applied sciences

MDPI

Article

Local Feature-Aware Siamese Matching Model for Vehicle Re-Identification

Honglie Wang [1,*], Shouqian Sun [1], Lunan Zhou [2], Lilin Guo [3], Xin Min [1] and Chao Li [1]

[1] College of Computer Science and Technology, Zhejiang University, No. 38 Zheda Road, Hangzhou 310027, China; ssq@zju.edu.cn (S.S.); minx@zju.edu.cn (X.M.); superli@zju.edu.cn (C.L.)
[2] Institute of Advanced Digital Technology and Instrument, Zhejiang University, No. 38 Zheda Road, Hangzhou 310027, China; 11315008@zju.edu.cn
[3] Alibaba Group, Ali Yun Feitian Park, Zhuan Tang Street, West Lake District , Hangzhou 310000, China; lilin.gll@alibaba-inc.com
* Correspondence: wanghonglie@zju.edu.cn

Received: 17 March 2020; Accepted: 30 March 2020; Published: 3 April 2020

Abstract: Vehicle re-identification is attracting an increasing amount of attention in intelligent transportation and is widely used in public security. In comparison to person re-identification, vehicle re-identification is more challenging because vehicles with different IDs are generated by a unified pipeline and cannot only be distinguished based on the subtle differences in their features such as lights, ornaments, and decorations. In this paper, we propose a local feature-aware Siamese matching model for vehicle re-identification. A local feature-aware Siamese matching model focuses on the informative parts in an image and these are the parts most likely to differ among vehicles with different IDs. In addition, we utilize Siamese feature matching to better supervise our attention. Furthermore, a perspective transformer network, which can eliminate image deformation, has been designed for feature extraction. We have conducted extensive experiments on three large-scale vehicle re-ID datasets, i.e., VeRi-776, VehicleID, and PKU-VD, and the results show that our method is superior to the state-of-the-art methods.

Keywords: vehicle re-identification; attention mechanism; Siamese neural networks

1. Introduction

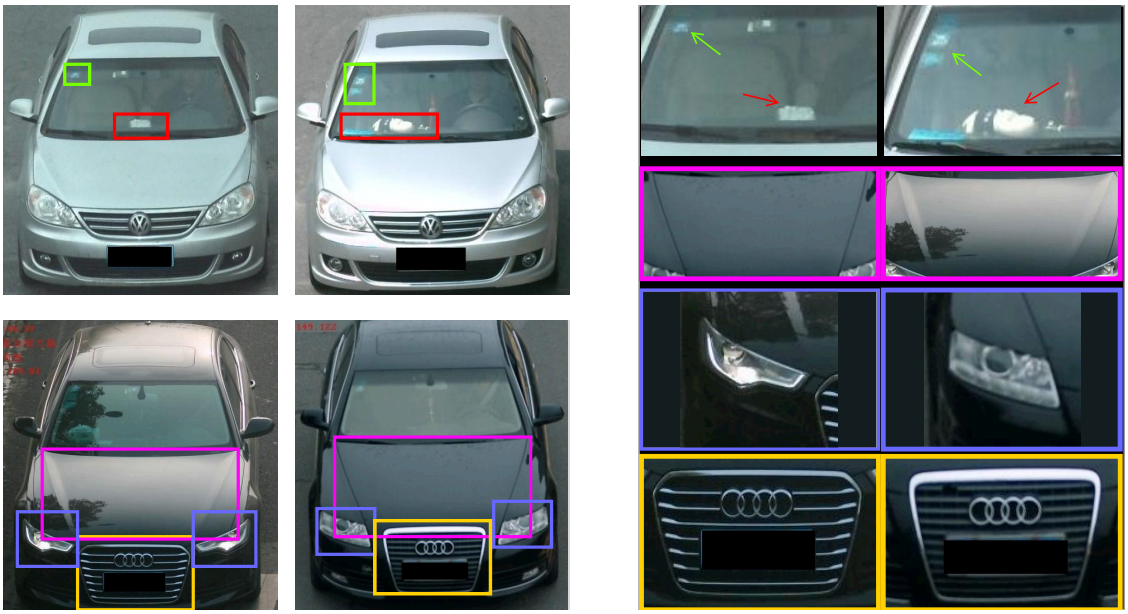
Vehicle re-identification (re-ID) returns a series of images containing the same vehicle ID as that of an image from a database. It is widely used in intelligent transportation, public security, and urban computing [1–3]. The straightforward way of vehicle re-ID is license-plate recognition [4]; however, a license plate is not always visible. Figure 1 shows the cases in which a vehicle ID cannot be determined based on the license plate. For example, the license plates of vehicles are sometimes occluded, illegally used, or invisible in some views. In particular, the number plates, models, and colors of genuine cars are sometimes used in other vehicles for performing illegal activities, such as smuggling, assembling, and scrapping and stealing vehicles, and the act of doing so is known as "car cloning." Therefore, vehicle IDs cannot be distinguished only based on license plates in several scenarios, and therefore, vehicle re-ID through other image-based features is urgently needed (Code available at https://github.com/WangHonglie/LFASM_pytorch).

(a) (b) (c)

Figure 1. Cases in which vehicle IDs cannot be determined according to the license plate: (**a**) License plate occlusion (e.g., dirt and reflection); (**b**) illegal use of license plate (e.g., License plate does not match the car); and (**c**) invisible license plate.

In recent years, deep learning [5], person re-ID [6–8] and fine-grained retrieval [9,10] have gained remarkable success. Vehicle re-ID datasets, such as VeRi-776 [3], VehicleID [11], and PKU-VD [12], have been released, thereby facilitating research on deep learning. However, because of the inconspicuous divergences among vehicles, vehicle re-ID is still difficult.

The main challenge of vehicle re-ID is distinguishing between two vehicles of the same or similar types. Images of different IDs which are captured from the same view may be more alike than those with the same ID but captured from different angles. Owing to camera resolution and shooting angle, obtaining a very high-quality vehicle image is sometimes difficult. Thus, vehicles always have inconspicuous differences, as shown in Figure 2. Different vehicles of the same model are similar in global appearance, and thus difficult to distinguish. Most existing studies focus on the entire image, and such subtle differences cannot be easily distinguished.



(a) (b)

Figure 2. (**a**) The two vehicles above are of the same type but have different IDs. They can be distinguished based on the windshield stickers (green box) and ornaments (red box). The two vehicles below are of a similar type but have different engine hoods (pink box), headlights (blue box), and air intakes (yellow box). (**b**) It is obviously easier to distinguish the vehicles based on their key parts.

Unlike general classification problems, the number of categories in a re-ID problem is uncertain. Therefore, some metric learning methods are committed to reducing the distance between images of the same vehicle, and enlarging the distance between images of different vehicles. Schroff et al. [13] proposed triplet loss, which directly optimizes the feature embedding. Bai et al. [14] combined the local structural

constraints to generate feature embedding more effectively. He et al. [15] proposed the Triplet-Center loss, which jointly considers the distances inside a class and relationships between different classes.

In this paper, we propose an effective feature extractor to find more fine-grained features by training it using two supervising methods. The first is an end-to-end classification module, in which a local net is aimed at selecting the regions of interest and another extractor, transposed convolutional layer (CTL), is proposed to find more implicit features. The other supervising method is a Siamese net, which matches the local features of two images and supervises attention better. Inspired by the spatial transformer network (STN) [16], we propose a perspective transformer network (PTN), which has greater degrees of freedom and can eliminate the deformation in images. To demonstrate an improved accuracy of retrieval, we have re-ranked the re-ID results given by Zhong et al. [17], thereby effectively ranking more true images at the top of the ranking list.

In summary, our major contributions to the literature of this field are threefold.

- We propose a local feature-aware Siamese matching model (LFASM) that can learn the local feature matching of different images. This is done by providing additional supervision so that the network is better trained, increasing the distance between classes, and reducing the distance within classes.
- To focus on the informative parts, we propose a local feature net that provide supervised attention to the regions of interest, thereby assigns different weights to different parts of the input. Unlike some methods [18–20] based on additional information (such as spatial, temporal, and part labels), our method is only based on the images of vehicles.
- We also propose a PTN, which can project a picture to a new view plane and eliminate the deformation of images. Compared to STN [16], PTN has greater flexibility for image transformation.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 describes the proposed local feature-aware Siamese matching model for vehicle re-ID and some details about our experiment. In Section 4, we discuss the experimental results, and Section 5 gives our conclusions.

2. Related Work

In this section, we review the existing studies on vehicle re-ID.

2.1. Vehicle Re-ID

Vehicle re-ID has become a major research area over the past decade. Owing to the development of the convolutional neural network (CNN) [21,22], the extraction of deeper features of images has become easier. Liu et al. [3] released the VeRi-776 dataset, which includes ultiview vehicle images, and Liu et al. [11] released VehicleID on a large scale. Yan et al. [12] contributed two rich annotated vehicle datasets, VD1 and VD2, obtained in real time from two cities, and containing high-resolution images. Wang et al. [19] utilized 20 key-point locations of vehicles to extract orientation information and proposed an orientation invariant feature embedding module. De et al. [23] proposed a two-stream Siamese classification model for vehicle re-ID, and Wei et al. [24] proposed an recurrent neural network-based hierarchical attention (RNN-HA) network, which combines a large number of attributes for vehicle re-ID. Bai et al. [14] proposed a group sensitive triplet embedding approach that can model the interclass differences. Recently, He et al. [20] considered both local and global representations to propose a valid learning framework for vehicle re-ID, however, their method depends on the labeled parts and is therefore labor-intensive. Krizhevsky et al. [21] first proposed the use of triplet loss to help the model directly learn feature embedding. The effect of triplet loss largely depends on the choice of training samples. Therefore, Hermans et al. [25] proposed hard mining to choose the hard positive and negative samples to train the network better. Furthermore, Chen et al. [26] proposed a quadruplet network for a greater impact of training.

2.2. Fine-Grained Visual Recognition

Although the identification of the main categories of objects is easy (such as computers, mobile phones, and water cups), determining highly refined object classification names (such as the type of bird and model of computers) is even more challenging. The greatest challenge is that the visual differences between the different subcategories of the same main category are minimal. Vehicle re-ID is a typical example of fine-grained recognition, the classification of which is mainly conducted using a part-based model and a representation learning model. Zhang et al. [27] employed the approach of learning of the entire object as well as the use of part detectors for fine-grained object recognition. Fully convolutional network (FCN) attention [28] can adaptively select the attention area and efficiently position multiple object parts. Lin et al. [29] proposed a bilinear structure comprising two feature extractors that can model pairwise feature interactions in an invariant manner.

2.3. Attention Mechanisms

The attention mechanism stems from the study of human vision. To make rational use of the limited visual-information-processing resources, humans must select specific parts of the visual area, and then focus on these parts. For example, when reading, only a few words are noticed at one time and then processed. The basic idea of visual-attention mechanisms is to enable a model to ignore irrelevant information and focus on the significant one. The attention mechanism has various forms of implementation; these mainly include soft and hard attention. Typical examples of soft attention include the STN [16], residual attention network [30], and two-level attention [31]. Although the hard attention model is required to predict the region of interest, it usually learns through reinforcement learning [32].

3. Proposed Method

We propose a local feature-aware Siamese matching (LFASM) model for vehicle re-ID. In this section, we provide a brief overview of the problem of vehicle re-ID and put forward our framework (Section 3.1). Then, we present the local feature-aware module, which is capable of learning more significant information (Section 3.2 and describe how we match the corresponding parts (Section 3.3). Finally, we propose our feature extractor in Section 3.4 and its implementation in Section 3.5.

3.1. Framework and Overview

Given a query vehicle image, the target of vehicle re-ID is to obtain a set of images from the gallery with the same ID as that of the candidate image. At present, we believe that vehicles with the same ID have more similar image feature embeddings. Therefore, these feature embeddings must be extracted and the similarity score between the embeddings of this candidate image and those of other images in the gallery must be calculated. The training set is then defined as $\{x_i, y_i\}_{i=1}^{N}$, where y_i represents the identification label of image x_i and N represents the number of training images. The similarity between query image q and gallery image g is defined as $D(\phi(q;\boldsymbol{\theta}), \phi(g;\boldsymbol{\theta}))$, where $\phi(\cdot;\boldsymbol{\theta})$ is the feature extractor and $D(\cdot)$ is a metric function. To obtain a better feature extractor, the parameter θ must be learned through gradient descent:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} L\left(\phi(\boldsymbol{x};\boldsymbol{\theta})^{\top}\boldsymbol{w}, \boldsymbol{y}\right), \tag{1}$$

where L is the loss function, and $\boldsymbol{w}$ is the weight vector.

Figure 3 shows the framework of the proposed LFASM model for vehicle re-ID. It comprises two branches: one in charge of the ID classification, and the other used for Siamese local feature matching to better supervise our attention module. Each branch comprises two modules including a local net to output an attention descriptor, $\boldsymbol{m} \in \mathbb{R}^{C\times H\times W}$. The score from the array $\boldsymbol{m}$ represents the amount of

attention required in this area. An attention-based feature extractor is used to extract the deep features of input images.

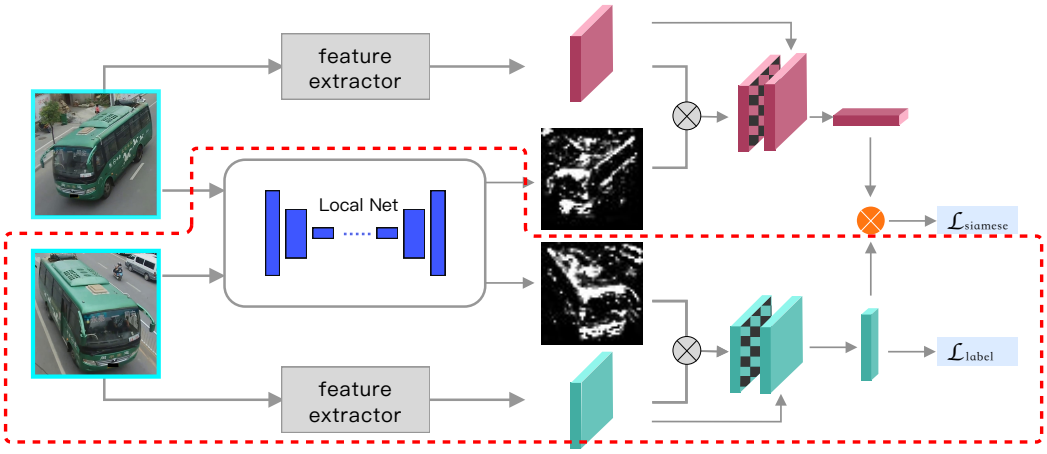


Figure 3. Framework of the proposed local feature-aware Siamese matching model (LFASM). The dotted line highlights the end-to-end model for feature extraction. The area outside the dotted line is the Siamese-feature-matching module.

3.2. Local Feature

This module aims to determine which informative parts deserve the greatest attention (e.g., outline, lights, windshield stickers, engine hood, and ornament), as shown in Figure 2. The goal of this study is to make our system more responsive to differences in these parts in order to effectively distinguish vehicle identities. The local feature net is an additional neural network that assigns different weights to different parts of the input. Our local net outputs an attention descriptor, $\boldsymbol{m} \in \mathbb{R}^{H\times W}$, representing the values of different parts of features. To prevent $\boldsymbol{m}$ from being negative, we used softplus [33] as our activation function. We project the attention descriptor, $\boldsymbol{m} \in \mathbb{R}^{H\times W}$, into the first feature map, $\boldsymbol{f}_1 \in \mathbb{R}^{C\times H\times W}$, by element-wise multiplication and obtain the masked feature, $\boldsymbol{f}'_1 \in \mathbb{R}^{C\times H\times W}$. For each tensor, $\boldsymbol{f}_{i,j} \in \mathbb{R}^C$ and $\boldsymbol{m}_{i,j} \in \mathbb{R}$, where (i,j) is the spatial location in $\boldsymbol{f}_1 \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{m} \in \mathbb{R}^{H\times W}$, the corresponding output tensor, $\boldsymbol{f}'_{i,j} \in \mathbb{R}^C$, can be determined as follows:

$$a_{(i,j)} = \boldsymbol{m}_{(i,j)} \times \boldsymbol{f}_{(i,j)}. \tag{2}$$

To limit the value of $\boldsymbol{m}_{(i,j)}$ between 0–1, we normalize the attention map $\boldsymbol{m}$ by

$$\boldsymbol{m} = \frac{\boldsymbol{m}}{\max\left(\|\boldsymbol{m}\|_p, \epsilon\right)}, \tag{3}$$

where $p = 2$ and $\epsilon = 1 \times 10^{-12}$.

Figure 4 shows the key parts of the vehicle images selected by our attention model. It can be seen that this model filters most of the background and some parts of the vehicle with poor information. The white part of Figure 4 represents the value of $\boldsymbol{m}$ close to 1, on the contrary, the black part represents the value of $\boldsymbol{m}$ close to 0. This module can better find the noteworthy part of the images and reduce the noise impression in the remainder of the image, so that the model can be more focused, and can more readily distinguish different vehicles.

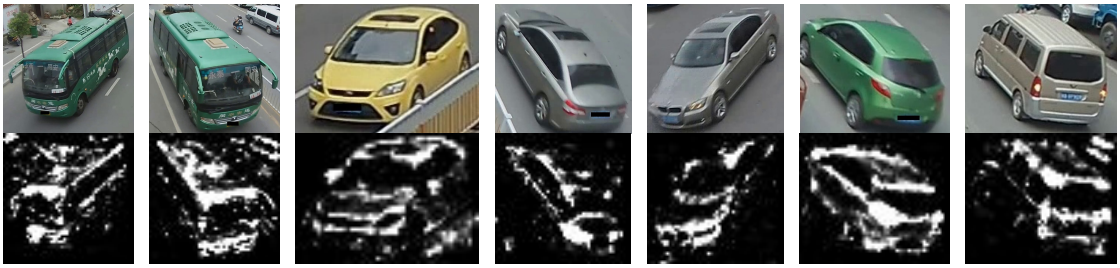


Figure 4. Visualizations of input images and their corresponding attention descriptor.White areas represent larger weights to the images.

3.3. Siamese Match

Although local features are always used in image retrieval [34,35], their use is not sufficient to distinguish images only according to class labels. To enhance the training of local features in the network, we propose a Siamese feature matching module. This module allows the network to know whether the two input pictures belong to the same ID. This is done by providing additional supervision so that the network is better trained, increasing the distance between classes, and reducing the distance within classes.

Given two images, $\{p, y_p\}$ and $\{q, y_q\}$, where y represents the identification label. The features of these two images can be denoted as $\phi(p;\boldsymbol{\theta})$ and $\phi(q;\boldsymbol{\theta})$, respectively. We measure the similarity of the two feature embeddings through a dot product:

$$s\left(p_i, q_j\right) = (W_\theta p_i)^T (W_\phi q_j), \tag{4}$$

where i and j are the positions of p and q in the feature map, respectively. The target label $\mathbf{y}$ can be computed as

$$\mathbf{y} = \frac{1}{\mathcal{C}(p)} \sum_{\forall j} s\left(p_i, q_j\right) g\left(p_j\right), \tag{5}$$

where $\mathcal{C}(p) = \sum_{\forall j} s\left(p_i, q_j\right)$ aims to normalize the result; $g = W_g p_i$ and W_g are the weights to be learned for this pair of features. While y_p=y_q, target label $\mathbf{y}$ converges to 1, else it converges to 0.

3.4. Attention-Based Feature Extractor

PTN. Vehicle pictures are taken by surveillance cameras, which essentially show the projection of the real scene on the camera chip, as in Figure 5. Owing to the different camera parameters and environmental factors, the obtained vehicle pictures often contain varying degrees of distortion. To eliminate the effects of projection transformations in different scenes, we propose a PTN, which predicts the transformation θ to apply to the input image using Equation (6), as shown in Figure 6. The main structure of the PTN comprise two convolutional networks, both of which output a 3×3 transformational matrix. We apply this transformational matrix to the features after the first block. The first two rows of the transformational matrix are identical to the affine matrix, which implements linear transformation and translation, and the third row is used to implement perspective transformation.

$$\begin{pmatrix} x_i' \\ y_i' \\ w_i' \end{pmatrix} = \mathcal{T}_\theta\left(G_i\right) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}, \tag{6}$$

where (x_i^t, y_i^t) are the target coordinates of the regular grid in the output feature map, and $(x_i^s = x'/w', y_i^s = y'/w')$ are the source coordinates in the input feature map that define the sample points. The main purpose of PTN is to eliminate the deformation by perspective transformations of vehicles in the images.

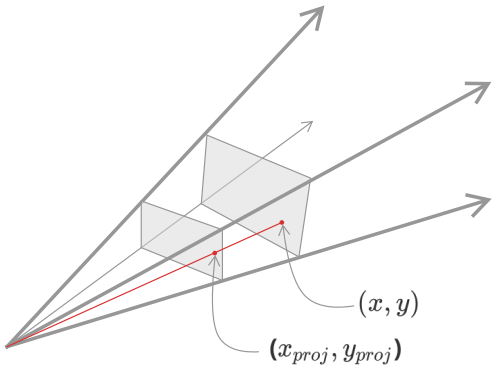


Figure 5. Schematic of projection transformation. (x, y) are the coordinates in the real scene, and (x_{proj}, y_{proj}) are the corresponding coordinates in the camera chip. Different cameras exhibit different degrees of distortion.

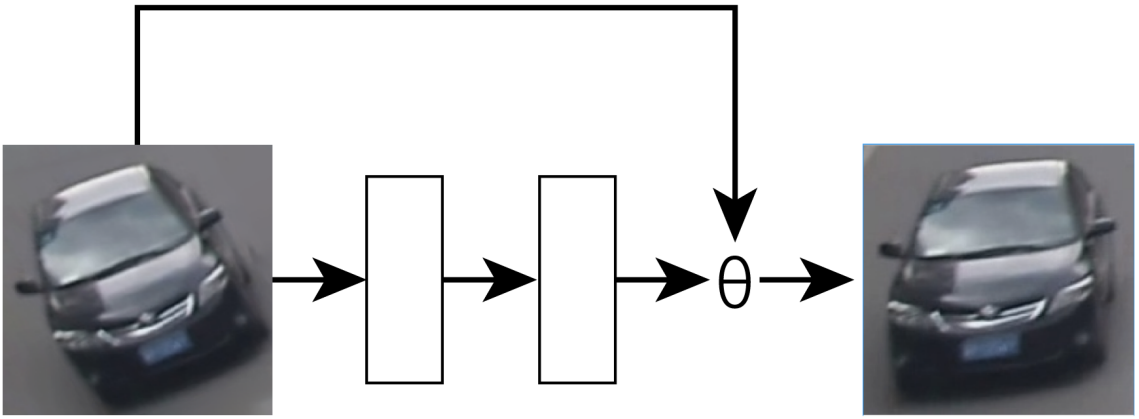


Figure 6. The perspective transformer network(PTN) architecture, composed of two convolutional networks, is used to transform the image. θ is a 3×3 transformation matrix.

CTL. We used ResNet-50 [36] as the base model of the feature extractor after PTN. As mentioned earlier, a component in the model was dedicated to extracting explicit key areas of images; some implicit features that play an important role in the re-ID task could not be extracted at the pixel level. Therefore, we applied the attention map $\mathbf{M} \in \mathbb{R}^{1 \times H \times W}$ to the intermediate feature map. The activated feature, $f' \in \mathbb{R}^{C \times H \times W}$, can be expressed as follows:

$$f' = \mathbf{M} \otimes f, \tag{7}$$

where $\otimes$ and f denote the element-wise multiplication and input feature, respectively. We obtained an attention map $\mathbf{M}$ through a transposed convolutional layer after a convolutional layer (CTL), as shown in Figure 7. The main purpose of the CTL is to extract the more informative part of the feature.

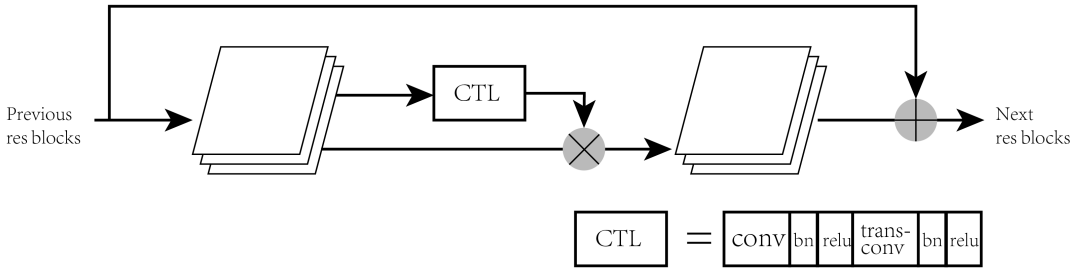


Figure 7. Diagram of a ResBlock in the Feature Extractor. To compute attention map **M**, we applied a convolutional layer and a transposed convolutional layer (both equipped with batch normalization and ReLU) on the convolution outputs in each block.

3.5. Implementation Details

In our experiments, ResNet-50 was used as the backbone network for feature extraction. The output of class block, $x \in \mathbb{R}^d$, was used as the acquired image representation, and $d = 512$ in our experiment. We measured the feature distances of two images by calculating the cosine distances. The stochastic gradient descent [37] with hyper-parameters ($weight\_decay = 5 \times 10^{-4}, momentum = 0.9, nesterov = True$) was adopted for model optimization. We set the learning rate of the fully connected layer to 0.005 and the other layers to 0.001 with a gradual decrease. All the images were scaled to 256×256 pixels.

Even if the features could be effectively clustered, if our query lies at the edge of the space in its category, we inevitably obtain a considerable amount of true negatives, as shown in Figure 8. One of the solutions to retrieve more true-positives is to enlarge the distance between different clusters. For this purpose, we set the arcFace loss [38] to measure the distance between the images; it uses angular distance to represent the distances between features. Furthermore, the scaling factor s was set to 10 in our experiments. Algorithm 1 depicts the whole pseudo code algorithm employed to train the proposed neural network architecture.

$$L = \frac{1}{N}\sum_i -\log\frac{e^{s\left(\cos\left(\theta_{y_i,i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i,i}+m\right)\right)} + \sum_{j\neq y_i} e^{s\cos\left(\theta_j,i\right)}} \tag{8}$$

Algorithm 1 Framework of LFASM.

Input: Two labeled images $\{p, y_p\}$ and $\{q, y_q\}$; Local Net $\theta\left(\cdot, \mathbf{w}_l\right)$; Feature extractor $\varphi\left(\cdot, \mathbf{w}_f\right)$
Output: Feature embedding f_{emb}

1: **while** maximum iterations not reached **do**
2: Extract features: f_i=$\varphi\left(i, \mathbf{w}_f\right)$ for all i=p,q;
3: Extract attention map: a_i=$\theta\left(i, \mathbf{w}_l\right)$ for all i=p,q
4: Deep feature map:$\phi(i;\boldsymbol{\theta}) = concat(a_i \times f_i, a_i)$for all i=p,q
5: $\mathbf{y} \leftarrow \frac{1}{C(p)}\sum_{\forall j} s\left(p_i, q_j\right) g\left(p_j\right)$ // i and j are the positions of p and q in the feature map,
6: $f_{emb} \leftarrow \phi(q;\boldsymbol{\theta})$
7: Fine tuning:min$(L(f_{emb}, y_q) + L(y, if\ y_q = y_p))$
8: **end while**

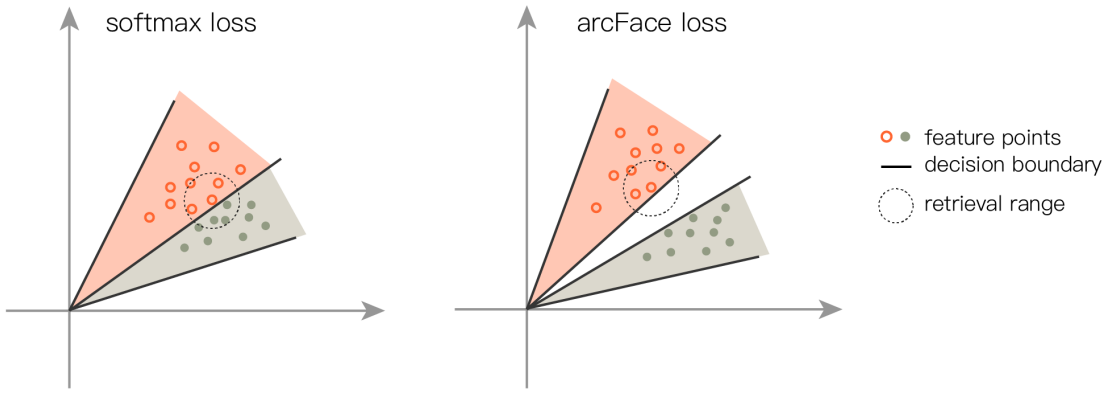


Figure 8. When the query is at the edge of the space in its category (**left sample**), it is more easily recalled as a false-positive. One of the methods to avoid this is to enlarge the inter class distance (**right sample**).

4. Experiments

4.1. Dataset and Metric

To verify the effectiveness of the proposed LFASM method, we conducted experiments on three important datasets, namely VehicleID, VeRi-776, and PKU-VD, and compared our results with those of the state-of-the-art vehicle methods for re-ID.

VeRi-776 [3] contains roughly 50,000 images of 776 vehicles captured by 2–18 cameras from different view angles. Every image in the query set contained 678 images of 200 vehicles, in which the images were captured by all the cameras in the cars.

VehicleID [11] comprises 221,763 images of 26,267 vehicles captured by different cameras and provides three test subsets of different sizes, with 800, 1600, and 2400 gallery images, respectively, such that we can evaluate our model on different data scales. The dataset contains images captured from two view angles: front and back.

PKU-VD [12] contains a large number of images with rich annotations (vehicle model and color). So far, it is the largest dataset for vehicle re-ID and is divided into two subsets: VD1 and VD2. The images in VD1 and VD2 were captured from surveillance videos and traffic cameras, respectively. They comprise approximately 1,098,649 and 807,260 images, respectively.

We computed the mean average precision (mAP) to evaluate the performance of our model. Average precision (AP) is a measure that considers both recall and precision. The AP for image q can be expressed as

$$AP(q) = \frac{\sum_k P(k) \times rel(k)}{N_{gt}(q)},$$

where $N_{gt}(q)$ is the number of ground truths, P(k) is the precision at rank k, and $rel(k) = 1$ when the matching of query image q to a test image is satisfied at rank k.

The mAP is the mean value of APs of all queries and can be expressed as

$$mAP = \frac{\sum_q AP(q)}{Q},$$

where Q is the number of query images. The mAP combines both precision and recall and is a comprehensive evaluation criterion.

4.2. Main Result

We present our results on three benchmark datasets: VeRi-776 [3], VehicleID [11], and PKU-VD [12] and compare the results with those of state-of-the-art vehicle re-ID methods. Table 1 shows the flops counter for each parts in LFASM.

Table 1. Flops counter for each parts in LFASM.

Module	Input Resolution	Params (M)	MACs (G)
LAF	234 × 234	51.69	19.46
PTN	234 × 234	0.97	0.09
SFM	234 × 234	8.40	0.008
LFASM	234 × 234	60.09	19.47

VeRi-776: The total numbers of query and gallery images were 1678 and 11,579, respectively. We compared the proposed LFASM with the state-of-the-art vehicle re-ID methods. First, we considered LOMO [39], which utilizes a handcrafted local feature for person re-ID; it solves the problems associated with view and illumination variations. The GoogLeNet fine-tuned on the CompCars dataset [40] can extract high-level semantic attributes of the vehicle appearance, while VAMI [41] is a viewpoint-aware attention model used to extract the core area from different views through an adversarial network, and QD-DLF [42] has different directional feature pooling layers. Siamese-CNN + Path-LSTM [18] is a two-stage framework that combines complex spatiotemporal information and effectively regularizes the re-ID results.

The comparison results on the VeRi-776, presented in Table 2 show that our proposed LFASM model achieves accuracies of 61.92%, 90.11%, and 92.91% mAP, top-1, and top-5, respectively. The ROC curves for the VeRi-776 are plotted in Figure 9, and the Area Under the Curve (AUC) is 0.974. The standard deviation values (std) of ap is 0.236. To determine the effect of each component in our model, we conducted an ablation study on VeRi-776. Our framework comprises three components: a PTN, local aware features (LAF), and Siamese feature match (SFM). We removed one component at a time and retrained the remaining network to evaluate the model performance in the absence of the removed component. Our model was able to achieve accuracies of 52.69%, 83.41%, and 90.81% for mAP, top-1, and top-5, respectively, without either PTN, SFM, or LAF, and the results were considered to be the baseline. The results of other comparative experiments are detailed in Table 3. The performance shows that the attention module has the most significant influence on the learning process; the other modules were also found to have improve the experimental results.

Table 2. Performance (%) comparison of different methods on the VeRi-776 dataset.

Method	mAP	Top-1	Top-5
BOW-SIFT [43]	1.51	1.91	4.53
LOMO [39]	9.78	23.9	39.1
GoogleNet [40]	17.8	50.7	67.5
VAMI [41]	50.1	77	90.8
QD-DLF [42]	61.8	88.5	**94.5**
Siamese-CNN + Path-LSTM [18]	58.3	83.5	90.0
LFASM (Ours)	61.92	90.11	92.91

Table 3. Ablation study on the VeRi-776 dataset.

Method	mAP	Top-1	Top-5
baseline	52.69	83.41	90.81
PTN	55.27	86.18	89.96
LAF + PTN	58.8	87.25	91.2
SFM + LAF +PTN	61.92	90.11	92.91

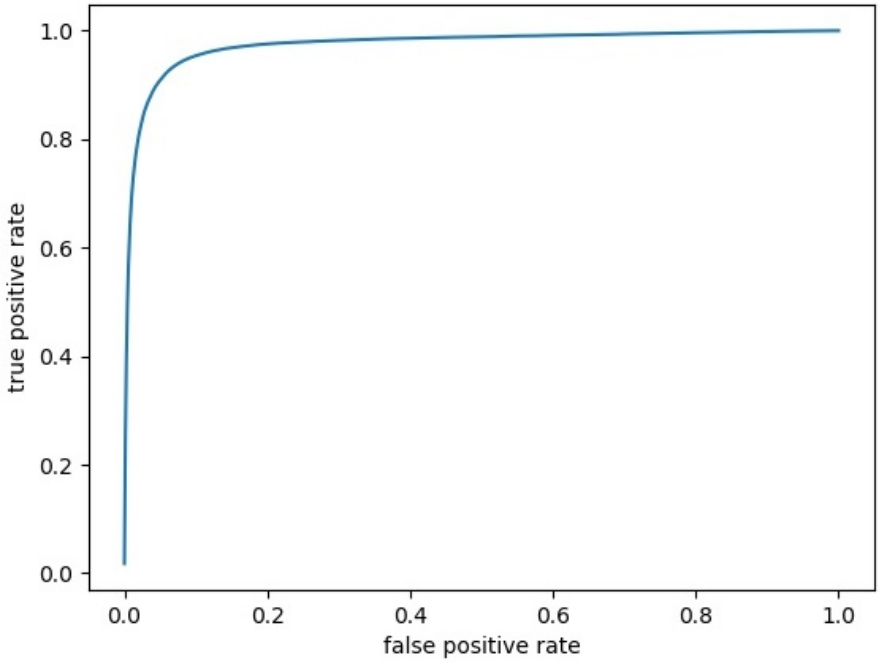


Figure 9. The receiver operating characteristic(ROC) curve of LFASM on VeRi.

VehicleID: VehicleID has a larger number of images than that of VeRi-776, with both front and rear views of the vehicles. The testing data of VehicleID were split into three subsets, as detailed in Table 4.

Table 4. Number of images and IDs in different sizes of VehicleID subsets.

ID	800	1600	2400
images	6493	13,377	19,777

Table 5 presents the comparison results on the VehicleID dataset. As shown, our model achieves the highest top-1 rate and exhibits robust performance with respect to other evaluation indices.

PKU-VD: Furthermore, we tested our method on the PKU-VD dataset, in turn, the two subdatasets: VD1 and VD2. Each subdataset is further divided into test sets of the following three sizes: small, medium, and large. Table 6 presents the number of test images in each subdataset. We followed the official setting provided by [12] for our model evaluation. Both VD1 and VD2 comprise 2000 query images, and the number of gallery images is listed in Table 6. Our method was also able to achieve good performance on a large-scale dataset, as detailed in Tables 7 and 8.

Table 5. Performance (%) comparison of different methods on the VehicleID dataset.

Method	Small			Medium			Large			Mean		
	mAP	Top-1	Top-5	mAP	Top-1	Top-5	mAP	Top-1	Top-5	mAP	Top-1	Top-5
DenseNet121	68.8	66.1	77.8	69.4	67.3	75.4	65.3	63.1	72.6	67.8	65.5	75.3
QD-DLF	**76.5**	72.3	92.5	**74.6**	70.7	88.9	68.4	64.1	83.4	73.2	69.0	88.3
Ours	76.2	**91.3**	**93.6**	72.2	**88.6**	**92.4**	**71.9**	**89.8**	**93.5**	**73.4**	**89.9**	**93.2**

Table 6. Total number of test images in VD1 and VD2.

Dataset	Small	Medium	Large
VD1	106,887	604,432	1,097,649
VD2	105,550	457,910	807,260

Table 7. Performance (%) comparison of different methods on the PKU-VD1 dataset.

Method	mAP
MGR [12]	79.1
QD-DLF [42]	87.5
LFASM (Ours)	**89.3**

Table 8. Performance (%) comparison of different methods on the PKU-VD2 dataset.

Method	mAP
MGR [12]	74.7
QD-DLF [42]	84.6
LFASM (Ours)	**86.2**

Figure 10 shows the results returned by the LFASM. Each row indicates a query image and its top-5 retrievals. As shown, the model performs effectively on most data except for those containing vehicles with a dim background.

Figure 10. Top-5 re-ID results. Green boxes denote true positives, while red boxes denote false positives. The five rows on the left are sampled from Vehicle-ID, and those on the right are sampled from VeRi-776.

Feature correspondences. The main feature of LFASM is its focus on the informative parts of images. Furthermore, we demonstrate the feature correspondence between the query and gallery images to reveal the function of LFASM when retrieving images. We extracted the local descriptor using an attention map and utilizing the nearest neighbor search (NNS) to find the best matches in each image. As shown in Figure 11, our model can effectively match the key parts (i.e., lights, windshield stickers, and engine hood). Therefore, our method can be used to retrieve images according to the number of matches in some other scenarios. However, using the distance between the feature vectors directly, accurate results can be obtained on the three datasets.

Figure 11. Visualization of the local-feature matches with the highest responsiveness among various pictures obtained by extracting and comparing local features.

5. Conclusions

In this paper, we proposed a model that combines the LAFs of vehicle images. In addition to global features, LFASM emphasizes the significant parts that are most likely to be different in vehicles with different IDs. This encourages the model to focus on more details in local regions. Furthermore, we applied local-feature matching, which compares the local features of two embeddings and helps the local net to better learn an attention map. Moreover, the PTN allows images to be aligned directly without the need to match key points, thereby facilitating image identification by the model. The experimental results on three large vehicle datasets show that LFASM can extract discriminative features and achieve excellent performance. On the other hand, as shown in Figure 10, the model performs well on most data except for those containing a dim background. In some other scenarios, such as in the case of different views of two cars or in the absence of shared parts in the two cars, it is difficult for our model to achieve effective identification. Improving recognition of vehicles with different views is the focus of future work.

Author Contributions: Conceptualization, H.W., L.G., L.Z. and C.L.; Funding acquisition, S.S.; Investigation, H.W., X.M. and C.L.; Methodology, L.G., L.Z. and C.L.; Supervision, L.G., L.Z., X.M. and C.L.; Visualization, L.G. and X.M.; Writing , H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D program of China under grant No. 2017YFB1402104.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, Y.; Capra, L.; Wolfson, O.; Yang, H. Introduction to the Special Section on Urban Computing. *ACM Trans. Intell. Syst. Technol.* **2014**, *5*, 1–2. [CrossRef]
2. Zhang, J.; Wang, F.Y.; Wang, K.; Lin, W.H.; Xu, X.; Chen, C. Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 1624–1639. [CrossRef]
3. Liu, X.; Wu, L.; Tao, M.; Ma, H. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2016.
4. Selokar, A.R.; Jain, S. Automatic number plate recognition system using a fast stroke-based method. *IEEE Trans. Multimed.* **2014**, *1*, 1–5.

5. Hao, X.; Zhang, G.; Ma, S. Deep Learning. *Int. J. Semant. Comput.* **2016**, *10*, 417–439. [CrossRef]
6. Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; Sun, J. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. *Comput. Sci.* **2017**, arXiv:1711.08184.
7. Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; Hu, J. Pose transferrable person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
8. Huang, H.; Li, D.; Zhang, Z.; Chen, X.; Huang, K. Adversarially occluded samples for person re-identification. In Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
9. Yao, H.; Zhang, S.; Zhang, Y.; Li, J.; Tian, Q. One-shot fine-grained instance retrieval. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; ACM: New York, NY, USA, 2017; pp. 342–350.
10. Zhang, X.; Zhou, F.; Lin, Y.; Zhang, S. Embedding label structures for fine-grained feature representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1114–1123.
11. Liu, H.; Tian, Y.; Wang, Y.; Lu, P.; Huang, T. Deep relative distance learning: Tell the difference between similar vehicles. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
12. Yan, K.; Tian, Y.; Wang, Y.; Zeng, W.; Huang, T. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 562–570.
13. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
14. Bai, Y.; Lou, Y.; Gao, F.; Wang, S.; Wu, Y.; Duan, L.Y. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Trans. Multimed.* **2018**, *20*, 2385–2399. [CrossRef]
15. He, X.; Zhou, Y.; Zhou, Z.; Bai, S.; Bai, X. Triplet-center loss for multi-view 3d object retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1945–1954.
16. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. *Comput. Sci.* **2015**, arXiv:1506.02025.
17. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1318–1327.
18. Shen, Y.; Xiao, T.; Li, H.; Yi, S.; Wang, X. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1900–1909.
19. Wang, Z.; Tang, L.; Liu, X.; Yao, Z.; Yi, S.; Shao, J.; Yan, J.; Wang, S.; Li, H.; Wang, X. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 379–387.
20. He, B.; Li, J.; Zhao, Y.; Tian, Y. Part-regularized near-duplicate vehicle re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Boston, MA, USA, 2012; pp. 1097–1105.
22. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
23. de Oliveira, I.O.; Fonseca, K.V.O.; Minetto, R. A Two-Stream Siamese Neural Network for Vehicle Re-Identification by Using Non-Overlapping Cameras. *Comput. Sci.* **2019**, arXiv:1902.01496.
24. Wei, X.S.; Zhang, C.L.; Liu, L.; Shen, C.; Wu, J. Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification. In Proceedings of the Asian Conference on Computer Vision, Perth, WA, Austria, 2–6 December 2018; Springer: Cham, Switzerland, 2018; pp. 575–591.
25. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737 .

26. Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond triplet loss: A deep quadruplet network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 403–412.
27. Zhang, N.; Donahue, J.; Girshick, R.; Darrell, T. Part-based R-CNNs for fine-grained category detection. European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 834–849.
28. Liu, X.; Xia, T.; Wang, J.; Yang, Y.; Zhou, F.; Lin, Y. Fully convolutional attention networks for fine-grained recognition. *Comput. Sci.* **2016**, arXiv:1603.06765.
29. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
30. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual Attention Network for Image Classification. *Comput. Sci.* **2017**, arXiv:1704.06904v1.
31. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification. *Comput. Sci.* **2014**, arXiv:1411.6447.
32. Rao, Y.; Lu, J.; Zhou, J. Attention-aware deep reinforcement learning for video face recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
33. Arribas, J.I.; Cid-Sueiro, J.; Adali, T.; Figueiras-Vidal, A.R. Neural architectures for parametric estimation of a posteriori probabilities by constrained conditional density functions. In Proceedings of the Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No. 98TH8468), Madison, WI, USA, 25–25 August 1999; IEEE: Hoboken, NJ, USA, 1999; pp. 263–272.
34. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the CVPR 2010—23rd IEEE Conference on Computer Vision & Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE Computer Society: Washington, DC, USA, 2010; pp. 3304–3311.
35. Jegou, H.; Perronnin, F.; Douze, M.; Sánchez, J.; Perez, P.; Schmid, C. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1704–1716. [CrossRef] [PubMed]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
37. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010, Paris, France, 22–27 August 2010; Springer: Cham, Switzerland, 2010; pp. 177–186.
38. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
39. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
40. Yang, L.; Luo, P.; Change Loy, C.; Tang, X. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3973–3981.
41. Zhou, Y.; Shao, L. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6489–6498.
42. Zhu, J.; Zeng, H.; Huang, J.; Liao, S.; Lei, Z.; Cai, C.; Zheng, L. Vehicle re-identification using quadruple directional deep learning features. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 410–420. [CrossRef]
43. Zheng, L.; Wang, S.; Zhou, W.; Tian, Q. Bayes merging of multiple vocabularies for scalable image retrieval. In Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1955–1962.