47
+ IFVD	EN-B0	74.43	72.93
+ Ours	EN-B0	**75.79**	**74.92**
Teacher	MN2-1.0	71.34	71.40
Student	MN2-0.5	63.34	63.89
+ KD	MN2-0.5	64.60	66.03
+ SKD	MN2-0.5	65.06	65.84
+ IFVD	MN2-0.5	65.31	66.78
+ Ours	MN2-0.5	**67.28**	**67.58**

with the training patch sizes (i.e., 473×473 for ADE20K [49] and PASCAL-Context [19], and 713×713 for Cityscapes [5]), which is different from the formal evaluation phase when the sliding windows inference strategy is adopted. The center cropping for the intermediate validation and the sliding window inference for the final evaluation are both implemented according to the official PyTorch implementation of PSPNet.

From Fig. 5, we can observe that APD consistently outperforms other methods by a large margin on both three benchmark datasets throughout the entire training process, which manifests the robustness of our method.

Visual comparison. We present the qualitative comparison between SKD and IFVD on Cityscapes, ADE20K and PASCAL-Context in Fig. 6 where it is observed that our predictions are generally better than the others by capturing more local contextual information for distillation.

4.4 Ablation Study

In this Section, we first verify that $\mathcal{L}_{ob}$ and $\mathcal{L}_{rec}$ are important to align teacher's observations and perspectives respectively. Then, as two projectors $\mathcal{P}_t$ and $\mathcal{P}_s$ are introduced during student training, we show that the improvement brought by $\mathcal{L}_{ob}$ and $\mathcal{L}_{rec}$ is not originated from these additional learnable modules. Besides, we provide a sensitivity analysis of λ_{rec} and τ to show the robustness of our method.

Effectiveness of $\mathcal{L}_{ob}$ and $\mathcal{L}_{rec}$. The proposed Adaptive Perspective Distillation (APD) has two components $\mathcal{L}_{ob}$ and $\mathcal{L}_{rec}$. $\mathcal{L}_{ob}$ accomplishes the alignment between auxiliary predictions $\boldsymbol{p}_{a,t}$ and $\boldsymbol{p}_{a,s}$ (i.e., observations) obtained from the adaptive perspectives, while $\mathcal{L}_{rec}$ rectifies student view $\mathcal{A}_s$, making it similar to $\mathcal{A}_t$ of teacher. Because the adaptive $\mathcal{A}_s$ encodes more specific semantic details for each image than the fixed $\mathcal{C}_s$, the produced $\boldsymbol{p}_{a,s}$ are generally more accurate than $\boldsymbol{p}_s$ obtained from

TABLE 3: Efficiency comparison on the test set of Cityscapes. Teacher model is PSPNet [47] with ResNet-101. RN, MN2 and EN represent ResNet [8], MobileNet-V2 [26] and EfficientNet [28] respectively.

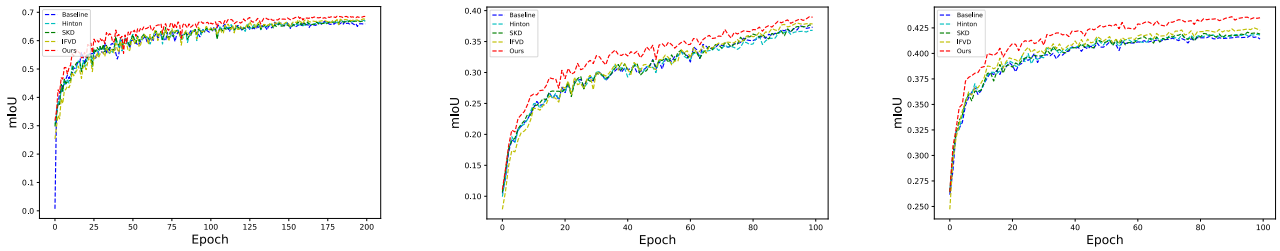
Methods	*test* mIoU	Params (M)	FLOPS (G)
ENet [22]	58.3	0.3580	3.612
ESPNet [18]	60.3	0.3635	4.422
FCN [27]	65.3	134.5	333.9
ERFNet	68.0	2.067	25.60
ICNet [46]	69.5	26.50	28.30
RefineNet	73.6	118.1	525.7
PSPNet [47]	78.4	70.43	574.9
RN-18 + SKD	72.9	16.31	148.2
RN-18 + IFVD	73.2	16.31	148.2
RN-18 + Ours	**74.9**	16.31	148.2
MN2-1.0 + SKD	72.1	4.840	39.44
MN2-1.0 + IFVD	72.0	4.840	39.44
MN2-1.0 + Ours	**73.5**	4.840	39.44
EN-B0 + SKD	73.0	13.44	95.86
EN-B0 + IFVD	73.6	13.44	95.86
EN-B0 + Ours	**75.2**	13.44	95.86

TABLE 4: Performance comparison with state-of-the-art methods using PSPNet on three popular benchmark datasets: Cityscapes [5], ADE20K [49] and PASCAL-Context [19]. Teacher and student models adopt ResNet-101 and ResNet-18 as their backbones.

Methods	Cityscapes	ADE20K	PASCAL-Context
Teacher	78.15	43.44	48.50
Student	74.15	37.19	42.29
+ KD	74.81	37.69	42.45
+ SKD	74.56	37.61	42.53
+ IFVD	74.10	37.89	42.74
+ Ours	**75.68**	**39.25**	**43.96**

TABLE 5: Cross-model distillation results on Cityscapes with PSPNet [47] and DeepLab-V3 [2]. RN, MN2 and EN represent ResNet [8], MobileNet-V2 [26] and EfficientNet [28] respectively. PSPNet $\rightarrow$ DL-V3 means the teacher network is PSPNet and the student is DeepLab-V3, and vise versa.

Method	Backbone	PSPNet $\rightarrow$ DL-V3	DL-V3 $\rightarrow$ PSPNet
Teacher	RN-101	78.15	78.47
Student-I	RN-18	74.15	74.47
+ KD	RN-18	75.13	73.50
+ SKD	RN-18	75.65	73.67
+ IFVD	RN-18	75.42	74.29
+ Ours	RN-18	**76.01**	**75.90**
Student-III	MN2-1.0	71.34	71.40
+ KD	MN2-1.0	71.81	71.57
+ SKD	MN2-1.0	72.45	71.74
+ IFVD	MN2-1.0	70.97	72.54
+ Ours	MN2-1.0	**73.22**	**73.66**
Student-IV	EN-B0	72.30	71.54
+ KD	EN-B0	72.66	73.73
+ SKD	EN-B0	72.29	73.69
+ IFVD	EN-B0	72.87	74.06
+ Ours	EN-B0	**75.03**	**75.51**
Teacher	MN2-1.0	71.34	71.40
Student	MN2-0.5	63.34	63.89
+ KD	MN2-0.5	64.42	64.88
+ SKD	MN2-0.5	64.11	64.47
+ IFVD	MN2-0.5	64.27	64.36
+ Ours	MN2-0.5	**67.14**	**66.90**



(a) Cityscapes (b) ADE20K (c) PASCAL-Context

Fig. 5: Validation mIoU curves on Cityscapes, ADE20K and PASCAL-Context. Our proposed APD (colored in red) consistently outperforms other methods throughout the training process. The teacher is PSPNet with ResNet-101 and the student is PSPNet with ResNet-18.

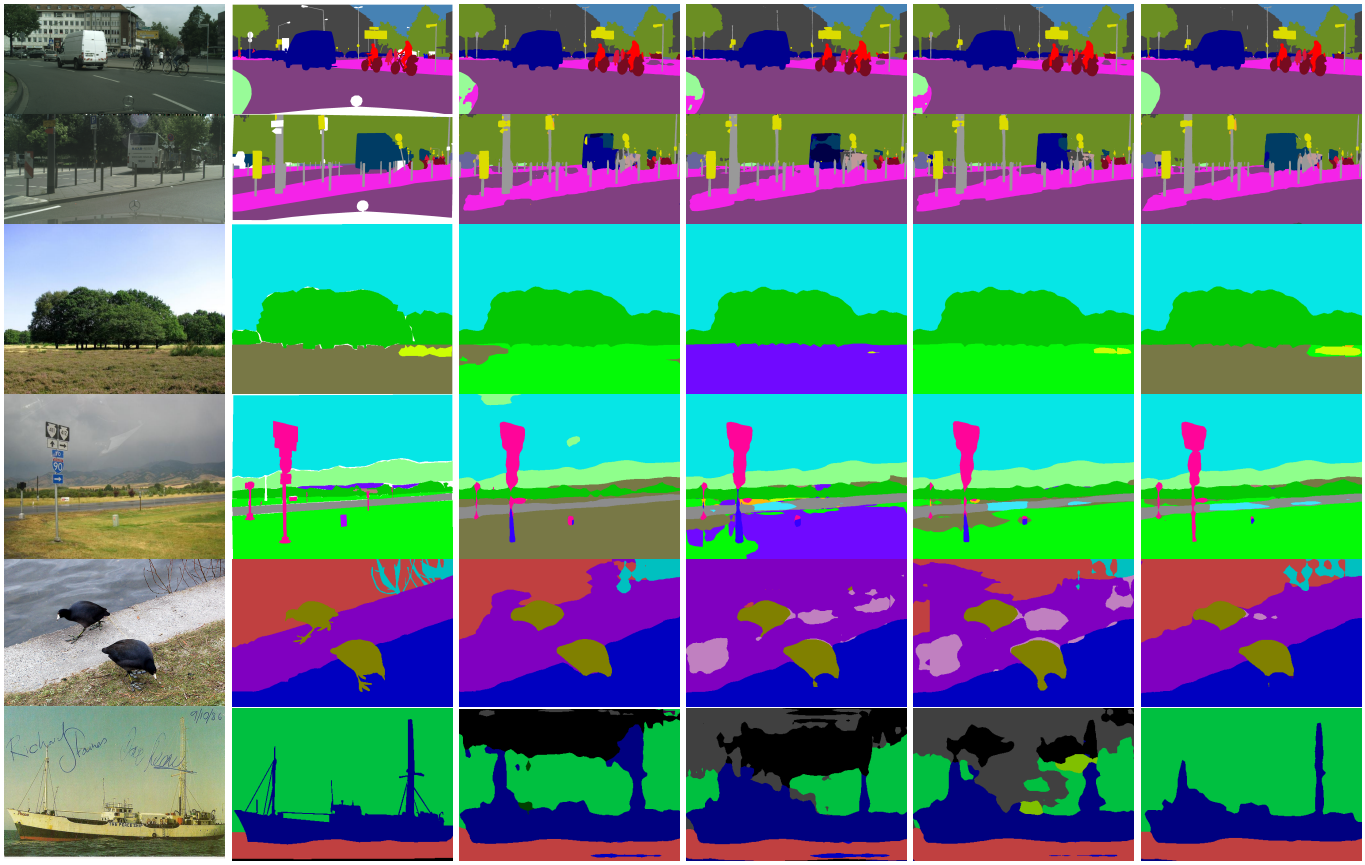


Image GT KD IFVD SKD Ours

Fig. 6: Visual comparison on Cityscapes, ADE20K and PASCAL-Context. White regions in GT are ignored during evaluation.

$\mathcal{C}_s$, as demonstrated in Fig. 3. Results in Table 6 show that the observation alignment and perspective rectification are both indispensable.

Different perspectives result in varying observations. Thus perspective rectification is helpful for the observation alignment as proved by Exp.III & Exp.VI and Exp.IV & Exp.VII. However, without observation alignment $\mathcal{L}_{ob}$, implementing $\mathcal{L}_{rec}$ alone with $\mathcal{L}_{kd}$ in Exp.V only slightly improves the performance of Exp.II. On the other hand, merely applying observation alignment via $\mathcal{L}_{ob}$ achieves decent improvement as shown by Exp.II & Exp.III. When perspectives are rectified by $\mathcal{L}_{rec}$, $\mathcal{L}_{ob}$ boosts performance from 42.88 from 43.96 as shown in Exp.V & Exp.VI.

In Eq. (9), $\boldsymbol{p}_{a,t}$ is used as soft targets to distill knowledge from teacher to student in the proposed APD. An alternative is to replace the soft targets with one-hot labels, denoted as $\mathcal{L}_{ce}^{Local}$ in Table 6, thus Kullback-Leibler divergence in Eq. (9) equals to the standard Cross Entropy Loss. Soft targets encode the "dark knowledge" of teacher and are more informative than one-hot hard labels. Therefore, superior performance has been achieved by $\mathcal{L}_{ob}$ (Exp.III & Exp.VI) compared to $\mathcal{L}_{ce}^{Local}$ (Exp.IV & Exp.VII) in Table 6. While bringing $\mathcal{L}_{ob}$ and $\mathcal{L}_{ce}^{Local}$ together in Exp.VIII is slightly worse than Exp.VII, implying that the benefits of $\mathcal{L}_{ce}^{Local}$ do not outweigh that of $\mathcal{L}_{ob}$. Also, by comparing Exp.VI and Exp.VIII, we can conclude that the hard one-hot label used by $\mathcal{L}_{ce}^{Local}$ might adversely affect the knowledge transfer that is accomplished by $\mathcal{L}_{ob}$ with the soft labels that are more

TABLE 6: Ablation study on PASCAL-Context. Teacher is PSPNet with ResNet-101 and student is PSPNet with ResNet-18. The first column denotes the experiment IDs. $\mathcal{L}_{ob}$ uses with soft targets $\boldsymbol{p}_{a,t}$ as shown in Eq. (9), while $\mathcal{L}_{ce}^{Local}$ means directly applying Cross Entropy loss with one-hot hard targets.

Exp.	$\mathcal{L}_{kd}$	$\mathcal{L}_{ob}$	$\mathcal{L}_{ce}^{Local}$	$\mathcal{L}_{rec}$	mIoU
I	-	-	-	-	42.29
II	✓	-	-	-	42.48
III	✓	✓	-	-	43.32
IV	✓	-	✓	-	42.92
V	✓	-	-	✓	42.88
VI	✓	✓	-	✓	**43.96**
VII	✓	-	✓	✓	43.38
VIII	✓	✓	✓	✓	43.87
IX	-	✓	-	✓	43.81

TABLE 7: Ablation study on PASCAL-Context with PSPNet. Teacher is built upon ResNet-101 and student is with ResNet-18. $\mathcal{L}_{ifv}$ and $\mathcal{L}_{skd}$ are the intra-class feature variation distillation and pair-wise distillation of IFVD and SKD . We reproduce them according to their official implementations. $\mathcal{P}$ means $\mathcal{L}_{ifv}$ and $\mathcal{L}_{skd}$ are applied to the projected features $\boldsymbol{f}_{a,t}$ and $\boldsymbol{f}_{a,s}$.

Exp.	$\mathcal{L}_{kd}$	$\mathcal{L}_{ifv}$	$\mathcal{L}_{skd}$	$\mathcal{P}$	$\mathcal{L}_{ob}$	$\mathcal{L}_{rec}$	mIoU
I	✓	-	-	-	-	-	42.48
II	✓	✓	-	-	-	-	42.74
III	✓	✓	-	✓	-	-	43.02
IV	✓	✓	-	✓	✓	✓	44.05
V	✓	-	✓	-	-	-	42.53
VI	✓	-	✓	✓	-	-	42.39
VII	✓	-	✓	✓	✓	✓	43.98

informative [10]. Besides, Exp.IX shows that even without the normal KD loss $\mathcal{L}_{kd}$, the proposed $\mathcal{L}_{ob}$ and $\mathcal{L}_{rec}$ still achieve decent improvement compared to the baseline results in Exp.I.

Effect of Projectors $\mathcal{P}_t$ and $\mathcal{P}_s$. To generalize our method to different teacher & student models whose output features are with different channels, we use projectors $\mathcal{P}_t$ and $\mathcal{P}_s$ to process the feature maps of teacher and student to the same channels, satisfying the requirement of the similarity calculation in Eq. (8). Otherwise, the perspectives cannot be rectified. Two projectors are only used for training and are simply discarded during inference, boosting student models without structural change.

To show that the improvement of $\mathcal{L}_{rec}$ and $\mathcal{L}_{ob}$ is not caused by the two additional projectors, we implement SKD and IFVD on the projected features ($\boldsymbol{f}_{a,t}$ and $\boldsymbol{f}_{a,s}$) to compare with the performance obtained from the features without projection ($\boldsymbol{f}_t$ and $\boldsymbol{f}_s$). We note that $\mathcal{P}_t$ is still optimized by $\mathcal{L}_t$ in the following experiments for a fair comparison.

Experimental results are presented in Table 7 where the results of IFVD and SKD implemented on the projected features are comparable to that without projectors as shown in Exp.II & Exp.III and Exp.V & Exp.VI. Besides, the proposed $\mathcal{L}_{rec}$ and $\mathcal{L}_{ob}$ are still complementary to the models implemented with IFVD and SKD, proved by Exp.IV and Exp.VII in Table 7.

Sensitivity analysis. Different hyper-parameters may cause performance variation. Thus we conduct sensitivity analysis in Table 8 where the best performance is robust to different values of λ_{rec} and $1/\tau$ within the range of 5-20.

TABLE 8: Sensitivity analysis with different values of λ_{rec} and τ. Experimental results are obtained with PSPNet-RN101 (teacher) and PSPNet-RN18 (student) on PASCAL-Context.

Values	0.1	1	5	10	20	50	100
λ_{rec}	43.32	43.54	43.70	43.96	43.95	43.71	43.48
$1/\tau$	42.05	42.92	43.79	43.96	43.62	43.36	43.09

TABLE 9: Comparison on PASCAL-Context between cosine similarity and dot product for observation generation. The teacher is PSPNet with ResNet-101 and the student is PSPNet with ResNet-18. "Main-Cos" means the main perspective (classifier) adopts cosine similarity for prediction and "Adapt-Cos" means the adaptive one uses cosine similarity. Thus "Adapt-Cos" can only be adopted by APD.

Method	Main-Cos	Adapt-Cos	mIoU
Baseline-I (Default)		N/A	42.29
Baseline-II	✓	N/A	41.90
KD-I (Default)		N/A	42.48
KD-II	✓	N/A	42.03
APD-I			38.04
APD-II (Default)		✓	**43.96**
APD-III	✓	✓	43.24

TABLE 10: Different values of τ_m for the baseline model implemented with "Main-Cos".

$1/\tau_m$	10	20	30	40	50
Baseline-II	40.09	41.56	41.88	**41.90**	41.62
KD-II	40.22	41.79	41.92	**42.03**	41.88
APD-III	40.99	43.15	43.09	**43.24**	43.11

4.5 Cosine Similarity in APD

In segmentation models, the universal perspective $\mathcal{C}$ applies dot product on the features $\boldsymbol{f}$ yielded by the feature generator $\mathcal{G}$ to produce the observation $\boldsymbol{p}$, while in the proposed APD, the adaptive perspective $\mathcal{A}$ generates observations $\boldsymbol{p}_a$ via cosine similarity. The difference between cosine similarity and dot product is that the former measures the angle between two vectors and the latter takes both the angle and magnitude into account.

Experimental results. Both cosine similarity and dot product seem to be feasible for yielding observation, while we find that cosine similarity is more suitable for the adaptive perspective in APD. Results are shown in Table 9. Specifically, by comparing models of "Baseline" and "KD", it can be found that applying cosine similarity to the main universal perspective $\mathcal{C}$ (i.e., Main-Cos) is detrimental to the overall performance. On the other hand, "Main-Cos" also causes performance deduction on the proposed APD, shown by "APD-II" and "APD-III". As for "Adapt-Cos" that can only be adopted by the proposed APD, it is necessary for APD since the performance drops from 43.96 ("APD-II") to 38.04 ("APD-I") if the adaptive perspective does not exploit the cosine similarity for yielding observations.

In summary, through the experiments in Table 9, we empirically find that the dot-product is more suitable for the universal perspective (i.e., normal classifier) and the cosine similarity is better for the proposed adaptive perspective.

Analysis. The performance discrepancy between "Main-Cos" and "Adapt-Cos" might be related to the formation processes of

TABLE 11: Object detection results on COCO 2017 *val*.

Method	Backbone	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Teacher	ResNet101	42.04	62.48	45.88	25.22	45.55	54.60
Student	ResNet18	33.26	53.61	35.26	18.96	35.68	43.16
KD	ResNet18	33.68	54.10	35.93	19.65	36.17	43.22
FitNet	ResNet18	34.13	54.16	36.71	18.88	36.50	44.69
FGFI	ResNet18	34.16	54.43	36.60	18.79	36.57	44.97
SKD	ResNet18	33.97	54.66	36.62	18.71	36.67	44.14
IFVD	ResNet18	34.20	54.63	36.66	19.16	36.65	44.71
Ours	ResNet18	**35.47**	**56.68**	**38.00**	**20.41**	**38.17**	**46.14**
Teacher	ResNet50	40.22	61.02	43.81	24.16	43.53	51.98
Student	MobileV2	29.47	48.87	30.90	16.33	30.77	38.86
KD	MobileV2	30.13	50.28	31.35	16.69	31.91	39.56
FitNet	MobileV2	30.20	49.80	31.69	16.39	31.64	39.69
FGFI	MobileV2	30.27	49.87	31.60	17.03	31.82	40.06
SKD	MobileV2	31.52	50.72	33.35	17.66	33.52	40.75
IFVD	MobileV2	30.67	50.30	32.43	17.09	33.62	38.38
Ours	MobileV2	**32.58**	**53.23**	**34.41**	**19.12**	**34.66**	**42.35**

TABLE 12: Instance segmentation results on COCO 2017 *val* set. The results are measured in box mAP and mask mAP.

Method	Backbone	mAP^{box}	mAP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{s}	AP^{mask}_{m}	AP^{mask}_{l}
Teacher	ResNet101	42.90	38.63	60.45	41.28	19.48	41.33	55.29
Student	ResNet18	33.98	31.25	51.07	33.10	14.18	32.80	45.53
KD	ResNet18	34.53	31.66	51.85	33.59	14.80	33.38	45.73
FitNet	ResNet18	34.69	31.75	51.46	33.82	14.50	33.25	46.76
FGFI	ResNet18	34.73	31.85	51.59	33.72	14.95	33.25	46.94
SKD	ResNet18	34.53	31.62	51.90	33.54	14.48	33.44	46.10
IFV	ResNet18	34.59	31.64	52.06	33.38	14.93	33.49	46.34
Ours	ResNet18	**35.90**	**32.84**	**53.70**	**34.71**	**15.77**	**34.79**	**47.81**

the universal perspective that is shared by all training images and the adaptive perspective that is created individually. The shared universal perspective approaches to an optimal magnitude by well-fitting the entire training set. The magnitude values of features serve as additional descriptors, revealing more information for individual feature vectors. Therefore, the universal perspective, with well-learned class-wise magnitude, achieves better performance by adopting the dot product. However, the magnitude of the adaptive perspective is determined by the individual feature map and thus the magnitude might be biased towards the feature vectors with large magnitude, causing inappropriate representation for those features with low magnitude. Also, the magnitude values of features belonging to the same category vary in different images due to the varying co-occurrent contextual information. Thus we instead only focus on the semantic relation by adopting cosine similarity to alleviate the issues caused by the magnitude instability of the adaptive perspective that is formed merely based on individual samples.

Besides, it is worth noting that, since the purposes of "Main-Cos" and "Adapt-Cos" are different, we have carefully tuned the values of the scaler τ_m for "Main-Cos" to have a fair comparison with "Adapt-Cos" in Table 9. Specifically, according to the sensitive analysis in Table 8, τ of "Adapt-Cos" is set to 0.1 (i.e., $1/\tau = 10$) , while directly applying $\tau_m = 0.1$ to "Main-Cos" significantly worsens the performance as shown in Table 10 where $1/\tau_m = 40$ (i.e., $\tau_m = 0.025$) achieves the best performance. Thus models with "Main-Cos" in Table 9 are implemented with $\tau_m = 0.025$.

4.6 Extensions

Although our method is motivated from the perspective of semantic segmentation tasks, it also generalizes well to the tasks of object detection and instance segmentation. Implementation details and results are presented as follows.

4.6.1 Object Detection

Implementation details. We use the most popular Faster-RCNN-FPN detector in Detectron2 [34] with different backbones as our strong baselines. We use the standard training policies provided in Detectron2 excepts for the number of GPUs. The original models in Detectron2 are trained using 8 GPUs. The official $1\times$ training policy is to train 90,000 iterations with 16 images per batch. The learning rate is initialized as 0.02 and decayed by 10 at 60,000 and 80,000 iterations. The baseline and other models are trained on 4 GPUs, thus we halve the batch size to 8 and double the total iterations to 180,000. The initial learning rate is 0.01 and it decays by 10 at 120,000 and 160,000 iterations. Our reproduction yields similar baseline performance and costs the same overall GPU time. We use the standard multi-scale training augmentations. The input images are randomly resized to one of the sizes $\{640, 672, 704, 736, 768, 800\}$ and then images are randomly horizontal fliped with a probability of 0.5. We do NOT use any augmentations during the inference.

We reimplement the KD loss proposed by Hinton *et al.*on the logits of the classification branch in the RoI head. The loss weight is also set to 10. We notice that, in object detection, the teacher and student may have different proposals, causing a mismatch between the features after the RoI Align operation as well as the final predicted logits. To address this issue, since only the

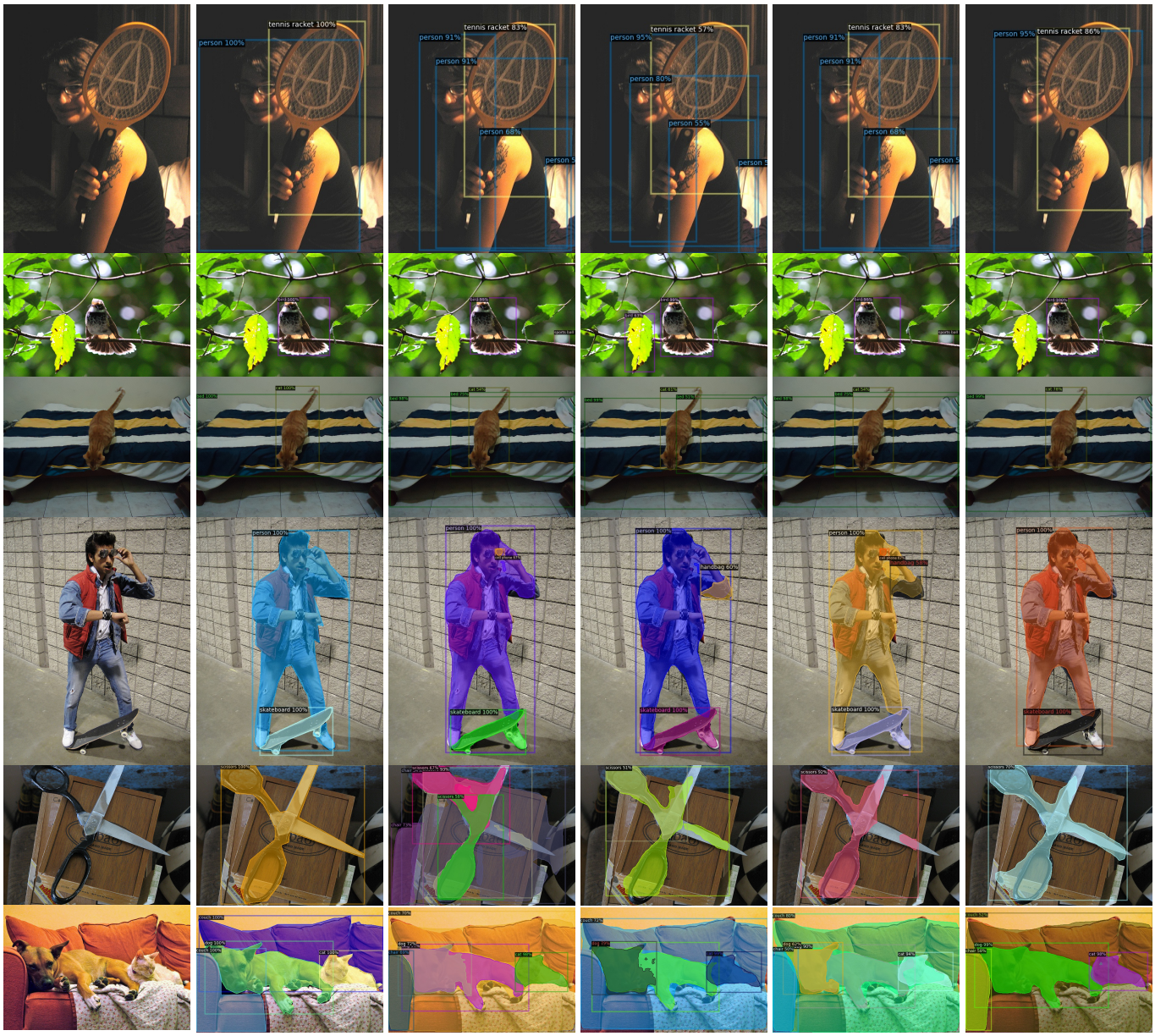


Fig. 7: Visual comparison of object detection (first three rows) and instance segmentation (last three rows) on COCO2017 *val* set.

student's proposals are used for generating the final task losses, we let the teacher network adopt the proposals yielded by the student, thus the teacher's features and logits are aligned with that of the student.

FGFI is a distillation method specifically designed for object detection. However, the baseline method used in FGFI is relatively weaker than the popular ones. So we reimplement FGFI on our stronger baseline according to the official code provided by the authors.

We apply the proposed APD to the features after the RoI Align operation. We simulate the scenario in the semantic segmentation tasks and assume every feature vector in the feature map belongs to the class of the corresponding proposal. Then we consider all proposals in a mini-batch as a whole and generate adaptive perspectives in a batch-wise manner.

SKD on the detection task is reimplemented based on its official code. We apply the SKD loss on the features after the FPN structure with a 2×2 down-sampling, following its default configurations. Similarly, the IFVD loss is reimplemented using the official code. Since class labels are required by IFVD, we apply the IFVD loss on the features after the RoI Align operation. Our code for object detection will also be made publicly available.

Results. We summarize our results on COCO [15] with the Faster-RCNN-FPN [23] detector in Table 11. We re-implement the classic distillation methods KD and FitNet, as well as one recent method FGFI [31] that achieves state-of-the-art performance for distillation in object detection. Moreover, to comprehensively compare with the methods in semantic segmentation, we also apply SKD and IFVD to the object detection task, since both segmentation and detection tasks require structured dense prediction. It can be observed in Table 11 that our method still outperforms all other methods by a large margin on the detection task, including

FGFI that is specifically designed for detection. The results on detection further demonstrate the superiority and generalization ability of our method.

We present the qualitative comparison between SKD and IFVD on COCO2017 *val* set in Fig. 7 where it is observed that our predictions are generally better than the others.

4.6.2 Instance Segmentation

We further adapt our method to the instance segmentation task on COCO 2017 dataset. Instance segmentation is a more challenging task aiming to segment every object in each image. The Mask-RCNN with FPN in Detectron2 is adopted as our baseline. The training process of instance segmentation is similar to that of object detection, following the standard training policies provided in Detectron 2.

The results are summarized in Table 12 where our method improves the results of instance segmentation task by a large margin, while the other related methods barely improve the baseline performance. The challenging instance segmentation task further demonstrates the superiority of our proposed method. The qualitative comparison on COCO2017 *val* set is shown in Fig. 7.

5 Conclusion

We have presented the proposed Adaptive Perspective Distillation (APD). Different from the previous distillation methods that distill knowledge via pixel-wise predictions obtained by the fixed perspective (i.e., classifier), APD aims at creating adaptive perspectives for individual samples, revealing more details on the encoded feature for helping student models achieve better performance. APD has no structural constraints on the base model and thus can be easily applied to normal semantic segmentation frameworks. APD is also complementary to other existing knowledge distillation methods in segmentation. The extensive comparison with state-of-the-art knowledge distillation methods for semantic segmentation demonstrate the effectiveness and generalization ability of APD.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 2017.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[4] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[7] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *CVPR*, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. 2019.

[10] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv*, 2015.

[11] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*, 2020.

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[13] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.

[14] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019.

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[16] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *arXiv*, 2015.

[17] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *TPAMI*, 2020.

[18] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda G. Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *ECCV*, 2018.

[19] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[20] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[21] Nikolaos Passalis and Anastasios Tefas. Probabilistic knowledge transfer for deep representation learning. *arXiv*, 2018.

[22] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv*, 2016.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[24] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[26] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[27] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2017.

[28] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.

[30] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[31] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019.

[32] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[33] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *ECCV*, 2020.

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[35] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[36] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *ECCV*, 2020.

[37] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[38] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation. *arXiv*, 2020.

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018.

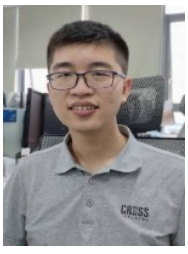
[40] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[41] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020.

[42] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv*, 2018.
[43] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.
[44] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018.
[45] Hengshuang Zhao. semseg. https://github.com/hszhao/semseg, 2019.
[46] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018.
[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.
[48] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018.
[49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

Zhuotao Tian received the B.Eng. degree (Honors) in Computer Science from the School of Computer Science and Technology, Harbin Institute of Technology (HIT) in 2018. He is currently a 3rd year Ph.D. student at the Chinese University of Hong Kong (CUHK), under the supervision of Prof. Jiaya Jia. He serves as a reviewer for IJCV, CVPR, ICCV, ECCV, AAAI. His research interests include few-shot learning, semi-supervised learning, semantic segmentation and scene text detection.

Pengguang Chen received the B.Eng. degree in Computer Science from the Department of Computer Science and Technology, Nanjing University in 2018. He is currently a 3rd year Ph.D. student at the Chinese University of Hong Kong (CUHK), under the supervision of Prof. Jiaya Jia. He serves as a reviewer for CVPR, ICCV, ECCV. His research interests include neural architecture search, self-supervised learning, knowledge distillation and semantic segmentation.



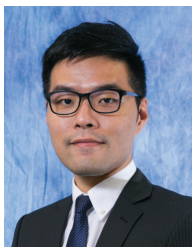
Xin Lai received the B.Eng. degree in computer science and technology from Harbin Institute of Technology (HIT) in 2020. He is currently a first-year Ph.D student in computer science and engineering department of the Chinese University of Hong Kong (CUHK). His research interests focus on computer vision and deep learning, especially on semi-supervised learning, few-shot learning and domain generalization techniques.

Li Jiang received her B.S. degree in Computer Science and Technology from Harbin Institute of Technology, China in 2017. She is currently a Ph.D. student in the Department of Computer Science and Engineering at The Chinese University of Hong Kong. She serves as a reviewer for CVPR, ICCV, ECCV. Her research interests include computer vision, semantic/instance segmentation and 3D scene understanding.

Shu Liu now serves as Co-Founder and Technical Head in SmartMore. He received the BS degree from Huazhong University of Science and Technology and the PhD degree from the Chinese University of Hong Kong. He was the winner of 2017 COCO Instance Segmentation Competition and received the Outstanding Reviewer of ICCV in 2019. He continuously served as a reviewer for TPAMI, CVPR, ICCV, NIPS, ICLR and etc. His research interests lie in deep learning and computer vision.

Hengshuang Zhao is a postdoctoral researcher at the University of Oxford. He received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2019. His team won champions of ImageNet Scene Parsing Challenge, LSUN Semantic Segmentation Challenge and WAD Drivable Area Segmentation Challenge at ECCV'16, CVPR'17, and CVPR'18 respectively. He is recognized as outstanding/top reviewers of ICCV'19 and NeurIPS'19. He received the rising star award at the world artificial intelligence conference 2020. His general research interests cover the broad area of computer vision and machine learning, with special emphasis on high-level scene recognition and pixel-level scene understanding. He is a member of IEEE.

Bei Yu received the Ph.D. degree from The University of Texas at Austin in 2014. He is currently an Associate Professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. He has served as TPC Chair of ACM/IEEE Workshop on Machine Learning for CAD, and in many journal editorial boards and conference committees. He is the Editor of IEEE TCCPS Newsletter. He received seven Best Paper Awards from ASPDAC 2021, ICTAI 2019, Integration, the VLSI Journal in 2018, ISPD 2017, SPIE Advanced Lithography Conference 2016, ICCAD 2013, ASPDAC 2012, and six ICCAD/ISPD contest awards. He is a member of IEEE.



Ming-Chang Yang received his B.S. degree in Department of Computer Science from National Chiao-Tung University, Hsinchu, Taiwan, in 2010. He received his Master and Ph.D. degrees in Graduate Institute of Networking and Multimedia from National Taiwan University, Taipei, Taiwan, in 2012 and 2016, respectively. Now he is an assistant professor at the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His primary research interests include emerging non-volatile memory and storage technologies, memory and storage systems, and next-generation memory/storage architecture designs. He is a member of IEEE.

Jiaya Jia received the Ph.D. degree in Computer Science from Hong Kong University of Science and Technology in 2004 and is currently a full professor in Department of Computer Science and Engineering at the Chinese University of Hong Kong (CUHK). He assumes the position of Associate Editor-in-Chief of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and is in the editorial board of International Journal of Computer Vision (IJCV). He continuously served as area chairs for ICCV, CVPR, AAAI, ECCV, and several other conferences for the organization. He was on program committees of major conferences in graphics and computational imaging, including ICCP, SIGGRAPH, and SIGGRAPH Asia. He is a Fellow of the IEEE.