Vehicle Type Recognition in Surveillance Images From Labeled Web-Nature Data Using Deep Transfer Learning

Jitian Wang, Han Zheng, Yue Huang, and Xinghao Ding, *Member, IEEE*

***Abstract*— Vehicle type recognition from surveillance images represents a challenging task in the domain of intelligent monitoring systems. Recently, deep learning methods have been applied to solve this problem. The existing deep learning methods, such as convolutional neural networks (CNN), assume that the training and test data are generated from the same or similar imaging systems. They also require a lot of manual annotations for each task. In this paper, we aim to create an improved deep learning method for vehicle type recognition from surveillance images and propose a system based on CNN and transfer learning. Labeled image data of different types of vehicles are easy to acquire from both vehicle manufacturers and Internet sources. Therefore, our proposed surveillance-based vehicle type recognition system is implemented using only labels from Web data. This allows us to overcome the task of manually labeling the data from surveillance images during the training phase. We need to overcome the gap in the types of vehicles between two different imaging systems. For this, a regularization technique in transfer learning is introduced to the objective function of the traditional convolutional neural network. The proposed method was verified through experiments with the public data set comprehensive cars. The experimental results demonstrate that our proposed recognition method outperforms existing deep learning methods when the training and test data are taken from different imaging systems.**

***Index Terms*— Vehicle type recognition, surveillance images, convolutional neural network, transfer learning, unsupervised domain adaptation.**

I. Introduction

IN RECENT years, developing countries have witnessed rapid expansions in their urban areas. This expansion has led to an increase in the demand for better public safety and improvement in traffic conditions. Surveillance images and videos, combined with computer-vision methods, have become the most effective low-cost technologies applied to intelligent traffic monitoring. The task of vehicle monitoring is challenging - vehicle plates can be easily removed or counterfeited, while vehicle logos are small and can be easily covered. Vehicle type is a strong and robust indicator which can be used for the recognition of vehicles suspected of traffic violations. Vehicle-type tracking can be used to identify the manufacturer and the specific model of a vehicle from the large number of images and videos from surveillance data captured daily by traffic monitoring systems. However, due to differences in imaging conditions and subtle variations in the appearances of the vehicles, vehicle type recognition is still a challenging task. Currently, in surveillance systems, cameras are usually located several meters above the ground. Moreover, cameras are intended for traffic control, and therefore capture the upper-frontal-view vehicle images, making the task of accurate vehicle identification challenging.

A. Related Works

The recent studies on vehicle type recognition and related areas can be divided into three categories: a) The first category addresses the recognition problem by detecting and recognizing the license plates and logos of vehicles. These features can be used to directly identify the vehicle type and the manufacturer [1]–[3]. However, license plates and logos do not contain fine-grained information on the vehicle type, which can lead to failures in recognizing the exact maker and model of the vehicle. Another problem is that recognition can also fail if the plates and logos are covered or faked. b) The second category classifies the vehicles directly using global and local low-level features and their combinations. These features include oriented contour points, scale-invariant feature transform (SIFT) features, pyramid histogram of oriented gradients (PHOG), Gabor transform features, Haar-like features, sobel edge response, edge orientation, direct normalized gradients, locally normalized gradients, Harris corner response, among others [4]–[13]. In this category of recognition, the hand-crafted features should be carefully designed for each recognition task. c) The third category adopts the convolutional neural networks (CNN) architecture, a deep learning method, which has recently proven to be highly effective in the

Manuscript received March 15, 2017; revised August 14, 2017 and October 7, 2017; accepted October 19, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 81671766, Grant 81301278, Grant 61571382, Grant 61571005, Grant 61172179, and Grant 61103121, in part by the Natural Science Foundation of Guangdong Province under Grant 2015A030313007, in part by the Fundamental Research Funds for the Central Universities under Grant 20720160075, in part by the National Natural Science Foundation of Fujian province, China, under Grant 2017J01126, and in part by CCF-Tencent open grant. The Associate Editor for this paper was J. Miller. *(Jitian Wang and Han Zheng contributed equally to this work.) (Corresponding author: Yue Huang.)*

The authors are with the Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Communication Engineering, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: yhuang2010@xmu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2765676

domain of computer vision [14]. Recently, we proposed a vehicle logo recognition method based on CNNs using an efficient pre-training strategy for surveillance images [1]. Additionally, a pre-trained convolutional neural network and unsupervised learning method using frontal view web-nature images with image reconstruction and sparse Laplacian filter learning method were proposed in [15]. Yang *et al.* [16] published a large dataset for fine-grained vehicle type recognition, and then proposed a CNN-based method for fine-grained vehicle model recognition capable of distinguishing vehicles from different viewpoints. Fang *et al.* [17] proposed another fine-grained vehicle type recognition method, where the classification was done based on a combination of local and global features extracted from the CNN model. The CNN model employed in [17], called pre-trained AlexNet [18], is one of the most widely-used models. Further, in [19], a 3D vehicle bounding box is used for unpacking the vehicle images from the video stream. Then these images were fed to a deep convolutional neural network in order to boost the recognition performance. Yu *et al.* [20] used a faster region CNN (FR-CNN) model to detect the vehicles, and then used another CNN model to identify the types. From our literature review, we see that recently, deep learning methods have been widely used for vehicle type recognition, and that they outperform the methods which use hand-crafted features [14]–[17], [19]. For example, the classification accuracy of the recent work in [17] is 98.63% for fine-grained vehicle type recognition, while the works with the hand-crafted features have been able to achieve the best accuracy of 93.1%, or lower when using smaller datasets.

B. Motivations

Traditional machine learning algorithms assume that the training and test set data always have the same distribution. Majority of classical works on vehicle type recognition, with both coarse and fine-grained recognition algorithms, rely on this assumption. Since, the training and test data are from the same dataset, they are supposed to be generated using the same imaging procedure [17], [19]. However, in real-life scenarios, this assumption may not be satisfied. In our paper, to perform the vehicle recognition task from the view of real-world applications, we forego this assumption, and come up with techniques which can deal with this discrepancy.

However, we know that vehicle manufacturers almost always release high quality images of vehicles either before or soon after the market-release of any new model. Compared to the low quality images obtained from surveillance data, these vehicle images, which are called web-nature images, are much cheaper to collect and label. Therefore, to address the lack of proper labelled data, it would be helpful to use these web-nature images as the training data for the task of vehicle type recognition or label generation. However, using these web images as training data would heavily affect the performance of traditional machine learning algorithms which work with the assumption that the train and test data are drawn from the same distribution. Training the algorithm on these high quality images would lead to low performances when tested using low-quality images from the surveillance data.

Since web-nature images can be easily collected either from the Internet or from the manufacturers, in this work we aim to design an improved CNN method which uses the principle of knowledge transfer to learn from the web-nature images and then detect the vehicles from the surveillance data. For this purpose, we introduce an unsupervised domain adaptation technique to a deep convolutional neural network as an additional regularization step. This is supposed to minimize the gap between the images of the same vehicle type from the two imaging systems. Using this technique, the classification system trained on web-nature data can be applied for directly recognizing the vehicles from the surveillance data.

The contributions of this paper can be summarized as follows: 1) to the best of our knowledge, this is the first report of a vehicle type recognition system from surveillance data, where the classification is only trained on labeled web-nature data. The features extracted in the proposed model are simultaneously representative and transferable, making it possible to take full advantage of the large-scale labeled data present on the web. Therefore, this reduces the burden of annotation of the surveillance images. 2) This paper is the first work where the recognition of vehicle types is validated with degraded images with various imaging qualities. During validation, the proposed model outperforms other deep learning methods which do not use transfer learning.

The remaining part of the paper is organized as follows: in section II we provide a detailed description of the method. The description of the dataset, the implementation details, and the experimental results are provided in section III. In section IV we perform robustness validation and present the results and other related discussions. Our conclusions are presented in section V.

II. Methods

The architecture of the proposed deep CNN model is presented in Fig. 1. From Fig. 1, we can see that the proposed framework represents an end-to-end system, where the feature extraction and the classifier are globally optimized. The presented architecture has a structure similar to the widely-used deep convolutional neural network AlexNet, and consists of five convolutional layers, three pooling layers, and three fully connected layers [20]. Each unit of the proposed CNN is introduced in this section. The differences between our proposed model and AlexNet, which are the regulations in the last two layers are also highlighted in Fig.1. The detailed description of the neural network for our proposed vehicle type recognition is provided as follows.

The convolution layer in the proposed CNN is similar to the convolution layer in common CNN models [14], [20]. Namely, the inputs of convolution layer are either original images or outputs of the previous layer. If x_i^l denotes the ith input of lth layer, and $k_i^l j$ denotes the jth kernel that corresponds to x_i^l, then the feature map of output of lth layer is defined by:

$$x_j^{(l+1)} = f(k_{ij}^l \otimes x_j^l) \tag{1}$$

where $x_j^{(l+1)}$ represents the part of input of $(l+1)$th layer. In order to accelerate the training procedure of the CNN,

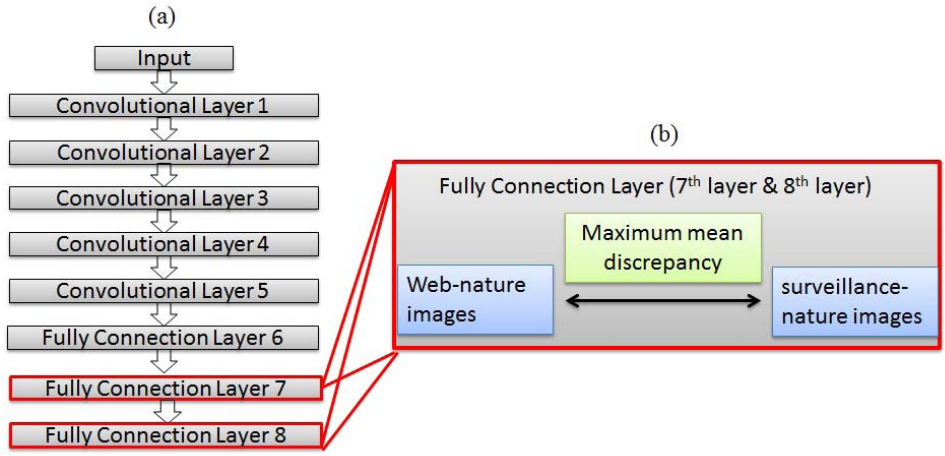


Fig. 1. Architecture of the proposed fine-grained vehicle type recognition system based on CNN and unsupervised domain adaptation (a). Fully connection layers. (b) the expansion of the last two layers.

TABLE I
PARAMETERS OF THE CONVOLUTIONAL LAYERS OF OUR PROPOSED CNN MODEL

Layer index	No. of kernel number	Kernel size
1	96	11×11×3
2	256	5×5×48
3	384	3×3×256
4	384	3×3×192
5	256	3×3×192

the function f is defined in terms of a nonlinear non-saturating mapping, called Rectified Linear Units (ReLUs) [20], instead of as a saturating nonlinearity, such as sigmoid function. The parameters of the convolutional layers are presented in Table I.

The max-pooling layers, which are always aligned after a convolutional layer, are introduced to detect the maximum response of the generated feature maps in order to simultaneously discern kernels as well as to reduce the resolution of the feature maps. There are three max-pooling layers in the proposed model. The max-pooling action creates position invariance over larger local regions and down-samples the input by a factor of 2 in each direction. In the proposed network, a pooling layer is assigned after each of the first, the second, and the fifth convolutional layers. The third, the fourth, and the fifth convolutional layers are connected to their next layers directly without any pooling operation.

The neurons of a fully connected layer are connected to all the neurons in the previous layer. There are three fully connected layers in the proposed model, with each containing 4,096 neurons. These layers are applied at the end of the network as nested linear classifiers. Moreover, in the last layer, the soft-max is applied as a classifier. The feature vector generated by the previous layer is fed to the soft-max, along with the probabilities of the categories representing the prediction results of the classifier. A domain adaptation is performed between the labeled training data and the unlabeled test data. Theoretically, the two basic domains in transfer learning are the source domain and the target domain, and it is always assumed that data in these domains have different distributions. In this study, the web-nature data and the surveillance data can be considered as the source and the target domains, respectively. Since, there is no annotation for the data in the target domain, it is desirable to transfer the knowledge from the fully-labeled large-scale data in the source domain to the unlabeled data in the target domain. This has been defined as the task of unsupervised domain adaptation in the theory of transfer learning [21]. In this work, a domain adaptation based on the maximum mean discrepancy is introduced to build the CNN model. Maximum mean discrepancy (MMD) represents a two-sample problem used to determine whether two distributions are the same or not [22]. If F is defined as a class of functions f, then, the distance between two distributions is defined by:

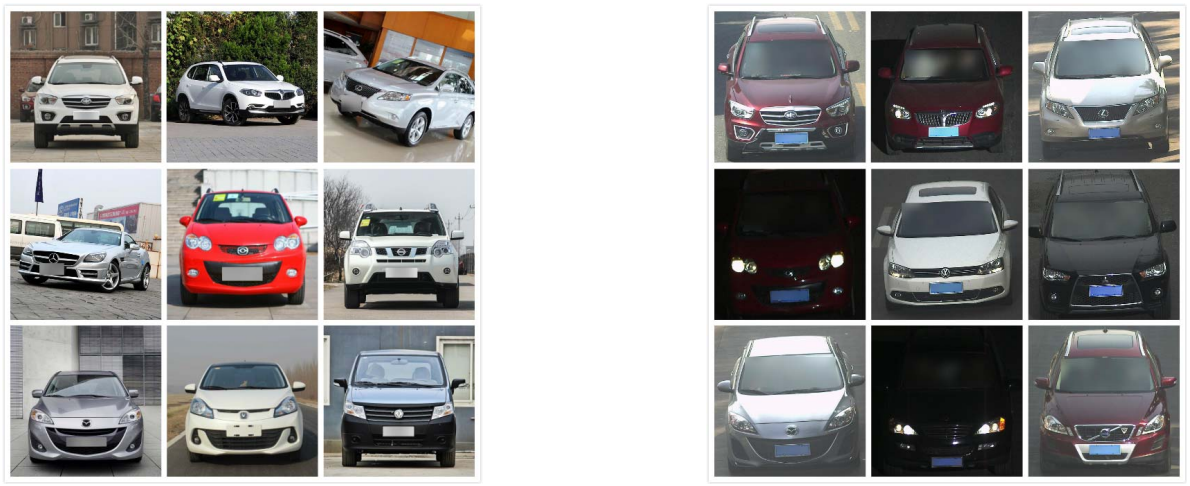
$$D(p,q) := \sup_{f\in F}(E_p[f(x)] - E_q[f(y)]) \tag{2}$$

where x and y are independently and identically distributed (IID) from p and q, respectively; sup denotes the supremum, and E represents the expectation [22]. In MMD, the distance between two distributions is measured by the mean of features that are mapped in the Hilbert space induced by a given kernel [23]. A smaller value of D represents larger probability that data in two domains have the same distribution [24]. Using the kernel operations, MMD can be expressed by expectation of kernel functions, so the function in (2) can be replaced by its square formulation as [25]:

$$D_k^2(p,q) = E_{x_p^s x_p^s}k(x_p^s,x_p^s) + E_{x_q^t x_q^t}k(x_q^t,x_q^t) - 2E_{x_p^s x_q^t}k(x_p^s,x_q^t) \tag{3}$$

where E denotes the expectation, x_p^s is the sample in source domain, x_p^t is the corresponding samples in target domain, and k is the kernel defined as $k(x_i,x_j) = e^{-\|x_i-x_j\|^2/\gamma}$.

In our proposed method, MMD is used as a regularization term to overcome the gap between the web-nature data and the surveillance data for the same type of vehicle. Considering that CNN achieves good performances during the extraction of high-level and high-dimensional features, $f(x)$ and $f(y)$ in (2) can be defined as hierarchical feature extraction procedures of



(a) (b)

Fig. 2. (a) Web-nature images and (b) surveillance images in the public dataset CompCars.

CNN [26]. If ϕ represents the features generated by AlexNet, then (2) can be rewritten as:

$$D(p,q) := \sup_{f \in F} \left(E_p[\phi(x)] - E_q[\phi(y)] \right) \tag{4}$$

If we denote $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ as the ith labeled sample in source domain, y_i^s as the labels, $D_t = \{x_j^t\}_{j=1}^{N_t}$ as the jth sample in the target domain, N_s and N_t as sizes of source and target domains, then, the objective function of the method in our work is written as:

$$\min_{\theta} \frac{1}{N_s} \sum_{i=1}^{N_s} J(\theta(x_i^s), y_i^s) + \lambda \sum_{\ell=\ell_1}^{\ell_2} D_k^2(\theta_l(D_s), \theta_l(D_t)) \tag{5}$$

where the first term J is the same as the typical CNN, e.g. AlexNet, θ is used to represent the set of parameters of a CNN model, where θ (x_i^s) denotes the conditional probability of assigning of sample x_i^s to label y_i^s. Since, there is no information regarding the labels in target domain, both x_i^s and y_i^s are from source domain. Further, in (5), D represents the MMD-based regularization term for calculation of distance of high-level representation between the domains, $\theta_l(D_s)$ and $\theta_l(D_t)$ denote the outputs of lth layer in the source and the target domain, respectively. λ is the trade-off coefficient, and the value $\lambda > 0$. From (5), we can derive that the objective function can benefit from both transfer learning and deep learning. The variable θ is optimized in order to gain a strong representative power in the data of both source domain and to reduce the discrepancies between the two domains. In this way, the transferable features between the two domains can be learned, and suing supervised learning, labeled samples in the source domain can be used to recognize the samples in the target domain.

From Fig.1 we can be notice that domain adaptation is applied only in the last two layers, while the other layers remain fixed. This is also demonstrated in (5) where $l1$ and $l2$ are equal to 7 and 8, respectively. This is consistent with the discussions in [27], where the authors showed that features from the lower layers are more transferrable across different domains, while features extracted from higher layers are more domain-specific. In this study, the mini-batch stochastic gradient descent (SGD) algorithm is employed to optimize the objective function [26].

III. Experimental Results

A. Dataset Descriptions

The Comprehensive Cars (CompCars) dataset [16] is used for evaluating the proposed model.[1] We choose this particular dataset for validation due to the following reasons: 1) it is one of the largest datasets for fine-grained vehicle type recognition; 2) it includes both web-nature images (Fig.2(a)) and surveillance images (Fig.2(b)). Our proposed work only focuses on the vehicle type recognition rather than on the task of vehicle detection, so if there are more than one vehicles in the images, each of them can be segmented using a vehicle detection method. For our purpose, we remove the images of rear-view vehicles from the web-nature data. From Fig. 2(b), we observe that some surveillance images suffer from large variations in illumination due to the differences in the traffic imaging conditions, making the recognition of vehicles from frontal-view surveillance data more challenging. The average size of an original surveillance image is 800×850 pixels. The web-nature class contains standard photos of cars from different viewpoints, e.g. rear, front and side. However, it should be mentioned that the frontal-view images in web-nature data are quite different from the frontal-view images obtained from surveillance data. This is due to the factor that, the cameras in surveillance monitoring systems are always located several meters above the ground, thus, the exact description of the viewpoint in surveillance images should be top-frontal instead of standard-frontal as used in the web-nature data. .

B. Implementation and Results

In our proposed implementation, the basic server settings are: a 56 Intel(R)Xeon(R) CPU E5-2683 V3@ 2.00GHz,

[1] http://mmlab.ie.cuhk.edu.hk/datasets/comp_cars/index.html

TABLE II
DETAILS OF 20 VEHICLE TYPES IN C1

Besturn-X80	Zhonghua-V5	Lexus-GX	Benz-SLK Class	Haima-Prince
Volkswagen-Beetle	Volvo-XC60	Toyota-FJ Cruiser	Mitsubishi-Outlander	Honda-Spirior
Nissan-X Trail	MAZDA-5	Acura-ZDX	Changan-Benben	Volkswagen-Sagitar
Shuanglong-Kyron	Geely-GC7	Hyundai-Tucson	Jeep-Patriot	Dongfengfengdu-Shuaike

TABLE III
DETAILS OF D1 AND D2

Index	Data modality	Image number.	Category	Label or not	Roles
D1	Web-nature	1844	20 categories(C1)	Labeled	Training-Source
D2	Surveillance-nature	2221	20 categories(C1)	Unlabeled	Testing-Target

64G RAM, and NVIDIA GeForce 1080 GPU. All the images are resized to the same size of 227 $\times$ 227 as in [20]. Many pre-training strategies for CNNs, including both supervised [20] and unsupervised strategies [15], have been reported. However, the development of a new pre-training strategy was not our main contribution, and we applied one of the most widely used pre-training strategy reported in [17] and [20]. Here a CNN is pre-trained with a set of large-scale labeled natural images, e.g. ImageNet, and then it is fine-tuned with the vehicle image data. Shin *et al.* [28] stated that for image recognition tasks, pre-trained models have the ability to both enhance the performance and to speed up the training procedure of the CNN.

We used 1844 labeled images (D1) of 20 different types of vehicles from the web-nature dataset as the training data (source domain). We also extracted 2221 unlabeled surveillance-nature images (D2) to be used as the test data (target domain). There was no overlap between the images in the training and test datasets. All vehicle types in both the training and test data belonged to the 20 categories (called C1). During the training phase, the neural networks in our proposed method learned the labeled images in D1, together with the unlabeled images in D2. On the other hand, during the test phase, the neural network only predicted the labels of the images in D2. Details of C1, D1, and D2 are provided in the Tables II and III.

Existing vehicle type recognition systems use the surveillance-data as the training data, in order to maintain the similarity in the sources of the training and test data. To the best of our knowledge, this is the first attempt to perform vehicle recognition in surveillance images using a system trained only on web-nature labeled data. In our proposed work, we compared the proposed CNN model with two state-of-the-art CNN models without domain adaptation. The comparison in the results is obtained using the pre-trained AlexNet and ResNet [17], [18], [29] models. AlexNet is one of most popular CNN models [18], and has been used in [17] for fine-grained vehicle type recognition. In their implementation, the network is first pre-trained with a large-scale labeled natural dataset called ImageNet, and then it is fine-tuned with the specific data [17], [18]. We also implemented the pre-trained AlexNet only with the surveillance-nature data in D2, where half of D2 was used as the training data and the other half of D2 was used as the test data. For our experiments, we obtain a high recognition accuracy of 99.09%. This high accuracy demonstrates the strong recognition ability of the pre-trained AlexNet in the case where both the training data and test data are drawn from the same imaging system. This is one reason that only CNN methods are employed in the comparisons. The other reason is that the recent works in vehicle type recognition employ CNN-related methods. Besides AlexNet, ResNet is a recently developed state-of-the-art CNN model, proposed to facilitate network training and optimization, and to obtain a satisfactory accuracy with considerable neural network depth [29]. It won the first place in the classification task in the Large Scale Visual Recognition Challenge 2015 (ILSVRC 2015). The comparison between existing recognition methods and our proposed method was performed using the Caffe library with GPU acceleration. We implemented the pre-trained AlexNet and ResNet models with source codes published by Fang *et al.* [17] and Zhang *et al.* [29] on their websites. The number of layers in the ResNet was set to 20. For implementing the AlexNet and ResNet in our research, the neural networks were trained on labeled web-nature data and then were directly tested on the surveillance data.

TABLE IV
COMPARISON OF METHODS IN TERMS OF ACCURACY

Method	Accuracy
Pre-trained AlexNet[17,18]	42%
ResNet[29]	9.2%
Proposed	54.56%

The comparison of our proposed method and commonly used recognition methods in terms of accuracy is provided in Table IV. According to the results presented in Table IV, we can observe that ResNet has the lowest performance, meanwhile, performances of the pre-trained AlexNet decreases dramatically to 42% when the network is trained using the web-nature data and tested using the surveillance-nature data. Our proposed method achieves an improvement of 12.56% and 39.36% when it is compared with the pre-trained AlexNet and ResNet models. The results mean that the introduction of unsupervised domain adaptation enables us to overcome the gap between the training and test data to enhance the performance.

Fig. 3. Surveillance images for different down-sampling rates. From left to right are original image, low-resolution image with down-sampling rate of 4, 16 and 64 respectively. All images are zoom in to the same size.

TABLE V

COMPARISON OF METHODS IN TERMS OF ACCURACY AT EACH VEHICLE TYPE, N_1 AND N_2 DENOTE IMAGE NUMBER IN THE TRAINING DATA (D1) AND TESTING DATA (D2) RESPECTIVELY. ACCURACY 1 AND ACCURACY 2 DENOTE PERFORMANCES BY PRE-TRAINED ALEXNET AND PROPOSED METHOD

Vehicle type	N_1	N_2	Accuracy1	Accuracy2
Besturn X80	53	83	46.99%	90.36%
Zhonghua V5	56	84	38.09%	52.38%
Lexus GX	48	176	3.41%	22.73%
Benz SLK Class	49	24	8.33%	4.17%
Haima Prince	46	129	58.14%	42.64%
Volkswagen Beetle	52	61	70.49%	91.80%
Volkswagen Sagitar	58	367	58.46%	17.49%
Toyota FJ Cruiser	54	51	60.78%	58.82%
Mitsubishi Outlander	58	169	16.57%	40.24%
Honda Spirior	63	67	1.49%	56.72%
MAZDA 5	53	111	59.46%	51.35%
Nissan X-Trail	50	104	17.31	67.31%
Acura ZDX	74	54	16.67%	22.22%
Changan Benben	50	148	33.78%	4.054%
Volvo XC60	85	94	8.51%	29.79%
Shuanglong Kyron	55	50	38%	22%
Geely GC7	62	33	45.45%	36.36%
Hyundai Tucson	50	80	30%	35%
Jeep-Patriot	57	267	60.67%	73.41%
Dongfengfengdu Shuaike	41	69	36.23%	27.54%
Average	56	111	42%	**53.81%**

We want to further examine the recognition accuracies for each vehicle type in the C1 dataset. In Table V we especially compare the performances between our proposed model and the pre-trained AlexNet. To do this, we calculate the average value of the accuracies across the 20 vehicle types. It can be observed that the average accuracy of our proposed method achieves an improvement of 11.81% compared to the pre-trained AlexNet model.

The process of network learning from D2 is a typical CNN model, which the focus is on extracting features with strong representative power from the surveillance data. As a comparison, the process of network learning from D1 and D2 is used to train a network that has the ability to extract features which are simultaneously representative and transferable across domains (web-nature and surveillance). In the view of transfer learning, the regularization term of our proposed model is able to take the advantage of the data from both the domains. This can be considered as additional knowledge, which, for the task of recognition in the target domain, can then be transferred to the target domain. This transfer can be used to reduce the gap between the source and target domains, which in turn improves the classification performance.

Computational cost is a prerequisite for measuring the efficiency of real-world applications. The computing time of our proposed method and AlexNet during training is 2878 seconds and 3807 seconds respectively. The results are calculated based on 1844 labeled web-nature images and 2221 unlabeled surveillance images, as in the Experiment section. The similar values in computing times denote that there is very little difference in the computational cost caused by the domain adaptation step during training. During the testing procedure, the computing time is calculated based on the averaged total testing time across the 2221 unlabeled surveillance images. The times are 0.7088 seconds and 0.7042 seconds for our proposed method and AlexNet respectively. The results demonstrate that there is no significant difference in the computational cost between AlexNet and our proposed model during the test phase.

IV. DISCUSSIONS

In this section, a detailed analysis of the performance of our proposed method is presented, including the validation of its robustness against simulated resolutions and blurring. In view of transfer learning, we also present discussions on the pre-training procedure.

A. Robustness Validations

In this subsection, we validate the robustness of our proposed method for images with different resolutions and blurring effects. In order to simulate degraded low-resolution images generated in real-world traffic applications, the images intended for testing are degraded with various down-sampling rates and blurring kernels. This is a validation for the robustness of our system in terms of different image qualities caused by various imaging conditions in surveillance systems. The robustness validation is performed using the technique described in the previous section, where the web-nature data and surveillance data belonging to 20 vehicle categories were employed as training/source and test/target data. The accuracy in the robustness validation is calculated in the same way as shown in Table IV. In Fig. 3, we present an example of a surveillance image with different down-sampling rates.

(a) (b)

Fig. 4. Fig.5 Two examples in the simulated surveillance images which both contains more than one vehicle. The vehicles in the images are with different resolutions. In each image, the vehicles are detected by Faster R-CNN [30] at first with the bounding boxes, and then are recognized by the proposed framework.

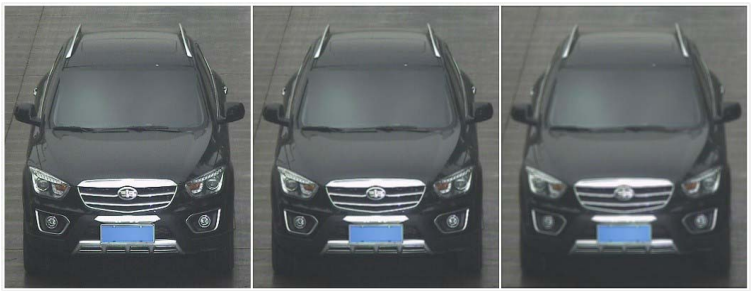


Fig. 5. Surveillance images for different blurring kernels: (a) original image, (b) image processed by a low-pass Gaussian filter with standard deviation of 0.5, and (c) image processed by a low-pass Gaussian filter with standard deviation of 1.

TABLE VI
CLASSIFICATION ACCURACY FOR DIFFERENT DOWN-SAMPLING RATES

Down-sampling rate	1	4	16	64
AlexNet [18]	42%	41%	39%	17%
Proposed	54.56%	52.5%	51.72%	31.5%

A down-sampling rate of 4 denotes that the image has 1/4 the resolution of the original image, with 1/2 down-sampled length. Comparison of the classification accuracies of our proposed method and the pre-trained AlexNet for different down-sampling rates is provided in Table VI. From the table we can observe that the accuracy of our proposed method is higher than the accuracy of the pre-trained AlexNet at all down-sampling rates.

Considered the fact that in the real-world applications, there are more than one vehicle in one some surveillance image, and the vehicles are in various resolutions as well, the recognition results are also presented in two simulated images, as shown in Fig. 4. The vehicles in the images are detected by Faster R-CNN [30] at first (bounding boxes in the images), and then are recognized by proposed framework.

In order to validate the robustness of our method for various blurring effects, the test images are degraded with different Gaussian kernels. In the experiment, the test images are first degraded with a down-sampling rate of 4, and then the blurred images are generated using Gaussian kernels with different standard deviations. As shown in Fig. 5, a larger standard deviation denotes a stronger blurring effect. The comparison of the performance of the recognition using our proposed method and AlexNet for different blurring kernels is presented in Table VII. From this table, we can observe that our proposed method is robust even when images are heavily blurred. Furthermore, the accuracy of our proposed method is higher than the accuracy of AlexNet for all blurring kernels. Besides, the accuracy is 54% when there is no blurring effect, and decreases to 50% for the strongest

TABLE VII
CLASSIFICATION ACCURACY FOR DIFFERENT BLURRING EFFECTS (STD DENOTES STANDARD DEVIATION VALUE)

Std of Gaussian function	0	0.5	1
AlexNet [18]	42%	40%	23%
Proposed	54.56%	54.3%	50%

TABLE VIII
CLASSIFICATION ACCURACY FOR DIFFERENT TRANSFER LEARNING STRATEGIES

Pre-training in ImageNet	Domain Adaptation	Accuracy
×	×	7.3%
×	√	11.7%
√	×	42%
√	√	54.56%

TABLE IX
DETAILS OF 10 NEW GENERATION VEHICLE TYPES (C2) IN THE PRE-TRAINING VALIDATION

Audi-Q5	Yiqi-Weizhi	Lexus-ES	Volkswagen-Jetta	Toyota-Huaguan	MAZDA-6	KIA-Cerato	Jinbei-Haishi	Wuling-Hongguang	BYD-F6

TABLE X
DETAILS OF THREE VEHICLE DATASETS IN THE PRE-TRAINING VALIDATION

Index	Data modality	Image Number	Category	Label or not	Roles
D2	Surveillance-nature	2221	20 categories(C1)	Labeled	Training-Source
D3	Web-nature	1114	10 categories(C2)	Labeled	Training-Source
D4	Surveillance-nature	2700	10 categories(C2)	Unlabeled	Testing-Target

blurring effect. As a comparison, the accuracy of a pre-trained AlexNet decreases dramatically from 42% to 23% for the same increase in blurring. This evaluation demonstrates that our proposed method is more robust and better than the pre-trained AlexNet across different blurring effects. It can be concluded that the performance improvement obtained from our method is enhanced when the recognition task performed on the surveillance data is more challenging, for example, has strong blurring effects.

B. Analysis on the Pre-Training Process in the View of Transfer Learning

Experimental results in the previous section have demonstrated that the proposed model performs better than the traditional pre-trained AlexNet for the task of recognizing vehicle types from surveillance images labeled with web-nature data only. Transfer learning can be categorized as inductive learning and transductive learning according to the label information in the source and the target domains. For example, unsupervised domain adaptation belongs to transductive learning, where the source domain is fully labeled and there is no label in the target domain. According to the transfer learning theory, the process of identifying vehicle types from a pre-trained large-scale natural image dataset, such as ImageNet, can be explained as an inductive transfer learning strategy, where ImageNet is the source domain, and the proposed vehicle data is the target domain [21]. The performance of CNNs can be improved using knowledge transfer from auxiliary data (ImageNet) from the source domain to the vehicle data in the target domain. The knowledge transfer process can be explained as that the discriminative patterns in the natural images can also be used to provide benefits to the recognition process for classifying different vehicle types. Therefore, there are two transfer learning strategies applied in our proposed method. They are: a) pre-training using the ImageNet and b) unsupervised domain adaptation between the surveillance data and the web-nature data. In order to further validate our proposed method, we analyzed the impact of the two transfer learning strategies on the resultant accuracy of our proposed method. The results obtained are presented in Table VIII, where we can see that both transfer learning strategies enhance the performance of our proposed recognition task.

Besides, it has been reported in other references [17], as well as observed during our experiments, that AlexNet achieves a satisfactory classification performance in the presence of sufficient labeled surveillance data. Inspired by the fact that a network pre-trained with ImageNet is a technique which is widely used to improve the performances of AlexNet, in our experiment, we test whether a classification system trained using labeled surveillance images can further enhance the performances of our model during the recognition of newer generation vehicles.

TABLE XI
COMPARISON OF PROPOSED MODEL WITHOUT AND WITH PRE-TRAINING WITH D2, IN TERMS OF ACCURACY AT EACH VEHICLE TYPE, N_3 AND N_4 DENOTE IMAGE NUMBER IN D3 AND D4 RESPECTIVELY. ACCURACY 1 AND ACCURACY 2 DENOTE PERFORMANCES BY PROPOSED METHOD WITHOUT AND WITH PRE-TRAINING AT D2 RESPECTIVELY

Vehicle type	N_3	N_4	Accuracy1	Accuracy2
Audi Q5	63	309	85.44%	91.87%
Yiqi-Weizhi	44	68	38.24%	48.52%
Lexus ES	54	83	38.55%	37.34%
Volkswagen-Jetta	52	807	80.45%	79.81%
BYD F6	53	108	12.04%	3.70%
Toyota-Huaguan	49	546	17.39%	19.04%
Mazda 6	52	342	21.35%	45.03%
KIA-Cerato	57	222	50%	48.65%
JingbeiHaishi	53	75	69.34%	86.67%
Wulinghongguang	51	141	46.81%	43.97%
Average	53	270	51.17%	**55.13%**

For this experiment, we use 1178 labeled web-nature vehicle images from an additional 10 categories (C2) for training. The vehicle images in C2/D3 are a simulation of new generation vehicle types collected from the Internet. We use 2700 surveillance-nature images from C2 as the test data. The train dataset is called D3 and the test dataset is called D4. Besides, images in the surveillance dataset D2 are still used with their labels. Moreover, there is no overlap among D2, D3 and D4. As shown in Table II and Table IX, the vehicle types in C1 and C2 do not overlap either.

Besides pre-training using ImageNet, in this experiment, we also pre-train our proposed model using labeled surveillance-nature data (D2) from the 20 categories (C1). Next, the model is fine-tuned using labeled web-nature images (D3) as an update for introducing information of the new generation vehicle types described in C2. Finally, the proposed model is validated with the test surveillance data in D4. The details of C2, D3 and D4 are provided in Table IX and Table X and the experimental results are provided in Table XI. As in the experimental section, we also calculate the value of the recognition accuracy across the 10 test vehicle types in C2. The average accuracies of our proposed model without and

with pre-training in D2 are 51.17% and 55.13% respectively. We can observe that our proposed method with pre-training in labeled surveillance data has the ability to improve the performance of recognition.

Here we also include another experiment to demonstrate the results when some labeled surveillance data are available to perform transfer learning with our proposed model. The description of the dataset for this experiment is the same as mentioned in the second paragraph of the section III-B. The network is first trained with the proposed model using 1844 labeled web-nature images and 2221 unlabeled surveillance images from 20 categories. Then, the network is fine-tuned using 111 labeled surveillance images as an update. The rest of the 2100 surveillance images is used as the test set. The average accuracy over the 20 categories for vehicle type recognition is 65.55%. This is better than the performance of our original proposed system (54.56%) shown in the last row of Table IV. The experimental results demonstrate that the recognition performances can be further enhanced when some labeled data in the target domain is used. However, the requirement of obtaining 111 labeled surveillance images can sometimes still be a challenge for some practical uses of our application. This is because, since there are large quantities of surveillance-nature images or videos every day, it is very time-consuming to manually recognize and annotate the vehicle-of-interest from the surveillance data.

V. Conclusions

Although deep learning methods have demonstrated satisfactory performances for vehicle type recognition in the literature, the manual labeling of the vehicle types from surveillance images to generate the training data is a very difficult and time-consuming task. In this paper, we present a novel approach to solve the problem of fine-grained vehicle type recognition from surveillance images gathered from real-world intelligent monitoring systems. By introducing the idea of unsupervised domain adaptation in transfer learning, our work proposes a vehicle type recognition system that can extract transferable features across web-nature data and surveillance data. Thus, large-scale annotated web-nature data, which is much easier to collect, can be applied to the task of vehicle type recognition from surveillance images. The system strongly reduces the burden of collecting labeled data for each specified surveillance system. The proposed method was verified using fine-grained real-world vehicle type data obtained from surveillance images. The proposed method was compared with widely-used deep learning methods, and the obtained results demonstrate that our proposed method performs better than the commonly used learning methods. Lastly, the validation of our proposed method on a public dataset demonstrates that our proposed method is highly suitable for real-world vehicle type recognition.

References

[1] Y. Huang, R. Wu, Y. Sun, W. Wang, and X. Ding, "Vehicle logo recognition system based on convolutional neural networks with a pretraining strategy," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1951–1960, Aug. 2015.

[2] A. P. Psyllos, C.-N. E. Anagnostopoulos, and E. Kayafas, "Vehicle logo recognition using a SIFT-based enhanced matching scheme," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 322–328, Jun. 2010.

[3] H. Pan and B. Zhang, "An integrative approach to accurate vehicle logo detection," *J. Elect. Comput. Eng.*, vol. 2013, Sep. 2013, Art. no. 391652.

[4] L. Liao, R. Hu, J. Xiao, Q. Wang, J. Xiao, and J. Chen, "Exploiting effects of parts in fine-grained categorization of vehicles," in *Proc. Int. Conf. Image Process.*, Sep. 2015, pp. 745–749.

[5] H. He, Z. Shao, and J. Tan, "Recognition of car makes and models from a single traffic-camera image," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3182–3192, Dec. 2015.

[6] X. Ma and W. E. L. Grimson, "Edge-based rich representation for vehicle classification," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1185–1192.

[7] V. S. Petrovic and T. F. Cootes, "Analysis of features for rigid structure vehicle type recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2004, pp. 587–596.

[8] F. Kazemi, H. Pourreza, R. Moravejian, and E. Kazemi, "Vehicle recognition using curvelet transform and thresholding," in *Advances in Computer and Information Sciences and Engineering*, T. Sobh, Ed. Dordrecht, The Netherlands: Springer, 2008, pp. 142–146.

[9] J. Li, W. Zhao, and H. Guo, "Vehicle type recognition based on harris corner detector," in *Proc. 2nd Int. Conf. Transp. Eng.*, 2009, pp. 3320–3325.

[10] B. Zhang, "Reliable classification of vehicle types based on cascade classifier ensembles," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 322–332, Mar. 2013.

[11] B. Zhang and Y. Zhou, "Vehicle type and make recognition by combined features and rotation forest ensemble," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 26, no. 3, p. 1250004, 2012.

[12] P. Negri, X. Clady, M. Milgram, and R. Poulenard, "An oriented-contour point based voting algorithm for vehicle type classification," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2006, pp. 574–577.

[13] Y. Tang, C. Zhang, R. Gu, P. Li, and B. Yang, "Vehicle detection and recognition for intelligent traffic surveillance system," *Multimedia Tools Appl.*, vol. 76, no. 4, pp. 5817–5832, 2017.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[15] 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.

[16] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3973–3981.

[17] J. Fang, Y. Zhou, Y. Yu, and S. Du, "Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1782–1792, Jul. 2017.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[19] J. Sochor, A. Herout, and J. Havel, "BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3006–3015.

[20] S. Yu, Y. Wu, W. Li, Z. Song, and W. Zeng, "A model for fine-grained vehicle classification based on deep learning," *Neurocomputing*, vol. 257, pp. 97–103, Sep. 2017.

[21] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[22] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[23] A. Iyer, S. Nath, and S. Sarawagi, "Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 530–538.

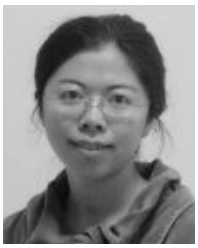
[24] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 136–144.

[25] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *Ann. Stat.*, vol. 41, no. 5, pp. 2263–2291, 2013.

[26] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 97–105.
[27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 3320–3328.
[28] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.
[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

Jitian Wang received the B.S. degree from Xiamen University, Xiamen, China, in 2016,where he is currently pursuing the master's degree with the Department of Communication Engineering, School of Information Science and Engineering. His main research interests include machine learning and image processing.

Han Zheng received the B.S. degree from the Chongqing University of Post and Telecommunication, Chongqing, China, in 2015. He is currently pursuing the master's degree with the Department of Communication Engineering, School of Information Science and Engineering, Xiamen University. His main research interests include machine learning and image processing.



Yue Huang received the B.S. degree from Xiamen University, Xiamen, China, in 2005, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2010. She was a Visiting Scholar with Carnegie Mellon University from 2015 to 2016. She is currently an Associate Professor with the Department of Communication Engineering, School of Information Science and Engineering, Xiamen University. Her main research interests include machine learning and image processing.

Xinghao Ding (M'99) was born in Hefei, China, in 1977. He received the B.S. and Ph.D. degrees from the Hefei University of Technology, Hefei, in 1998 and 2003, respectively. From 2009 to 2011, he was a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, NC, USA. Since 2011, he has been a Professor with the Department of Communication Engineering, School of Information Science and Engineering, Xiamen University, Xiamen, China. His main research interests include image processing, and machine learning.