This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

Digital Object Identifier

Robust Semantic Segmentation with Multi-Teacher Knowledge Distillation

ABDOLLAH AMIRKHANI[1], (Member, IEEE), AMIR KHOSRAVIAN[1], MASOUD MASIH-TEHRANI[1], HOSSEIN KASHIANI[2]
[1]School of Automotive Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran
[2]School of Electrical Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran

Corresponding author: Abdollah Amirkhani (e-mail: amirkhani@iust.ac.ir).

ABSTRACT Recent studies have recently exploited knowledge distillation (KD) technique to address time-consuming annotation task in semantic segmentation, through which one teacher trained on a single dataset could be leveraged for annotating unlabeled data. However, in this context, knowledge capacity is restricted, and knowledge variety is rare in different conditions, such as cross-model KD, in which the single teacher KD prohibits the student model from distilling information using cross-domain context. To fix this concern, we have looked into learning a lightweight student from a group of teachers. To be more specific, we train five distinct lightweight convolutional neural networks (CNNs) for semantic segmentation on several datasets. Several state-of-the-art augmentation transformations have also been utilized in our training phase. The impacts of such training scenarios are then assessed in terms of student robustness and accuracy. As the main contribution of this paper, our proposed multi-teacher KD paradigm endows the student with the ability to amalgamate and capture a variety of knowledge illustrations from different sources. Results demonstrated that our method outperforms the existing studies on both clean and corrupted data in the semantic segmentation task while benefiting from our proposed score weight system. Experiments validate that our multi-teacher framework results in an improvement of 9% up to 32.18% compared to the single-teacher paradigm. Moreover, it is demonstrated that our paradigm surpasses previous supervised real-time studies in the semantic segmentation challenge.

INDEX TERMS Autonomous vehicles, Convolutional neural networks, Knowledge distillation, Semantic segmentation, Semi-supervised learning

I. INTRODUCTION

Convolutional neural networks (CNNs) are now regarded as the state-of-the-art solutions for the bulk of computer vision applications, such as object detection [1], [2] and semantic segmentation [3], [4], thanks to the advancements in their deep learning designs. CNNs can conduct excellent feature extraction and pattern recognition by virtue of their powerful feature parameter sharing and dimensionality reduction. However, most of these models are trained by employing an entirely supervised learning approach and relying on human-labeled data. Compiling full human annotations for a large-scale database is an exceedingly time-consuming procedure [5]. This restriction has motivated the researchers to substitute the supervised learning perspective with various other techniques, such as the semi-supervised learning [6] approach, multi-source domain adaptation [7], and weakly supervised semantic segmentation [8], to take advantage of unlabeled data.

Recently, several works have employed knowledge distillation (KD) to transfer knowledge from a deep CNN, considered a teacher, to a compact student model. With this approach in mind, a teacher network is initially trained on a database containing labeled images. The trained network is then used to annotate a large amount of unlabeled data, called pseudo labels. Finally, the student network learns the desired features from the teacher-annotated database. The student-teacher KD performs in a variety of ways. For example, in several studies, the weights are updated by means of pseudo labels generated from the same self-training model [9]. In other papers, networks are simultaneously trained with both pseudo-labeled and labeled images for various tasks such as image classification [10] and object detection [11]. Additionally, in some iteration-based research works [12], a trained student becomes the new teacher and generates a new set of pseudo labels. These new labels are then leveraged to train the next student.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

The majority of KD student networks rely solely on the expertise of a single teacher [13], [14]. Even with outstanding deep learning architectures such as ESPNetv2 [15], Ghost-UNet [16], UNet++ [17], DeepLab [18], and DeepLabv3+ [19], the pseudo labels generated by a single teacher may fail to offer the high-level information required to train a robust student. This conclusion is based on the examination of two critical criteria: generalization and robustness. Compared to the single-teacher framework, multi-teacher KD has a higher chance of generalization to any desired image domain that the student network is expected to learn. If there is a significant domain shift between training data and new data, the model may not operate reliably on such data.

To address this issue, the presence of multiple informative and instructive teachers to aggregate their knowledge could effectively weaken the domain-shift related to model generalization. Autonomous vehicles are a great example of this application since they travel in various environments and consequently experience severe domain shift. Moreover, these images could contain specific noises against which the model may not be robust; resulting in an inferior segmentation performance. Since annotating new data by an expert teacher is mainly without human supervision; therefore, a student could promptly learn from falsely-annotated images, resulting in erroneous weight updates.

In this paper, we propose a new multi-teacher knowledge distillation (MTKD) method for semi-supervised learning, in which several teachers share their information on every pixel. As such, when one teacher fails to yield a high-precision segmentation, the remaining teachers could refine and enhance the label index and cancel out the effect of adverse errors. Our approach involves five DABNet [20] teacher architectures that have been trained using a variety of datasets, including Berkeley Deep Drive (BDD100K) [21], Mapillary [22], and Indian Driving Dataset (IDD) [23]. These databases cover numerous and varied domains to ensure that the teacher models can be adequately generalized to unseen environments. Besides, a unique technique is implemented to boost the robustness of each teacher against various synthetic image corruptions.

Although the teacher architectures are identical, the training phase for each teacher aims to cover a different domain. Having trained the teachers with different training scenarios, we leverage such teachers to instruct the FastSCNN model [24] in KD procedure using the Cityscapes database [25]. It is mandatory to note that we do not use any determined annotation in the Cityscapes database; instead, the teachers provide pseudo labels with the FastSCNN model on fine and coarse images. The performance of our teachers and students is evaluated on clean (without any corruption) as well as corrupted data in order to find out the impact of each training scenario on network robustness.

Finally, in our proposed multi-teacher approach, all the teachers simultaneously distill their information for the semantic segmentation challenge in a new lightweight student. Using a special score weight (SW) system, different scores are assigned to the teachers in a specified class based on their performances on clean and corrupted data. The output of this scoring system will determine the label of the desired pixel. Since multiple teachers are involved in generating these scores, a student will no longer acquire fine-to-coarse spatial features from a single restricted network. Therefore, auto-generated labels are more informative and instructive due to the knowledge aggregation of multiple teachers.

The main contributions of the present work can be summarized as follows:

- Implementing different training scenarios on teacher models and employing each teacher model to train individual students.
- Using clean and corrupted images to investigate the effects of these training scenarios on both teacher and student networks in order to find out which scenario yields a more robust model for KD in a semantic segmentation task.
- Proposing a multi-teacher method for semi-supervised learning, which is a simple yet practical approach to train a robust student network based on our SW system.

The rest of this paper is organized as follows. First, a brief literature review focusing on KD and model robustness is presented in Section II. Subsequently, Section III describes the selected databases, networks, and each scenario used in this paper. Our evaluation results and the outputs of the multi-teacher learning approach are given in Section IV. Finally, Section V concludes the paper.

II. RELATED WORK

This section presents the literature review related to central aspects of this paper, including KD in deep learning, multi-teacher KD, and the robustness of CNNs against different corruptions.

A. KNOWLEDGE DISTILLATION

Deep neural networks (DNNs) require large-scale and high-quality data to perform well, especially when dealing with various perturbations. Transferring knowledge from one expert source to another untrained source is a typical policy for coping with labeled data scarcity in the training stage [13]. In semi-supervised learning, a small set of labeled data is used in conjunction with a huge set of unlabeled data to train a network. KD is a type of semi-supervised learning that aims to train a student network so that it can compete with the accuracy of deep teacher models requiring less computational resources. Such a network would be appropriate for applications used in mobile devices [26]. Many studies have taken advantage of KD in a variety of computer vision tasks [14], [27]–[35]. For example, class probabilities generated by a sophisticated model have been employed to train a lightweight model [36]. KD has also been implemented to transfer the intermediate feature maps in an image classification task [37].

Furthermore, KD can be found in other tasks such as 3D object detection [37], pedestrian re-identification [38],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

and semantic segmentation [39]. Peng et al. [40] considered cross-modal and cross-domain challenges of KD and investigated a visual-textual life-long KD to address these issues. A one-step training procedure with a unified ensemble structure has been proposed in the KD field [41]. Another challenge in KD is incremental learning, where models need to learn new tasks while remembering the old ones. This challenge is analyzed in semantic segmentation, and the network updates its weights so that it learns new features from the new set of data while maintaining the knowledge of the prior key features [42]. The teacher model also needs to be dependable in order for the student network to learn from it effectively. This is investigated by Tan et al. [43], where authors presented an inter-class correlation regularization for training the mentioned reliable teacher.

Regarding single teacher KD, Liu et al. [44] adopted pixel-wise distillation in the semantic segmentation task and analyzed the efficiency of pair-wise and holistic distillation on different databases. By pointing to computational complexity and domain shift issues, Kothandaraman et al. [45] employed domain adaptive KD for autonomous vehicles application. Another category of studies in KD aim to tackle the issue of model efficiency while maintaining high accuracy [28], [46]–[48]. He et al. [46] apply KD on a compact student network to address the mentioned problem and take advantage of a pretrained autoencoder for feature similarity optimization. There also exist other variations of single teacher KD algorithms, which propose novel approaches to distilling knowledge, such as intra-class feature variation [49] or contextual-relation consistent domain adaptation [50].

B. MULTI-TEACHER KD

The student-teacher KD framework has made significant progress in transferring knowledge from one sophisticated teacher architecture to a lightweight student network. However, knowledge capacity and diversity may be constrained in some instances, such as cross-model KD [51]. To cope with this issue, the training of a portable student network by several teachers has been investigated [52]. In this study, a student learns to execute the same or different task from several teachers, rather than just one. With this strategy, students can assimilate different forms of knowledge gained from diverse teacher networks and build a comprehensive knowledge system.

In 2019, Mirzadeh et al. [53] demonstrated that any desired teacher network is only capable of distilling knowledge to a student model with a specific threshold of parameters. If the size of the model is less than that threshold, the KD procedure may not be compelling. Therefore, the authors take advantage of a multi-step KD context to mitigate the stated issue. In contrast to two-phase distillation approaches, on-the-fly Native Ensemble [54] is proposed, which encounters training procedure as a one-stage online distillation. In 2020, Xiang et al. [55] proposed a novel framework called Learning From Multiple Experts. In the definitions proposed by authors, 'Experts' refer to the models which extract features on less imbalanced data distribution. They exploit Self-paced Expert Selection and Curriculum Instance Selection as learning schedules for a reliable knowledge transfer from Experts to the student networks. Various works take advantage of KD with multiple teachers in different applications of computer vision, such as person re-identification [56], skin disease classification [57], and video action recognition [58].

C. MODEL ROBUSTNESS

In the wild, various corruptions concerning rotation and blurring and different adverse weather conditions might occur and impact the performance of trained CNNs. Owing to the vulnerability of CNNs, this can cause a noticeable drop in network performance. Su et al. [59] provided an in-depth investigation regarding the accuracy and robustness trade-off of numerous CNN models. So far, different benchmarks have been proposed with a focus on model robustness and robustness enhancement [60]. A simple way to improve the robustness of a machine learning model is data augmentation [61]. However, even though corrupted data can enhance the robustness of CNN models in semantic segmentation [62], introducing corrupted data into the training domain largely diminishes the model accuracy on uncorrupted data.

It worth noting that various types of corruption influence one another. One research perspective is to enhance model robustness with uncertainty estimation through various ways, such as style-transfer [63]. Furthermore, recent studies have assessed the robustness of CNNs against domain shift effect [64]. Hassaballah et al. [65] reported a detailed benchmark related to the performance of CNN models when facing various weather conditions, including snow and fog. In addition, they introduced a new dataset regarding challenges of vision-based self-driving cars in adverse weather conditions called DAWN. Kamann and Rother [66] demonstrated that model robustness is influenced by elaborate architectural designs as well as common image corruptions. There is also more variety of research activities regarding the subject of model robustness against adverse weather conditions [67], night scenes [68], and rain [69].

While there are many single-teacher KD standpoints, as discussed in subsection II-A, we look into KD using a multi-teacher approach. This allows us to combine the expertise of several different teacher networks into a single student model. Our student model enjoys key features collected from teacher networks that have trained by means of different large-scale databases. Despite the fact that plenty of studies also utilized multi-teacher KD frameworks, the research community has way less attention to the robustness of CNNs against image distortions. This paper aims to enhance the performance of the student model in noisy environments by means of a novel, easy-yet-effective multi-teacher KD approach. Furthermore, our proposed MTKD does not need high memory GPUs for the training procedures compared to other multi-teacher methods. This is due to the fact that we train and benchmark each teacher network on an individual basis. In the next step, utilizing our MTKD frame-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

work (which will be discussed in section III), we train our student CNN based on the knowledge of all teacher models. Therefore, since only one CNN is trained in each step, there is no need for high-end GPUs to process large data tensors from multiple teachers simultaneously. Another distinction between our technique and other multi-teacher approaches is that we educate each teacher with a distinct training scenario to robustify them against various noises in the wild. Our student network learns convolutional feature maps from all of these models, allowing it to generalize effectively and retain its performance in semantic segmentation tasks despite different kinds of corruptions.

III. THE PROPOSED APPROACH

In this section, we propose our central idea and elucidate the details of our MTKD method. Our approach includes four easy-to-implement operations, which can be summarized as follows: 1) Training teacher models on labeled data, 2) Evaluating the performance of each teacher on both clean and noisy data to determine which network is more robust in segmenting specific classes, 3) Using the teacher evaluation results to establish the SW system, and 4) Training a student network based on the SW system to determine the corresponding labels. Fig. 1 depicts an overview of our proposed method. As demonstrated in Fig. 1, first, a unique training scenario will be used to train each of the five individual teachers as thoroughly as feasible.

In the next step, these teachers will be employed to segment the unlabeled data. Then, exploiting our SW system, which will be discussed in subsection III-A, optimized labels will be generated for the unlabeled dataset and used to train a robust student.

In this experiment, the teacher database should be in a certain form to ensure the generalization of models. Despite this, there could be a noticeable domain shift between unlabeled data and training images, compromising the generalization of CNNs. This motivates us to select a mixture of databases that can provide large-scale domains for our teachers while yet allowing adequate domain shift for unlabeled data to challenge the MTKD technique.

The training datasets for the teachers include BDD100K, Mapillary, and IDD. Afterward, we will exploit the Cityscapes urban database for the semantic segmentation task; which contains unlabeled data and offers the desired domain shift compared to teacher training images. A number of selected images in Fig. 2 illustrate the severity of domain shift between labeled and unlabeled datasets. It is crucial to point out that the database for each teacher varies regarding different circumstances to enhance its robustness against various noises. Table 1 summarizes the databases employed for different scenarios. The databases have been downsized to the 480×360 resolution to speed up the training process due to the large number of training images.

Even though any desired CNN architecture can be easily implemented in the multi-teacher technique, we have used DABNet and FastSCNN as teacher and student networks,

TABLE 1: The databases and methods used in each scenario to enhance the robustness of teacher models

Scenario Index	Database			Scenario Technique			
	Mapillary	BDD100K	IDD	AdaIN	Augmix	Pretrain	Painting-by-Numbers
Scenario #1	✓	✓					
Scenario #2	✓	✓					✓
Scenario #3	✓	✓		✓			
Scenario #4	✓	✓			✓		
Scenario #5		✓	✓			✓	

respectively. Containing 0.76 and 1.14 million parameters, respectively, both of these networks are in the category of lightweight models. In this study, we have not utilized deep CNN architectures such as DeepLabv3+ to examine the practicality of the multi-teacher technique since KD still occurs between lightweight teacher and student architectures. Considering an ensemble of teachers with deep CNN architecture for KD consumes lots of resources and delays the training stage, which can confine the application of KD. In addition, it is proved that deeper CNNs do not necessarily make better teachers in that lightweight students fail to mimic deep teachers due to capacity mismatches. However, by taking a lightweight architecture such as DABNet, training the multi-teacher framework with a wide range of large-scale datasets would be more feasible in comparison with deep CNNs such as DeepLabv3+, and we manage to alleviate the mismatch between student and teacher capacities. Based on the conducted experiments, DABNet consistently achieves a higher accuracy even with fewer parameters, and alternatively, FastSCNN leads to a substantial reduction in run-time, making it more suitable for mobile device applications like autonomous vehicles.

To train a robust student on unlabeled data, we need to train each teacher model by means of a distinct training scenario. These training scenarios vary based on our needs and criteria. In our experiment, we adopted five different scenarios to accomplish multi-teacher learning. In our study, various teachers provide the student networks with multiple sources of information. By doing so, the student will be able to observe multiple forms of knowledge and generalize more effectively while also enjoy complementary information from each of the teachers. The reasoning for this may be described by the cognitive process of human learning. In fact, a student does not learn from a single teacher; rather, a student learns from different illustrations of knowledge better when numerous teachers on the same task or separate tasks are adopted.

Mentioned training scenarios are as follows: 1) consolidated BDD100K and Mapillary databases (No data augmentation), 2) painting-by-number data augmentation [70], 3) AdaIN fast style transfer [71], 4) Augmix augmentation [72], and 5) pretraining a teacher on an unstructured database (in our case, IDD). Each scenario will be discussed separately in subsequent sections.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

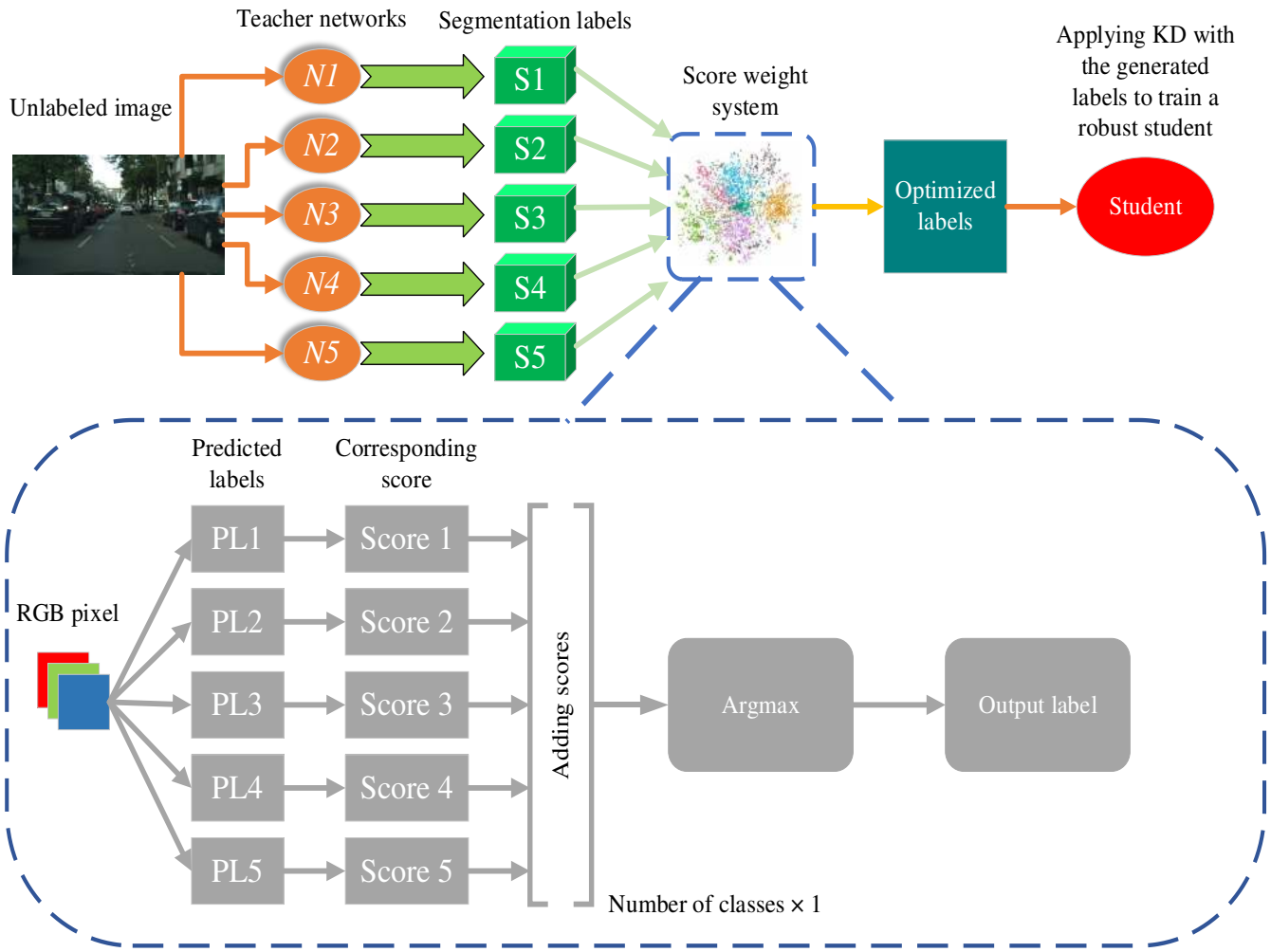

FIGURE 1: Overview of the multi-teacher technique and SW system.

a: Simple BDD100K and Mapillary databases (Sc1)

Scenario 1 is the baseline of this study. We train the DAB-Net model with a combination of BDD100K and Mapillary databases without any unique augmentations. However, we adopt basic augmentation techniques like flipping or random image scaling to improve the performance of the model. In this scenario, 25000 images are used in the training process; out of which 7000 images are related to the BDD100K, and the rest belong to the Mapillary database.

b: Painting-by-number (Sc2)

Painting-by-number is an easy-to-use augmentation technique for robustness enhancement in semantic segmentation tasks. However, because this technique requires accessing the ground-truth of images first, it may not be the ideal solution for a new database with totally unlabeled photos. Having the ground-truth labels, we can construct a texture-free representation of an image by attributing random colors to each label; which is unlikely to occur in the wild. In the next step, we alpha blend the original image with the texture-free version, based on Eq. 1.

$$P_{(h,w)}(aug) = \alpha P_{(h,w)}(pbd) + (1-\alpha)P_{(h,w)}(orig) \quad (1)$$

In the above equation, $P_{(h,w)}(aug)$, $P_{(h,w)}(pbd)$, and $P_{(h,w)}(orig)$ respectively denote the augmented, texture-free, and the initial intensities of a particular pixel of width w and height h. As stated in [70], variable α (with a random value between 0.7 and 0.99) is a blend parameter for reducing the texture bias of CNN models.

We apply painting-by-number augmentation to 25% of our training database. However, for an unbiased comparison between the considered scenarios, we do not add the augmented images to the training database in order to eliminate the enhancement effect of additional training data. Instead, we replace the original images with the augmented ones (in all training scenarios).

c: AdaIN fast style transfer (Sc3)

The AdaIN method [71] duplicates the content of an input image in the style of another. By adopting the AdaIN style transfer technique and alleviating the bias of a model, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

FIGURE 2: Sample images of each database; (a) BDD100K, (b) Mapillary, (c) IDD, and (d) Cityscapes.

can prevent our network to learn the features from image textures and make it focus better on image shapes. Similar to scenario 2, we replace 25% of our database with AdaIN style transfer outputs. Our style source for this scenario is the Kaggle painter-by-number dataset [73], and the stylization coefficient has been set to 0.5 for all image samples. Yim et al. [74] stated that AdaIN could not be found beneficial for semantic segmentation tasks since it distorts image structures. Although it is crucial in semantic segmentation to retain image content structure during the style transfer process, we believe that AdaIN could be practically used since it does not necessarily distort image patterns. From our perspective, by choosing a proper stylization coefficient and an appropriate style source, the AdaIN approach can be reliably applied to style a database in semantic segmentation task. In this work, we set the stylization coefficient to 0.5 since higher values increase the stylization strength and massively distort the image structure. Meanwhile, lower values decrease the model robustness since it does not make enough changes to the texture of the image.

d: Augmix augmentation (Sc4)

In the Augmix method [72], various augmentations are sampled arbitrarily and applied to an image. Since Augmix is based on random augmentations, it creates a dissimilarity between augmented images and, therefore, inhibits a model from memorizing the augmentation pattern. Like previous scenarios, we augment 25% of our database with the Augmix augmentation approach. For implementing the Augmix procedure in this scenario, we have randomly selected one to three individual operations, including brightness, posterize, sharpness, auto contrast, equalize, solarize, and contrast.

e: Unstructured pretrain (Sc5)

In contrast to the considered training domains, autonomous vehicles will eventually be driven in entirely diverse environments. In unstructured environments, such as off-road trails, there is a significant domain shift, and therefore model may face a severe performance drop. In this scenario, we seek to pretrain a model on an unstructured database and then finetune its weights on a structured one. In this regard, we take advantage of IDD as the unstructured database and complete the training procedure by using merely the BDD100K database. Note that we do not finetune the Mapillary database as other scenarios because the total number of images is not comparable with other scenarios. To be more specific, IDD and BDD100K databases contain 12872 and 7000 images, respectively.

Fig. 3 exhibits some sample images subjected to painting-by-number, AdaIN, and Augmix training scenarios. To investigate the effects of various scenarios on a student network, each teacher directly trains its relevant students, and then

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

their results are compared.

We train each network until the convergence, ensuring that the student model learns sufficiently and can avoid simultaneous overfitting. The model convergence is verified on a regular basis by comparing the accuracy and loss plots of the train and test sets. In this paper, the number of classes has been set to 15, which is the same as in all four databases. These classes include road, sidewalk, building, traffic light, traffic sign, terrain, vegetation, sky, person, rider, bus, car, caravan, motorcycle, and truck. Random scales and random mirror augmentations have been applied in all training procedures to guarantee that the models are as generalizable as possible. The initial learning rate has been set to 0.0005, which diminishes during training. We adopt the Adam optimizer [75] with batch sizes of 32 and 64 for teacher and student networks, respectively. Note that except for scenario 5, which includes pretraining and finetuning steps, the teacher and student networks are trained from scratch in the remaining scenarios. Our loss function is the 2D cross-entropy function presented in Eq. 2.

$$l = -(g_i \log(p_i) + (1 - g_i) \log(1 - p_i)) \tag{2}$$

where g_i and p_i represent the ground-truth and the model prediction for pixel i, respectively. The experiments were also repeated with the Focal loss function [76]; however, there was not much difference between the outcomes.

A. SCORE WEIGHT SYSTEM

In our MTKD framework, the SW system facilitates teacher communication and determines which label the student model should be trained on for a specific pixel. Since robustness enhancement in KD is the primary goal of our MTKD approach, we first need to analyze the robustness of each model separately. We employ various synthetic image corruptions to corrupt our test data. Such corruptions include shot noise, impulse noise, Gaussian noise, elastic transform, Gaussian blur, motion blur, defocus blur, pixelate, spatter, saturate, frost, glass blur, and zoom blur. In addition, four severity levels are considered for image corruptions in our experiments, ranging from level 1 (light corruption) to level 4 (severe corruption).

Each class gets an exclusive SW in each network based on its segmentation accuracy on both clean and corrupted data. Eq. 3 illustrates how to calculate SW using mean intersection over union (mIoU) as the accuracy measure.

$$SW_{(i,j)} = \sqrt{\sum_{n=0}^{N} mIoU_n^2} \tag{3}$$

In Eq. 3, $SW_{(i,j)}$ is the score weight for network i and class j. N and n are maximum and current severity levels, respectively. Since our work examines four different levels of corruption, N is set to 4. A severity level of $n = 0$ represents clean and uncorrupted data evaluation. In addition, $1 \leqslant n \leqslant 4$ is related to corrupted images and noisy environment in which the higher value of n indicates a greater degree of corruption. Fig. 4 demonstrates several clean and corrupted images to visualize each severity level in detail. Since the applied noises in this study are all synthetic corruptions, therefore the choice of severity levels should have a higher chance of occurrence in the wild. As a result, we only evaluate our student models on severity levels ranging from 1 to 4, as depicted in Fig. 4. It is worth highlighting that the SW system is not limited to the mIoU accuracy metric and could be defined by means of any other metrics. Having determined all the SW entries, we define the SW matrix according to Eq. 4. We consider a $C \times 1$ vector (score map) with zero elements for all given pixels, where C is the number of classes. When network i predicts class j, we add the $SW_{(i,j)}$ value to row j of the score map vector. After repeating the mentioned method for all adopted teachers, the SW system chooses the maximum argument of the score map vector as the label output.

$$SW = \begin{bmatrix} SW_{(1,1)} & SW_{(1,2)} & \cdots & SW_{(1,j-1)} & SW_{(1,j)} \\ SW_{(2,1)} & \ddots & \cdots & SW_{(2,j-1)} & SW_{(2,j)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ SW_{(i-1,1)} & SW_{(i-1,2)} & \cdots & \ddots & SW_{(i-1,j)} \\ SW_{(i,1)} & SW_{(i,2)} & \cdots & SW_{(i,j-1)} & SW_{(i,j)} \end{bmatrix} \tag{4}$$

IV. EVALUATION RESULTS

All the teacher networks are evaluated on clean and corrupted data to identify which scenario boosts the robustness of models the most. Moreover, we repeat the same procedure to determine the impact of each scenario on the robustness of the trained student model. Ultimately, we assemble a SW matrix for each scenario by which we train the robust students on unlabeled data. Comprehensive evaluations regarding our MTKD framework are presented in this section to prove the efficiency of our approach in both clean and corrupted data. Additionally, assessments are replicated by means of different real-time architectures to substantiate the compatibility of our method with other student networks.

A. EVALUATION OF TEACHER MODELS

Tables 2 and 3 illustrate the accuracies of teacher networks related to Mapillary and BDD100K validation sets, respectively. Also, Tables 4 and 5 show how teacher models perform in terms of robustness across all severity levels. Note that shrinking our teacher networks to save training and assessment time would lead to information loss in the training phase, thereby hurting the final accuracy. Fig. 5 displays the segmentation outputs of the teacher models for all five scenarios. Below, we will discuss the performance of each scenario in detail.

a: Scenario 1

Tables 2 and 3 depict that the basic model (scenario 1) performs well on clean data, with overall mIoU values of 60.8% for Mapillary and 57.3% for BDD100K databases,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

FIGURE 3: Sample images of our different training scenarios for enhancing the robustness of teacher models. The images from top to bottom belong to AdaIN fast style transfer (first row), Augmix augmentation (second row), and painting-by-number (last row) scenarios.

respectively. However, when faced with corrupted images, the performance of the baseline model drops dramatically; indicating that the model is not robust against these synthetic corruptions. At maximum severity level ($s = 4$), the accuracy of this model reduces to 22.54% (a drop of 62.93%) for the Mapillary and 20.61% (a drop of 64.03%) for the BDD100K validation set. The model learns better features in the Mapillary domain since the number of training examples is more extensive and hence achieves 3.5% higher accuracy on clean data.

b: Scenario 2

The painting-by-number approach can substantially adjust the texture of an image since the alpha-blend coefficient has a minimum value of 0.7. Although this is an effective method for emphasizing the shape-bias of a model, it leads to a slight performance drop on clean data. According to Tables 2 and 3, by applying this method, model accuracies on clean validation sets of Mapillary and BDD100K decline by 4% and 3.9%, respectively, compared to the baseline model. These modest accuracy drops are accompanied by a noticeable gain in robustness. For severity levels of 1 to 4, this model yields performance increases of 6.7%, 5.29%, 6.17%, and 5.44% on BDD100K and 4.5%, 4.56%, 4.82%, and 4.52% on Mapillary evaluation.

c: Scenario 3

Based on Tables 2 and 3, AdaIN surpasses the painting-by-number approach and attains superior accuracy on clean data. According to Tables 4 and 5, AdaIN enhances model robustness; supporting our assertion in relation to the effectiveness of the proposed method. Despite the fact that maintaining the pattern and structure of an image during style transfer is the top concern in semantic segmentation tasks, AdaIN could still be utilized in a manner that does not lead to undesirable distortion. This, of course, is in contrast to the assertion made by Yim et al. [74] related to AdaIN fast style transfer. Compared to the painting-by-number scenario (Sc2), AdaIN exhibits an average accuracy gain of 1.65% on clean data. Regarding robustness evaluation, at the maximum severity level ($s = 4$), AdaIN improves mIoU values for the Mapillary and BDD100K domains by 9.4% and 10.09%, respectively, at the highest severity level ($s = 4$).

d: Scenario 4

There is a trade-off between the Augmix augmentation method (Sc4) and AdaIN fast style transfer (Sc3). Augmix outperforms AdaIN on the clean dataset, and it also has a higher accuracy gain at severity level 1. However, AdaIN is more robust than the Augmix scenario at severity levels of 2, 3, and 4. This is due to the fact that Augmix is a type of augmentation algorithm, and AdaIN is a style transfer method. When an image is augmented with Augmix, the style remains the same as the original image, allowing the Augmix to perform better on clean data. However, when high-level image corruptions distort an image structure, AdaIN can deal better with such perturbations. Compared to the baseline teacher, AdaIN has a slightly higher performance (0.6% higher) on clean data. When it comes to the model's robustness, there are more apparent gains of 7.7 and 7.8% for the Mapillary and BDD100K databases, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

FIGURE 4: Visualization of image corruptions in each severity level applied on the cityscapes validation set.

e: Scenario 5

Our evaluation database does not contain unstructured images. Thus, we do not expect scenario 5 to compete with other scenarios. The IDD pretrain scenario has a performance reduction of 2% on BDD100K clean data; however, its performance for the Mapillary test increases to 14.1% since it is not trained in the Mapillary domain. The primary objective of this scenario is to keep the multi-teacher approach effective in unstructured environments such as off-road routes.

B. EVALUATION OF STUDENT MODELS

In this subsection, the student models are evaluated, and their robustness is compared to the similar models trained via supervised learning to determine whether our teacher models

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

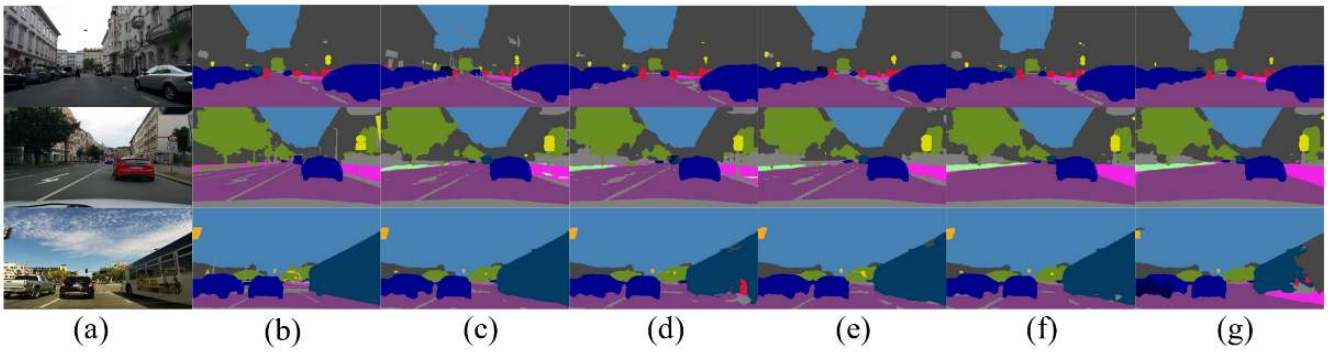

FIGURE 5: Sample images from the Mapillary validation set and the corresponding segmentation outputs of teachers; (a) original image, (b) ground-truth, and the outputs of (c) scenario 1, (d) scenario 2, (e) scenario 3, (f) scenario 4, (g) scenario 5

TABLE 2: Accuracies of teacher networks in all training scenarios for Mapillary validation set

Scenario	Total mIoU	Road	Sidewalk	Building	Traffic Light	Traffic Sign	Terrain	Vegetation	Sky	Bus	Caravan	Person	Car	Motorcycle	Truck
Sc1	60.8	82.2	54.3	79.4	43.3	52.0	54.9	84.3	96.7	48.3	22.9	60.5	82.7	36.4	53.3
Sc2	56.8	80.0	51.9	77.7	37.4	47.4	51.3	82.5	96.4	41.0	16.1	52.9	80.2	29.7	50.9
Sc3	57.5	81.1	51.7	77.9	38.7	48.2	50.1	82.5	96.5	45.0	15.1	54.5	80.7	31.1	52.0
Sc4	60.3	82.2	55.3	79.6	43.1	52.5	52.8	84.0	96.7	47.6	21.8	53.8	82.6	35.9	56.3
Sc5	46.7	68.7	25.9	70.3	28.6	37.8	30.5	76.7	94.5	42.0	6.5	37.1	75.9	21.3	37.6

TABLE 3: Accuracies of teacher networks in all training scenarios for BDD100K validation set

Scenario	Total mIoU	Road	Sidewalk	Building	Traffic Light	Traffic Sign	Terrain	Vegetation	Sky	Bus	Caravan	Person	Car	Motorcycle	Truck
Sc1	57.3	89.1	53.6	78.7	38.3	30.4	42.7	82.2	92.2	49.6	21.2	64.9	84.9	33.2	41.8
Sc2	53.4	87.3	50.7	76.9	30.9	26.1	39.9	80.9	92.0	41.8	12.4	65.0	82.8	24.4	36.8
Sc3	56.0	88.7	51.8	77.4	33.4	27.4	40.6	80.6	92.7	47.3	19.7	68.0	84.0	30.3	41.6
Sc4	57.9	88.0	51.3	78.8	36.8	30.9	39.0	82.0	92.5	50.0	36.1	67.5	84.1	32.3	42.2
Sc5	55.3	88.3	51.9	76.8	33.5	29.9	39.0	80.8	91.5	48.3	9.2	64.1	84.2	29.5	47.1

could be an appropriate substitute for human annotation and relieve the labeling burden. Table 6 presents the performance of the FastSCNN model on the Cityscapes dataset for each scenario. The severity level of zero in Table 6 stands for clean data evaluation.

With mIoU values of 58.4% and 58.3%, respectively, painting-by-number (Sc2) and Augmix (Sc4) approaches attain first-rate accuracies on clean data. They demonstrate 1.3% and 1.2% improvements over the mIoU of the same model trained using a supervised method, respectively. AdaIN fast style transfer (Sc3) also beats the models trained on the ground-truth annotations with a modest performance gain of 0.5%.

Expectedly, Scenario 1 and IDD pretrain (Sc5) both exhibit

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

TABLE 4: Teacher robustness evaluations related to the Mapillary validation set

Scenario	Severity	Total mIoU	Road	Sidewalk	Building	Traffic Light	Traffic Sign	Terrain	Vegetation	Sky	Bus	Caravan	Person	Car	Motorcycle	Truck
Scenario 1	1	45.7	68.5	38.4	67.9	23.1	38.4	36.2	62.5	85.1	34.4	11.4	42.7	70.7	20.0	40.8
	2	37.5	60.3	30.2	58.2	11.9	31.8	29.9	50.6	76.5	29.6	6.8	30.5	60.0	15.2	33.1
	3	30.5	52.3	24.4	45.0	6.0	24.5	26.1	39.6	69.4	24.5	6.0	24.4	49.8	9.5	25.1
	4	22.5	41.6	17.4	34.0	3.1	16.5	19.5	31.2	60.1	15.3	2.6	15.3	35.8	3.2	20.1
Scenario 2	1	50.2	74.8	40.9	69.7	34.6	45.4	42.6	65.5	82.2	41.0	15.3	43.0	76.3	26.6	45.4
	2	42.0	68.1	32.9	59.8	27.4	37.4	31.2	51.9	75.9	34.8	11.9	32.4	67.0	21.2	36.5
	3	35.3	61.3	27.0	49.5	21.4	30.0	27.6	41.8	72.6	28.8	10.4	25.4	55.0	16.1	27.0
	4	27.1	50.5	21.0	39.4	14.1	17.3	22.3	34.4	68.6	20.0	6.5	14.2	41.4	10.0	19.3
Scenario 3	1	52.2	76.1	44.4	73.3	33.6	44.5	43.1	77.1	94.4	39.8	12.1	45.7	76.0	25.7	45.4
	2	46.3	70.0	37.6	66.7	29.2	38.7	36.9	71.3	90.1	34.9	8.2	34.0	69.2	23.1	37.5
	3	39.7	64.6	33.0	54.4	23.5	29.9	34.1	64.7	79.0	31.0	8.2	19.9	62.2	19.3	32.5
	4	31.9	57.0	25.2	40.4	15.9	22.1	27.3	52.3	74.6	23.1	3.0	13.5	51.9	15.2	24.6
Scenario 4	1	53.1	76.2	43.7	72.4	36.2	47.0	43.1	75.6	92.5	42.0	13.1	45.6	77.7	28.1	49.7
	2	45.6	70.6	36.3	64.7	30.9	41.6	34.5	64.7	88.0	34.7	9.0	30.6	69.2	23.9	39.9
	3	38.1	65.3	31.0	55.4	24.0	34.2	28.5	50.5	82.0	27.8	6.2	21.7	61.1	16.8	29.3
	4	30.2	55.9	25.4	43.9	16.6	24.1	27.0	39.3	74.3	21.4	6.0	11.6	49.6	8.8	19.2
Scenario 5	1	39.4	62.1	19.9	60.1	19.8	30.6	24.1	67.5	87.8	34.1	5.9	27.2	66.0	14.6	31.8
	2	33.1	56.4	16.8	47.7	15.4	25.1	20.2	58.4	82.4	26.9	3.4	21.7	51.3	16.1	22.1
	3	26.8	52.0	14.4	36.0	11.4	19.2	15.4	51.1	76.8	18.0	2.7	12.0	41.1	9.1	16.6
	4	20.1	42.6	10.0	26.8	6.0	11.5	11.8	40.0	69.6	11.2	0.3	6.2	29.2	5.8	10.2

accuracy decreases of 1.1% and 11.1%, respectively. Scenario 1, in particular, cannot compete with a supervised approach since it is trained using a semi-supervised method and no special methodology is taken to enhance its performance. Scenario 5 is identical to the baseline model; however, since the Cityscapes database concentrates on structured scenes, it fails to train a reliable student using this data due to the significant domain shift.

In terms of robustness performance, scenarios 2 and 4 rank first. Compared to the supervised learning technique, the mIoU values for these scenarios are 2.4% and 2.3% higher, respectively. With the assistance of AdaIN fast style transfer, scenario 3 also outperforms the baseline model. At severity levels 0 to 3, it demonstrates accuracy gains of 1.6%, 3.8%, 2.8%, and 2.7%, respectively. However, at severity level 4, the model's mIoU value is 0.5% lower than the baseline model, attributed to severe synthetic image corruptions.

In terms of particular classes, the most prevalent classes considered in the Cityscapes database are roads, buildings, greenery, automobiles, and the sky. This prevalence gives rise to superior accuracy and robustness in all the models in that it provides a considerable amount of pixels with these labels, compared to other classes. With a mIoU accuracy of 91.2%, the Augmix scenario beats the other models. The supervised method, with 81.5% accuracy on clean data, ranks last in this class; and as for robustness evaluation, there is a noticeable gap of 4.5% in robustness evaluation between the supervised method and the Augmix model. Results indicate that the semi-supervised approach equipped with the appropriate augmentation technique can train a more versatile student capable of predicting challenging domains and being more robust against natural and synthetic distortions in the outdoors.

The performance of the supervised model in the sky class supports this assertion. According to Table 6, the ground-truth and scenario 1 models reach 96.4% and 96.2% accuracies for the reference class, respectively, which are greater than the remaining scenarios. Nevertheless, the performances of these models drop to 42.8% and 47.5% on synthetically corrupted data since none of these scenarios contain any robustness approach. Even though scenarios 2 to 4 initially experience a slight performance drop on clean data, they gain greater robustness evaluation accuracy. The painting-by-number approach attains an accuracy enhancement of 9.9% (23.13% gain) in comparison with the supervised model. Fig. 6 shows the overall accuracy of each model evaluated in this subsection.

C. EVALUATION OF THE MULTI-TEACHER METHOD

The results of the multi-teacher learning method are presented and discussed here. As described in subsection III-A, having computed all the SW elements and constructed the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

TABLE 5: Teacher robustness evaluations related to the BDD100K validation set

Scenario	Severity	Total mIoU	Road	Sidewalk	Building	Traffic Light	Traffic Sign	Terrain	Vegetation	Sky	Bus	Caravan	Person	Car	Motorcycle	Truck
Scenario 1	1	41.4	75.8	39.9	66.1	14.0	19.3	29.2	62.2	79.9	28.1	8.7	46.0	73.3	7.0	30.8
	2	35.6	65.8	34.5	57.4	5.8	14.9	25.3	49.5	72.6	25.9	10.5	40.9	64.8	5.2	25.4
	3	27.4	57.1	25.9	45.2	3.4	11.9	19.1	36.5	63.2	17.1	1.8	28.9	54.3	2.0	17.3
	4	20.6	44.5	17.8	33.8	1.6	9.6	14.1	28.0	52.5	11.5	2.2	22.8	40.5	0.0	9.7
Scenario 2	1	48.1	81.7	45.0	69.6	30.3	25.7	32.9	67.3	80.6	38.4	15.4	56.7	78.7	15.5	36.1
	2	40.9	74.1	37.9	60.0	25.0	21.1	27.6	53.1	74.5	29.7	16.6	44.8	70.4	6.1	31.9
	3	33.6	66.7	31.3	50.3	19.7	17.2	24.3	39.9	70.5	17.8	16.1	31.5	60.1	0.4	24.6
	4	26.1	54.9	24.9	40.8	13.3	8.9	18.4	31.8	65.7	12.2	9.5	19.5	47.6	0.1	17.3
Scenario 3	1	49.9	82.1	45.6	72.3	27.4	23.6	36.6	75.1	90.1	37.8	17.9	53.6	78.9	19.3	37.8
	2	44.2	75.6	40.1	65.1	24.9	20.4	32.3	68.7	85.6	31.6	9.8	45.1	73.5	14.4	32.5
	3	36.3	69.5	33.7	53.2	20.5	17.1	28.4	58.9	74.5	24.1	4.3	29.6	64.5	3.3	26.2
	4	30.8	62.0	29.1	41.4	13.6	12.6	21.6	47.3	69.9	16.7	15.0	26.8	56.3	2.6	16.0
Scenario 4	1	49.9	81.3	43.4	71.2	29.5	25.1	32.9	72.4	86.5	39.6	29.1	57.1	79.3	15.1	37.6
	2	43.3	74.9	37.6	63.6	26.4	20.9	25.8	60.8	81.5	32.3	24.4	45.3	72.3	7.3	32.9
	3	35.4	68.8	31.9	54.1	19.9	16.9	21.9	46.9	75.3	21.1	12.7	34.8	64.1	2.3	25.3
	4	28.4	59.8	26.7	43.6	12.9	11.2	20.2	36.4	68.4	14.9	9.1	26.6	53.7	0.6	13.8
Scenario 5	1	44.5	79.5	42.9	66.8	27.6	23.0	31.0	71.8	85.3	35.5	3.1	36.5	76.9	5.4	37.7
	2	38.2	69.9	35.3	54.5	22.8	19.1	23.9	62.2	79.0	29.0	7.1	32.4	65.2	8.5	26.2
	3	30.8	61.2	27.7	41.4	16.3	15.3	14.1	51.4	73.1	19.1	23.1	19.4	51.9	0.4	16.6
	4	21.8	48.5	20.1	31.7	6.9	9.1	10.3	38.1	65.9	14.6	0.1	13.6	39.4	0.2	7.5

SW matrix, we take the SW matrix to train a robust student network with all the five teacher models at the same time. The performances of the multi-teacher technique on clean and corrupt data have been listed in Table 7. In addition, Fig. 7 displays several segmentation outputs of our proposed method.

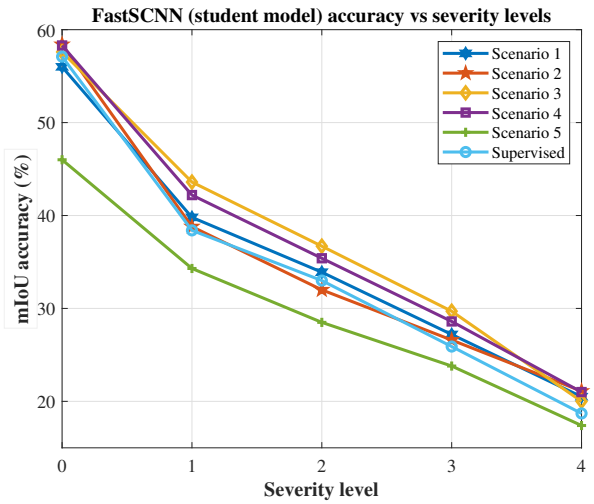

FIGURE 6: mIoU accuracies of FastSCNN for all the considered KD scenarios as well as the supervised approach. The results are associated with the Cityscapes dataset with 480×360 image resolution.

Table 7 reflects that our proposed multi-teacher learning approach surpasses the supervised and the student models trained in different scenarios. On clean data, the multi-teacher learning approach achieves a 6.67% gain in the mIoU metric compared to scenario 2; which indicates the best performance among all our experiments. Compared to the supervised model, this gain rises to 9.1%; proving the advantage of the multi-teacher learning approach.

In terms of robustness criterion, the MTKD still maintains its merits at all severity levels. To be more specific, for scenarios 1 to 5, Table 7 indicates accuracy gains of 12.19%, 9%, 15%, 9.5%, 32.18%, and 23% in robustness metric, at the maximum severity level ($s = 4$). This corroborates that our proposed SW matrix plays a substantial role in enhancing network communication, increasing the burden on computing resources. Armed with the SW matrix, we can determine whether a network's predictions are valid and whether an inaccurate annotation should be rejected.

D. EVALUATING MTKD WITH DIFFERENT NUMBER OF TEACHERS

In this paper, the total number of teacher models has been set to 5 concerning the presented robustness scenarios. Accord-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

TABLE 6: Performances of student models on clean and corrupted data for Cityscapes dataset

Scenario	Severity	Total mIoU	Road	Sidewalk	Building	Traffic Light	Traffic Sign	Terrain	Vegetation	Sky	Bus	Caravan	Person	Car	Motorcycle	Truck
Scenario 1	0	56.0	81.7	51.9	77.2	34.3	43.6	50.1	82.1	96.2	39.4	25.2	46.1	78.5	29.9	47.1
	1	39.8	72.7	41.0	62.9	12.8	23.6	29.1	66.8	66.7	33.7	8.1	34.5	68.6	7.4	28.9
	2	33.9	63.7	36.1	56.6	11.8	17.8	23.5	49.6	62.0	29.0	3.5	22.9	60.6	4.6	33.5
	3	27.2	57.1	31.9	47.3	8.0	13.4	22.2	36.6	57.1	23.0	2.0	16.1	49.2	2.1	14.5
	4	20.5	52.2	24.3	38.5	4.2	5.6	13.9	29.1	47.5	16.2	1.4	11.2	37.2	1.8	3.9
Scenario 2	0	58.4	90.2	63.3	81.0	27.6	40.1	46.0	84.3	86.9	47.4	25.6	60.1	84.1	25.4	55.7
	1	38.8	71.0	43.4	66.1	18.3	6.4	30.3	69.4	69.7	35.2	11.0	33.9	64.7	11.1	13.5
	2	32.0	63.8	37.3	58.7	12.5	3.7	25.3	55.3	64.6	31.2	6.7	23.0	54.3	6.7	4.3
	3	26.6	58.0	34.4	50.9	6.9	2.7	19.6	46.4	63.0	23.7	3.8	13.3	44.5	2.8	2.6
	4	21.1	50.8	27.3	43.9	2.7	1.9	12.2	38.9	52.7	17.1	2.7	9.7	31.6	2.2	1.1
Scenario 3	0	57.6	91.0	62.8	80.7	29.2	39.2	45.3	84.0	87.0	44.5	17.8	54.7	84.8	27.6	57.3
	1	43.6	79.6	46.5	68.5	18.0	24.5	32.8	71.3	68.4	34.9	8.9	34.9	68.2	8.8	44.8
	2	36.7	70.4	40.5	58.6	10.9	15.3	29.9	61.4	63.5	32.1	9.0	23.5	57.6	7.7	34.0
	3	29.7	62.7	35.5	50.6	8.0	8.0	24.3	48.9	59.7	25.2	3.4	16.6	45.8	4.3	22.9
	4	20.0	51.7	27.5	40.5	5.7	3.7	7.9	35.2	45.5	17.7	2.1	9.2	28.1	0.4	5.0
Scenario 4	0	58.3	91.2	65.5	80.6	31.8	40.4	47.9	84.3	87.0	45.8	25.4	52.9	84.4	25.6	53.0
	1	42.2	73.8	48.8	68.9	9.8	20.2	34.3	72.2	69.5	34.9	17.2	30.2	68.1	5.8	37.2
	2	35.4	65.8	43.3	57.5	10.5	5.9	30.8	54.2	63.3	31.0	14.5	19.7	57.5	7.2	35.1
	3	23.8	61.0	24.4	47.1	3.7	7.4	16.3	51.8	51.3	12.3	0.6	2.5	40.4	0.8	13.0
	4	17.4	49.2	19.5	38.8	1.7	3.8	8.9	42.3	42.7	6.9	0.2	1.7	25.0	0.2	2.6
Scenario 5	0	56.0	84.7	50.5	73.4	12.8	20.6	38.4	79.4	79.1	37.7	6.8	33.1	79.6	9.1	38.8
	1	34.3	73.4	33.3	63.8	5.8	12.3	25.5	68.0	64.1	27.7	2.9	12.5	62.1	1.8	26.4
	2	28.5	66.0	23.6	54.3	4.4	9.8	18.5	57.6	56.3	18.9	1.9	8.3	50.2	1.3	27.4
	3	23.8	61.0	24.4	47.1	3.7	7.4	16.3	51.8	51.3	12.3	0.6	2.5	40.4	0.8	13.0
	4	17.4	49.2	19.5	38.8	1.7	3.8	8.9	42.3	42.7	6.9	0.2	1.7	25.0	0.2	2.6
Supervised	0	57.1	81.5	54.2	77.1	35.6	44.0	52.6	82.0	96.4	41.3	24.2	50.3	78.3	30.4	51.7
	1	38.4	69.2	35.8	66.5	16.1	17.5	16.0	65.8	66.3	32.5	13.8	33.3	66.6	4.1	34.4
	2	33.0	61.8	31.7	58.5	12.6	12.4	16.2	54.0	61.5	27.3	11.9	22.6	57.7	2.1	32.2
	3	25.9	54.5	25.9	47.1	10.2	9.4	15.5	40.4	54.5	21.0	7.0	13.4	42.1	1.6	20.7
	4	18.7	49.6	19.1	37.1	6.9	4.8	9.5	30.9	42.8	13.0	4.9	6.3	26.2	2.2	9.1

TABLE 7: Results of applying the multi-teacher learning approach on the Cityscapes database. Severity level zero indicates the evaluation of clean data.

Severity	Total mIoU	Road	Sidewalk	Building	Traffic Light	Traffic Sign	Terrain	Vegetation	Sky	Bus	Caravan	Person	Car	Motorcycle	Truck
0	62.3	94.2	69.1	84.3	33.6	44.1	51.0	87.6	89.9	50.7	27.7	54.7	87.7	33.9	63.8
1	44.5	78.5	49.6	65.2	22.0	29.3	33.4	57.5	73.2	39.1	14.3	30.1	70.2	16.5	44.9
2	36.5	72.2	44.1	57.1	16.3	22.8	21.3	46.9	67.6	33.5	10.4	16.3	61.0	7.3	34.7
3	31.0	66.1	40.3	49.7	12.6	18.3	20.3	38.8	64.6	25.9	6.1	11.9	48.3	6.5	24.3
4	23.0	58.0	31.6	41.2	7.7	12.6	10.6	32.8	52.9	17.5	4.1	6.0	31.3	8.7	7.6

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

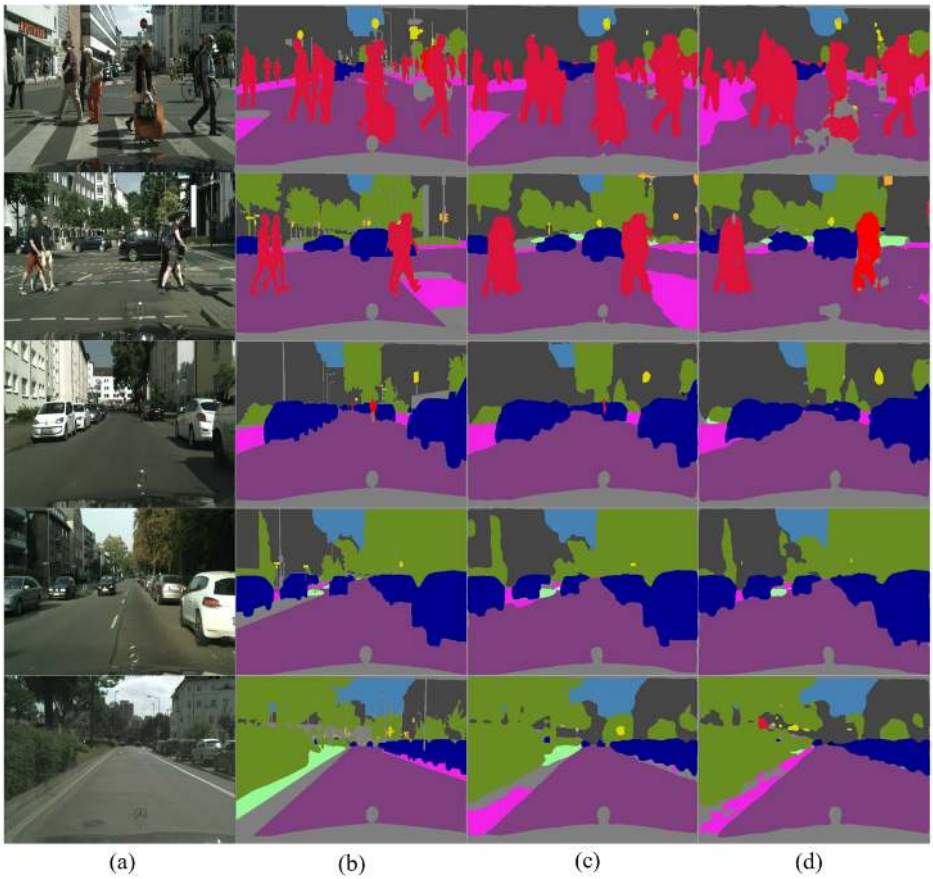

FIGURE 7: Random sample images of Cityscapes validation set and their corresponding segmentation outputs; (a) original image, (b) ground-truth, (c) multi-teacher learning segmentation, (d) supervised approach segmentation

ing to section III, the total number of teacher models is one of the key factors in MTKD. This subsection aims to analyze the performance of the proposed MTKD technique with different numbers of teacher models. We experiment with different settings for the teacher network quantity parameter and reconstruct the relevant SW matrices. By doing so, we manage to conduct an in-depth analysis of the influence of this parameter in the MTKD framework.

We replicate the student training procedure with two, three, and four teacher models and compare the evaluation results with five teacher MTKD discussed in subsection IV-C. Furthermore, the permutation of teacher networks is also taken into consideration to investigate the impact of each proposed scenario. Concerning the results demonstrated in Fig 6, it is worth noting that Scenario 5 has not been employed in this subsection since its unstructured environment is subjected to a large domain shift compared to our test domain. Therefore, scenario 5 is only adopted by the five teachers in KD. The accuracy of the student network in relation to the MTKD framework with varied teacher models number is shown in Table 8.

As demonstrated in Table 8, MTKD, even with two compact teacher models, is able to outperform the similar student network trained on ground-truth labels. This is due to the reliable SW system, making MTKD generate pseudo labels for unseen images constantly. With respect to the average accuracies reported in Table 8, the MTKD framework has enjoyed the minimum gains of 4.04%, and 8.96% and maximum gains

TABLE 8: Effect of the number of teacher networks on the performance of the student model.

Number of Teachers	Scenarios	Student Accuracy					
		0	1	2	3	4	Average
2	1, 2	58.8	39.9	33.4	26.9	21	36
	1, 3	58	44	36.1	30	20.5	37.7
	1, 4	58.5	43.7	36	29.4	20.9	37.7
	2, 3	60.4	43.6	35.8	29.8	20.6	38
	2, 4	61.1	42.1	35.8	28.2	21.3	37.77
	3, 4	61	44.2	36.2	30.1	20.8	38.4
3	1, 2, 3	58.9	43.5	39.1	30	21.9	38.6
	1, 2, 4	59.4	43.2	35.4	28.8	22.3	37.8
	1, 3, 4	58.6	43.8	35.2	30.3	20.6	37.7
	2, 3, 4	60.6	43.3	35.4	30.2	21.5	38.2
4	1, 2, 3, 4	62.1	44	35.8	30.4	22.3	38.9
5	1, 2, 3, 4, 5	62.3	44.5	36.5	31	23	39.4
Supervised		57.1	38.4	33	25.9	18.7	34.6

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

of 10.98%, and 11.56% compared to the supervised approach with two and three teacher models, respectively. Adopting MTKD with four teacher networks raises this gain to 12.42%. Even with the existence of severe domain shift in structured environment, the adopted scenario 5 pertaining to the teacher model still improved the gain from 12.42% to 13.87%. This demonstrates that the SW system can retain the robustness of CNN student even when a teacher model is not applicable in the specific target domain. This is due to the fact that the SW matrix continuously weighs down the corresponding label of the erroneous teacher network by utilizing the other teacher models. Nonetheless, this teacher can produce true positive labels on random pixels of the unseen image, which assists the SW system in choosing the correct label.

Analyses reveal that the choice of scenarios also dramatically impacts the robustness of the student model against synthetic corruptions. For instance, the accuracy of a student network trained with two teacher models ranges from 36% to 38.4%. This is due to the fact that scenario 1 suffers from the lack of a robustness enhancement approach. The combination of scenarios 3 and 4 can train a more robust network since both scenarios include a distinct approach to improve the robustness of the CNNs. Note that when utilizing the SW system with only two teacher networks, teacher selection causes a larger variety on student accuracy. To be more precise, when a teacher network with the higher impact in the SW matrix outputs an erroneous label, another network is not able to refine the label to the corresponding ground-truth. This variety decreases from 2.4% to 0.9% in MTKD with three teacher networks, as reported in Table 8, in that the other two networks have a greater chance to change the label to the desired one.

E. MTKD RESULTS WITH OTHER ARCHITECTURES

In previous sections, FastSCNN is adopted in training scenarios as the student model. For the sake of compatibility, in this subsection, the MTKD framework is utilized with different CNN architectures to analyze the efficacy of the approach in other relevant models. It should be noted that the model selection is based on compact real-time networks due to the computational complexity criterion, which should be satisfied in autonomous vehicles. To this end, ESPNetV2 [15] and LEDNet [77] are utilized for student networks. The mentioned architectures are trained on the Cityscapes dataset via the MTKD with all five teachers. For fair comparison, in all experiments, knowledge is aggregated with the same teacher architecture. The learning policy and training hyperparameters for these networks are similar to FastSCNN, discussed in Section III. Furthermore, training and evaluation phases are carried with 480×360 image resolution to ensure an unbiased comparison. Each model is also trained with a supervised approach with the same setting in a manner that the comparison can represent the efficiency of the proposed MTKD approach.

Table 9 illustrates the overall results of all networks. As it can be observed, the proposed framework proved to be impactful for knowledge aggregation in all evaluated models. This demonstrates the functionality of the SW system and validates that this approach can be effectively adopted to enhance the robustness of any desired CNN architecture. Occasionally, robustness improvement with data augmentation may contribute to the reduction of accuracy in the noise-free environment. However, the experimental evaluations indicate that MTKD can ameliorate the model robustness against image distortions and enhance the segmentation accuracy on the clean data simultaneously.

TABLE 9: Results of adopting different student architectures via the MTKD framework. All models with the training procedure label of MTKD are trained with knowledge aggregation of five DABNet teachers.

Architecture	Training Procedure	Student Accuracy					
		0	1	2	3	4	Average
FastSCNN [24]	MTKD	62.3	44.5	36.5	31	23	39.46
	Supervised	57.1	38.4	33	25.9	18.7	34.62
ESPNetV2 [15]	MTKD	60.9	42.3	31.8	26.6	20.8	36.48
	Supervised	55.6	36.5	28.7	22.3	16.8	31.98
LEDNet [77]	MTKD	66.8	49.2	40	34.5	26.1	43.32
	Supervised	61.1	42.7	36.3	27.9	22.4	38.08

Based on the results of Table 9, the trained models by means of our proposed MTKD framework outperform the similar model with a supervised training procedure. In overall evaluations, MTKD leads to improvements of 4.84, 4.5, and 5.24% in FastSCNN, ESPNetV2, and LEDNet architectures, respectively. Therefore, our evaluations validate the efficiency of SWS for KD in the semantic segmentation task. In terms of synthetic noises, all student networks also verify robustness enhancement in all severity levels. More specifically, performance improvements of 3.7 up to 4.3% are yielded for each network in the maximum severity level. The results validate that in the realm of vision-based autonomous vehicles, MTKD is a valuable framework for robustness enhancement of CNNs in noisy environments.

V. CONCLUSION

This paper proposed a novel, yet easy-to-implement, multi-teacher-based KD method for training a robust student network in semantic segmentation challenge. We initially train each teacher individually and then desired evaluation conditions are adopted for the evaluation phase. In our experiments, these conditions were constructed for benchmarking the robustness of our distilled students on both clean and corrupted data. Then, the SW matrix can be formed through specific evaluation metrics. When a teacher accomplishes a pseudo label assignment, the SW matrix adds the corresponding score to the score map, by which the label with the highest score would be picked for training the student model. Extensive experiments demonstrate that the proposed method beats all single student-teacher strategies, even when different augmentation and style transfer techniques are employed. Furthermore, compared to a fully-supervised model trained

on ground-truth labels, the MTKD has superior accuracy and robustness. We also investigate the influence of our MTKD framework on a variety of teachers and various permutations of teacher selections. The promising results demonstrate that: 1) Defining distinct training scenarios yield higher robustness of student network against a variety of synthetic image corruptions, 2) MTKD remains working satisfactorily even when the teacher fails to generalize well in the new target domain, 3) SW system verifies to be a highly effective way to exploit massive unlabeled data through semi-supervised learning fusion in that it constantly chooses appropriate corresponding labels concerning the evaluation of all teacher networks, 4) Even with a limited number of lightweight teacher networks, MTKD can easily surpass its competitors in supervised learning, and 5) Our framework yields impressive performance on both clean and corrupted data without requiring high computational resources. Moreover, due to the general structure of our framework, it can be applied to other vision tasks such as object detection and image classification.

REFERENCES

[1] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, 2020.

[2] A. Amirkhani and M. P. Karimi, "Adversarial defenses for object detectors based on gabor convolutional layers," *The Visual Computer*, 2021. [Online]. Available: doi.org/10.1007/s00371-021-02256-6

[3] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019.

[4] S. W. Cho, N. R. Baek, J. H. Koo, M. Arsalan, and K. R. Park, "Semantic segmentation with low light images by modified cyclegan-based image enhancement," *IEEE Access*, vol. 8, pp. 93 561–93 585, 2020.

[5] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and K. Keutzer, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2020.

[6] K.-K. Tseng, Y. Zhang, Q. Zhu, K. Yung, and W. Ip, "Semi-supervised image depth prediction with deep learning and binocular algorithms," *Applied Soft Computing*, vol. 92, p. 106272, 2020.

[7] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," in *Advances in Neural Information Processing Systems (NIPS)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 7285–7298.

[8] O. T. Nartey, G. Yang, S. K. Asare, J. Wu, and L. N. Frempong, "Robust semi-supervised traffic sign recognition via self-training and weakly-supervised learning," *Sensors*, vol. 20, no. 9, p. 2684, 2020.

[9] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: towards omni-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4119–4128.

[10] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 687–10 698.

[11] N.-V. Nguyen, C. Rigaud, and J.-C. Burie, "Semi-supervised object detection with unlabeled data," in *14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019, pp. 289–296.

[12] L.-C. Chen, R. G. Lopes, B. Cheng, M. D. Collins, E. D. Cubuk, B. Zoph, H. Adam, and J. Shlens, "Naive-student: leveraging semi-supervised learning in video sequences for urban scene segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 695–714.

[13] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: a survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[14] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[15] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "Espnetv2: a light-weight, power efficient, and general purpose convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9190–9200.

[16] I. A. Kazerouni, G. Dooly, and D. Toal, "Ghost-unet: an asymmetric encoder-decoder architecture for semantic segmentation from scratch," *IEEE Access*, vol. 9, pp. 97 457–97 465, 2021.

[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[18] 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, 2018.

[19] 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.

[20] G. Li and J. Kim, "Dabnet: depth-wise asymmetric bottleneck for real-time semantic segmentation," in *British Machine Vision Conference (BMVC)*, 2019, p. 259.

[21] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: a diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2636–2645.

[22] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4990–4999.

[23] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar, "Idd: a dataset for exploring problems of autonomous navigation in unconstrained environments," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1743–1751.

[24] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: fast semantic segmentation network," in *British Machine Vision Conference (BMVC)*, 2019, p. 289.

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[26] T. Ran, L. Yuan, and J. Zhang, "Scene perception based visual navigation of mobile robot in indoor environment," *ISA Transactions*, vol. 109, pp. 389–400, 2021.

[27] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[28] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2021.

[29] K. Zhang, C. Zhanga, S. Li, D. Zeng, and S. Ge, "Student network learning via evolutionary knowledge distillation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[30] Q. Zhao, J. Dong, H. Yu, and S. Chen, "Distilling ordinal relation and dark knowledge for facial age estimation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3108–3121, 2021.

[31] Q. Dou, Q. Liu, P. A. Heng, and B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2415–2425, 2020.

[32] X. Zhang, X. Li, Y. Yang, and R. Dong, "Improving low-resource neural machine translation with teacher-free knowledge distillation," *IEEE Access*, vol. 8, pp. 206 638–206 645, 2020.

[33] M. Zhu, J. Li, N. Wang, and X. Gao, "Knowledge distillation for face photo-sketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2020.

[34] 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.

[35] X. Li, S. Li, B. Omar, F. Wu, and X. Li, "Reskd: residual-guided knowledge distillation," *IEEE Transactions on Image Processing*, vol. 30, pp. 4735–4746, 2021.

[36] U. Gregor, G. Krzysztof J., E. K. Samira, A. S. W. Ozlem, C. Rich, M. Abdelrahman, P. Matthai, and R. Matt, "Do deep convolutional nets really need to be deep (or even convolutional)?" in *Neural Information Processing Systems (NIPS)*, 2016, pp. 2654–2662.

[37] P. YUN, Y. LIU, and M. LIU, "In defense of knowledge distillation for task incremental learning and its application in 3d object detection," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2012–2019, 2021.

[38] Y. Chen, N. Wang, and Z. Zhang, "Darkrank: accelerating deep metric learning via cross sample similarities transfer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[39] 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.

[40] Y. Peng, J. Qi, Z. Ye, and Y. Zhuo, "Hierarchical visual-textual knowledge distillation for life-long correlation learning," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 921–941, 2021.

[41] A. Zaras, N. Passalis, and A. Tefas, "Improving knowledge distillation using unified ensembles of specialized teachers," *Pattern Recognition Letters*, vol. 146, pp. 215–221, 2021.

[42] U. Michieli and P. Zanuttigh, "Knowledge distillation for incremental learning in semantic segmentation," *Computer Vision and Image Understanding*, vol. 205, p. 103167, 2021.

[43] C. Tan, J. Liu, and X. Zhang, "Improving knowledge distillation via an expressive teacher," *Knowledge-Based Systems*, vol. 218, p. 106837, 2021.

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

[45] D. Kothandaraman, A. Nambiar, and A. Mittal, "Domain adaptive knowledge distillation for driving scene semantic segmentation," in *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2021, pp. 134–143.

[46] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 578–587.

[47] L. Zhang, C. Bao, and K. Ma, "Self-distillation: towards efficient and compact neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[48] R. Takahashi, T. Matsubara, and K. Uehara, "A novel weight-shared multi-stage cnn for scale robustness," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1090–1101, 2019.

[49] 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 (ECCV)*, 2020, pp. 346–362.

[50] J. Huang, S. Lu, D. Guan, and X. Zhang, "Contextual-relation consistent domain adaptation for semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 705–722.

[51] C. Zhang and Y. Peng, "Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 1135–1141.

[52] Y. Liu, W. Zhang, and J. Wang, "Adaptive multi-teacher multi-level knowledge distillation," *Neurocomputing*, vol. 415, pp. 106–113, 2020.

[53] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5191–5198, 2020.

[54] X. Lan, X. Zhu, and S. Gong, "Knowledge distillation by on-the-fly native ensemble," in *Neural Information Processing Systems (NIPS)*, 2018, pp. 7528–7538.

[55] L. Xiang, G. Ding, and J. Han, "Learning from multiple experts: self-paced knowledge distillation for long-tailed classification," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 247–263.

[56] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1187–1196.

[57] S. Back, S. Lee, S. Shin, Y. Yu, T. Yuk, S. Jong, S. Ryu, and K. Lee, "Robust skin disease classification by distilling deep neural network ensemble for the mobile diagnosis of herpes zoster," *IEEE Access*, vol. 9, pp. 20 156–20 169, 2021.

[58] M.-C. Wu and C.-T. Chiu, "Multi-teacher knowledge distillation for compressed video action recognition based on deep learning," *Journal of Systems Architecture*, vol. 103, p. 101695, 2020.

[59] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, "Is robustness the cost of accuracy? a comprehensive study on the robustness of 18 deep image classification models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[60] J. Yim and K.-A. Sohn, "Enhancing the performance of convolutional neural networks on quality degraded datasets," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017, pp. 1–8.

[61] K. Alex, S. Ilya, and H. Geoffrey E, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[62] C. Kamann and C. Rother, "Benchmarking the robustness of semantic segmentation models with respect to common corruptions," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 462–483, 2021.

[63] S. Cygert and A. Czyżewski, "Toward robust pedestrian detection with data augmentation," *IEEE Access*, vol. 8, pp. 136 674–136 683, 2020.

[64] A. Khosravian, A. Amirkhani, H. Kashiani, and M. Masih-Tehrani, "Generalizing state-of-the-art object detectors for autonomous vehicles in unseen environments," *Expert Systems with Applications*, vol. 183, p. 115417, 2021.

[65] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4230–4242, 2021.

[66] C. Kamann and C. Rother, "Benchmarking the robustness of semantic segmentation models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8828–8838.

[67] A. Kerim, U. Celikcan, E. Erdem, and A. Erdem, "Using synthetic data for person tracking under adverse weather conditions," *Image and Vision Computing*, vol. 111, p. 104187, 2021.

[68] G. Li, Y. Yang, X. Qu, D. Cao, and K. Li, "A deep learning based image enhancement approach for autonomous driving at night," *Knowledge-Based Systems*, vol. 213, p. 106617, 2021.

[69] M. Tremblay, S. S. Halder, R. de Charette, and J.-F. Lalonde, "Rain rendering for evaluating and improving robustness to bad weather," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 341–360, 2021.

[70] C. Kamann and C. Rother, "Increasing the robustness of semantic segmentation models with painting-by-numbers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 369–387.

[71] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1501–1510.

[72] H. Dan, E. D. C. Norman, Mu, Z. Barret, G. Justin, and L. Balaji, "Augmix: a simple data processing method to improve robustness and uncertainty," in *International Conference on Learning Representations (ICLR)*, 2019.

[73] K. Nichol, "Painter by Numbers," https://www.kaggle.com/c/painter-by-numbers/, 2016.

[74] M. Kim and H. Byun, "Learning texture invariant representation for domain adaptation of semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 975–12 984.

[75] K. Diederik P and B. Jimmy, "Adam: a method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[76] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[77] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki, "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1860–1864.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107841, IEEE Access

ABDOLLAH AMIRKHANI received the M.Sc. and Ph.D. degrees (with honors) in electrical engineering from Iran University of Science and Technology (IUST), Tehran, in 2012 and 2017, respectively. He earned the Outstanding Student Award (2015) form the First Vice President of Iran. In 2016, he was conferred award by the Ministry of Science, Research and Technology. He is an Assistant Professor in the school of automotive engineering at IUST. He is the Associate Editor of the "Engineering Science and Technology, an International Journal". He has been actively involved in several National R&D projects, related to the development of new methodologies and learning algorithms based on AI techniques. His research interests are in machine vision, fuzzy cognitive maps, data mining and machine learning.

AMIR KHOSRAVIAN received his B.Sc. in mechanical engineering from Qom University of Technology (QUT) in 2017. He also received his M.Sc. in mechanical engineering from IUST in 2019. He is currently a Ph.D. student at the school of automotive engineering, IUST, Tehran, Iran. He is interested in autonomous vehicles, computer vision, deep learning, convolutional neural networks, and machine learning.

MASOUD MASIH-TEHRANI He received his B.Sc. degree in mechanical engineering at Tehran university, Tehran, Iran in 2004. He also received his M.Sc. in mechanical engineering from Yazd university, Yazd, Iran, in 2006 and Ph.D. from Tehran University, Tehran, Iran, in 2012. He is currently a faculty member and assistant professor at the school of automotive engineering, IUST, Tehran, Iran. His research interests are hybrid energy storage systems, hybrid flywheel vehicles, and vehicle control systems.

HOSSEIN KASHIANI received the B.Sc. degree in Electrical Engineering from Imam Khomeini International University, Ghazvin, Iran, in 2015, and the M.Sc. degree in Electrical Engineering - Digital Electronic Systems from IUST, Tehran, Iran, in 2018. He is currently a Research Assistant at the Computer Vision Center, Electrical Engineering Department, IUST. His current research interests include computer vision, deep learning, machine learning, and medical image analysis.