Computers, Materials & Continua CMC, vol.63, no.1, pp.337-351, 2020

Cold Start Problem of Vehicle Model Recognition under Cross-Scenario Based on Transfer Learning

Hongbo Wang[1, *], Qian Xue[1], Tong Cui[1], Yangyang Li[2] and Huacheng Zeng[3]

Abstract: As a major function of smart transportation in smart cities, vehicle model recognition plays an important role in intelligent transportation. Due to the difference among different vehicle models recognition datasets, the accuracy of network model training in one scene will be greatly reduced in another one. However, if you don't have a lot of vehicle model datasets for the current scene, you cannot properly train a model. To address this problem, we study the problem of cold start of vehicle model recognition under cross-scenario. Under the condition of small amount of datasets, combined with the method of transfer learning, load the DAN (Deep Adaptation Networks) and JAN (Joint Adaptation Networks) domain adaptation modules into the convolutional neural network AlexNet and ResNet, and get four models: AlexNet-JAN, AlexNet-DAN, ResNet-JAN, and ResNet-DAN which can achieve a higher accuracy at the beginning. Through experiments, transfer the vehicle model recognition from the network image dataset (source domain) to the surveillance-nature dataset (target domain), both Top-1 and Top-5 accuracy have been improved by at least 20%.

Keywords: Vehicle model recognition, transfer learning, cold start, and artificial intelligence.

1 Introduction

With the development of artificial intelligence and Internet of Things (IoT) technologies, smart cities emerged as the times require. Smart transportation, as an important public resource, provides real-time traffic monitoring, vehicle management, travel information services, and vehicle auxiliary control functions. The vehicle management system usually includes three processes. First, analyze the video image under the road surveillance camera, extract the vehicle image from it, then use the vehicle image to identify and classify the vehicle model, and finally perform statistics and analysis on the vehicle data, thereby effectively manage urban vehicles. Vehicle model recognition is also helpful for vehicle re-identification task [Szegedy, Liu, Jia et al. (2015); Yang, Luo, Chen et al.

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

[2] National Engineering Laboratory for Public Safety Risk Perception and Control, China Academy of Electronics and Information Technology, Beijing, 100041, China.

[3] University of Louisville, Louisville, KY, 40292, US.

[*] Corresponding Author: Hongbo Wang. Email: hbwang@bupt.edu.cn.

Received: 13 May 2019; Accepted: 10 June 2019.

(2015)] and vehicle tracking task [Liu, Liu, Mei et al. (2016)], and is of great significance for public safety and crime prevention.

Vehicle model identification includes identification of the vehicle brand and model. This problem becomes more complicated and has greater challenge. Because too many brands of various vehicles, and there are many sub-brands in each brand. There are many different models under each sub-brand. It is estimated that there are thousands of models of vehicles on the road, and the appearance of vehicles of many brands is very small. Therefore, it is a very challenging problem to achieve correct identification of so many models.

Regarding the problem of vehicle model recognition, researchers have conducted a lot of researches. The current solutions for vehicle model recognition can be classified in the following categories: the traditional methods are mainly based on the appearance of the car and other inherent features to identify, such as size, shape, texture, etc., as well as some advanced features of edge, color features [Abdelmaseeh, Badreldin, Abdelkader et al. (2012); Dule, Gokmen and Beratoglu (2010); Kim, Park and Choi (2008); Hu, Yang, Bai et al. (2013); Baek, Park, Kim et al. (2007)], Histogram of Oriented Gradient (HoG) [Csurka, Dance, Fan et al. (2004); Behley, Steinhage and Cremer (2013); Lee, Gwak and Jeon (2013)] features, pyramid Histogram of Oriented Gradient (PHOG) [Zhang (2013)] features, etc. There are also some of the most natural and intuitive 3D features used to represent the appearance and location features of the target for fine-grained classification [Krause, Stark and Deng (2013); Lin, Morariu, Hsu et al. (2014); Buch, Orwell and Velastin (2008); Xia, Feng and Zhang (2016); Yu, Zhao, Zheng et al. (2018); Yang, Luo, Wang et al. (2018); Sun, Yuan, Zhou et al. (2018)].

Most datasets currently used in the research are high quality pictures selected from the Internet, such as car forums, most of which are provided by car dealers. The resolution of the pictures is very high. The angle is mainly positive, and the light is good. There is little interference from other objects in the picture. In reality, the image comes from the surveillance camera. These images are of low quality, the resolution is relatively low, the angle is varied, the light changes with the daylight conditions, and there is object obstruction. Fig. 1 shows some contrast pictures. The diversity of the environment makes the vehicle dataset unique in its own context. The above pictures are in the CompCars dataset [Yang, Luo, Chen et al. (2015)], captured from the network and car forums. These pictures are clear and the resolution of the pictures is very high. The pictures below show the images from our surveillance camera. The lighting conditions are not good and the resolution of the pictures is very low, and there is occlusion in images.

In different scenes, such as a different background, the effect of the model will drop a lot. For example, after a model is trained on the A dataset, the accuracy on the A dataset is high, but the accuracy is much lower than A on the new B dataset, and the B dataset has the same task with A dataset. Tafazzoli et al. [Tafazzoli, Frigui and Nishiyama (2017)] performed experiments with CompCars-51 and VMMRdb-51 datasets, respectively, and CompCars-51 as the training dataset. The Top-1 accuracy on this dataset is 96.88%, while in VMMRdb-51 the accuracy is only 36.10%. With VMMRdb-51 as the training set, the accuracy of Top-1 on this dataset is 90.26%, and the accuracy on CompCars-51 is only 40.28%.

We also do experiments to verify this point. Specifically, we collected and produced a

vehicle model identification data set under the surveillance camera, collected 71 days of data from 178 surveillance cameras, and marked 431 models with a total of 170,304 images, as shown in the lower part of Fig. 1. Then we use AlexNet [Krizhevsky, Sutskever and Hinton (2012)] and ResNet [He, Zhang, Ren et al. (2016)] models, and use CompCars [Yang, Luo, Chen et al. (2015)] dataset as the training dataset, and our dataset as the test dataset. We use two evaluation criteria, Top-1 accuracy and Top-5 accuracy. Top-1 accuracy refers to the accuracy of first category of the model output same with the actual result. Top-5 accuracy refers to the accuracy of top five categories of the model output contain the actual results. Our result shows that the accuracy after transfering has been significantly improved. AlexNet's Top-1 accuracy on the training set was dropped from 78.99% to 31.39% and Top-5 was dropped from 92.67% to 59.09%. ResNet's Top-1 accuracy also dropped from 66.34% to 30.34%, and Top-5 accuracy dropped from 86.28% to 55.46%.

Figure 1: Data set comparison

The difference between datasets in different scenarios poses a big problem for practical engineering applications. For a new smart transportation project, when the system is just put into operation, it is always desirable to get a higher recognition accuracy in a shorter time, because there is little labeled data. Therefore, it can only rely on the model that was previously trained in other scenarios which is similar to the current task, reaching a better result. In view of the uniqueness of the data set in different scenarios, in order to quickly realize the recognition of the vehicle brand in a new scene when lacking the annotated data, it is necessary to solve an important and challenging problem, cold start. This is the cold start problem, in a new scenario, lacking label data, to realize the vehicle model recognition and get a good result on the task.

In order to solve the cold start problem, this paper resorts to the transfer learning method. By reducing the difference between the source domain (one scenario) and the target domain (another scenario), even in different actual scenarios in case of small labeling images of target domains, to achieve vehicle model recognition and classification problems in different scenarios. In this paper, the DAN (Deep Adaptation Networks) Long et al. [Long, Cao, Wang et al. (2015)] and JAN (Joint Adaptation Networks) [Long, Zhu, Wang et al. (2016)] domain adaptation modules in transfer learning are added to the AlexNet [Krizhevsky, Sutskever and Hinton (2012)] and ResNet [He, Zhang, Ren et al. (2016)] networks to generate four network structures, AlexNet-JAN, AlexNet-DAN, ResNet-50-JAN, and ResNet-50-DAN. We realized the transfer from the CompCars [Yang, Luo, Chen

et al. (2015)] dataset (source domain) to our dataset (target domain) in the vehicle model recognition using AlexNet-DAN model, and obtained 62.09% Top-1 accuracy, 85.44% Top-5 accuracy, before transferring the Top-1 accuracy is only 31.39% and the Top-5 accuracy is 59.09%, both Top-1 and Top-5 accuracy have been improved by at least 20%-30%. Overall, we provide a method for cold start problems in vehicle model recognition cross-scenario, with an acceptable result in the absence of annotated data. And we also collect and produce a large data quantity surveillance camera sourced vehicle image dataset marked by vehicle models to fill the gaps in the vehicle brand identification field where there are few vehicle image datasets for surveillance cameras.

The rest of this paper is organized as follows. The Section 2 discusses the related work. In Section 3, we introduce the transfer learning method and the network structure used in the experiment. The details of dataset used in the experiment, experimental setup, experimental results and analysis are introduced in Section 4. We give our conclusions in Section 5.

2 Related work

There are many ways about the fine-grained classification problem. Krause et al. [Krause, Stark and Deng (2013)] used a 3D geometric classifier HOG-SVM to classify 196 models, achieving an accuracy of 67.6%. In recent years, the development of convolutional neural networks has brought a revolutionary breakthrough in image recognition. AlexNet, ResNet, VGG [Simonyan and Zisserman (2014)], Densenet [Huang, Liu and Weinberger (2016)], and other networks have achieved good results in ILSVRC every year. More and more classification problems have also started to use deep learning methods. Vehicle model recognition is no exception, it belongs to the fine-grained classification problem [Lin, Morariu, Hsu et al. (2014); Buch, Orwell and Velastin (2008); Xia, Feng and Zhang (2016); Yu, Zhao, Zheng et al. (2018)]. Yang et al. [Yang, Luo, Chen et al. (2015)] use the GoogLeNet network to classify 431 vehicle models. The Top-1 accuracy is 76.70%, and the Top-5 accuracy is 91.70%. Similarly, this paper uses the two commonly used deep learning networks AlexNet and ResNet to conduct experiments.

There are many public datasets in the field of vehicle model recognition. The Tang team Yang et al. [Yang, Luo, Chen et al. (2015)] published the CompCars dataset, which contains images from the network and surveillance cameras. The network images are collected from car forums, public websites and search engines, covering most commercial vehicle models in the past decade. A total of 136,727 vehicle images contain various angles of the vehicle; surveillance images are collected by surveillance cameras and contain 50,000 front view car pictures. Krause et al. [Krause, Stark and Deng (2013)] established two datasets of 10-BMW and 197-car, which are also pictures taken from the Internet. Tafazzoli et al. [Tafazzoli, Frigui, and Nishiyama (2017)] disclosed that the VMMR dataset contains 9,170 categories with a total of 291,752 images, covering models from 1950 to 2016. Although these data sets are numerous and have many categories, there are certain differences between the different datasets of vehicle models. When one dataset is used for training and the other dataset is used to test, the effect will be much lower. Therefore, using these public datasets to solve the problem of vehicle model recognition in another scene is not ideal.

Different from most current vehicle model recognition studies, this paper mainly focuses

on the cold start problem in vehicle model recognition proposed above. We use the method of transfer learning to achieve domain adaptation from network images to surveillance images. Transfer learning is widely used in many fields, such as machine learning and data mining. When data distribution changes from one domain to another, many models need to be rebuilt with new training data, while transfer learning can avoid a lot of expensive data tagging work.

Yi et al. [Yi, Lei, Liao et al. (2014)] applied transfer learning to deep learning and proposed a depth transfer metric learning (DTML) method. It has achieved good results in cross-dataset face recognition and human re-identification. Shen et al. [Shen, Qu, Zhang et al. (2017)] proposed a new method called Wasserstein Distance Guided Representation Learning (WDGRL) for the learned domain invariant feature representations, which has been well validated on the adaptive datasets of emotion and image classification. Their work proves that transfer learning has a better role in deep learning and image classification. This paper applies transfer learning to vehicle model recognition. By adding the DAN and JAN domain adaptation modules in transfer learning to the AlexNet and ResNet-50 networks, we achieve transfering from CompCars (source domain) to our data set in vehicle model recognition (target Domain).

3 Network model

3.1 Transfer learning

Transfer learning is not only widely used in the fields of machine learning and data mining. With the development of deep learning, many theories of transfer learning are gradually applied to the field of deep learning, and have achieved good results. Yosinski et al. [Yosinski, Clune, Bengio et al. (2014)] demonstrated through experiments that the first few layers of the neural network are basically general features, and that the effect of transferring the first few layers will be better. If fine-tune is added in the depth transfer network, the effect will be improved. It may be better than the original network, and the transfer of network layers can accelerate the learning and optimization of the network.

Domain adaptation is an important research direction in the field of transfer learning. Common domain adaptation includes instance-based adaptation, feature representation-based adaptation, classifier-based adaptation, where in the unsupervised case, because there are no target labels, so classifier based adaptation is not feasible.

3.2 DAN and JAN

There are many ways of domain adaptation, such as SHL-MDNN, DAN, JAN. DAN maps hidden layers related to learning tasks in CNN to the reconstructed nuclear Hilbert space, and minimizes the distance between different domains through multicore optimization. The DAN module is based on the DDC (Deep Domain Confusion) method proposed by Tzeng et al. [Tzeng, Hoffman, Zhang et al. (2014)] at the University of California, Berkeley. The DAN method adds three adaptive layers to the field adaptation learning in the three layers before the classifier. And DAN adopts the multiple kernel maximum mean discrepancy (MK-MMD), moreover the DAN method integrates the parameter learning of MK-MMD into the training of convolutional neural networks, but

does not increase the training time of the network. DAN method is show in Fig. 2.

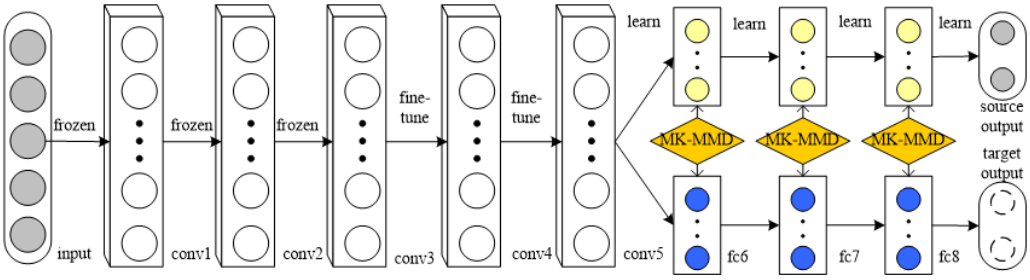


Figure 2: DAN method

The formula for calculating MK-MMD is:

$$\mathcal{K} \triangleq \{k = \sum_{u=1}^{m} \beta_u k_u : \beta_u \geq 0, \forall u\} \tag{1}$$

DAN's optimization goal consists of two parts: loss function and adaptive loss. The distribution distance is the MK-MMD distance we mentioned above. Therefore, DAN's optimization goal is:

$$\min_{\Phi} \frac{1}{n_a} \sum_{i=1}^{n_a} J(\theta(x_i^a), y_i^a) + \lambda \sum_{l=l_1}^{l_2} d_k^2(\mathcal{D}_s^l, \mathcal{D}_t^l) \tag{2}$$

JAN proposed a new joint distribution distance measurement relationship, using this relationship to generalize the transfer learning ability of the depth model to adapt the data distribution in different fields, and simultaneously performs adaptive and confrontational learning of joint distribution in deep networks. JAN method is show in Fig. 3.

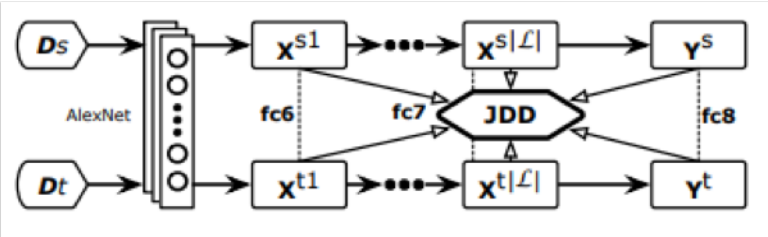


Figure 3: JAN method

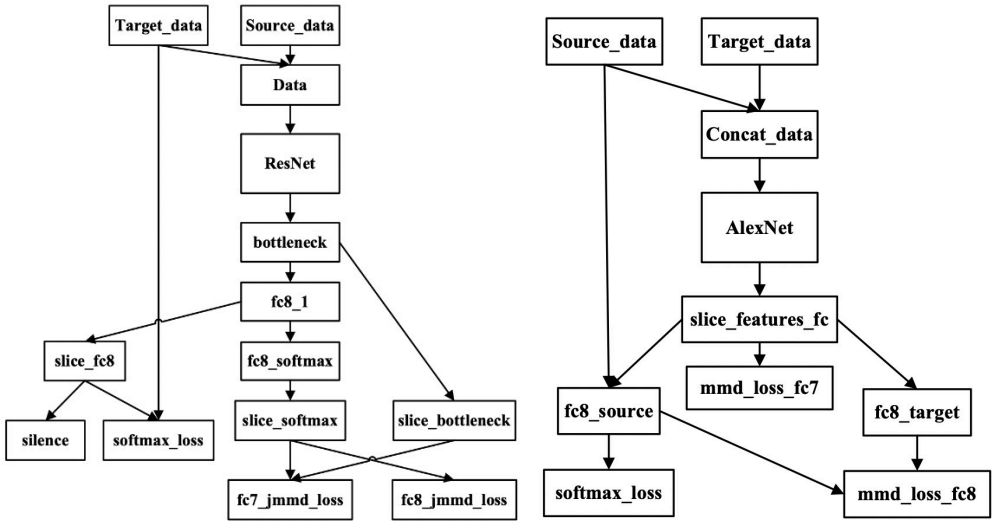
The JAN method extends the adaptive method of data to the adaptation of categories, and proposes JMMD metrics (Joint MMD):

$$\widehat{D_{\mathcal{L}}}(P, Q) = \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} \prod_{\ell \in \mathcal{L}} k^{\ell}\left(z_i^{sl}, z_j^{sl}\right) + \frac{1}{n_t^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \prod_{\ell \in \mathcal{L}} k^{\ell}\left(z_i^{tl}, z_j^{tl}\right)$$
$$- \frac{1}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \prod_{\ell \in \mathcal{L}} k^{\ell}(z_i^{sl}, z_j^{tl}) \tag{3}$$

Similar to DAN, JAN's optimization goal is the sum of the loss function of the difference between the predicted and actual values of JMMD and its own loss function. The optimization goals of JAN are as follows:

$$\min_{f} \max_{\theta} \frac{1}{n_s} \sum_{i=1}^{n_s} J\left(f(x_i^s), y_i^s + \lambda \widehat{\mathcal{D}_l}(P, Q, \theta)\right) \tag{4}$$

In this paper, DAN and JAN were applied to experiments in AlexNet and ResNet respectively. Using other similar transfer learning methods, similar experimental results should be obtained. We just selected DAN and JAN, because they have stable performance in the field of deep transfer learning. The JAN module was added to the first three layers of the AlexNet and ResNet-50 classifiers to form AlexNet-JAN and ResNet-50-JAN. JAN is more complicated than DAN. It calculates the JMMD loss by the knowledge of the training in the previous layers through multiple bottleneck layers, and cross-processes the JMMD losses with the subsequent layers, plus the softmax error calculation of the classification optimization itself to optimize the network together. At the last few layers of the network, added the domain adaptation algorithm. Generated four network structures, AlexNet-JAN, AlexNet-DAN, ResNet-50-JAN, and ResNet-50-DAN. Fig. 4 shows ResNet-50-JAN and AlexNet-DAN network structure. For ResNet-50-JAN, the domain adaptation module is added after the last bottleneck and fc-8 layer. For the AlexNet-DAN network, the domain adaptation module is added after the fc-7 layer.



(a) ResNet-50-JAN (b) AlexNet-DAN

Figure 4: ResNet-50-JAN and AlexNet-DAN

4 Experiments

In this section, we use the two datasets, CompCars and our surveillance dataset collected under the surveillance camera to conduct experiments. First, we use AlexNet and ResNet-50 to perform fine-tune training on CompCars's network images to obtain a model with better classification effect in the source domain. Then, we added the domain adaptation modules JAN and DAN to AlexNet and ResNet-50, respectively, to improve the classification effect of the surveillance dataset on networks models and achieve the purpose of cold start.

4.1 Datasets

At present, there are many problems in the data set used in the field of vehicle model recognition. Take CompCars as an example: CompCars data set has three main shortcomings in the data set under the surveillance camera. First, their surveillance cameras have a total of 50,000 images, but they are mostly taken from the same surveillance camera in three time periods. The illumination conditions are very poor and the clarity is very low. It is difficult for people to distinguish the vehicle model of these two models of photos, let alone the classification task of vehicle brand recognition. These two kinds of pictures account for a large proportion of the total 50,000 pictures, and it is impossible to judge whether the classification of these pictures is correct. Secondly, some of the 50,000 images in CompCars dataset have obvious classification errors. It can be seen that their datasets are only filtered by machine. Thirdly, most of the pictures in the same classification are taken repeatedly by several vehicles. Some rare models even have only one car taken repeatedly, which leads to poor diversity of the pictures under the same classification and makes it difficult to extract rich vehicle brand features.

We collected 71 days of data from 178 surveillance cameras in the actual scene of the smart city. We made a dataset, VMRSD1. Our pictures have different weather, different lighting conditions (day, night), different angles. Then we used the object detection model to intercept the vehicle picture, and according to the standard of the CompCars dataset, we classified the images into 431 categories. We finally got 170,304 pictures, due to the position of the camera, most of the pictures are front or back. Some of these classifications did not capture the corresponding images due to differences in geographic area and time. As the Fig. 5 shows, N is the number of pictures for each category. The vertical axis represents the number of categories in which the number of pictures is within a certain range. There are 169 categories with fewer than 50 images, 195 categories with fewer than 100 images, and 50 categories with greater than 1,000 images.

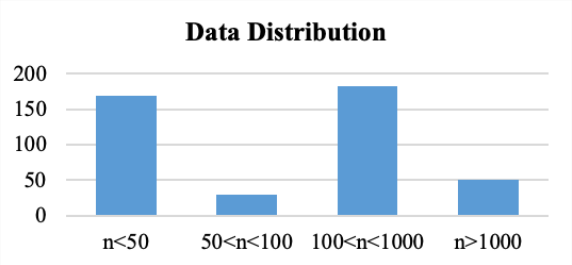


Figure 5: The distribution of datasets

The dataset in our experiment is divided into two parts, one is the datasets of the network pictures, mainly used in the source domain for fine-tune training and field adaptation training; and the dataset from the surveillance camera pictures, is mainly used for field adaptation, and are adapted to the accuracy test before and after training. The dataset of the network image has selected from the CompCars dataset proposed in Yang et al. [Yang, Luo, Chen et al. (2015)].

4.2 Fine-tune experiments

In order to verify the deep network model in the Imagenet classification problem and fine-tune the pre-training parameters, it has a good effect in solving the problem of fine-grained classification of network vehicle images. we chose two classification models in deep learning, AlexNet and ResNet-50. Verified the effectiveness of the deep learning classification network model in solving the fine-grained classification problem of vehicle classification.

The hardware condition we use is Nvidia GTX1080Ti with 11 G memory capacity. In our experimental conditions, AlexNet's batch-size can be set very large. After a comprehensive consideration, we used 256 as the batch-size, which accounted for about 4 G in the training process and 60,000 iterations for about 6 hours. The ResNet-50 model is much more complex than AlexNet, and the batch-size can only be set to 32. At this time, the memory is about 10 G or more, and 600,000 iterations take about 100 hours.

Observe the first 20,000 iterations of AlexNet, we can see that due to the large batch-size, the recognition accuracy of the model increases very quickly, and the loss and accuracy curve during the entire iteration changes smoothly, shown in Fig. 6. When the iteration reaches 20,000 times, the Top-1 accuracy is 76.16%, Top-5 accuracy is 90.45%. After 60,000 iterations, our Top-1 accuracy reached 78.99% and Top-5 reached 92.67%, exceeding GoogLeNet's 76.70% Top-1 accuracy and 91.70% Top-5 accuracy. The recognition accuracy in subsequent iterations did not change significantly, so we ended our training experiment.

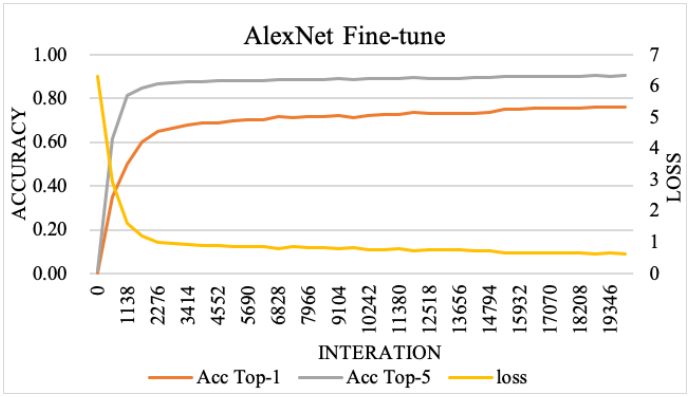


Figure 6: AlexNet fine-tune, accuracy and loss curve of the train process

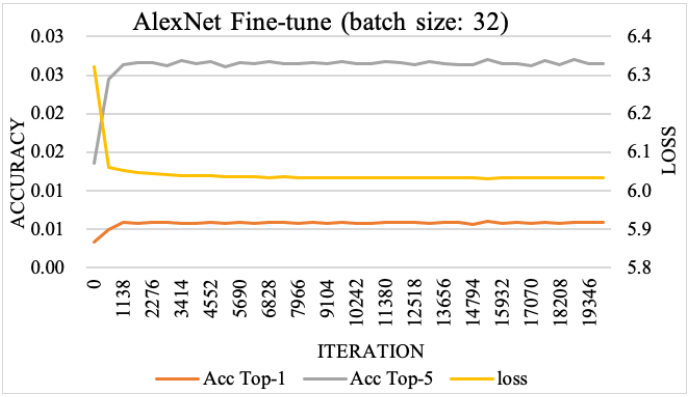


Figure 7: AlexNet fine-tune, accuracy and loss curve of the train process

The ResNet theoretically can achieve better results than AlexNet, but in actual training, we find that when the batch size is 32, the recognition accuracy of ResNet-50 after first 40,000 iterations is improved, but is not faster than AlexNet, the curve jitter is severe during the iteration, as the Fig. 7 shows. When iterated to 40,000 times, the Top-1 accuracy is 53.81%, and Top-5 accuracy is 77.98%. At 600,000 iterations, our Top-1 accuracy reached 66.34% and Top-5 accuracy reached 86.28%, which is lower than the GoogLeNet and AlexNet.

To verify that ResNet's poor recognition is related to the smaller batch-size, we added a fine-tune training experiment on AlexNet with the same 32 batch-size. The results after 20,000 iterations are shown in Fig. 8. It can be seen that AlexNet did not even converge at the lower batch-size. The reason is that when calculating MMD, it is best to use all data in the dataset, but in practice, people usually only use the value of one batch to calculate. If the batch-size is too small, the MMD cannot show the real distance, which affects the network convergence.

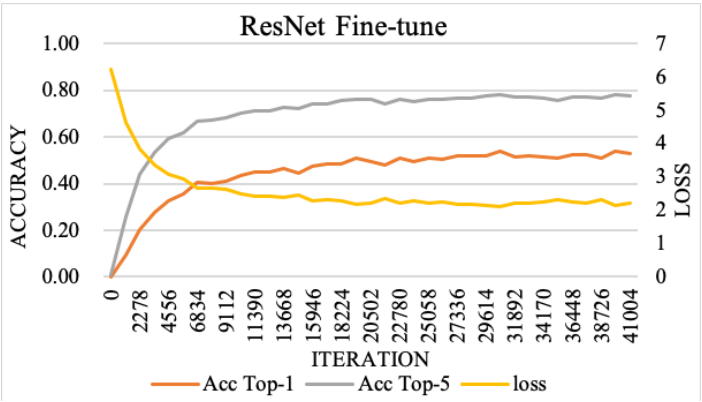


Figure 8: ResNet fine-tune, accuracy and loss curve of the train process

4.3 Adaptation experiments

In the domain adaptation experiments, we use the dataset in fine-tune training as the source domain, and use the 170,304 images collected from our surveillance camera as the target domain and test set part. We rewritten the network structures of AlexNet and ResNet-50, added the JAN module and the DAN module, and generated four network structures: AlexNet-JAN, AlexNet-DAN, ResNet-50-JAN, and ResNet-50-DAN. We conducted some experiments. Similar to the fine-tune part, we set the parameters of the source and target domains with a batch-size of 128 for two AlexNet experiments. The number of training iterations is still 60,000. In the two ResNet experiments, the parameter batch-size is 16 in both source and target domain, and the number of training iterations is still 60,000.

From the results of Tab. 1, using JAN and DAN to perform domain adaptation training on the trained ResNet-50 and AlexNet models, both Top-1 and Top-5 accuracy have been improved by 20%-30%. Among them, ResNet-50 has a smaller promotion than AlexNet due to the smaller batch-size. AlexNet-JAN received 87.95% Top-5 accuracy, which is 2.66% lower than the Top-5 accuracy of the source network image. AlexNet-DAN received 62.09% Top-1 accuracy, which is 11.04% lower than the Top-1 accuracy of the source network image. At the same time, due to the narrowing of the gap between the two different domains, the recognition accuracy of the source domain with network picture is slightly lower than before the field adaptation.

Table 1: Accuracy on source domain and target domain

Model	Target Domain		Source Domain	
	Top-1	Top-5	Top-1	Top-5
GoogLeNet	-	-	76.70	91.70
AlexNet	31.39	59.09	**78.99**	**92.67**
AlexNet+JAN	57.38	87.95	72.18	90.61
AlexNet+DAN	**62.09**	**85.44**	73.13	90.97
ResNet	30.34	55.46	66.34	86.28
ResNet+JAN	50.80	75.54	64.98	85.69
ResNet+DAN	52.39	77.62	64.44	86.12

We also selected some characteristic pictures and conducted a single picture recognition test. Fig. 9 shows the Top-5 predicted classes of the classification model for six cars in our surveillance-nature data. Below each image is the ground truth class and the probabilities for the Top-5 predictions with the correct class labeled in red, on the contrary, it is blue. We can see when the picture's quality is better, the picture is well lit, unobstructed, and the angle is positive, the result of the recognition is better as expected.

Figure 9: Some examples

Among them, the first picture has excessive exposure. Although the picture of the tail of the vehicle is taken, the shape of the vehicle logo and the overall vehicle tail are difficult to distinguish by the naked eye due to overexposure, but the model still recognizes with 64.59% confidence. The Cadillac XTS is out, and the Confidence Top-3 is the Cadillac series. The confidence level is more than 90%, which means that the characteristics learned by deep learning do not depend on the specific logo or other details. The second picture is more ambiguous, with slight occlusion on the top, and the light is darker. However, because the angle is relatively correct, the vehicle face information is more accurate, and the model is not only accurate but also highly reliable. The third picture belongs to the picture quality is poor, and there is time occlusion above, the car face information is fuzzy, although the confidence of the overall recognition is not very high, but the model identifies the correct vehicle brand with Top-1 confidence.

In addition, some cases of inaccurate identification were selected. There is a common situation in which the recognition accuracy is low. The vehicle brand is relatively unpopular, and the corresponding training set has fewer pictures. At this time, because the neural network extracts fewer features, it is prone to classification errors, but it is still similar to the fourth picture. The situation identified in Top-5. As in the fifth picture, when the overall quality of the picture is poor, the overall recognition confidence is low, and the correct model does not appear in the Top-5 confidence level. The sixth picture is a typical case where a number of similar models lead to misidentification. Although the vehicle picture is clear and complete, because the Volkswagen models are too similar, the highest recognition confidence is not the correct result. At the same time, the model also has a certain recognition effect on the situation of shelter, poor light, and overexposure. The reason for recognition poor pictures is mainly due to interference from several similar vehicles, blurry pictures, and some categorizations contain very few pictures.

4.4 Train with our dataset

Of course, if there are enough annotated images, we still recommend using the pre-trained model in ImageNet directly for finetune. In this regard, we also randomly assigned the 170,304 images we collected to the training set and the test set at a ratio of 70% and 30%. Conduct fine-tune test, after 300,000 iterations, the Top-1 accuracy is 89.5%, Top-5 accuracy is 95.4%. Since the pictures are mostly based on the front and

back, the same car is photographed with similar angles, the complexity of the data set is lower, and the model will be better. It also verifies that for different datasets, due to the differences in the distribution of datasets, the trained models will have different biases, and it is difficult to meet the needs of various scenarios.

5 Conclusions

In this paper, to address the cold start problem of vehicle model recognition under cross-scene, we combined transfer learning, using only a small amount of annotation data to realize vehicle model recognition under real surveillance camera. We load the DAN and JAN domain adaptation modules into the AlexNet and ResNet-50, and train the networks. We also collect and produce a vehicle model identification data set under the surveillance camera, and mark 431 vehicle models with a total of 170,304 images. Then transfer the dataset (source domain) to our surveillance-nature picture data set (target domain), the Top-1 accuracy is 62.09% and Top-5 accuracy is 85.44%, both Top-1 and Top-5 accuracy have been improved by 20%-30% after transferring.

Acknowledgement: This work was supported by CETC Joint Research Program under Grant 6141B08020101, 6141B08080101, National Key R&D Program of China under Grant 2018ZX09201014, and the National Natural Science Foundation of China under Grant 61002011.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Abdelmaseeh, M.; Badreldin, I.; Abdelkader, M. F.; El Saban, M. (2012): Car make and model recognition combining global and local cues. *Pattern Recognition*, pp. 910-913.

Baek, N.; Park, S. M.; Kim, K. J. (2007): Vehicle color classification based on the support vector machine metho. *International Conference on Intelligent Computing*, pp. 1133-1139.

Behley, J.; Steinhage, V.; Cremers, A. (2013): Laser-based segment classify-cation using a mixture of bag-of-words. *International IEEE/RSI Conference on Intelligent Robots and Systems*.

Ben-david, S.; Blitzer, J.; Crammer, K. (2006): Analysis of representations for domain adaptation. *International Conference on Neural Information Processing Systems*, pp. 137-144.

Buch, N.; Orwel, J.; Velastin, S. A. (2008): Detection and classification of vehicles for urban traffic scenes. *Visual Information Engineering*, pp. 182-187.

Csurka, G.; Dance, C.; Fan, L. (2004): Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision ECCV*, vol. 44, pp. 1-22.

Dule, E.; Gokmen, M.; Beratoglu, M. S. (2010): A convenient feature vector construction for vehicle color recognition. *Proceedings of 11th WSEAS International Conference on*

Neural Networks, Evolutionary Computing and Fuzzy Systems, pp. 250-255.

Fu, J.; Zheng, H.; Mei, T. (2017): Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Hu, W.; Yang. J.; Bai, L. (2013): A new approach for vehicle color recognition based on specular-free image. *Proceedings of SPIE-the International Society for Optical Engineering*, pp. 906-917.

Huang, G.; Liu, Z.; Weinberger, K. Q. (2016): Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708.

Kim, K. J.; Park, S. M.; Choi, Y. J. (2008): Deciding the number of color histogram bins for vehicle color recognition. *Asia-Pacific Services Computing Conference*, pp. 134-138.

Krause, J.; Stark, M.; Deng, J.; Li, F. F. (2013): 3D object representations for fine-grained categorization. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554-561.

Krizhevsky, A.; Sutskever, I.; Hinton, G. (2012): Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

Lee, S.; Gwak, S.; Jeon, M. (2013): Vehicle model recognition in video. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 2.

Leotta, M. J.; Mundy, J. L. (2011): Vehicle surveillance with a generic, adaptive, 3D vehicle model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1457-1469.

Lin, Y. L.; Morariu, V. I.; Hsu, W.; Davis, L. S. (2014): Jointly optimizing 3D model fitting and fine-grained classification. *Computer Vision-ECCV*, pp. 466-480.

Liu, X.; Liu, W.; Mei, T. (2016): A deepLearning-based approach to progressive vehicle re-identification for urban surveillance. *European Conference on Computer Visio*, pp. 869-884.

Long, M.; Cao, Y.; Wang, J.; Jordan, M. I. (2015): Learning transferable features with deep adaptation networks. *Proceeding ICML'15 Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 97-105.

Long, M.; Zhu, H.; Wang, J.; Jordan, M. I. (2016): Deep transfer learning with joint adaptation networks. *International Conference on Machine Learning*.

Shen, J.; Qu, Y.; Zhang, W. (2017): Wasserstein distance guided representation learning for domain adaptation. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Sun, M.; Yuan, Y.; Zhou, F.; Ding, E. (2018): Multi-attention multi-class constraint for fine-grained image recognition. *Computer Vision and Pattern Recognition*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. et al. (2015): Going deeper with

convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*.

Tafazzoli, F.; Frigui, H.; Nishiyama, K. (2017): A large and diverse dataset for improved vehicle make and model recognition. *Computer Vision and Pattern Recognition Workshops*, pp. 874-881.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; Darrell, T. (2014): Deep domain confusion: maximizing for domain invariance. *Computer Vision and Pattern Recognition*, arXiv:1412.3474

Wang, Y.; Morariu, V. I.; Davis, L. S. (2016): Learning a discriminative filter bank within a CNN for fine-grained recognition. *Computer Vision and Pattern Recognition*.

Wang, Z.; Tang, L.; Liu, X. (2017): Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. *IEEE International Conference on Computer Vision*, pp. 379-387.

Xia, Y.; Feng, J.; Zhang, B. (2016): Vehicle logo recognition and attributes prediction by multi-task learning with CNN. *International Conference on Natural Computation and 13th Fuzzy Systems and Knowledge Discovery*.

Yang, L. J.; Luo, P.; Chen, C. L.; Tang, X. O. (2015): A large-scale car dataset for fine-grained categorization and verification. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3973-3981.

Yang, Z.; Luo, T.; Wang, D. (2018): Learning to navigate for fine-grained classification. *European Conference on Computer*, pp. 420-435.

Yi, D.; Lei, Z.; Liao, S.; Li, S. Z. (2014): Deep metric learning for person re-identification. *IEEE International Conference on Pattern Recognition*.

Yosinski, J.; Clune, J.; Bengio, Y. (2014): How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, vol. 27, pp. 3320-3328.

Yu, C.; Zhao, X.; Zheng, Q. (2018): Hierarchical bilinear pooling for fine-grained visual recognition. *European Conference on Computer Vision*, pp. 574-589.

Zhang, B. (2013): Reliable classification of vehicle types based on cascade classifier ensembles. *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 322-332.