Structured Knowledge Distillation for Semantic Segmentation

Yifan Liu[1*] Ke Chen[2] Chris Liu[2] Zengchang Qin[3,4] Zhenbo Luo[5] Jingdong Wang[2†]
[1]The University of Adelaide [2]Microsoft Research Asia [3]Beihang University
[4]Keep Labs, Keep Inc. [5]Samsung Research China

Abstract

In this paper, we investigate the knowledge distillation strategy for training small semantic segmentation networks by making use of large networks. We start from the straightforward scheme, pixel-wise distillation, which applies the distillation scheme adopted for image classification and performs knowledge distillation for each pixel separately. *We further propose to distill the* structured *knowledge from large networks to small networks, which is motivated by that semantic segmentation is a structured prediction problem. We study two structured distillation schemes: (i)* pair-wise *distillation that distills the pairwise similarities, and (ii)* holistic *distillation that uses GAN to distill holistic knowledge. The effectiveness of our knowledge distillation approaches is demonstrated by extensive experiments on three scene parsing datasets: Cityscapes, Camvid and ADE20K.*

1. Introduction

Semantic segmentation is the problem of predicting the category label of each pixel in an input image. It is a fundamental task in computer vision and has many real-world applications, such as autonomous driving, video surveillance, virtual reality, and so on. Deep neural networks have been the dominant solutions for semantic segmentation since the invention of fully-convolutional neural networks (FCNs) [38]. The subsequent approaches, e.g., DeepLab [5, 6, 7, 48], PSPNet [56], OCNet [50], RefineNet [23] and DenseASPP [46] have achieved significant improvement in segmentation accuracy, often with cumbersome models and expensive computation.

Recently, neural networks with small model size, light computation cost and high segmentation accuracy, have attracted much attention because of the need of applications on mobile devices. Most current efforts have been devoted to designing lightweight networks specially for segmentation or borrowing the design from classification networks, e.g., ENet [31], ESPNet [31], ERFNet [34] and ICNet [55]. The interest of this paper lies in compact segmentation networks, with a focus on training compact networks with the help of cumbersome networks for improving the segmentation accuracy.

We study the knowledge distillation strategy, which has been verified valid in classification tasks [15, 35], for training compact semantic segmentation networks. As a straightforward scheme, we simply view the segmentation problem as many separate pixel classification problems, and then directly apply the *knowledge distillation* scheme to pixel-level. This simple scheme, we call *pixel-wise distillation*, transfers the class probability of the corresponding pixel produced from the cumbersome network (teacher) to the compact network (student).

Considering that semantic segmentation is a structured prediction problem, we present structured knowledge distillation and transfer the structure information with two schemes, *pair-wise distillation* and *holistic distillation*. The *pair-wise distillation* scheme is motivated by the widely-studied pair-wise Markov random field framework [22] for enforcing spatial labeling contiguity, and the goal is to align the pair-wise similarities among pixels computed from the compact network and the cumbersome network.

The *holistic distillation* scheme aims to align higher-order consistencies, which are not characterized in the pixel-wise and pair-wise distillation, between segmentation maps produced from the compact segmentation network and the cumbersome segmentation network. We adopt the adversarial training scheme, encouraging the holistic embeddings of the segmentation maps produced from the compact segmentation network not to be distinguished from the output of the cumbersome segmentation network.

To this end, we optimize an objective function that combines a conventional multi-class cross-entropy loss with the distillation terms. The main contributions of this paper can be summarized as follows.

- We study the knowledge distillation strategy for training accurate compact semantic segmentation networks.

*Part of this work was done when Y. Liu was an intern at Microsoft Research, Beijing, China.

†Corresponding author.

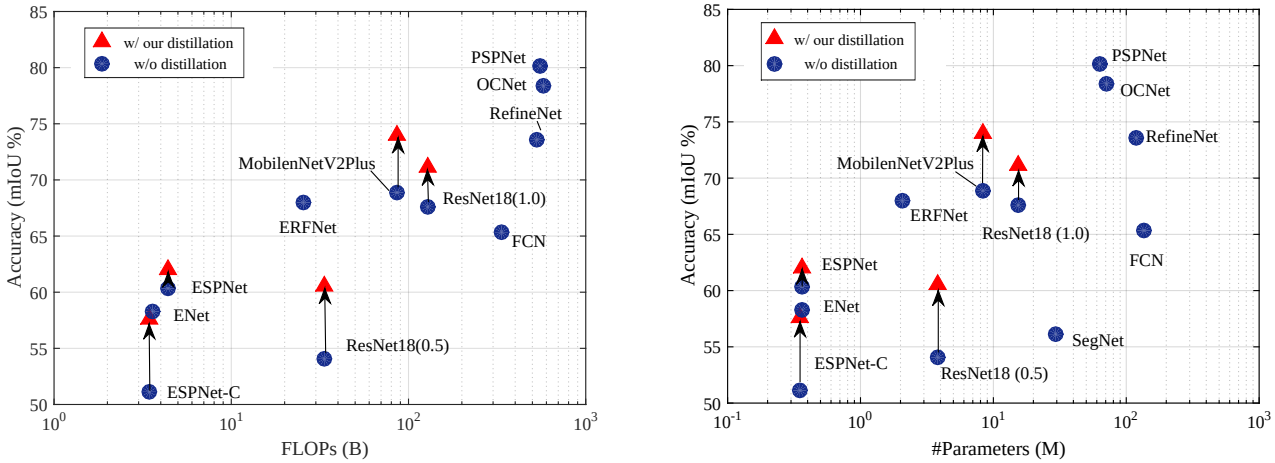


Figure 1: The complexity, parameters and the mIoU for different networks on the Cityscapes test set. The FLOPs is calculated with the resolution of 512×1024. The red triangles are the results of our distillation method while others are without distillation. Blue circles are collected from FCN* [38], RefineNet [23], SegNet [3], ENet [31], PSPNet [56], ERFNet [34], ESPNet [28], MobileNetV2Plus [25], and OCNet [50]. We can see that with our proposed distillation method, we can achieve a higher mIoU, but no extra FLOPs and #Parameters.

- We present two structured knowledge distillation schemes, pair-wise distillation and holistic distillation, enforcing pair-wise and high-order consistency between the outputs of the compact and cumbersome segmentation networks.

- We demonstrate the effectiveness of our approach by improving recently-developed state-of-the-art compact segmentation networks, ESPNet, MobileNetV2-Plus and ResNet18 on three benchmark datasets: Cityscapes [10], CamVid [4] and ADE20K [58], which is illustrated in Figure 1.

2. Related Work

Semantic segmentation. Deep convolutional neural networks have been the dominant solution to semantic segmentation since the pioneering works, fully-convolutional network [38], DeConvNet [30], U-Net [36]. Various schemes [47] have been developed for improving the network capability and accordingly the segmentation performance. For example, stronger backbone networks, e.g., GoogleNets [39], ResNets [14], and DenseNets [17], have shown better segmentation performance. Improving the resolution through dilated convolutions [5, 6, 7, 48] or multi-path refine networks [23] leads to significant performance gain. Exploiting multi-scale context, e.g., dilated convolutions [48], pyramid pooling modules in PSPNet [56], atrous spatial pyramid pooling in DeepLab [6], object context [50], also benefits the segmentation. Lin et al. [24] combine deep models with structured output learning for semantic segmentation.

In addition to cumbersome networks for highly accurate segmentation, highly efficient segmentation networks have been attracting increasingly more interests due to the need of real applications, e.g., mobile applications. Most works focus on lightweight network design by accelerating the convolution operations with factorization techniques. ENet [31], inspired by [40], integrates several acceleration factors, including multi-branch modules, early feature map resolution down-sampling, small decoder size, filter tensor factorization, and so on. SQ [41] adopts the SqueezeNet [18] fire modules and parallel dilated convolution layers for efficient segmentation. ESPNet [28] proposes an efficient spatial pyramid, which is based on filter factorization techniques: point-wise convolutions and spatial pyramid of dilated convolutions, to replace the standard convolution. The efficient classification networks, e.g., MobileNet [16], ShuffleNet [54], and IGCNet [53], are also applied to accelerate segmentation. In addition, ICNet (image cascade network) [55] exploits the efficiency of processing low-resolution images and high inference quality of high-resolution ones, achieving a trade-off between efficiency and accuracy.

Knowledge distillation. Knowledge distillation [15] is a way of transferring knowledge from the cumbersome model to a compact model to improve the performance of compact networks. It has been applied to image classification by using the class probabilities produced from the cumbersome model as soft targets for training the compact model [2, 15, 42] or transferring the intermediate feature maps [35, 51].

There are also other applications, including object detection [21], pedestrian re-identification [9] and so on. The very recent and independently-developed application for semantic segmentation [45] is related to our approach. It mainly distills the class probabilities for each pixel separately (like our pixel-wise distillation) and center-surrounding differences of labels for each local patch

(termed as a local relation in [45]). In contrast, we focus on distilling structured knowledge: pairwise distillation, which transfers the relation among all pairs of pixels other than the relation in a local patch [45], and holistic distillation, which transfers the holistic knowledge that captures high-order information.

Adversarial learning. Generative adversarial networks (GANs) have been widely studied in text generation [43, 49] and image synthesis [12, 20]. The conditional version [29] is successfully applied to image-to-image translation, including style transfer [19], image inpainting [32], image coloring [26] and text-to-image [33].

The idea of adversarial learning is also adopted in pose estimation [8], encouraging the human pose estimation result not to be distinguished from the ground-truth; and semantic segmentation [27], encouraging the estimated segmentation map not to be distinguished from the ground-truth map. One challenge in [27] is the mismatch between the generator's continuous output and the discrete true labels, making the discriminator in GAN be of very limited success. Different from [27], in our approach, the employed GAN does not have this problem as the ground truth for the discriminator is the teacher network's logits, which are real valued. We use adversarial learning to encourage the alignment between the segmentation maps produced from the cumbersome network and the compact network.

3. Approach

Image semantic segmentation is a task of predicting a category label to each pixel in the image from C categories. A segmentation network takes an RGB image $\mathbf{I}$ of size $W \times H \times 3$ as the input, then it computes a feature map $\mathbf{F}$ of size $W' \times H' \times N$, where N is the number of channels. Finally, a classifier is applied to compute the segmentation map $\mathbf{Q}$ of size $W' \times H' \times C$ from $\mathbf{F}$, which is upsampled to the spatial size $W \times H$ of the input image as the segmentation results.

3.1. Structured Knowledge Distillation

We apply the knowledge distillation [15] strategy to transfer the knowledge of the cumbersome segmentation network T to a compact segmentation network S for better training the compact segmentation network. In addition to a straightforward scheme, pixel-wise distillation, we present the two structured knowledge distillation schemes, pair-wise distillation and holistic distillation, to transfer structured knowledge from the cumbersome network to the compact network. The pipeline is illustrated in Figure 2.

Pixel-wise distillation. We view the segmentation problem as a collection of separate pixel labeling problems, and directly use knowledge distillation to align the class probability of each pixel produced from the compact network.

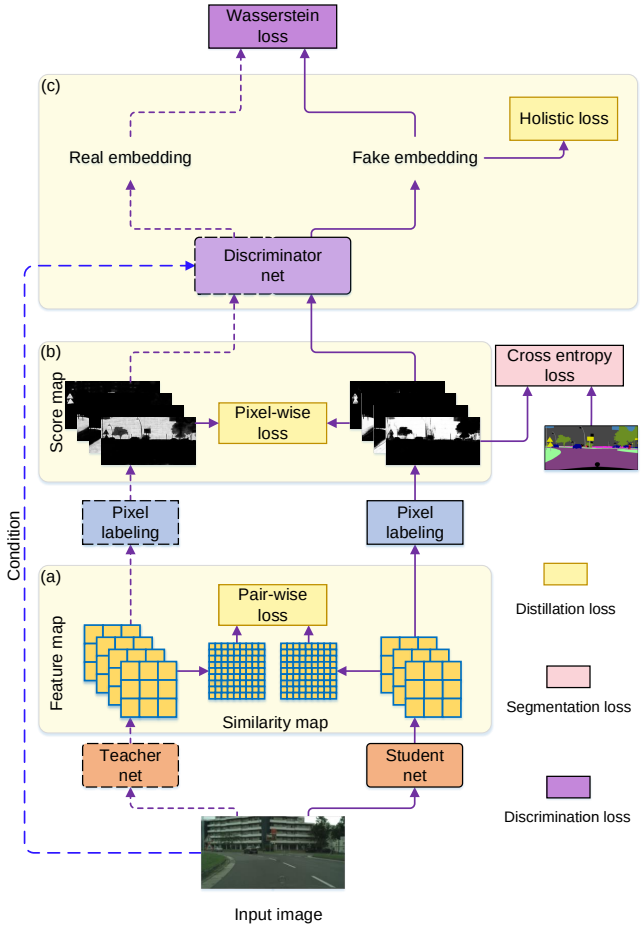


Figure 2: Our distillation framework. (a) Pair-wise distillation. (b) Pixel-wise distillation. (c) Holistic distillation. In the training process, we fix the cumbersome network as our teacher net, and only the student net and the discriminator net will be optimized. The student net with a compact architecture will be trained with three distillation terms and a cross-entropy term.

We adopt an obvious way [15]: use the class probabilities produced from the cumbersome model as soft targets for training the compact network.

The loss function is given as follows,

$$\ell_{pi}(\mathsf{S}) = \frac{1}{W' \times H'} \sum_{i \in \mathcal{R}} \mathrm{KL}(\mathbf{q}_i^s \| \mathbf{q}_i^t), \tag{1}$$

where $\mathbf{q}_i^s$ represent the class probabilities of the ith pixel produced from the compact network S, $\mathbf{q}_i^t$ represent the class probabilities of the ith pixel produced from the cumbersome network T, $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence between two probabilities, and $\mathcal{R} = \{1, 2, \ldots, W' \times H'\}$ denotes all the pixels.

Pair-wise distillation. Inspired by the pair-wise Markov random field framework that is widely adopted for improving spatial labeling contiguity, we propose to transfer the

pair-wise relations, specially pair-wise similarities in our approach, among pixels.

Let a^t_{ij} denote the similarity between the ith pixel and the jth pixel produced from the cumbersome network T and a^s_{ij} denote the similarity between the ith pixel and the jth pixel produced from the compact network S. We adopt the squared difference to formulate the pair-wise similarity distillation loss,

$$\ell_{pa}(\mathsf{S}) = \frac{1}{(W' \times H')^2} \sum_{i \in \mathcal{R}} \sum_{j \in \mathcal{R}} (a^s_{ij} - a^t_{ij})^2. \quad (2)$$

In our implementation, the similarity between two pixels is simply computed from the features $\mathbf{f}_i$ and $\mathbf{f}_j$ as

$$a_{ij} = \mathbf{f}_i^\top \mathbf{f}_j / (\|\mathbf{f}_i\|_2 \|\mathbf{f}_j\|_2),$$

which empirically works well.

Holistic distillation. We align the high-order relations between the segmentation maps produced from the cumbersome and compact networks. The holistic embeddings of the segmentation maps are computed as the representations.

We adopt conditional generative adversarial learning [29] for formulating the holistic distillation problem. The compact net is regarded as a generator conditioned on the input RGB image $\mathbf{I}$, and the predicted segmentation map $\mathbf{Q}^s$ is regarded as a fake sample. We expect that $\mathbf{Q}^s$ is as similar to $\mathbf{Q}^t$, which is the segmentation map predicted by the teacher and is regarded as the real sample, as possible. Wasserstein distance [13] is employed to evaluate the difference between the real distribution and fake distribution, which is written as the following,

$$\begin{aligned} \ell_{ho}(\mathsf{S}, \mathsf{D}) = &\mathbb{E}_{\mathbf{Q}^s \sim p_s(\mathbf{Q}^s)}[\mathsf{D}(\mathbf{Q}^s|\mathbf{I})] \\ &- \mathbb{E}_{\mathbf{Q}^t \sim p_t(\mathbf{Q}^t)}[\mathsf{D}(\mathbf{Q}^t|\mathbf{I})], \end{aligned} \quad (3)$$

where $\mathbb{E}[\cdot]$ is the expectation operator, and $\mathsf{D}(\cdot)$ is an embedding network, acting as the discriminator in GAN, which projects $\mathbf{Q}$ and $\mathbf{I}$ together into a holistic embedding score. The Lipschitz requirement is satisfied by the gradient penalty.

The segmentation map and the conditional RGB image are concatenated as the input of the embedding network D. D is a fully convolutional neural network with five convolutions. Two self-attention modules are inserted between the final three layers to capture the structure information [52, 57]. Such a discriminator is able to produce a holistic embedding representing how well the input image and the segmentation map match.

3.2. Optimization

The whole objective function consists of a conventional multi-class cross-entropy loss $\ell_{mc}(\mathsf{S})$ with pixel-wise and structured distillation terms [1]

$$\begin{aligned} \ell(\mathsf{S}, \mathsf{D}) = &\ell_{mc}(\mathsf{S}) + \lambda_1(\ell_{pi}(\mathsf{S}) + \ell_{pa}(\mathsf{S})) \\ &- \lambda_2 \ell_{ho}(\mathsf{S}, \mathsf{D}), \end{aligned} \quad (4)$$

where λ_1 and λ_2 are set as 10 and 0.1, making these loss value ranges comparable. We minimize the objective function with respect to the parameters of the compact segmentation network S, while maximize it with respect to the parameters of the discriminator D, which is implemented by iterating the following two steps:

- **Train the discriminator** D. Training the discriminator is equivalent to minimizing $\ell_{ho}(\mathsf{S}, \mathsf{D})$. D aims to give a high embedding score for the real samples from the teacher net and a low embedding score for the fake samples from the student net.

- **Train the compact segmentation network** S. Given the discriminator network, the goal is to minimize the multi-class cross-entropy loss and the distillation loss relevant to the compact segmentation network:

  $$\ell_{mc}(\mathsf{S}) + \lambda_1(\ell_{pi}(\mathsf{S}) + \ell_{pa}(\mathsf{S})) - \lambda_2 \ell^s_{ho}(\mathsf{S}),$$

  where

  $$\ell^s_{ho}(\mathsf{S}) = \mathbb{E}_{\mathbf{Q}^s \sim p_s(\mathbf{Q}^s)}[\mathsf{D}(\mathbf{Q}^s|\mathbf{I})]$$

  is a part of $\ell_{ho}(\mathsf{S}, \mathsf{D})$ given in Equation 3, and we expect S to achieve a higher score under the evaluation of D.

4. Implementation Details

Network structures. We adopt state-of-the-art segmentation architecture PSPNet [56] with a ResNet101 [14] as the cumbersome network (teacher) T.

We study recent public compact networks, and employ several different architectures to verify the effectiveness of the distillation framework. We first consider ResNet18 as a basic student network and conduct ablation studies on it. Then, we employ an open source MobileNetV2Plus [25], which is based on a pretrained MobileNetV2 [37] model on the ImageNet dataset. We also test the structure of ESPNet-C [28] and ESPNet [28] that are very compact and have low complexity.

Training setup. Most segmentation networks in this paper are trained by mini-batch stochastic gradient descent (SGD) with the momentum (0.9) and the weight decay (0.0005) for 40000 iterations. The learning rate is initialized as 0.01 and is multiplied by $(1 - \frac{iter}{max-iter})^{0.9}$. We random cut

[1] The objective function is the summation of the losses over the mini-batch of training samples. For description clarity, we ignore the summation operation.

the the images into 512×512 as the training input. Normal data augmentation methods are applied during training, such as random scaling (from 0.5 to 2.1) and random flipping. Other than this, we follow the settings in the corresponding publications [28] to reproduce the results of ESPNet and ESPNet-C, and train the compact networks under our distillation framework.

5. Experiments

5.1. Datasets

Cityscapes. The Cityscapes dataset [10] is collected for urban scene understanding and contains 30 classes with only 19 classes used for evaluation. The dataset contains $5,000$ high quality pixel-level finely annotated images and $20,000$ coarsely annotated images. The finely annotated $5,000$ images are divided into $2,975$/ 500/ $1,525$ images for training, validation and testing. We only use the finely annotated dataset in our experiments.

CamVid. The CamVid dataset [4] is an automotive dataset. It contains 367 training and 233 testing images. We evaluate the performance over 11 different classes such as building, tree, sky, car, road, etc. and ignore the 12th class that contains unlabeled data.

ADE20K. The ADE20K dataset [58] is used in ImageNet scene parsing challenge 2016. It contains 150 classes and under diverse scenes. The dataset is divided into $20K$/$2K$/ $3K$ images for training, validation and testing.

5.2. Evaluation Metrics

We use the following metrics to evaluate the segmentation accuracy, as well as the model size and the efficiency.

The *Intersection over Union (IoU)* score is calculated as the ratio of interval and union between the ground truth mask and the predicted segmentation mask for each class. We use the mean IoU of all classes (*mIoU*) to study the distillation effectiveness. We also report the class IoU to study the effect of distillation on different classes. *Pixel accuracy* is the ratio of the pixels with the correct semantic labels to the overall pixels.

The *model size* is represented by the number of network parameters. and the *Complexity* is evaluated by the sum of floating point operations (FLOPs) in one forward on a fixed input size.

5.3. Ablation Study

The effectiveness of distillations. We look into the effect of enabling and disabling different components of our distillation system. The experiments are conduct on ResNet18 with its variant ResNet18 (0.5) representing a width-halved version of ResNet18 on the Cityscapes dataset. In Table 1, the results of different settings for the student net are the average results from three runs.

Table 1: The effect of different components of the loss in the proposed method. PI = pixel-wise distillation, PA = pair-wise distillation, HO = holistic distillation, ImN = initial from the pretrain weight on the ImageNet.

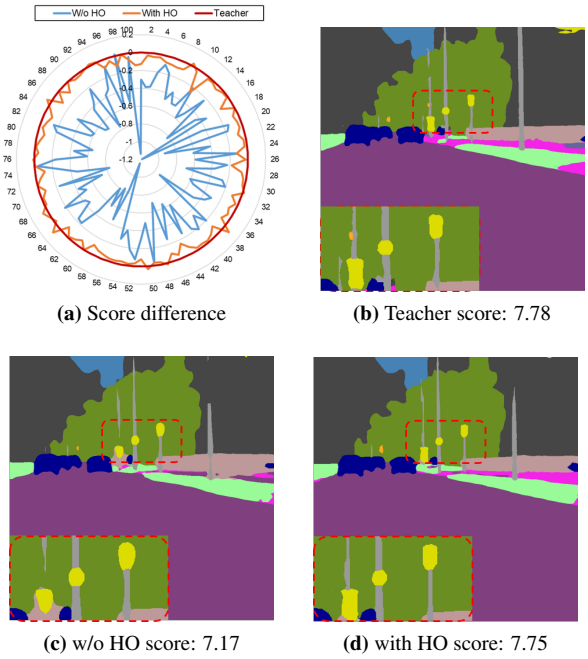
Method	Validation mIoU (%)	Training mIoU (%)
Teacher	78.56	86.09
ResNet18 (0.5)	55.37 ± 0.25	60.67 ± 0.37
+ PI	57.07 ± 0.69	62.33 ± 0.66
+ PI + PA	61.03 ± 0.49	65.73 ± 0.38
+ PI + PA + HO	$\mathbf{61.63 \pm 0.99}$	$\mathbf{66.13 \pm 0.70}$
ResNet18 (1.0)	57.50 ± 0.49	62.98 ± 0.45
+ PI	58.63 ± 0.31	64.32 ± 0.32
+ PI + PA	62.48 ± 0.23	68.77 ± 0.37
+ PI + PA + HO	$\mathbf{63.24 \pm 0.74}$	$\mathbf{69.93 \pm 0.86}$
+ ImN	69.10 ± 0.21	74.12 ± 0.19
+ PI + ImN	70.51 ± 0.37	75.10 ± 0.37
+ PI + PA + ImN	71.37 ± 0.12	76.42 ± 0.20
+ PI + PA + HO + ImN	$\mathbf{72.67 \pm 0.57}$	$\mathbf{78.03 \pm 0.51}$

From Table 1, we can see that distillation can improve the performance of the student network, and distilling the structure information helps the student learn better. With the three distillation terms, the improvements for ResNet18 (0.5), ResNet18 (1.0) and ResNet18 (1.0) with weights pretrained from the ImageNet dataset are 6.26%, 5.74% and 2.9%, respectively, which indicates that the effect of distillation is more pronounced for the smaller student network and networks without initialization with the weight pre-trained from the ImageNet. Such an initialization is also a way to transfer the knowledge from other source (ImageNet). The best mIoU of the holistic distillation for ResNet18 (0.5) reaches 62.7% on the validation set.

On the other hand, one can see that each distillation scheme lead to higher mIoU score. This implies that the three distillation schemes make complementary contributions for better training the compact network.

Furthermore, we illustrate that GAN is able to distill the holistic knowledge. For each image, we feed three segmentation maps, output by the teacher net, the student net w/o holistic distillation, and the student net w/ holistic distillation, into the discriminator D, and compare the embedding scores of the student net to the teacher net. Figure 3a shows the difference of embedding scores, with holistic distillation, the segmentation maps produced from student net can achieve a similar score to the teacher, indicating that GAN helps distill the holistic structure knowledge. Figure 3b, 3c and 3d are segmentation maps and their corresponding embedding scores of a randomly-selected image. The well-trained D can assign a higher score to a high quality segmentation maps, and the student net with the holistic distillation can generate segmentation maps with higher scores and better quality. The self-attention modules in the discriminator are useful for capturing the structure information and benefit the holistic distillation. The gain of using two

self-attention modules is around 1%, from 71.6% to 72.67% for ResNet18 (1.0).



(a) Score difference **(b)** Teacher score: 7.78

(c) w/o HO score: 7.17 **(d)** with HO score: 7.75

Figure 3: Illustrations of that GAN is able to distill the holistic structure with ResNet18 (1.0) as an example student net. (a) shows the score difference of 100 samples between the teacher and the student with and without the adversarial holistic distillation. (b), (c) and (d) present the segmentaion maps and the embedding scores of a randomly-selected sample.

Feature and local pair-wise distillation. We compare the variants of the pair-wise distillation:

- Feature distillation by MIMIC [35, 21]: We follow [21] to align the features of each pixel between T and S through a 1×1 convolution layer to match the dimension of the feature
- Feature distillation by attention transfer [51]: We aggregate the response maps into a so-called attention map (single channel), and then transfer the attention map from the teacher to the student.
- Local pair-wise distillation [45]: We distill a local similarity map, which represents the similarities between each pixel and the 8-neighborhood pixels.

We replace our pair-wise distillation by the above three distillation schemes to verify the effectiveness of our global pair-wise distillation. From Table 2, we can see that our pair-wise distillation method outperforms all the other distillation methods. The superiority over feature distillation schemes: MIMIC [21] and attention transfer [51], which transfers the knowledge for each pixel separately, comes from that we transfer the structured knowledge other than aligning the feature for each individual pixel. The superiority to the local pair-wise distillation shows the effectiveness of our global pare-wise distillation which is able to transfer the whole structure information other than a local boundary information [45].

Table 2: Empirical comparison of feature transfer MIMIC [35, 21], attention transfer [51], and local pair-wise distillation [45] to our global pair-wise distillation. The segmentation is evaluated by mIoU (%). PI: pixel-wise distillation. MIMIC: using a 1×1 convolution for feature distillation. AT: attention transfer for feature distillation. LOCAL: The local similarity distillation method. PA: our pair-wise distillation. ImN: initializing the network from the weights pretrained on ImageNet dataset.

Method	ResNet18 (0.5)	ResNet18 (1.0) + ImN
w/o distillation	55.37	69.10
+ PI	57.07	70.51
+ PI + MIMIC	58.44	71.03
+ PI + AT	57.93	70.70
+ PI + LOCAL	58.62	70.86
+ PI + PA	**61.03**	**71.37**

Table 3: The segmentation results on the testing, validation (Val.), training (Tra.) set of Cityscapes.

Method	#Params (M)	FLOPs (B)	Test §	Val.	Tra.
Current state-of-the-art results					
ENet [31] †	0.3580	3.612	58.3	n/a	n/a
ERFNet [48] ‡	2.067	25.60	68.0	n/a	n/a
FCN [38] ‡	134.5	333.9	65.3	n/a	n/a
RefineNet [23] ‡	118.1	525.7	73.6	n/a	n/a
OCNet [50]‡	62.58	548.5	80.1	n/a	n/a
PSPNet [56] ‡	70.43	574.9	78.4	n/a	n/a
Results w/ and w/o distillation schemes					
MD [45] ‡	14.35	64.48	n/a	67.3	n/a
MD (Enhanced) [45] ‡	14.35	64.48	n/a	71.9	n/a
ESPNet-C [28] †	0.3492	3.468	51.1	53.3	65.9
ESPNet-C (ours) †	0.3492	3.468	57.6	59.9	70.0
ESPNet [28] †	0.3635	4.422	60.3	61.4	n/a
ESPNet (ours) †	0.3635	4.422	62.0	63.8	73.8
ResNet18 (0.5) †	3.835	33.35	54.1	55.4	60.7
ResNet18 (0.5) (ours) †	3.835	33.35	60.5	61.6	66.1
ResNet18 (1.0) †	15.24	128.2	56.0	57.5	63.0
ResNet18 (1.0) (ours) †	15.24	128.2	62.1	63.2	69.9
ResNet18 (1.0) ‡	15.24	128.2	67.6	69.1	74.1
ResNet18 (1.0) (ours) ‡	15.24	128.2	71.4	72.7	77.4
MobileNetV2Plus [25] ‡	8.301	86.14	68.9	70.1	n/a
MobileNetV2Plus (ours) ‡	8.301	86.14	74.0	74.5	83.1

† Train from scratch
‡ Initialized from the weights pretrained on ImageNet
§ We test all our models on single scale. Some cumbersome networks are test on multiple scales, such as OCNet and PSPNet.

5.4. Results

Cityscapes. We apply our structure distillation method to several compact networks: MobileNetV2Plus [25] which is based on a MobileNetV2 model, ESPNet-C [28] and ESPNet [28] which are carefully designed for mobile applications. Table 3 presents the segmentation accuracy, the

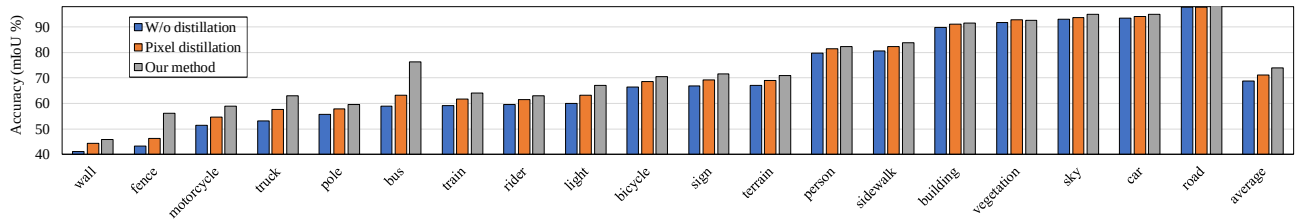
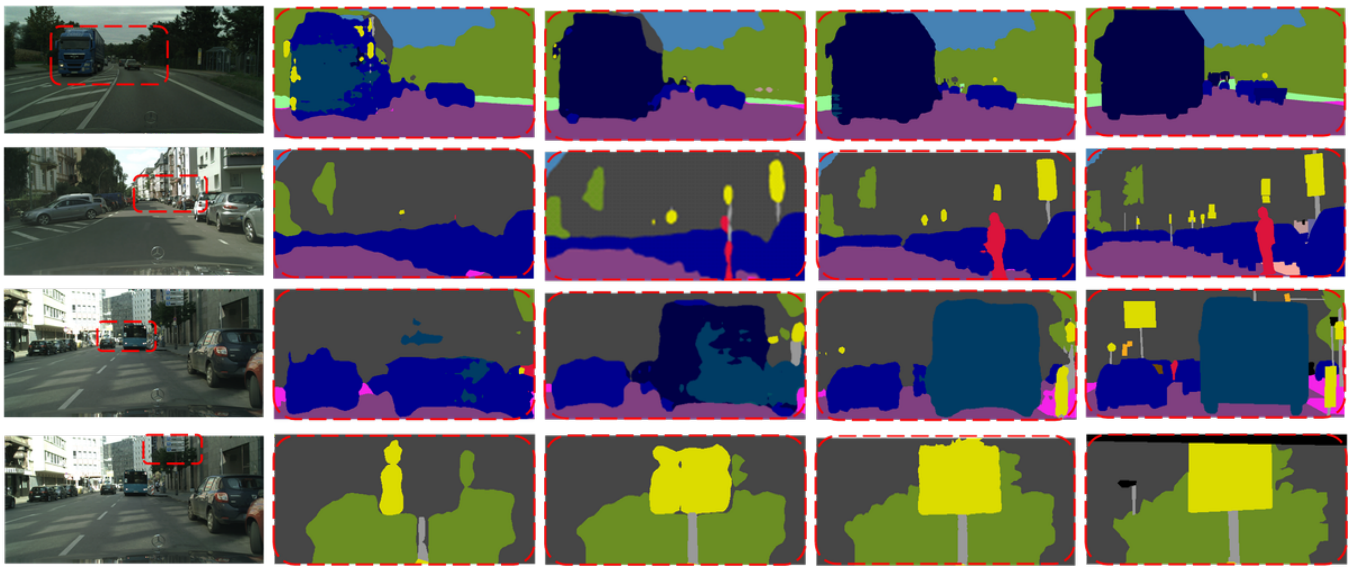


Figure 4: Illustrations of the effectiveness of pixel-wise and structured distillation schemes in terms of class IoU scores on the network MobileNetV2Plus [25] over the Cityscapes test set. Both pixel-level and structured distillation are helpful for improving the performance especially for the hard classes with low IoU scores. The improvement from structured distillation is more significant for structured objects, such as bus and truck.



(a) Image **(b)** W/o distillation **(c)** Pixel-wise distillation **(d)** Our method **(e)** Ground truth

Figure 5: Qualitative results on the Cityscapes testing set produced from MobileNetV2Plus: (a) initial images, (b) w/o distillation, (c) only w/ pixel-wise distillation, (d) Our distillation schemes: both pixel-wise and structured distillation schemes. The segmentation map in the red box about four structured objects: trunk, person, bus and traffic sign are zoomed in. One can see that the structured distillation method (ours) produces more consistent labels.

model complexity and the model size. GLOPs[2] is calculated on the resolution 512×1024 to evaluate the complexity. #parameters is the number of network parameters. We can see that our distillation approach can improve the results over 5 compact networks: ESPNet-C and ESPNet [28], ResNet18 (0.5), ResNet18 (1.0), and MobileNetV2Plus [25]. For the networks without pre-training, such as ResNet18 (0.5), ResNet18 (1.0) and ESPNet-C, the improvements are very significant with 6.2%, 5.74% and 6.6%, respectively. Compared with MD (Enhanced) [45] that uses the pixel-wise and local pair-wise distillation schemes over MobileNet, our approach with the similar network MobileNetV2Plus achieves higher segmentation quality (74.5 vs 71.9 on the validation set) with a little higher computation complexity and much smaller model size.

Figure 4 shows the IoU scores for each class over MobileNetV2Plus. Both the pixel-wise and structured distillation schemes improve the performance, especially for the categories with low IoU scores. In particular, the structured distillation (pair-wise and holistic) has significant improvement for structured objects, e.g., 17.23% improvement for Bus and 10.03% for Truck. The qualitative segmentation results in Figure 5 visually demonstrate the effectiveness of our structured distillation for structured objects, such as trucks, buses, persons, and traffic signs.

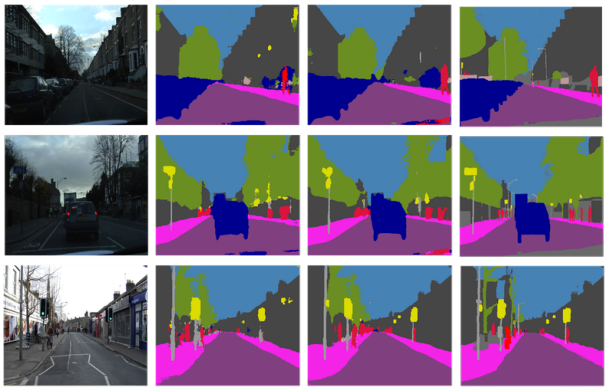
CamVid. Table 4 shows the performance of the student networks w/o and w/ our distillation schemes and state-of-the-art results. We train and evaluate the student networks w/ and w/o distillation at the resolution 480×360 following the setting of ENet. Again we can see that the distillation scheme improves the performance. Figure 6 shows some samples on the CamVid test set w/o and w/ the distillation produced from ESPNet.

We also conduct an experiment by using an extra unla-

[2]The FLOPs is calculated with the pytorch version implementation [1]

Table 4: The segmentation performance on the test set of CamVid. ImN = ImageNet dataset, and unl = unlabeled street scene dataset sampled from Cityscapes.

Method	Extra data	mIoU (%)	#Params (M)
ENet[31]	no	51.3	0.3580
FC-DenseNet56[11]	no	58.9	1.550
SegNet[3]	ImN	55.6	29.46
DeepLab-LFOV[5]	ImN	61.6	37.32
FCN-8s[38]	ImN	57.0	134.5
ESPNet-C[28]	no	56.7	
ESPNet-C (ours)	no	60.3	0.3492
ESPNet-C (ours)	unl	64.1	
ESPNet[28]	no	57.8	
ESPNet (ours)	no	61.4	0.3635
ESPNet (ours)	unl	65.1	
ResNet18	ImN	70.3	
ResNet18 (ours)	ImN	71.0	15.24
ResNet18 (ours)	ImN+unl	72.3	



(a) Image **(b)** W/o dis. **(c)** Our method **(d)** Ground truth

Figure 6: Qualitative results on the CamVid test set produced from ESPNet. W/o dis. represents for the baseline student network trained without distillation.

beled dataset, which contains 2000 unlabeled street scene images collected from the Cityscapes dataset, to show that the distillation schemes can transfer the knowledge of the unlabeled images. The experiments are done with ESPNet and ESPNet-C. The loss function is almost the same except that there is no cross-entropy loss over the unlabeled dataset. The results are shown in Figure 7. We can see that our distillation method with the extra unlabeled data can significantly improve mIoU of ESPNet-c and ESPNet for 13.5% and 12.6%.

ADE20K. The ADE20K dataset is a very challenging dataset and contains 150 objects. The frequency of objects appearing in scenes and the pixel ratios of different objects follow a long-tail distribution. For example, the stuff classes like wall, building, floor, and sky occupy more than 40% of all the annotated pixels, and the discrete objects, such as vase and microwave at the tail of the distribution, occupy only 0.03% of the annotated pixels.

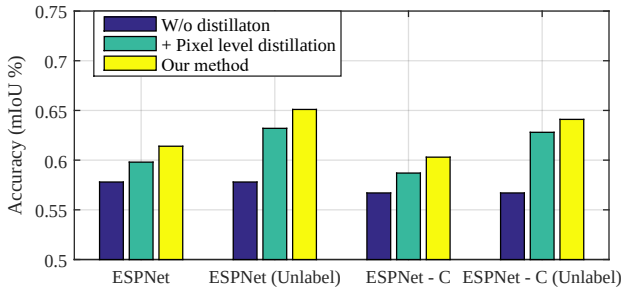


Figure 7: The effect of structured distillation on CamVid. We can see that distillation can improve the results in two cases: trained over only the labeled data and over both the labeled and extra unlabeled data.

Table 5: mIoU and pixel accuracy on validation set of ADE20K.

Method	mIoU(%)	Pixel Acc. (%)	#Params (M)
SegNet [3]	21.64	71.00	29.46
DilatedNet50 [44]	34.28	76.35	62.74
PSPNet (teacher) [56]	42.19	80.59	70.43
FCN [38]	29.39	71.32	134.5
ESPNet [28]	20.13	70.54	0.3635
ESPNet (ours)	23.91	73.94	0.3635
MobileNetV2Plus [25]	33.64	74.38	8.301
MobileNetV2Plus (ours)	35.51	76.20	8.301
ResNet18 [44]	33.82	76.05	15.24
ResNet18 (ours)	36.55	77.77	15.24

We report the results for ResNet18 and the MobileNetV2Plus which are trained with the initial weights pretrained on the ImageNet dataset, and ESPNet which is trained from scratch in Table 5. All the results are tested on single scale. For ESPNet, with our distillation, we can see that the mIoU score is improved by 3.78%, and it achieves a higher accuracy with samller #parameters compared to SegNet. For ResNet18, after the distillation, we have a 2.73% improvement over the one without distillation reported in [44]. We check the result for each class and find that the improvements are mainly from the discrete objects.

6. Conclusion

We study knowledge distillation for training compact semantic segmentation networks with the help of cumbersome networks. In addition to the pixel-level knowledge distillation, we present two structural distillation schemes: pair-wise distillation and holistic distillation. We demonstrate the effectiveness of our proposed distillation schemes on several recently-developed compact networks on three benchmark datasets.

References

[1] https://github.com/warmspringwinds/pytorch-segmentation-detection/blob/master/pytorch_segmentation_detection/utils/flops_benchmark.py, 2018. 7

[2] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Proc. Advances in Neural Inf. Process. Syst.*, pages 2654–2662, 2014. 2

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, (12):2481–2495, 2017. 2, 8

[4] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *Proc. Eur. Conf. Comp. Vis.*, pages 44–57. Springer, 2008. 2, 5

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proc. Int. Conf. Learn. Representations*, 2015. 1, 2, 8

[6] 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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 1, 2

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proc. Eur. Conf. Comp. Vis.*, 2018. 1, 2

[8] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial PoseNet: A structure-aware convolutional network for human pose estimation. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1212–1221, 2017. 3

[9] Y. Chen, N. Wang, and Z. Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *Proc. Eur. Conf. Comp. Vis.*, 2018. 2

[10] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 2, 5

[11] S. J. M. Drozdzal, D. Vazquez, and A. R. Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *Proc. Workshop of IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 8

[12] I. J. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Z. Ghahramani, and M. Welling. Generative adversarial nets. *Proc. Advances in Neural Inf. Process. Syst.*, 3:2672–2680, 2014. 3

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 5767–5777, 2017. 4

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016. 2, 4

[15] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 1, 2, 3

[16] 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: Comp. Res. Repository*, abs/1704.04861, 2017. 2

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *CVPR*, pages 2261–2269, 2017. 2

[18] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *arXiv: Comp. Res. Repository*, abs/1602.07360, 2016. 2

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *Proc. Eur. Conf. Comp. Vis.*, pages 694–711, 2016. 3

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *Proc. Int. Conf. Learn. Representations*, 2018. 3

[21] Q. Li, S. Jin, and J. Yan. Mimicking very efficient network for object detection. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7341–7349, 2017. 2, 6

[22] S. Z. Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009. 1

[23] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *CVPR*, pages 5168–5177, 2017. 1, 2, 6

[24] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Exploring context with deep structured models for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 2

[25] H. Liu. Lightnet: Light-weight networks for semantic image segmentation. https://github.com/ansleliu/LightNet, 2018. 2, 4, 6, 7, 8

[26] Y. Liu, Z. Qin, T. Wan, and Z. Luo. Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks. *Neurocomputing*, 311:78–87, 2018. 3

[27] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *arXiv: Comp. Res. Repository*, abs/1611.08408, 2016. 3

[28] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. *Proc. Eur. Conf. Comp. Vis.*, 2018. 2, 4, 5, 6, 7, 8

[29] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv: Comp. Res. Repository*, abs/1411.1784, 2014. 3, 4

[30] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1520–1528, 2015. 2

[31] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv: Comp. Res. Repository*, abs/1606.02147, 2016. 1, 2, 6, 8

[32] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2536–2544, 2016. 3

[33] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *Proc. Int. Conf. Mach. Learn.*, pages 1060–1069, 2016. 3

[34] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo. Efficient convnet for real-time semantic segmentation. In *IEEE Intelligent Vehicles Symp.*, pages 1789–1794, 2017. 1, 2

[35] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv: Comp. Res. Repository*, abs/1412.6550, 2014. 1, 2, 6

[36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention, MICCAI*, pages 234–241, 2015. 2

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 4

[38] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640, 2017. 1, 2, 6, 8

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–9, 2015. 2

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2818–2826, 2016. 2

[41] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich, et al. Speeding up semantic segmentation for autonomous driving. In *Proc. Workshop of Advances in Neural Inf. Process. Syst.*, 2016. 2

[42] G. Urban, K. J. Geras, S. E. Kahou, O. Aslan, S. Wang, R. Caruana, A. Mohamed, M. Philipose, and M. Richardson. Do deep convolutional nets really need to be deep (or even convolutional)? In *Proc. Int. Conf. Learn. Representations*, 2016. 2

[43] H. Wang, Z. Qin, and T. Wan. Text generation based on generative adversarial nets with latent variables. In *Proc. Pacific-Asia Conf. Knowledge discovery & data mining*, pages 92–103, 2018. 3

[44] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proc. Eur. Conf. Comp. Vis.*, 2018. 8

[45] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W.-S. Zheng. Improving fast segmentation with teacher-student learning. *Proc. British Machine Vis. Conf.*, 2018. 2, 3, 6, 7

[46] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 1

[47] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 2

[48] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *Proc. Int. Conf. Learn. Representations*, 2016. 1, 2, 6

[49] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proc. AAAI Conf. Artificial Intell.*, pages 2852–2858, 2017. 3

[50] Y. Yuan and J. Wang. Ocnet: Object context network for scene parsing. In *arXiv: Comp. Res. Repository*, volume abs/1809.00916, 2018. 1, 2, 6

[51] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *Proc. Int. Conf. Learn. Representations*, 2017. 2, 6

[52] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *CoRR*, volume abs/1805.08318, 2018. 4

[53] T. Zhang, G. Qi, B. Xiao, and J. Wang. Interleaved group convolutions. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 4383–4392, 2017. 2

[54] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 2

[55] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. *Proc. Eur. Conf. Comp. Vis.*, 2018. 1, 2

[56] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2881–2890, 2017. 1, 2, 4, 6, 8

[57] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, pages 267–283, 2018. 4

[58] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 2, 5