Efficient Semantic Segmentation via Self-attention and Self-distillation

Shumin An, Qingmin Liao*, Zongqing Lu, and Jing-Hao Xue, *Senior Member, IEEE*

***Abstract*—Lightweight models are pivotal in efficient semantic segmentation, but they often suffer from insufficient context information due to limited convolution and small receptive field. To address this problem, we propose a tailored approach to efficient semantic segmentation by leveraging two complementary distillation schemes for supplementing context information to small networks: 1) a self-attention distillation scheme, which transfers long-range context knowledge adaptively from large teacher networks to small student networks; and 2) a layer-wise context distillation scheme, which transfers structured context from deep layers to shallow layers within student networks for promoting semantic consistency of the shallow layers. Extensive experiments on the ADE20K, Cityscapes, and Camvid datasets well demonstrate the effectiveness of our proposal.**

***Index Terms*—Semantic segmentation, self-attention distillation, layer-wise context distillation.**

I. Introduction

SEMANTIC segmentation is a significant component of visual scene understanding. As a dense predicting task to assign category labels for every pixel, semantic segmentation has been widely applied in autonomous driving [1], [2].

Fully convolutional network (FCN) [3] is a pioneering work in semantic segmentation and has achieved remarkable performance. Based on FCN, several improvements have been proposed. In addition to applying stronger backbone networks [4], the most common idea is to enhance the sensitivity of the model to global information via aggregating image context, e.g. using dilated convolutions [5], multi-scale features [6], conditional random field [7], and attention mechanism [8], [9]. These methods could achieve good segmentation results, but they often have large amounts of parameters and slow inference, hindering their applications in real life.

In contrast, lightweight networks [10]–[13] have drawn increasing attention for their efficiency. The characteristics of small size, low delay and easy embedding make lightweight networks attractive in resource-constrained applications, such as mobile devices. These models, however, often have lower accuracy than heavy networks.

To balance model accuracy and efficiency, knowledge distillation (KD) [14] was proposed as an effective approach, by training a small student network with the supervision of a large teacher network. Useful information, such as intra-class similarity and inter-class difference, is distilled from the teacher to the student, which could improve the student network's performance without increasing its inference load. Following KD, several distillation methods [15]–[18] have been developed and significantly improved the performance of image-level classification. However, these methods produced limited improvement in semantic segmentation, since they only focus on separate pixel-wise information transferring but ignore the dependency relationship among pixels, which is of great importance for pixel-level segmentation tasks.

Therefore, in this paper, to address the issue of insufficient context knowledge with lightweight networks and the issue of lack of long-range context transferring with current KD methods, we propose a tailored approach to efficient semantic segmentation by leveraging two complementary distillation schemes for supplementing context information to small student networks: a *self-attention distillation* scheme, which transfers long-range context knowledge adaptively from large teacher networks to small student networks; and a *layer-wise context distillation* scheme, which transfers structured context from deep layers to shallow layers within student networks for promoting semantic consistency of the shallow layers.

The self-attention distillation [19]–[23] selectively aggregates the context information at each feature position, for both teacher and student networks, via a weighted sum of features among all positions, with the weight determined by the feature similarity between two positions. Therefore, similar features could benefit each other regardless of their spatial distances, achieving spatial labeling contiguity and semantic consistency. The layer-wise context distillation scheme is motivated by self-distillation [24]–[26], and the goal is to enrich semantic information in shallow layers via obtaining long-range structured context information from deep layers within the same network, as deep layers usually contain richer spatial context information than shallow layers due to more convolution and larger abstracting power.

Through the leverage of complementary self-attention and self-distillation in knowledge distillation, our proposed method can effectively transfer the long-range context information for efficient semantic segmentation by a small student network. As shown in Fig. 1, our method is effective in improving efficient segmentation models without increasing any computation. For traffic scene dataset, whose scene is complicated and contains rich semantic content, there are some difficult segmentation areas, e.g. small size objects (thin telegraph pole, people in the distance) and obscured objects. Our proposed method could enhance the feature representation of these areas, promote

This work was partly supported by the National Natural Science Foundation of China (No.61771276), and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (No.JCYJ20170817161056260). (Corresponding author: Qingmin Liao)

S. An, Q. Liao, Z. Lu are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and also with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK

semantic consistence via transferring context knowledge from the teacher network, and obtain more significant improvement.

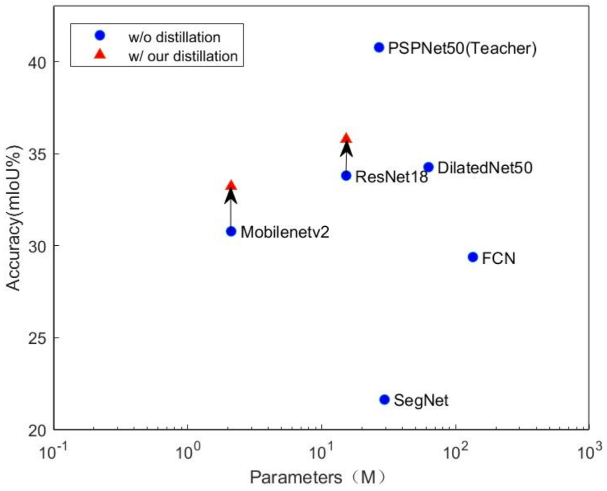

Fig. 1: The number of parameters and accuracy of different models on the ADE20K val dataset. Blue dot denotes the models without distillation. Red triangle denotes the models distilled by our proposed distillation method. The proposed method can increase segmentation accuracy significantly, with no extra parameters.

In summary, our main contributions are as follows:

- We propose a tailored approach to efficient semantic segmentation by leveraging two complementary distillation schemes for supplementing context information to small student networks: a *self-attention distillation* scheme and a *layer-wise context distillation* scheme.
- We propose a self-attention distillation scheme, which can help lightweight student network capture long-range context knowledge adaptively from heavy teacher network.
- We propose a layer-wise context distillation scheme, which can enrich structured context information of shallow layers, promote their semantic consistency, and in return benefit deeper layers.

In the rest of this paper, Section II provides a review of related work in semantic segmentation and knowledge distillation, as well as the main difference between these work and our proposed method. Section III details the proposed knowledge distillation strategy, which includes self-attention distillation and layer-wise context distillation. Section IV presents experiments and analysis to validate the effectiveness of proposed method, and Section V concludes the paper.

II. Related Work

A. Semantic Segmentation

Recent works in semantic segmentation are mainly based on FCN [3] and gain improvements through context aggregation. For example, CRF [7], [27] is introduced as a post-processing module to refine segmentation boundary; dilated convolutions and multi-scale features [6], [28] are used to enlarge receptive field; Recurrent Neural Network (RNN) [29]–[31] is adopted to model relationship among pixels; and attention mechanism [32]–[35] is used to make models explore the correlation knowledge of channels or spatial pixels and benefit segmentation. These methods, however, are often with large amounts of parameters and computation.

Lightweight models have been proposed for high efficiency. Some works [11], [36], [37] simplify models via decomposing the standard convolution into a more efficient form, e.g. ESPNet [11] combines point-wise convolutions and spatial pyramid of dilated convolutions. There is also a way [12] to reduce the complexity of the entire network via adopting channel split and shuffle operation in each residual block. In addition, lightweight models often have fewer convolutional layers, e.g. ResNet18, compared with complex models, e.g. ResNet50, but the receptive field and the ability to gain long-range context information are limited.

B. Knowledge Distillation

Knowledge distillation (KD) [14] is an effective method to improve the performance of a small student network without increasing its inference load. Some methods [15]–[18] transfer feature representation knowledge from hidden layers. FitNet [15] and Mimic [16] align feature maps between teacher and student networks, used as initialization weight or loss function for training the student network. AFD [17] trains multiple networks simultaneously and makes them learn the feature map distribution from each other via adversarial learning. AT [18] aligns the feature attention map of two networks. However, these methods have limited performance in semantic segmentation, since they focus on separate pixel-wise value transferring but ignore the relationship among pixels.

Several KD methods [38]–[41] are specially tailored for efficient semantic segmentation, mainly by adding context information transferring. For example, [38] considers the difference of center pixel and eight adjacent pixels as local context information, and keep it aligned between the large teacher network and the small student network. Knowledge adaptation [39], structured KD [40] and CSC [41] extract and transfer long-range context information in a fixed pair-wise similarity from the teacher network to the student network. MINILM [42] proposes self-attention distillation, but the self-attention mechanism is a component of basic segmentation network and the relation knowledge is still transferred in a fixed form, similar to [39]–[41], and could not get transferred adaptively. In addition, MINILM [42] is specially designed for Transformer based models and could not be applied in general to CNN models which is commonly used in semantic segmentation.

Different from [39]–[41], which align pair-wise similarity directly in a fixed pattern, our work adopts a self-attention mechanism to transfer long-range context information in an adaptive way, in which context information of all pixels are aggregated to current pixel and the connection weights are not fixed but determined by the feature similarities, hence similar semantic features would benefit each other without considering spatial distance. Furthermore, the context transferring obstacles due to different model structures would be

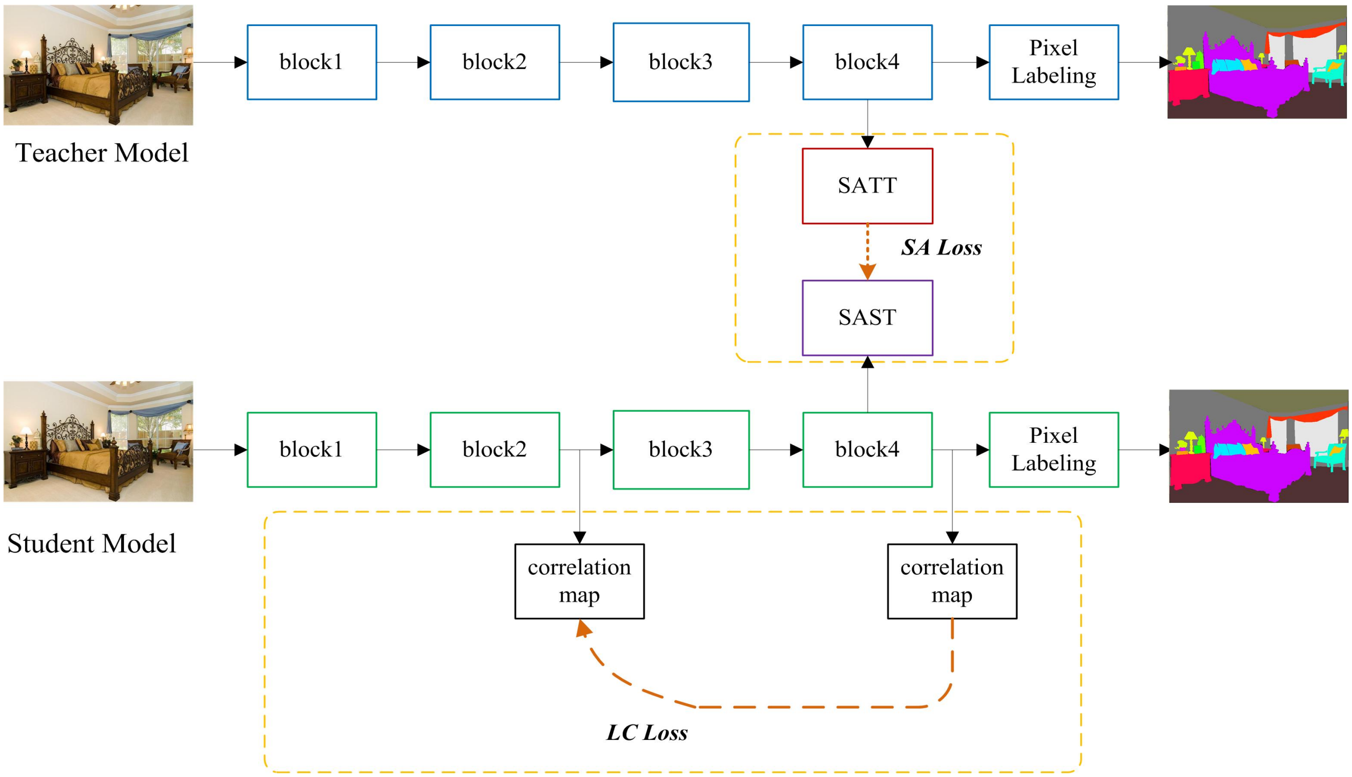

Fig. 2: The architecture of proposed framework. Between the (upper) large teacher network and the (lower) small student network, the self-attention (SA) distillation module consists of two components: self-attention teacher transform (SATT) and self-attention student transform (SAST). The layer-wise context (LC) distillation module transfers a correlation map from deep layers to shallow layers.

mitigated thanks to the adaptive context aggregating weights. There are several works tailored for decreasing this obstacles: [43] specially tailors different training data for the teacher and student networks; and [44] additionally builds the teacher assistant model to help training. Our method does not need to design training data or assistant network, it could bridge the model size gap via adaptive learning.

There are also some recent KD methods [24]–[26] transferring knowledge from the layer-wise point of view. However, they distill separate pixel-wise knowledge without considering context information: [24], [25] directly align feature maps among shallow layers and the deepest layer with deep supervision; and [26] aligns attention maps among shallow layers and deep layers. Long-range structured context information is crucial to semantic segmentation, especially for lightweight models with limited receptive field.

Different from [24]–[26], our proposed layer-wise context distillation focuses on transferring semantic context knowledge across layers, which would enrich semantic information of shallow layers, e.g. by enhancing intra-class similarity and inter-class difference, and improve the segmentation performance.

III. Proposed Method

In this section, we propose a new KD strategy tailored for efficient semantic segmentation. It contains two complementary distillation schemes: self-attention distillation and layer-wise context distillation. The whole framework is shown in Fig. 2. There are two separate networks: one is the large teacher network, which has high segmentation accuracy, and the other is the small student network, which is simple and has high efficiency.

A. Self-Attention Distillation

1) Motivation: Semantic segmentation is not a separate pixel classification task. In semantic segmentation, it is crucial to consider the dependency among pixels for generating discriminant feature representations and avoiding misclassification of pixels. However, small networks have less ability to gain this contextual dependencies compared with large networks, leading to weaker segmentation. Motivated by the self-attention mechanism [20], which could exploit long-range spatial semantic inter-dependency to improve the model's awareness of global contextual information, this paper proposes a self-attention module to help supply the student network with context knowledge. Different from general works [33] that straightly embed self-attention in the interior of model, which would largely increase the amount of parameters, we use a self-attention module as a bridge to help small network train better via context information supplement. As illustrated in Fig. 2, the self-attention (SA) module consists of two components: self-attention student transform (SAST) and self-attention teacher transform (SATT).

2) Self-Attention Student Transform (SAST): SAST is designed for the student network to adaptively aggregate context information from the student feature map. SAST is based on self-attention mechanism and illustrated in Fig. 3.

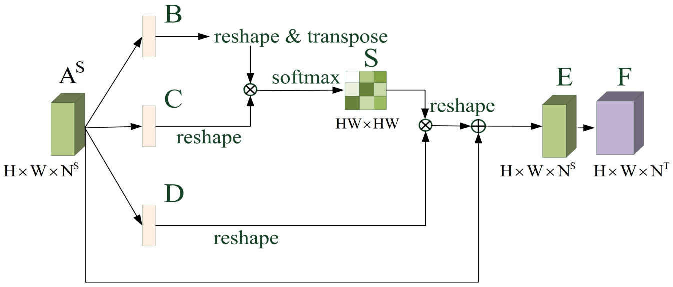

Fig. 3: The architecture of SAST.

Assume the student feature map is A^S of size $H \times W \times N^S$ representing the height, weight and channels, respectively. It gets through three convolution layers respectively to obtain three feature maps B, C and D of size $H \times W \times N^S$, and then is reshaped as $M \times N^S$, where M equals $H \times W$. The transposition of B is multiplied with C, and an attention map S of size $M \times M$ is obtained after softmax:

$$S_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^{M} \exp(B_i \cdot C_j)}, \quad (1)$$

where i and j denote the indices of pixels in the feature map, $i, j \in \{1, \ldots, M\}$. The dot product is adopted to determine the correlation of two positions. S is a correlation matrix, where S_{ji} measures the impact of position i on position j in the feature map. The more similar of two positions, the greater correlation. Matrix D is multiplied with S and reshaped as the original size $H \times W \times N^S$, and the output is added to A with scale factor α:

$$E_j = \alpha \sum_{i=1}^{M} (S_{ji} D_i) + A_j^S. \quad (2)$$

The feature map E has a global view of context information since each position value is a weighted sum of all positions in the original feature map. Then, E gets through a convolution layer to obtain feature map F, with the view of adjusting channels to the same as the teacher feature map A^T. The loss function could be written as

$$L_{SA} = \frac{1}{M} \sum_{j=1}^{M} \left\| \frac{F_j}{\|F_j\|_2} - \frac{A_j^T}{\|A_j^T\|_2} \right\|_2, \quad (3)$$

The L_2 loss is adopted to formulate the SA distillation loss. F_j, after acquiring knowledge from the teacher network A_j^T, not only transmits knowledge to the jth pixel A_j^S, but also to other relevant locations A_i^S. If A_i^S is more relevant to A_j^S (S_{ji} is larger), it would gain greater knowledge from F_j.

The mapping relationship is shown in Fig. 5(b). For traditional KD method designed for image-level classification, e.g. Mimic, the mapping relationship is illustrated in the first row of Fig. 5(b), which only aligns individual pixel value without considering the relationship among pixels. The SAST method is illustrated in the second row of Fig. 5(b), which adaptively aggregates global context information to current pixel with similar features benefiting each other regardless of their spatial distances, thus achieving intra-class compactness and semantic consistency.

3) Self-Attention Teacher Transform (SATT): SATT is designed to transfer the latent rich context information from the teacher network into an explicit form that can be easily learned by the student network. The structure of SATT is illustrated in Fig. 4.

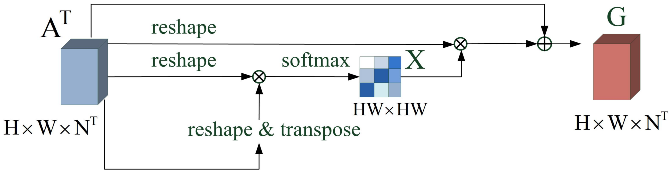

Fig. 4: The architecture of SATT.

The teacher feature map A^T is reshaped as $M \times N^T$, where M equals $H \times W$, and then the transposition of reshaped A^T is multiplied by itself to obtain a correlation map X of size $M \times M$ after softmax:

$$X_{ji} = \frac{\exp(A_i^T \cdot A_j^T)}{\sum_{i=1}^{M} \exp(A_i^T \cdot A_j^T)}, \quad (4)$$

where X_{ji} represents the impact of position i on position j in the teacher feature map, $i, j \in \{1, \ldots, M\}$.

SATT aims to use correlation weight X to aggregate all associated pixels to current pixel. That is, the reshaped A^T is multiplied with X and reshaped as the original size $H \times W \times N^T$, and then the result is added to original teacher feature map A^T to obtain feature map G:

$$G_j = \sum_{i=1}^{M} (X_{ji} A_i^T) + A_j^T. \quad (5)$$

Feature map G also has a global view of context information since each position value is a weighted sum of all positions in the original teacher feature map. The loss function could be written as

$$L_{SA} = \frac{1}{M} \sum_{j=1}^{M} \left\| \frac{F_j}{\|F_j\|_2} - \frac{G_j}{\|G_j\|_2} \right\|_2. \quad (6)$$

Compared with SAST, F_j obtains knowledge from not only the jth position in A^T but also its relevant positions which is determined by correlation map X. The mapping relationship is shown in the third row of Fig. 5(b), which makes the student feature map gain more context information from the teacher network and further improves semantic consistency.

4) Visualization of Results with and without SA: To verify the effectiveness of SA module, we present some results via adding SAST and SATT step by step. One point is selected in pillow area marked by red dot in Fig. 5(a). The correlation map of this point and all other spatial positions in the feature map is shown in Fig. 5(c). It can be seen that both the inter-class difference, e.g. from wall and bed area, and intra-class similarity, e.g. to the other pillow area, get enhanced by adding SAST and SATT. The student feature map

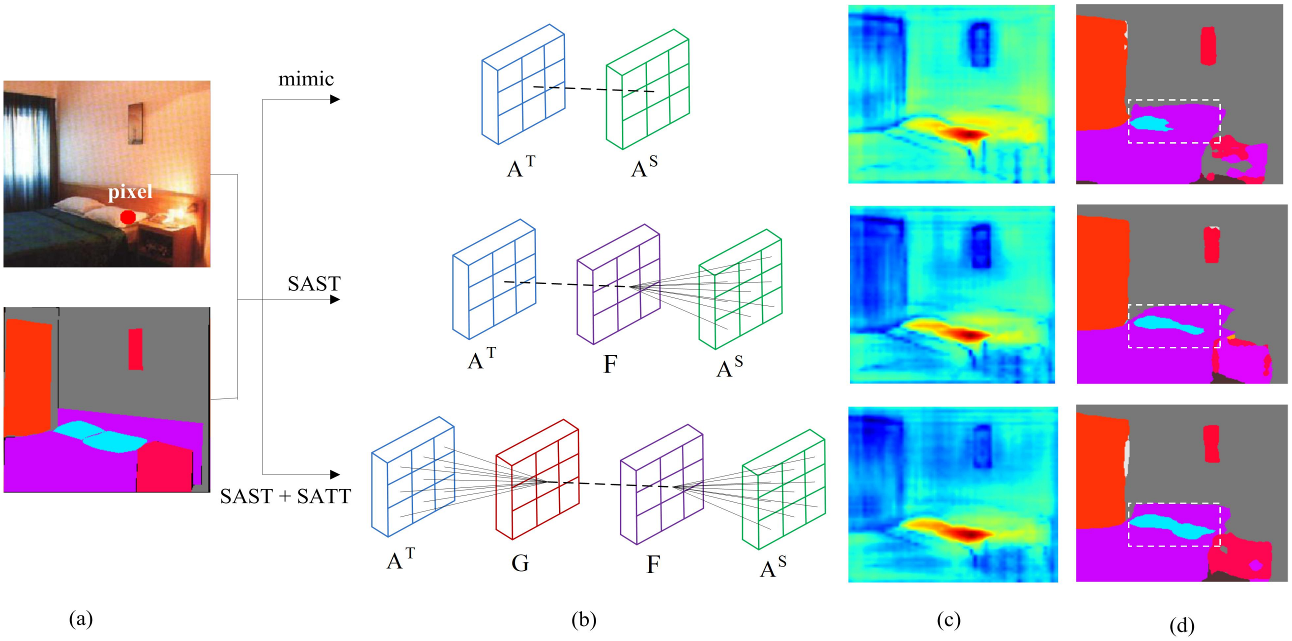

Fig. 5: The effectiveness of SA: (a) input image with random selected point marked by red dots and ground-truth label; (b) the mapping relationship of mimic, SAST, and SAST+SATT; (c) the visualization of correlation map of given pixel point in three methods; and (d) segmentation results of three methods.

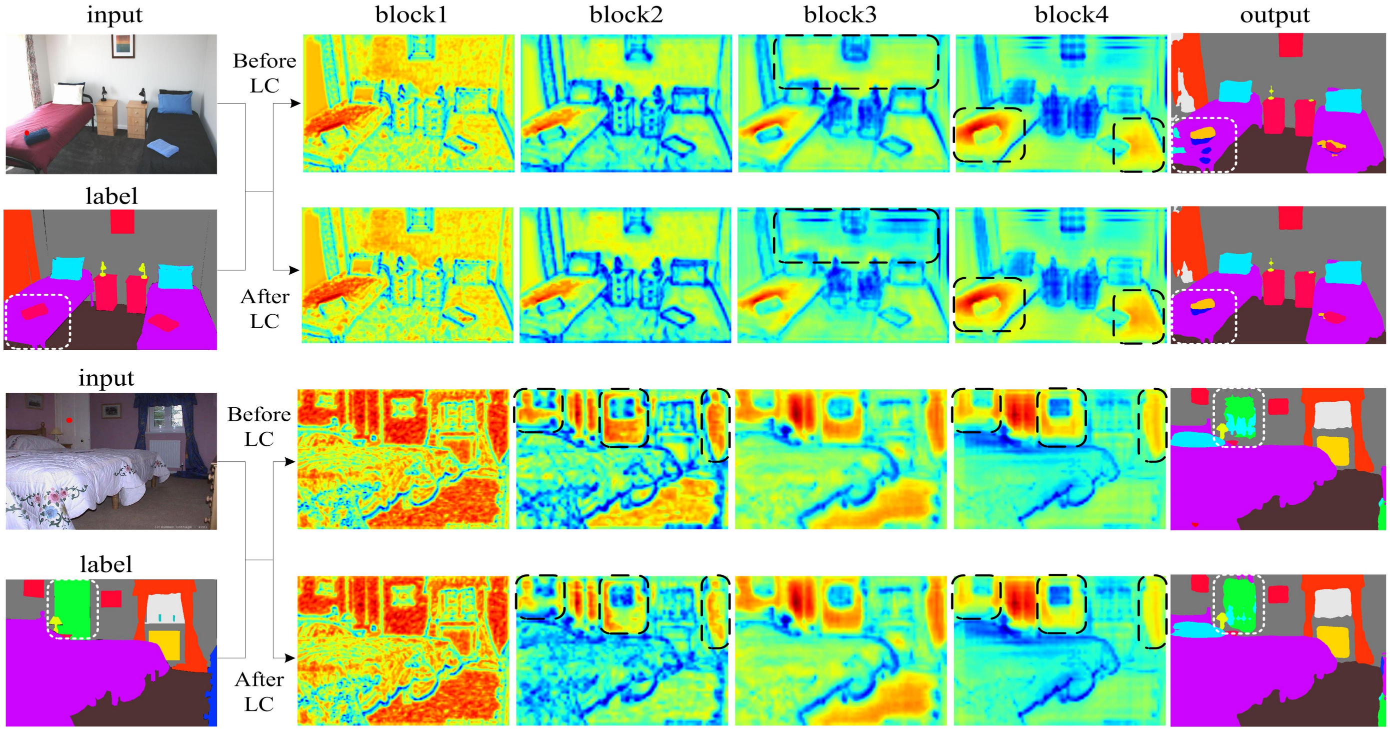

Fig. 6: Two examples of segmentation results and correlation information of each block feature map before and after applying layer-wise context distillation.

fails to capture long-range context knowledge via mimicking the teacher feature map straightly without considering the relationship among pixels. Self-attention module could make up for the deficiency and improve segmentation performance, as shown in Fig. 5(d).

B. Layer-Wise Context Distillation

1) Motivation: Deep layers contain richer context information than shallow layers in a same network due to more convolution and larger abstracting power. For a network with four blocks (like the one in Fig. 2), we visualize the correlation information of one random selected point, marked by red dot in input image, and other spatial positions in the feature map of different blocks, as shown in Fig. 6. The feature maps in preceding blocks more focus on local correlation information and gain less long-range dependency among pixels compared with deeper blocks.

Motivated by self-distillation methods [24]–[26] which transfer useful knowledge from the layer-wise point of view, layer-wise context (LC) distillation is proposed to enrich long-range context information of shallow layers by distilling it from deep layers, with the aim of promoting semantic consistency of the shallow layers, and then in return benefit deep layers and final segmentation results.

2) Layer-Wise Context Distillation: The LC distillation only needs to distill long-range context information, so it directly computes correlation between any two feature positions via bilinear pooling function as correlation map, which is the same as with [39], [40]. The main difference between the LC distillation and [39], [40] is that the LC distillation transfers correlation map from layer-wise point of view within the student network, while [39], [40] get it transferred from the teacher network to the student network. The LC distillation applies bilinear interpolation at each feature map of different layers in a network to ensure the same spatial size. The feature map is normalized before applying bilinear pooling.

Assume the feature map is A with size $H \times W \times N$, $K(A)$ denotes the correlation map of A, and $i, j \in \{1, \dots, H \times W\}$. The correlation of two random selected positions i, j in feature map A could be written as

$$K(A)_{ij} = \frac{1}{H \times W} \cdot \frac{A_i^T A_j}{\|A_i\|_2 \|A_j\|_2}. \tag{7}$$

Let A_l represent the feature map of lth layer in network after interpolation, with $l, \Delta \in \{1, \dots, L-1\}$. The L_2 loss is adapted to formulate the layer-wise context distillation loss as

$$L_{LC}(A_l, A_{l+\Delta}) = L_2(K(A_l), K(A_{l+\Delta})). \tag{8}$$

It is noteworthy that the connection paths between shallow layers and deep layers could be diverse. For instance, knowledge flow in block4 $\longrightarrow$ block3, block4 $\longrightarrow$ block1, block3 $\longrightarrow$ block1, etc. The number of possible paths among two blocks for a network with L blocks is $\frac{L(L-1)}{2}$. Dense connections could also be applied by combining above paths. We shall present an ablation study to find out which type of connection could improve performance significantly.

3) Total Loss: The total loss comprises three terms: the cross entropy loss L_{ce} with ground-truth label, the self-attention distillation loss L_{SA} in Eq.(6), and the layer-wise context distillation loss L_{LC} in Eq.(8):

$$L = L_{ce} + \beta L_{SA} + \gamma L_{LC}, \tag{9}$$

where the effect of three losses is balanced by the loss weights β and γ, which are set to 10 and 20 respectively, making these loss value ranges comparable.

4) Visualization of Correlation Information with and without LC Distillation: The effectiveness of LC distillation is validated by investigating the correlation information of different layer features with and without the LC distillation. For a given randomly selected point marked with red dot in input image, the correlation with all other pixel positions in the feature map is shown in Fig. 6. The intra-class similarity and inter-class difference of shallow features are weaker than deep features, but they could get enhanced after the LC distillation, as shown in black dashed box. Furthermore, as for the deepest feature, its intra-class similarity (shown in block4 of the first image) and inter-class difference (shown in block4 of the second image) are also enhanced due to the better semantic consistency learned in the shallower features. With the help of LC distillation, long-range context information among pixels could be enhanced to make more robust decision, as reflected in segmentation results (see the rightmost column of Fig. 6).

IV. Experiments

In this section, we first introduce the datasets and implementation details, then discuss ablation studies to verify the effectiveness of our proposed method, and finally present results on three publicly available datasets: ADE20K, Cityscapes and Camvid.

A. Datasets

1) ADE20K: The ADE20K dataset [45] is a challenging dataset with 150 objects and diverse scenarios. The pixel ratios of different objects follow a long tail distribution. The dataset contains 20k images for training, 2k images for validation, and 3k images for testing.

2) Cityscapes: The Cityscapes dataset [2] contains urban street scene images with 30 classes of dense pixel annotations, 19 of which are used for evaluation. It includes 5k finely annotated images and 20k coarsely annotated images. We only use high quality finely annotated images, and they are divided into 2975 for training, 500 for validation, and 1525 for testing.

3) Camvid: The Camvid dataset [46] is a driving scenarios dataset with 12 classes, 11 of which are used for evaluation since 12th class includes unlabeled data. The dataset contains 367 images for training, 100 images for validation, and 233 images for testing.

B. Implementation Details

To demonstrate the effectiveness of proposed method, we choose two public lightweight models as the student network: MobileNetV2 and ResNet18. PSPNet50 is used as the teacher network for generating features with rich context information.

1) Training: The mini-batch stochastic gradient descent (SGD) is used for training the student network with momentum (0.9) and weight decay (0.0005). We set initial learning rate as 0.01 and use poly learning rate strategy with power 0.9. The training dataset is augmented by randomly scaling (from 0.5 to 2.0), randomly flipping and cropping images into 713×713 as input data. The batch-size is set to 8 in training stage. The experiments employ 180K iterations for ADE20K, 120K iterations for Cityscapes, and 120K iterations for Camvid.

2) Evaluation Metrics: The segmentation performance is measured by the mean Intersection Over Union(mIOU) and the pixel accuracy. IOU is the ratio of intersection to union between predicted results and ground truth for every object category, while mIOU is the average of IOU in all categories. The pixel accuracy is the ratio of correct predicted pixels to all pixels. In addition, the experiments also evaluate the efficiency of model, including model size and computation complexity, which are measured by the amount of model parameters and floating point operations (FLOPs), respectively.

C. Ablation Studies

1) Distillation Paths of LC: This section summarizes the performance of segmentation when different connection paths are conducted on the LC distillation, as shown in Table I, where P_{ij} denotes knowledge flow from the jth block to the ith block.

TABLE I: Performance of different connection paths of ResNet18 on the ADE20K val dataset. The best is in bold.

Path	mIOU(%)	Path	mIOU(%)	Path	mIOU(%)
P_{12}	35.72	P_{23}	35.72	$P_{23} + P_{24}$	35.74
P_{13}	35.71	$\mathbf{P_{24}}$	**35.82**	$P_{23} + P_{34}$	35.78
P_{14}	35.77	P_{34}	35.80	$P_{24} + P_{34}$	35.78

From Table I, we can make the following observations. First, if distillation paths are conducted among shallow layers, e.g. P_{12}, the accuracy is not improved and even lower than baseline accuracy 35.83%, since shallow layers lack rich scene context. The accuracy is improved when extracting the context knowledge from the deepest block feature, e.g. P_{14}, P_{24} and P_{34}, as the deepest block feature encodes the most global information. Secondly, single path works better than multiple paths, e.g. P_{24}, P_{34} outperform $P_{24} + P_{34}$. We conjecture that multiple paths may over-constrain the network and have an adverse effect on the learning process. Therefore, single path is shown in Fig. 2.

2) The Effectiveness of SA and LC Distillation Modules: The SA and LC distillation modules are proposed to transfer long-range context information from the teacher network to the student network and from deep layers to shallow layers within student network, respectively. To demonstrate the effectiveness of each distillation module, this section enables and disables different components of the proposed method. The experiments are conducted in two student networks: ResNet18 and MobileNetV2 on ADE20K. The initial model pretrained on ImageNet is adopted to help training. The segmentation results are summarized in Table II, all results are tested in a single scale on the ADE20K val set, and FLOPs is calculated with input size 713×713.

TABLE II: Ablation studies on different components of the loss in our method. T: Teacher; S: Student.

Method	mIOU(%)	Acc(%)	Params(M)	FLOPs(B)
T: PSPNet50	40.79	79.65	44.62	140.6
S1: ResNet18	33.82	76.05	11.38	48.18
S1+SAST	35.58	77.06	11.38	48.18
S1+SAST+SATT	35.73	77.10	11.38	48.18
S1+SAST+SATT+LC	**35.82**	**77.13**	11.38	48.18
S2: MobileNetV2	30.79	75.22	1.96	6.53
S2+SAST	33.06	76.22	1.96	6.53
S2+SAST+SATT	33.18	76.30	1.96	6.53
S2+SAST+SATT+LC	**33.24**	**76.36**	1.96	6.53

As can be seen from Table II, each distillation scheme gains higher mIOU and pixel accuracy, which implies the effectiveness on training a better student network. The proposed distillation strategy boosts the performance from 33.82 to 35.82 for ResNet18, from 30.79 to 33.24 for MobileNetV2 in mIOU (%), without extra parameters or computation. Furthermore, the effectiveness of the SA and LC distillation modules is visualized in Fig. 5 and Fig. 6, respectively. It can be seen that, with our proposed method, more context information and semantic consistency can be captured to help the student network make a better decision. In addition, we present some visualized segmentation results in Fig. 7 and Fig. 8 to show the effect of each module. The results via the SAST module are much closer to ground truth compared with the student network without distillation, while the SATT and LC modules could further promote semantic consistency within the objects (e.g. the areas of internal misclassification are reduced in curtain, ground, chair), and make the object segmentation more complete (e.g. pillows and lamps are closer to the real shapes).

3) Comparison with Other Distillation Methods: This section demonstrates the superiority of our proposed method to other common distillation methods: KD [14], Mimic [16], knowledge adaptation [39] and CSC [41].

a) KD: This method defines the probability output of each category for every pixel in the teacher network as soft label, and it is used to supervise the learning procedure of the student network. The soften degree is controlled by a temperature parameter, which is set 2 empirically.

b) Mimic: This method defines the feature map of teacher network as knowledge. The feature maps are aligned between the student network and the teacher network via a 3×3 convolution layer to match the dimension of channels.

c) knowledge adaptation: This method launches two branches to gain knowledge for the student network at the feature map level. It makes the student network mimic both the feature map and affinity information of the teacher network.

d) CSC: This method considers both spatial and channel correlations at the feature map level via pair-wise dot product.

The proposed method is compared with the above three methods on ADE20K. Table III shows that the proposed method promotes the student network most and achieves the highest segmentation accuracy. KD and Mimic only align separate pixels without considering the transfer of context knowledge. Knowledge adaptation, CSC and our proposed

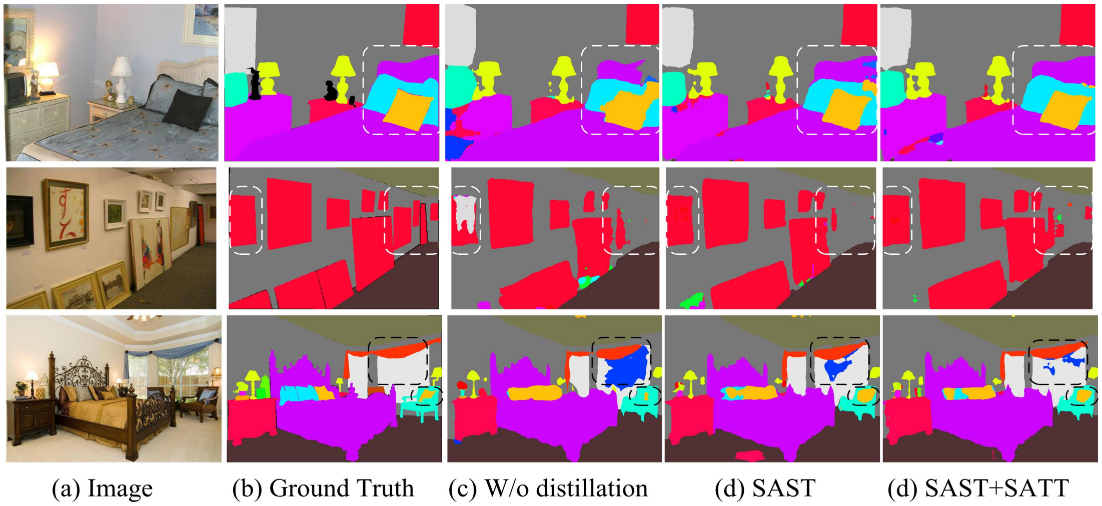

Fig. 7: Visualized results before and after applying each component of SA.

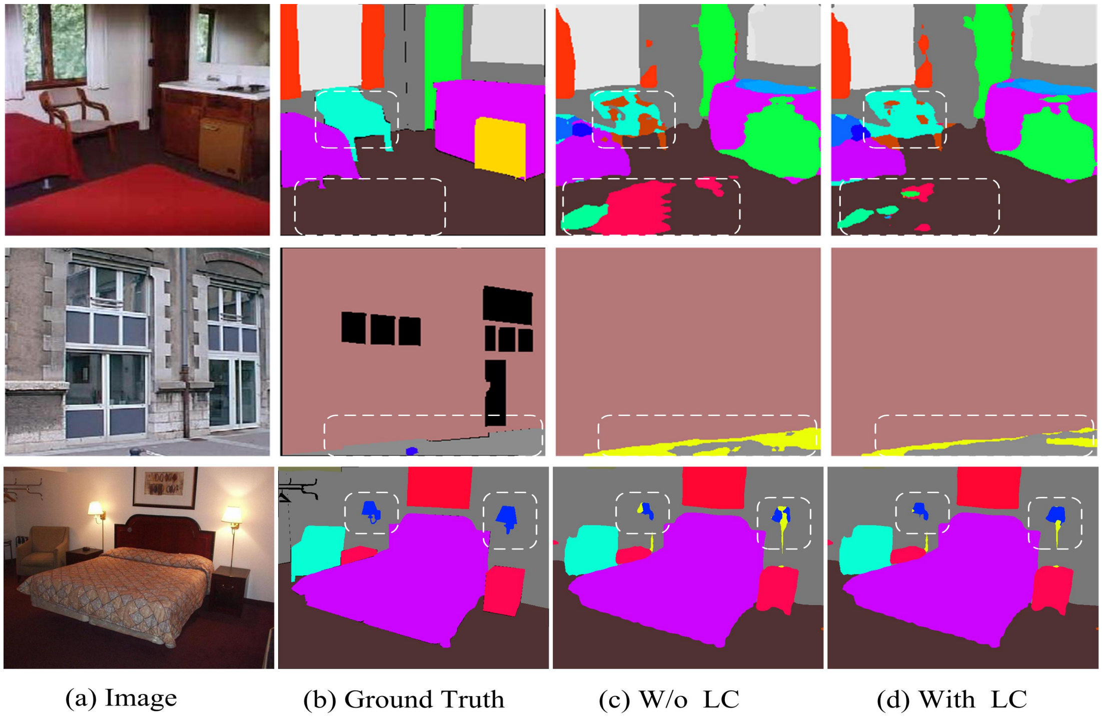

Fig. 8: Visualized results before and after applying layer-wise context distillation.

TABLE III: Comparison of KD [14], Mimic [16], knowledge adaptation [39] and CSC [41] with our proposed method. The performance is evaluated on ADE20K val with one single scale.

Method	mIOU(%)	Acc(%)
T: PSPNet50	40.79	79.65
S: ResNet18	33.82	76.05
S+KD	34.27	76.17
S+Mimic	35.06	76.97
S+knowledge adaptation	35.12	76.98
S+CSC	35.25	77.00
S+SA(ours)	35.73	77.10
S+SA+LC(ours)	**35.82**	**77.13**

method make up for the deficiency and thus outperform KD and mimic methods. Knowledge adaptation defines correlation map via calculating dot product between any two feature positions as context knowledge, and gets it aligned between two networks. CSC considers channel correlation more than knowledge adaptation does, and performs better in results, but it still represents and transfers correlation information in a fixed pattern via aligning pair-wise similarity directly. Our proposed SA module transfers context information in an adaptive way, compared with the fixed context representation pattern in knowledge adaptation and CSC, it makes use of adaptive context aggregating weights and could mitigate the transferring obstacles caused by different model structures. In addition, our proposed LC module could further promote semantic consistency via transferring context information from layer-wise point of view within the student network, while knowledge adaptation merely considers knowledge transferring from teacher network to student network. Therefore, our proposed method achieves higher segmentation accuracy than knowledge adaptation and CSC.

D. Segmentation Results on ADE20K, Cityscapes and Camvid

TABLE IV: Segmentation performance on the validation set of ADE20K.

Method	Acc(%)	mIOU(%)	Params(M)
SegNet [47]	71.00	21.64	47.52
DilatedNet50 [5]	76.35	34.28	62.74
FCN [3]	71.32	29.39	134.5
PSPNet50 [6] (teacher)	79.65	40.79	44.62
MobileNetV2 [10]	75.22	30.79	1.96
MobileNetV2(ours)	**76.36**	**33.24**	1.96
ResNet18 [4]	76.05	33.82	11.38
ResNet18(ours)	**77.13**	**35.82**	11.38

1) ADE20K: We evaluate the proposed distillation method on ADE20K with two student networks: ResNet18 and MobileNetV2. The performances of student networks and other public models on the dataset are listed in Table IV. For MobileNetV2, the proposed method improves 2.45% mIOU, making it outperform SegNet and FCN, with higher segmentation accuracy and less model complexity. For ResNet18, it improves 2% mIOU, and outperforms DilatedNet50 after distillation. Fig. 9 visualizes some segmentation results obtained from without distillation and traditional knowledge distillation method KD. The results from our proposed method are much closer to the ground truth and better in semantic consistency.

TABLE V: Segmentation performance on the validation and test sets of Cityscapes.

Method	val mIOU(%)	test mIOU(%)	Params(M)	FLOPs(GB)
PSPNet50 [6] (teacher)	78.45	77.93	44.62	351.2
FCN [3]	-	65.3	128.23	311.36
DeepLabv2-CRF [7]	-	70.4	62.7	499.7
Dilation10 [5]	-	67.1	25.6	192.7
SegNet [47]	-	57.0	47.52	464.8
ENet [48]	-	58.3	0.358	3.612
DANet50 [33]	78.32	-	45.16	397.4
ShuffleNetV2 [49]	70.85	-	12.6	132.85
ShuffleNetV2 [49] +HANet [50]	71.52	-	13.7	132.9
ResNet18 +HANet [50]	71.34	-	14.9	117.37
ResNet18	70.11	69.16	11.38	117.26
ResNet18(ours)	**73.10**	**71.79**	11.38	117.26

2) Cityscapes: Then, the proposed method is evaluated on Cityscapes. We choose ResNet18 as the student network and the performances are presented in Table V. FLOPs is calculated with input size 512×1024 for all methods. The proposed distillation method improves segmentation accuracy by 2.99% and 2.63% in mIOU for validation and test data, respectively. Distilled student network outperforms DeepLabv2-CRF and ShuffleNetV2 significantly, although it is weaker than these models before distillation. Compared with HANet, which is a lightweight add-on module and could improve the performance by extracting the height-wise context information, our method achieves higher improvement while not increasing the amount of parameters. The segmentation results are visualized in Fig. 10. The proposed distillation method can help the student network classify, and provide details for, small objects, such as telegraph pole and people in the distance, thanks to transferring rich context information.

TABLE VI: Segmentation performance on the test set of Camvid.

Method	mIOU(%)	Params(M)
PSPNet50 [6] (teacher)	73.54	44.62
FCN [3]	57.10	134.5
DeepLab-CRF-LargeFOV [51]	61.60	20.5
SegNet [47]	55.6	47.52
ENet [48]	51.3	0.358
ResNet18	66.12	11.38
ResNet18(ours)	**70.16**	11.38

3) Camvid: The performances on the test set of Camvid are listed in Table VI. The proposed method improves the student network segmentation accuracy by a large amplitude 4% in mIOU. The segmentation results are shown in Fig. 11. The objects, e.g. car and telegraph pole, can be labelled correctly after distillation.

V. Conclusion

This paper proposes a novel approach to efficient semantic segmentation, by leveraging two complementary knowledge distillation strategies. Through a self-attention distillation scheme, student models could gain long-range context information adaptively from teacher models; through a layer-wise context distillation scheme from deep layers to shallow layers, semantic consistency of segmentation results can be further enhanced. Extensive experiments have demonstrated the effectiveness of our proposed method that lightweight student networks can gain a large margin in segmentation accuracy without any extra burden of inference.

References

[1] A. Ess, T. Müller, H. Grabner, and L. Van Gool, "Segmentation-based urban traffic scene understanding." in *British Machine Vision Conference*, vol. 1, 2009, p. 2.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[3] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation." *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

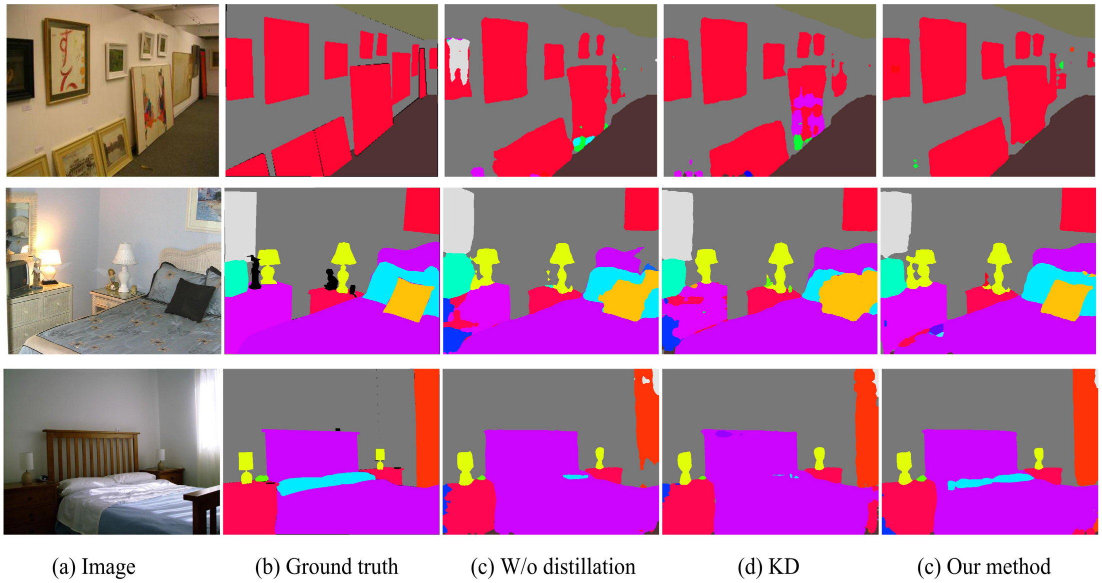

Fig. 9: Visualized results on the ADE20K dataset produced from ResNet18.

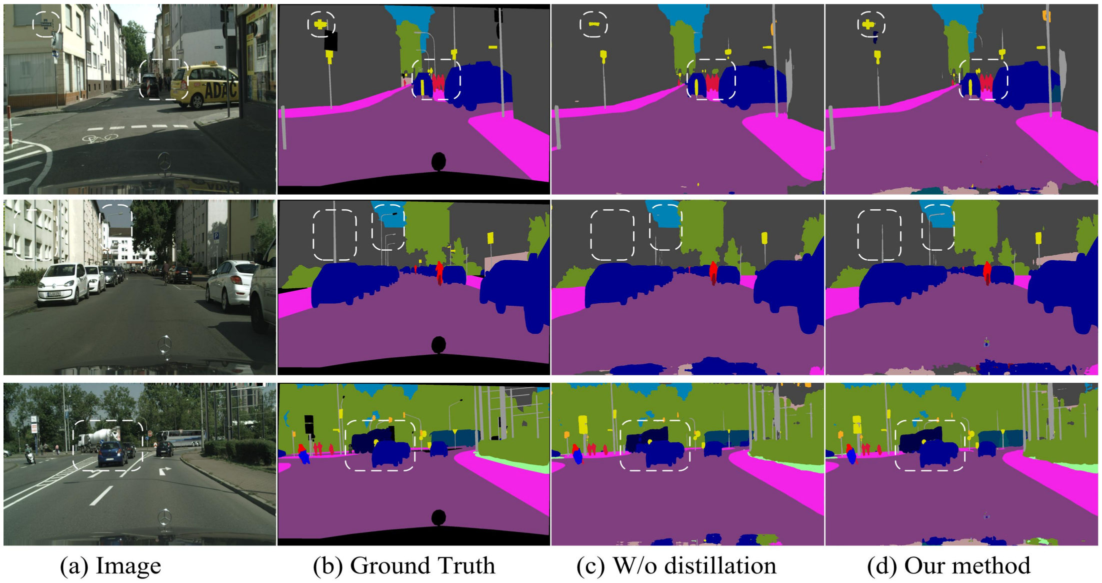

Fig. 10: Visualized results on the Cityscapes dataset produced from ResNet18.

[9] X. Li, L. Zhang, G. Cheng, K. Yang, Y. Tong, X. Zhu, and T. Xiang, "Global aggregation then local distribution for scene parsing," in *IEEE Transactions on Image Processing*, vol. 30, 2021, pp. 6829–6842.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[11] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 552–568.

[12] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki, "LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1860–1864.

[13] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4373–4382.

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

[15] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, "FitNets: Hints for thin deep nets," *Proc. ICLR*, 2015.

[16] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 6356–6364.

[17] I. Chung, S. Park, J. Kim, and N. Kwak, "Feature-map-level online adversarial knowledge distillation," in *International Conference on Machine Learning*, 2020, pp. 2006–2015.

[18] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *Proc. ICLR*, 2017.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[20] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

[21] M. Rangwala and R. Williams, "Learning multi-agent communication through structured attentive reasoning," in *NeurIPS*, 2020.

[22] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you

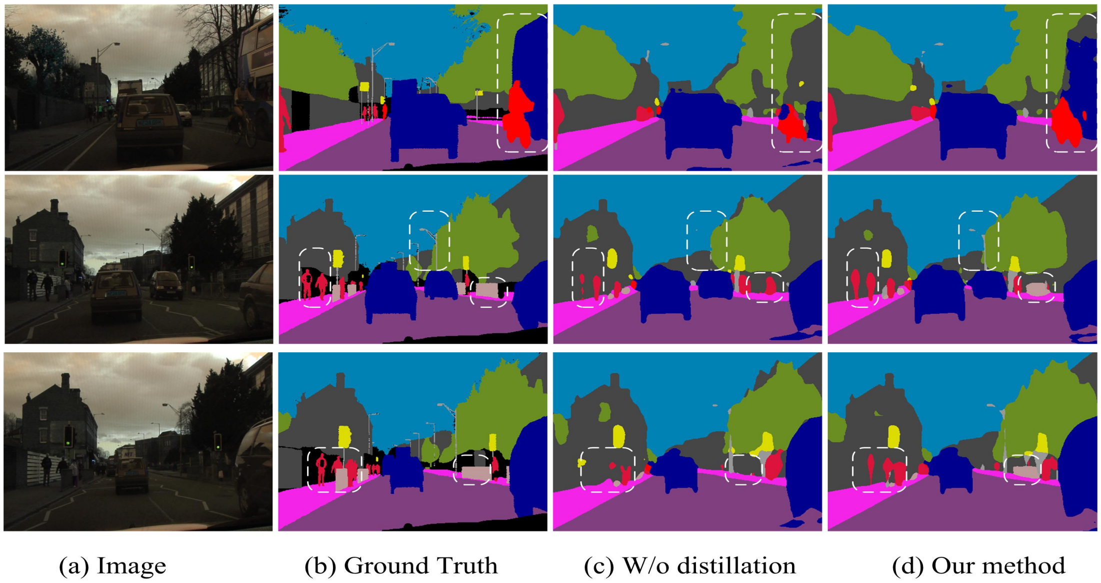
(a) Image (b) Ground Truth (c) W/o distillation (d) Our method

Fig. 11: Visualized results on the Camvid dataset produced from ResNet18.

need for video understanding?" in *International Conference on Machine Learning*, 2021.

[23] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," *Proc. ICLR*, 2020.

[24] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3713–3722.

[25] Y. Luan, H. Zhao, Z. Yang, and Y. Dai, "MSD: Multi-self-distillation learning via multi-classifiers within deep neural networks," *arXiv preprint arXiv:1911.09418*, 2019.

[26] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection CNNs by self attention distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1013–1021.

[27] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[28] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[29] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 41–48.

[30] H. Fan and H. Ling, "Dense recurrent neural networks for scene labeling," *arXiv preprint arXiv:1801.06831*, 2018.

[31] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "DAG-recurrent neural networks for scene labeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3620–3629.

[32] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[33] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[34] Y. Yuan and J. Wang, "OCNet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[35] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603–612.

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[37] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[38] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W.-S. Zheng, "Improving fast segmentation with teacher-student learning," *arXiv preprint arXiv:1810.08476*, 2018.

[39] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 578–587.

[40] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[41] S. Park and Y. S. Heo, "Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy," in *Sensors*, vol. 20, no. 16, 2020, p. 4616.

[42] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *NeurIPS*, 2020.

[43] J. Fu, X. Geng, Z. Duan, B. Zhuang, X. Yuan, A. Trischler, J. Lin, C. Pal, and H. Dong, "Role-wise data augmentation for knowledge distillation," *Proc. ICLR*, 2020.

[44] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[45] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[46] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. Eur. Conf. Comp. Vis.*, 2008, pp. 44–57.

[47] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

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

[49] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient cnn architecture design," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 116–131.

[50] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[51] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *arXiv preprint arXiv:1412.7062*, 2014.

Shumin An received her M.Eng. degree in electronic and communication engineering from University of Science and Technology of China in 2018. She is currently pursuing the Ph.D. degree in the Department of Electronic Engineering, Tsinghua University. Her research interests include image processing, pattern recognition and artificial intelligence.

Qingmin Liao received the Ph.D. degree in signal processing and telecommunications from the University of Rennes 1 in 1994. He is a Professor in the Department of Electronic Engineering and the Shenzhen International Graduate School, Tsinghua University. His research interests include image/video processing, transmission, analysis, biometrics and their applications.

Zongqing Lu received the Ph.D. degree in signal processing from Xidian University in 2007. He is an Assistant Professor in the Department of Electronic Engineering, Tsinghua University. His research interests include image processing and machine learning.

Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a Professor in the Department of Statistical Science, University College London. His research interests include statistical classification, high-dimensional data analysis, pattern recognition and image processing. He was/is an Associate Editor of Digital Signal Processing, Neurocomputing, the IEEE Transactions on Circuits and Systems for Video Technology, and the IEEE Transactions on Neural Networks and Learning Systems.