IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 18, 2019, accepted October 9, 2019, date of publication October 22, 2019, date of current version November 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948965

Multi-Label-Based Similarity Learning for Vehicle Re-Identification

SAGHIR ALFASLY[1,4], (Student Member, IEEE), YONGJIAN HU[1,2], (Senior Member, IEEE), HAOLIANG LI[3], TIANCAI LIANG[4], XIAOFENG JIN[4], BEIBEI LIU[1,2], (Member, IEEE), AND QINGLI ZHAO[4]

[1]School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China
[2]Sino-Singapore International Joint Research Institute, Guangzhou 510700, China
[3]Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore 639798
[4]GRG Intelligent Security Institute, Guangzhou 510006, China

Corresponding author: Yongjian Hu (eeyjhu@scut.edu.cn)

This work was supported in part by the Sino-Singapore International Joint Research Institute under Grant 206-A018001 and Grant 206-A017023, in part by the Science and Technology Foundation of Guangzhou Huangpu Development District under Grant 2017GH22, in part by the Science and Technology Foundation of Guangdong Province under Grant 2017A050501002 and Grant 2017A030310320, and in part by the EU Horizon 2020 Project entitled Computer Vision Enabled Multimedia Forensics and People Identification under Grant 690907.

ABSTRACT The massive attention to the surveillance video-based analysis makes the vehicle re-identification one of the current hot areas of interest to study. Extracting discriminative visual representations for vehicle re-identification is a challenging task due to the low-variance among the vehicles that share same model, brand, type, and color. Recently, several methods have been proposed for vehicle re-identification, that either use feature learning or metric learning approach. However, designing an efficient and cost-effective model is significantly demanded. In this paper, we propose multi-label-based similarity learning (MLSL) for vehicle re-identification obtaining an efficient deep-learning-based model that derives robust vehicle representations. Overall, our model features two main parts. First, a multi-label-based similarity learner that employs Siamese network on three different attributes of the vehicles: vehicle ID, color, and type. The second part is a regular CNN-based feature learner that employed to learn feature representations with vehicle ID attribute. The model is trained jointly with both parts. In order to validate the effectiveness of our model, a set of extensive experiments has been conducted on three of the largest well-known datasets VeRi-776, VehicleID, and VERI-Wild datasets. Furthermore, the parts of the proposed model are validated by exploring the influence of each part on the entire model performance. The results prove the superiority of our model over multiple state-of-the-art methods on the three mentioned datasets.

INDEX TERMS Deep convolutional neural network, discriminative features, multi-label-based similarity learning, metric learning, vehicle re-identification.

I. INTRODUCTION

The task of extracting robust visual representations is the cornerstone of building all effective algorithms for computer vision applications. This task differs from one application to another in its complexity, where some applications consider it as a challenging task due to the minimal variations that can be extracted to distinguish instance from another, which can be found apparently in fine-grained classification, re-identification and face recognition. Recently, vehicle image analysis has widely attracted the attention of researchers due to the revolution in artificial intelligence techniques particularly convolutional neural networks (CNNs). This revolution leads to a massive improvement in intelligence public security and public transportation systems. Based on the purpose of vehicle image analysis, many algorithms and deep-learning-based models have been proposed either for vehicle classification [1]–[4], vehicle detection and tracking [5]–[7], vehicle license plate verification [8] or for vehicle retrieval and re-identification [3], [8]–[17]. Nevertheless, many challenges are being met while dealing with these problems such as partial/heavy occlusion in vehicle detection,

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

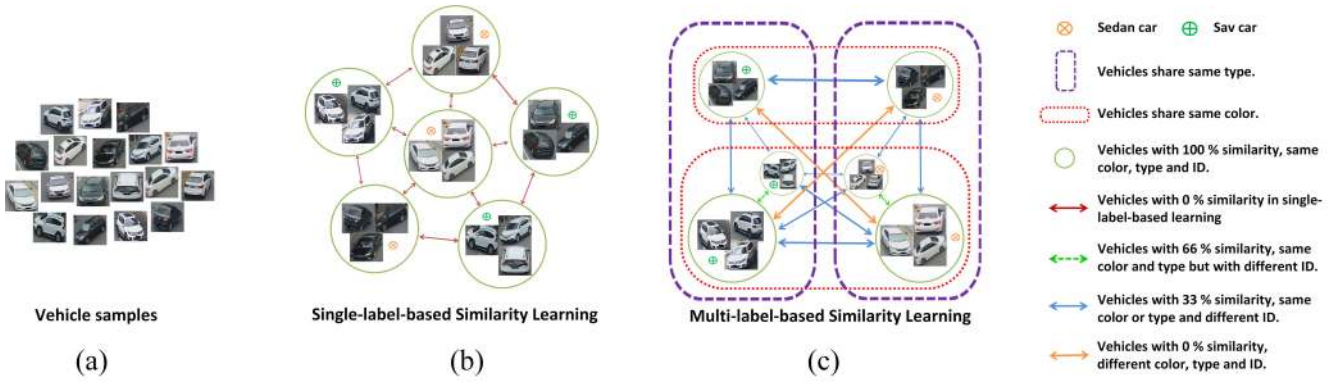

FIGURE 1. Vehicle re-identification in single-label-based similarity and multi-label-based similarity learning. The input raw identities in (a), in (b) the single-label-based similarity in terms of distance learning. The proposed Multi-Label-Based Similarity Learning (MLSL) is illustrated in (c) where the additional vehicle attributes, i.e., color and type, are considered for similarity calculation. As an example, the proposed framework pushes the vehicle images with different IDs, color and type apart more than those vehicle images which they differ in ID while they are sharing same color/type.

and vehicle viewpoint variation in vehicle re-identification. These challenges indicate that there is still broad room for the improvement on vehicle image analysis methods. What makes re-identification problem more challenging in deep learning is that the algorithms are required to achieve accurate re-identification performance for unseen identities in training phase. Unlike classification and detection models which are trained and tested on a defined number of categories, in the re-identification problem the models are trained on a defined number of categories but these models are still required to re-identify undefined number of identities that are unseen on training phase.

Several vehicle datasets have been built such as [1], [18] for vehicle classification, [19] for event-based classification. [20], [21] for vehicle detection. Many well annotated datasets [1], [3], [22]–[26] have been built in order to facilitate the training of the deep-learning-based models for vehicle re-identification.

Several methods have been proposed for vehicle re-identification. These methods can be categorized based on the learning type into semi-supervised/supervised learning methods [3], [8], [9], [11], [12], [14], [27] and unsupervised learning methods [15], [16]. The supervised models are trained either with feature learning [14], [28] or metric learning [3], [9], [11], [27]. Some supervised feature-learning-based models are trained with only the vehicle ID label such as in [17], [27], whereas some models [3], [8], [11], [14] tend to use different vehicle attributes including color and vehicle type or vehicle view-point.

Most of the state-of-the-art methods use metric (Similarity) learning scheme either as the cornerstone of their models or as the most important part. This learning scheme pushes the neural network to generate more discriminating features. However, the performance of the most recent models is still unsatisfactory either in terms of speed or in the re-identification accuracy. That motivated us to design a new model which uses a new metric learning strategy, we call it Multi-Label-Based Similarity Learning (MLSL), to boost the vehicle re-identification performance. Furthermore, we employ a low-cost base CNN feature extractor in terms of number of parameters and computational complexity, which in turn facilitates using the model for the real-time processing in real-world applications. Our contribution of this work can be summarized as follows:

1) Introduce a multi-label-based similarity learning for vehicle re-identification that jointly learns three different similarities of the vehicle pairs with the attributes: vehicle ID, color, and type.
2) Design an efficient model that jointly learns features and similarities, leading to outperforming multiple recent state-of-the-art models.
3) Extensive experiments have been conducted to validate the proposed model's parts, as well as evaluate the proposed model against most recent methods.

Unlike the literature methods, where the assigned label of similarity for each pair of vehicle images should be either 0 or 1 based on only the vehicle ID, our proposed model is inspired by human visual attention mechanism, where it is designed to minimize the distance of the vehicles with same identity to 0, whereas the distance between dissimilar vehicles is maximized and contributed by each unshared attribute, i.e., ID, color, and type. The overall idea of the multi-label-based similarity learning scheme is illustrated in Fig. 1. We transferred this idea into an efficient model depicted in detail in Fig. 2. To validate our learning strategy, our model is evaluated on three well-known datasets VeRi-776 [22], VehicleID [3] and VERI-Wild [24]. The results prove the effectiveness of our model on these datasets.

The rest of this paper is structured as follows: In Section II we discuss the related work. In Section III we describe our model in detail. The utilized datasets and the proposed model training are discussed in Section IV. In the experimental results Section V we analyze the impact of each part of the proposed model and evaluate the model performance against state-of-the-art models.

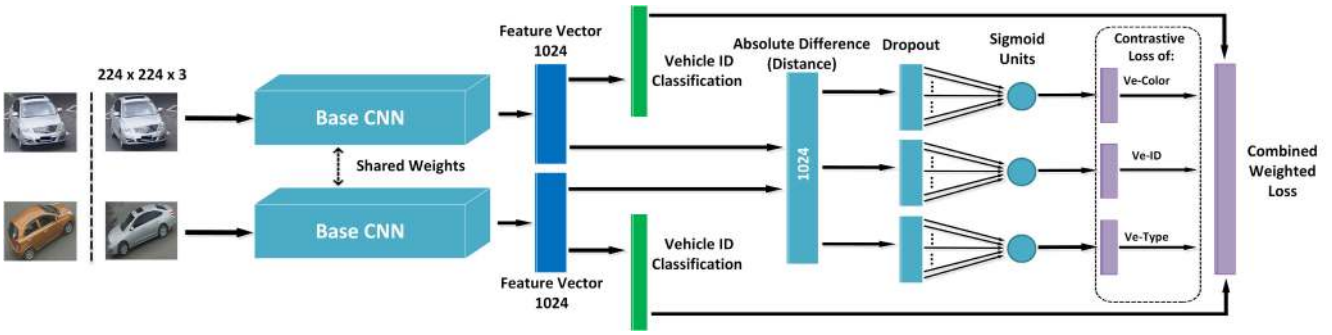

FIGURE 2. The proposed vehicle re-identification model. We employ the categorization, i.e., Softmax classifiers in green color, with the proposed multi-label-based similarity learning to derive high discriminating features (feature vectors in blue color).

II. RELATED WORK

Since a notable advancement is accomplished in intelligent public security and public transportation systems, the vehicle re-identification becomes a highly demanded task. Several vehicle re-identification methods have been developed with deep learning techniques. Generally, these methods can be grouped based on the learning approach into two main groups: feature-based learning [14], [17], [28] and similarity-based learning [3], [9], [11], [27].

Liu *et al.* [8] fused the hand-crafted low-level features with the high-level semantic features to represent each vehicle. A Siamese network is employed for plate verification. They also employed the temporal data that generated from different surveillance cameras [22], with assumption that the vehicle images, with same ID, have a small temporal distance, whereas the images of vehicles with different IDs have large distance.

In another way, Shen *et al.* [9] took advantage of the data that describe the source surveillance cameras to generate a set of visual-spatio-temporal path information, seeking an improvement in vehicle re-identification. Then they employed Long Short-Term Memory units (LSTM) to process the candidate path along with Siamese network to estimate the vehicle images similarity score.

Zhu *et al.* [27] train, Siamese network with classification and similarity learning jointly on vehicle ID labels. Moreover, they utilize a hybrid similarity function to calculate the similarities which combine absolute difference and multiplication in elementwise mode. They claim that this hybrid similarity function boosts the re-identification performance.

Another work in [3] employed metric learning but with the triplet loss function. Each input contains two vehicle image sets, one set contains vehicle images of same ID and the second set contains several vehicle images with different IDs. The triplet loss function pulls the vehicle images of same vehicle ID together and pushes the vehicle of different IDs apart.

Zhou and Shao [11] employed attention mechanism in order to pay attention to the shared regions of all defined vehicle viewpoints, (i.e., front, rear, side, front-side, rear-side). Moreover, they utilized a generative adversarial network (GAN) to transform each single-view feature representation into multi-view representation.

Zhu *et al.* [28] designed a quadruple CNN-based model that extracts visual features of the vehicle images using different CNNs which share the same structure. Each base CNN learns separately, followed by one of the four directional average pooling (i.e., horizontal, vertical, diagonal, or anti-diagonal), and guided by different Softmax classifiers.

LSTM units were utilized in [29] for the purpose of learning the multi-view vehicle representation. A CNN is used as a feature extractor followed by LSTM to derive the multi-view representations. For the same purpose, LSTM layer is employed in [17]. The authors proposed a variational feature learning (VFL) by employing KL divergence on the extracted CNN features to derive the Gaussian distribution by two fully connected layers. They claimed that re-identification performance on transformed features to Gaussian distribution outperforms the re-identification performance on the raw CNN features.

Although, multi-label learning is commonly employed in different classification problems [30], [31]. Rossi *et al.* [32] used a similarity-based learning for multi-label classification. However, our approach in this study resorts to employing multi-labels for similarity distance learning.

Similar to our purpose in some aspects, Wang et al. proposed in [33] a multi-similarity loss for metric learning, which uses two similarities: self and relative similarities. This loss is calculated by running sampling and weighting in an iterative way.

Despite all improvements in object re-identification, vehicle re-identification still does not receive an equivalent attention that has been paid to person re-identification. Thus, there is still room for improvement, particularly in the efficiency of the feature extraction for vehicle re-identification purpose.

III. THE PROPOSED MODEL

The vehicle images of the same vehicle identity are pulled together while the vehicle images with different IDs are pushed apart. Several vehicle identities share the same color or/and type/model, so the limit of pushing these identities

apart considers these two attributes. Where if two vehicle images belong to different IDs but they share color or type/model or even both, then the similarity should be greater than the similarity of vehicle images with different colors and different types/models.

Our model is composed of two main parts that help to derive efficient vehicle features. The output feature representation is learned with two types of supervision, multi-label-based similarity learning and vehicle categorization learning. For vehicle categorization learning we employ Softmax classifier to categorize the vehicles based on their IDs. This helps the model to derive more efficient features for each vehicle identity from different viewpoints. The multi-label-based similarity learning is the second part of our supervised model, which learns the distance between vehicles based on their IDs, color, and type. The model jointly learns both features and multi-label similarity. In this section we explain the modules of our model in detail.

A. BASE NETWORK

The utilized baseline CNN differs from model to another in the literature methods. For instance, ResNet feature extractor is employed in [23], VGG-CNN-M in [3], [24], and [34], Inception-based feature extractor in [1], [13], and [22], MobileNet v1 in [17] and [35], DenseNet121 in [23], whereas in some models, new base networks have been designed such as in [11] and [27].

In this work, we employ MobileNet v1 [36] as base-CNN that pre-trained on ImageNet [37]. We have chosen this neural network due to two main reasons: First, MobileNet shows competitive performance against large and complex neural networks, because it is small in terms of the number of parameters and computational complexity. Utilization of Depthwise Separable Convolution is the main reason behind that sharp reduction of parameters and computational cost, which in turn makes it applicable for real-time processing and feasible in real-world applications. Second, the huge reduction in parameters prevents MobileNet from overfitting and makes it appropriate for different customized classification, detection, and re-identification problems.

By eliminating the classification part from MobileNet, the resulting CNN-based feature extractor consists of one convolutional layer followed by 13 depthwise separable convolutional layers. With five stride steps of (2, 2) through different layers, it progressively downsizes the spatial dimension of the input image I_i of (W, H, D), here W, H and D denoting width, height, and depth respectively. The last output feature map has a spatial dimension of $(\hat{W}, \hat{H}, \hat{D})$. Finally, a global average pooling is applied to end up with an output feature vector $\vec{X}_i$ of size $\hat{D}$ for the input image I_i. The output feature vector of the base feature extractor is shown in blue color in Fig. 2.

B. VEHICLE FEATURE LEARNING

For features learning, we employ Softmax classifier on top of the base network, described in III-A, to learn vehicle identity representation from different viewpoints. Vehicle ID labels are used to supervise the categorization training. Figure 2 shows the Softmax classifiers in green color. For the predicted output vector $\vec{O}$ from the input image I with target label T which represents the vehicle ID, we calculate the classification loss $\mathcal{L}^{cls}$ as in (1).

$$L^{cls}(\vec{O}, T) = -\sum_{j=1}^{C} T_j(log\vec{O}_j),$$
$$\vec{O}_j = \frac{e^{\vec{O}_j}}{\sum_i^C e^{\vec{O}_i}} \tag{1}$$

where $\vec{O}_j$ is the Softmax activation from the baseline neural network. We use one-hot labels that make the loss function simply formulated as (2).

$$L^{cls}(\vec{O}, T) = -log\left(\frac{e^{\vec{O}_t}}{\sum_i^C e^{\vec{O}_i}}\right) \tag{2}$$

where $\vec{O}_t$ is the hot unit in the target vehicle ID label T.

C. VEHICLE MULTI-LABEL-BASED SIMILARITY LEARNING

1) SIAMESE NETWORKS

In 1993, the first Siamese network was used for signature verification [38]. Later on, this type of network learning is employed for image similarity matching in many re-identification and verification applications, e.g., face verification [39], [40], person re-identification [41], vehicle re-identification [9], [27], and image recognition [42]. According to the similarity learning principle, the output vehicle representations are similar for all images of the same vehicle ID, regardless of their viewpoints.

Given a set of vehicle image pairs $\mathcal{P} = \{P_1, ..., P_z\}$. The Siamese base network receives input image pair $P_i = (I_i^a, I_i^b)$ resulting in a corresponding feature representation pair $(\vec{X}_i^a, \vec{X}_i^b)$. The similarity distance between the output feature vectors is then calculated, which is explained in Sect. III-C.2.

2) ABSOLUTE SIMILARITY DISTANCE

There are different types of similarity distance measures, e.g., Euclidean distance, Manhattan distance, Minkowski distance, and Cosine similarity. For our model, we simply extract the absolute difference between the input vector pair $(\vec{X}_i^a, \vec{X}_i^b)$ in elementwise mode. The process of mapping the obtained distance vector $\vec{D}$ into 0 or 1 is achieved by sigmoidal units described in next Section III-C3. The distance calculating is formulated as in (3).

$$\vec{D}(\vec{X}^a, \vec{X}^b)_i = \|\vec{X}_i^a - \vec{X}_i^b\| \tag{3}$$

3) SIGMOIDAL MAPPING UNITS

A sigmoidal unit σ is added on top of the similarity distance layer that maps the similarity distance-vector $\vec{D}$ into 0 or 1 as a binary classification with logistic prediction. Figure 2 illustrates these layers, where we add three sigmoidal units each

TABLE 1. Statistics of the three well-known large-scale datasets for vehicle re-identification.

Dataset	Entire Dataset		Training		Testing		Testing Subsets	# Cameras
	Images	Identities	Images	Identities	Images	Identities		
Veri-776 [22]	49,357	776	37,778	576	11,579	200	1	18
VehicleID [3]	221,763	26,267	113,346	13,164	108,221	13,103	6	12
VERI-Wild [24]	406,314	40,671	277,797	30,671	128,517	10,000	3	174

unit is assigned to one of the three used labeled attributes: vehicle ID, color, and type. This can be formulated as in (4).

$$d_i^m = \sigma \vec{D}(\vec{X}^a, \vec{X}^b)_i \tag{4}$$

where $m \in M, M = \{id, clr, typ\}$.

4) MULTI-LABEL SIMILARITY LOSS

We have adopted the Contrastive Loss introduced in [43] on top of the similarity mapping layer. We apply this loss jointly for three attributes, (i.e., vehicle ID, color, and type/model). Let Y be a binary label assigned to the pair $P_i = (I_i^a, I_i^b)$ and the corresponding feature vector pair $(\vec{X}_i^a, \vec{X}_i^b)$. For each pair of feature vector $(\vec{X}_i^a, \vec{X}_i^b)$, a vector of the absolute element-wise difference $\vec{D}_i$ is computed by (3), which in turn mapped by (4) into d_i^m. For each d_i^m a corresponding label Y_i^m. If the feature vectors $\vec{X}_i^a$ and $\vec{X}_i^b$ are similar in terms of the attribute m, then the $Y_i^m = 0$. Contrarily, if they are dissimilar, then $Y_i^m = 1$. The loss can be calculated for the i-th input pair by (5).

$$L^{ver}(d_i, (Y, \vec{X}^a, \vec{X}^b)_i) = \sum_{m \in M} \lambda^m \Big((1 - Y_i^m) L_S(d_i^m) + (Y_i^m) L_D(d_i^m)) \Big) \tag{5}$$

where $(Y, \vec{X}^a, \vec{X}^b)_i$ is the i-th labeled vehicle pair. Both partial loss functions, L_S of a similar vehicle pair and dissimilar vehicles pair L_D, are constructed to minimize the L^{ver} to obtain small values of d^m for similar vehicle pairs and large values of d^m for dissimilar vehicle pairs. We can formulate the final verification loss function as in (6).

$$L^{ver}(d_i, (Y, \vec{X}^a, \vec{X}^b)_i) = \sum_{m \in M} \lambda^m \Big((1 - Y_i^m) \frac{1}{2} (d_i^m)^2 + (Y_i^m) \frac{1}{2} max(0, margin - d_i^m)^2 \Big) \tag{6}$$

where $margin > 0$ and λ^m is the contribution values of each corresponding loss term of the vehicle attribute in M. In our experiments, we set it by default to $\lambda^m = 1$ for each m.

Overall, the totaled loss $\mathcal{L}$ for each vehicle input pair P_i, which combines the categorization loss L^{cls_a}, L^{cls_b} and the verification loss L^{ver}

$$\mathcal{L}(P_i) = \alpha(L^{cls_a} + L^{cls_b}) + L^{ver} \tag{7}$$

where $\alpha \geq 0$ is constant to weight the classification loss contributions in the total loss.

Based on a set of extensive experiments, we found the best practice for all the modules of our model. In the following sections, we analyze our model features as well as compare its performance against most recent state-of-the-art deep-learning-based models.

IV. TRAINING AND SETTINGS

A. DATASETS

We have used VeRi-776 [22] dataset for evaluating our model's modules. Whereas the evaluation against the state-of-the-art models is conducted on three well-known vehicle datasets, VeRi-776 [22], VehcielID [3], and VERI-Wild [24]. Table 1 lists the main characteristics of these datasets.

1) VERI-776

VeRi-776 [22] is a large-scale image dataset for vehicle re-identification. It was captured by 20 cameras in real-world urban surveillance environments with unconstrained surveillance scenarios. Each image set of each vehicle ID is captured by 2-18 surveillance cameras in different viewpoints, illuminations, occlusions and resolutions. Each vehicle is labeled with three attributes, i.e., ID, type, color, and camera ID. The VeRi-776 dataset consists of a training set containing 37, 778 images of 576 vehicles and testing set with 11, 579 images of 200 vehicles. The proposed evaluation protocol employs mean average precision (mAP) and Top-K Cumulative Match Characteristic (CMC) scores for Top-1, and Top-5 for evaluating the performance of re-identification.

2) VEHICLEID

VehicleID [3] is another large-scale dataset for vehicle re-identification that was captured in a small city in China. Several real-world surveillance cameras are used to collect the vehicle images with two viewpoints, i.e., front and rear. VehicleID dataset consists of 221, 763 images of 26, 267 vehicles. VehicleID dataset consists of the training set with 113, 346 images of 13, 164 different vehicles and testing set with 108, 221 images. Each vehicle image is labeled with vehicle-id, color, and vehicle model. For the performance evaluation of the re-identification, the testing set is organized in six subsets, i.e., Test-800, Test-1600, Test-2400, Test-3200, Test-6000, and Test-13164. The first three sets are the most common to be used in several recent related studies. The Top-K CMC scores for Top-1 and Top-5 are used for evaluating the performance of re-identification on this dataset.

3) VERI-WILD

VERI-Wild [24] is the largest image dataset for vehicle re-identification and tracking. It was captured by a real surveillance camera network of 174 cameras that makes this dataset more diverse as well as more challenging. VERI-Wild dataset covers different viewpoints, resolutions, illuminations, weather conditions and occlusions. In contrast to the Veri-776 and VehicleID, VERI-Wild dataset contains images captured at night. It contains in total 416, 314 images of 40, 671 identities divided into a training set with 277, 797 images of 30, 671 and testing set of 10, 000 vehicle identities with 128, 517 images. The testing set is further organized in three different-sized subsets: a small testing subset of 3, 000 identities with 41, 816 images, medium-size testing subset with 5, 000 identities and 69, 389 images, the large testing subset with 138, 517 images of 10, 000 identities. Similarly to VeRi-776 dataset, the mean Average Precision (mAP) and Top-K CMC scores for Top-1, and Top-5 for evaluating the performance of re-identification.

B. TRAINING

In this section we explain in detail the proposed model training on the mentioned vehicle datasets. The proposed model is trained in two stages: Firstly, we train the base network with vehicle IDs as a supervised classification problem. Then the obtained trained model from this stage is used to fine-tune the proposed model (in Fig. 2) to jointly learn the feature representations and the similarity distances. Training the base network and the entire proposed model share some training hyperparameters. For both stages, the Adam optimizer [44] is used with the initial learning rate of 0.001 and a momentum of 0.9. In the first stage, the learning rate is multiplied by 0.1 every 100 epochs, whereas in the second stage the learning rate is multiplied by 0.1 every 50 epochs. In order to minimize the effects of model overfitting, we employ input image augmentation and dropout regularization. The proposed model receives images in dimension of 224×224 pixels, that randomly augmented with cropping, brightening, and rotation.

1) BASE NETWORK TRAINING

Three instances of the base network, described in Sect.III-A, are trained. Each base network instance is trained on one of the three used datasets VeRi-776, VehicleID, or VERI-Wild. We add a Softmax classifier on top of the base network as a convolutional layer with a number of 1×1 kernels corresponds to the number of vehicles, 576 in VeRi-776, 13, 164 in VehicleID, and 30, 671 in VERI-Wild. We add a dropout layer prior to the classification layer with the rate of 0.001.

2) JOINT TRAINING OF FEATURE AND MULTI-SIMILARITY

In the second training stage, we fine-tuned the proposed model 2 with the base network parameters trained on the first stage. We apply dropout of rate 0.5 prior to each sigmoidal unit. The model in this stage is trained jointly by totaled 5 weighted loss functions: two vehicle classification loss functions of both Siamese branches a and b, and three contrastive loss functions each for one of the vehicle attributes (ID, color, type/model), all are weighted with $(0.1, 0.1, 1, 1, 1)$ respectively. The input image batches are generated in online mode with an input size of 224×224 pixels (described in detail in Sect. IV-C). Similar to the training of the base network, same hyperparameters are used in this stage.

C. ONLINE BATCH GENERATOR

In our work, we built an online batch sampler, that prepares the image pairs, aiming to ensure that each single vehicle image has the same probability to pair any image in the training set. Batch generating procedure picks up, randomly, three images. Two images with the same vehicle identity, whereas the third image has another vehicle identity, resulting in two pairs. For each vehicle image pair $P_i = (I_i^a, I_i^b)$, two types of labels are generated. The first type is prepared for categorization training which represents the vehicle IDs for T_i^a and T_i^b. On the other hand, three pair-based binary labels Y_i^{id}, Y_i^{clr}, Y_i^{typ} are generated representing the similarity labels of the vehicle ID, color, and the vehicle type respectively. Batch sampler iterates these steps until the number of generated pairs match the defined batch size, which in turn needs to be an event number.

Most Siamese models in literature [9], [27], [39]–[42] label the similarity of an image pair of the same vehicle with 1, whereas 0 is the pair label with different vehicle IDs. However, in this paper, we consider the labels Y_i^{id}, Y_i^{clr}, and Y_i^{typ} as the difference distance labels. That leads to assign the label 0 to the pair with images of a similar attribute, and 1 to the pair of images with dissimilar attribute.

V. EXPERIMENTAL RESULTS

In this section, we study the influence of each part of our model and its contribution on the entire performance on VeRi-776 [22] dataset. We begin by studying the influence of using the proposed multi-label similarity learning against the regular single-label learning. Then, we analyze the performance of the model that is separately trained in feature-based, similarity-based and joint learning approaches. Next, we study the impact of the absolute distance comparing to other distance metrics, the impact of the sigmoidal units, and the impact of the dropout regularization. Finally, we evaluate our model performance against several state-of-the-art vehicle re-identification models on VeRi [22], VehicleID [3], and VERI-Wild [24] datasets.

A. MODULES ANALYSIS OF THE PROPOSED MODEL

When we train the proposed model with randomly online batch generating, we use the number of iterations to stop the training instead of the number of epochs. For each instance of the proposed model in the following subsections, we use a mini-batch of 24 for a total number of 48, 000 iterations. The learning rate is initialized with 0.001, decreasing it after

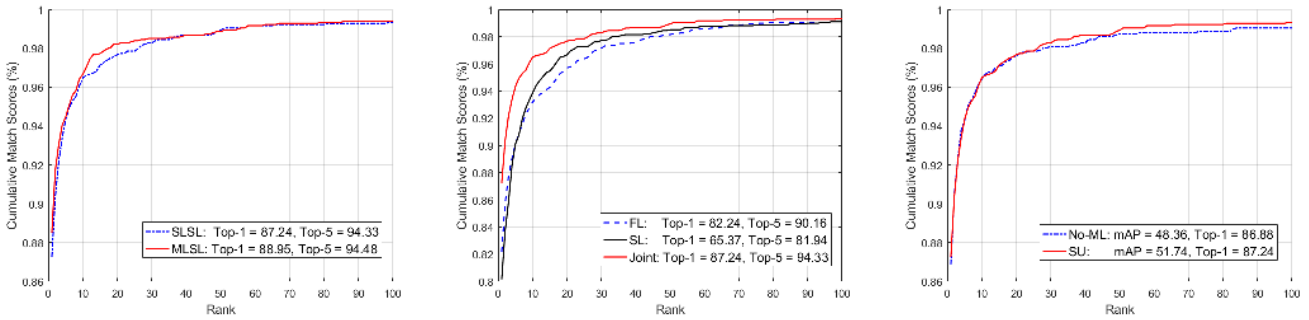

FIGURE 3. The performance comparison of different instances of learning, Plot (a) shows the performance comparison of the Joint Single-Label-Based Similarity Learning and Classification Learning (SLSL) against the Joint Multi-Label-Based Similarity Learning and Classification Learning (MLSL). Plot (b) shows the performance comparison of the Feature Learning (FL), the Similarity Learning (SL), and the joint similarity and classification learning (SLSL). In (c) the performance of the proposed model with sigmoidal unit (SU) and without mapping layer (No-ML) is shown on VeRi-776 dataset.

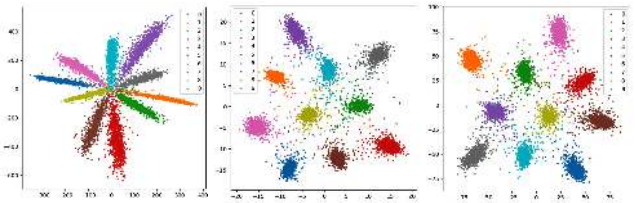

FIGURE 4. Feature distributions of feature-based learning (a), similarity-based learning (b), and joint learning (c) of the MNIST's testing set for digit classification.

TABLE 2. Performance of the single-label against multi-label similarity learning, on VeRi-776 dataset.

Similarity Learning	mAP %	Top-1 %	Top-5 %
Single-label-based Similarity SLSL	51.74	87.24	94.33
Multi-label-based Similarity MLSL	**55.49**	**88.95**	**94.48**

24, 000 iterations by multiplying it with 0.1, then decreasing it for two more times each after 12, 000 iterations.

1) SINGLE-LABEL AGAINST MULTI-LABEL SIMILARITY LEARNING

In this part, the performance of the regular single-label learning is compared with the proposed multi-label similarity learning. The utilized base network is MobileNet V1 that is pre-trained on the VeRi-776 dataset (see Section IV-B1). Our default proposed model, illustrated in Fig. 2, is trained with the vehicle ID attribute for categorization learning, jointly with the similarity learning that uses vehicle color and vehicle type along with the vehicle ID. Moreover, we trained the same model but with only the vehicle ID labels for both vehicle classification and similarity learning as single-label learning. We have summarized the mAP, Top-1, and Top-5 in Table 2, which proves the superior performance of the multi-metric learning (MLSL) against Single-Label-Based Similarity Learning (SLSL). Fig. (3,a) shows the CMC plot of both models.

2) FEATURE LEARNING, SIMILARITY LEARNING, AND JOINT LEARNING

In this part, we compare the performance of three different learning approaches: feature-based learning, similarity-based learning, and joint learning. For the approaches, we use MobileNet V1 [36] pre-trained on VeRi-776 dataset as a base network. In feature-based learning, we follow the procedure of base network training as a categorization classifier explained in Sect. III-B. To test the performance of the similarity-based learning, we train one instance of the proposed model after eliminating the categorization component, (i.e., Softmax classifiers), as well as eliminating the vehicle color and vehicle type similarity learning components in order to use a single-label-based similarity learning which only uses the binary labels of the vehicle ID attribute. Finally, with the vehicle ID labels, we jointly trained an instance of our model which combines the categorization and the similarity learning components.

The result on VeRi-776, summarized in Table 3, demonstrates the effectiveness of the joint-based learning, which outperforms both feature-based learning and similarity-based leaning by more than 4% and 21% in terms of Top-1 respectively. Fig. (3,b) shows the CMC plot of the feature-based, single-metric-based, and joint-based learning approaches.

Since the number of vehicle IDs in the testing set of Veri-776 dataset is 200, it is hard to visualize their feature distribution. However, we trained the three mentioned model instances on MNIST [45] which has only 10 classes.

TABLE 3. Performance of feature learning, metric learning, and Joint learning on VeRi-776 dataset.

Learning Scheme	mAP %	Top-1 %	Top-5 %
Feature-based learning	41.24	82.24	90.16
Similarity-based learning	38.43	65.37	81.94
Joint learning	**51.74**	**87.24**	**94.33**

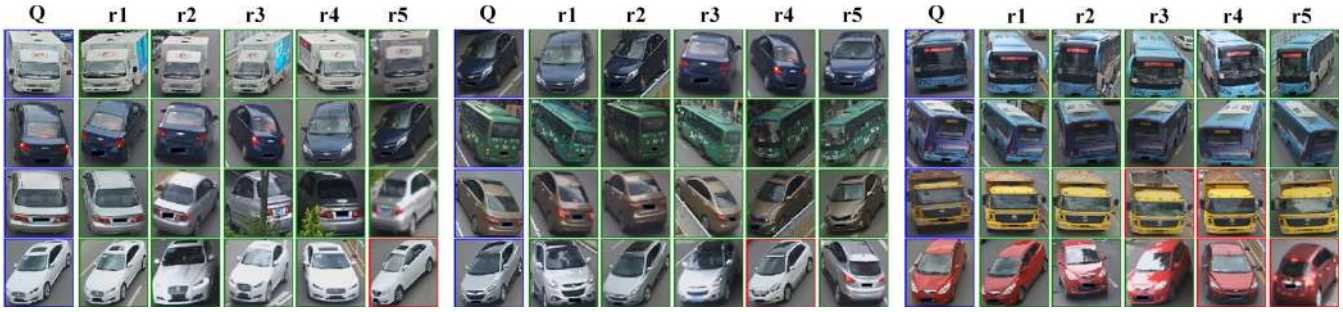

FIGURE 5. Samples of the retrieval performance of the proposed Multi-Label-Based Similarity Learning (MLSL) on testing set of VeRi-776 dataset. The true positive retrieved vehicle images are bounded with boxes in green color while the mis-retrieved images are bounded in red color.

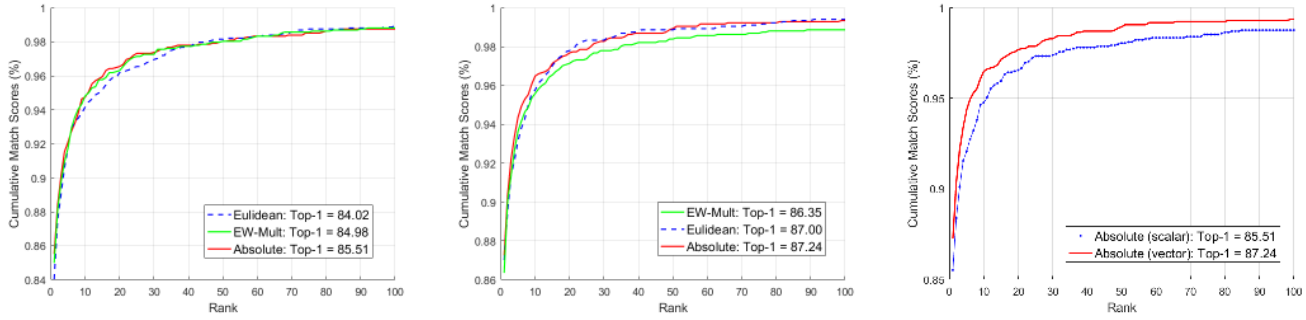

FIGURE 6. Impact of distance type between two vehicles' feature vectors on VeRi-776 dataset. Plot (a) shows the performance of the proposed model with scalar-based distance. Plot (b) shows the performance of the same model but when it uses the proposed distance calculation (vector-based distance). Plot (c) shows the default utilized distance in our model (i.e., absolute distance), with both distance calculation schemes.

TABLE 4. Comparison of the supervised feature learning against supervised similarity learning and jointly learning for digit retrieval on MNIST [45] dataset.

Retrieval Acc	Feature-based	Similarity-based	Joint Learning
Top-1	92.34	97.94	**98.79**
Top-2	99.13	99.40	**99.63**

We replaced the base network with a tiny network of 4 convolutional layers with 128 filters of size 3×3 except the last convolutional layer that is prior to classification layer which is created with 2 filters. The Top-1 and Top-2 retrieval accuracies of the three models on MNIST are summarized in Table 4. Note that the feature vector used for matching the query and gallery images is the convolutional layer with a size of 2. In order to visualize the feature distribution of the 10 digits as shown in Fig. 4, we use the mentioned layer with size 2, which allows us to visualize digit features (i.e., one filter represents x dimension and the second for y dimension on a graph).

3) SIMILARITY MAPPING

The output vector of the similarity distance, obtained from the distance metric function, is mapped into one of the binary similarity labels (0 or 1). In this part, we evaluate the impact of using the similarity mapping layer. On one hand, we have trained the proposed model without the similarity mapping layer. In order to do so, it is required to extract the similarity distance as a scalar (i.e., as in regular distance calculation) rather than extracting a similarity distance as a vector. On the other hand, we evaluate the similarity distance mapping layer that contains a single Sigmoidal Unit (SU). As it is shown in Fig. (3,c), the model instance that employs the sigmoidal unit for similarity mapping outperforms the model instance that does not employ a mapping layer. The results, in terms of Top-1, Top-5 and mAP on VeRi-776 dataset, are listed in Table 5.

TABLE 5. Comparison of the proposed model performance with and without similarity mapping unit, on VeRi-776 dataset.

Similarity Mapping Layer	mAP %	Top 1 %	Top 5 %
Without Mapping layer	48.36	86.88	94.27
Single Sigmoidal unit	**51.74**	**87.24**	**94.33**

4) IMPACT OF DISTANCE CALCULATION SCHEME

Similarity distance calculation is achieved in two different schemes. As is illustrated in Fig. 7, in (a) the distance calculation results in a vector of the elementwise distance between the two input vectors $\vec{X}^a$ and $\vec{X}^b$, whereas in (b) the output is a single value. Using these two schemes, we have evaluated the impact of different similarity distances including the absolute distance, the Euclidean distance, and the Element-Wise Multiplication. By skimming the Table 6 and Fig. 6, we can realize that the vector-based similarity calculation Fig. (6,b) performs better than the scalar-based similarity calculation Fig. (6,a). In Fig. (6,c), we plot the CMC of both vector-based

TABLE 6. Comparison of the proposed model with different distance types between two vehicle feature vectors on VeRi-776 dataset.

Distance Metric	Scalar-Based Distance			Vector-Based Distance		
	mAP %	Top-1 %	Top-5 %	mAP %	Top-1 %	Top-5 %
Elementwise Multiplication	40.40	61.70	35.40	48.50	86.35	93.62
Euclidean Distance	**42.55**	84.02	91.71	**51.99**	87.00	93.14
Absolute Distance	42.42	**85.51**	**92.07**	51.74	**87.24**	**94.33**

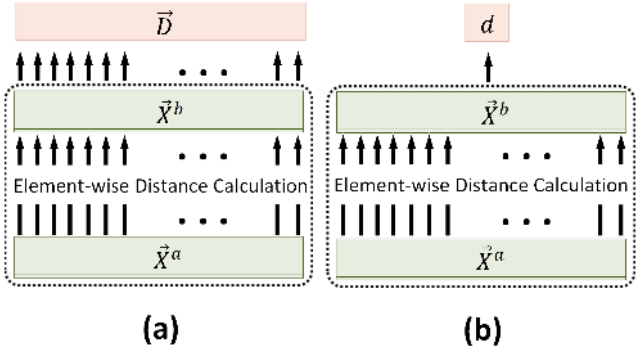

FIGURE 7. Similarity distance calculation, in (a) the output distance is a vector of the elementwise distance calculation that we use in our model, in (b) the regular scalar-based distance calculation.

TABLE 7. Performance comparison of the proposed model with different dropout rates on VeRi-776 dataset.

Dropout Rate	mAP %	Top-1 %	Top-5 %
0.2	37.58	78.42	87.84
0.5	**51.74**	**87.24**	**94.33**
0.7	48.36	86.88	94.27
0.9	40.15	77.94	88.79

and scalar-based absolute distance calculation. It is notably observed that the similarity distance extracted as a vector helps the mapping layer to guide the model for generating better discriminative features. As reported in Table 6, among the three tested similarity calculation functions, the absolute similarity distance function outperforms the Euclidean and Elementwise Multiplication functions in terms of Top-1 and Top-5.

5) IMPACT OF DROPOUT

Although MobileNet V1 [36] is less prone to the problem of overfitting comparing to those neural networks which use regular convolution, the model still overfits the datasets. This is because most objects used to re-identify are semantically similar. In this part, we evaluate the impact of using different rates for the dropout layer prior to the similarity mapping layer. Fig. 8, shows the mAP, Top-1, and Top-5 against the dropout rate. Table 7 lists the reported results of Top-1, Top-5, and mAP on VeRi-776 dataset. Apparently, a positive effect is obtained by applying the dropout for vehicle re-identification, where the trained model with a dropout of 0.5 gains the best performance. This is because the ability of the network to learn a discriminating representation for %50 of the output feature vector is much easier than to learn %100 of the output vector in same batch iterations. However, the process of training with large dropout rates requires more time.

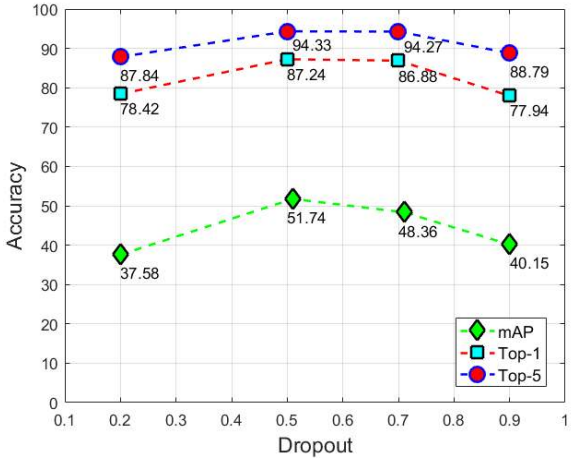

FIGURE 8. Performance of the proposed model with different dropout rates on VeRi-776 dataset.

B. PERFORMANCE OF THE PROPOSED MODEL AGAINST RECENT RELATED WORKS

Many methods have been proposed for vehicle re-identification. We have compared our model with most recent methods. The reported results show our model superiority over multiple state-of-the-art models on VeRi-776 and VERI-Wild datasets while obtaining a competitive performance on VehicleID dataset.

1) PERFORMANCE EVALUATION ON VERI-776 DATASET

To validate the effectiveness of our model, we have compared its performance with many recent related works, including: Siamese-Visual [9], GoogLeNet [1], FACT [22], Chain MRF model [9], SCCN-Ft [29], CLBL-8-Ft [29], XVGAN [46], OIF [13], Siamese-CNN [9], VAMI [11], NuFACT [8], PROVID [8], VR-PROUD [16], Path-LSTM [9], JFSDL [27], D-DLF [28], and Mob.VFL-LSTM [17]. Table 8 summarizes their performance in terms of mAP, Top-1, and Top-5. Obviously, our model provides the best performance among all methods.

In many state-of-the-art methods, a combination of multiple networks is used to boost the performance. Some examples of these combinations are listed in the lower right part of the Table 8 including: SCCN-Ft + CLBL-8-Ft [29], FACT + Plate-SNN + STR [22], OIF + ST [13], NuFACT + Plate-REC [8], NuFACT + Plate-SNN [8], Siamese-CNN + Path-LSTM [9]. The authors in Mob.VFL-LSTM [17] combined Gaussian modeling with LSTM, a remarkable boost can be observed on the performance. However, our model outperforms their performance. Retrieved samples by the

TABLE 8. The performance comparison of different methods on VeRi-776 dataset.

Model	mAP %	Top-1	Top-5	Model	mAP %	Top-1	Top-5
Siamese-Visual [9]	29.48	41.12	60.31	VR-PROUD [16]	40.05	83.19	91.12
GoogLeNet [1]	17.89	52.32	72.17	Path-LSTM [9]	54.49	82.89	89.81
FACT [22]	18.49	50.95	73.48	JFSDL [27]	53.53	82.90	91.60
Chain MRF model [9]	44.31	54.41	61.50	D-DLF [28]	53.26	84.92	93.03
SCCN-Ft [29]	20.13	55.46	70.02	Mob.VFL-LSTM [17]	58.08	87.18	94.63
CLBL-8-Ft [29]	23.25	57.95	77.16	SCCN-Ft+CLBL-8-Ft [29]	25.12	60.83	78.55
XVGAN [46]	24.65	60.20	77.03	FACT+Plate-SNN+ STR [22]	27.77	61.44	78.78
OIF [13]	48.00	65.92	87.66	OIF + ST [13]	51.42	68.30	89.70
Siamese-CNN [9]	54.21	79.32	88.92	NuFACT + Plate-REC [8]	48.55	76.88	91.42
VAMI [11]	50.13	77.03	90.82	NuFACT + Plate-SNN [8]	50.87	81.11	92.79
NuFACT [8]	48.47	76.76	91.42	Siamese-CNN+ Path-LSTM [9]	58.27	83.49	90.04
PROVID [8]	53.42	81.56	95.11	MLSL (Ours)	**61.13**	**90.04**	**96.00**

TABLE 9. The performance comparison of different methods on VehicleID dataset.

Model	Test size 800		Test size 1600		Test size 2400	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
VGG+Triplet Loss [3]	40.40	61.70	35.40	54.60	31.90	50.30
VGG+CLL [3]	43.60	64.20	37.00	57.11	32.90	53.30
GoogLeNet [1]	47.88	67.18	43.40	63.86	38.27	59.39
FACT [22]	49.53	68.07	44.59	64.57	39.92	60.32
Mixed Diff+CLL [3]	49.00	73.50	42.80	66.80	38.20	61.60
XVGAN [46]	52.87	80.83	49.55	71.39	44.89	66.65
JFSDL [27]	54.80	85.26	48.29	78.79	41.29	70.63
C2F-Rank [47]	61.10	81.70	56.20	76.20	51.40	72.20
VAMI [11]	63.12	83.25	52.87	75.12	47.34	70.29
Mob.VFL [17]	73.37	85.52	**69.52**	81.00	**67.41**	78.48
MLSL (Ours)	**74.21**	**88.38**	69.23	**81.48**	66.55	**78.67**

TABLE 10. The performance comparison of different methods on VERI-Wild dataset.

Model	Small			Medium			Large		
	mAP	Top-1	Top-5	mAP	Top-1	Top-5	mAP	Top-1	Top-5
GoogLeNet [1]	24.27	57.16	75.13	24.15	53.16	71.10	21.53	44.61	63.55
Triplet [48]	15.69	44.67	63.33	13.34	40.34	58.98	9.93	33.46	51.36
Softmax [8]	26.41	53.40	75.03	22.66	46.16	69.88	22.66	37.94	59.89
CCL [3]	22.66	56.96	75.0	19.28	51.92	70.98	14.81	44.60	60.95
HDC [49]	29.14	57.10	78.93	29.14	49.64	72.28	18.30	43.97	64.89
GSTE [34]	31.42	60.46	80.13	26.18	52.12	74.92	19.50	45.36	66.50
FDA-Net [24]	35.11	64.03	82.80	29.80	57.82	78.34	22.78	49.43	70.48
MLSL (Ours)	**46.32**	**86.03**	**95.10**	**42.37**	**83.00**	**93.54**	**36.61**	**77.51**	**91.44**

proposed model on the testing set of the VeRi-776 dataset are shown in Fig. 5.

2) PERFORMANCE EVALUATION ON VEHICLEID DATASET

The second evaluation of our model against several state-of-the-art models is conducted on VehicleID dataset. Following the evaluation protocol proposed in [3], the three testing subsets of size 800, 1600 and 2400 are used to evaluate the Top-1 and Top-5. Table 9 lists the comparison between our model and several latest leading models including: VGG + Triplet Loss [3], VGG + CLL [3], GoogLeNet [1], FACT [22], Mixed Diff + CLL [3], XVGAN [46], JFSDL [27], C2F-Rank [47], VAMI [11], and Mob.VFL [17]. The complex model, VAMI of [11], which employs the attention mechanism and generative adversarial network (GAN), obtains competitive performance in terms of the Top-1 and Top-5 on the Testing-800, while C2F-Rank [47] gains better performance on Testing-1600 and Testing-2400. Moreover, Mob.VFL [17] gains the best performance comparing to other models except our model which exceeds its performance in terms of Top-1 and Top-5 on Testing-800 subset and Top-5 of other testing subsets. For example, on Testing-800, the Top-1 accuracy is increased from 73.37 of Mob.VFL to 74.21.

3) PERFORMANCE EVALUATION ON VERI-WILD DATASET

The evaluation of our model on the third dataset VERI-Wild is conducted against several models including GoogLeNet [1], Triplet [48], Softmax [8], CCL [3], HDC [49], GSTE [34], and FDA-Net [24]. We follow the evaluation protocol proposed in [24]. Three testing identity subsets of 3000, 5000, and 10,000 correspond to 41,816, 69,389, and 138,517 image subsets respectively. Table 10 lists the performance comparison between our model and several models in the related works in terms of Top-1, Top-5, and mAP. In this dataset, we obtain the best performance with impressive results. A superior performance is given by the proposed model which improves the Top-1 accuracy in the small testing

subset by about 22% which is increased from 64.03, obtained by FDA-Net [24], to 86.03 and mAP from 35.11 to 46.32.

VI. CONCLUSION

In this work, we introduced an efficient deep-learning-based model that jointly learns vehicle feature and multi-label-based similarity for vehicle re-identification. The multi-label-based similarity learning (MLSL) employs three-vehicle attributes: ID, color, and type/model. The effectiveness of our model is validated by extensive experiments. Furthermore, the experiments on three well-known vehicle datasets show the superior performance of our model against several recent state-of-the-art methods. Our purpose of designing this model is to derive high discriminating features by learning the intra/inter vehicles-id features. This discrimination is obtained by multi-label-based training. The proposed model is cost-effective to be applied to real-time vehicle re-identification applications.

Finally, it is worthy to employ the proposed model for other re-identification problems particularly for the person re-identification, which we plan to explore in the future.

REFERENCES

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

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

[3] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[4] J. T. Lee and Y. Chung, "Deep learning-based vehicle classification using an ensemble of local expert and global networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 920–925.

[5] S. Ram and J. J. Rodriguez, "Vehicle detection in aerial images using multiscale structure enhancement and symmetry," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3817–3821.

[6] J.-N. Xin, X. Du, and J. Zhang, "Deep learning for robust outdoor vehicle visual tracking," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 613–618.

[7] M. Y. Yang, W. Liao, X. Li, and B. Rosenhahn, "Deep learning for vehicle detection in aerial images," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3079–3083.

[8] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018.

[9] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1900–1909.

[10] Y. Li, Y. Li, H. Yan, and J. Liu, "Deep joint discriminative learning for vehicle re-identification and retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 395–399.

[11] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.

[12] C.-W. Wu, C.-T. Liu, C.-E. Chiang, W.-C. Tu, and S.-Y. Chien, "Vehicle re-identification with the space-time prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 121–1217.

[13] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 379–387.

[14] X. Liu, S. Zhang, Q. Huang, and W. Gao, "Ram: A region-aware deep model for vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[15] P. A. Marín-Reyes, L. Bergamini, J. Lorenzo-Navarro, A. Palazzi, S. Calderara, and R. Cucchiara, "Unsupervised vehicle re-identification using triplet networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 166–1665.

[16] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, "VR-PROUD: Vehicle re-identification using progressive unsupervised deep architecture," *Pattern Recognit.*, vol. 90, pp. 52–65, Jun. 2019.

[17] S. A. S. Alfasly, Y. Hu, T. Liang, X. Jin, Q. Zhao, and B. Liu, "Variational representation learning for vehicle re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3118–3122.

[18] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. 4th Int. IEEE Workshop 3D Represent. Recognit. (3dRR)*, Sydney, NSW, Australia, Dec. 2013, pp. 554–561.

[19] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "HATS: Histograms of averaged time surfaces for robust event-based object classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sydney, NSW, Australia, Jun. 2018, pp. 1731–1740.

[20] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1475–1490, Nov. 2004.

[21] A. Geiger, P. Lenz, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[22] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[23] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8797–8806.

[24] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3235–3243.

[25] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 1475–1490, Jul. 2018.

[26] A. Kanacı, X. Zhu, and S. Gong, "Vehicle re-identification in context," in *Proc. German Conf. Pattern Recognit.*, 2018, pp. 377–390.

[27] J. Zhu, H. Zeng, Y. Du, Z. Lei, L. Zheng, and C. Cai, "Joint feature and similarity deep learning for vehicle re-identification," *IEEE Access*, vol. 6, pp. 43724–43731, 2018.

[28] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Vehicle re-identification using quadruple directional deep learning features," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[29] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3287, Jul. 2018.

[30] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[31] H. Gouk, B. Pfahringer, and M. Cree, "Learning distance metrics for multi-label classification," in *Proc. 8th Asian Conf. Mach. Learn.*, vol. 63, 2016, pp. 318–333.

[32] R. A. Rossi, N. K. Ahmed, H. Eldardiry, and R. Zhou, "Similarity-based multi-label learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1–8, Jul. 2017.

[33] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5022–5030.

[34] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[35] R. Kumar, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle re-identification: An efficient baseline using triplet embedding," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–12.

[36] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[38] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.
[39] S. Chopra, R. Hadsell, and Y. Lecun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.
[40] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1–9.
[41] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Deep hybrid similarity learning for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3183–3193, Nov. 2018.
[42] R. Z. G. Koch and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Workshop*, 2015, pp. 1–30.
[43] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1735–1742.
[44] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[45] 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.
[46] Y. Zhou and L. Shao, "Cross-view gan based vehicle generation for re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–12.
[47] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[48] F. Schroff, K. Dmitry, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[49] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 815–823.

SAGHIR ALFASLY received the B.S. degree in computer information systems from Sana'a University, Yemen, in 2010, and the M.Sc. degree in computer science from Kuvempu University, India, in 2015. He is currently pursuing the Ph.D. degree in information and communication engineering with the School of Electronic and Information Engineering, South China University of Technology, China. His current research interests include surveillance image and video analysis, person/vehicle re-identification, computer vision, and machine learning.

YONGJIAN HU received the Ph.D. degree in communication and information systems from the South China University of Technology, China, in 2002. From 2000 to 2004, he visited the Department of Computer Science, City University of Hong Kong four times, as a Research Assistant, a Senior Research Associate, and a Research Fellow, respectively. From 2005 to 2006, he was a Research Professor with the School of Information and Communication Engineering, SungKyunkwan University, South Korea. From 2006 to 2008, he was a Research Professor with the Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), South Korea. From 2011 to 2013, he was a Marie Curie Fellow with the Department of Computer Science, The University of Warwick, U.K. He is currently a Full Professor with the School of Electronic and Information Engineering, South China University of Technology. He has published more than 70 peer reviewed papers. His current research interests include information hiding, multimedia security, and machine learning. He is a Senior Member of the Chinese Institute of Electronics (CIE) and the China Computer Federation (CCF).

HAOLIANG LI received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China (UESTC), in 2013, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2018. He is currently the Wallenburg-NTU Presidential Postdoctoral Fellow with the Rapid-Rich Object Search Lab, NTU, Singapore. His current research interests include information forensics and security, and transfer learning.

TIANCAI LIANG received the Ph.D. degree in control theory and control engineering from the South China University of Technology, in 2008. From 2008 to 2011, he held a postdoctoral position with the Guangzhou Radio Group. From 2011 to 2015, he was a Technical Manager with GRG Banking Equipment Company Ltd. From 2015 to 2018, he was a Chief Researcher with GRG Banking Equipment Company Ltd. He has published more than 200 patents engineer. His current research interests include pattern recognition, machine learning, and video understanding. He received the Professor-Level Senior Engineers, in 2015.

XIAOFENG JIN received the bachelor's degree in information engineering from the Beijing Institute of Technology, China, in 2007, and the Ph.D. degree in optical engineering from the Chinese Academy of Sciences, China, in 2012. His current research interests include video processing, big data, and machine learning.

BEIBEI LIU received the Ph.D. degree in communication and information system from Sun Yat-sen University, China, in 2009. She was a Researcher with the Korea Advanced Institute of Science and Technology and Newcastle University, U.K. She is currently an Assistant Professor with the School of Electronic and Information Engineering, South China University of Technology. Her current research interests include multimedia information security and machine learning.

QINGLI ZHAO received the B.E. degree in information engineering and the Ph.D. degree in communication and information system from the South China University of Technology, China, in 2004 and 2014, respectively. He is currently an Algorithm Engineer and the Team Leader of intelligent video processing with the GRG Intelligent Security Institute. His current research interests include video analysis and deep learning.