

Received May 7, 2020, accepted May 18, 2020, date of publication May 25, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997286

A Super-Learner Ensemble of Deep Networks for Vehicle-Type Classification

MOHAMED A. HEDEYA¹, AHMAD H. EID¹, AND REHAB F. ABDEL-KADER¹

Electrical Engineering Department, Faculty of Engineering, Port-Said University, Port-Said 42523, Egypt

Corresponding author: Mohamed A. Hedeya (mohamed.hedeya@eng.psu.edu.eg)

ABSTRACT Automatic vehicle-type classification plays an imperative role in the development of efficient Intelligent Transportation Systems (ITS). In this paper, a super-learner ensemble is proposed for the vehicle-type classification problem. A densely connected single-split super learner is utilized to exploit the strengths and diminish the weaknesses of the individual base learners ResNet50, Xception, and DenseNet. The super learner aims to learn fusion weights in a data-adaptive manner to obtain the optimal combination of the base learners. The proposed method is simple, robust, and enhances the discrimination capabilities among the similarly-looking classes without requiring any hand-crafted features or logical reasoning. The proposed method is evaluated using two of the most challenging publicly available traffic surveillance datasets: the MIOvision Traffic Camera Dataset (MIO-TCD) and the Beijing Institute of Technology's (BIT) vehicle classification dataset. Three variants of the super learner ensemble: RXD-CV-CW, RXD-CV-CW-NCW and Augmented-RXD, were examined on the MIO-TCD dataset with variations in applying class weights and data augmentation during training. RXD-CV-CW-NCW and Augmented-RXD share the third place among the published state-of-the-art methods reported in the MIO-TCD classification challenge. Augmented-RXD generalizes to the classes in common between the two datasets without degrading its performance on the MIO-TCD dataset. Both variants achieved an overall accuracy of 97.94%, and a Cohen Kappa score of 96.78%. In addition, the super-learner variants that we trained on the BIT-Vehicle dataset images achieved overall accuracies of up to 97.62%.

INDEX TERMS Deep learning, ensemble learning, intelligent transport systems, vehicle classification.

I. INTRODUCTION

Developing Intelligent traffic surveillance systems (ITSS) has become an important research area as it provides an innovative tool to improve transportation safety, efficiency, and driver satisfaction. Automatic vehicle type classification plays an imperative role in ITSS as it has various applications, such as Electronic Toll Collection (ETC), traffic control, intelligent parking systems, and traffic flow analysis.

As opposed to using intrusive installments of radars, loop detectors, or road tubes for traffic data acquisition, recent advances in machine learning gives a significant advantage to vision-based vehicle detection and classification methods. Automatic vehicle type classification is a challenging problem particularly when the images are captured by traffic surveillance cameras. Traffic surveillance images are usually low-resolution and subject to different illumination,

The associate editor coordinating the review of this manuscript and approving it for publication was Amr Tolba¹.

occlusion, and weather conditions. In addition, vehicle types introduce a lot of inter- and intra-class similarities. Although several vehicle datasets are currently publicly-available, not all of them are suitable for training traffic surveillance methods. Some datasets are targeted at autonomous driving with images taken by on-board cameras [1]–[3]. Other datasets contain high-resolution images taken by non-surveillance cameras and are typically used for fine-grained vehicle analysis [4], [5]. The Beijing Institute of Technology's (BIT)-Vehicle Dataset [6] contains 9,850 high-quality top-frontal view images that were captured by surveillance cameras. The dataset possesses many challenges, such as various lighting conditions, background confusion, and a variety of vehicle models and colors. The CompCars Dataset [5] is another surveillance dataset which contains 44,481 images. Although Yang *et al.* [5] used the CompCars dataset to prove the effectiveness of deep convolutional networks in classifying many car models; the dataset contains only frontal view images taken in daylight and clear weather. Furthermore,

it focuses on the fine-grained model categorization of cars, mini-vans and pickup trucks excluding large trucks, buses, motorcycles, and pedestrians. The MIOvision Traffic Camera Dataset (MIO-TCD) [7] is the largest traffic surveillance dataset available to date. The classification dataset consists of 648,959 low-resolution images, divided into 11 categories: Articulated Truck, Bicycle, Bus, Car, Motorcycle, Non-Motorized Vehicle, Pedestrian, Pickup Truck, Single-unit Truck, Work Van and Background. The images were captured at different time periods during the day and under different weather conditions. The captured images contain vehicles in diverse orientations. The classification task using the MIO-TCD is extremely challenging. This is due to the high imbalance nature of the dataset, the inter-class similarity between the categories that have similar visual characteristics, and the heavy compression artifacts in some images.

Ensembles of artificial neural networks have gained popularity in many image classification and localization applications due to their exceptional adaptive prediction performance [8], [9]. Ensembles combine several baseline models that have different architectures to improve the stability and predictive capability of the model. The performance of the individual base-learners depends mostly on the data-dimensionality, model-hypothesis and the bias-variance trade-offs of the model. Consequently, it is unfeasible to know beforehand which learner would attain the best performance given a specific prediction problem and a particular dataset. Ensembles can effectively harness the complementary strength of the different base learners as some base learners might have a weak overall prediction but can be effective at discriminating specific subclasses. Different merging strategies were reported in the literature such as majority voting, unweighted average, Bayesian voting...etc. However, these methods are vulnerable to weak learners, sensitive to over-confident learners and may lead to information loss.

The super learner is a loss-based supervised-learning ensemble framework that minimizes the cross-validation risk for combination by finding the optimal combination of a group of prediction algorithms [9]. This is achieved by optimizing the weights of the base learners on the validation set in an adaptive manner. The Super Learner could be considered as an extra 1×1 convolution layer over the validation set stacked on the outputs of the base learners.

The main contributions of this paper can be summarized as follows:

- We present a super-learner ensemble model for vehicle-type classification in surveillance frames. The super learner consists of a fully-connected layer added to the fused outputs of three base learners: ResNet50 [10], Xception [11], and DenseNet [12].
- The different networks were trained and tested using two of the most challenging and largest publicly available traffic surveillance datasets; the MIO-TCD and the BIT-vehicle datasets.

- While our method is simple, easy to train, does not include any handcrafted features or any logic reasoning components, the experimental results demonstrate its effectiveness. In terms of the overall evaluation metrics, the ensemble performs better than each of the base learners and is on a level comparable to the state-of-the-art methods.

The rest of this paper is organized as follows: Section II provides an overview of the related work. The technical details and the framework of the proposed system are presented in Section III. Experimental results of the proposed system and comparisons to existing algorithms are reported in Section IV. Finally, concluding remarks are summarized in section V.

II. RELATED WORK

A. TRADITIONAL VEHICLE-TYPE CLASSIFICATION METHODS

Traditional vehicle-type classification models integrate different types of sensors and image-processing methods that incorporate essential hand-crafted features depending on the application context and the granularity of the required classification. Cho *et al.* [13] applied a Kalman filter to fuse radar and LIDAR systems for object detection and classification. They switched between two motion models for tracking pedestrians, bicyclists, and cars. Thakoor and Bhanu [14] proposed a feature called structural signature to classify vehicles into sedan, pickup truck and SUV/minivan from their rear-view videos observed on highways. They used support vector machines (SVM) for the classification. Kafai and Bhanu [15] presented another rear-view based classification method using the spatial information among the landmarks of the vehicle (e.g. taillights and license plates) and a Hybrid Dynamic Bayesian Network (HDBN) classifier with multiple time slices corresponding to multiple video frames. The main limitation of the methods described in [14] and [15] was that they could not differentiate between SUV and minivan because these two vehicle categories look similar to the rearview. Theagarajan *et al.* [16] were able to discriminate between SUV and minivan from the rearview. They presented a method to compute the Visual Rear Ground Clearance of a vehicle from its rear-view video and classify it into two classes namely Low Visual Rear Ground Clearance Vehicles and High Visual Rear Ground Clearance Vehicles.

B. VEHICLE-TYPE CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS AND DEEP LEARNING

Image classification started to shift towards convolutional neural networks after Krizhevsky *et al.* [17] introduced unprecedented performance in the ImageNet LSVRC (ILSVRC-2010) competition [18]. Dong *et al.* [6] used a pre-trained Convolutional Neural Network together with multi-task learning to classify the vehicles into Bus, Microbus, Minivan, Sedan, SUV, and Truck from vehicle

frontal-view images. They introduced the Sparse Laplacian Filter Learning (SLFL), an unsupervised learning method, to learn the filter bank of the convolutional layer. They used their BIT-Vehicle dataset, which includes 9850 high-resolution vehicle frontal-view images. Khaled *et al.* [19] used the BIT dataset to study the effect of color and spatial resolutions of the vehicle images on the classification results of a variety of classification methods. Huval *et al.* [20] used the OverFeat [21] architecture along with a mask detector to detect vehicles and highway lanes in real-time. Wang *et al.* [22] used CNN together with Fisher feature encoding algorithms for vehicle type classification. The datasets used in the above-mentioned approaches did not contain enough samples that can represent real-world traffic surveillance images.

C. VEHICLE-TYPE CLASSIFICATION USING THE MIO-TCD DATASET

As emphasized in Section I, the MIO-TCD dataset is one of the largest datasets prepared for traffic surveillance purposes. The MIO-TCD traffic surveillance challenge was introduced in conjunction with CVPR 2017. Several ensemble methods were designed to address the MIO-TCD Classification Challenge. Kim and Lim [23] implemented a bagging system by training several CNN models with several random subsets of the MIO-TCD dataset. To compensate for the imbalanced data distribution, they applied weighted voting that depends on the error rate of each class. Lee and Chung [24] proposed an ensemble method that combines local and global expert networks. The local expert networks were all GoogLeNet, and they were trained using subsets of the dataset depending on the aspect ratio and the size of the input images. The global expert networks comprised of three convolutional nets (AlexNet, GoogleNet, and ResNet18) that were trained on the entire dataset. At the test time, the local experts are selected using a gating function and the network outputs are combined using a softmax layer. Jung *et al.* [25] proposed an ensemble model that they called Joint Fine-tuning with DropCNN that enabled them to train several ResNets simultaneously. Theagarajan *et al.* [8] proposed an ensemble of three ResNet models. A weighted loss function was applied to handle the imbalanced distribution of the dataset. They also implemented patch-based logical reasoning to address the genuine dual-class misclassification problem. To address the imbalanced data challenge, Liu *et al.* [26] proposed a method that integrates deep neural networks with balanced sampling in two stages: data augmentation with balanced sampling and an ensemble of convolutional neural networks trained on the augmented data. Their method was able to enhance the mean precision of all categories while preserving high overall accuracy. Later, Liu *et al.* [27] proposed a method that applied generative adversarial nets (GANs) for data augmentation. Their proposed approach consists of three stages: training several GANs on the original dataset to generate adversarial samples for the rare classes, training an ensemble of different-architectures of CNN models on the

original imbalanced dataset, and finally refining the ensemble model on the augmented dataset after filtering out the low-quality adversarial samples. This resulted in increasing the mean performance of some categories while maintaining high overall accuracy. Although deep model-based methods can achieve very promising performance, a number of challenges remain such as: distinguishing similarly-looking vehicles, unbalanced datasets, false detections and small vehicles [7].

III. PROPOSED WORK

Each model has its strengths and weaknesses. The aim of ensemble learning is to supervise the strengths and weaknesses of multiple models, leading to better classification decisions in general. Our proposed method for vehicle-type classification is a stacking ensemble of three deep neural networks inspired by the super-learner ensemble method proposed by Ju *et al.* in [9]. As opposed to the thoughtful weighted average ensemble that was presented in [8], instead of using pre-set fusion weights, our proposed super learner aims to learn fusion weights in a data-adaptive manner. The proposed ensemble is a cross-validation ensemble framework that acquires a non-linear fusion function that can better exploit the individual base learners' strengths and reduce their weaknesses, and hence enhance the discrimination capabilities among the similarly-looking vehicles. The proposed network architecture is shown in Figure 1.

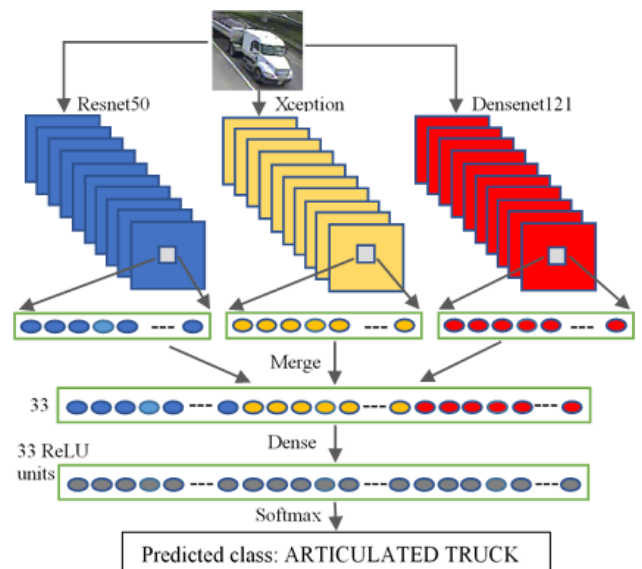


FIGURE 1. Architecture of the super-learner ensemble.

A. BASE LEARNERS

We used three powerful deep convolutional neural network models as the base learners: ResNet50 [10], Xception [11], and DenseNet [12]. ResNet introduced a residual learning framework to ease the training of deep networks. It reformulated the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced

functions. DenseNet introduced several advantages: it avoids the vanishing gradient problem, strengthens feature propagation, improves feature reuse, and substantially reduces the number of parameters. Xception introduced a novel deep convolutional neural network architecture inspired by Inception, where Inception modules have been replaced with depth-wise separable convolutions. Compared to Inception V3 [28], Xception achieved performance gains due to using the model parameters more efficiently. These three models proved to be the best individually-performing networks on the MIO-TCD as reported in [7]. Being three different powerful networks should provide the opportunity to exploit the strengths of each network through the super learner. Each of the three models takes 224×224 RGB input images and has an 11-output softmax layer corresponding to the 11 categories of the MIO-TCD classification dataset.

B. THE SUPER-LEARNER ENSEMBLE

The proposed super learner was designed to attain a non-linear fusion function of the outputs of the base learners in order to enhance its discrimination capabilities considering the imbalanced nature of the MIO-TCD classification dataset and its inter-class similarities. Therefore, instead of applying a linear stacking of the base learners, stacking on the logit scale, or just stacking a 1×1 convolution layer on the output of the base learners as explained in [9], we added a fully-connected layer with ReLU activation units between the merged output of the base learners and the output softmax layer. The fully connected layer consists of 33 ReLU activation units.

C. CROSS-VALIDATION

Wolpert introduced the idea of stacking in [29]. As an extension of stacking, van der Laan *et al.* introduced the super learner in [30] as a cross-validation based stacking. It combines the base learners by optimizing the v -fold cross-validated loss to compute the optimal ensemble weight vector. V -fold cross-validation is best-suited and optimal for small datasets. It was applied to a variety of topics, such as survival analysis [31] clinical trial [32] and mortality prediction [33]. For large classification datasets, optimizing the v -fold cross-validated loss would require a huge time. Instead, we applied the single-split super learner, in which only the set-aside validation set is used to train the super-learner ensemble in addition to its original purpose of assessing and tuning the base learners. Therefore, the weights of the super learner are calculated by minimizing the single-split cross-validated loss as suggested in [9]. Ju *et al.* [34] show the success of the single-split super learner on three large healthcare databases.

D. DATA AUGMENTATION

We performed some of our experiments with data augmentation. In those experiments, we used the images of the Sedan and SUV classes of the BIT-Vehicle dataset to

augment the Car class of the MIO-TCD dataset. We also augmented the MIO-TCD's Bus class with the Bus images of the BIT-Vehicle dataset.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATASETS

We performed our experiments on 2 large traffic surveillance datasets: The MIO-TCD dataset and the BIT-vehicle dataset. We applied our super learner variants and compared them to other methods on the MIO-TCD dataset first, and then extended the idea to the BIT vehicle dataset.

The MIO-TCD Dataset is highly imbalanced. The size of each class is shown in Table 1.

Four metrics are used for the evaluation of the MIO-TCD classification challenge. The first one is the overall accuracy *Acc*, which is defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

where *TP* is the number of true positive images regardless of their category, *TN* is the number of True negative images, *FP* is the number of false positive and *FN* is the number of false negative images.

Dominating categories such as 'Car' and 'Background' have a strong influence on the accuracy metric. The other three metrics, which are the mean recall (*mRec*), the mean precision (*mPre*), and the Cohen Kappa Score (*Kappa*) [35] account for this imbalance. The mean precision and mean recall are defined as follows:

$$mPre = \frac{\sum_{i=1}^{11} Pre_i}{11}, \quad mRec = \frac{\sum_{i=1}^{11} Rec_i}{11}, \quad (2)$$

where












$$Pre_i = \frac{TP_i}{TP_i + FP_i} \quad \text{and} \quad Rec_i = \frac{TP_i}{TP_i + FN_i}. \quad (3)$$

The Cohen Kappa Score measures the agreement between two annotators: the first annotator is a method under evaluation and the second annotator is the ground truth. It is defined as follows:

$$Kappa = \frac{Acc - P_e}{1 - P_e}, \quad (4)$$

where P_e is the probability of agreement when the two annotators assign random labels. It is a good measure for both multi-class and imbalanced class problems. It basically measures how much better a specific classifier is performing than a classifier that guesses randomly according to the frequency of each class. That said, there is controversy surrounding Cohen Kappa due to the difficulty in interpreting indices of agreement. Stein *et al.* [36] applied Bradley-Terry model, suggesting that it may serve as an extension to Kappa that can provide more detail upon strength and direction of disagreement. Pontius *et al.* [37] suggested that it is conceptually simpler and more informative to evaluate quantity and allocation disagreement between items.

TABLE 1. Size of each category in the MIO-TCD dataset.

Category	Training	Samples of training images	Testing
Articulated Truck	10,346		2,587
Bicycle	2,284		571
Bus	10,316		2,579
Car	260,518		65,131
Motorcycle	1,982		495
Non-Motorized Vehicle	1,751		438
Pedestrian	6,262		1,565
Pickup Truck	50,906		12,727
Single-Unit Truck	5,120		1,280
Work Van	9,679		2,422
Background	160,000		40,000
Total	519,164		129,795

The final ranking of the MIO-TCD classification methods is calculated by taking the average of the ranks of the 4 metrics: *Rank (Acc)*, *Rank (mPre)*, *Rank (mRe)* and *Rank (Kappa)*.

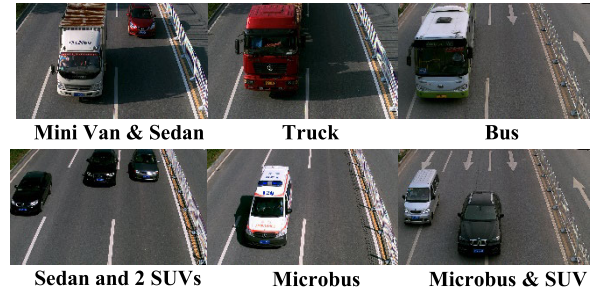


FIGURE 2. Sample images of the BIT-vehicle dataset.

The BIT-vehicle dataset is highly imbalanced as well. The dataset contains 9850 high-resolution vehicle frontal-view images. All vehicles in the dataset are divided into six categories: Bus, Microbus, Minivan, Sedan, SUV, and Truck. Each image in the dataset contains one or more vehicles, so the location of each vehicle is pre-annotated. The numbers of vehicles in each category are 558, 883, 476, 5,922, 1,392, and 822, respectively. The actual image sizes of the BIT-vehicle dataset are 1600×1200 and 1920×1080 . Figure 2 shows downsized sample images from the BIT-vehicle dataset.

B. PREPROCESSING

Each image was resized such that the shorter side has a length of 256 pixels and the other side has the required length to maintain the aspect ratio. So, if we have an $M \times N$ image with an aspect ratio $AR = N/M$, the resized image will have the following dimensions:

$$\begin{cases} \text{if } M < N; M = 256, & N = AR \times M \\ \text{if } M > N; N = 256, & M = N/AR, \end{cases} \quad (5)$$

Then, during each training epoch, a randomly cropped 224×224 patch from each input image is extracted and used for training.

C. EXPERIMENTAL SETUP

We performed our experiments on an ASUS ROG STRIX with Intel Core i7-6700HQ CPU, 16GB of RAM, and an NVIDIA GeForce GTX 1060 GPU with 6GB of GPU memory. Keras with Tensorflow backend was utilized in the experiments.

The training set of the MIO-TCD dataset was split into 80% data for training and 20% data for validation. In addition to validating the base learners, the validation set was used for training the super-learner ensemble. The base learners were all initialized with ImageNet pre-trained weights.

To handle the imbalanced nature of the MIO-TCD dataset, we used the class-weighted categorical cross-entropy loss function for most of the training epochs of the base learners, and for training one of the super-learner ensemble variants. We set the class weights such that the weight of each class is equal to the total number of training images divided by the number of images of that class.

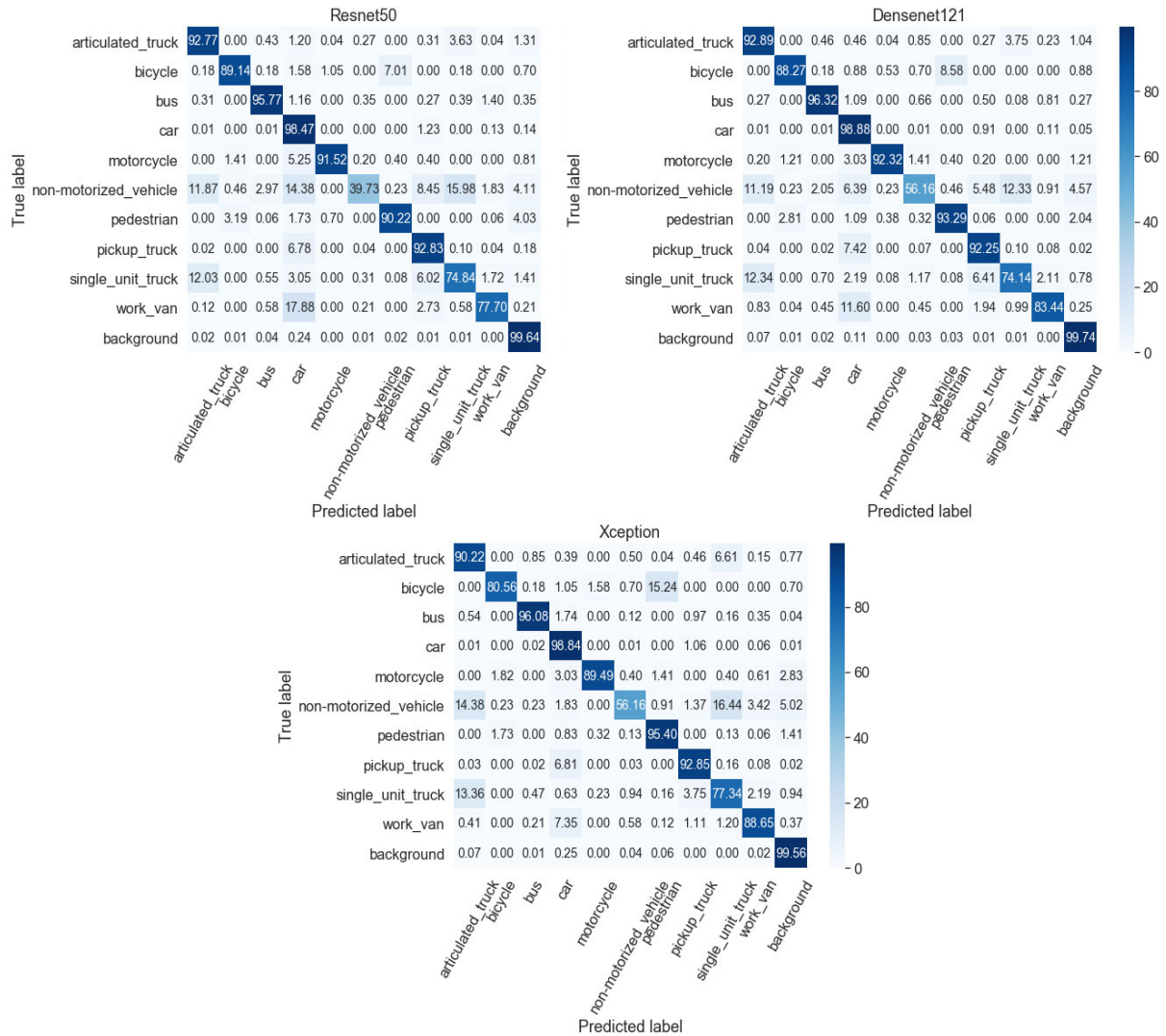


FIGURE 3. Confusion matrices of the base learners evaluated on the test set.

TABLE 2. Size of each category in the BIT-vehicle dataset.

	Training	Validation	Testing	Total
Bus	334	112	112	558
Microbus	529	177	177	883
Minivan	284	96	96	476
Sedan	3,551	1,185	1,185	5,921
SUV	834	279	279	1,392
Truck	493	165	165	823
Total	6,025	2,014	2,014	10,053

There is no separate testing set for the BIT-vehicle dataset, so we randomly split the data into 60% for training, 20 % for validation, and 20% for testing. The random splits took into consideration to maintain the same proportion of the number of vehicles per category as the original dataset. Table 2 shows the size of each category in the 3 splits of the BIT-vehicle dataset.

As suggested in [7], we used the Adam [38] optimizer with a learning rate of 10^{-3} . However, the learning rate was reduced in the later epochs of training the base learners.

During each epoch, the data is randomly shuffled. We adjusted the training batch size used for each of the individual models as well as the ensemble model to allow data to suit the 6GB GPU memory.

To avoid overfitting, we applied early stopping. The training is stopped if the validation loss does not improve after 5 consecutive training epochs.

D. EXPERIMENTAL RESULTS ON THE MIO-TCD DATASET

We fine-tuned the Resnet50, Densenet121, and Xception networks on the MIO-TCD dataset until they reached testing accuracies of 97.13%, 97.51%, and 97.54%, respectively. Fig. 3 shows the confusion matrices of each of the 3 base learners evaluated on the testing set.

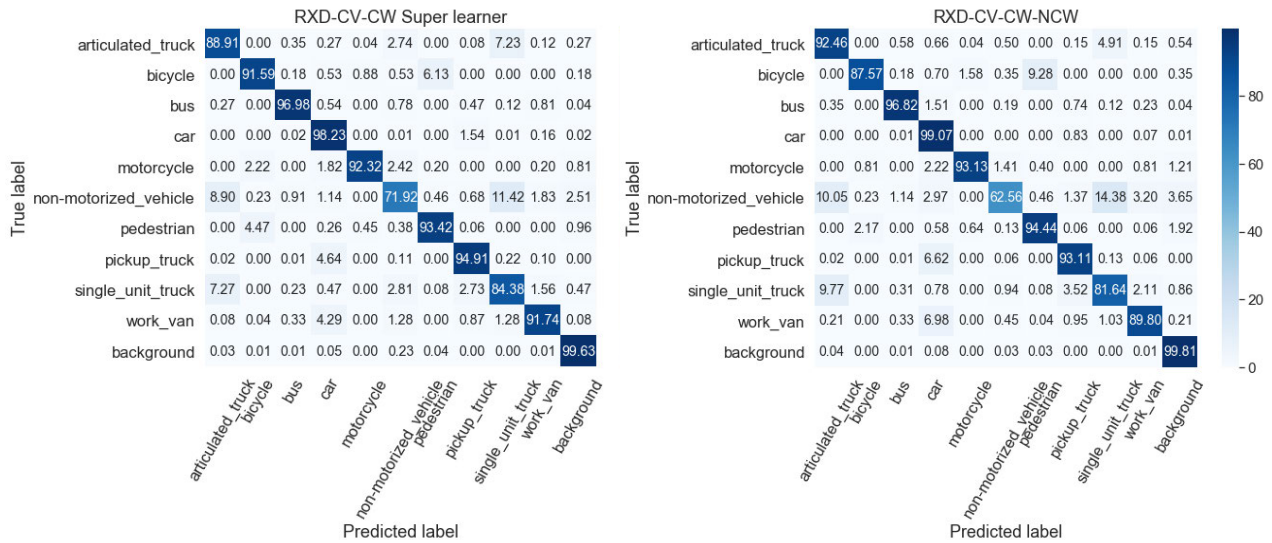


FIGURE 4. Confusion matrices of “RXD-CV-CW” vs. “RXD-CV-CW-NCW” evaluated on the test set.

Subsequently, we trained the super learner using the validation set for one epoch with the class weights applied. We call this method “RXD-CV-CW”. Then, we trained the super learner for more epochs without applying the class weights. We call this method “RXD-CV-CW-NCW”.

To get more accurate predictions on the testing set, we applied the standard 10-crop method [17] for evaluating the 3 base learners as well as the super-learner methods. Therefore, after resizing each test image such that the shorter side is 256 pixels, we extracted 10 patches which are the central crop, the four corners, and their horizontal flips and averaged the predictions made by each model. Fig. 4 shows the confusion matrices of “RXD-CV-CW” and “RXD-CV-CW-NCW” evaluated on the testing set.

Although both proposed super-learner methods were trained only on the images of the validation set for just a few epochs, they attained high accuracy on the testing set images. Table 3 -A demonstrates that with the exception of the mean precision of “RXD-CV-CW”, both of the proposed super-learner methods achieved better evaluation-metric scores compared to the base learners. Table 4 and 5 show the Recall and Precision scores of our base learners compared to those of our super-learner methods. Although the recall and precision scores of some of the base learners for few individual classes outperform the super learner ensembles, table 3 demonstrates that the super learners achieve a significantly better scores in the four overall performance metrics. This supports the statement mentioned earlier in the introduction that ensembles can effectively harness the complementary strength of the different base learners. Though some base learners might have a weak overall prediction it can be effective at discriminating specific subclasses.

Using the class-weighted loss function in “RXD-CV-CW” resulted in an improvement in the recall scores of some rare-sample classes, such as the Bicycle, Work Van, Single Unit

TABLE 3. Comparison of testing results on the MIO-TCD dataset a) with augmentation and b) without augmentation. Boldface indicates achievement of the best result.

		Cohen Kappa Score	Acc.	Mean Prec.	Mean Rec.	
A)	Base Learners (No Augmentation)	Resnet50	0.9551	0.9713	0.9239	0.857
		Densenet121	0.961	0.9751	0.9159	0.8797
		Xception	0.9617	0.9754	0.9156	0.8774
Super-Learner Ensembles (No Augmentation)	RXD-CV-CW	0.9638	0.9767	0.895	0.9127	
	RXD-CV-CW-NCW	0.9678	0.9794	0.9298	0.9004	
B)		Cohen Kappa Score	Acc.	Mean Prec.	Mean Rec.	
Augmented Base Learners	Resnet50	0.9546	0.9709	0.9177	0.8576	
	Densenet121	0.9552	0.9713	0.8827	0.8702	
	Xception	0.961	0.975	0.9132	0.8771	
Augmented Super-Learner Ensemble	Augmented-RXD Super Learner	0.9678	0.9794	0.9215	0.9027	

Truck, Motorcycle, and Non-Motorized Vehicle. However, using the class-weighted loss function resulted in a relatively low mean precision score. On the other hand, training “RXD-CV-CW-NCW” for few epochs with un-weighted loss function considerably increased the mean precision, overall accuracy, as well as the Cohen Kappa score.

Fig. 5 presents samples of the different testing images that were correctly classified by either of the proposed

TABLE 4. The recall scores of the base learners vs. the proposed super learners (The classes are denoted as AT: Articulated Truck, BI: Bicycle, Bus: Bus, Car: Car, MO: Motorcycle, NMV: Non-motorized vehicle, PE: Pedestrian, PT: Pickup Truck, SUT: Single-unit Truck, WV: Work Van, BG: Background). Boldface indicates achievement of the best result.

		Recall Scores										
		AT	BI	Bus	Car	MO	NMV	PE	PT	SUT	WV	BG
Base Learners	Resnet50	0.9277	0.8914	0.9577	0.9847	0.9152	0.3973	0.9022	0.9283	0.7484	0.7770	0.9964
	Densenet121	0.9289	0.8827	0.9632	0.9888	0.9232	0.5616	0.9329	0.9225	0.7414	0.8344	0.9974
	Xception	0.9022	0.8056	0.9608	0.9884	0.8949	0.5616	0.9540	0.9285	0.7734	0.8865	0.9956
Proposed Super-Learner Ensembles	RXD-CV-CW	0.8891	0.9159	0.9698	0.9823	0.9232	0.7192	0.9342	0.9491	0.8438	0.9174	0.9963
	RXD-CV-CW-NCW	0.9246	0.8757	0.9682	0.9907	0.9313	0.6256	0.9444	0.9311	0.8164	0.898	0.9981

TABLE 5. The precision scores of the base learners vs. the proposed super learners (The classes are denoted as AT: Articulated Truck, BI: Bicycle, Bus: Bus, Car: Car, MO: Motorcycle, NMV: Non-motorized Vehicle, PE: Pedestrian, PT: Pickup Truck, SUT: Single-unit Truck, WV: Work Van, BG: Background). Boldface indicates achievement of the best result.

		Precision scores										
		AT	BI	Bus	Car	MO	NMV	PE	PT	SUT	WV	BG
Base Learners	Resnet50	0.9094	0.8899	0.9728	0.9754	0.9577	0.8325	0.9658	0.9218	0.8216	0.9230	0.9933
	Densenet121	0.8980	0.9016	0.9760	0.9787	0.9723	0.6910	0.9574	0.9383	0.8303	0.9352	0.9962
	Xception	0.8875	0.9237	0.9802	0.9810	0.9630	0.7664	0.9222	0.9359	0.7663	0.9483	0.9972
Proposed Super-Learner Ensembles	RXD-CV-CW	0.9357	0.8602	0.9835	0.9883	0.9723	0.5189	0.9644	0.9181	0.7781	0.9270	0.9984
	RXD-CV-CW-NCW	0.9221	0.9259	0.9827	0.9825	0.9584	0.7896	0.9548	0.9486	0.8145	0.9506	0.9978

super-learner methods or misclassified by both. The MIO-TCD dataset contains a lot of challenging images. Due to the blurry nature, the low resolution and compression artifacts in some images, they are hard to be classified even by humans. Although our super-learner methods were robust in accurately predicting the classes of many challenging images of the MIO-TCD dataset, they still fail in classifying some images as shown in the column of the suspected misclassified images in Fig. 5.

Table 6 lists the evaluation results of the proposed super learners vs. state-of-the-art methods that participated in the MIO-TCD classification challenge. “RXD-CV-CW” achieved the best classification accuracy of the Bicycle (91.59%) and Work Van (91.74%) classes. “RXD-CV-CW-NCW” achieved the second-best overall accuracy (97.94%) and Cohen Kappa score (96.78%). “RXD-CV-CW-NCW” comes at the third rank after the methods of [25] and [8], which got the first- and second-best mean precision scores respectively. Our mean precision score is relatively lower than those achieved in [25] and [8].

E. SUPER-LEARNER ENSEMBLES VS. WEIGHTED-AVERAGE ENSEMBLE

We compare the performance of the super-learner ensemble with the performance of a simple weighted average ensemble of the base learners. The three base learners were combined using weighted prediction vectors. We used the same weighing approach of [8]. So, the weight vectors were the average of the precision and recall of each individual class

as follows:

$$W_{in} = Average(Pre_{in}, Rec_{in}), \tag{6}$$

where i refers to the base learner, n refers to the class index, $Pre_{in} = TP_{in}/(TP_{in}+FP_{in})$ and $Rec_{in} = TP_{in}/(TP_{in}+FN_{in})$. The weights for each network are obtained by evaluating the precision and recall scores of that network on the validation set. The final prediction is then calculated by averaging W_1X_1, W_2X_2 and W_3X_3 which are the weighted predictions of the 3 base learners. We called this network “RXD Weighted-Average Ensemble”. Table 6 demonstrates that the “RXD-CV-CW-NCW” ensemble achieves better scores than the “RXD Weighted-Average ensemble” in all the performance metrics except for the mean precision score. As a result, “RXD-CV-CW-NCW” achieves a better average rank than that of the “RXD Weighted-Average Ensemble”.

F. ENSEMBLES WITH DIFFERENT FUSION METHODS

In the proposed super learner ensembles “RXD-CV-CW” and “RXD-CV-CW-NCW” we used a simple concatenation on the outputs of the individual base learners. In [39] and [40], T. Akilan et al. explored fusion approaches other than concatenation that can improve classification accuracy. We examined the use of the product fusion and max fusion approaches that were introduced in [39]. We call them “RXD Multiplication Super Learner” and “RXD Max Super Learner”. The results presented in Table 6 reveals that the product and max fusion approaches excel in the recall or the precision scores of some of the individual classes and the “RXD Multiplication Super Learner” achieves the highest recorded mean recall score.



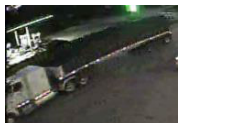
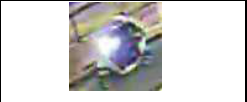
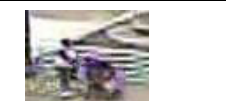


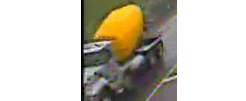


	Correctly classified by both models	Correctly classified by “RXD-CV-CW” only	Correctly classified by “RXD-CV-CW-NCW” only	Suspected misclassification by both models
Articulated Truck				
Bicycle				
Bus				
Car				
Motorcycle				
Non-motorized Vehicle				
Pedestrian				
Pickup Truck				
Single Unit Truck				
Work Van				
Background				

FIGURE 5. Examples from the classification results of “RXD-CV-CW” and “RXD-CV-CW-NCW” on the MIO-TCD classification dataset.



FIGURE 6. Examples of the classification results on the BIT-vehicle dataset (The green caption is the ground truth from the BIT-Vehicle dataset, the red caption is the predicted class from the MIO-TCD dataset, and the blue caption indicates a correct equivalent label).

TABLE 6. The results of the MIO-TCD super learners vs. the state-of-the-art methods designed for the MIO-TCD dataset (The classes are denoted as AT: Articulated Truck, BI: Bicycle, Bus: Bus, Car: Car, MO: Motorcycle, NMV: Non-motorized vehicle, PE: Pedestrian, PT: Pickup Truck, SUT: Single-unit Truck, WV: Work Van, BG: Background). Boldface indicates achievement of the best result.

	Cohen Kappa Score	Acc.	Mean Prec.	Mean Rec.	AT	BI	Bus	Car	MO	NMV	PE	PT	SUT	WV	BG
Jung et al. [25]	0.9681	0.9795	0.9530	0.8970	0.9324	0.8949	0.9779	0.9853	0.9111	0.5228	0.9406	0.9539	0.8336	0.9166	0.9984
Theagarajan et al. [8]	0.9658	0.9780	0.9439	0.9190	0.9451	0.8984	0.9794	0.9790	0.9374	0.7237	0.9348	0.9624	0.8445	0.9059	0.9980
RXD-CV-CW-NCW	0.9678	0.9794	0.9298	0.9004	0.9246	0.8757	0.9682	0.9907	0.9313	0.6256	0.9444	0.9311	0.8164	0.8980	0.9981
Augmented -RXD Super Learner	0.9678	0.9794	0.9215	0.9027	0.9219	0.8722	0.9779	0.9894	0.9172	0.6621	0.9374	0.9370	0.8219	0.8947	0.9979
Kim and Lim [23]	0.9666	0.9786	0.9355	0.9041	0.9412	0.8739	0.9593	0.9866	0.9131	0.7078	0.9610	0.9510	0.8273	0.8258	0.9980
Lee and Chung [24]	0.9675	0.9792	0.9298	0.9024	0.9358	0.8774	0.9620	0.9889	0.9212	0.6872	0.9425	0.9507	0.8289	0.8353	0.9966
Liu et al. [27]	0.9657	0.9780	0.9355	0.9074	0.9324	0.9089	0.9891	0.9862	0.9333	0.6164	0.9259	0.9455	0.8383	0.9116	0.9933
RXD Weighted-Average Ensemble	0.9660	0.9783	0.9422	0.8827	0.9451	0.8792	0.9698	0.9912	0.9253	0.5114	0.9489	0.9283	0.7578	0.8547	0.9986
RXD Multiplication	0.9637	0.9769	0.8669	0.9462	0.9169	0.8634	0.9531	0.9923	0.8929	0.4269	0.9361	0.9249	0.8008	0.8303	0.9985
RXD-CV-CW	0.9638	0.9767	0.8950	0.9127	0.8891	0.9159	0.9698	0.9823	0.9232	0.7192	0.9342	0.9491	0.8438	0.9174	0.9963
Liu et al. [26]	0.9651	0.9776	0.9201	0.8844	0.9312	0.9037	0.9663	0.9889	0.9010	0.5594	0.9022	0.9402	0.7898	0.8468	0.9984
RXD Max	0.9624	0.9760	0.9318	0.8736	0.9227	0.8704	0.9624	0.9895	0.9354	0.4680	0.9367	0.9309	0.7836	0.8126	0.9979

However, their overall ranks are low compared to the other ensemble methods listed in Table 6.

G. TESTING THE PERFORMANCE OF OUR MIO-TCD NETWORKS ON THE BIT-VEHICLE DATASET

This experiment was conducted to examine how the “RXD-CV-CW-NCW” super learner would generalize to the images of the BIT-vehicle dataset without performing tailored training on the BIT-vehicle images. There are considerable differences between the MIO-TCD dataset and the BIT-vehicle dataset. The MIO-TCD images introduce many challenges because they are recorded during daytime/nighttime, different seasons, diverse weather conditions, various camera positions and orientations and have strong compression artifacts. On the other side, the BIT-vehicle images are high-resolution top-frontal view images that are taken in clear weather conditions and most of them are taken during the daytime. Also,

while the Car class of the MIO-TCD dataset contains vehicles of type sedan, SUV and family van, the BIT-vehicle dataset dedicates separate classes for the sedan and SUV vehicles. The mini-van category of the BIT-vehicle dataset is different from the work van category of the MIO-TCD dataset. It looks more like the single-unit truck.

As the vehicle locations in the BIT-vehicle dataset are pre-annotated, we did not have to apply the 10-crop method for testing. We just cropped the vehicle object at the pre-annotated location, resized the cropped object image so that the shorter side is 256 pixels, and then made the prediction based on the center-cropped 224 × 224 patch.

Table 7 shows the confusion matrix of the 2,014 images that were randomly selected as the test sample from the BIT-vehicle dataset and how they were classified to the MIO-TCD classes without training on the BIT-vehicle dataset.

TABLE 7. Confusion matrix of ‘RXD-CW-CW-NCW’ on the BIT-vehicle test set without training on the BIT-vehicle images. Boldface indicates accurate classifications.

		Predicted						
		AT	Bus	Car	PT	SUT	WV	BG
True	Bus	36.6 %	29.5 %	0.9%	0.0 %	17.9 %	0.9 %	14.3 %
	Micro bus	3.4%	46.3 %	0.6%	0.6 %	20.3 %	28.8 %	0.0 %
	Minivan	10.4 %	39.6 %	0.0%	0.0 %	45.8 %	4.2 %	0.0 %
	Sedan	1.9%	15.6 %	15.3 %	14.9 %	32.7 %	19.3 %	0.3 %
	SUV	5.0%	22.6 %	6.8 %	3.2 %	31.2 %	30.8 %	0.4 %
	Truck	40.6 %	30.9 %	0.0%	0.0 %	27.3 %	0.0 %	1.2 %

Despite the obvious differences between the two datasets, the results were reasonable. None of the 2,014 BIT-vehicle test images were misclassified as Bicycle, Motorcycle, Non-motorized vehicle or Pedestrian classes of the MIO-TCD dataset. This is a sensible result because none of these classes exist or have equivalent classes in the BIT-vehicle dataset. Around 68% of the truck test samples were classified either as Articulated truck or Single-unit Truck (40.6% and 27.3% respectively). The remaining 30.9% were misclassified as Bus. This is an expected result due to the similarity among the 3 classes from the frontal view, knowing that some of the BIT-vehicle images show only the vehicle front (or a partial view of the vehicle front) without showing the body of the vehicle. On the other hand, 29.5% of the Bus BIT-vehicle test samples were correctly classified as Bus, while 54.5% were classified as either Articulated Truck or Single-Unit Truck. 46.3% of the Microbus class was classified as Bus, and 28.8% were classified as Work Van, and these are the 2 classes that are most similar to the Microbus class which doesn't exist in the MIO-TCD dataset. 45.8% of the Minivan class were classified as Single-Unit truck, which is the most similar one to the Minivan class. Only 6.8% of the SUV test images were correctly classified as the equivalent Car class but 30.8% of the SUV test images were classified as Work Van, which is a similarly-looking class. As for the Sedan, only 15.3% were correctly classified as Car. Fig. 6 shows examples of the classification results of “RXD-CV-CW-NCW” on the test set of the BIT-Vehicle dataset.

In the following experiment, we considered augmenting the MIO-TCD training set with some training samples of the Sedan, SUV, and Bus classes from the BIT-vehicle dataset. We chose these 3 BIT classes because they are the classes that can map with no doubt to equivalent classes in the MIO-TCD dataset, namely the MIO-TCD Car and Bus class. The Sedan, SUV, and Bus classes comprise 78.3% of the BIT-vehicle dataset. We fine-tuned the 3 base learners as well as the RXD-CV-CW-NCW using the augmented training set. Table 3 -B shows the evaluation metrics of the 3 base learners as well as the super learner after augmentation. We called the resulting super learner Augmented-RXD. Fig. 7 shows the

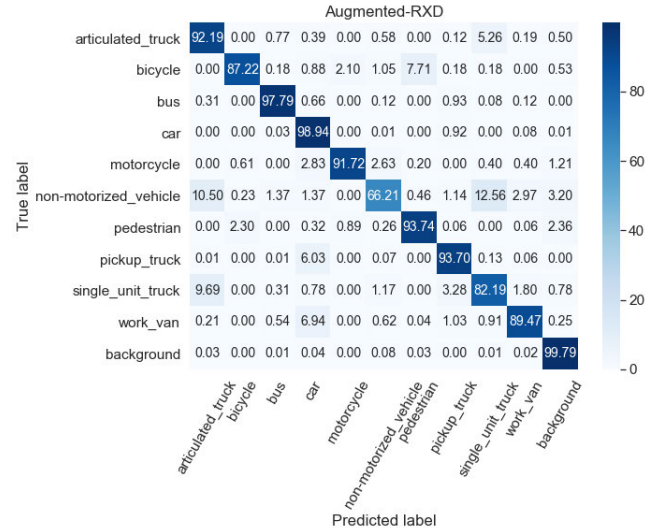


FIGURE 7. Confusion matrix of Augmented-RXD super learner.

TABLE 8. Confusion matrix of ‘Augmented-RXD super learner’ on the BIT-vehicle test set. Boldface indicates accurate classifications.

		Predicted						
		AT	Bus	Car	PT	SUT	WV	BG
True	Bus	0.0%	96.4 %	3.6%	0.0 %	0.0 %	0.0 %	0.0 %
	Microbus	0.0%	2.3%	97.2%	0.0 %	0.0 %	0.6 %	0.0 %
	Minivan	12.5 %	0.0%	85.4%	2.1 %	0.0 %	0.0 %	0.0 %
	Sedan	0.0%	0.0%	99.9 %	0.0 %	0.0 %	0.0 %	0.1 %
	SUV	0.0%	0.0%	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
	Truck	0.0 %	48.5 %	49.7%	1.8 %	0.0 %	0.0 %	0.0 %

TABLE 9. The evaluation metrics of the base learners vs. the super learners evaluated on the BIT-vehicle test set. Boldface indicates achievement of the best result.

		Cohen Kappa Score	Acc.	Mean Prec.	Mean Rec.
Base Learners	Resnet50	0.9341	0.9593	0.9295	0.9418
	Densenet121	0.9518	0.9702	0.9557	0.9533
	Xception	0.9574	0.9737	0.9551	0.9585
Proposed Super-Learner Ensembles	BIT-RXD	0.9574	0.9737	0.9557	0.9643
	BIT-XD	0.9615	0.9762	0.9624	0.9676

confusion matrix of the Augmented-RXD super learner as evaluated on the MIO-TCD testing set. Table 6 shows that this augmented super learner achieved as good metrics scores as the un-augmented super learner “RXD-CV-CW-NCW”, and both of them share the third rank together with the super learner of [23]. Compared to “RXD-CV-CW-NCW”, the mean precision of the Augmented-RXD super learner decreased by 0.83%, while the mean recall was increased

TABLE 10. The recall scores of the base learners vs. the proposed super learners evaluated on the BIT-vehicle test set. Boldface indicates achievement of the best result.

		Recall Scores					
		Bus	Microbus	Minivan	Sedan	SUV	Truck
Base Learners	Resnet50	0.9732	0.9096	0.9167	0.9789	0.9211	0.9515
	Densenet121	0.9821	0.9266	0.8854	0.9823	0.9677	0.9758
	Xception	1.0000	0.9379	0.8958	0.9857	0.9677	0.9636
Proposed Super-Learner Ensembles	BIT-RXD	1.0000	0.9435	0.9375	0.9831	0.9642	0.9576
	BIT-XD	0.9911	0.9435	0.9375	0.9831	0.9749	0.9758

by 0.23%. We again tested the augmented super learner on the BIT-vehicle testing set, and the confusion matrix is shown in Table 8.

After augmentation, the super learner was able to classify 100% of the Sedan and SUV testing images into the MIO-TCD’s Car class. 1,184 out of the 1,185 Sedan images were correctly classified. The remaining Sedan images were misclassified as Background. However, this incorrect prediction is in most due to erroneously annotated images that should be annotated as background. Furthermore, 96.4% of Bus images were classified correctly. However, since only the Sedan, SUV, and Bus classes were used for augmentation and fine-tuning, the augmented super learner seems to have learned that these high-resolution frontal-view images should only be one of these 3 classes. This may explain that out of the 2014 BIT test images 1803 images were classified as Car (89.5%), and 192 images were classified as Bus (9.5%). So, as for the MIO-TCD dataset, the data augmentation is not technically sound because it did not improve the performance on the MIO-TCD dataset. It just helped the model to generalize well to the Bus, Sedan, and SUV images of the BIT-vehicle dataset.

H. EXPERIMENTAL RESULTS ON THE BIT-VEHICLE DATASET

Finally, we performed a customized training on the 6 vehicle classes of the BIT-vehicle dataset. Consequently, the softmax output layers of the base learners and the super learner became 6-unit layers. Similarly, we revised the number of units of the ReLU fully connected layer between the concatenated outputs of the base learners and the softmax output layer of the super learner to be $(n \times 6)$, where n is the number of base learners.

In [6], the original paper of the BIT-vehicle dataset, Dong et al. used a large number of unlabeled vehicle images from the BIT-vehicle dataset to learn the filters of the network using unsupervised pre-training. Subsequently, they trained the softmax output layer with randomly selected 200 samples from each vehicle category. Also, they kept 200 samples from each vehicle category for testing.

Since we apply the supervised learning approach, we used most of the dataset images to train the base learners.

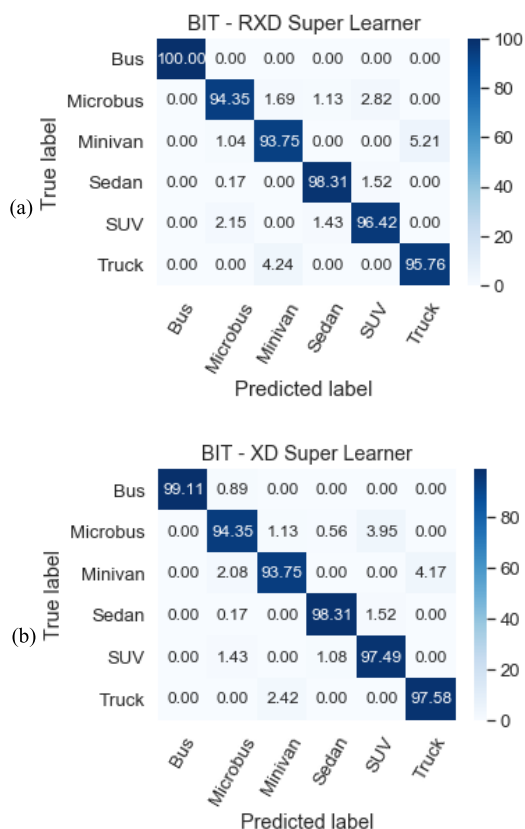


FIGURE 8. Confusion matrices of the (a) RXD and (b) XD super learners on the BIT-vehicle dataset.

We explained in IV-C how we split the BIT-Vehicle dataset into training, validation, and testing sets. So, we used the training set (60% of the dataset images) to train the base learners. Then, the validation set (20% of the dataset images) was used to train the super learner. The performance of the base learners and the super learner was evaluated using the testing set, which is the remaining 20% of the images.

Although the common evaluation metric of the BIT-vehicle dataset in literature is the accuracy, we evaluated the mean recall, mean precision, and Cohen Kappa scores to make the results more indicative and more comprehensive.

TABLE 11. The precision scores of the base learners vs. the proposed super learners evaluated on the BIT-vehicle test set. Boldface indicates achievement of the best result.

		Precision scores					
		Bus	Microbus	Minivan	Sedan	SUV	Truck
Base Learners	Resnet50	0.9909	0.8944	0.8224	0.9872	0.9245	0.9573
	Densenet121	1.0000	0.9371	0.9444	0.9940	0.9000	0.9583
	Xception	1.0000	0.9595	0.9053	0.9949	0.9247	0.9464
Proposed Super-Learner Ensembles	BIT-RXD	1.0000	0.9489	0.9000	0.9949	0.9212	0.9693
	BIT-XD	1.0000	0.9489	0.9375	0.9966	0.9158	0.9758



FIGURE 9. Examples from the classification results of BIT-RXD and BIT-XD for vehicle images that are (a) Correctly classified by BIT-RXD and misclassified by BIT-XD (b) Correctly classified by BIT-XD and misclassified by BIT-RXD (c) Incorrectly classified by both super learners (d) Correctly classified by both super learners (The Red caption is the ground truth, and the red caption indicates a misclassification).

We fine-tuned the Resnet50, Densenet121, and Xception base learners on the BIT-vehicle dataset until they reached testing accuracies of 95.93%, 97.02%, and 97.73%, respectively. In the first BIT super learner, we combined the outputs of the 3 base learners. We called this super learner ‘BIT-RXD’. As shown in Table 9, BIT-RXD achieved better mean precision and mean recall scores than the base learners. However, it had the same Cohen Kappa and Accuracy scores as that of the Xception model.

It was noticed that the scores of the Resnet base learner were relatively low compared to those of Densenet and Xception. Therefore, we trained another super learner which ensembles the outputs of Xception and Densenet only. We called it ‘BIT-XD’. As shown in Table 9, BIT-XD achieved better scores in the four metrics not only compared to the base learners but also compared to BIT-RXD. Tables 10 and 11 show the recall and precision scores of the base learners and super learners. Although the base learners

are better than the super learners in the recall or precision scores of some classes, the improvement that the super learners achieved in the recall and precision scores of the remaining classes resulted in a better mean recall and mean precision scores.

The confusion matrices of BIT-RXD and BIT-XD are shown in Fig. 8. Examples of the classification results of BIT-RXD and BIT-XD are shown in Fig. 9. Figure 9 shows clearly the inter-class similarities between the categories of the BIT-vehicle dataset and how each of the super learners dealt with them. The BIT-RXD achieved 100% accuracy in the Bus class. The BIT-XD misclassified only 1 bus image as microbus as shown in the top left image of Figure 9. BIT-XD achieved equal or better accuracies in the remaining classes.

V. CONCLUSION

In this paper, a super-learner ensemble of deep networks for vehicle classification in traffic surveillance images was proposed. We introduced a densely connected single-split super learner and applied variants from it to two of the most challenging and largest publicly available traffic surveillance datasets, the MIO-TCD dataset and the BIT-vehicle dataset. While our method is simple, easy to train, does not include any handcrafted features or any logic reasoning, it achieved fantastic results that compare to those of the state-of-the-art methods that were designed for the two datasets. Three variants of the super learner ensemble: RXD-CV-CW, RXD-CV-CW-NCW and Augmented-RXD, were examined on the MIO-TCD dataset with variations in applying class weights and data augmentation during training. RXD-CV-CW-NCW and Augmented-RXD share the third place among the published state-of-the-art methods reported in the MIO-TCD classification challenge. The applied data augmentation did not yield a significant performance improvement on the MIO-TCD dataset. However, it helped the network to generalize well to the Bus, Sedan and SUV images of the BIT-vehicle dataset.

In addition, the super-learner variants that we trained on the BIT-Vehicle dataset images performed very well and achieved overall accuracies of up to 97.62%.

In our future work, we will consider extending our work to the MIO-TCD localization challenge as well.

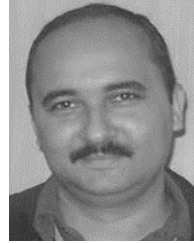
REFERENCES

- [1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
- [2] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
- [3] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2110–2118.
- [4] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
- [5] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3973–3981.
- [6] Z. Dong, Y. Wu, M. Pei, and Y. Jia, "Vehicle type classification using a semisupervised convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2247–2256, Aug. 2015.
- [7] Z. Luo, F. Branchaud-Charron, C. Lemaire, J. Konrad, S. Li, A. Mishra, A. Achkar, J. Eichel, and P.-M. Jodoin, "MIO-TCD: A new benchmark dataset for vehicle classification and localization," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5129–5141, Oct. 2018.
- [8] R. Theagarajan, F. Pala, and B. Bhanu, "EDeN: Ensemble of deep networks for vehicle classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 906–913.
- [9] C. Ju, A. Bibaut, and M. van der Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *J. Appl. Statist.*, vol. 45, no. 15, pp. 2800–2818, Nov. 2018.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Arxiv.Org*, vol. 7, no. 3, pp. 171–180, 2015.
- [11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *SAE Int. J. Mater. Manuf.*, vol. 7, no. 3, pp. 01–2014-0975, 2014.
- [12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*. [Online]. Available: <https://arxiv.org/abs/1608.06993>
- [13] H. Cho, Y.-W. Seo, B. V. K. V. Kumar, and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1836–1843.
- [14] N. S. Thakoor and B. Bhanu, "Structural signatures for passenger vehicle classification in video," *IEEE Trans. Intell. Transport. Syst.*, vol. 14, no. 4, pp. 1796–1805, Dec. 2013.
- [15] M. Kafai and B. Bhanu, "Dynamic Bayesian networks for vehicle classification in video," *IEEE Trans. Ind. Informat.*, vol. 8, no. 1, pp. 100–109, Feb. 2012.
- [16] R. Theagarajan, N. S. Thakoor, and B. Bhanu, "Robust visual rear ground clearance estimation and classification of a passenger vehicle," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst.*, Nov. 2016, pp. 2539–2544.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.
- [18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [19] K. F. Hussain, M. Affi, and G. Moussa, "A comprehensive study of the effect of spatial resolution and color of digital images on vehicle classification," *IEEE Trans. Intell. Transport. Syst.*, vol. 20, no. 3, pp. 1181–1190, Mar. 2019.
- [20] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, F. Mujica, A. Coates, and A. Y. Ng, "An empirical evaluation of deep learning on highway driving," 2015, *arXiv:1504.01716*. [Online]. Available: <http://arxiv.org/abs/1504.01716>
- [21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*. [Online]. Available: <http://arxiv.org/abs/1312.6229>
- [22] S. Wang, Z. Li, H. Zhang, Y. Ji, and Y. Li, "Classifying vehicles with convolutional neural network and feature encoding," in *Proc. IEEE 14th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2016, pp. 784–787.
- [23] P.-K. Kim and K.-T. Lim, "Vehicle type classification using bagging and convolutional neural network on multi view surveillance image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 914–919.
- [24] J. T. Lee and Y. Chung, "Deep learning-based vehicle classification using an ensemble of local expert and global networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 920–925.
- [25] H. Jung, M.-K. Choi, J. Jung, J.-H. Lee, S. Kwon, and W. Y. Jung, "ResNet-based vehicle classification and localization in traffic surveillance systems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 934–940.
- [26] W. Liu, M. Zhang, Z. Luo, and Y. Cai, "An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors," *IEEE Access*, vol. 5, pp. 24417–24425, 2017.
- [27] W. Liu, Z. Luo, and S. Li, "Improving deep ensemble vehicle classification by using selected adversarial samples," *Knowl.-Based Syst.*, vol. 160, no. 422, pp. 167–175, Nov. 2018.

- [28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [29] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.
- [30] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Stat. Appl. Genet. Mol. Biol.*, vol. 6, no. 1, Jan. 2007.
- [31] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan, "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, Dec. 2005.
- [32] S. E. Sinisi, E. C. Polley, M. L. Petersen, S. Y. Rhee, and M. J. Van Der Laan, "Super-learning: An application to the prediction of HIV-1 drug resistance," *Stat. Appl. Genet. Mol. Biol.*, vol. 6, no. 1, pp. 1–24, 2007.
- [33] R. Pirracchio, M. L. Petersen, M. Carone, M. R. Rigon, S. Chevret, and M. J. van der Laan, "Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): A population-based study," *Lancet Respiratory Med.*, vol. 3, no. 1, pp. 42–52, Jan. 2015.
- [34] C. Ju, M. Combs, S. D. Lendle, J. M. Franklin, R. Wyss, S. Schneeweiss, and M. J. van der Laan, "Propensity score prediction for electronic health-care databases using super learner and high-dimensional propensity score methods," *J. Appl. Statist.*, vol. 46, no. 12, pp. 2216–2236, Sep. 2019.
- [35] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.
- [36] A. Stein, J. Aryal, and G. Gort, "Use of the bradley-terry model to quantify association in remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 852–856, Apr. 2005.
- [37] R. G. Pontius and M. Millones, "Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment," *Int. J. Remote Sens.*, vol. 32, no. 15, pp. 4407–4429, Aug. 2011.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
- [39] T. Akilan, Q. M. J. Wu, A. Safaei, and W. Jiang, "A late fusion approach for harnessing multi-CNN model high-level features," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 566–571.
- [40] T. Akilan, Q. J. Wu, and H. Zhang, "Effect of fusing features from multiple DCNN architectures in image classification," *IET Image Process.*, vol. 12, no. 7, pp. 1102–1110, Jul. 2018.



MOHAMED A. HEDEYA received the B.S. degree in computer engineering from Suez-Canal University, in 2002. He is currently pursuing the master's degree with the Electrical Engineering Department, Faculty of Engineering, Port-Said University, Port-Fouad, Egypt. His research interests include artificial intelligence and computer vision.



AHMAD H. EID received the B.S. and M.Sc. degrees in computer engineering from Suez-Canal University, in 1999 and 2004, respectively, and the Ph.D. degree from Port-Said University, Port-Fouad, Egypt, in 2010. He is currently an Assistant Professor with the Electrical Engineering Department, Faculty of Engineering, Port-Said University. His research interests include ensemble classifier systems and image processing.



REHAB F. ABDEL-KADER received the B.S. degree in computer engineering from Suez-Canal University, in 1996, the M.Sc. degree in electrical engineering from Tuskegee University, in 1999, and the Ph.D. degree from Auburn University in 2003. She is currently a Professor with the Electrical Engineering Department, Faculty of Engineering, Port Said University, Port-Fouad, Egypt. Her current research interests include artificial intelligence and computer vision.

• • •