An Efficient Knowledge Distillation Architecture for Real-time Semantic Segmentation

Amir M. Mansourian
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
amir.mansurian@sharif.edu

Nader Karimi Bavandpour
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
nkarimi@ce.sharif.edu

Shohreh Kasaei, IEEE, Senior Member
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
kasaei@sharif.edu

***Abstract*—Recently, Convolutional Neural Networks have made significant progress in segmentation. When it comes to semantic segmentation, accuracy and efficiency are equally as crucial. Although these deep networks have achieved high accuracy, they suffer from low inference speed, which makes them impractical for use in real-time settings. In this paper, a simple yet efficient knowledge distillation approach is investigated as a means of transferring knowledge from the feature maps of the cumbersome model (teacher) to guide the compact model (student) learning. This is in contrast to some existing computationally expensive methods in training time. In order to address this issue, we resort to the straightforward approach known as pixel-wise distillation to distill the feature maps of the last Convolution layer of the teacher model to the student model. Furthermore, pair-wise distillation is utilized to distill pair-wise similarities of the intermediate layers. To validate the effectiveness of the proposed method, extensive experiments have been conducted on the PascalVoc 2012 dataset using a state-of-the-art DeepLabV3+ segmentation network with different backbone architectures. Experiments show that the proposed method has balanced mIoU and training time well.**

***Index Terms*—Convolutional Neural Networks, Semantic Segmentation, Knowledge Distillation**

I. Introduction

A pixel-wise classification problem is referred to as semantic segmentation. This task aims to assign a specified class (or label) to each pixel in an image. It is a fundamental topic in the field of computer vision, and it can be applied to a wide variety of applications in the real world, including virtual reality, autonomous driving, and video surveillance. Various semantic segmentation approaches have recently emerged based on deep neural networks that have shown superior performance. The succeeding methods, such as DeepLab [1] and PSPNet [2], have achieved a considerable improvement in the accuracy of segmentation, albeit they frequently include cumbersome models and costly computation.

Some real-time architectures for semantic segmentation have been proposed to tackle this problem, e.g., ENet [3], ESPNet [4], ICNet [5] and BiSeNet [6] [7]. On the other hand, several strategies have been proposed to reduce the model's size and improve its cost-effectiveness, including model pruning, model quantization, and knowledge distillation. Among these, knowledge distillation is currently being researched in depth. Knowledge distillation refers to the method that helps the training process of a smaller network (student) under the supervision of a more extensive network (teacher) which was first proposed in [8]. Unlike other compression methods, it can reduce the size of a network without consideration of the structural differences that exist between a teacher network and a student network.

For the purpose of training compact semantic segmentation networks, the knowledge distillation technique is investigated, which has already been tested and shown to be effective in classification tasks [8] [9]. Like most existing methods, in this work, the semantic segmentation problem is approached by viewing it as a collection of distinct pixel classification problems. Then the knowledge distillation is applied to the pixel level. Different from the classification task, semantic segmentation has a structured output. Long-range dependencies are significant for semantic segmentation, and the teacher and student models typically capture different long-range contextual information due to the differences in the receptive fields.

In this work, we also present structured knowledge distillation and transfer the structure information with pair-wise distillation using intermediate feature maps, which are proven to contain rich information [10] [11]. This pair-wise distillation, along with the pixel-wise distillation, provides rich information for the student network from the teacher network. To this end, 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:

- 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.

II. Related Work

In the following, we review works of literature that are relevant to this work, including state-of-the-art researches on

semantic segmentation and knowledge distillation.

Semantic Segmentation: Semantic segmentation is known as a challenging task, which is about to combine global information with detailed local information to predict the structure of an input image in terms of classifying pixels to categories. Semantic segmentation networks are generally larger than classification ones, as they have to extract additional information beside the information needed for classification. The fully convolutional framework, first introduced in [14], added several important improvements to segmentation network design. It can use pretrained weights of classification networks, perform on variable input size, and be trained end-to-end. DeepLabV3+ [13] and PSPNet [2] are two of the most powerful and popular existing segmentation networks. Due to their flexible design, one can choose big and powerful or small and efficient classifier networks as their backbones. They use Atrous convolution and pyramid spatial pooling to capture global context while preserving feature maps' resolution and details. In this work, DeepLab with ResNet101 [15] backbone is adopted as the teacher and DeepLab with ResNet18 and MobileNet backbones are chosen as the student networks.

In addition to cumbersome networks for highly accurate segmentation, real-time segmentation networks have been attracting increasingly more interest due to the need for real applications, such as mobile applications. This is because highly efficient segmentation networks can segment data in a fraction of the time as cumbersome networks. Most works concentrate on creating lightweight networks by speeding up the convolution operations using factorization methods. ENet [6], inspired by [1], incorporates multiple acceleration factors, such as multi-branch modules, early feature map resolution down-sampling, minimal decoder size, filter tensor factorization, etc. ESPNet [4] replaced conventional convolution layers with a spatial pyramid of dilated convolutions. ICNet [5] utilized cascading multi-resolution branches to increase efficiency. BiSeNet [6] uses two branches, one for learning spatial information and the other for obtaining a large receptive field: spatial and context paths.

Knowledge Distillation: The idea of knowledge distillation first appeared in [8], where the student network uses the teacher's predictions as soft labels (compared to zero and one hard label of ground-truth). Soft labels hold useful information about the structure of a problem and relationships between the categories and provide useful information for training the student. The teacher and student framework is widely used for helping to train compact students. There are also numerous other scenarios where it comes useful. For instance, here is a list of some of the recent related works:

- [16] trains a sequence of identical networks in such a way that each network distills from the previously trained one and this leads to 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.

Most of the discussed methods are designed for image classification, but [21] applied its method for object detection as well. Other examples of successful work on object detection and classification include [22] [23] [24] [25] [26] [27]. After classification and detection, one of the first applications of distillation to semantic segmentation was introduced in [28]. They used the prediction of the teacher instead of the ground-truth for training the student. It led to better results because the teacher's output is an easier distribution 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. [32] introduced the consistency loss between the student and the teacher to make their segmentation boundary similar. It also takes the L2 norm of the difference between the student's output probabilities and the teacher's as another 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 to distill meaningful information between the teacher and student using marginal ReLU. [35] introduced two novel distillation losses in segmentation. Pairwise loss is defined as the mean square distance between elements of affinity matrices of the teacher and student networks (affinity matrix contains inner products between every pair of features which encode pixels). The second loss is called holistic distillation, which uses adversarial learning to make feature maps of student similar to its teacher's, using a discriminator convolutional network. [36] is another relevant work that was developed parallel with [35]. It uses an affinity loss which is almost the same as the pairwise loss in [35], except that they train an auto-encoder for the last convolutional layer of their teacher network before computing its affinity matrix. They also use the direct L2 norm distance between the student's last convolutional features and the teacher's encoded features as an additional loss.

In this paper, we use the idea of pre-activation pixel-wise distillation introduced in [34] to distill the knowledge of the last convolution layer of the teacher to the student. Also, for transferring long-range information, a pair-wise distillation method is utilized, similar to [35] on the intermediate layers of teacher and student.

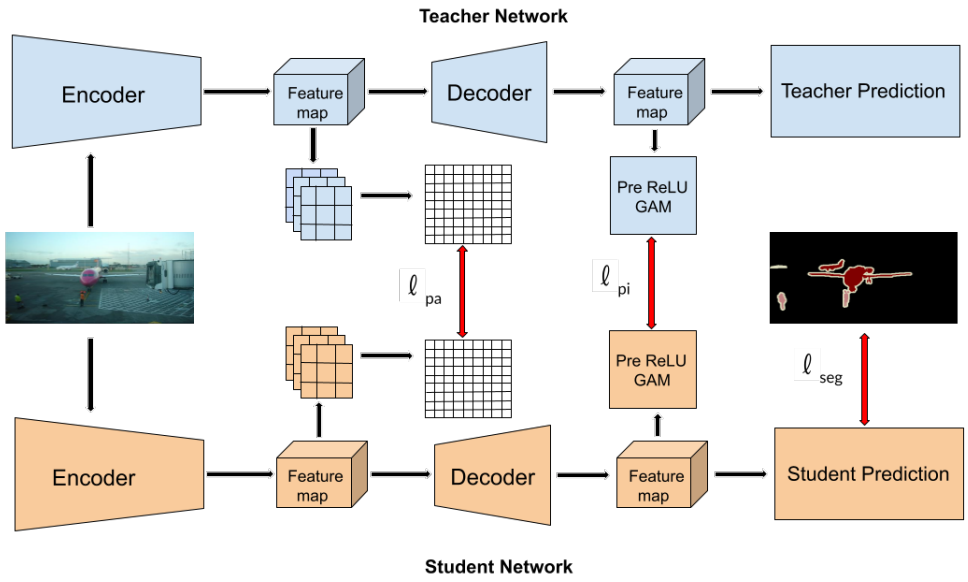

Fig. 1. Our proposed framework for knowledge distillation. The architecture of both the teacher and student networks is Deeplab-V3 + [13], although their encoders are different. Student network encoder depth is shallower than that of teacher network encoder depth. Teacher network is fixed the during 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 preReLU 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.

III. PROPOSED METHOD

In this section, the proposed distillation method between two segmentation CNNs is explained. As it is mentioned in the previous sections, [9] introduced a method to distill the knowledge between two feature maps using a distance which is applied directly to all of their elements. [37] showed that using an attention map which is created by taking summation of feature map channels in a specific layer with uniform weights can improve the performance boost with respect to the method of [8]. The method of [37] is called Global Attention Map(GAM) distillation in this paper, because each feature map is mapped to a single attention matrix. In this section, the idea of transferring attention maps is investigated further by distilling pre-activation attention maps using the idea of [34], and adding pair-wise loss similar to [35]. This will provides the student network with rich knowledge from the teacher network to mimic. The high-level architecture of using the proposed method and standard cross-entropy loss for semantic segmentation is depicted in Fig. 1. In the remainder of this section, first, a mathematical notation is presented, and then the proposed method to create feature maps and affinity matrixes and a loss function to distill them between student and teacher networks is formally introduced. Suppose $A \in R^{c\times w\times h}$ is an intermediate feature map from a segmentation network with spatial dimensions $h \times w$ and number of channels c. The notation $A^k(x)$ is used to show the element at depth k and spatial dimensions are indexed by the vector x, and different feature maps are indexed as A_i. The element-wise power operator on matrix A is denoted by $|A|^p$. The GAM attention matrix for the feature map at layer i is defined as [37]

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

where $p = 2$ in this paper's experiments. If the G_i computed from the teacher network is denoted by G_i^t, and G_i computed from the student network by G_i^s, then the GAT distillation loss function can be written as [37]

$$\ell_{GAT_i} = \| \frac{G_i^t}{\| G_i^t \|_2} - \frac{G_i^s}{\| G_i^s \|_2} \|_2 \tag{2}$$

which is then used in combination with segmentation loss with a weighted sum

$$\ell_{total} = \ell_{seg} + \sum_i \lambda_i \ell_{GAT_i}. \tag{3}$$

Here, the segmentation loss ℓ_{seg} is the widely used cross entropy function between the student network's normalized predictions and ground-truth labels. The loss in (2) was originally defined for image classification, but it can be readily used in semantic segmentation, as well. The distillation point in [9] is the end of an arbitrarily chosen intermediate layer, which has been demonstrated to have poor performance. ReLU allows the beneficial information (positive) to pass through and filters out the adverse information (negative). Therefore, knowledge distillation must be designed under the acknowledgment of this information dissolution. Similar to [34], the pre-activation position to distill knowledge is used because positive and negative values are preserved in the pre-ReLU position without deformation. A proper distance function is needed based on

the distillation point in the pre-ReLU position. In the teacher's feature, the positive responses are utilized for the network, which necessitates that the positive responses be transferred with their exact values. If the student response is higher than the target value, it should be decreased for a negative teacher response. However, it does not need to be increased if the student response is lower than the target value since negatives are blocked by ReLU regardless of their values. For an arbitrary feature map of the teacher and student, $T, S \in R^{c\times w\times h}$, let the i-th component of the tensor be $T_i, S_i \in R$. Partial L2 distance is defined as [34]

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

Then our pixel-wise loss between teacher and student is defined

$$\ell_{pi} = d_p(L^t, L^s). \quad (5)$$

Where L^t and L^s are GAM matrixes of the last convolution layer of the teacher and student, respectively. These matrixes are created based on equation (1) with $p = 1$, to preserve negative values. Although 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, We make use of a loss 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

$$f_i = Flatten(MaxPool(M^i)). \quad (6)$$

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

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

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

$$\ell_{pa}(E^t, E^s) = \frac{1}{(w\times h)^2} \sum_{i}^{w}\sum_{j}^{h} (e_{ij}^t - e_{ij}^s)^2. \quad (8)$$

Where E^t and E^s are the similarity matrix of teacher and student, respectively. The overall loss function of our method then is a weighted sum of ℓ_{seg}, ℓ_{pi}, and ℓ_{pa}, defined by

$$\ell_{total} = \ell_{seg} + \alpha\ell_{pi} + \beta\ell_{pa}. \quad (9)$$

Note that any possible difference of spatial dimensions between attention maps of teacher and student networks is compensated by a simple operation of bilinear upsampling. 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.

IV. EXPERIMENTAL RESULTS

The standard Pascal Voc 2012 dataset is used to validate the proposed method. It contains 1,464 labeled images for training, 1,449 for validation, and 1,456 for test. This dataset is widely used for the semantic segmentation task and measuring the mean Intersection over Union (mIoU) metric over the validation set is usually adopted for reporting the results. There are 21 classes present in this dataset, including background class, which must be included in computing the mIoU. The teacher networks is the Deeplab-V3+ with ResNet101 backbone which has 59,344,309 trainable parameters and the student networks are the Deeplab-V3+ with ResNet18 backbone with 16,608,181 and MobileNet-V2 with 5,816,053 trainable parameters. All of the weights defined as in loss functions (3), and (9) has been fine-tuned by trying values 100, 10, 1, and 0.1 and choosing the best one. Based on this, the best choice for λ, β, and α were 1, 1, and 10, respectively. All of the models are trained with a similar configuration of batch size of 6, total epochs of 120, and a starting learning rate of 0.007. Each training image is preprocessed by the operations of random scaling to 0.5 to 2 times of their original size, horizontal random flip, and finally a random crop of 513×513 and for validation, also, each image is resized to 513×513 pixels. In the experiments of this work, no augmentation is added to the standard Pascal Voc dataset (as some of other papers). The teacher and student networks use the ImageNet pretrained weights in their backbones and their segmentation parts are randomly initialized. The experiments in this section are performed on two different layers, and in the names of the methods the middle layer refers to the last layer of the decoder, and the end layer refers to the last convolutional layer of a segmentation network. In table I, comprehensive comparisons have been presented to validate the effectiveness of each distillation method. Results for two different backbones with different sizes show that the proposed method is architecture-independent and can be applied to each encoder/decoder-based segmentation network. On the other hand, one can see that each distillation module leads to a higher mIoU score. This implies that our two distillation modules contribute to better training of the student network. Table II shows the results of different approaches explained in the last sections. From Table II, can be seen that distillation can improve the performance of the student network, and the proposed distillation method performs better than methods of [37] and [34] without adding too much computational burden compared to the methods that try to generate more precise feature maps by exploiting channel information. In Fig.2 some examples of the output of the teacher, student, and student with distillation are used to show the effect of the proposed distillation method. As

TABLE I
EFFECTIVENESS OF THE PROPOSED DISTILLATION METHOD ON TWO STUDENT NETWORKS: MOBILENETV2 AND RESNET-18 WITH/WITHOUT PIXEL-WISE AND PAIR-WISE DISTILLATION MODULES. RESULTS ARE AVERAGE OF 3 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

mentioned earlier, pixel-wise distillation works better on the last layer because it is closer to probability scores and may be a better candidate for pixel-wise distillation. For pair-wise distillation, the intermediate layer has a better performance than the last layer. The results in table III validate these claims. The training time of each method in table III also shows the simplicity of both distillation methods. As can be seen, the training time for the last layers is more than intermediate layers. This is because the feature map size of the intermediate layers(output of the encoder) is less than the last layers(output of the decoder). For this reason, using the GAM matrix for the last layers will reduce the training time. In the end, extensive experiments show that using pixel-wise distillation on the last layers and pair-wise distillation on the intermediate layers will lead to a good balance between accuracy and training time.

TABLE II
AVERAGE AND STANDARD DEVIATION OF MIOU METRIC OF 3 RUNS WITH DIFFERENT RANDOM SEEDS AND THEIR TRAINING TIME FOR DIFFERENT TRAINING METHODS ON THE VALIDATION SET OF PASCAL VOC 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 III
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. RESULTS ARE AVERAGE OF 3 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

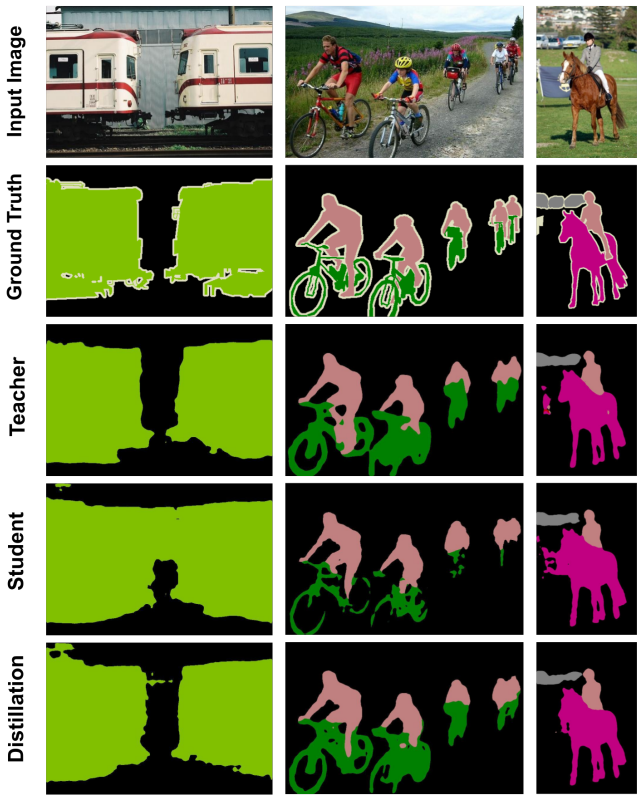

Fig. 2. Comparison of segmentation results between ground-truth, teacher prediction, student prediction and prediction after distillation.

V. CONCLUSION AND FUTURE WORK

In this work, two methods for distilling knowledge from a cumbersome network to a compact model was 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 a 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 decrease as the depth of layer of representation increase. 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 matrixes 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 use the 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 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, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

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

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

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

[5] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 405–420.

[6] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[7] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3051–3068, 2021.

[8] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

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

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

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

[16] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1607–1616.

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

[18] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5008–5017.

[19] I. Sarridis, C. Koutlis, S. Papadopoulos, and 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, and R. Stiefelhagen, "Transformer-based knowledge distillation for efficient semantic segmentation of road-driving scenes," *arXiv preprint arXiv:2202.13393*, 2022.

[21] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[22] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[23] K. Yue, J. Deng, and F. Zhou, "Matching guided distillation," in *European Conference on Computer Vision*. Springer, 2020, pp. 312–328.

[24] S. Tang, Z. Zhang, Z. Cheng, J. Lu, Y. Xu, Y. Niu, and F. He, "Distilling object detectors with global knowledge," *arXiv preprint arXiv:2210.09022*, 2022.

[25] C. Yang, M. Ochal, A. Storkey, and E. J. Crowley, "Prediction-guided distillation for dense object detection," *arXiv preprint arXiv:2203.05469*, 2022.

[26] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen, "Knowledge distillation with the reused teacher classifier," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 933–11 942.

[27] H.-J. Ye, S. Lu, and D.-C. Zhan, "Generalized knowledge distillation via relationship matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[28] G. Ros, S. Stent, P. F. Alcantarilla, and 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, and X. Liang, "Exploring inter-channel correlation for diversity-preserved knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8271–8280.

[30] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–362.

[31] Y. Feng, X. Sun, W. Diao, J. Li, and X. Gao, "Double similarity distillation for semantic image segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 5363–5376, 2021.

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

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

[34] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1921–1930.

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

[36] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 578–587.

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