AICSD: Adaptive Inter-Class Similarity Distillation for Semantic Segmentation

Amir M. Mansourian, Rozhan Ahmadi, Shohreh Kasaei

***Abstract*—In recent years, deep neural networks have achieved remarkable accuracy in computer vision tasks. With inference time being a crucial factor, particularly in dense prediction tasks such as semantic segmentation, knowledge distillation has emerged as a successful technique for improving the accuracy of lightweight student networks. The existing methods often neglect the information in channels and among different classes. To overcome these limitations, this paper proposes a novel method called Inter-Class Similarity Distillation (ICSD) for the purpose of knowledge distillation. The proposed method transfers high-order relations from the teacher network to the student network by independently computing intra-class distributions for each class from network outputs. This is followed by calculating inter-class similarity matrices for distillation using KL divergence between distributions of each pair of classes. To further improve the effectiveness of the proposed method, an Adaptive Loss Weighting (ALW) training strategy is proposed. Unlike existing methods, the ALW strategy gradually reduces the influence of the teacher network towards the end of training process to account for errors in teacher's predictions. Extensive experiments conducted on two well-known datasets for semantic segmentation, Cityscapes and Pascal VOC 2012, validate the effectiveness of the proposed method in terms of mIoU and pixel accuracy. The proposed method outperforms most of existing knowledge distillation methods as demonstrated by both quantitative and qualitative evaluations. Code is available at:** https://github.com/AmirMansurian/AICSD

***Index Terms*—Deep Neural Networks, Semantic Segmentation, Knowledge Distillation, Inter-class Similarity, Intra-class Distribution, Adaptive Loss Weighting.**

I. Introduction

Semantic Segmentation, as an essential element for understanding visual scenes, is a fundamental and challenging task in computer vision. It is a member in the group of dense prediction tasks with the objective of generating a labeling map, in which a particular class label is assigned to each pixel of the input image. Semantic segmentation has found numerous real-world applications in many fields; such as autonomous driving, video surveillance, scene and human-body parsing and many other areas.

Pioneered by the pivotal work of Fully Convolutional Network (FCN) [1], deep neural networks have significantly advanced the field of semantic segmentation. Since then, numerous FCN-based methods have been introduced to enhance segmentation performance by generating accurate segmentation maps. Those methods have improved segmentation accuracy through various approaches; such as employing stronger backbone networks [2], deeper network architectures with higher capacity compared to FCNs [3], incorporating multi-scale image contexts [4], and refining segmentation details [5]. These methods exhibit significant effectiveness in enhancing the performance and accuracy of semantic segmentation. However, the efficacy comes at the cost of efficiency; since the complexity in their model design requires substantial hardware resources, including significant computational power and memory requirements. As a result, their application in real-world is constrained, particularly on resource-limited devices such as mobile and other edge devices. In such real-world scenarios, the demand for lightweight and resource-efficient models becomes essential. Given the mentioned concerns, lightweight neural networks with small model size and light computation cost have received significant attention. Quantization [6]–[8], pruning [9], and decomposition [10] of

The authors are with the Department of Computer Engineering, Sharif University of Technology, Tehran 11155, Iran (e-mail: amir.mansurian@sharif.edu; roz.ahmadi@sharif.edu; kasaei@sharif.edu).

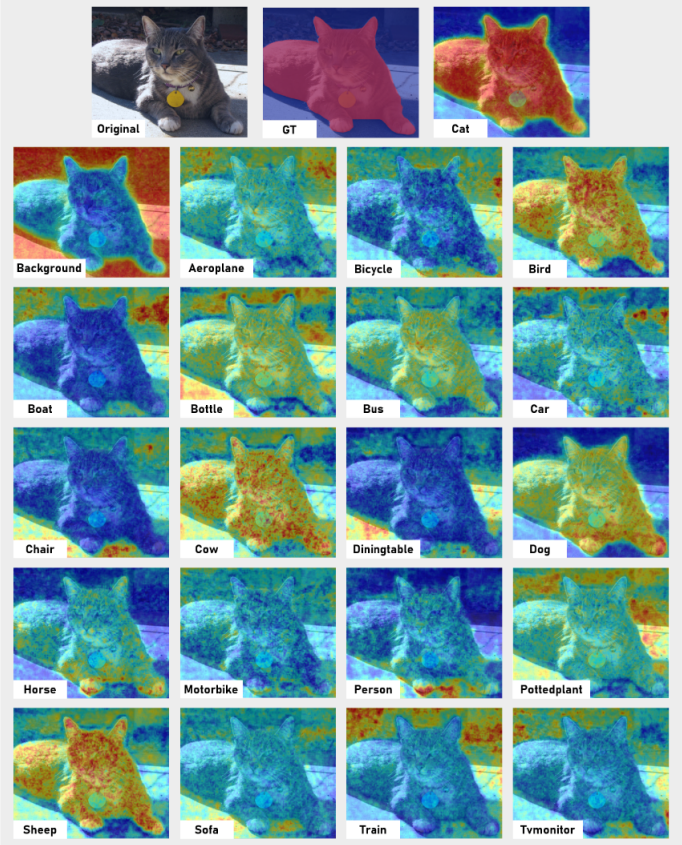

Fig. 1. Intra-class distributions for each class. Distributions are created by applying softmax to spatial dimension of output prediction of last layer. Similarities between each pair of intra-class distributions have good potential for distillation. Distributions are created from the PASCAL VOC 2012 dataset with 21 class categories.

weights, present some of the approaches aiming to achieve lightweight networks. Those methods compress networks by using smaller precision for weights, removing redundant layers from the networks, and replacing large backbones with lighter versions, respectively. However, they have not fully closed the segmentation performance gap between compact networks and more complex ones.

Knowledge Distillation, first introduced by Bucila et al. [11] and popularized by [12], has served as a successful strategy for achieving a better trade-off between performance and efficiency of deep neural networks by using the knowledge of a more complex network (the teacher) to assist the training of a lighter network (the student). Methods based on knowledge distillation have greatly improved the accuracy of lightweight networks, performing tasks; such as image classification [13]–[17], object detection [18]–[20], and face recognition [21]–[23]. The knowledge distilled in the pioneering work of [12] for the task of image classification provided soft labels from a heavy teacher network with more beneficial information (e.g., intra-class similarity and inter-class difference), than the hard labels originally provided to the small network in the form of one-hot class label vectors. The fundamental idea of those methods is that soft labels offer the knowledge that hard ground truth labels are unable to convey. The student network is then supervised to mimic the teacher model using both hard and soft labels.

Knowledge Distillation has also been introduced to the field of semantic segmentation [24]–[26]. Liu et al. [25] viewed segmentation as the problem of classifying each pixel in an input image and, as a solution, applied knowledge distillation on pixel level (known as pixel-wise knowledge distillation). Semantic segmentation, being a dense prediction task, is used to predict dense structured outputs. While pixel-wise knowledge is effective for image classification, it is not enough to improve the performance of semantic segmentation. This is because aligning the pixel-level class distribution between teacher and student networks has the potential to disregard the contextual relationships among pixels and may even lead to performance degradation due to existing noise in activation maps.

Considering these limitations, the adoption of complementary approaches such as pair-wise and holistic knowledge distillation has shown positive efficacy in enhancing the performance of semantic segmentation by capturing spatial structural context and complementary information beyond pixel-level knowledge. However, there are several constraints limiting the performance of these methods. To begin with, many prior works use the spatial relationships between feature maps or channels and the inter-class relations are ignored. Secondly, the negative effects of teacher on student's training procedure are not thoroughly considered. Despite the progress made in defining different pair-wise distillation methods for transferring structured knowledge from teacher to student, existing methods do not consider inter-class similarities, which can be a valuable source of information for distillation. Moreover, as the teacher network itself has a lot of error in its predictions, it would not be reasonable to force the student to mimic the teacher's outputs throughout the training. Most existing works ignore this fact and set a constant scale for their distillation losses, and pixel-wise and pair-wise losses have a fixed impact on the training of the student network, which may lead to negative effects on the training process.

To address the aforementioned challenges, this paper proposes a pair-wise distillation method that leverages intra-class distributions and inter-class similarity matrices. The method defines class-specific intra-class distributions that capture the network's attention for each class and constructs inter-class similarity matrices that highlight similarities between these distributions. Figure 1 illustrates the intra-class distributions for a given image. Additionally, an adaptive loss weighting strategy is proposed that adjusts the scale of losses during student training process to emphasize the positive teacher knowledge and reduce the impact of negative information.

In summary, the main contributions of this work are as follows:

- Proposing a new pair-wise distillation method, called Inter-Class Similarity Distillation (ICSD), to transfer structured information from teacher to student.
- Proposing a training strategy to control the impact of distillation in training phase of the student network by changing the scale of losses in an adaptive manner (ALW).
- Validating the effectiveness of the proposed method with extensive experiments on the Cityscapes and PASCAL VOC 2012 datasets with state-of-the-art DeepLab V3+ segmentation network.

The remaining sections of this paper are structured as follows. Some related work relevant to the proposed method are reviewed in Section II. This is followed by a detailed explanation of the proposed method in Section III. Lastly, extensive experiments and ablation studies are discussed in Section IV.

II. RELATED WORK

In the following, the literature most relevant to this work is reviewed. This includes state-of-the-art research surrounding semantic segmentation and knowledge distillation.

A. Semantic Segmentation

Semantic segmentation is a fundamental and challenging task in the field of computer vision. It allows for a deep, fine-grained understanding and analysis of the visual content. The problem is defined as assigning a particular class label to each pixel of an input image. The methods based on deep convolutional neural networks include pioneering work of FCN [1] followed by U-net [33], SegNet [34], and DeconvNet [35], in which managing to capture spatially structured context has been the key to their success.

Several approaches have been suggested in order to further improve the performance and accuracy. Incorporating boundary information [36] to better define object edges within the scene, extracting contextual information using multi-context aggregation [3] to capture spatial context in various scales from the image, enlarging the receptive field to gather more details regarding the long-rang relationships within pixels

TABLE I
LOSS IMPLEMENTATION IN EXISTING PAIR-WISE KNOWLEDGE DISTILLATION METHODS.

Method	Alignment Order	Distance	Formulation	Information
MD [27]	Pixel Level	L2	$L_2(y, x)$	8-Neighbourhood Consistence Map
SKD [25]	Pixel Level	L2	$\frac{f_i^T f_j}{(\|f_i\|_2 \|f_j\|_2)}$	Pixel Similarity
CSC [28]	Channel Level	L2	$\bigoplus_{c=1}^{C} f_c(i,j) \otimes f_{s_c}(i,j)$	Set of Channel Correlations
ICKD [29]	Channel Level	L2	$G^{F^T} = f(F^T) \cdot f(F^T)^T$	Inter-Channel Correlation Matrix
CSD [30]	Class Level	L2	$CM(q)_{ij} = N(q^{i\cdot}) \cdot N(q^{j\cdot})$	Cross-Category Correlation Matrix
FSP [31]	Feature map Level	L2	$\sum_{s=1}^{h} \sum_{t=1}^{w} \frac{F^1_{s,t,i}(x;W) \times F^2_{s,t,j}(x;W)}{h \times w}$	Cross-Layer FSP Matrix
SP [32]	Instance Level	L2	$\frac{\tilde{G}_T^{(l)}}{\left\|\tilde{G}_T^{(l)}\right\|_2}$	Cross-Instance

through the usage of multi-scale features and dilated convolutions [37]–[39], and attention-based methods [4], [40] to explore the connection between pixels and channels are some of the strategies incorporated to improve the semantic segmentation performance. This, however, comes with the cost of requiring high computational resources. The more complex the model design gets, larger and deeper networks are employed which demand considerable hardware and computational capabilities.

Various strategies have been proposed to make semantic segmentation models suitable for real-world applications, especially mobile applications. Some methods achieve higher efficiency by replacing heavy networks such as Deeplab-V3+ [38] and PSPNet [3] with lighter models where the encoder module is less complex; e.g., Mobilenet-V2 [41] and Resnet18 [10]. Other approaches have proposed lowering the cost of convolutional operations to increase efficiency. Enet [6] uses filter factorization, lighter encoder/decorder, and early downsampling to create an efficient network. ESPNet [8] replaces standard convolutions with a module that is a combination of efficient spatial pyramid of convolutions and point-wise convolutions. ICNet [7] is an image cascade network where features from low and high resolution images are fused to maintain a balance between efficiency and accuracy. BiSeNet [42] achieves this goal by using the combination of a spatial and contextual path to increase feature processing efficiency.

B. Knowledge Distillation

Knowledge Distillation is another common approach used in semantic segmentation to balance the trade-off between accuracy and efficiency. Introduced by Bucila et al. [11] and popularized by [12], knowledge distillation trains a smaller and compact network called the student, using the knowledge distilled from a heavier cumbersome network called the teacher. The fundamental approach is to minimize the KL-divergence between the logits outputed by the teacher and student by getting the student to imitate the actions of the teacher. These logits, considered as soft labels, provide the student network with substantial information from the cumbersome teacher network that the original hard labels (i.e., one-hot class label vectors) are unable to collect.

Other methods have explored the distilling knowledge based on corresponding features across teacher and student networks. FitNet [43] aligns feature maps extracted from intermediate hidden layers. RKD [44] distills distance and angle wise correlations between features. In AT [45], the student is trained to imitate the corresponding intermediate attention map from teacher. [46] calculates weights for each channel of a feature map, which is the importance of each channel, and then distills these weights. Also, decreases the impact of distillation loss with the increase of epochs. [47] employs the Gram Matrix [48] to distill the correlation between feature maps activated and selected considering the distributions of neuron selectivity patterns. These strategies have managed to improve the performance of lightweight networks without causing any increase in their inference load, in many cases even making these student networks considerably faster.

C. Knowledge Distillation for Semantic Segmentation

Knowledge distillation has been employed in creating fast and compact semantic segmentation networks [49]–[53]. SKDS [25] applied knowledge distillation in pixel-level. Its strategy, however successful in other tasks, is limited in improving the performance of semantic segmentation. The focus on individual pixel-wise relations in these methods, fail to capture the structural context needed in dense prediction. Taking these restrictions into account, SKDS has also demonstrated that incorporating complementary strategies (i.e., pair-wise and holistic knowledge distillation alongside pixel-wise knowledge distillation) can be effective in improving the performance of semantic segmentation by taking cross-pixel relationship dependencies into consideration. Holistic knowledge distillation aligns high-order relations between the segmentation maps distilled from the teacher into the student network.

This work is mostly focused on pair-wise knowledge distillation. Based on pair-wise Markov random field framework [54], SKDS proposes to incorporate pair-wise similarity relations among pixels to improve spatial labeling contiguity. Xie et al. [27] have also employed a pixel-level feature map-based pair-wise distillation method by capturing local pixel

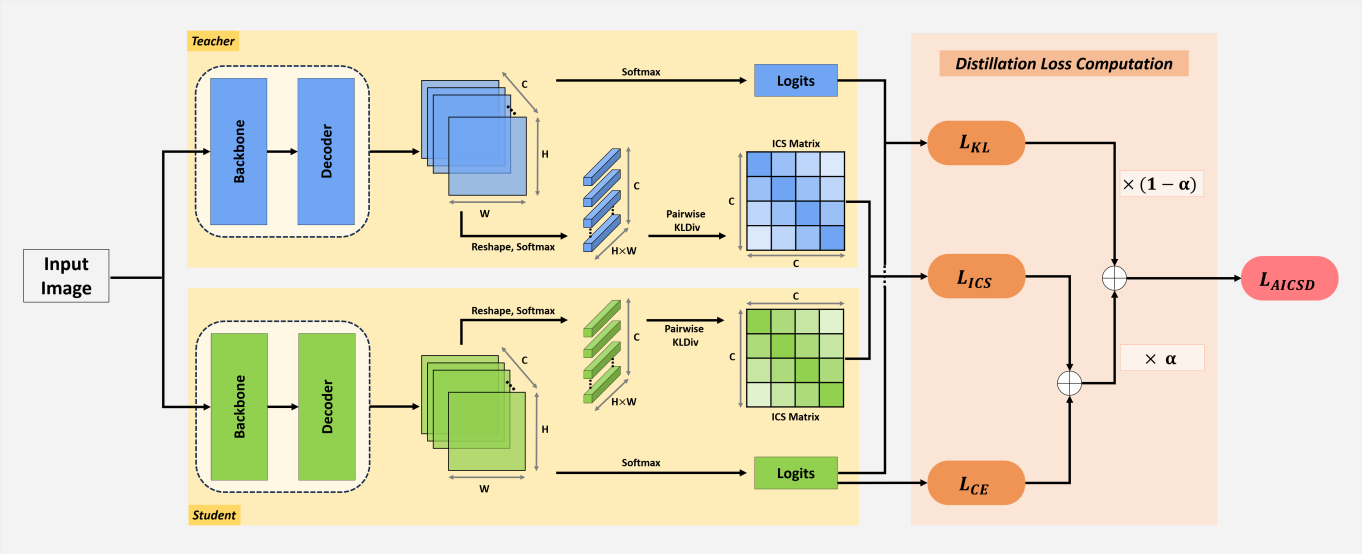

Fig. 2. **Overall diagram of the proposed AICSD**. Network outputs are flattened into 1D vectors, followed by application of a softmax function to create intra-class distributions. KL divergence is then calculated between each distribution to create inter-class similarity matrices. An MSE loss function is then defined between the ICS matrices of the teacher and student. Also, KL divergence is calculated between the logits of the teacher and student for pixel-wise distillation. To mitigate the negative effects of teacher network, an adaptive weighting loss strategy is used to scale two distillation losses and cross-entropy loss of semantic segmentation. During training, hyperparameter α undergoes adaptive changes and progressively increases with epoch number.

probability differences among each pixel and its eight neighbours. Motivated by that, several subsequent proposed studies have achieved notable improvements. Feng et al. [30] distill class-level pair-wise relationships through similarity within category dimensions. Liu et al. [29] capture the similarities and variations across the channels of a feature map. Yim et al. [31] defined the flow between network layers by calculating the inner product between features maps from consecutive pair of levels. Tung et al. [32] consider instance-level pair-wise distillation. This is done by computing the (dis)similarities between samples in a batch and train the student network to preserve the relations in its own representation space. Channel-level pair-wise distillation is explored in [28], [29]. Park et al. [28] propose a channel and spatial correlation (CSC) loss to extract and transfer the full long-range relationship in the feature map. Table I shows a summary of pair-wise losses for knowledge distillation implemented in the mentioned methods.

In this study, a pair-wise distillation method is proposed, similar to [29]. The method involves computing intra-class distributions for each class, followed by the calculation of an inter-class similarity matrix for distillation purposes. Additionally, an adaptive training strategy for the student network using a distillation loss is investigated, drawing inspiration from [46].

III. PROPOSED METHOD

In this section, first, the basic knowledge distillation technique [12] is reviewed. Subsequently, the proposed inter-class similarity distillation method is explained. Finally, the training strategy for adding distillation losses to the Cross Entropy (CE) loss of semantic segmentation is examined. The method diagram of the proposed approach is shown in Figure 2.

A. Preliminary

Let $Z^T, Z^S \in \mathbb{R}^{C\times H\times W}$ denote the outputs of the last layer of the teacher (T) and student (S) networks, respectively. Notation $Z_k(i,j)$ is used to show the element at depth k, and spatial dimensions are indexed by i and j. The first knowledge distillation method, originally proposed by Hinton et al. [12], aimed to use the class probabilities generated by the teacher network as soft labels for training the student network. This was done by minimizing the Kullback-Leibler (KL) divergence between the class probabilities of the teacher and student networks. In the context of semantic segmentation, as a collection of separate classification tasks, the distillation loss can be written as

$$\ell_{kd} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} KL(\sigma(\frac{z^T(i,j)}{\tau})||(\sigma(\frac{z^S(i,j)}{\tau})), \quad (1)$$

where $\sigma(.)$ and τ denote the softmax function and temperature factor, respectively, and $KL(.||.)$ represents the KL divergence between the logits of the T and S networks. In this work, this distillation method is implemented in our own method, called KD method, and it's loss is written as ℓ_{KD}.

B. Inter-Class Similarity Distillation

In this section, the formulations for creating an inter-class similarity matrix from intra-class distributions are discussed in detail. Given Z^T and Z^S as the outputs of the teacher and student networks for a specific image, intra-class distributions are created by applying softmax at the spatial dimension of Z^t and Z^s. This leads to intra-class information for each class, regardless of other classes, and a probability distribution is provided for a specific class. Considering $G \in \mathbb{R}^{C\times HW}$ as intra-class distributions, the formulations can be written as

$$G_i = \sigma(Flatten(Z_i)), \tag{2}$$

where Flatten(.) function vectorizes a 2D output map ($H \times W$) into a 1D vector with length HW. Then, the Inter-Class Similarity (ICS) matrix can be created from the intra-class distributions by calculating the KL divergence between each pair of intra-class distributions as

$$ICS(i,j) = KL(G_i||G_j). \tag{3}$$

Finally, the inter-class similarity loss between teacher and student is defined by

$$\ell_{ICS} = \frac{1}{C^2} \parallel ICS^T - ICS^S \parallel_2^2, \tag{4}$$

where ICS^T and ICS^S denote ICS matrices of teacher and student and C is the number of classes. As last layer outputs of the both teacher and student networks have the same number of channels and spatial dimension, there is no need to change the spatial size or number of channels of student to match the teacher network and the proposed pair-wise distillation method can be applied on any segmentation network.

This distillation method is called Inter-Class Similarity Distillation (ICSD), and its final objective is given by

$$\ell_{total} = \ell_{CE} + \lambda \ell_{ICSD} \tag{5}$$

where ℓ_{CE} denotes the conventional Cross Entropy (CE) loss of semantic segmentation task and λ is a hyperparameter that controls the scale of proposed distillation loss.

C. Adaptive Loss Weighting

Although distillation improves the accuracy of the student, the teacher network itself can still have error in its predictions. [55] investigated the negative impacts of distillation and [46] proposed an early decay strategy to decrease the effect of their channel-wise distillation loss in the last epochs of training phase. Inspired by those methods, an Adaptive Loss Weighting (ALW) process is utilized to mitigate the negative impacts of distillation. In this strategy, the standard distillation method is used to train the student network in the first epochs, as the soft logits provided by the teacher are easier to learn from than zero-one labels. However, towards the end of training phase, the focus shifts to the CE loss and ICSD loss, which transfer the structure of the teacher to the student, allowing the student to take control of the training process. The final loss of the proposed method with the ALW strategy is defined as

$$\ell_{AICSD} = \alpha(\ell_{CE} + \ell_{ICSD}) + (1-\alpha)\ell_{kd}, \tag{6}$$

where the hyperparameter α controls the weighting of losses according to the ALW strategy and can be adjusted either linearly

$$\alpha = \frac{e-1}{N_e}, \tag{7}$$

where e is the epoch number and N_e is the number of all epochs, or adjusted exponentially

$$\alpha = \beta^{e-1}, \tag{8}$$

where β is a hyperparameter.

In fact, this process is, in general, analogous to the human learning system in which a child, in early stages, cannot learn on its own without the guidance of a teacher. As the student network becomes more trained, it can gradually learn on its own, by using labels, with only a structural guidance (pair-wise loss) from the teacher network. This training strategy can be used in any other distillation methods with pixel-wise or pair-wise distillation losses.

IV. EXPERIMENTS

This section begins by introducing the datasets, evaluation metrics, and implementation details. Next, the results of the proposed method are reported and compared with those of some existing distillation methods. Finally, ablation studies are discussed to more validate the proposed method.

A. Datasets

1) Pascal VOC 2012: The Pascal VOC dataset is a widely used computer vision dataset for object recognition and segmentation, containing 1,464 labeled images for training, 1,449 for validation, and 1,456 for testing, with 21 foreground object categories including background class. In this work, an augmented version of the dataset is used that includes extra annotations, as provided by [38].

2) Cityscapes: The Cityscapes dataset was created for the purpose of urban scene understanding and includes 30 object classes, although only 19 of these classes are used for evaluation purposes. The dataset consists of 5,000 high-quality images that have been finely annotated at the pixel level, as well as an additional 20,000 images that have been coarsely annotated. The finely annotated images are divided into three sets: 2,975 for training, 500 for validation, and 1,525 for testing. In this work, only the subset of 5,000 finely annotated images is used.

B. Evaluation Metrics

The segmentation performance is evaluated using two metrics: mean Intersection-over-Union (mIoU) and pixel accuracy. The IoU metric measures the ratio of intersection to union between the predicted results and ground truth for each object category, and the mIoU is the average of the IoUs across all categories. Additionally, the pixel accuracy metric measures the ratio of correctly predicted pixels to all the pixels. To represent the model size, the number of network parameters is reported.

C. Implementation Details

1) Network Architectures: For a fair evaluation, the experiments are conducted using the same teacher and student network as in [29]. Specifically, the teacher network used in all of the experiments is Deeplab V3+ with ResNet101 as the backbone. For the student network, Deeplab V3+ segmentation with different backbones including ResNet18 and MobileNetv2 is used.

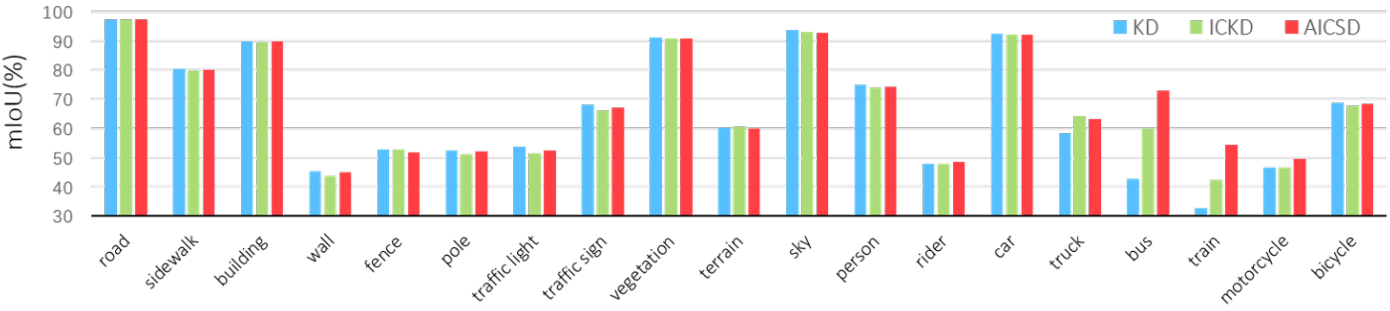

Fig. 3. Comparison of mIoU per class between KD, ICKD, and proposed AICSD on the validation set of Cityscapes dataset. Student network uses a ResNet18 backbone.

TABLE II
PERFORMANCE COMPARISON OF AICSD WITH OTHER DISTILLATION METHODS FOR TWO DIFFERENT BACKBONES ON PASCAL VOC 2012 VALIDATION SET.

Method	mIoU(%)	Params(M)
Teacher: Deeplab-V3 + (ResNet-101)	77.85	59.3
Student1: Deeplab-V3 + (ResNet-18)	67.50	16.6
Student2: Deeplab-V3 + (MobileNet-V2)	63.92	5.9
Student1 + KD	69.13 ± 0.11	16.6
Student1 + AT	68.95 ± 0.26	16.6
Student1 + SP	69.04 ± 0.10	16.6
Student1 + ICKD	69.13 ± 0.17	16.6
Student1 + AICSD (ours)	**70.03 ± 0.13**	16.6
Student2 + KD	66.39 ± 0.21	5.9
Student2 + AT	66.27 ± 0.17	5.9
Student2 + SP	66.32 ± 0.05	5.9
Student2 + ICKD	67.01 ± 0.10	5.9
Student2 + AICSD (ours)	**68.05 ± 0.24**	5.9

TABLE III
PERFORMANCE COMPARISON ON CITYSCAPES VALIDATION SET.

Method	mIoU(%)	Accuracy(%)
T: ResNet101	77.66	84.05
S1: ResNet18	64.09	74.8
S2: MobileNet v2	63.05	73.38
S1 + KD	65.21 (+1.12)	76.32 (+1.74)
S1 + AT	65.29 (+1.20)	76.27 (+1.69)
S1 + SP	65.64 (+1.55)	76.90 (+2.05)
S1 + ICKD	66.98 (+2.89)	77.48 (+2.90)
S1 + AICSD (ours)	**68.46 (+4.37)**	**78.30 (+3.72)**
S2 + KD	64.03 (+0.98)	75.34 (+1.96)
S2 + AT	63.72 (+0.67)	74.79 (+1.41)
S2 + SP	64.22 (+1.17)	75.28 (+1.90)
S2 + ICKD	65.55 (+2.50)	76.48 (+3.10)
S2 + AICSD (ours)	**66.53 (+3.48)**	**76.96 (+3.58)**

2) Training Details: The student networks are trained using a similar configuration, which includes a batch size of 6 and total number of 120 epochs for pascal dataset and batch size of 4 and total number of 50 epochs for the Cityscapes dataset. The stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.007 (pascal), and 0.01 (cityscapes) is used and it is reduced according to the cosine annealing scheduler. Prior to training, each image is preprocessed using random scaling to 0.5 to 2 times of their original size, horizontal random flipping, and a random crop of 513×513 pixels for pascal dataset, and 512×1024 for the Cityscapes dataset. The teacher and student networks use pre-trained weights from the ImageNet dataset for their backbones, while their segmentation parts are initialized randomly. The hyperparameters defined in the equations 5 and 8 were fine-tuned by testing different values and selecting the optimal ones. After experimentation, the values that produced the best results were λ=9500 and β=0.985. For inference, the performance is evaluated on a single scale and original inputs and results are average of three runs. The implementation is done using the PyTorch framework. All networks are trained on a single NVIDIA GeForce RTX 3090 GPU.

D. Experimental Results

To evaluate the performance of the proposed pair-wise method, extensive experiments were conducted and compared to several existing distillation methods including: KD [12], Attention Transfer (AT) [45], Similarity preserving (SP) [32] and Inter-channel Correlation Knowledge Distillation (ICKD) [29].

All of the aforementioned methods were tested on both intermediate layers and final output maps to find the best results. Table II shows that the proposed AICSD method outperforms all of the mentioned methods with different student architectures on the validation set of the Pascal VOC 2012 dataset.

Table III represents a comparison between the proposed method and other distillation methods in terms of mIoU and pixel accuracy on the validation set of the Cityscapes dataset. Although the degree of improvement achieved by the proposed method varies depending on the student network architecture, it consistently outperforms the other methods across different student backbones.

Moreover, Figure 3 compares the per-class mIoU of our proposed AICSD with KD and ICKD methods on the Cityscapes dataset. As shown in this figure, proposed method achieves significant improvements on some classes while maintaining similar performance on other classes. This suggests that the proposed method is able to capture and transfer new information from the teacher to the student that other methods may not capture.

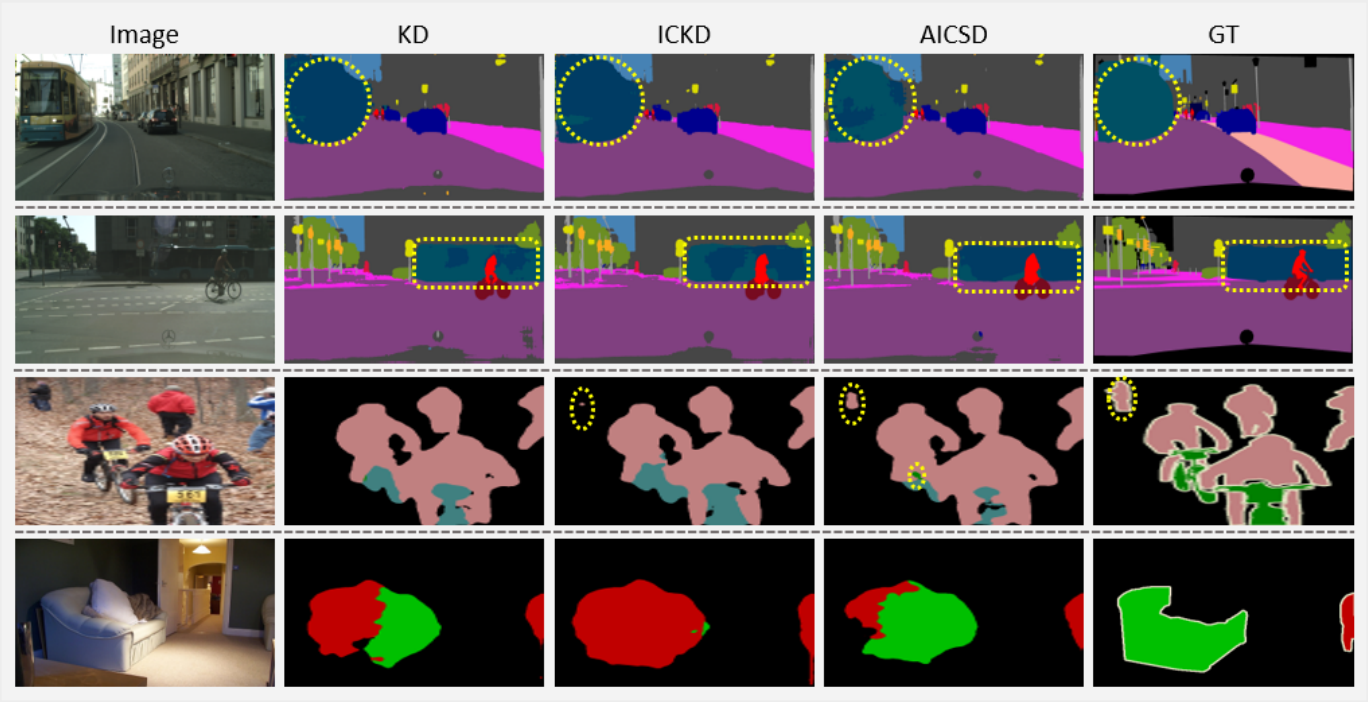

Fig. 4. Qualitative comparison of results on Cityscapes and Pascal VOC validation sets. First two rows show results on Cityscapes and last two rows show results on Pascal VOC. Performance improvement of proposed AICSD method is compared with KD and ICKD methods.

E. Ablation Studies

As introduced in the previous section, the proposed method includes the ICSD loss and the known KD loss, which are used together with the ALW training strategy. Table IV presents an analysis of the impact of each of these methods on the improvement of the student network. The experiments conducted on the Pascal VOC 2012 dataset with two different backbone architectures demonstrate that the ALW strategy results in better performance than separately using the KD or ICSD losses. Moreover, with the MobileNet backbone, it can improve both the KD and ICSD methods by approximately 1.5 %, in terms of mIoU.

Table V compares two different variants of the ALW training strategy, namely linear and exponential loss weighting, and demonstrates that both methods can improve the results of the student network. Additionally, depending on the architecture of the student network, one method may achieve better results than the other.

TABLE IV
ABLATION STUDY ON PASCAL VOC 2012. VALIDATING EFFECTIVENESS OF PROPOSED ICSD AND ALW TRAINING STRATEGY.

Method	KD	ICSD	ALW	*val* mIoU(%)
Teacher:ResNet101				77.85
Student1:ResNet18	n/a	n/a	n/a	67.50
Student1:ResNet18	✓	✗	✗	69.13 (+1.63)
Student1:ResNet18	✗	✓	✗	69.20 (+1.70)
Student1:ResNet18	✓	✗	✓	69.48 (+1.94)
Student1:ResNet18	✓	✓	✓	70.03 (+2.53)
Student2:MbileNetV2	n/a	n/a	n/a	63.92
Student2:MbileNetV2	✓	✗	✗	66.39 (+2.47)
Student2:MbileNetV2	✗	✓	✗	66.58 (+2.66)
Student2:MbileNetV2	✓	✗	✓	67.02 (+3.10)
Student2:MbileNetV2	✓	✓	✓	68.05 (+4.13)

TABLE V
ABLATION FOR PROPOSED ALW TRAINING STRATEGY FOR TWO DIFFERENT APPROACHES, LINEAR AND EXPONENTIAL, ON PASCAL VOC 2012 VALIDATION SET.

Method	Linear ALW	Exponential ALW
S1: ResNet18	69.74 (+2.24)	70.03 (+2.53)
S2: MobileNetV2	68.05 (+4.13)	67.78 (+3.87)

F. Qualitative Assessment

To further validate the effectiveness of the proposed method, some qualitative assessments of the model's performance are conducted. Figure 5 visualizes the intra-class distributions and inter-class similarity matrices for a given image in the Cityscapes and Pascal VOC datasets. The top rows show two images from the Cityscapes dataset, their corresponding pair-wise matrices, and the intra-class distributions of road, rider, bus, and motorcycle classes (from top left to bottom right). The bottom rows show the same for images from the Pascal VOC dataset and the intra-class distributions of sheep, person, dog, and car classes. The figure compares the results of student network without distillation, student network trained with our proposed AICSD method, and the teacher network. The qualitative results show that our proposed method helps the student preserve the structure from the teacher and leads to both intra-class distributions and pair-wise matrices that are more similar to the teacher.

Figure 4, on the other hand, represents output masks for KD, ICKD, and our proposed AICSD method on the Cityscapes

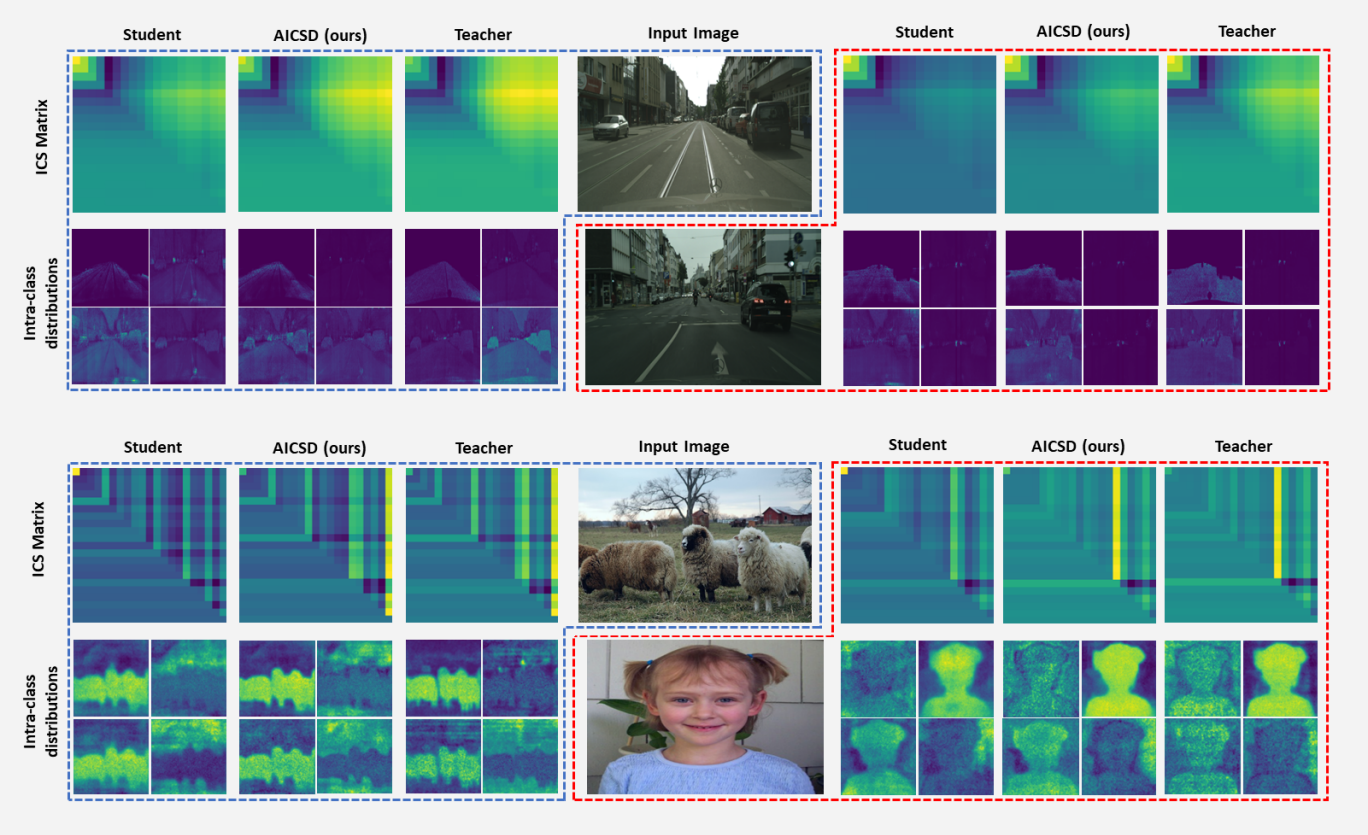

Fig. 5. Illustration of the intra-class distributions and inter-class similarity matrices for selected images, from the validation sets of Cityscapes and Pascal VOC datasets. For each image, the top row shows the inter-class similarity matrix and the bottom row shows the intra-class distributions of four classes. Proposed distillation method trains the student network to mimic the structures of the teacher network. Student backbone is MobileNet for Pascal VOC dataset and ResNet18 for Cityscapes images.

and Pascal VOC datasets. As can be seen from this figure, and validated by Figure 3, the two other distillation methods have poor performance for some classes (such as train and bus). In contrast, the proposed method can address this issue and improve the results for these two classes. The same improvement is observed for images from the Pascal VOC dataset, where the proposed method improves the results of classes such as bicycle, boat, and sofa by a good margin compared to the two other distillation methods.

V. CONCLUSION

This paper pressents the effective usage of knowledge distillation strategy to improve the performance of a lightweght network with the help of a larger and more complex network. In addition to the pixel-wise knowledge distillation method, the paper introduces a pair-wise method and a training strategy to enhance the knowledge distillation process. The proposed pair-wise method enhances the results of the student network by transferring inter-class similarities created from the outputs of the networks, making it appliable to any semantic segmentation network architecture. The training strategy also boosts results by combining the proposed ICSD method with the pixel-wise distillation in an adaptive manner that weights the losses that control the model.

Extensive experiments were conducted on two challenging datasets, using two different student networks, to validate the effectiveness of the proposed method. Ablation studies were also performed to demonstrate the impact of the ALW strategy, which can be combined with other distillation methods as well.

ACKNOWLEDGMENTS

The High Performance Center (HPC) of Sharif University of Technology is acknowledged by the authors for the provision of computational resources for this work.

REFERENCES

[1] 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.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[3] 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.

[4] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[5] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, “Efficient piecewise training of deep structured models for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3194–3203.

[6] 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.

[7] 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.

[8] 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.

[9] J. M. Alvarez and M. Salzmann, "Learning the number of neurons in deep networks," *Advances in neural information processing systems*, vol. 29, 2016.

[10] 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.

[11] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 535–541.

[12] 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.

[13] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.

[14] K. Yue, J. Deng, and F. Zhou, "Matching guided distillation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 312–328.

[15] 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.

[16] H.-J. Ye, S. Lu, and D.-C. Zhan, "Generalized knowledge distillation via relationship matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1817–1834, 2022.

[17] 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.

[18] C. Yang, M. Ochal, A. Storkey, and E. J. Crowley, "Prediction-guided distillation for dense object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 123–138.

[19] S. Tang, Z. Zhang, Z. Cheng, J. Lu, Y. Xu, Y. Niu, and F. He, "Distilling object detectors with global knowledge," in *European Conference on Computer Vision*. Springer, 2022, pp. 422–438.

[20] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7842–7851.

[21] Y. Feng, H. Wang, H. R. Hu, L. Yu, W. Wang, and S. Wang, "Triplet distillation for deep face recognition," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 808–812.

[22] Y. Huang, J. Wu, X. Xu, and S. Ding, "Evaluation-oriented knowledge distillation for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 740–18 749.

[23] J. Li, Z. Guo, H. Li, S. Han, J.-w. Baek, M. Yang, R. Yang, and S. Suh, "Rethinking feature-based knowledge distillation for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 156–20 165.

[24] 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.

[25] 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.

[26] Y. Shan, "Distilling pixel-wise feature similarities for semantic segmentation," *arXiv preprint arXiv:1910.14226*, pp. 1–12, 2019.

[27] 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.

[28] 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.

[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. 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.

[31] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

[32] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 184–192.

[35] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[36] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6819–6829.

[37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[38] 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.

[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

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

[41] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *arXiv preprint arXiv:1801.04381*, 2018.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, pp. 3, 5, 2017.

[48] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2414–2423.

[49] 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.

[50] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 346–362.

[51] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5311–5320.
[52] S. An, Q. Liao, Z. Lu, and J.-H. Xue, "Efficient semantic segmentation via self-attention and self-distillation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 256–15 266, 2022.
[53] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 319–12 328.
[54] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. Springer Science & Business Media, 2009.
[55] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

Amir Mohammad Mansourian received the B.Sc. degree from the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran, in 2021, and the M.Sc. degree from the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, in 2023.
He is a member of Image Processing Laboratory (IPL) since 2021. His current research interests include Computer Vision, Image/Video Processing, and Deep Learning.

Rozhan Ahmadi received the B.Sc. degree from the Department of Computer Engineering, Shahid Beheshti University, Tehran, Iran, in 2021, and the M.Sc. degree from the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, in 2023. She is a member of Image Processing Laboratory (IPL) since 2021. Her current research interests include Computer Vision, Image/Video Processing, and Deep Learning.

Shohreh Kasaei (M'05–SM'07) received the B.Sc. degree from the Department of Electronics, the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran, in 1986, the M.Sc. degree from the Graduate School of Engineering, Department of Electrical and Electronic Engineering, University of the Ryukyus, Okinawa, Japan, in 1994, and the Ph.D. degree from Signal Processing Research Centre, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, Queensland, Australia, in 1998.