Multi-Granularity Distillation Scheme Towards Lightweight Semi-Supervised Semantic Segmentation

Jie Qin[1,2,3†⋆], Jie Wu[2†‡], Ming Li[2], Xuefeng Xiao[2],
Min Zheng[2], and Xingang Wang[3‡]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] ByteDance Inc
[3] Institute of Automation, Chinese Academy of Sciences

Abstract. Albeit with varying degrees of progress in the field of Semi-Supervised Semantic Segmentation, most of its recent successes are involved in unwieldy models and the lightweight solution is still not yet explored. We find that existing knowledge distillation techniques pay more attention to pixel-level concepts from labeled data, which fails to take more informative cues within unlabeled data into account. Consequently, we offer the first attempt to provide lightweight SSSS models via a novel multi-granularity distillation (MGD) scheme, where multi-granularity is captured from three aspects: i) complementary teacher structure; ii) labeled-unlabeled data cooperative distillation; iii) hierarchical and multi-levels loss setting. Specifically, MGD is formulated as a labeled-unlabeled data cooperative distillation scheme, which helps to take full advantage of diverse data characteristics that are essential in the semi-supervised setting. Image-level semantic-sensitive loss, region-level content-aware loss, and pixel-level consistency loss are set up to enrich hierarchical distillation abstraction via structurally complementary teachers. Experimental results on PASCAL VOC2012 and Cityscapes reveal that MGD can outperform the competitive approaches by a large margin under diverse partition protocols. For example, the performance of ResNet-18 and MobileNet-v2 backbone is boosted by **11.5%** and **4.6%** respectively under 1/16 partition protocol on Cityscapes. Although the FLOPs of the model backbone is compressed by **3.4-5.3×** (ResNet-18) and **38.7-59.6×** (MobileNetv2), the model manages to achieve satisfactory segmentation results.

Keywords: Semi-supervised Semantic Segmentation, Lightweight, Distillation, Multi-granularity

1 Introduction

Semantic segmentation is a fundamental and crucial task due to extensive vision applications such as scene understanding and autonomous driving. However,

---

†Equal contribution. ⋆This work was done while Jie Qin interned at ByteDance.
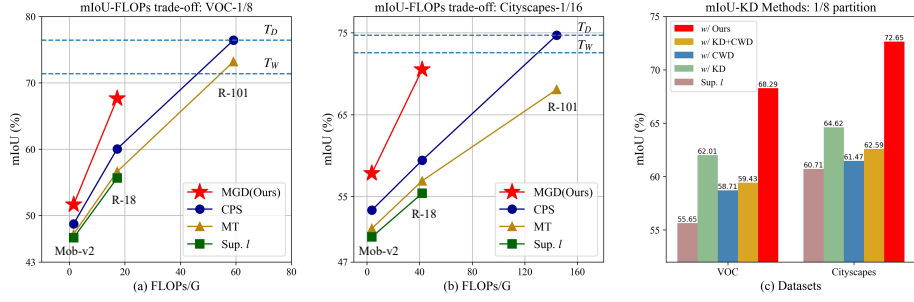‡Corresponding author. Code is available at github.com/JayQine/MGD-SSSS.

Fig. 1: (a)/(b) The mIoU-FLOPs trade-off between MGD and existing methods including Sup. l , MT [?] and CPS [?]. T_D and T_W denote two complementary teacher models. Sup. l denotes the model trained only with the labeled data under the corresponding partition. (c) The results of existing distillation methods (KD [?] and CWD [?]) for optimizing lightweight model of ResNet-18

this task is still extremely dependent on adequate granular pixel-level annotations [?, ?, ?, ?], which requires a huge amount of manual effort. To alleviate such expensive and unwieldy annotations, some researchers attempt to address this task in the semi-supervised paradigm, where the model is merely accessible to a small amount of image-label pairs joint with abundant unlabeled images. This is an exceedingly favorable setting since such coarse unlabeled images are more readily available on the internet. The main challenge under this paradigm is how to take good advantage of unlabeled data to alleviate the drastic drop in performance when the labeled data is reduced. The recent state-of-the-art semi-supervised semantic segmentation approaches [?,?,?,?] mostly benefit from pseudo labeling and consistency regularization.

Despite semi-supervised semantic segmentation (SSSS) having witnessed prevailing success in a series of computer vision applications, most previous approaches [?, ?, ?] have resorted to unwieldy and cumbersome models. It is arduous to popularize these models that require tremendous computational costs, which becomes a critical bottleneck when such models are required to deploy on resource-constrained server platforms or even more lightweight mobile devices. As shown in Fig. ??, the performance of the existing competitive approaches suffers from drastic performance degradation when replacing the cumbersome backbone with a lightweight one such as ResNet-18 [?] or MobileNet-v2 [?]. We also leverage the existing distillation methods [?, ?, ?] to optimize lightweight models but also get unimpressive results. It may due to the fact that existing distillation methods mainly transfer pixel-level knowledge, which lacks diverse and hierarchical distillation signals. Furthermore, these approaches are designed for labeled data and fail to take the data characteristic of unlabeled data into account, hence they are not well adapted to the SSSS task. These observations motivate us to raise a question: *how to obtain a lightweight SSSS model with*

a lighter computation cost joint with high performance? Compared with traditional semi-supervised semantic segmentation, lightweight solution confronts three additional challenges: (i) How to design teacher models to efficiently assist in boosting the capacity of the lightweight model? (ii) How to fully leverage both limited labeled and sufficient unlabeled data in the semi-supervised distillation setting? (iii) Are there more granularity concepts that can be resorted to refine the lightweight model for this fine-grained task?

To answer these questions, we propose a novel *Multi-Granularity Distillation (MGD)* scheme to obtain a lightweight SSSS model with impressive performance and lite computational costs. This scheme elaborately designs triplet-branches to leverage complementary teacher models and hierarchical auxiliary supervision to distill the lightweight model. Specifically, we capture multi-level concepts to formulate the hierarchical optimization objectives, which can be regarded as diverse granularity supervision signals and play a key role to break through the bottleneck of the capacity of the lightweight student model. The contributions of this work are summarized into four folds:

– To the best of our knowledge, we offer the first attempt to obtain the lightweight model for semi-supervised semantic segmentation. We provide efficient solutions for the challenges of this newly-raised task and alleviate the issue of performance degradation during model compression. We believe our work provides the foundation and direction for lightweight research in the field of semi-supervised semantic segmentation.
– We propose a novel multi-granularity distillation (MGD) scheme that employs triplet-branches to distill task-specific concepts from two complementary teacher models into a student one. The deep-and-thin and shallow-and-wide teachers help to provide comprehensive and diverse abstractions to boost the lightweight model.
– The labeled-unlabeled data cooperative distillation is designed to make full use of the characteristics of extensive unlabeled data in the semi-supervised setting. Then we design a hierarchical loss paradigm to provide multi-level (image-level/region-level/pixel-level) auxiliary supervision for unlabeled data, which supervises the lightweight model to grasp the high-level categorical abstraction and low-level contextual information.
– Extensive experiments on PASCAL VOC2012 and Cityscape demonstrate that MGD can outperform the competitive approaches by a large margin under diverse partition protocols. Although the FLOPs of the model backbone is compressed by **3.4-5.3×** (ResNet-18) and **38.7-59.6×** (MobileNetv2), the model still can achieve satisfactory performance.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation is a fundamental and crucial task in the field of computer vision [?,?,?,?]. It aims to perform a pixel-level prediction to cluster parts of an

image together that belong to the same object class. Many approaches [?, ?, ?] employ the encoder-decoder structure to resume the resolution of features step by step and accomplish pixel-level classification simultaneously. In order to capture the long-range relationships, DeepLab v3+ [?] make use of dilated convolution [?] to increase the receptive field of networks. [?, ?, ?] use the pyramid pooling to capture multi-scale features and integrate pyramid features to obtain global context information. [?, ?, ?, ?, ?] use the attention mechanism to aggregate the global contexts that benefit dense prediction. HRNet [?] maintains the context of images by holding the resolution of features.

2.2   Semi-Supervised Learning

Semi-supervised learning aims to utilize limit labeled data and a large amount of unlabeled data to improve the model capacity [?, ?, ?, ?]. The consistency regularization technique applies different perturbations or augmentations and encourages the model to produce invariant outputs [?, ?, ?]. For example, VAT [?] applies adversarial perturbations to the output and Temporal Model [?] incorporates the outputs over epochs and makes them consistent. Data augmentations or mixing are performed to generate the different perturbed data to train networks [?, ?]. FixMatch [?] takes advantage of the weak augmentation to generate the pseudo labels and supervise the images with strong augmentation. MT [?] generates the predictions by exponential moving average (EMA). Self-supervised learning [?], GAN-based methods [?, ?, ?] and self-training [?, ?] are also exploited to assist semi-supervised learning.

2.3   Semi-Supervised Semantic Segmentation

Semi-supervised semantic segmentation is proposed to accomplish the segmentation task with limited image-label pairs and ample unlabeled images. GAN-based methods [?, ?, ?] are proposed to synthesize additional training data or generate the pseudo labels by using adversarial loss on predictions. From the perspective of consistency regularization, data augmentations [?, ?], feature space pertubations [?] and network initialization [?, ?] are applied to generate the different types of predictions and compute consistency losses. CCT [?] exploits the consistent supervision on the outputs of the main decoder and several auxiliary decoders with different perturbations. GCT [?] employs two segmentation networks with the same structure but different initialization and enforces the consistency between the predictions. PC^{2}Seg [?] and [?, ?] take advantage of the contrastive learning to improve the representation ability on the unlabeled data. PseudoSeg [?] uses the weak augmented images to generate the pseudo labels, which are used to supervise the strong augmented images.

2.4   Knowledge Distillation

Knowledge Distillation [?, ?] is a knowledge transfer technique that optimizes a lightweight student model with effective information transfer and supervision

of a larger teacher model or ensembles. Besides the knowledge transfer in the outputs, the feature maps [**?**, **?**] in the intermediate layers of networks are used to improve the performance of the student networks. Moreover, some methods attempt to transfer the attention concepts [**?**, **?**, **?**] of feature maps from each channel in intermediate layers. Some approaches try to exploit the knowledge distillation on the dense prediction tasks [**?**,**?**,**?**,**?**]. In addition, [**?**] indicates that multiple teacher networks can provide more effective information for training a lightweight and ascendant student network. In this paper, we design a deeper and a wider model to play as the complementary teachers and provide multi-granularity auxiliary supervision.

3   Methodology

In this section, we first illustrate the overview of the proposed multi-granularity distillation (MGD) framework. Then the complementary teacher distillation and the novel labeled-unlabeled data cooperative distillation are presented in Sec. **??** and Sec. **??**, respectively. The hierarchical and multi-level distillation for unlabeled data is introduced Sec. **??**.

3.1   Multi-granularity Distillation Scheme

Due to limited labeled data and restricted learning capacity of lightweight models, conventional consistency regularization approaches [**?**,**?**,**?**] suffer from a drastic degradation in performance when they are directly applied on the lightweight models (refer to Fig. **??**). Furthermore, we also tried a series of existing distillation methods to relieve significant performance degradation but found that they are not so effective under the setting of semi-supervised segmentation. Therefore, we propose a multi-granularity distillation scheme to leverage task-oriented knowledge distillation techniques to guide the learning procedure of the lightweight model, as shown in Fig. **??**. Specifically, multi granularity denotes that the distillation process is carried out synergistically from diverse aspects. Firstly, we design dual teachers to provide complementary distillation abstractions. Secondly, we formulate a labeled-unlabeled data cooperative distillation to optimize the network via the data characteristics of semi-supervision. Thirdly, we employ image-level, region-level and pixel-level visual cues from complementary teacher models to formulate a hierarchical and multi-level optimization objective.

3.2   Complementary Teacher Distillation

Solution for "How to design teacher models to efficiently assist in boosting the capacity of the lightweight model?"

Although [**?**] has proved that multiple teacher models can provide more effective information for students, the research on diverse teacher structures is still insufficient. For the fine-grained task of semantic segmentation, complementary
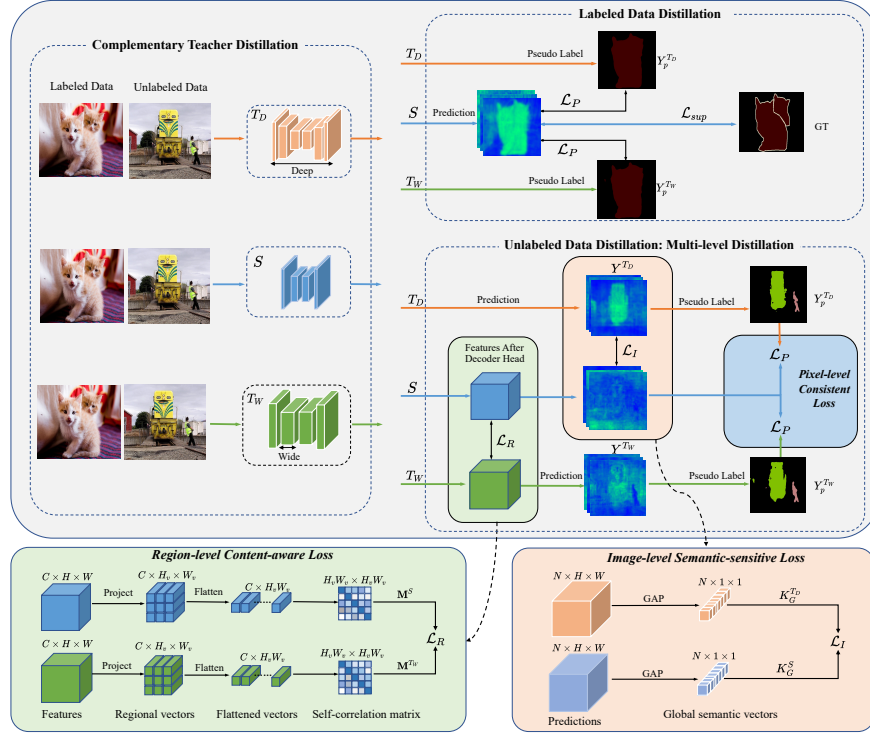
Fig. 2: The pipeline of Multi-Granularity Distillation (MGD) Scheme. MGD is formulated as triplet-branches, where there are two structurally-complementary teacher models (*i.e.*, T_D and T_W) and a target student model S. T_D indicates the deep-and-thin teacher model, T_W denotes the shallow-and-wide teacher model, and the student network S is a shallow and thin model. Image-level semantic-sensitive loss, region-level content-aware loss, and pixel-level consistency loss are designed to form the hierarchical and multi-level loss setting

teacher networks can bring two advantages: i) help to capture more comprehensive visual cues and boost the performance of the student model from different aspects; ii) suppress over-fitting issue that is common in the distillation. Hence in this paper, the teacher model is expanded from two complementary dimensions compared to the student model, *i.e.*, depth and width. Specifically, the complementary teachers contain a deep-and-thin teacher model T_D (such as ResNet-101 [**?**]) and a wide-and-shallow teacher model T_W (such as Wide ResNet-34 [**?**]). These two complementary teacher models can achieve comparable performance while tending to contain diverse potential characteristics.

T_D has proven its ability in extracting high-level semantic and global categorical abstractions [**?**], which helps to achieve promising results in the classification-oriented task. Hence, we attempt to extract the global semantic knowledge $K_G^{T_D}$

from T_D to enhance the global perception ability of the student model. Furthermore, T_W is conducive to capturing the diverse content-aware information due to the sufficient channels, which is beneficial for modeling the contextual relationship between regions. In this paper, we first follow CPS [?] to pre-train two teacher models and freeze their parameters to conduct an offline distillation procedure for refining the student model.

3.3   Labeled-unlabeled Data Cooperative Distillation

Solution for "How to fully leverage both limited labeled and sufficient unlabeled data in the semi-supervised distillation setting?"

The existing distillation techniques are mainly carried out under the full supervision setting and lack of distillation practice for unlabeled data. So we involve a novel labeled-unlabeled data cooperative distillation scheme in MGD. As for labeled data (X_l) distillation, the student model is optimized by ground truth labels $\hat{Y}$ with the supervision loss $\mathcal{L}_{Sup}^l$:

$$\mathcal{L}_{Sup}^l(Y, \hat{Y}) = -\frac{1}{H \times W} \sum_{i=1}^{H \times W} (\hat{y}_i log(y_i)) , \tag{1}$$

where y_i represents the prediction of the student model and the $\hat{y}_i$ illustrates the ground truth label. Superscript l represents labeled data.

Furthermore, we design the pixel-level consistency loss to maintain that the multiple predictions of the same input are consistent. Specifically, both output of the teacher model T_D and T_W can be regarded as pseudo label Y_p for the student model S. Pixel-level consistency loss is set as:

$$\begin{aligned} \mathcal{L}_P^l(Y, Y_p) &= \mathcal{L}_{ce}(Y, Y_p^{T_D}) + \mathcal{L}_{ce}(Y, Y_p^{T_W}) \\ &= -\frac{1}{H \times W} \sum_{i=1}^{H \times W} [y_{pi}^{T_D} log(y_i) + y_{pi}^{T_W} log(y_i)], \end{aligned} \tag{2}$$

where $y_{pi}^{T_D}$ and $y_{pi}^{T_W}$ are the pseudo labels generated by T_D and T_W respectively. $H \times W$ denotes the number pixels of prediction masks. The $\mathcal{L}_{Sup}^l$ and $\mathcal{L}_P^l$ are combined to form the labeled distillation loss $\mathcal{L}_{label}$. Besides labeled data distillation, we further design a hierarchical and multi-level loss setting for unlabeled data distillation $\mathcal{L}_{unlabel}$ in Sec. **??**.

3.4   Hierarchical and Multi-level Distillation

Solution for "Are there more granularity concepts that can be resorted to refine the lightweight model for this fine-grained task?"

To capture the multi-granularity features oriented for segmentation task, we design a hierarchical and multi-level distillation scheme for unlabeled data, which helps to provide students with holistic and local concepts in various aspects.

Because there is no ground-truth supervision for unlabeled data, we introduce image-level semantic-sensitive loss, region-level content-aware loss, and pixel-level consistency loss to achieve a more comprehensive distillation procedure on unlabeled data.

Image-level Semantic-sensitive Loss. Image-level semantic-sensitive loss is introduced to distill the high-dimensional semantic abstraction from the deeper teacher model T_D into the lightweight one. After obtaining the predictions $Y \in \mathcal{R}^{N \times H \times W}$ from T_D, we employ global average pooling (GAP) operation to calculate the global semantic vector of each category, which is illustrated as:

$$K_G^{T_D} = G(Y^{N \times H \times W}), \tag{3}$$

where $K_G^{T_D} \in \mathcal{R}^{N \times 1 \times 1}$ summarizes the global semantic knowledge of Nth categories. G denotes the global average pooling operation in each channel. We further calculate the global semantic representation K_G^S of student model and employ image-level semantic-sensitive loss $\mathcal{L}_I^u$ to supervise it:

$$\mathcal{L}_I^u \left(K_G^S, K_G^{T_D} \right) = \frac{1}{N} \sum_{i=1}^{N} \|k_{Gi}^S - k_{Gi}^{T_D}\|_1, \tag{4}$$

where k_{Gi}^S and $k_{Gi}^{T_D}$ denote the semantic category of the student S and the teacher T_D respectively. N represents the number of categories and u represents unlabeled data. In this way, the student model attempts to learn higher-dimensional semantic representations, which helps to provide global guidance on distinguishing task-specific categories with respect to the semantic segmentation task.

Region-level Content-aware Loss. Region-level content-aware loss aims to take advantage of channels advantage of wider teacher model T_W to provide plentiful regional context information. It leverages correlation between image patches from T_W to guide the student S to model contextual relationships between regions, which is effective to better understand image regional details and relationships. The decoders output features (F^{T_W} and F^S) from T_W and S are extracted to calculate the content-aware loss. Firstly, $F \in \mathcal{R}^{C \times H \times W}$ are projected into the regional content vectors $\mathbf{V} \in \mathcal{R}^{C \times H_v \times W_v}$ via projection heads. Each vector $\boldsymbol{v} \in \mathcal{R}^{C \times 1 \times 1}$ in $\mathbf{V}$ represents the regional content of the patch features with size $C \times H/H_v \times W/W_v$. The projection heads are employed with the adaptive average pooling. Then, the regional content vectors $\mathbf{V}$ are used to calculate the self-correlation matrix $\mathbf{M} \in \mathcal{R}^{H_v W_v \times H_v W_v}$:

$$\mathbf{M} = sim(\mathbf{V}^T \mathbf{V}), \ with \ m_{ij} = \frac{\boldsymbol{v}_i^T \boldsymbol{v}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{v}_j\|}, \tag{5}$$

where m_{ij} denotes the value of self-correlation matrix at the coordinate of row i and column j produced with the cosine similarity $sim()$. The $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the ith and j^{th} vectors of flattened $\mathbf{V} \in \mathcal{R}^{C \times H_v W_v}$. The self-correlation matrix represents the regional relationship among features F, which can reflect the correlation between image regions. The region-level content-aware loss is set up

to minimize the difference between $\mathbf{M}^{Tw}$ and $\mathbf{M}^S$:

$$\mathcal{L}_R^u\left(\mathbf{M}^S, \mathbf{M}^{Tw}\right) = \frac{1}{H_v W_v \times H_v W_v} \sum_{i=1}^{H_v W_v} \sum_{j=1}^{H_v W_v} (m_{ij}^S - m_{ij}^{Tw})^2, \qquad (6)$$

where m_{ij}^S and m_{ij}^{Tw} denote the self-correlation values of student model and teacher model. $H_v W_v \times H_v W_v$ is the size of self-correlation matrices.

Pixel-level Consistency Loss. The pixel-level consistency loss is also employed on the unlabeled data X_u, which is the same as Eq.(2) and denoted as $\mathcal{L}_P^u$. To sum up, the total loss function in our multi-granularity distillation scheme is summarized as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{Sup}^l + \mathcal{L}_P^l}_{\mathcal{L}_{label}} + \underbrace{\mathcal{L}_P^u + \lambda_1 \mathcal{L}_I^u + \lambda_2 \mathcal{L}_R^u}_{\mathcal{L}_{unlabel}}, \qquad (7)$$

where the λ_1 and λ_2 are the trade-off parameters for the loss functions.

4   Experiment

4.1   Datasets and Evaluation Metrics

Datasets. PASCAL VOC2012 [?] is a widely used benchmark for semantic segmentation, which consists of 20 foreground objects classes and one background class. Following [?, ?], we adopt 10,582 augmented images as the full training set and 1,449 images for validation. Cityscapes [?] is a real urban scene dataset that consists of 2,974 images for training, 500 images for validation, and 1,525 for testing with 19 semantic classes. We follow [?, ?] to use the public 1/16, 1/8, 1/4, and 1/2 subsets of the training set as the labeled data and the remaining images in the training set as the unlabeled data.

Evaluation Metrics. To verify the performance of the segmentation model, we calculate the mean Intersection-over-Union (mIoU) of all classes. We also compare the parameters and FLOPs of the model backbone for fair evaluation.

4.2   Implementation Details

We initialize the weights of the lightweight backbones (ResNet-18 [?] and Mobile Netv2 [?]) with pre-trained weights on ImageNet [?]. We employ DeepLabv3+ [?] segmentation architecture in our framework, which consists of a visual backbone and an ASPP based segmentation heads. The deep-and-thin teacher model ResNet-101 and the wide-and-shallow teacher model Wide ResNet-34 [?] are pre-trained via CPS [?] and do not update their parameters in the procedure of optimizing student models. We leverage SGD algorithm for network optimization with a 0.0005 weight decay and a 0.9 momentum. We set the hyper-parameters of the coefficients of the loss functions as $\lambda_1 = 0.002$ and $\lambda_2 = 100$. There is currently no researches on lightweight semi-supervised semantic segmentation.

Table 1: Comparison with SOTAs on the VOC2012 Val set with 512×512 input size under different partition protocols. The first split reports the results of the complementary teacher models. The second and third splits denote the results of employing ResNet-18 and MobileNetv2 as the lightweight backbone, respectively. Sup. l denotes training with only labeled data under the corresponding partition.

Method	Backbone	Params	FLOPs	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
CPS [?]	ResNet-101 (T_D)	42.63M	59.13G	73.98	76.43	77.61	78.64
CPS [?]	Wide ResNet-34 (T_W)	66.83M	91.12G	68.17	71.38	74.78	75.85
Sup. l	ResNet-18			42.17	55.65	62.52	65.39
MT [?]	ResNet-18			44.61	56.62	61.79	66.48
CPS [?]	ResNet-18	11.19M	17.20G	47.13	60.05	65.89	68.69
KD [?]	ResNet-18	($3.81\times$ w.r.t. T_D)	($3.44\times$ w.r.t. T_D)	56.37	62.01	66.23	68.55
CWD [?]	ResNet-18	($5.97\times$ w.r.t. T_W)	($5.30\times$ w.r.t. T_W)	50.45	58.71	63.96	65.82
Ours (MGD)	ResNet-18			**66.86**	**68.29**	**69.71**	**71.59**
Sup. l	MobileNetv2			46.02	46.66	48.19	49.61
MT [?]	MobileNetv2			45.11	47.20	49.15	49.63
CPS [?]	MobileNetv2	1.81M	1.52G	46.57	48.75	49.94	50.52
KD [?]	MobileNetv2	($23.55\times$ w.r.t. T_D)	($38.90\times$ w.r.t. T_D)	47.99	49.30	50.07	50.68
CWD [?]	MobileNetv2	($36.92\times$ w.r.t. T_W)	($59.95\times$ w.r.t. T_W)	46.13	47.45	48.89	49.70
Ours (MGD)	MobileNetv2			**50.95**	**52.27**	**53.92**	**55.88**

We replace the backbone of competitive methods *i.e.*, CPS [?] and MT [?] with ResNet-18 and MobileNetv2 to make a comparison with our approach. The classical distillation methods KD [?] (KL Divergence) and CWD [?] are employed on our complementary teacher models to distill the lightweight models.

4.3   Comparison with state-of-the-art methods

In this section, we compare MGD with several state-of-the-art methods in terms of computation cost and accuracy under all partition protocols. We employ ResNet-18 and MobileNetv2 as the lightweight backbones and report the results on PASCAL VOC2012 and Cityscapes in Tab. **??**, **??**, **??** and Fig. **??**. We summarize the FLOPs and parameters of the model backbone, and the corresponding segmentation heads in these models are compressed proportionately.

PASCAL VOC2012. The experimental results on PASCAL VOC2012 are shown in Tab. **??**, which can be summarized as the following observations: 1) Replacing the large model with a small model directly will result in obvious performance degradation. For example, it decreased from 73.97% (ResNet-101) to 47.13% (ResNet-18) under 1/16 partition protocols, with a serious decline of 26.8%. 2) Employing the existing distillation method obtain minor improvement, while our approach contributes to alleviating the performance drop effectively. As shown in Fig. **??**, MGD shows significant performance improvement under diverse partition protocols, and the growth trend is more obvious when the amount of labeled data is less. Compared with CPS, MGD helps to boost the performance from 47.13 to 66.86 in the 1/16 partition setting. In more detail, the improvement is 19.73%, 8.24%, 3.82%, and 2.90% under 1/16, 1/8, 1/4, and 1/2 partition protocols, respectively. 3) Furthermore, although the model with ResNet-18 backbone is compressed by **3.44×** (**5.30×**) FLOPs and **3.81×** (**5.97×**) parameters, it only exists a small gap from the result of T_W. It fully demonstrates
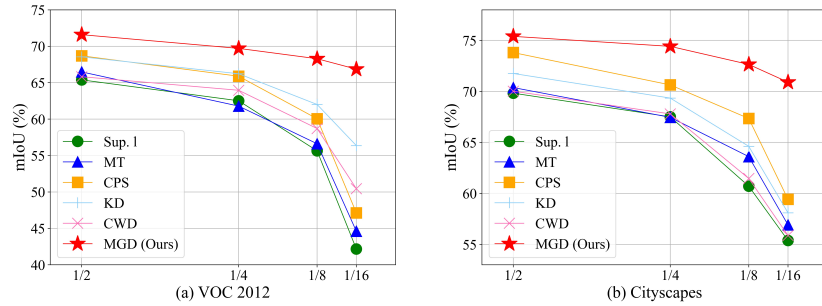
Fig. 3: Performance of different methods under different partition protocols with ResNet-18 backbone

Table 2: Comparison for few-supervision on PASCAL VOC2012. We follow the same partition protocols as PseudoSeg [?].

Method	Backbone	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)
CPS [?]	ResNet-101	64.07	67.42	71.71	75.88
CPS [?]	Wide ResNet-34	60.01	63.32	67.84	72.03
Sup. l	ResNet-18	15.54	22.43	35.78	51.47
MT [?]	ResNet-18	12.58	18.15	33.71	50.02
CPS [?]	ResNet-18	24.53	36.19	47.61	56.85
Ours	ResNet-18	**58.03**	**61.94**	**64.97**	**67.40**

the effectiveness of our method. 4) Although the lightweight model with Mo-bileNetv2 compresses more calculational costs (**38.90× (59.95×)** FLOPs and **23.55× (36.92×)** parameters), MGD manages to achieve an impressive performance. For example, MGD with MobileNetv2 (50.95) even surpasses the CPS method with ResNet-18 backbone (47.13) under 1/16 partition protocol.

Few Supervision on PASCAL VOC2012. To evaluate the robustness of our method on the few labeled case, we conduct the experiments on PAS-CAL VOC 2012 with few supervision by following the same partition protocols adopted in PseudoSeg [?]. As shown in Tab. **??**, it can observe that our method still can achieve impressive results with fewer labeled samples, which far exceed MT [?] and CPS [?]. Surprisingly, the performance of MGD (58.03) trained with only 1/16 partition data can surpass CPS (56.85) under 1/2 partition proto-cols. The few supervision experiments further reveal that MGD is suitable for lightweight model optimization with few or even very few samples.

Cityscapes. We conduct the experiments under different partition proto-cols on Cityscapes and summarize the result in Tab. **??**. From the Tab. **??** and Fig. **??**, we observe a similar performance gain and trend under different parti-tion protocols as PASCAL VOC2012. Our method can establish state-of-the-art results among all partition protocols with both ResNet-18 and MobileNetv2 backbones. For example, MGD obtains 70.91/58.00 under the 1/16 partition

Table 3: Comparison with state-of-the-arts on the Cityscapes val set with the input size of 800×800 under different partition protocols.

Method	Backbone	Params	FLOPs	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
CPS [?]	ResNet-101(T_D)	42.63M	144.32G	74.72	77.62	78.85	79.58
CPS [?]	Wide ResNet-34(T_W)	66.83M	222.45G	72.56	75.41	77.68	78.37
Sup. l	ResNet-18			55.39	60.71	67.51	69.83
MT [?]	ResNet-18	11.19M	42.01G	56.92	63.61	67.46	70.38
CPS [?]	ResNet-18	($\mathbf{3.81}\times$ w.r.t. T_D)	($\mathbf{3.43}\times$ w.r.t. T_D)	59.43	67.35	70.64	73.81
KD [?]	ResNet-18	($\mathbf{5.97}\times$ w.r.t. T_W)	($\mathbf{5.30}\times$ w.r.t. T_W)	58.10	64.62	69.35	71.76
CWD [?]	ResNet-18			55.88	61.47	67.80	70.02
Ours (MGD)	ResNet-18			**70.91**	**72.65**	**74.42**	**75.40**
Sup. l	MobileNetv2			50.06	51.97	54.03	56.86
MT [?]	MobileNetv2	1.81M	3.73G	51.10	53.85	56.21	57.25
CPS [?]	MobileNetv2	($\mathbf{23.55}\times$ w.r.t. T_D)	($\mathbf{38.69}\times$ w.r.t. T_D)	53.34	55.61	57.98	59.20
KD [?]	ResNet-18	($\mathbf{36.92}\times$ w.r.t. T_W)	($\mathbf{59.64}\times$ w.r.t. T_W)	53.45	54.89	55.28	57.90
CWD [?]	ResNet-18			51.52	52.64	54.37	57.01
Ours (MGD)	MobileNetv2			**58.00**	**59.27**	**61.36**	**62.55**

Table 4: Analysis of different teacher setting on VOC2012 and Cityscapes under 1/8 partition with student ResNet-18. T_D: ResNet-101, T_W: Wide ResNet-34.

First Teacher		Second Teacher		VOC	Cityscapes
T_D	T_W	T_D	T_W		
$\checkmark$				64.85	69.63
	$\checkmark$			64.36	68.98
$\checkmark$		$\checkmark$		66.14	71.19
	$\checkmark$		$\checkmark$	65.78	70.92
$\checkmark$			$\checkmark$	**68.29**	**72.65**

protocol with ResNet-18/MobileNetv2 backbone, which outperforms CPS by 11.48%/4.66%.

Qualitative Results. As illustrated in Fig. **??**, we compare our method with the competitive approach CPS [**?**] and Sup. l in terms of segmentation results on PASCAL VOC2012 Val [**?**]. From the figure, we find that Sup. l and CPS fail to capture the internal pixels of the target or mistake the object categories. On the contrary, our approach manages to obtain the integral masks closer to the ground truth labels, which further proves the superiority of the MGD scheme and its multi-level setting.

4.4   Ablation Study

Analysis of Teacher Setting. To evaluate the significance of the complementary teacher scheme, we employ different teacher model combinations to optimize the lightweight model. As shown in Tab. **??**, employing two teacher models in the distillation procedure is better than a single model. Furthermore, adopting different structural-based teachers can achieve more impressive performance than leveraging the same combination, which indicates that teacher models with different architectures help to provide complementary concepts and success in optimizing lightweight models from different dimensions.

Fig. 4: Qualitative results on the PASCAL VOC2012 Val. (a) Input images. (b) Ground truth labels. (c) The segmentation results from Sup. l. (d) The segmentation results from CPS [**?**]. (e) The segmentation results from MGD

Table 5: The performance with different loss setting on the VOC2012 and Cityscapes under 1/8 partition protocol with student ResNet-18.

(a) Results of different levels distillation.

Pixel-level		Image-level		Region-level		VOC	City
$\mathcal{L}_{Sup}$	$\mathcal{L}_P$	$\mathcal{L}_I^{T_D}$	$\mathcal{L}_I^{T_W}$	$\mathcal{L}_R^{T_D}$	$\mathcal{L}_R^{T_W}$		
√						55.65	60.71
√	√					63.61	69.05
√	√	√				66.01	71.03
√	√				√	66.02	71.15
√	√	√			√	**68.29**	**72.65**
√	√	√	√	√	√	67.51	71.96

(b) Performance of labeled and unlabeled data distillation losses.

$\mathcal{L}_{label}$		$\mathcal{L}_{unlabel}$			VOC	City
$\mathcal{L}_{Sup}$	$\mathcal{L}_P^l$	$\mathcal{L}_P^u$	$\mathcal{L}_I^{T_D}$	$\mathcal{L}_R^{T_W}$		
√	√				58.24	63.98
√	√	√	√	√	**68.29**	**72.65**

Analysis of Loss Setting. To investigate the distillation loss setting, we conduct extensive experiments and analyze the components in the multi-level distillation loss and the cooperative distillation loss in Tab. **??**. As shown in Tab. **??**, the variant with the $\mathcal{L}_P^{u,l}$ can obtain a notable improvement than the supervised baseline. $\mathcal{L}_I^u$ and $\mathcal{L}_R^u$ further boost the performance of the lightweight model to 66.01/71.03% and 66.02/71.15%, which proves the effectiveness of region-level content-aware loss and image-level semantic-sensitive loss. We employ $\mathcal{L}_I^u$ and $\mathcal{L}_R^u$ on different branches to further explore the optimal configuration for these two losses. The experimental results show that when applying $\mathcal{L}_I^u$ to the deeper teacher model and $\mathcal{L}_R^u$ to a wider model can achieve the best results, *i.e.*, 68.29%/72.65%. It may be because the deeper teacher is beneficial to provide the corresponding global semantic characteristics while the wider model has the advantage in capturing local context information. The results in Tab. **??**

Table 6: Different size of H_v and W_v on VOC2012 under 1/8 partition protocol with student ResNet-18.

$H_v \times W_v$	3×3	5×5	7×7	9×9	11×11
mIoU	66.83	67.01	**68.29**	67.75	67.38

also reveal that the distillation losses from image-level, region-level and pixel-level are all essential for breaking through the bottleneck of the capacity of the lightweight model. As summarized in Tab. **??**, MGD can achieve 10.05% and 8.67% performance improvement via unlabeled data distillation, which demonstrates the effectiveness of the unlabeled distillation loss. And it further reveals the labeled-unlabeled data cooperative distillation is efficient for optimizing the lightweight SSSS model.

Analysis of Size of Regional Content Vectors. The regional content vectors $\mathbf{V} \in \mathcal{R}^{C \times H_v \times W_v}$ are generated by the adaptive average pooling operation. To determine the optimal size of $H_v \times W_v$, we follow the dataset characteristic to set $H_v = W_v$ and conduct the experiments on PASCAL VOC2012. As shown in Tab. **??**, the MGD can achieve the best result when $H_v \times W_v = 7 \times 7$.

5   Discussion

In this paper, we provide an effective solution on how to obtain a lightweight segmentation model in a semi-supervised setting. Furthermore, there are still many aspects worthy of further study: i) How to formulate the whole optimization process into a one-stage procedure? ii) How to further narrow the performance gap between the lightweight model and the large model? iii) Is there any data selection strategy such as the active learning algorithm that can improve the performance? We hope that our work can provide inspired insight and encourage more researchers to involve in the field of lightweight semi-supervised models.

6   Conclusion

In this paper, we offer the first attempt to obtain lightweight SSSS models via a novel multi-granularity distillation (MGD) scheme. MGD employs triplet-branches to distill diverse granularity concepts from two complementary teacher models into a student one. Within MGD, the labeled-unlabeled data cooperative distillation scheme is set up to take full advantage of semi-supervised data characteristics. Furthermore, MGD designs a hierarchical and multi-level loss setting for unlabeled data, which consists of an image-level semantic-sensitive loss, a region-level content-aware loss, and a pixel-level consistency loss. This novel hierarchical loss setting plays a key role to break through the bottleneck of the capacity of the lightweight model. We conduct extensive experiments on two benchmarks to demonstrate the effectiveness of the proposed approach and analyze the key factors that contribute more to addressing this task.

References

1. Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. arXiv preprint arXiv:2104.13415 (2021)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. TPAMI (2017)
3. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 (2019)
4. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
5. 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. TPAMI (2017)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
7. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR (2021)
8. 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: CVPR (2016)
9. Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R.: Good semi-supervised learning that requires a bad gan. arXiv preprint arXiv:1705.09783 (2017)
10. Dong-DongChen, W., WeiGao, Z.H.: Tri-net for semi-supervised deep learning. In: IJCAI (2018)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
12. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR (2019)
13. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
14. Grandvalet, Y., Bengio, Y., et al.: Semi-supervised learning by entropy minimization. CAP (2005)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
17. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: ICCV (2019)
18. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934 (2018)
19. Ke, Z., Qiu, D., Li, K., Yan, Q., Lau, R.W.: Guided collaborative training for pixel-wise semi-supervised learning. In: ECCV (2020)
20. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS (2012)

22. Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., Jia, J.: Semi-supervised semantic segmentation with directional context-aware consistency. In: CVPR (2021)
23. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
24. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICMLW (2013)
25. Li, C., Xu, K., Zhu, J., Zhang, B.: Triple generative adversarial nets. arXiv preprint arXiv:1703.02291 (2017)
26. Liu, B., Wu, Z., Hu, H., Lin, S.: Deep metric transfer for label propagation with limited annotated data. In: ICCVW (2019)
27. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2604–2613 (2019)
28. Liu, Y., Shu, C., Wang, J., Shen, C.: Structured knowledge distillation for dense prediction. TPAMI (2020)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
30. Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high-and low-level consistency. TPAMI (2019)
31. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. TPAMI (2018)
32. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV (2015)
33. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: CVPR (2020)
34. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer (2015)
36. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018)
37. Shu, C., Liu, Y., Gao, J., Yan, Z., Shen, C.: Channel-wise knowledge distillation for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5311–5320 (2021)
38. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
39. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: ICCV (2017)
40. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 (2015)
41. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
42. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
43. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)

44. Wang, X., Zhu, Z., Qin, F., Ye, Y., Huang, G., Chi, X., He, Y., Wang, X.: Mvster: Epipolar transformer for efficient multi-view stereo. arXiv preprint arXiv:2204.07346 (2022)
45. Wang, Y., Zhou, W., Jiang, T., Bai, X., Xu, Y.: Intra-class feature variation distillation for semantic segmentation. In: European Conference on Computer Vision. pp. 346–362. Springer (2020)
46. Wu, J., Hu, H., Wu, Y.: Image captioning via semantic guidance attention and consensus selection strategy. TOMM (2018)
47. Wu, J., Hu, H., Yang, L.: Pseudo-3d attention transfer network with content-aware strategy for image captioning. TOMM (2019)
48. Wu, J., Xie, S., Shi, X., Chen, Y.: Global-local feature attention network with reranking strategy for image caption generation. In: PRCV (2017)
49. Xie, J., Shuai, B., Hu, J.F., Lin, J., Zheng, W.S.: Improving fast segmentation with teacher-student learning. arXiv preprint arXiv:1810.08476 (2018)
50. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)
51. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR (2020)
52. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: CVPR (2018)
53. You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: SIGKDD (2017)
54. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
55. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
56. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
57. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: ICCV (2019)
58. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
59. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: ECCV (2018)
60. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
61. Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: ECCV (2018)
62. Zhong, Y., Yuan, B., Wu, H., Yuan, Z., Peng, J., Wang, Y.X.: Pixel contrastive-consistent semi-supervised semantic segmentation. In: ICCV (2021)
63. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: ICCV (2019)
64. Zou, Y., Zhang, Z., Zhang, H., Li, C.L., Bian, X., Huang, J.B., Pfister, T.: Pseudoseg: Designing pseudo labels for semantic segmentation. arXiv preprint arXiv:2010.09713 (2020)