Improving Fast Segmentation With Teacher-student Learning

Jiafeng Xie[1]
xiejf6@mail2.sysu.edu.cn
Bing Shuai[2]
bshuai001@ntu.edu.sg
Jian-Fang Hu[134]
hujf5@mail.sysu.edu.cn
Jingyang Lin[1]
linjy47@mail2.sysu.edu.cn
Wei-Shi Zheng[14]
wszheng@ieee.org

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Nanyang Technological University, Singapore
[3] Guangdong Key Laboratory of Information Security Technology
[4] Key Laboratory of Machine Intelligence and Advanced Computing, MOE

Abstract

Recently, segmentation neural networks have been significantly improved by demonstrating very promising accuracies on public benchmarks. However, these models are very heavy and generally suffer from low inference speed, which limits their application scenarios in practice. Meanwhile, existing fast segmentation models usually fail to obtain satisfactory segmentation accuracies on public benchmarks. In this paper, we propose a teacher-student learning framework that transfers the knowledge gained by a heavy and better performed segmentation network (i.e. teacher) to guide the learning of fast segmentation networks (i.e. student). Specifically, both zero-order and first-order knowledge depicted in the fine annotated images and unlabeled auxiliary data are transferred to regularize our student learning. The proposed method can improve existing fast segmentation models without incurring extra computational overhead, so it can still process images with the same fast speed. Extensive experiments on the Pascal Context, Cityscape and VOC 2012 datasets demonstrate that the proposed teacher-student learning framework is able to significantly boost the performance of student network.

1 Introduction

Recently, segmentation performance has been lamdrastically improved in deep learning era, where end-to-end segmentation networks [3, 21, 39] are developed to generate high-fidelity segmentation maps on challenging real-world images [2, 5, 22]. For example, deeper and higher-capacity ConvNets [8, 21, 40] are adopted in Fully Convolutional Network (FCN) to enhance its segmentation accuracies. Some researchers are dedicated to aggregating informative contexts for local feature representation, which leads to advancement of segmentation network architectures for boosting the segmentation accuracies. Meanwhile, a

large body of research works focus on refining local segmentation results to obtain non-trivial performance enhancement of segmentation networks. Some representative works include DeepLab-v2 [3], DilatedNet [36], CRF-CNN [16], DAG-RNN [29], PSPNet [39], etc. In addition, some researchers also explore refining the low-level segmentation details (e.g. sharper object boundaries and better segmentation accuracies for small-size objects) by either adopting fully connected CRF as a post-processing module [3, 40] or by learning a deconvolutional network on top of coarse prediction maps [23]. It's interesting to see that segmentation accuracies are progressively improved on public benchmarks, however, these models become less likely to be applicable in real-world scenarios where short latency is essential and computational resources are limited.

With the above concerns in mind, recent researchers start to ingest the recent advancement of network architectures, and integrate them to develop faster segmentation networks. Representative examples are ENet [24] and SegNet [2]. Although promising, their segmentation accuracies are still bounded by the model capacity. In general, they fail to obtain comparable segmentation accuracies compared to heavy and deep segmentation networks. In this paper, we propose a novel framework to improve the performance of fast segmentation network without incorporating extra model parameters or incurring extra computational overhead, and thus can keep the inference speed of the fast segmentation network to be unchanged. To this end, we propose a novel teacher-student learning framework to make use of the knowledge gained in a teacher network. Specifically, our framework intend to regularize the student learning by the zero-order and first-order knowledge obtained from teacher network on fine annotated training data. To distill more knowledge from teacher, we further extend our framework by integrating the teacher-student learning on fine annotated training data and unlabeled auxiliary data. Our experiments show that the proposed teacher-student learning framework can boost the performance of student network by a large margin.

Our contribution is three-fold: (1) a novel teacher-student learning framework for improving fast segmentation models, with the help of an auxiliary teacher network; (2) a joint learning framework for distilling knowledge from teacher network based on both fine annotated data and unlabeled data; and (3) extensive experiments on three segmentation datasets demonstrating the effectiveness of the proposed method.

2 Related work

In the following, we review the literatures that are most relevant to our work, including researches of architecture evolvement for semantic segmentation and knowledge distillation.

Accuracy Oriented Semantic Segmentation. This line of research covers most of published literatures in semantic segmentation. In a nutshell, the goal is to significantly improve the segmentation accuracy on public segmentation benchmarks. Following the definition of a general segmentation network architecture in Shuai et al. [30], we categorize the literatures to three aspects that improve the segmentation performance. In one aspect, performance enhancement is largely attributed to the magnificent progress of pre-trained ConvNet [15] [32] [31][8] [12], which is simply adapted to be the local feature extractor in segmentation networks. The core of this progress is to obtain better ConvNet model on large-scale image datasets (e.g. ImageNet [27]) by training deeper or more complicated networks. Meanwhile, many researchers are dedicated to developing novel computational layers that are able to effectively encode informative context into local feature maps. This research direction plays a significant role to enhance the visual quality of prediction label maps as well

as to boost the segmentation accuracy. Representative works, such as DPN [20], CRF-CNN [16] , DAG-RNN [30], DeepLab-v2 [3] , RecursiveNet [28], ParseNet [19], DilatedNet [36] , RefineNet [17], PSPNet [39] formulated their computational layers to achieve effective context aggregation, and they can significantly improve the segmentation accuracy on Pascal VOC benchmarks. In addition, research endeavours have also been devoted to recovering the detailed spatial information by either learning a deep decoder network [2][23] [34] or applying a disjoint post-processing module such as fully connected CRF [14] [3][40]. These techniques have collectively pushed the segmentation performance saturating on Pascal VOC benchmarks[1]. The steady progress also calls for the unveiling of new and more challenging benchmarks (e.g. Microsoft COCO [18] dataset). Although significant progress has been made regarding to the visual quality of segmentation predictions, these models are usually computationally intensive. Thus, they are problematic to be directly applied to resource constrained embedded devices and can not be used for real-time applications.

Fast Semantic Segmentation. Recently, another line of research emerges as of state-of-the-art models achieve saturating segmentation accuracies on urban street images [9]. Its goal is to develop fast segmentation models that has the potential to be applied in real-world scenarios. For example, Paszke et al.[24] adopted a light local feature extraction network in their proposed ENet, which can be run in real-time for moderate sized images (e.g. 500 x 500). Zhao et al. [38] developed the ICNet that only fed the heavy model with downsampled input images, so the inference speed of ICNet remains competitively fast. One problem of these works are that the performance of these models are not satisfactory due to their lower capacity. In this paper, we propose to improve the performance of fast segmentation networks by regularizing their learning with the knowledge learned by a heavy and accurate teacher model. In this regard, this line of research is orthogonal and complementary to our teacher-student learning framework.

Knowledge Distillation via Teacher-Student Learning. In image classification community, knowledge distillation [6, 9, 13, 35, 37] has been widely adopted to improve the performance of fast and low-capacity neural networks. Hinton et al. [9] pioneered to propose transfer the "dark knowledge" from an ensemble of networks to a student network, which leads to a significant performance enhancement on ImageNet 1K classification task [27]. Romero et al. [25] further extended the knowledge transfer framework to allow it happens in intermediate feature maps. Their proposed FitNet[25] allowed the architecture of student network to go deeper and more importantly to achieve better performance. Huang et al. [13] proposed to regularize the student network learning by mimicking the distribution of activations of intermediate layers in a teacher network. In addition, knowledge distillation has also been sucessfully in pedestrian detection [11] and face recognition [33] as well. Recently, Ros et al. [26] explored and discussed different knowledge transfer framework based on the output probability of a teacher deconvolutional network, and they observed segmentation accuracy improvement of student networks. Our methods differ from it in the following aspects: (1) both zero-order and first-order knowledge from teacher models are transferred to student; and (2) unlablled auxiliary data are used to encode the knowledge of teacher models, which is further transferred to the student models and improve their performance.

[1] More than 85% mean IOU has been achieved by state-of-the-art segmentation models.

3 Approach

Our approach involves two kind of deep networks: student network and teacher network. The student network is a deep network for segmentation with a shallower architecture. Thus it can segment images with a fast speed. In contrast, the teacher network is a deeper network with more complex architectures. Thus, it typically performs better than the student network in the term of segmentation accuracy, but has a slower segmentation speed. In this work, we propose a teacher-student learning framework to improve the student learning with the guidance of a teacher network. The proposed overall framework is summarized in Figure 1. In the following, we discuss it in detail.

3.1 Teacher-student learning for fine annotated data

Here, we describe how to facilitate the learning of student network with the help of a teacher network based on the provided fine annotated training data. Let's denote the student network and teacher network as S and T, respectively. In order to transfer enough informative knowledge from teacher network for learning a robust student network, we formulate the objective function for our student-teacher learning as

$$L = L_S + r(S,T) \tag{1}$$

where L_S indicates the traditional segmentation (cross entropy) loss for the employed student network. $r(S,T)$ is a function indicating the knowledge bias between the learned student network and teacher network. It serves as a regularization term for regularizing our student learning. In this term, the student and teacher networks are connected together and the knowledge can be distilled from teacher network T to student network S by minimizing L. Here, we define $r(S,T)$ as

$$r(S,T) = \alpha L_p(S,T) + \beta L_c(S,T) \tag{2}$$

L_p is the probability loss defined as $L_p(S,T) = \sum_{i=1,2,\ldots,I} \sum_{\mathbf{x}\in G} \|\mathbf{p}_S^i(\mathbf{x}) - \mathbf{p}_T^i(\mathbf{x})\|_2^2$, where I is the batch size of model's input and $\mathbf{p}_S^i(\mathbf{x}), \mathbf{p}_T^i(\mathbf{x}) \in R^n$ are the probability outputs of the student and teacher network at pixel $\mathbf{x}$ in the image region G. It is defined in the way such that the probability output of the student network is similar with that of the teacher network. This function can capture the zero-order knowledge between different segmentation outputs.

In complement to L_p, term L_c is used to capture the first-order knowledge between outputs of student and teacher network. We formulate it as $\sum_{i=1,2,\ldots,I} \sum_{\mathbf{x}\in G} \|\mathbf{c}_S^i(\mathbf{x}) - \mathbf{c}_T^i(\mathbf{x})\|_2^2$, where I is the batch size of model's input and the consistence map $\mathbf{c}$ is defined as $\mathbf{c}_\cdot^i(\mathbf{x}) = \sum_{\mathbf{y}\in B(\mathbf{x})} \|l(\mathbf{y}) - l(\mathbf{x})\|_2^2$. Here, $B(\mathbf{x})$ indicates the 8-neighborhood of pixel $\mathbf{x}$ and l is the logits output of the corresponding network. This term is employed to ensure that the segmented boundary information obtained by student and teacher networks can be closed with each other. By this way, the teacher network provides some useful fist-order knowledge for regularizing our student network learning.

Overall, the above two loss terms (i.e., L_p and L_c) constrain the student network learning from different perspectives. They complement well with each other for improving the learning of shallowed student network. Our scheme has the following characteristics for segmentation. First, it can improve the student segmentation network without incurring extra computational overhead. Second, both the zero-order and first-order knowledge are transferred from teacher to guide our student learning.

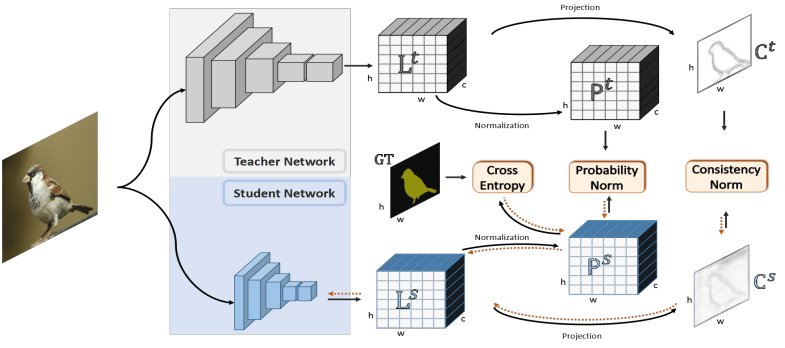

Figure 1: The detailed architecture of our teacher-student learning framework. In the framework optimization, only the student networks are updated by back-propagation (indicated by red dashed lines) and the teacher network are fixed. For transferring zero-order and first-order knowledge, we generate the probability maps and consistency maps using the logits output of the teacher and student networks.

3.2 Teacher-student learning with unlabeled data

In addition to the fine annotated images in the training set, we can also obtain a large number of unlabeled images from the Internet for the network training. However, it is unrealistic to annotate all the available images as manually annotating image for segmentation in a pixel-level is quite time consuming. Here, we illustrate that our teacher-student learning framework can be easily extended to make use of these unlabeled images for further improving the learning of student network. In the framework, we treat the segmentation results obtained by the teacher network as the ground truth segmentations for the unlabeled images and then conduct our teacher-student learning on the unlabeled data. Therefore, here we have a total of two teacher-student learning, one is conducted on the manually labeled training set with fine annotations and the other is conducted on the unlabeled data with noisy annotations generated by the teacher network. Both the teacher-student learning can be learned jointly. Specifically, the objective function for our teacher-student learning with both labeled and unlabeled data can be formulated as

$$L = L_{LabeledData} + \lambda L_{unlabeledData} \tag{3}$$

where $L_{LabeledData}$ is the loss for the teacher-student learning on the fine annotated training data. $L_{unlabeledData}$ indicates the loss for the teacher-student learning on the unlabeled data. Here, we use parameter λ to control the balance of teacher-student learning for different data. Finally, our teacher-student learning with unlabeled data is achieved by minimizing the loss L defined in (3).

4 Experiments

4.1 Ablation study

In this section, we perform ablation studies on Pascal Context [22] to justify the effectiveness of our technical contributions in Section 3. We adopt state-of-the-art segmentation

architecture DeepLab-v2[3] as our teacher and student network in the ablation analysis. In detail, DeepLab-v2 is a stack of two consecutive functional components: (1), a pre-trained ConvNet for local feature extraction (feature backbone network); and (2), Atrous Spatial Pyramid Pooling (ASPP) network for context aggregation. In general, the model capacity of DeepLab-v2 largely correlates with that of feature backbone network. Thus in our ablation experiments, we instantiate our teach network with a higher capacity feature backbone network – ResNet-101 [8], and employ a recent more computational efficient network MobileNet[10] in student network.[2]

Dataset: Pascal Context [22] has 10103 images, out of which 4998 images are used for training. The images are from Pascal VOC 2010 datasets, and they are annotated as pixelwise segmentation maps which include 540 semantic classes (including the original 20 classes). Each image has approximately 375×500 pixels. Similar to Mottaghi et al. [22], we only consider the most frequent 59 classes in the dataset for evaluation.

Implementation Details: The segmentation networks are trained by batch-based stochastic gradient descent with momentum (0.9). The learning rate is initialized as 0.1, and it is dropped by a factor of 10 after 30, 40 and 50 epoches are reached (60 epoches in total). The images are resized to have maximum length of 512 pixels, and they are zero padded to have square size to allow for batch processing. Besides, general data augmentation methods are used in network training, such as randomly flipping the images, randomly performing scale jitter (scales are between 0.5 to 1.5), etc. α and β in Equation 2 are empirically set to 4 and 0.4 respectively.we have validate that alpha and beta need to make the probability loss,consistency loss and crose entropy on the same order of magnitude. We randomly take 10k unlabeled images from COCO unlabel 2017 dataset[18], and use the teacher segmentation network to generate their pseudo ground truth pixelwise maps. To reduce noises, pixels will not be annotated if their corresponding class likelihood is less confident than 0.7. We implement the proposed network architecture in Tensorflow [1] framework and the algorithms are run on a GPU 1080Ti device.

Results: The results of ablation studies are shown in Table 1, where we can observe that the teacher network achieves 48.5% mIoU [3] at 16.7 fps and the student network yields 40.9% mIoU at 46.5 fps. Not surprisingly, teacher network significantly outperforms its student counterpart in terms of segmentation accuracy. In contrast, student network can run in real-time, and it has the potential to be applied in real-time application scenario. They are a reasonably appropriate teacher-student setting in our ablation experiments. As expected, the segmentation accuracy of our student network is improved by 1.4% to 42.3% mIOU if we transfer the possibility knowledge from teacher network by only considering L_p loss in Equation 2 i.e., the case of $\alpha = 4.0, \beta = 0$. It demonstrates that the probability distribution output by teacher network indeed carry informative knowledge for improving the learning of our student network. We can observe another promising 0.5% mIOU gain if the consistency loss L_c is further included. This encouraging result indicates that the proposed consistency loss is able to positively guiding the student network learning. Finally, when we use the 10K unlabeled images to further facilitate our teacher-student learning, we can observe a significant mIOU improvement (1.0%). This demonstrates that the knowledge gained by the teacher network can be embedded in unseen data, via which the knowledge is implicitly distilled to

[2]This simply represents a typical teacher-student pair, where teacher is a heavy and accurate segmentation network and student, in the contrary, is an efficient and less-accurate network.

[3]We note that the reported mIoU (48.5%) is much higher than that in [3] (44.7%). This is because that our model has been pre-trained on the MS-COCO dataset. We also find that freezing the batch normalization statistics can benefit our model training significantly, which mainly contributes to the performance superiority.

Model	mIOU(%)	speed (FPS)
ResNet-101-DeepLab-v2 (teacher) [3]	48.5	16.7
MobileNet-1.0-DeepLab-v2	40.9	46.5
MobileNet-1.0-DeepLab-v2 (L_p)	42.3	46.5
MobileNet-1.0-DeepLab-v2 (L_p +L_c)	42.8	46.5
MobileNet-1.0-DeepLab-v2 (L_p +L_c+$UnlabeledData$)	43.8	46.5
FCN-8s [21]	37.8	N/A
ParseNet [20]	40.4	N/A
UoA-Context + CRF [16]	43.3	< 1
DAG-RNN [30]	42.6	9.8
DAG-RNN + CRF [30]	43.7	< 1

Table 1: Comparison results on Pascal Context dataset. 'MobileNet-1.0-DeepLab-v2 (L_p)' indicates that probability loss is considered in the loss for knowledge transfer; 'MobileNet-1.0-DeepLab-v2 (L_p +L_c)' indicates that probability loss and consistency loss are both used for knowledge transfer; 'MobileNet-1.0-DeepLab-v2 (L_p +L_c+$UnlabeledData$)' represents that the unlabeled images are used in the knowledge transfer.

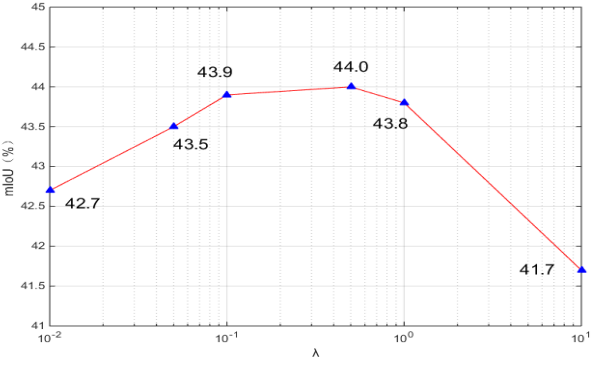

Figure 2: Evaluation on the influence of parameter λ.

the student network. Overall, the performance of student has been largely improved after digesting the knowledge from teacher with the proposed teacher-student learning framework. In comparison with state-of-the-arts, our enhanced student networks achieve very competitive results both in terms of inference efficiency as well as segmentation accuracy. We also experimentally find that our system is quite robust to the setting of some parameters like the rate of $L_{unlabeledData}$ in equation 3. For example, if we set λ to 0.1 or 1, the performance will only drop slightly (no more than 0.1%), which is illustrated in Figure 2.

In Figure 3, we present several interesting qualitative results. We can easily observe that additionally considering the information bias $r(S,T)$ in student loss function significantly improves the segmentation quality of student network. Specifically, the semantic predictions for "stuff" classes are smoother, and boundaries for "thin" classes (e.g. objects) are slightly shaper. Moreover, those prediction errors can be further decreased when more unlabeled images are incorporated into student network training.

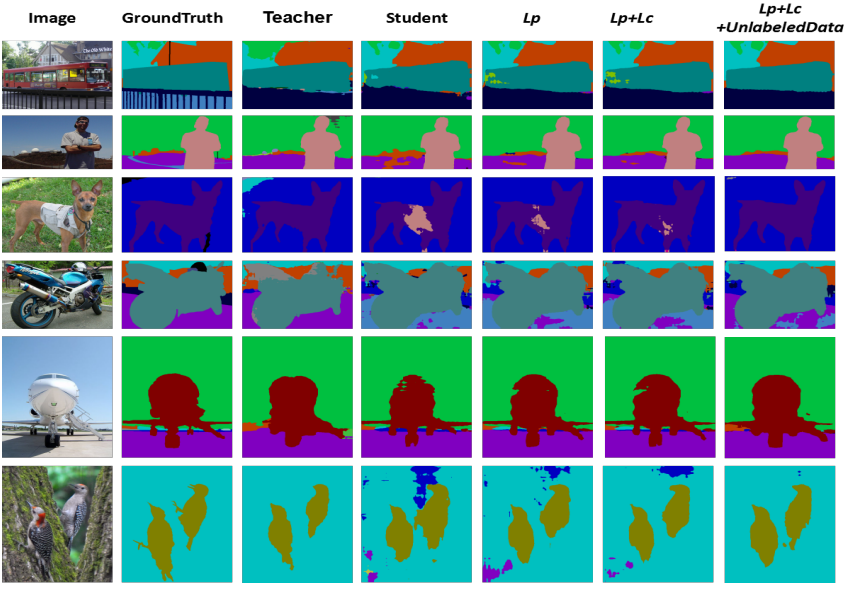

Figure 3: Some qualitative results on the Pascal Context dataset. The results illustrate that the proposed teacher-student learning framework can efficiently mine informative knowledge from teacher network to guide the learning of student network, and thus improve the performance of student network for segmenting objects. The figure is best viewed in color.

Backbone (Student)	Base(mIOU(%))	Enhanced(mIOU(%))	speed (FPS)
MobileNet-0.5	36.7	41.1	77.7
MobileNet-0.75	39.2	43.0	61.7
MobileNet-1.0	40.9	43.8	46.5

Table 2: Results of using different student networks on Pascal Context validation dataset. The same teacher network (ResNet-101-DeepLab-v2) is used to all three student networks.

4.2 More evaluations on the student network.

Given a pair of teacher and student networks, our teacher-student learning framework aims at mining informative knowledges from teacher network to improve the student learning. Here, we provide some experimental clues on how the students affects our teacher-student learning. Thus, we fix the teacher network, and instantiate three student networks whose performances (segmentation accuracy and speed) are significantly different. Specifically, three different DeepLab-v2 instances with MobileNet-1.0, MobileNet-0.75, and MobileNet-0.5 as the backbone are employed to form the student networks.

The detailed evaluation results are presented in Table 2, where the evaluated three teacher-student network settings have different performance gaps (7.6%, 9.3% and 11.8% mIoU). As expected, transferring knowledge from teacher network always improve the performances of students. Specifically, the segmentation accuracies of these three students are non-trivially improved by 2.9%, 3.8% and 4.4% mIOU, respectively. We can observe that the larger performance gaps between the teacher and student is, the more knowledge is gained, and thus

Model	mIoU (%)	speed (FPS)
SegNet [2]	56.3	19.6
ResNet-DeepLab-v2 [3]	70.9	7.4
PSPNet [39]	**80.1**	6.6
MobileNet-1.0-DeepLab-v2	67.3	20.6
MobileNet-1.0-DeepLab-v2 (Enhanced)	71.9	**20.6**

Table 3: Comparison with state-of-the-arts on Cityscapes validation set. ALL the inference speeds on this set are evaluated for the segmentation of 1024×512-sized images.

the higher performance improvement is observed for the student network. This observation suggests that the performance of a student network (e.g. MobileNet-1.0-DeepLab-v2) can be further improved by using a stronger teacher model.

4.3 Comparison with state-of-the-arts

To further demonstrate the effectiveness of the proposed teacher-student learning approach, we test our methods on Cityscape dataset and Pascal VOC2012 dataset. In the following experiments, we use MobileNet-1.0-DeeplabV2 and ResNet-101-Deeplab-v2 as our student and teacher, respectively. It's important to note that our learning framework is versatile to different teacher and student models, so the reported results can be simply improved by using either a stronger teacher or a better student model.

4.3.1 Experiments on Cityscapes dataset

The Cityscapes[4] is a large-scale dataset for semantic segmentation. This set is captured from the urban streets distributed in 50 cities for the purpose of understanding urban streets. The captured images have a large resolution of 2048×1024. For evaluation, a total of 5000 images are selected to be annotated in a fine scale and 20000 images are selected to be annotated coarsely. We follow the evaluation protocol in [4], where 2975, 500, and 1525 of the fine annotated images are selected to train, validate and test the model, respectively.

The detailed comparison results are presented in Table 3. As expected, by employing the proposed teacher-student learning framework, the performance of student network is improved to 71.9% mIoU, which is about 4.6% mIoU higher than the original student network. We can also note that the student network enhanced by our teacher-learning algorithm can even perform better than the employed teacher network (71.9 vs. 70.9), which demonstrates the effectiveness of our proposed teacher-student learning framework for mining informative knowledge to guide the learning of student network. We also observe that the enhanced student network can segment images at a fast speed with a good accuracy and it outperforms the SegNet [2] in terms of segmentation accuracy and speed.

4.3.2 Experiments on Pascal VOC2012

The Pascal VOC2012 dataset [5] consists 4369 images of 21 objects classes and a background class. For evaluation, the whole dataset is divided into training, validation, and test sets, each of which has 1446, 1449, and 1456 images, respectively. Following the experiment setup in SDS [7], the training set are extended to a set with 10,582.

Model	mIoU(%)	speed (FPS)
CRF-RNN[40]	72.9	7.6
Multi-scale[36]	73.9	16.7
ResNet-101-DeepLab-v2[3]	**75.2**	16.7
MobileNet-1.0-DeepLab-v2	67.3	46.5
MobileNet-1.0-DeepLab-v2 (Enhanced)	69.6	**46.5**

Table 4: Comparison with state-of-the-arts on VOC 2012 validation set.

The detailed comparison results are presented in Table 4. As can be seen, the student model enhanced by the proposed teacher-student learning framework can obtain a mIoU of 69.6% on this set, which outperforms the original student network by a margin of 2.3%. As compared with other state-of-the-arts [36, 40], our enhanced model has large advantage in terms of speed.

5 Conclusion

In this paper, we have proposed a teacher-student learning framework for improving the performance of existing fast segmentation models. In the framework, both the zero-order and first-order knowledges gained by a teacher network is distilled to regularize our student learning through both fine annotated and unlabeled data. Our experiments show that the proposed learning framework can largely improve the accuracy of student segmentation network without incurring extra computational overhead. The proposed framework mainly mine knowledge from one single teacher network for the student learning. In the future, we would explore multiple teachers based teacher-student learning framework.

Acknowledgment

This work was supported partially by the NSFC (No. 61702567, 61522115, 61661130157). The corresponding author for this paper is Jian-Fang Hu.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[5] M Everingham, L Van Gool, CKI Williams, J Winn, and A Zisserman. The pascal visual object classes challenge 2012 (voc2012). *Results*, 2012.

[6] ÇaÇğlar Gülçehre and Yoshua Bengio. Knowledge matters: Importance of prior information for optimization. *The Journal of Machine Learning Research*, 17(1):226–257, 2016.

[7] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[10] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[11] Qichang Hu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Fatih Porikli. Pushing the limits of deep cnns for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[12] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3, 2017.

[13] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[14] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[17] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[19] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[20] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 1377–1385. IEEE, 2015.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[22] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam Gyu Cho, Seong Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[23] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[24] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[25] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[26] German Ros, Simon Stent, Pablo F Alcantarilla, and Tomoki Watanabe. Training constrained deconvolutional networks for road scene semantic segmentation. *arXiv preprint arXiv:1604.01545*, 2016.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[28] Abhishek Sharma, Oncel Tuzel, and Ming-Yu Liu. Recursive context propagation network for semantic scene labeling. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2014.

[29] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Dag-recurrent neural networks for scene labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3620–3629, 2016.

[30] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Scene segmentation with dag-recurrent neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[33] Ying Tai, Jian Yang, Yigong Zhang, Lei Luo, Jianjun Qian, and Yu Chen. Face recognition with pose variations and misalignment via orthogonal procrustes regression. *IEEE Transactions on Image Processing*, 25(6):2673–2683, 2016.

[34] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. *arXiv preprint arXiv:1702.08502*, 2017.

[35] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[37] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[38] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. *arXiv preprint arXiv:1704.08545*, 2017.

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[40] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.