AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 55(1) (2023) 99-108
DOI: 10.22060/miscj.2023.21949.5307

An Efficient Knowledge Distillation Architecture for Real-time Semantic Segmentation

A. M. Mansourian, N. Karimi Bavandpour, Sh. Kasaei*

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

ABSTRACT: In recent years, Convolutional Neural Networks (CNNs) have made significant strides in the field of segmentation, particularly in semantic segmentation where both accuracy and efficiency are crucial. However, despite their high accuracy, these deep networks are not practical for real-time use due to their low inference speed. This issue has prompted researchers to explore various techniques to improve the efficiency of CNNs. One such technique is knowledge distillation, which involves transferring knowledge from a larger, cumbersome (teacher) model to a smaller, more compact (student) model. This paper proposes a simple yet efficient approach to address the issue of low inference speed in CNNs using knowledge distillation. The proposed method involves distilling knowledge from the feature maps of the teacher model to guide the learning of the student model. The approach uses a straightforward technique known as pixel-wise distillation to transfer the feature maps of the last convolution layer of the teacher model to the student model. Additionally, a pair-wise distillation technique is used to transfer pair-wise similarities of the intermediate layers. To validate the effectiveness of the proposed method, extensive experiments were conducted on the PascalVoc 2012 dataset using a state-of-the-art DeepLabV3+ segmentation network with different backbone architectures. The results showed that the proposed method achieved a balanced mean Intersection over Union (mIoU) and training time.

Review History:

Received: Nov. 15, 2022
Revised: Apr. 28, 2023
Accepted: Apr. 30, 2023
Available Online: Oct. 10, 2023

Keywords:

Computer Vision
Convolutional Neural Networks
Deep Learning
Semantic Segmentation
Knowledge Distillation
Deep Neural Networks

1- Introduction

Semantic segmentation is a pixel-wise classification problem that involves assigning a specific class or label to each pixel in an image. It is an essential topic in computer vision and has numerous real-world applications such as virtual reality, autonomous driving, and video surveillance. In recent years, several deep neural network-based approaches have been developed for semantic segmentation, resulting in superior performance. However, these methods, such as DeepLab [1] and PSPNet [2], come with cumbersome models and require costly computation, even though they have substantially improved the accuracy of segmentation. To address the issue of low inference speed in semantic segmentation, real-time architectures have been proposed, such as Enet [3], ESPNet [4], ICNet [5], and BiSeNet [6, 7]. Additionally, several strategies have been developed to reduce the size of models and improve their cost-effectiveness, including model pruning, model quantization, and knowledge distillation. Among these, knowledge distillation is currently receiving significant research attention. This technique involves training a smaller network under the supervision of a more extensive network, which was first proposed in [8]. Unlike other compression methods, knowledge distillation reduces the size of a network without considering the structural differences between the teacher and student networks. In order to train compact semantic segmentation networks, this paper investigates the effectiveness of the knowledge distillation technique, which has demonstrated success in classification tasks [8, 9]. Similar to most existing methods, the semantic segmentation problem is approached by treating it as a collection of individual pixel classification problems. The knowledge distillation approach is then applied to the pixel level. However, unlike classification task, semantic segmentation requires a structured output. As a result, long-range dependencies play a significant role in semantic segmentation, and the teacher and student models typically capture different long-range contextual information due to differences in their receptive fields. This research explores these differences and aims to develop a knowledge distillation approach that effectively transfers this contextual information across different models. Additionally, this work presents structured knowledge distillation, which transfers the structural information with pair-wise distillation using intermediate feature maps that are known to contain rich information [10, 11]. This pair-wise distillation, along with the pixel-wise distillation, gives the student network a wealth of information from the teacher network. To achieve this, an objective function that combines a conventional cross-entropy loss with the distillation losses is optimized. In summary, the main contributions of this work are as follows:

*Corresponding author's email: kasaei@sharif.edu

- Investigating a knowledge distillation strategy for training accurate compact semantic segmentation networks.
- Defining pixel-wise and pair-wise distillation approaches to transfer spatial information and long-range dependencies from teacher to student network.
- Validating the effectiveness of the method on the Pascal VOC 2012 [12] dataset with a state-of-the-art segmentation network; namely, DeepLabV3+ [13] with different backbones.

2- Related Works

This section provides an overview of the relevant literature related to the topic of this work. In particular, it reviews the state-of-the-art research on semantic segmentation and knowledge distillation.

Semantic Segmentation: Semantic segmentation is widely recognized as a challenging task, which involves combining global information with detailed local information to accurately predict the structure of an input image by classifying pixels into categories. Semantic segmentation networks are usually larger than classification networks, as they need to extract additional information beyond what is required for classification. The fully convolutional framework, which was first introduced in [14], provides several crucial improvements to segmentation network design. It can utilize pre-trained weights of classification networks, perform on variable input sizes, and be trained end-to-end. Two of the most powerful and popular segmentation networks in existence are DeepLabV3+ [13] and PSPNet [2]. Due to the flexible design of these networks, one can choose either a big and powerful or a small and efficient classifier network as their backbone. They both employ Atrous convolution and pyramid spatial pooling techniques to capture global context while preserving feature maps' resolution and details. In this work, the teacher network is DeepLab with ResNet101 [15] backbone, while the student networks are DeepLab with ResNet18 and MobileNet backbones.

In addition to cumbersome networks for highly accurate segmentation, real-time segmentation networks have become increasingly popular due to the need for highly accurate segmentation in real applications such as mobile applications. This is because highly efficient segmentation networks can segment data in a fraction of the time compared to cumbersome networks. Most works focus on creating lightweight networks by accelerating convolution operations using factorization methods. Enet [7], inspired by [1], incorporates multiple acceleration factors, such as multi-branch modules, early feature map resolution down-sampling, minimal decoder size, and filter tensor factorization, among others. ESPNet [4] replaces conventional convolution layers with a spatial pyramid of dilated convolutions to achieve lightweight segmentation networks. ICNet [5] employs cascading multi-resolution branches to increase efficiency, while BiSeNet [7] uses two branches to learn both spatial information and obtain a large receptive field: spatial and context paths. These lightweight segmentation networks have a significant impact on real-time applications, as they provide highly accurate segmentation results in a fraction of the time compared to cumbersome networks.

Knowledge Distillation: The concept of knowledge distillation was first introduced in [8], where the student network uses the teacher's predictions as soft labels instead of the ground-truth hard labels. Soft labels provide additional information about the problem structure and category relationships, making them useful for training the student network. The teacher-student framework is widely used for training compact student networks, and there are many other scenarios where it can be useful. Several recent works have explored its applications, and here are some examples:

- [16] trains a sequence of identical networks in a way that each network distills knowledge from the previously trained one, resulting in improved performance.
- [17] makes use of a method of channel-wise distillation that enables students to mimic the correct outputs of the teacher.
- [18] utilizes a review mechanism to use past feature maps as a guide for the current feature map's distillation.
- [19] employs auxiliary models to hold pruned intermediate layers of teacher and student, then distills them using the curriculum learning approach.
- [20] proposes a relation-based knowledge distillation framework for transformers.

The majority of the methods discussed are primarily designed for image classification, but [21] applied its approach to object detection as well. Other successful examples of work on object detection and classification include [22-27]. Following classification and detection, one of the earliest applications of distillation to semantic segmentation was introduced in [28]. In this work, the prediction of the teacher network was used instead of the ground truth to train the student network. This approach yielded better results because the teacher's output represents an easier distribution for the student network to learn. Due to the structural nature of semantic segmentation, additional distillation methods are applied in addition to pixel-wise knowledge distillation to transfer more structural information from teacher to student. Similar to [17], [29] attempts to take advantage of channel-wise distillation while employing a divide-and-conquer strategy because channel-wise distillation is time-consuming. [30] and [31] try to transfer class-wise similarity by creating class prototypes and category-wise similarity by constructing the correlation matrix, respectively. In [32], a consistency loss was introduced between the student and teacher networks to ensure their segmentation boundaries were similar. This was achieved by calculating the L2 norm of the difference between the output probabilities of the student and teacher networks, which was used as an additional loss. Authors of [33] employed channel and spatial correlation loss function in addition to adaptive cross-entropy loss, which adaptively uses ground-truth labels and teacher predictions. [34] investigated the design aspects of the feature distillation method by reviewing the position of feature maps to distill and distillation losses. They proposed a new distance function

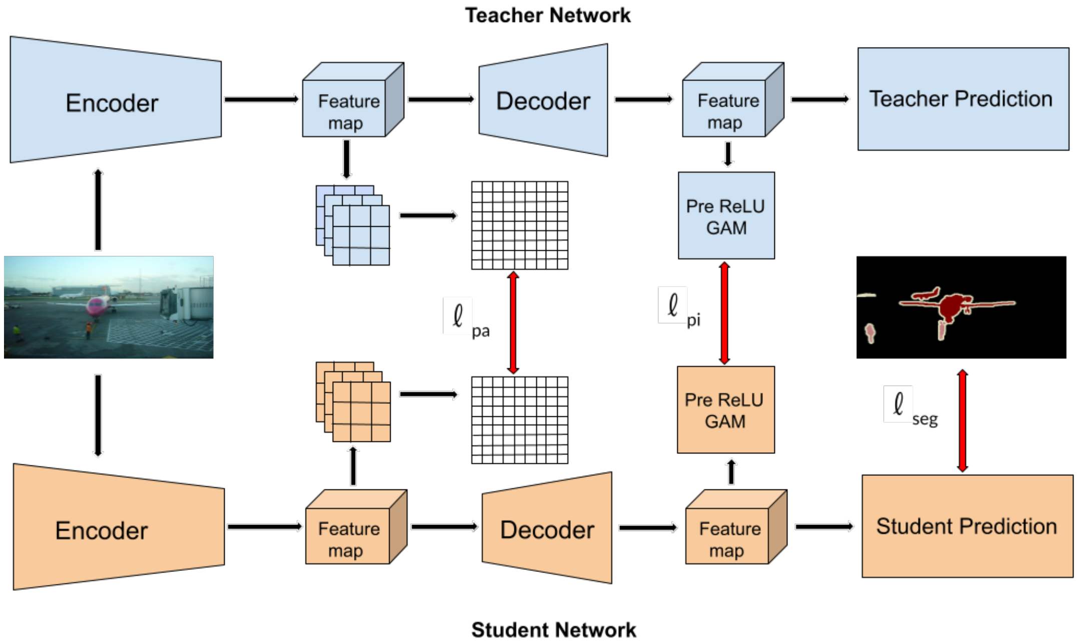

Fig. 1. Proposed framework for knowledge distillation. The architecture of both the teacher and student networks is Deeplab-V3 + [13], although their encoders are different. Student network encoder depth is shallower than that of teacher network encoder depth. Teacher network is fixed during the training process; only the student network will be trained with two distillation losses and cross-entropy loss. The pixel-wise distillation module uses the pre-ReLU feature map of the last convolution layer of the decoder before probability scores to transfer detailed spatial information. The pair-wise distillation module uses the feature map of the last layer of the encoder to create a pair-wise similarity matrix and transfer global information.

to distill meaningful information between the teacher and student using marginal ReLU. In [35], two novel distillation losses were introduced for segmentation. The pair-wise loss was defined as the mean square distance between elements of affinity matrices of the teacher and student networks. The affinity matrix contains inner products between every pair of features that encode pixels, and this loss aimed to encourage the student network to learn from the teacher network by minimizing the difference between their affinity matrices. The second loss, called holistic distillation, used adversarial learning to make feature maps of the student network similar to those of the teacher network, using a discriminator convolutional network. [36] is another relevant work that was developed parallel with [35]. This work also used an affinity loss, which is similar to the pair-wise loss in [35], except that an auto-encoder was trained for the last convolutional layer of the teacher network before computing its affinity matrix. The authors also used the direct L2 norm distance between the student's last convolutional features and the teacher's encoded features as an additional loss.

This paper utilizes the idea of pre-activation pixel-wise distillation, which was introduced in [34], to distill the knowledge of the last convolution layer of the teacher to the student. Additionally, a pair-wise distillation method, similar to [35], is used to transfer long-range information between the intermediate layers of the teacher and student networks.

3- Proposed Method

As mentioned before, [9] proposed a knowledge distillation approach to transfer information between two feature maps by applying the distance function pixel-by-pixel to all of their elements. [37] attempted to extract attention maps for the teacher and student networks from their feature maps in a specific layer by taking the summation over the channels of the feature maps. This allows the student network to see where the teacher network is concentrating and pushes the student to create an attention map that is similar to the teacher's. Because this method creates an attention matrix for each feature map, it is referred as Global Attention Map (GAM) distillation in this paper. This section delves deeper into the concept of transferring attention maps by distilling pre-activation attention maps using the idea of [34] and adding pair-wise loss similar to [35]. This will provide the student network with rich knowledge to mimic from the teacher network. Fig. (1) illustrates the diagram of the proposed approach and losses. The remainder of this section presents the mathematical notation, followed by the proposed method for creating attention maps and similarity matrices.

Finally, a loss function for distillation between the teacher and student networks is explained.

Let $A \in \mathbb{R}^{c \times h \times w}$ be an intermediate feature map obtained from a segmentation network with spatial dimensions of $h \times w$ and c channels. To index a specific element at depth k and spatial dimensions, we use the notation $A^k(x)$ where different feature maps are indexed as A_i. The matrix *A*'s element-wise power operation is represented by $| A^p |$. The Global Attention Map (GAM) attention matrix for the feature map at layer *i* is defined as per [37]

$$G_i = \sum_k | A_i^k |^p \tag{1}$$

where $p = 2$. The GAT matrix of the teacher network and the student network can be denoted as G_i^t and G_i^s, respectively. Then the loss function of the GAT distillation method can be expressed as described in [37]

$$\mathcal{L}_{GAT_i} = \left\| \frac{G_i^t}{\| G_i^t \|_2} - \frac{G_i^s}{\| G_i^s \|_2} \right\|_2 \tag{2}$$

where the loss function of the GAT distillation method is then used in combination with segmentation loss, where a weighted sum is applied.

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \sum_i \lambda_i \mathcal{L}_{GAT_i}. \tag{3}$$

in this context, the well-known cross-entropy loss function, $\mathcal{L}_{seg}$, is employed as the segmentation loss between the predictions of the student network and the ground-truth labels. Although the loss function in Eq. (2) was originally defined for image classification, it can be easily adapted for semantic segmentation. The distillation point in [9] is the end of an arbitrarily chosen intermediate layer, which has been demonstrated to have poor performance. The ReLU activation function selectively permits positive information to pass through and eliminates negative information, indicating that any knowledge distillation approach must account for this information loss. To address this issue, the pre-activation position is used for knowledge distillation, as it preserves both positive and negative values without any alteration, as demonstrated in [34]. To ensure effective distillation, an appropriate distance metric must be used based on the pre-ReLU position. In the teacher's feature map, positive responses are critical for the network, and they must be transferred with their exact values. If the student response is greater than the target value, it should be reduced for a negative teacher response. However, if the student response is lower than the target value, it need not be increased since negative values are already blocked by ReLU, regardless of their magnitudes. For any teacher and student feature maps, $T, S \in \mathbb{R}^{c \times w \times h}$, where the *i*-th component of the tensor is $T_i, S_i \in \mathbb{R}$, the partial L2 distance is defined as [34]

$$d_p(T,S) = \sum_i^{c \times w \times h} \begin{cases} 0 & if S_i \le T_i \le 0 \\ (T_i - S_i)^2 & otherwise. \end{cases} \tag{4}$$

then proposed pixel-wise loss between teacher and student is defined

$$\mathcal{L}_{pi} = d_p(L^t, L^s). \tag{5}$$

where L^t and L^s are GAM matrices of the last convolution layers of the teacher and student, respectively. These matrices are created based on Eq. (1) with $p = 1$, to preserve negative values. Although the GAM matrix neglects the information in the channels of the feature maps, simply summing over channels will reduce the training time of the method while still allowing the transfer of useful information. In addition to the pixel-wise loss, a loss is used that is pair-wise and analogous to [35]. Let *F* be a global feature produced by max pooling an intermediate feature map with proper stride size to create 3×3 features. These features are then flattened to create a feature vector of size 9 for each channel of the feature map as

$$f_i = Flatten(MaxPool(M^i)). \tag{6}$$

where $M \in \mathbb{R}^{c \times w \times h}$ is an intermediate feature map from the last convolution layer of the encoder and $f_i \in \mathbb{R}^9$; $1 \le i \le c$ is a new global feature created from *M*. Then similarity between the *i*-th and *j*-th pixel is calculated to create a similarity matrix, $E \in \mathbb{R}^{9 \times 9}$, as

$$e_{i,j} = \frac{f_i^T f_j}{\| f_i \|_2 \| f_j \|_2}. \tag{7}$$

finally, the squared difference is the basis for formulating the pair-wise similarity distillation loss in [35] as

$$\mathcal{L}_{pa}(E^t, E^s) = \frac{1}{(w \times h)^2} \sum_i^w \sum_j^h (e_{ij}^t - e_{ij}^s)^2. \tag{8}$$

where E^t and E^s are the similarity matrices of teacher and student, respectively. The overall loss function of our method then is a weighted sum of $\mathcal{L}_{seg}$, $\mathcal{L}_{pi}$, and $\mathcal{L}_{pa}$, defined by

Table 1. Effectiveness of the proposed distillation method on two student networks: MobileNet v2 and Resnet-18 with/without pixel-wise and pair-wise distillation modules. The results are average of three runs on the PascalVoc 2012 validation set.

Method	Pixel-wise	Pair-wise	mIoU(%)	Params(M)
Teacher: Deeplab-V3 + (ResNet-101)			74.78	59.3
Student: Deeplab-V3 + (ResNet-18)	n/a	n/a	66.59	16.6
Student: Deeplab-V3 + (ResNet-18)	✓		69.04	16.6
Student: Deeplab-V3 + (ResNet-18)		✓	68.47	16.6
Student: Deeplab-V3 + (ResNet-18)	✓	✓	69.21	16.6
Student: Deeplab-V3 + (MobileNet-V2)	n/a	n/a	62.92	5.8
Student: Deeplab-V3 + (MobileNet-V2)	✓		64.48	5.8
Student: Deeplab-V3 + (MobileNet-V2)		✓	63.56	5.8
Student: Deeplab-V3 + (MobileNet-V2)	✓	✓	64.71	5.8

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \alpha \mathcal{L}_{pi} + \beta \mathcal{L}_{pa}. \tag{9}$$

It should be noted that any potential differences in the spatial dimensions of the attention maps between the teacher and student networks can be rectified through a straightforward bilinear upsampling operation. As experiments of this research show, pixel-wise distillation achieves better results on the last layers, whereas pair-wise distillation can perform better on the intermediate layers.

4- Experiments

To validate the proposed method, the standard PascalVoc 2012 dataset was used, which includes 1,464 labeled images for training, 1,449 for validation, and 1,456 for testing. This dataset is widely used for the semantic segmentation task, and the mean Intersection over Union (mIoU) metric over the validation set is typically used for reporting results. The dataset contains 21 classes, including the background class, which must be included in computing the mIoU. The teacher network used in this study is the Deeplab-V3+ with ResNet101 backbone, which has 59,344,309 trainable parameters. The student networks are the Deeplab-V3+ with ResNet18 backbone, with 16,608,181 trainable parameters, and MobileNet-V2, with 5,816,053 trainable parameters. All of the weights defined in Eq. (3) and Eq. (9) were fine-tuned by testing values of 100, 10, 1, and 0.1, and selecting the best one. Based on this, the best choices for λ, β, and α were found to be 1, 1, and 100, respectively. All of the models were trained with a similar configuration, with a batch size of 6, a total of 120 epochs, and a starting learning rate of 0.007. Each training image was preprocessed by random scaling to 0.5 to 2 times of their original size, horizontal random flip, and finally, a random crop of 513×513 pixels. For validation, each image was resized to 513×513 pixels. By using this dataset and training configuration, the performance of the proposed approach can be evaluated and compared to other methods.

In the experiments presented in this study, the standard PascalVoc dataset was not augmented, unlike some other papers that have employed this technique. The teacher and student networks used ImageNet pre-trained weights in their backbones, and their segmentation parts were randomly initialized. The experiments were performed on two different layers: the middle layer, which refers to the last layer of the decoder, and the end layer, which refers to the last convolutional layer of the segmentation network. Table 1 presents comprehensive comparisons that validate the effectiveness of each distillation approach. The results for two different backbones with different sizes demonstrate that the proposed method is architecture-independent and can be applied to any encoder/decoder-based segmentation network. Each distillation module contributes to a higher mIoU score, indicating that two distillation modules improve the training of the student network. Table 2 shows the results of different approaches explained in the previous sections. It can be observed that distillation can improve the performance of the student network, and the proposed distillation method outperforms the methods of [37] and [34] without adding too much computational burden. In Fig. 2, some examples of the output of the teacher, student, and student with distillation are presented to demonstrate the effect of the proposed distillation method. As discussed earlier, pixel-wise distillation works better on the last layer, as it is closer to probability scores and a better candidate for pixel-wise distillation. For pair-wise distillation, the intermediate layer has better performance than the last layer. The results in Table 3 validate these claims. The training time of each method in Table 3 also shows the simplicity of both distillation methods. The training time for the last layers is higher than that of the intermediate layers, as the feature map size of the intermediate layers (output of the

Table 2. The average and standard deviation of mIoU metric of three runs with different random seeds and their training time for different training methods on the validation set of PascalVoc 2012.

Network	Avg. of mIoU	Std. of mIoU	Time(msecond)
Teacher	74.48	0.44	760
No Distillation	66.59	0.29	250
GAT [37]	67.11	0.28	450
Method of [34]	68.53	0.25	560
Proposed Method	69.21	0.63	740

Table 3. Comparison of the results and training time of each distillation method with different positions of feature maps. The middle and end refer to the last convolution layer of the decoder and the last convolution layer before the probability scores of the deepplab-v3+, respectively. The results are average of three runs on the PascalVoc 2012 validation set.

Distillation Method	Avg. of mIoU	Time(msecond)
Pixel-wise(MIDDLE)	68.46	700
Pixel-wise(END)	69.04	720
Pair-wise(MIDDLE)	68.47	680
Pair-wise(END)	68.54	760

encoder) is smaller than that of the last layers (output of the decoder). Therefore, using the GAM matrix for the last layers will reduce the training time.

In conclusion, the extensive experiments conducted in this study show that using pixel-wise distillation on the last layers and pair-wise distillation on the intermediate layers leads to a good balance between accuracy and training time. The proposed distillation approach can improve the performance of the student network and is applicable to a wide range of segmentation networks.

5- Conclusion and Future Work

In this work, two methods for distilling knowledge from a cumbersome network to a compact model were introduced by considering the pixel-wise and pair-wise similarity between the two networks. Experiments showed that it can successfully boost the student network's performance. Higher levels of deep networks contain more abstract information. In an extreme example, the normalized prediction layer is trained to have pure information about the structure of the problem and forget as much as possible about the details of instances of the objects. Even two identical network architectures might find two different local optimums in their training stages, and the chance of having distant representations for each input decreases as the depth of the layer of representation increases. This fact has attracted researchers to invent methods that can distill information from deeper and near last feature maps of two networks. The proposed method solved this problem by taking the intermediate feature maps and transforming them into similarity matrices and using the last layers to create meaningful representations that wash out restrictive details for distillation and hold helpful information that can guide the student in the optimization space. In the future, the community may want to pay more attention to the use of information in channels to create more meaningful feature maps to invent more novel distillation functions. Several works exploiting channel-wise information have been proposed but suffer

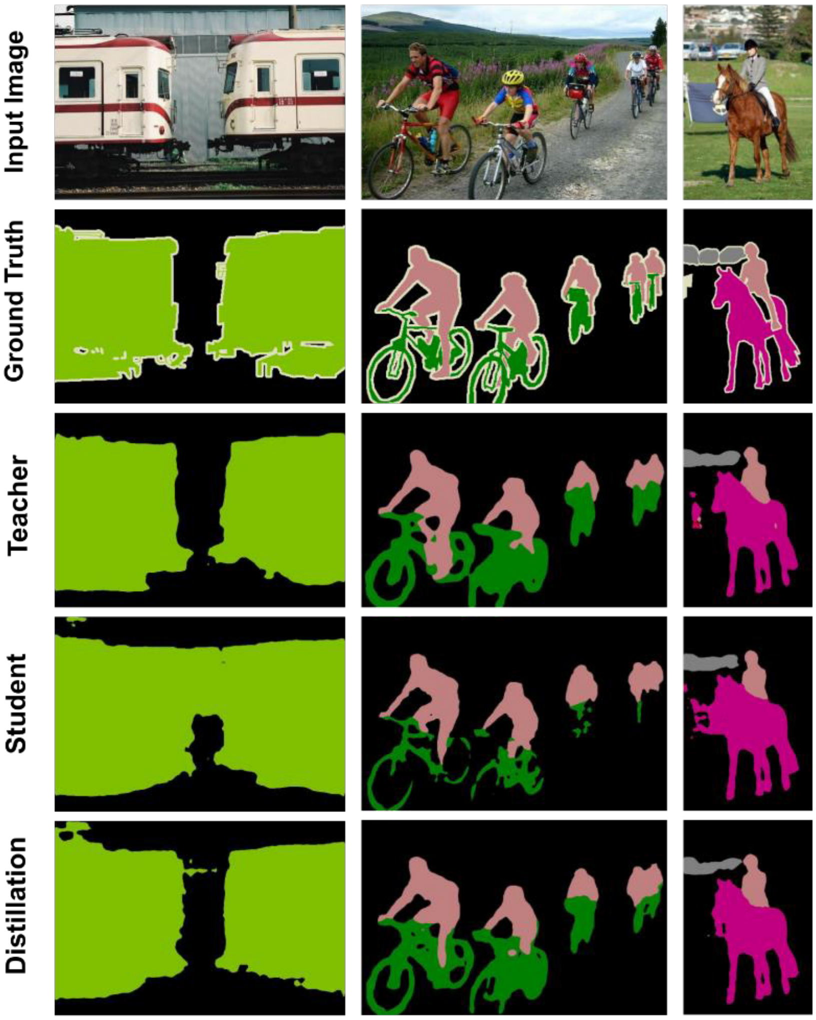

Fig. 2. Comparison of segmentation results between ground-truth, teacher prediction, student prediction, and prediction after distillation.

from expensive computation for distilling channel-wise information. In this work, simple and efficient methods were employed, but exploiting the information in the feature maps channels may have good potential for knowledge distillation.

References

[1] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587, (2017).

[2] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in, 2017, pp. 2881-2890.

[3] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv preprint arXiv:1606.02147, (2016).

[4] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, H. Hajishirzi, Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation, in, 2018, pp. 552-568.

[5] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, Icnet for real-time semantic segmentation on high-resolution images, in, 2018, pp. 405-420.

[6] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, N. Sang, Bisenet

v2: Bilateral network with guided aggregation for real-time semantic segmentation, International Journal of Computer Vision, 129 (2021) 3051-3068 %@ 0920-5691.

[7] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in, 2018, pp. 325-341.

[8] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531, (2015).

[9] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, arXiv preprint arXiv:1412.6550, (2014).

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in, 2016, pp. 2921-2929.

[11] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in, 2017, pp. 618-626.

[12] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision, 88 (2010) 303-338 %@ 0920-5691.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in, 2018, pp. 801-818.

[14] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in, 2015, pp. 3431-3440.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in, 2016, pp. 770-778.

[16] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, A. Anandkumar, Born again neural networks, in, PMLR, 2018, pp. 1607-1616 %@ 2640-3498.

[17] Z. Zhou, C. Zhuge, X. Guan, W. Liu, Channel distillation: Channel-wise attention for knowledge distillation, arXiv preprint arXiv:2006.01683, (2020).

[18] P. Chen, S. Liu, H. Zhao, J. Jia, Distilling knowledge via knowledge review, in, 2021, pp. 5008-5017.

[19] I. Sarridis, C. Koutlis, S. Papadopoulos, I. Kompatsiaris, InDistill: Transferring Knowledge From Pruned Intermediate Layers, arXiv preprint arXiv:2205.10003, (2022).

[20] R. Liu, K. Yang, H. Liu, J. Zhang, K. Peng, R. Stiefelhagen, Transformer-based knowledge distillation for efficient semantic segmentation of road-driving scenes, arXiv preprint arXiv:2202.13393, (2022).

[21] Z. Li, D. Hoiem, Learning without forgetting, IEEE transactions on pattern analysis and machine intelligence, 40(12) (2017) 2935-2947 %@ 0162-8828.

[22] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in, 2019, pp. 3967-3976.

[23] K. Yue, J. Deng, F. Zhou, Matching guided distillation, in, Springer, 2020, pp. 312-328 %@ 3030585549.

[24] S. Tang, Z. Zhang, Z. Cheng, J. Lu, Y. Xu, Y. Niu, F. He, Distilling Object Detectors with Global Knowledge, in, Springer, 2022, pp. 422-438.

[25] C. Yang, M. Ochal, A. Storkey, E.J. Crowley, Prediction-guided distillation for dense object detection, in, Springer, 2022, pp. 123-138.

[26] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, C. Chen, Knowledge distillation with the reused teacher classifier, in, 2022, pp. 11933-11942.

[27] H.-J. Ye, S. Lu, D.-C. Zhan, Generalized knowledge distillation via relationship matching, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2) (2022) 1817-1834 %@ 0162-8828.

[28] G. Ros, S. Stent, P.F. Alcantarilla, T. Watanabe, Training constrained deconvolutional networks for road scene semantic segmentation, arXiv preprint arXiv:1604.01545, (2016).

[29] L. Liu, Q. Huang, S. Lin, H. Xie, B. Wang, X. Chang, X. Liang, Exploring inter-channel correlation for diversity-preserved knowledge distillation, in, 2021, pp. 8271-8280.

[30] Y. Wang, W. Zhou, T. Jiang, X. Bai, Y. Xu, Intra-class feature variation distillation for semantic segmentation, in, Springer, 2020, pp. 346-362 %@ 3030585700.

[31] Y. Feng, X. Sun, W. Diao, J. Li, X. Gao, Double similarity distillation for semantic image segmentation, IEEE Transactions on Image Processing, 30 (2021) 5363-5376 %@ 1057-7149.

[32] J. Xie, B. Shuai, J.-F. Hu, J. Lin, W.-S. Zheng, Improving fast segmentation with teacher-student learning, arXiv preprint arXiv:1810.08476, (2018).

[33] S. Park, Y.S. Heo, Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy, Sensors, 20(16) (2020) 4616 %@ 1424-8220.

[34] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, J.Y. Choi, A comprehensive overhaul of feature distillation, in, 2019, pp. 1921-1930.

[35] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, Structured knowledge distillation for semantic segmentation, in, 2019, pp. 2604-2613.

[36] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, Y. Yan, Knowledge adaptation for efficient semantic segmentation, in, 2019, pp. 578-587.

[37] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, arXiv preprint arXiv:1612.03928, (2016).

HOW TO CITE THIS ARTICLE

A. M. Mansourian, N. Karimi Bavandpour, Sh. Kasaei, An Efficient Knowledge Distillation Architecture for Real-time Semantic Segmentation, AUT J. Model. Simul., 55(1) (2023) 99-108.

DOI: 10.22060/miscj.2023.21949.5307