IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

received September 18, 2017, accepted October 16, 2017, date of publication October 25, 2017,
date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2766203

An Ensemble Deep Learning Method for Vehicle Type Classification on Visual Traffic Surveillance Sensors

WEI LIU[1,2], MIAOHUI ZHANG[3], ZHIMING LUO[4], AND YUANZHENG CAI[5]

[1]Virtual Reality and Interactive Techniques Institute, East China Jiaotong University, Nanchang 330013, China
[2]Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, Yi Chang 443002, China
[3]Institute of Energy, Jiangxi Academy of Sciences, Nanchang 330096, China
[4]Cognitive Science Department, Xiamen University, Xiamen 361005, China
[5]Department of Computer Science, Minjiang University, Fuzhou 350108, China

Corresponding author: Miaohui Zhang (zhangmiaohui@jxas.ac.cn).

This work was supported in part by the Nature Science Foundation of China under Grant 61662024 and Grant 61602220, the Natural Science Foundation of Jiangxi Province under Grant 20171BAB212013, Grant 20161BBI90004, and Grant 20161BAB212057, and in part by the Open Fund of the Hubei Province Key Laboratory under Grant 2016KLA03.

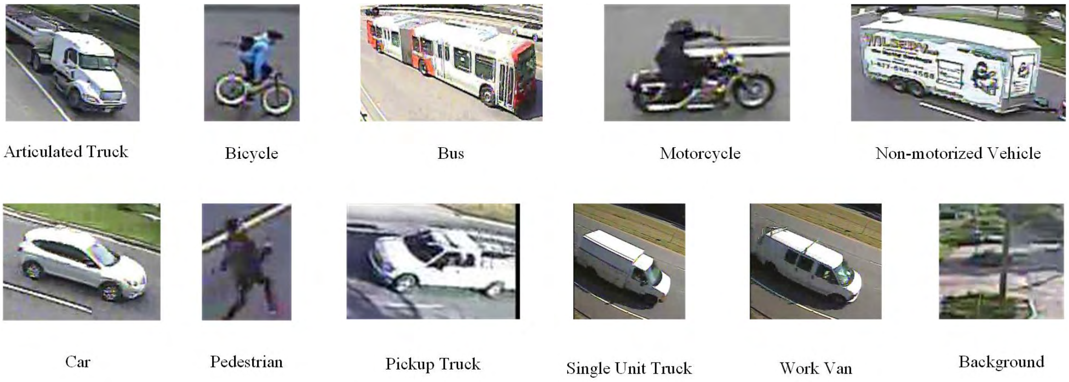
ABSTRACT Visual traffic surveillance systems play important roles in intelligent transport systems nowadays. The first step of a visual traffic surveillance system usually needs to correctly detect objects from images or videos and classify them into different categories (e.g., car, truck, and bus). This paper aims to introduce a new vehicle type classification scheme on the images acquired from multi-view visual traffic surveillance sensors. Most image classification algorithms focus on maximizing the percentage of the correct predictions, which have a deficiency that the images from minority categories are prone to be misclassified as the dominant categories. To address this challenge of classifying imbalanced data acquired from visual traffic surveillance sensors, we propose a method, which integrates deep neural networks with balanced sampling in this paper. The proposed method consists of two main stages. In the first stage, data augmentation with balanced sampling is applied to alleviate the unbalanced data set problem. In the second stage, an ensemble of convolutional neural network models with different architectures is constructed with parameters learned on the augmented training data set. Experiments on the MIOvision traffic camera dataset classification challenge data set demonstrate that the proposed method is able to enhance the mean precision of all categories, in the condition of high overall accuracy, compared with the baseline algorithms.

INDEX TERMS Traffic data, traffic surveillance systems, intelligent transport systems, image classification, ensemble learning, imbalanced data.

I. INTRODUCTION

In the last decade, we have seen a worldwide rise of using visual traffic surveillance systems, due to the rapidly growth of storage power, computation speed and the innovations in video compression standards. For the first step, a visual traffic surveillance system usually needs to correctly detect objects from images or videos and classify them into different categories *(e.g. car, truck, bus)*. Efficient and robust classification can lead to many semantic results, such as "pedestrian no.1 is moving, car no.3 stopped" or some more advanced results such as "van no.8 is turning right, bicycle no.5 is moving at a speed of 10 kilometers per hour." However, such high-level information is possible only if we can correctly detect and classify the objects.

With the increasing amount of available data, image processing has emerged to be a hot spot in the field of artificial intelligence and image classification is one of fundamental tasks. As is shown in Figure 1, the goal of image classification is to assign a predefined category label to an image. Image classification has a wide application in the field of artificial intelligence, including self-driving, augment reality, etc [1]–[3]. Recently, image classification have attracted more and more research interest. Though image classification has been widely studied in the academia and deployed in the industry, it is not a trivial task, still a challenging task. For example, many practical image classification data are imbalanced, i.e., some of the categories are represented by only a few samples, while some others make up the majority.


Articulated Truck

Bicycle

Bus

Motorcycle

Non-motorized Vehicle

Car

Pedestrian

Pickup Truck

Single Unit Truck

Work Van

Background

FIGURE 1. MIO-TCD classification challenge dataset acquired from visual traffic surveillance sensors.

In the field of traffic surveillance, a visual traffic surveillance system needs to detect vehicles or pedestrians and classify them if possible. In the practical application, *Pedestrains*, *Bicycles* and *Motorcycles* often make up minority of the data set, in contrast with *Cars* and *Buses*. Consequently, to avoid the misclassification of images from majority categories as rare classes, it is also not appropriate to assume misclassification errors cost for all samples are equal. If misclassification errors cost are implicitly assumed to be equal, images from minority categories are prone to be misclassified to be the dominant categories. Therefore, to effectively reduce the number of fatalities, it is reasonable to focus on enhancing the mean precision of all categories, in the condition of high overall accuracy.

In the field of machine learning, a lot of learning methods [4]–[9] and data manipulation techniques [10]–[14] have been proposed for dealing with imbalanced data over the last two decades. The approaches to tackle the problem of extremely imbalanced data can be mainly categorized into two broad types. One is based on cost sensitive learning [15] which assigning a high misclassification cost of the minority classes and then minimize the overall train loss. The other way is to employ a sampling tactic mainly includes oversampling the minority class, undersampling the majority classes, and synthesizing new minority classes. Most researches have been focused on the approach based on balanced sampling.

To tackle the imbalanced problem for traffic data acquired from visual traffic surveillance sensors, we propose an convolutional neural networks (CNN) based deep learning framework which can increase the mean precision in this paper. We focus on integrating deep neural networks with balanced sampling. As is shown in Fig.2, the proposed approach consists of two stages. In the first stage, data augmentation with balanced sampling is applied to alleviate the unbalanced data set problem. In the second stage, an ensemble of convolutional neural network models with different architectures is constructed with parameters learned on the augmented training data set.

The outline of this paper is organized as follows. Section II surveys related work . The detailed of the proposed method is presented in Section III. Experimental results and comparison are provided in Section IV. Finally, the conclusion of this paper is in Section V.

II. RELATED WORK

A. IMBALANCED DATA CLASSIFICATION

In recent years, it has a spate of interest in learning from imbalanced data in data mining and machine learning. A vast number of techniques have been tried, and can be mainly categorized into the algorithm oriented approaches [4]–[9], [16] and data manipulation techniques [10]–[14]. The former category mostly modify the training algorithms by adjusting misclassification costs, and the latter category operates at the data level by using data re-sampling. A comprehensive review is presented in [13].

There are several types of data manipulation techniques, which can be mainly divided into two groups: oversampling and undersampling. The easiest way for resampling is to randomly replicates minority instances to increase their population, or to randomly downsample the majority class. Different re-sampling algorithms have been widely studied and tested to counter the effect of imbalanced data sets in the last two decades [17], [18]. To make the data set balanced, minority instances are generated by certain algorithms in oversampling. The positive consequence for replication-based random oversampling is that it duplicates the number of errors for minority instances. But replicating-based random oversampling makes variables appear to have lower variance which has a tendency to overfit and does not increase any information actually. To address this issue, Chawla *et al.* [10] proposed synthetic minority over-sampling technique(SMOTE), generating new non-replicated minority examples, and several improved version can be found in [12], [13], and [19].

Downsampling is to throw away part of majority samples to balance the dataset which is very efficient. The main disadvantage is that potentially valuable information may be removed by dropping part of the majority samples, but it is often preferred to use undersampling than oversampling [11].

To avoid the disadvantages of re-sampling, there are many studies focusing on algorithm oriented approaches.

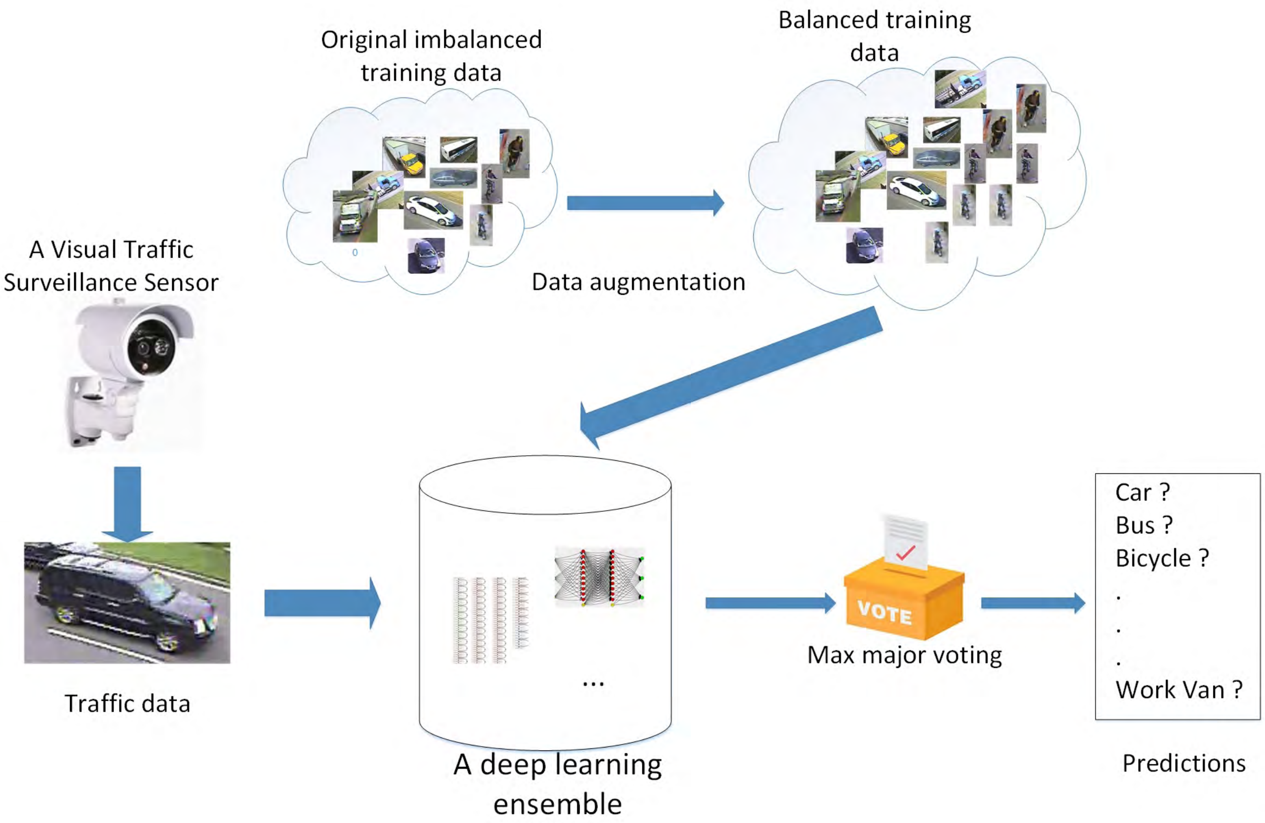


FIGURE 2. **The framework of our deep CNN ensemble model.**

Liu *et al.* [20] proposed two algorithms named *EasyEnsemble* and *BalanceCascade* respectively. Tang et.al [9] introduce the cost-sensitive into the classical SVM to improve performance of highly skewed datasets. Zadrozny et.al [4] proposed a family of methods for converting classifier learning algorithms and classification theory into cost-sensitive algorithms and theory, based on cost-proportionate weighting of the training examples. This method achieves better predictive performance, while drastically reducing the computation required by other baseline methods. To combat imbalance, Ting [16] studied how to improve our understanding of various cost-sensitive boosting algorithms and how variations in the boosting procedure affect misclassification cost and high cost error. Chen *et al.* [8] proposed two methods to to overcome the imbalanced data classification problem utilizing random forest. One is based on cost sensitive learning, and the other is based on a sampling technique. The two proposed methods are less vulnerable to noise than boosting.

B. DEEP LEARNING

In the past few years, deep neural networks have led to a series of breakthroughs on a variety of tasks, such as computer vision, machine translation and voice recognition, etc. One of the essential components bringing about these breakthrough results is convolutional neural networks. After the AlexNet proposed by Krizhevsky *et al.* [21], CNNs have shown superior performance for image classification compared with conventional ''shallow learning'' methods, and have also been successfully applied for object detection [22], video classification [23] and segmentation [24], etc. These successes spurred a new line of research that focused on designed higher performance CNNs, and the performance of network architectures has been significantly improved by utilizing deeper and wider structures. Simonyan *et al.* proposed VGGNet [25], facilitating the research on the use of deep architecture in computer vision. Szegedy *et al.* [26] presented GoogLeNet which contains Inception modules, setting the new state of the art for the ImageNet Challenge 2014.[1] To tackle this degradation problem, He *et al.* [27] presented a residual learning framework named ResNet that can train substantially deeper networks than those employed previously.

Although deep learning has been successful for a variety of tasks, only a few works [6], [7], [28]–[30] have addressed with the problem of imbalanced classification utilizing deep learning. Most of them rely on shallow models and handcrafted features. Khan *et al.* [29] proposed a cost-sensitive deep neural network to automatically learn robust features for both the dominant and rare classes. Jeatrakul *et al.* [28] proposed a method combined Synthetic Minority Oversampling Technique (SMOTE) and Complementary Neural

[1] http://image-net.org/challenges/LSVRC/2014/

Network (CMTNN) to handle the imbalance data. To learn discriminative representation, Huang *et al.* [7] proposed a deep learning framework through quintuplet instance sampling and the associated triple-header hinge loss. Yan *et al.* [30] proposed a learning framework to improve multimedia data classification, in which CNNs are integrated with a bootstrapping sampling algorithm.

Most of these methods can be treated as extensions of using traditional algorithms to handle imbalanced data classification. In this paper, we focus on tackling the problem of imbalanced data classification based on ensemble learning, combined with deep learning.

C. ENSEMBLE LEARNING

Ensemble Learning is another hot topic in machine learning, which utilize a set of learning algorithms to obtain better classification results than the constituent learning algorithms alone. Multiple classifiers are employed to learn the original dataset respectively during the training period, and then will be combined together to classify the unknown data. The single classifier tend to cause the bias in terms of a fixed set of parameters, and reduction of such bias can be obtained through the ensemble learning. The performance of ensemble learning depends on the precision of the constituent classifiers, which usually has stronger generalization ability than those base classifiers. A comprehensive review of ensemble learning can be found in [31].

Ensemble Learning can be mainly categorized into three types as follows:

- **Bagging** is the abbreviation of "bootstrap aggregating," which was proposed by Breiman [32] to improve the classification by combining prediction results of models trained independently on randomly generated training sets. The random forest [33] algorithm is a example of bagging, which combines a collection of random decision trees to achieve high classification accuracy.
- **Boosting** is an ensemble meta-algorithm which combine a set of weak classifiers to create a strong classifier. It incrementally build an ensemble by iteratively training a new model to emphasize those misclassified training samples from previous models. Although many newer algorithms have been proposed to yield better results, the Adaboost [34] still is the most widely implementation of boosting.
- **Bucket of models** is an ensemble learning technique in which a model selection algorithm is utilized to choose the best model in a set for different problems. The most common approach of model selection is through cross-validation.

III. THE PROPOSED SCHEME

A. DEEP ENSEMBLE MODEL

In this section, we presented the proposed deep learning framework for vehicle type classification on visual traffic surveillance sensors and the whole framework is showed in Figure 2. First, a balanced sampling data augmentation strategy is used to increase the number of samples of rare classes in the original dataset, which can reduce classification bias and use as much data as possible for training. Then, a set of convolutional neural networks models are trained on the balanced data set, all started from a good initialization *(pretrained on ImageNet)*. Finally, outputs of multiple models are combined together by maximum voting policy according to the predictions of single models. The details of the framework is presented as follows.

B. DATA AUGMENTATION WITH BALANCED SAMPLING

Data preparation is required when working with classification tasks such as vehicle classification based on neural network and deep learning models. Increasingly data augmentation for training data is also required on more complex deep learning models for vehicle classification. Therefore, we augment the training data set by data augmentation techniques for Deep Learning, including random rotations, shifts, flips, and cropping.

Although the regular augmentation techniques such as random cropping and rotations can enrich the training data, the extreme imbalanced data distribution are not changed essentially. To ease the problems caused by extreme imbalanced data distribution, we devised an over sampling scheme with random shuffling. That is, the size of the minority class is increased randomly by over-sampling. To avoid over-fitting, the size of the minority class is increased to a small number, compared with the size of the majority class in practical application. The details of the proposed balanced sampling scheme is shown in Fig.3, let $Tr = n$ be a threshold that denotes the size each rare class will be increased to. For the rare class c_i, firstly a random permutation $S = (S[0], S[1], \cdots, S[n])$is generated. Then, we get the actual identity for $S[j]$ with the following equation:

$$ind = S[j]\, mod\; n, \tag{1}$$

where *mod* denotes the modulo operation. After selecting samples by a random permutation and modulo operations, we get the expanded $\mathcal{D}_i^{'}$ for the rare class c_i based on the original data set $\mathcal{D}_i$. At last, the samples of all rare classes and the other classes are concatenated and reshuffled. The details of the proposed balanced sampling scheme is presented in Algorithm 1.

The threshold Tr is set to be 10,000 in this paper by cross-validation on the training set of MIOvision traffic camera dataset (MIO-TCD) classification challenge dataset.

C. REVISITING RESNETS

In this subsection, we briefly introduce the ResNets used in this paper. To ease the training of framework that are substantially deeper than those employed previously, He *et al.* [27] proposed a residual learning framework named ResNets. Deep residual networks consist of many stacked Residual Units as shown in Figure 4. Each unit can be expressed in

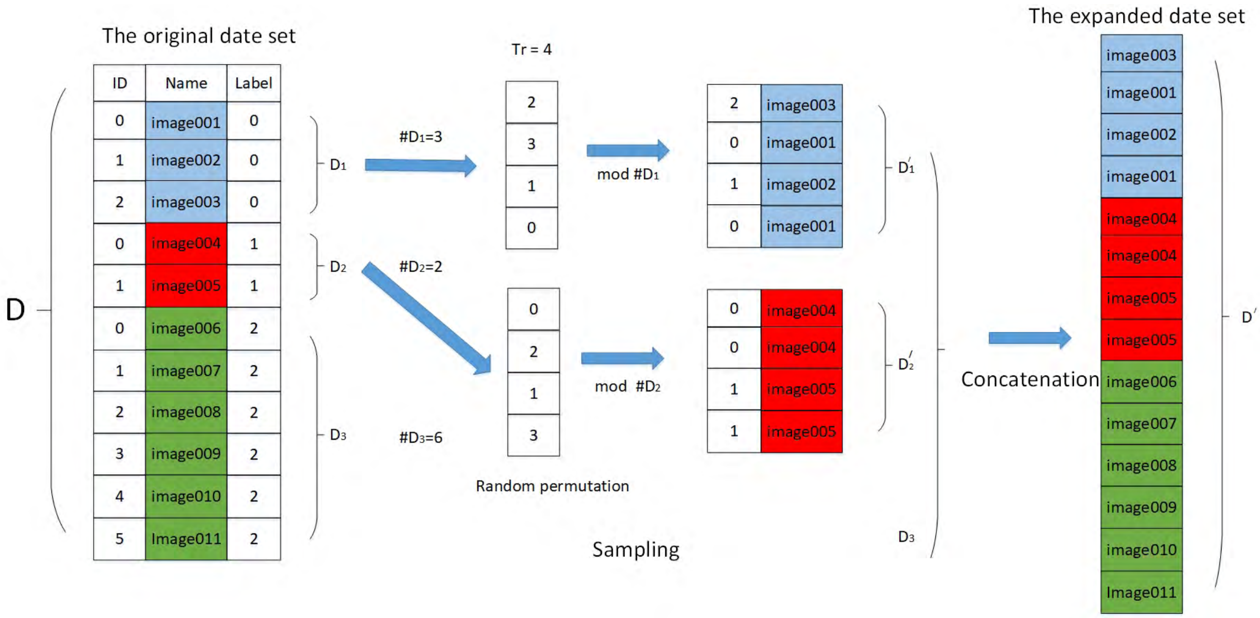


FIGURE 3. Balanced sampling.

```
Algorithm 1 Balanced Sampling
  Input:
  an original imbalanced data set $\mathcal{D}$;
  rare classes $\mathcal{C} = \{c_1, c_2, \cdots, c_m\}$;
  train data of rare classes $\mathcal{D}_r = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_m\}$;
  train data of rare classes after sampling $\mathcal{D}_r^{`} = \emptyset$;
  an threshold $n$;
  Output:
  an training data set after balanced sampling $\mathcal{D}^{`}$;
  for $i = 0$ to $m$ do
    $\mathcal{D}_i^{`} = \emptyset$
    $s = size(\mathcal{D}_i)$
    $\mathbf{s}$ = randperm($n$)
    for $j = 0$ to $n$ do
      $ind = \mathbf{s}[j] \bmod s$
      $\mathcal{D}_i^{`} = Concat(\mathcal{D}_i^{`}, \mathcal{D}_i(ind))$
    end for
    $\mathcal{D}_r^{`} = Concat(\mathcal{D}_r^{`}, \mathcal{D}_r^{`})$
  end for
  $\mathcal{D}^{`} = (\mathcal{D} - \mathcal{D}_r) \bigcup \mathcal{D}_r^{`}$
```

a general form [35]:

$$y_l = h(x_l) + F(x_l, W_l),$$
$$x_{l+1} = f(y_l), \tag{2}$$

where x_l and x_{l+1} are input and output of the l-th unit, and F is a residual function. In [27], f is a ReLU [36] function, and $h(x_l) = x_l$ is an identity mapping.

For ResNets with units in Figure 4(b) is much easier to train and has a better generalization than the original ResNets in [27], we use ResNets in [35] in this paper. All of the ResNets in this paper are started from a good initialization (pretrained on ImageNet).

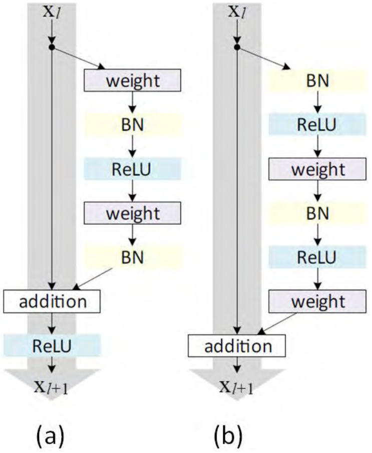


FIGURE 4. (a) Residual Unit in [27]; (b) Residual Unit in [35].

To tackle the imbalanced classification problem for Vehicle Type Classification, we proposed a deep CNN ensemble model. The deep CNN ensemble contains ResNet-50, ResNet-101 and ResNet-152. As is shown in 2, the proposed ensemble model comprises three key stages: starting CNN models from good initial parameters, fine tuning network parameters and averaging models. Concretely, firstly all of CNN models in the ensemble are pretrained on ImageNet. Next, the network parameters are refined using MIO-TCD data set enhanced by data augmentation. Finally, the outputs of refined CNN models are combined together by averaging their predictions.

As mentioned above, the proposed ensemble model contains multiple deep learning models. Therefore, the initial stage of the ensemble system generates many results for a single image to classify. The voting process is necessary to decide which class the image belongs to based on votes. In this paper, maximum majority voting is adopted to classify images based on initial predictions of single models. For the number of models in our ensemble is odd, consequently it doesn't have to be considered for cases with same votes in this paper.

IV. EXPERIMENTS AND RESULTS

A. DETAILS OF THE MIO-TCD CLASSIFICATION CHALLENGE DATASET

To demonstrate the effectiveness of our proposed framework, we use the MIO-TCD classification challenge dataset[2] for testing, which is a large benchmark traffic camera data set with a highly imbalanced data distribution. The dataset consists 648,959 samples in the classification dataset acquired at different times of the day and different periods of the year by traffic cameras deployed all over Canada and the United States. Those images have been selected to cover a wide range of challenges and are representative of typical visual data captured in urban traffic scenarios.

TABLE 1. Number of training samples for each category in MIO-TCD dataset.

Category	#
Articulated truck	10,346
Background	160,000
Bicycle	2,284
Bus	10,316
car	260,518
Motorcycle	1,982
Non-motorized vehicle	1,751
Pedestrian	6,262
Pickup truck	50,906
Single unit truck	5,120
Work van	9,679
Total	519,164

The classification challenge dataset contains 648,959 images divided into 11 categories, including *Articulated truck*, *Background*, *Bicycle*, *Bus*, *Car*, *Motorcycle*, *Non-motorized vehicle*, *Pedestrian*, *Pickup truck*, *Non-motorized vehicle*, *Single unit truck* and *Work van*. The size of training samples is 519,164. The number of training samples for each category is given in Table 1. As is shown in Table 1, number of samples for each category in MIO-TCD Dataset is in a range between 1,751 and 260,518. *Bicycle*, *Motorcycle* and *Vehicle* categories only contain a small number of training samples, while *Background* and *Car* make up the majority.

B. BASELINES

To indicate the effect of the proposed scheme, the state of art deep learning methods including ResNet-50, ResNet-101 and ResNet-152 [27] are used. The ResNet-50 trained with balance sampling is denoted as ResNet-50-BS, by analogy to ResNet-101-BS and ResNet-152-BS. We name the proposed method with DCEM-BS.

C. EVALUATION CRITERION

The prime goal of this paper is to introduce a new vehicle type classification scheme on the images acquired from multi-view Visual Traffic Surveillance Sensors. Let *TP* denote true positive, let *TN* denote true negative, let *FP* denote false positive, and let *FN* denote false negative. In order to objectively evaluate the performance of the introduced method and the baselines, we evaluate our approach by the following 6 metrics.

- **Precision of each category**

$$Pre_i = \frac{TP_i}{TP_i + FP_i}$$

- **Recall of each category**

$$Rec_i = \frac{TP_i}{TP_i + FN_i}$$

- **Accuracy**

$$Acc = \frac{TP}{\#ofTestingImages}$$

- **Mean Recall**

$$mRe = mean(Rec_i)$$

- **Mean Precision**

$$mPre = mean(Pre_i)$$

- **Cohen Kappa Score**

$$k = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e is the expected agreement when both annotators assign labels randomly [37].

The comparison experiment results are presented as follows.

D. RESULTS

We tested the proposed vehicle classification scheme and baselines with a TITAN X Pascal GPU on the deep learning framework MXNet. All of the pre-trained Resnets were downloaded from the MXNet Model Zoo,[3] a collection of pre-trained models ready for use.

Table 2 presents comparisons of precision for each category on the MIO-TCD classification challenge dataset. Table 2 indicates that the proposed method *DCEM-BS* got the best performance in term of precision for each category as a whole, in comparisons with the baselines. Moreover, networks with balanced sampling got better performances than the others. Table 2 shows that both ensemble learning

[2] http://tcd.miovision.com/challenge/dataset/

[3] https://mxnet.incubator.apache.org/model_zoo/index.html

TABLE 2. Comparisons of precision for each category on the MIO-TCD dataset. AT denotes articulated truck, MC denotes motorcycle, NV denotes non-motorized vehicle, PT denotes pickup truck, SUT denotes single unit truck, WV denotes work Van, and BG denotes background.

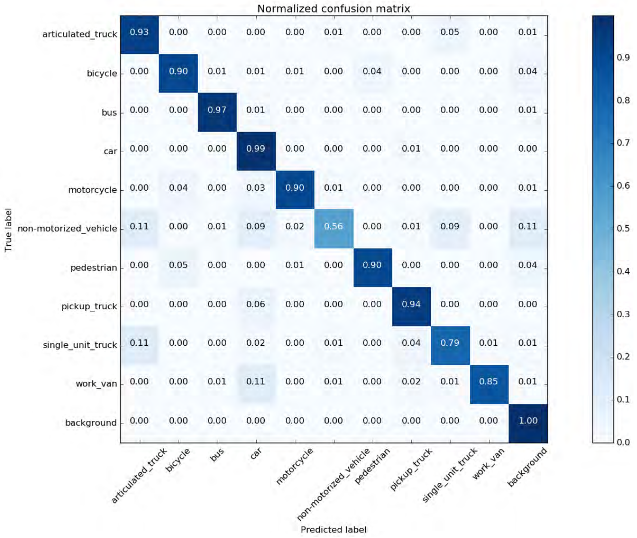
Model	AT	BG	Bicycle	Bus	Car	MC	NV	Pedestrian	PT	SUT	WV
ResNet-50	0.8748	0.9903	0.8135	0.9712	0.9718	0.8971	0.6516	0.9007	0.8644	0.7158	0.9013
ResNet-50-BS	0.8976	0.9930	0.8336	0.9754	0.9717	0.9180	0.5211	0.9048	0.8845	0.6814	0.9312
ResNet-101	0.8986	0.9923	0.8401	0.9772	0.9828	0.9320	0.7387	0.9313	0.8915	0.7450	0.9283
ResNet-101-BS	**0.9314**	0.9926	**0.8632**	0.9809	0.9806	0.9421	0.6466	0.9469	0.9089	0.7271	0.9315
ResNet-152	0.9050	0.9935	0.8471	0.9781	**0.9835**	0.9100	0.7390	0.9336	0.8870	0.7624	0.9368
ResNet-152-BS	0.9146	**0.9939**	0.8424	**0.9850**	0.9813	0.9287	0.6435	0.9421	0.9118	0.7403	0.9384
DCEM(ours)	0.8936	0.9923	0.8581	0.9765	0.9820	0.9439	**0.7812**	0.9392	0.9124	0.7844	0.9526
DCEM-BS(ours)	0.9115	0.9929	0.8269	0.9811	0.9830	**0.9550**	0.7538	**0.9718**	**0.9445**	**0.8328**	**0.9679**

TABLE 3. Comparisons of recall for each category on the MIO-TCD dataset.

Model	AT	BG	Bicycle	Bus	Car	MC	NV	Pedestrian	PT	SUT	WV
ResNet-50	0.8829	0.9939	0.8021	0.9411	0.9722	0.8808	0.3288	0.8754	0.8973	0.7063	0.7878
ResNet-50-BS	0.8469	0.9924	0.8861	0.9383	0.9759	**0.9272**	0.5913	0.9048	0.8937	0.7805	0.7655
ResNet-101	0.9076	0.9961	0.8739	0.9636	0.9761	0.9131	0.4840	0.9093	0.9364	0.7852	0.8390
ResNet-101-BS	0.8713	0.9962	0.8949	0.9581	0.9799	0.9212	0.5890	**0.9233**	0.9281	**0.8266**	0.8369
ResNet-152	0.9026	0.9956	0.8827	0.9686	0.9755	0.9192	0.4977	0.9157	**0.9408**	0.7898	**0.8625**
ResNet-152-BS	0.8740	0.9955	0.8984	0.9647	0.9809	0.9212	**0.6142**	0.9157	0.9287	0.8219	0.8551
DCEM(ours)	0.9192	0.9968	0.8687	0.9655	0.9816	0.9172	0.4566	0.9176	0.9346	0.7758	0.8456
DCEM-BS(ours)	**0.9312**	**0.9984**	**0.9037**	**0.9663**	**0.9889**	0.9010	0.5594	0.9022	0.9402	0.7898	0.8468

TABLE 4. The overall results on the MIO-TCD dataset.

Model	Mean Recall	Precision	Mean Precision	Cohen Kappa Score
ResNet-50	0.8244	0.9586	0.8684	0.9354
ResNet-50-BS	0.8639	0.9610	0.8648	0.9392
ResNet-101	0.8713	0.9691	0.8713	0.9520
ResNet-101-BS	0.8841	0.9705	0.8956	0.9540
ResNet-152	0.8773	0.9698	0.8978	0.9531
ResNet-152-BS	**0.8882**	0.9713	0.8929	0.9553
DCEM(ours)	0.8708	0.9723	0.9106	0.9568
DCEM-BS(ours)	0.8844	**0.9776**	**0.9201**	**0.9651**



	articulated_truck	bicycle	bus	car	motorcycle	non-motorized_vehicle	pedestrian	pickup_truck	single_unit_truck	work_van	background
articulated_truck	0.93	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.05	0.00	0.01
bicycle	0.00	0.90	0.01	0.01	0.01	0.00	0.04	0.00	0.00	0.00	0.04
bus	0.00	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01
car	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.01	0.00	0.00	0.00
motorcycle	0.00	0.04	0.00	0.03	0.90	0.01	0.00	0.00	0.00	0.00	0.01
non-motorized_vehicle	0.11	0.00	0.01	0.09	0.02	0.56	0.00	0.01	0.09	0.00	0.11
pedestrian	0.00	0.05	0.00	0.00	0.01	0.00	0.90	0.00	0.00	0.00	0.04
pickup_truck	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.94	0.00	0.00	0.00
single_unit_truck	0.11	0.00	0.00	0.02	0.00	0.01	0.00	0.04	0.79	0.01	0.01
work_van	0.00	0.00	0.01	0.11	0.00	0.01	0.00	0.02	0.01	0.85	0.01
background	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

FIGURE 5. The vehicle classification confusion matrix of the proposed scheme on the MIO-TCD classification challenge dataset.

and balanced sampling are effective to improve precision of vehicle classification.

Table 3 presents comparisons of precision for each category on the MIO-TCD classification challenge dataset. It indicates that balanced sampling is able to improve the recall for vehicle classification obviously, particularly for classes that is not dominant such as *Pedestrian* and *Work van*.

With regard to the overall performance, we got 0.8844 mean recall, 0.9776 classification accuracy, 0.9201 mean precision, and 0.9651 Cohen Kappa Score on verification data. The performance comparison with other deep learning models are shown in Table 4, which demonstrate the proposed scheme is able to increase mean precision to some extend, compared with the baseline algorithms. Concretely, the proposed DCEM-BS improves the mean precision by more than 2% in contrast to the single models. Moreover, DCEM-BS is better than DCEM in terms of performance, which indicates that our balanced sampling tactic is effective.

Figure 5 presents the vehicle classification confusion matrix of the proposed method. The confusion matrix demonstrates that the proposed vehicle classification scheme can classify rare classes such as *Pedestrian* and *Work van* accurately. Due to the overwhelmingly dominant positions in training data, *Background* and *Car* are easily classified correctly. Finally, we randomly visualized some results of proposed scheme in Fig 6 on the MIO-TCD classification challenge dataset. Our model sometimes failed to make good predictions especially when the images are blurred or details of objects are missing.

More results of the proposed scheme can be found in the MIO-TCD classification challenge related web page.[4]

[4] http://podoce.dinf.usherbrooke.ca/methods/classification/199/

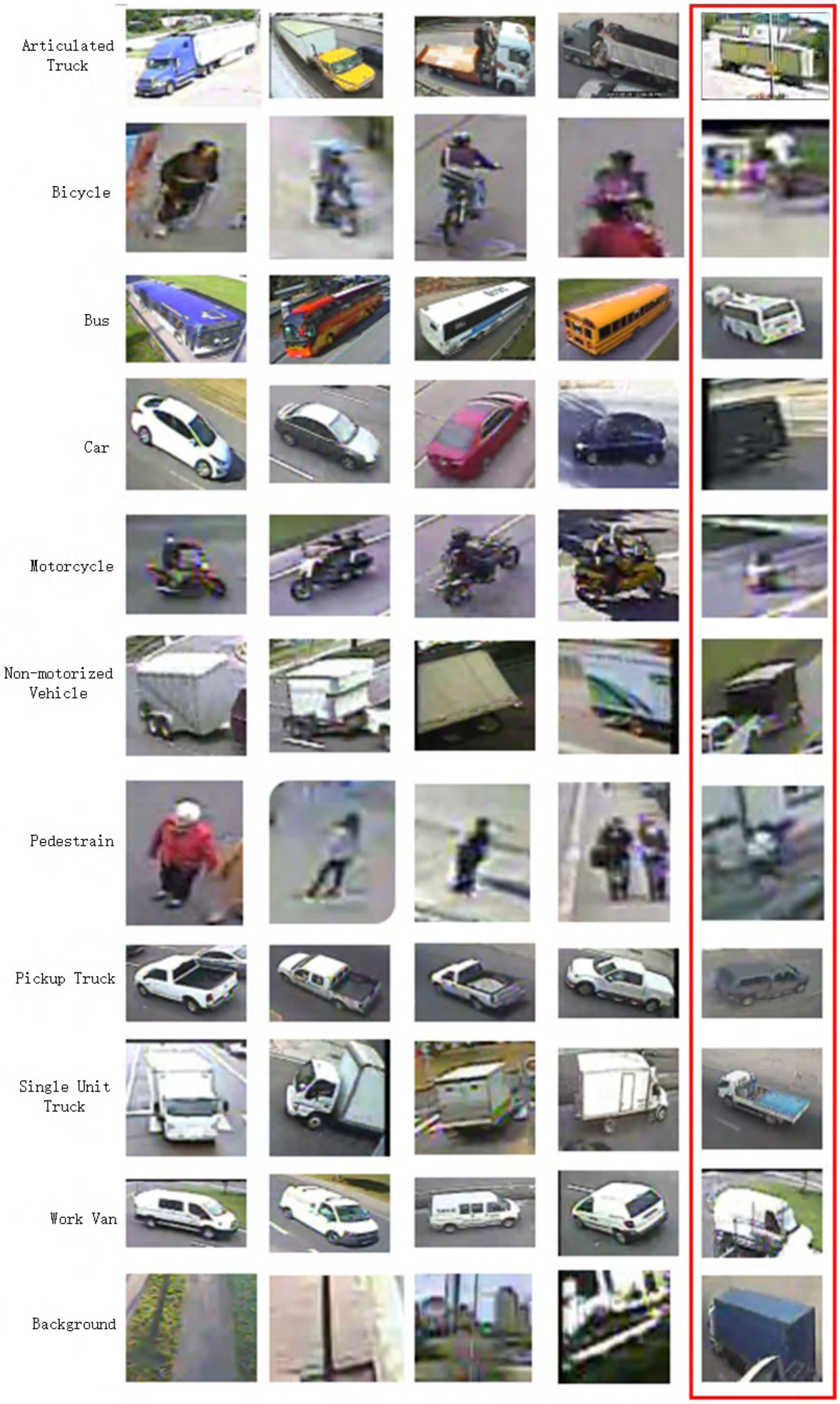


FIGURE 6. Visualization results of the proposed scheme on the MIO-TCD classification challenge dataset. The most right column shown in the red box are suspected misclassifications. For the convenience of showing, the aspect ratios of images have been changed.

V. CONCLUSION

To correctly classify vehicle type on images acquired from visual traffic surveillance sensors, we proposed an image classification scheme based on ensemble deep learning. The proposed vehicle classification scheme consists of two main stages. In the first stage, data augmentation with balanced sampling is applied to alleviate the unbalanced data set problem. In the second stage, an ensemble of convolutional neural network models with different architectures is constructed with parameters learned on the augmented training data set. Experiments on the MIO-TCD classification challenge dataset demonstrate that the proposed method is able to increase mean precision to some extend, compared with the baseline algorithms.

REFERENCES

[1] W. Liu, R. Ji, and S. Li, "Towards 3D object detection with bimodal deep Boltzmann machines over RGBD imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3013–3021.

[2] W. Liu, S. Li, X. Lin, Y. Wu, and R. Ji, "Spectral–spatial co-clustering of hyperspectral image data based on bipartite graph," *Multimedia Syst.*, vol. 22, no. 3, pp. 355–366, 2016.

[3] W. Liu, S. Li, D. Cao, S. Su, and R. Ji, "Detection based object labeling of 3D point cloud for indoor scenes," *Neurocomputing*, vol. 174, pp. 1101–1106, Jan. 2016.

[4] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proc. 3rd IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2003, pp. 435–442.

[5] C. Unsworth and G. Coghill, "Excessive noise injection training of neural networks for markerless tracking in obscured and segmented environments," *Neural Comput.*, vol. 18, no. 9, pp. 2122–2145, 2006.

[6] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006.

[7] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5375–5384.

[8] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep, 2004, vol. 110.

[9] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 281–288, Feb. 2009.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[11] C. Drummond *et al.*, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Proc. Workshop Learn. Imbalanced Datasets II*, Washington, DC, USA, 2003, vol. 11, pp. 1–8.

[12] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing* (Lecture Notes in Computer Science), vol. 3644, D. S. Huang, X. P. Zhang, and G. B. Huang, Eds. Berlin, Germany: Springer, 2005, pp. 878–887.

[13] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[14] T. Maciejewski and J. Stefanowski, "Local neighbourhood extension of smote for mining imbalanced data," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Apr. 2011, pp. 104–111.

[15] P. Domingos, "MetaCost: A general method for making classifiers cost-sensitive," in *Proc. 5th Int. Conf. Knowl. Discovery Data Mining*, San Diego, CA, USA, 1999, pp. 1–10.

[16] K. M. Ting, "A comparative study of cost-sensitive boosting algorithms," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 983–990.

[17] Y. Yan, Y. Liu, M.-L. Shyu, and M. Chen, "Utilizing concept correlations for effective imbalanced data classification," in *Proc. IEEE 15th Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2014, pp. 561–568.

[18] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.

[19] L. Zhang and W. Wang, "A re-sampling method for class imbalance learning with credit data," in *Proc. Int. Conf. Inf. Technol., Comput. Eng. Manage. Sci. (ICM)*, vol. 1. Aug. 2011, pp. 393–397.

[20] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, B, (Cybern.)*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[25] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[26] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[28] P. Jeatrakul, K. Wong, and C. Fung, "Classification of imbalanced data by combining the complementary neural network and smote algorithm," in *Proc. Neural Inf. Process. Models Appl.*, 2010, pp. 152–159.

[29] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. (Aug. 2015). "Cost sensitive learning of deep feature representations from imbalanced data." [Online]. Available: https://arxiv.org/abs/1508.03422

[30] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, "Deep learning for imbalanced multimedia data classification," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2015, pp. 483–488.

[31] Z.-H. Zhou, "Ensemble learning," in *Encyclopedia Biometrics*, S. Z. Li and A. Jain, Eds. New York, NY, USA: Springer, 2015, pp. 411–416.

[32] L. Breiman, "Bagging predictors," Stat. Dept. Univ. California, Oakland, CA, USA, Tech. Rep. 421, 1994.

[33] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1. Aug. 1995, pp. 278–282.

[34] L. Wang, M. Sugiyama, C. Yang, Z.-H. Zhou, and J. Feng, "On the margin explanation of boosting algorithms," in *Proc. COLT*, 2008, pp. 479–490.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[36] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[37] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Comput. Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

WEI LIU received the Ph.D. degree in artificial intelligence from Xiamen University, China, in 2015. He is currently an Assistant Professor with East China Jiaotong University, China. His research interests include machine learning and computer vision.

MIAOHUI ZHANG received the B.S. degree from the Huazhong University of Science and Technology, China, in 2009, and the Ph.D. degree in artificial intelligence from Xiamen University, China, in 2015. Since 2015, he has been an Assistant Researcher with the Jiangxi Academy of Sciences, China. His research interests include image retrieval, machine learning, and object recognition.

ZHIMING LUO received the B.S. degree in cognitive science from Xiamen University, China, in 2011. He is currently pursuing the Ph.D. degree in computer science with Xiamen University, China, and the University of Sherbrooke, Canada. His research interests include traffic surveillance video analytics, computer vision, and machine learning.

YUANZHENG CAI received the B.S. degree in software engineering from Fujian Normal University in 2010, the M.S. degree in computer science from Yunnan University in 2012, and the Ph.D. degree in artificial intelligence from Xiamen University in 2016. He is currently an Assistant Professor with Minjiang University. His research interests include image/video retrieval, machine learning, and object recognition.