MULTI-SCALE VEHICLE RE-IDENTIFICATION USING SELF-ADAPTING LABEL SMOOTHING REGULARIZATION

Yue Xu, Na Jiang, Lei Zhang, Zhong Zhou, Wei Wu*

State Key Laboratory of Virtual Reality Technology and Systems
Beihang University, Beijing, P.R.China
zz@buaa.edu.cn

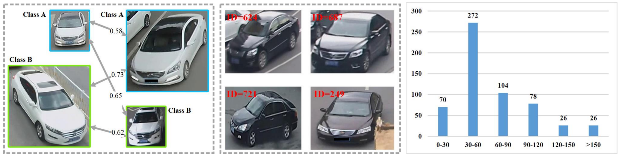
ABSTRACT

Vehicle re-identification (re-id) plays an important role in intelligent surveillance. Since difference vehicle models may have similar appearances, together with the problem of image scale variations, the vehicle re-id remains long-term challenging. We present a novel multi-scale vehicle re-id framework using self-adapting label smoothing regularization (SLSR). It integrates the appearance information from multi-scale images to alleviate the influence of scale changes caused by perspectives. To enhance the generalization ability in feature representations, we design the self-adapting label smoothing regulation in semi-supervised training process. It dynamically assigns labels to fake images to realize data augmentation. We validate the effectiveness of our proposed framework on popular VeRi and VehicleID datasets. Extensive experimental results demonstrate that our method outperforms most state-of-the-art methods on both datasets. Especially, we exceeds the latest method by 3.81% in mAP and 5.32% in rank-1 on VeRi dataset.

Index Terms— Vehicle Re-Identification, Semi-Supervised, Multi-Scale, Deep Neural Network

1. INTRODUCTION

Vehicle re-identification (re-id) is an important task in the field of computer vision, which refers to retrieve specified targets from large-scale gallery image set. In existing methods, license plate information [1] and appearance features [2, 3, 4, 5, 6, 7] are the key cues for fine-grained classification. Since that the license plate is easily to be occluded and falsified, this paper focuses on appearance-based vehicle re-id. With the rapid development of deep learning, DRDL [5], VAMI [6], and OIFE [3] are proposed and have achieved good performances on popular vehicle re-id datasets [1, 2]. However, the scale variations caused by the perspective and the low discrimination among vehicles with same model still need to be solved in practical applications.



(a) (b) (c)

Fig. 1. Difficulty analysis of VeRi dataset. (a) Error match caused by image scales variation. The values represent the similarity scores between image pairs. (b) Different cars with similar models. (c) Imbalanced distribution of data size in VeRi dataset. The horizontal axis indicates number of images in a class. The vertical axis denotes class number.

Taking popular VeRi dataset as an example, image size variations have obvious influence in similarity metric (Fig. 1(a)). The same/simliar model reduces the discrimination of the vehicle images (in Fig. 1(b)). To solve these problems, many methods have been proposed. Wang et al. [3] utilize 20 key point locations to locate ROIs for local feature extraction and then compose them into embedded features. Sockor et al. [6] estimate the image viewpoint and 3D bounding box to unpack the vehicle image into a plane for feature learning. Based on these methods, Jiang et al. [7] proposed intra-class similarity function and spatial-temporal re-ranking strategy, and get further improvement on Rank-1 accuracy. However, these methods are still limited by imbalance data distribution and insufficient data. As shown in Fig. 1(c), the number of images in each class varies greatly. Some classes in VeRi dataset have more than 150 images, while there are less than 30 images in some classes.

Because of the imbalanced data, the network pays more attention to the classes with enough samples, while ignoring the classes which lack samples. When the data distributions are quite different between the train set and the test set, the generalization ability of the trained network drops sharply. Regarding the three puzzled issues of vehicle re-id shown in Fig. 1, we propose a multi-scale vehicle re-identification framework exploiting self-adapting label smoothing regularization(SLSR). Our method integrates feature maps from different scale inputs by dense connections. The integrated fea-

This work is supported by the Natural Science Foundation of China under Grant No.61872024 and 61572061.

tures include the detailed information from the large images and the abstract information from the small images, which realizes the perception of images in different scales and alleviates the influence of scale variations on similarity metric. We also introduce SLSR as a semi-supervised training strategy. It can dynamically assign trained labels to the GAN generated images (fake images). The fake images and real images, thereby, can break the bottlenecks of insufficient data. To verify the effectiveness of our proposed approach, we conduct on a series of experiments on popular VeRi and VehicleID datasets. The results demonstrate that our approach achieve outstanding performance on Rank-1 and mAP.

2. PROPOSED METHODS

This section describes the detail of our proposed vehicle re-id framework, include the multi-scale feature extraction network (Sec. 2.1) and the semi-supervised learning strategy with SLSR (Sec. 2.2), as illustrated in Fig. 2.

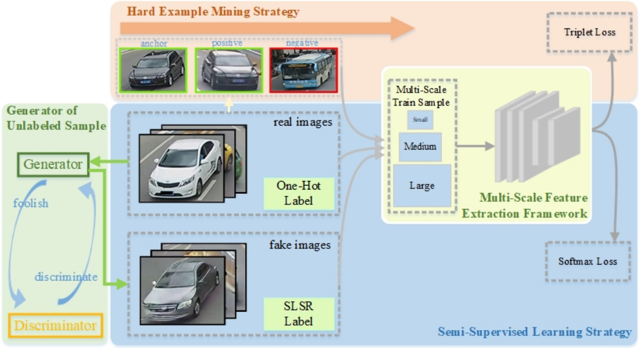


Fig. 2. Illustration of our proposed framework. We train generative adversarial network (GAN) with real images from the train set, then we generate fake images, which serve as auxiliary materials, and jointly trained with real images for data augmentation. Each training sample is resized to 3 scales to train the multi-scale feature extraction network. In addition, we use hard example mining strategy to decrease intra-class distance and increase the inter-class distance, it is not introduced in detail because it is just a trick and not our contribution.(Best viewed in color)

2.1. Multi-Scale Feature Extraction Network

In a real surveillance scenario, the sizes of the captured vehicle images are not fixed because of the different perspectives. The training strategies which resize the images to the same scale directly drop some useful information inevitably. In order to make use of the information contained in images of different scales and make it complementary, we design the scale aware module. We utilize three branches without sharing weights to extract feature maps from multi-scale inputs, and integrate features via densely connection. By adding supervision signal to the three branches, we regulate the classification behavior of each branch simultaneously.

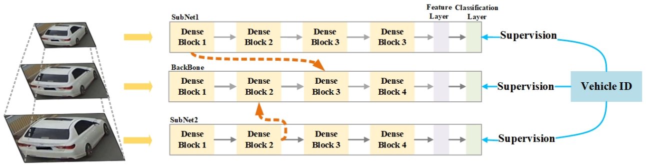


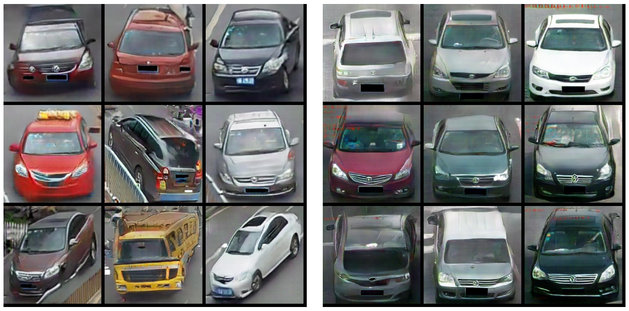
Fig. 3. Details of the proposed multi-scale feature extraction network. It contains three branches. Each images are resized to three scales (112×112, 224×224, 448×448) as inputs to the three branches. The dotted lines indicate the source and target blocks of feature map integration. After the last dense block in each branch, there are feature extraction layer and classification layer. The training of each branch is supervised under the same identity class label constraint concurrently.

The overall network structure of our proposed multi-scale vehicle re-identification network is illustrated in Fig. 3. It consists of three branches for scale-specific inputs, so as to learn the appearance features of the same vehicle image in different scales. The to-be-concatenate feature maps have the same dimension, because we expect feature of each scale has similar effect on the final feature. Instead of concentrating only on the output of backbone, we add supervisory signal on not only the backbone, but also the two subnets. Intuitively, we hope the network can identify a vehicle correctly both at a glance (image of small size) and at a close look (image of large size).

2.2. Semi-supervised Learning Strategy with SLSR

2.2.1. *Generation of Fake Images*

In order to generate fake images, we train DCGAN with real images in vehicle datasets. We show some fake images in Fig. 4. Compared with VeRi in Fig. 4(a), fake images from VehicleID in Fig. 4(b) need less training epochs and achieve substantially higher quality. There may be two reasons for this difference. On one hand, real images in VehicleID have higher native resolution than those in VeRi. On the other hand, VehicleID only contains pictures facing forward or backward, while the orientation of vehicles is various in VeRi.



(a) (b)

Fig. 4. Generated images from DCGAN.

2.2.2. *Self-Adapting Label Smoothing Regularization*

One-hot label only concentrates on the groundtruth class, while LSR [8] pays some attention to other classes. However, LSR cannot be used straightly on fake images, as the they do not have explicit labels. Pesudo-label [9] and LSRO [10] provide two ways of assigning groundtruth to fake images. To design a more appropriate label distribution function, we analyse the fake images.

Statistically, we found that the number of classes that get the highest predicted probability is usually not fixed. As shown in Table 1, where NC refers to the number of classes which have been predicted to have highest response for an image during 30 epochs, and NI refers to the number of images that have the given number of highest responding classes. The statistical result on 19200 images indicates that the predicted probability distribution is not stable. There is no indication that a fake image belongs to a particular class, so it is unconscionable to assign one-hot labels to fake images. What is more, it is always the few classes that get the highest responding. Therefore, rather than treat all classes equally as LSRO, it is reasonable to pay more attention on these classes.

Table 1. Distribution of fake images according to total number of the highest-response classes in 30 epochs.

NC	1	2	3	4	5	more
NI	2400	5740	5068	3449	1772	771

LSR gives a small and equal possibility to every non-groundtruth class, which is effective to avoid over-fitting. LSRO assigns equal possibilities to all classes for fake images, which enables the jointly training of fake images and real images. Inspired by them, we proposed our self-adapting label smoothing regularization (SLSR). During each training step, the class with maximum predicted possibility is treated as groundtruth. The probability of each non-groundtruth class is equal, and it is slightly higher for groundtruth class. In contrast to the one-hot in Eq. 1, label distribution of SLSR q_{SLSR} can be defined as Eq. 2.

$$q_{one-hot}(k) = \begin{cases} 0, & k \neq g \\ 1, & k = g \end{cases} \tag{1}$$

$$q_{SLSR}(k) = \begin{cases} \dfrac{1}{K} - \dfrac{\epsilon}{k}, & k \neq g \\ \dfrac{1}{K} + \dfrac{k-1}{k}\epsilon, & k = g \end{cases} \tag{2}$$

where $k \in \{1, 2, ..., K\}$ is one of the classes of vehicle dataset, and K is the total number of classes, g is the groundtruth of the input image. $\epsilon \in [0, 1]$ decides the probabilitic difference between groundtruth and non-groundtruth classes.

The SLSR is adopted based on the following reasons. It's obvious that the images belonging to the same class or same/similar model have similar appearance, which we refer to as "common features", and each of them has some particular attributes, which are called "unique features" in the following sections. On one hand, for different vehicles with same/similar model, the network may become confused if it focuses too much on the common features. After training DCGAN, the fake images show similar common features to the real images that DCGAN learns from. Giving a SLSR label to the fake images, the network will be punished if it attaches undue importance on common feature and output an extreme high possibility on one class. With the training going on, the network responses more to the unique feature gradually. Therefore, it becomes easy to distinguish vehicles with same/similar models. On the other hand, for the classes with small amounts of samples, network is difficult to understand the common features of them. On the contrary, it may focus too much on the unique features. For example, if there are few samples in a class, and one sample has an unusual color in train set coincidentally, the network will treat the color as a discriminative feature. However, in read world, vehicle images with such a color are not necessarily belonging to the same class. By importing fake samples with similar color into training, if the network makes a wrong prediction toward a labeled sample just because they have a similar and rare color, it will be punished. During back propagation stages, the network can pay attention to other features gradually, obtain a good understanding of the input images and apply relatively fair attentions for features in an image.

3. EXPERIMENTS

To validate the effectiveness of our proposed approach, we conduct experiments on two popular datasets, VeRi-776 [1] and VehicleID [2].

3.1. Dataset Descriptions and Implementation Details

VehicleID contains totally 26267 identities and is split into non-overlapping train/test sets of 111585/110178 images. VeRi includes 776 identities. Specially, there are 1678 queries, 11579 galleries and 37781 training images.

We perform all experiments on the Caffe [11] platform. We select DenseNet-121 as baseline, which connects each layer to every other layers in the same block to strengthen feature propagation and performs well in classification tasks. For the multi-scale feature extraction structure, in the training stage, the pre-training model of DenseNet-121 is used to initialize the parameters of the three branches. The three branches were trained simultaneously, and the losses are added weighted by 0.5, 1, 0.5. In the stage of generating fake images, we randomly initialize a 100 dimensional vector as inputs, the value of each neuron ranges in [-1, 1]. For VeRi dataset, all achieved images are selected as fake inputs. While for the VehicleID dataset, we only randomly select 50,000 images.

3.2. Evaluation of Vehicle Re-ID

We compare our methods with the state-of-the-art in Table 2 and Table 3. Table 2 illustrates the mAP and rank-1 of different methods on VeRi dataset, where we achieve mAP= 65.13% and rank1=91.24%. Table 3 shows rank-1, rank-5 match rate and mAP compared with other methods on VehicleID dataset. All those methods listed in Table 3 adopt the same strategy in [2] to split probe/gallery for testing sets of three scales. We can find that our approach achieves superior performance over other state-of-the art methods on all test sets.

The experimental results indicate that our proposed framework exceeds the latest methods by 3.81% in mAP and 5.32% in rank-1 in VeRi dataset, and exceeds them by 0.7%∼1.8% on rank-1 in different scaled test set of VehicleID dataset. This owes to that the extracted features contain information of multiple scales, and the data augmentation in semi-supervised training stage prevent over-fitting effectively.

Table 2. Comparison with state-of-the-art vehicle re-id methods on VeRi dataset.

Method	**mAP(%)**	**rank-1(%)**
LOMO[12]	9.64	25.3
BOW-CN[13]	12.20	33.9
PROVID[1]	22.77	61.4
KEPLER[14]	33.53	68.7
SiameseCNN+PathLSTM[15]	58.27	83.49
VAMI[6]	61.32	85.92
Ours	**65.13**	**91.24**

Table 3. Comparison with state-of-the-art vehicle re-id methods on VehicleID dataset.

Method	**Match rate**	**K=800**	**K=1600**	**K=2400**
VGG+Triplet Loss[16]	rank-1(%)	40.4	35.4	31.9
Mixed Diff+CLL[2]		49.0	42.8	38.2
DJDL[5]		72.3	70.8	68.0
Ours		**75.1**	**71.8**	**68.7**
VGG+Triplet Loss	rank-5(%)	61.7	54.6	50.3
Mixed Diff+CLL		73.5	66.8	61.6
DJDL		85.7	81.8	78.9
Ours		**89.7**	**86.1**	**83.1**
VGG+Triplet Loss	mAP(%)	44.4	39.1	37.3
Mixed Diff+CLL		54.6	48.1	45.5
DJDL		78.6	74.7	72.0
Ours		**79.3**	**75.4**	**73.3**

3.3. Ablation Studies

3.3.1. Comparison with other semi-supervised methods.

We compare our proposed SLSR with “all in one”, “pseudo-label” and “LSRO” methods, as shown in Table 4. During experiments, the network is always hard to convergence when all fake images are set the same label, is discribed in ”all in one” method. Our proposed SLSR strategy achieves the highest score when ϵ is set to 0.1. Note that this set of experiments is carried out based on the baseline.

Table 4. Comparison of “All in one”, “pseudo label”, “LSRO” and our proposed SLSR on VeRi dataset.

Method	**mAP(%)**	**rank1(%)**
all in one[17]	—	—
pseudo-label	54.90	85.70
LSRO	56.28	87.31
Ours	**57.82**	**88.49**

3.3.2. Contribution of adding different strategies

Table 5 shows the mAP and rank-1 after adding different strategies. Compared with the baseline, our proposed multi-scale framework with SLSR increases mAP by 8.85% and rank-1 by 3.93%. Baseline refers to training real images using DenseNet-121 with hard example mining strategy. The training of our proposed strategy could be divided into four stages. The entries beginning with “+” indicate training from the intermediate result of last step. We gradually add SubNet1 and SubNet2, which are described in Section 2.1, and put the fake images into training. We do not use any unmentioned strategies to ensure the fairness of the comparative experiments.

Table 5. Our results with different strategies on VeRi dataset.

Strategy	**mAP(%)**	**rank-1(%)**
Baseline	56.28	87.31
+SubNet1	60.01	88.64
+SubNet2	61.88	90.01
+Fake images	65.13	91.24

4. CONCLUSION

In this paper, a multi-scale vehicle re-id framework with SLSR is proposed. It contains a multi-scale structure with dense connections, which can alleviate the influence of scale variations by learning and integrating discriminative features from inputs of different scales. And it generates unsupervised samples using DCGAN for data augmentation. Meanwhile, SLSR is utilized to deal with the imbalanced data distribution and enhance the network generalization. Extensive experiments validate the effectiveness of each contribution and demonstrate the excellent performance of our method compared with other state-of-the-art approaches. In future work, we will further explore semi-supervised learning to improve vehicle re-id.

5. REFERENCES

[1] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *European Conference on Computer Vision*. Springer, 2016, pp. 869–884.

[2] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[3] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 379–387.

[4] Jakub Sochor, Jakub Špaňhel, and Adam Herout, "Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, 2018.

[5] Yuqi Li, Yanghao Li, Hongfei Yan, and Jiaying Liu, "Deep joint discriminative learning for vehicle re-identification and retrieval," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 395–399.

[6] Yi Zhou and Ling Shao, "Aware attentive multi-view inference for vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6489–6498.

[7] Na Jiang, Yue Xu, Zhong Zhou, and Wei Wu, "Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 858–862.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[9] Dong-Hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*, 2013, vol. 3, p. 2.

[10] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, vol. 3, 2017.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[12] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.

[13] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[14] Niki Martinel, Christian Micheloni, and Gian Luca Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5645–5658, 2015.

[15] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1918–1927.

[16] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[17] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.