electronics

Article

Vehicle Re-Identification with Spatio-Temporal Model Leveraging by Pose View Embedding

Wenxin Huang [1], Xian Zhong [2,3,*], Xuemei Jia [4], Wenxuan Liu [2], Meng Feng [2], Zheng Wang [4] and Shin'ichi Satoh [5]

[1] School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China; wenxinhuang_wh@163.com
[2] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China; lwxfight@whut.edu.cn (W.L.); 246172@whut.edu.cn (M.F.)
[3] School of Electronics Engineering and Computer Science, Peking University, Beijing 100091, China
[4] School of Computer Science, Wuhan University, Wuhan 430072, China; jiaxuemeiL@163.com (X.J.); wangzwhu@whu.edu.cn (Z.W.)
[5] National Institute of Informatics, Multimedia Information Research Division, Tokyo 101-8430, Japan; satoh@nii.ac.jp
* Correspondence: zhongx@whut.edu.cn

Abstract: Vehicle re-identification (Re-ID) research has intensified as numerous advancements have been made along with the rapid development of person Re-ID. In this paper, we tackle the vehicle Re-ID problem in open scenarios. This research differs from the early-stage studies that focused on a certain view, and it faces more challenges due to view variations, illumination changes, occlusions, etc. Inspired by the research of person Re-ID, we propose leveraging pose view to enhance the discrimination performance of visual features and utilizing keypoints to improve the accuracy of pose recognition. However, the visual appearance information is still limited by the changing surroundings and extremely similar appearances of vehicles. To the best of our knowledge, few methods have been aware of the spatio-temporal information to supplement visual appearance information, but they neglect the influence of the driving direction. Considering the peculiar characteristic of vehicle movements, we observe that vehicles' poses on camera views indicating their directions are closely related to spatio-temporal cues. Consequently, we design a two-branch framework for vehicle Re-ID, including a Keypoint-based Pose Embedding Visual (KPEV) model and a Keypoint-based Pose-Guided Spatio-Temporal (KPGST) model. These models are integrated into the framework, and the results of KPEV and KPGST are fused based on a Bayesian network. Extensive experiments performed on the VeRi-776 and VehicleID datasets related to functional urban surveillance scenarios demonstrate the competitive performance of our proposed approach.

Keywords: vehicle re-identification; spatio-temporal; features fusion; optimization

check for updates

Citation: Huang, W.; Zhong, X.; Jia, X.; Liu, W.; Feng, M.; Wang, Z.; Satoh, S. Vehicle Re-Identification with Spatio-Temporal Model Leveraging by Pose View Embedding. *Electronics* **2022**, *11*, 1354. https://doi.org/10.3390/electronics11091354

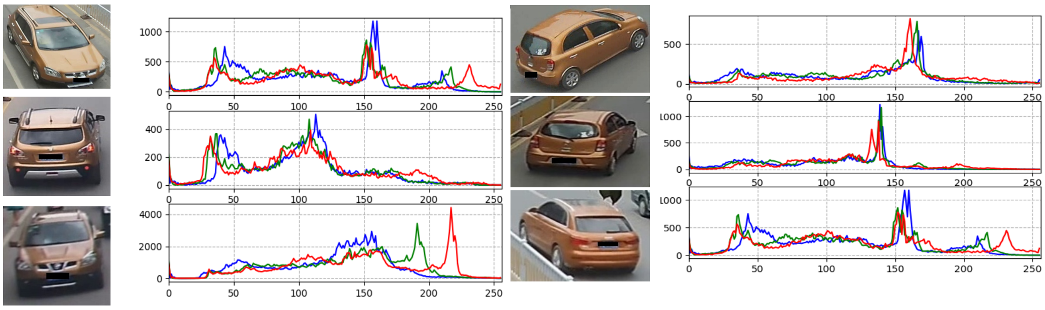
Academic Editor: Gwanggil Jeon

Received: 23 March 2022
Accepted: 20 April 2022
Published: 24 April 2022

1. Introduction

In recent years, there has been an explosive growth of massive surveillance camera installations that become an indispensable part of human life in urban public spaces with particular benefits for security services [1–5]. One of the significant functions of urban surveillance systems is to assist security officers in locating suspicious objects whereby the task entails an input query object's image, and the objective is to search for the same target in videos recorded by various cameras. With the rapid development and recent advances in person re-identification (Re-ID) techniques, vehicle Re-ID as the task of matching identical vehicles captured by different cameras distributed over non-overlapping scenes has attracted increasing attention. However, compared to person Re-ID [6–9], vehicle Re-ID is still a frontier topic along with the growing explosion in the use of urban surveillance cameras among the large-scale areas. It can be viewed as a definite image retrieval task in intelligent surveillance, which is different from traditional vehicle detection, recognition and categorization problems [10–12].

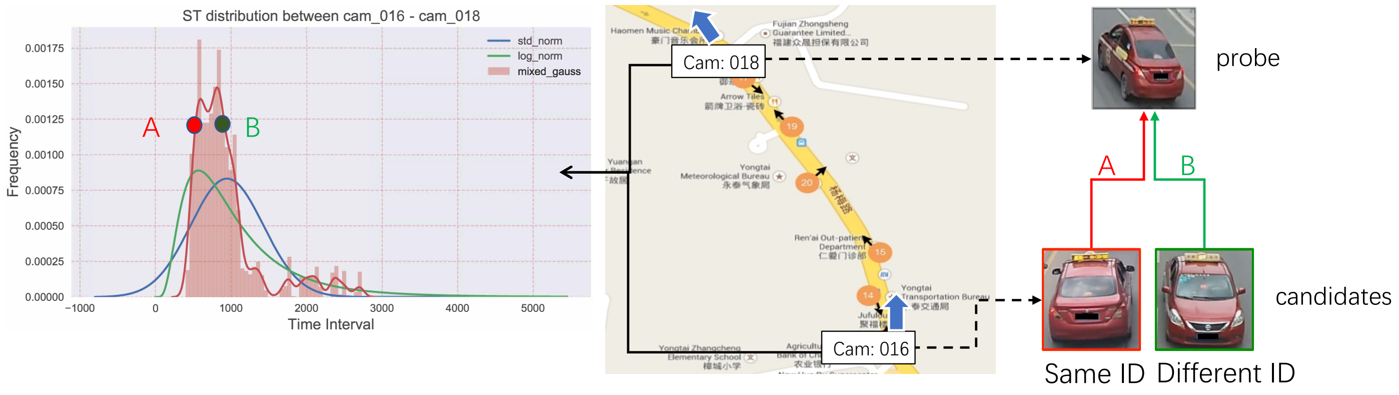
As object Re-ID technologies have developed, there have been considerable advances in vehicle Re-ID based on appearance features, such as texture, color and semantic properties, as well as orientation-invariant features inspired by person Re-ID research. Furthermore, there are few studies adopting license plate recognition. However, a vehicle license plate helps little in an open traffic environment where high-quality license plate images are difficult to obtain. Due to the impact of occlusions, illumination, poses indicating various orientations, and other factors, the task of vehicle Re-ID remains challenging in a large-scale road surveillance network. Moreover, the vehicle Re-ID task exhibits the specific characteristics that differ from person Re-ID. The general appearances of two different vehicles of the same color and type can be quite similar from the same viewpoint due to the characteristics of solid matter, while the same vehicle in different environments has a dramatically varying visual appearance in practical surveillance. Figure 1 illustrates various scenarios and demonstrates that satisfactory performance cannot be obtained by only relying on appearance features. On the other hand, distinctive characteristics in the vehicle Re-ID task not only cause problems but are also beneficial. The identities (IDs) of different vehicles driven by diverse individuals cannot be exactly the same. Additionally, driving behaviors are constrained by traffic rules, and vehicles move in certain directions that cannot be changed easily in continuous time in separate driving ways. Consequently, vehicles' spatio-temporal information is more effective than appearance for vehicle Re-ID. Accordingly, we propose to construct a two-step framework: (1) exploring appropriate appearance features for representing the vehicles, and (2) utilizing the spatio-temporal clues that assist in the retrieval process. The rationale is that in the physical world, the same vehicle cannot be seen by two different non-overlapping cameras at the same time, and different vehicles have different movement behaviors.



(a) The color distribution of vehicles with same identity (b) The color distribution of vehicles with different identity

Figure 1. (**a**) Illustration of the RGB color histogram distributions of the same vehicle. (**b**) Illustration of the RGB color histogram distributions of different vehicles. Different color lines denote different channels in RGB images.

Nevertheless, utilizing the proposed framework in a practical urban surveillance system faces three significant challenges. Firstly, in the early stage, the vehicle Re-ID task is explored in the specified view. While in open scenarios, appearance variations across different poses of a vehicle are far more important than those of a person. In contrast, different vehicles with the same color, type and pose always look more alike than different poses of a given vehicle. Secondly, although spatio-temporal information helps optimize the performance of vehicle Re-ID, existing spatio-temporal models neglect the problem that the time intervals of different vehicles driving through the viewable areas of a given pair of cameras from different directions are probably the same or close. This problem demonstrated by Figure 2 causes the spatio-temporal model not to perform as expected. Finally, spatio-temporal information and appearance features are heterogeneous and cannot be measured directly.



(a) Spatiotemporal Distribution (b) Camera Topology (c) Vehicle Images

Figure 2. (**a**) Illustration of the spatio-temporal distribution between cam_018 and cam_016. (**b**) The camera topology. The blue arrow indicates the camera shooting direction. (**c**) The probe image captured by cam_018 and the candidates images captured by cam_016. Branch A is an ID-matched pair, and the time interval between the probe and candidate corresponds to point A on the curve. Branch B is ID a non-matched pair, and the time interval between the probe and candidate corresponds to point B on the curve. Correct results cannot be obtained only by vehicle visual features and spatio-temporal information. It can be observed that a combination of camera shooting direction and vehicles' poses can be used to assist the judgment.

Existing vehicle Re-ID approaches also notice the vehicle Re-ID problem in a practical urban surveillance environment and make significant efforts to address the above challenges. Based on person Re-ID research, pose view has been explored to distinguish the appearances [13–16]. Local regions represented as different poses such as the front, sides and rear in vehicle images can be regarded as supplementary to the general visual appearance in recognizing vehicles. In fact, we explore the phenomenon that there are always some keypoints in a certain pose that are invisible in vehicle images captured by surveillance cameras, while some other keypoints in other local regions can be seen. Inspired by this observation, we consider exploiting simplified keypoints of a certain side view to define the poses. In addition, we utilize the integration of the general visual appearance and each pose's viewpoint appearance features with keypoints of an embedding to represent the visual features of vehicles. As Figure 3 illustrates, the fixed shooting directions of cameras in the camera network are different, and so are a vehicle's poses captured by cameras. By jointly considering the vehicles' poses and camera shooting directions, the relative movement directions of vehicles can be estimated. For example, if a vehicle's pose captured by the camera is the front, we can infer that the vehicle is approaching the camera and vice versa. The driving direction of the vehicle in the region can be determined based on the location of cameras and the relative direction. As vehicles' driving directions are generally stable in a real-world traffic environment, the vehicles with the same driving directions caught by each camera are most likely to be the same. Therefore, we construct a spatio-temporal model guided by pose view, which introduces an estimation of driving directions to optimize the spatio-temporal model. Finally, since visual features and spatio-temporal information refer to different data models that are difficult to fuse directly, we map them into the probability space to optimize the vehicles identification's performance. According to the above, we integrate pose view with keypoint embedding into both the visual appearance model and spatio-temporal model.

In conclusion, the contributions of this study can be summarized as follows:

- We propose a two-branch framework to optimize the performance of the vehicle Re-ID task, where one branch is a visual model combining general visual appearance and pose view with keypoint embedding, and the other is a spatio-temporal model guided by pose view, so that poses are involved in both branches' models.

- To the best of our knowledge, we take the lead in introducing a spatio-temporal model guided by pose view which optimizes the existing spatio-temporal methods, filtering the cases of similar vehicles appearing in the same section of the road but driving in different directions. Moreover, we account for this mechanism in deciding how pose view guides the spatio-temporal model.
- We design a fusion model with the generated visual model and a spatio-temporal model guided by pose view based on a Bayesian network. Extensive experiments on public vehicle Re-ID datasets demonstrate the effectiveness and superiority of the proposed approach. Our proposed model can also be easily extended to a practical urban surveillance environment.

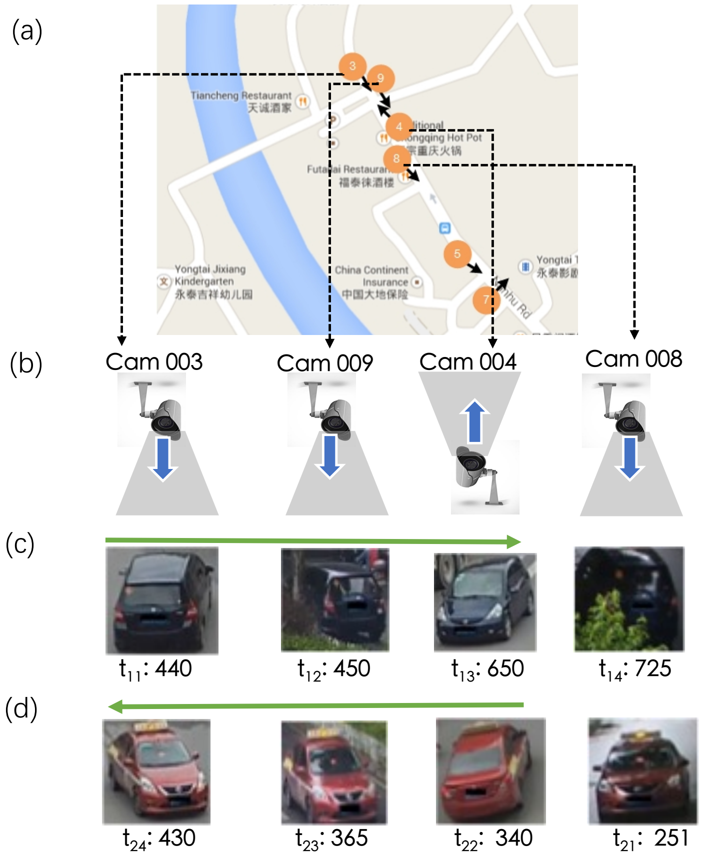


Figure 3. Relationship of vehicles' spatiotemporal information and vehicles' poses. (**a**) Map and camera topology. (**b**) Viewing directions of cameras. (**c**) Changes in spatiotemporal data and poses of a vehicle as it drives from cam_003 to cam_008. (**d**) Changes in spatiotemporal data and poses of a vehicle as it drives from cam_008 to cam_003.

Compared with our previous work [17], we reform the framework with keypoints-based pose view guiding—both the visual model (KPEV) and spatio-temporal model (KPGST), which significantly improves the accuracy for vehicle retrieval in the open scenarios. Additionally, we explain the mechanism of how the pose view guides the spatio-temporal model and design a fusion model based on Bayesian network, which makes the process more reasonable. To evaluate the effectiveness of our framework, we conduct extensive experiments on two large-scale vehicle Re-ID datasets, VeRi-776 [18] and VehicleID [19]. Comprehensive experiments demonstrate that the framework not only improves the accuracy but also remains efficient.

2. Related Works

2.1. Person Re-ID

Person Re-ID has been widely studied in the computer vision field, which has various important applications in recent years. Using features based on convolutional neural networks (CNN) and deep metric learning has led to significant progress being made in solving the person Re-ID problem [20–27], e.g., occlusions, clothes variations, domain gap. Pose-based approaches [28–31] based on deep neural networks have been applied to person Re-ID tasks and extensively explored. Furthermore, contextual information such as spatio-temporal information, object locations, the topology of 120 cameras, etc., has been widely exploited in multi-camera-based person Re-ID tasks [32–34].

2.2. Vehicle Re-ID

Vehicle Re-ID in a large-scale urban surveillance is a frontier area that has attracted more interest in recent years. Feris et al. [35] proposed a vehicle detection and retrieval framework. The vehicles were firstly classified by different types, sizes and colors. Recent works on the vehicle Re-ID task mainly concentrate on making breakthroughs based on deep neural networks. Some vehicle Re-ID research relies on a specific view and addresses the single-view case. Liu et al. [19] focused on precise vehicle retrieval while considering metric mapping to cluster positive samples. Yan et al. [36] exploited multi-grain ranking constraints and further optimize the ranking by the likelihood loss function. Some works began to tackle the vehicle Re-ID task on the road network. With global and partial multi-regional distance-based feature learning, Chen et al. [37] design a three-branch network to learn coarse-to-fine vehicle information. Zheng et al. [38] develop a two-stage architecture, aiming for learning robust vehicle representation progressively.

Liu et al. [18,39] released a high-quality vehicle Re-ID dataset VeRi-776 with 776 vehicle IDs captured by 20 cameras in a large-scale scenario, which contributed significantly to vehicle Re-ID research. The researchers explored an appearance-based model by integrating low-level and high-level semantic features based on CNNs. Additionally, they also utilized license plate information. Lou et al. [40] collected a new dataset captured by a large surveillance system containing 174 cameras covering a large urban district. It is the first vehicle Re-ID dataset that is collected from unconstrained conditions arising from the data collection in a real surveillance camera network of a city-scale district, covering a huge diversity of viewpoints, resolutions, illuminations, camera sources, weathers, occlusions, backgrounds, vehicle models in the wild, etc. However, it is difficult to correctly match vehicles among a large number of candidates in a real-world traffic environment by distinguishing vehicles only through visual features due to variations of image background, viewpoints and illumination. Inspired by the existing Re-ID methods focused on the person Re-ID problem, pose-based, multi-view and spatio-temporal relations-based approaches are taken into consideration when tackling the vehicle Re-ID task.

2.3. Multi-View and Contextual Models

Multi-view and contextual models have been widely exploited in multi-camera systems and nowadays are adopted into vehicle Re-ID research. Zapletal et al. [41] and Sochor et al. [42] proposed to use a 3D structure for aligning different vehicle vehicles' faces to extract accurate features. Zhou et al. [43] proposed to exploit the Spatially Concatenated ConvNet and a CNN-long short-term memory (CNN-LSTM) bidirectional loop to learn transformations across different viewpoints of vehicles and applied that approach to the Toy Car Re-ID dataset. Additionally, locations of keypoints are helpful, as the learned features can be aligned well by such keypoints. Wang et al. [44] used local region features of different orientations based on 20 keypoint locations and combined such features to learn more accurate features. Moreover, the authors utilized spatio-temporal information to optimize performance. Shen et al. [45], Liu et al. [46] and Li et al. [47] also exploited spatio-temporal relations. Although they also achieved good performances, they overlooked the influence of different driving directions on spatio-temporal models. Regarding vehicle–orientation–camera as a triplet and reforming shape similarity as orientation and camera Re-ID, Zhu et al. [14] utilize camera and orientation similarity as the penalty to obtain final similarity after training vehicle, orientation and camera Re-ID, respectively.

3. Proposed Methods

3.1. Problem Formulation

The task of vehicle Re-ID is to retrieve all vehicles in a camera network that have the same ID as the query vehicle. For the clarity of the problem definition, some notations used in the vehicle Re-ID problem are described as follows. In an urban surveillance system, we define a camera network C, which is composed of $M + 1$ cameras $C_0, C_1, \ldots, C_m$ with a non-overlapping field of view. The i-th vehicle at C_n is represented by O_n^i. The ID of the vehicle O_n^i is denoted by $\Upsilon(O_n^i)$. The moment it is captured is represented by t_n^i. For vehicle Re-ID, we expect to find the vehicles that have the same ID as a query vehicle in different camera views. The match probability between the probe O_n^i and the candidate O_m^j can be expressed as follows $P(\Upsilon(O_n^i) = \Upsilon(O_m^j) \mid m \neq n)$.

3.2. Model Overview

The architecture of the model is illustrated in Figure 4, which contains the following main steps:

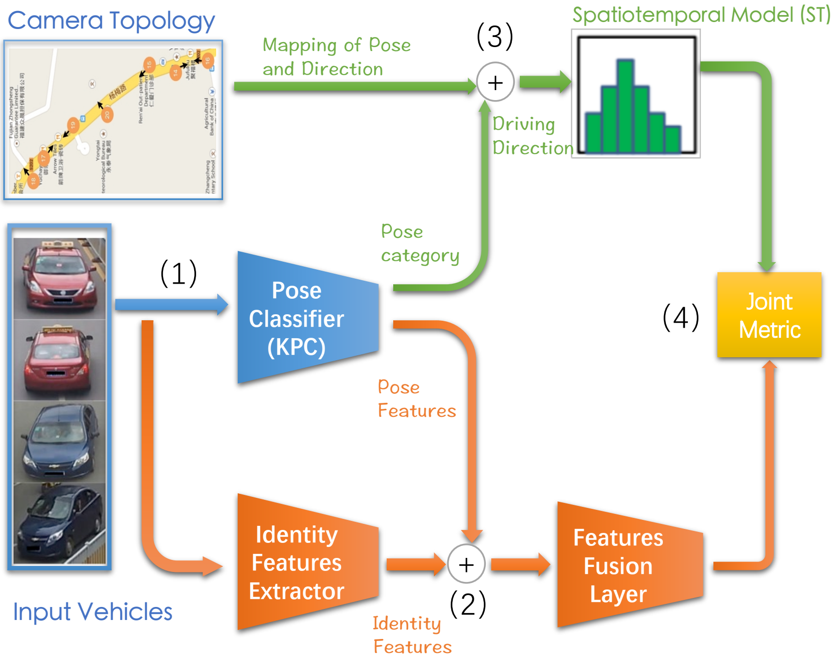


Figure 4. Illustrates the framework of the proposed model, consisting of (1) Extracting the vehicle's pose features and estimating pose category, (2) Integrating the vehicle's pose features and ID features, (3) Estimating the vehicle's driving direction based on pose and guiding the spatio-temporal model, (4) Joint metric of the visual features and spatio-temporal features.

step 1 Extracting the vehicle's pose features and estimating pose category. In this step, a Keypoint-based Pose Classifier (KPC) is proposed to extract the vehicle's poses features and estimate the pose category (Section 3.3).

step 2 Integrating the vehicle's pose features and ID features. In this step, a Keypoint-based Pose Embedding Visual model (KPEV) is proposed and composed of the above KPC, an ID feature extractor and a feature fusion module. The ID feature extractor is exploited to extract vehicle ID features. Afterwards, the feature fusion module is utilized to integrate the vehicle's ID features and pose features into pose-invariant features. The visual match probability of a probe and a candidate is assessed by feature distance (Section 3.4).

step 3 Estimating the vehicle's driving direction based on pose and guiding the spatio-temporal model. In this step, the relationship between the vehicle's driving direction and pose category is inferred from the camera topology and shooting directions. The spatio-temporal model is guided by the vehicle's driving direction and is called the Pose-Guided Spatio-Temporal model (PGST). The spatio-temporal match probability is inferred by the PGST model (Section 3.5).

step 4 Joint metric of the visual features and spatio-temporal features. In this step, we assume that the vehicle's visual occurrence probability and spatio-temporal occurrence probability are independent from each other. The vehicle's final match probability is calculated by combining the visual probability with spatio-temporal probability leveraging by pose view based on the Bayes' formula. We rank probabilities in the descending order and select the top rank (Section 3.6).

In the following sections, we will describe in detail the design of each key component of the model and analyze those components.

3.3. Keypoint-Based Pose Classifier

As shown in step 1 of Figure 4, a Keypoint-based Pose Classifier (KPC) is proposed to extract the vehicle's poses features and estimate pose category. Inspired by the OIFE [44], it contains four modules, i.e., pose keypoint regressor, global feature extractor, pose feature extractor and pose classifier module. The architecture of KPC is illustrated in Figure 5.

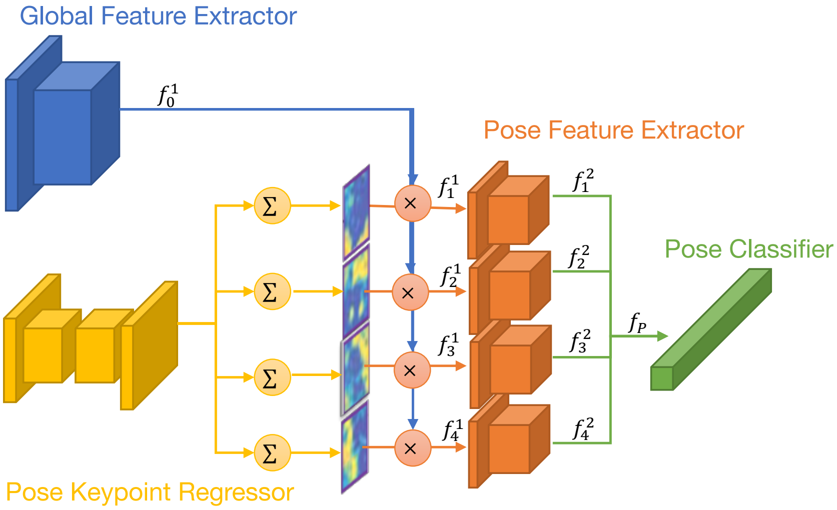


Figure 5. Illustrates the network architecture of the KPC model.

The pose keypoint regressor estimates the vehicle's 20 keypoint locations. The annotation of these keypoints is shown in Table 1. These keypoints are chosen as some principal vehicle components, e.g., the wheels, the lamps, the logos, the rear-view mirrors, and the license plates. The architecture of the pose keypoint regressor is a stacked hourglass network [48] that is usually used to generate response maps of human joints for human pose estimation. The pose keypoint regressor takes a vehicle image as input and yields 20 response maps of the vehicle's keypoints. As shown in Figure 6, poses are classified into four categories: front, rear, left side, and right side. The resulting 20 response maps are assigned to four clusters according to the visible points on each pose category: $C_1 = [5,6,7,8,9,10,13,14]$, $C_2 = [15,16,17,18,19,20]$, $C_3 = [1,2,6,8,11,14,15,17]$, and $C_4 = [3,4,5,7,12,13,16,18]$. The final output region masks are computed as the summation of all the feature maps belonging to each cluster:

$$R_I = \sum_{l \in C_I} F_l \tag{1}$$

where $i = 1,2,3,4$.

Table 1. The annotations of 20 keypoints on vehicle.

No.	Keypoints	No.	Keypoints
1	left-front wheel	11	left rear-view mirror
2	left-back wheel	12	right rear-view mirror
3	right-front wheel	13	right-front corner of vehicle top
4	right-back wheel	14	left-front corner of vehicle top
5	right fog lamp	15	left-back corner of vehicle top
6	left fog lamp	16	right-back corner of vehicle top
7	right headlight	17	left rear lamp
8	left headlight	18	right rear lamp
9	front auto logo	19	rear auto logo
10	front license plate	20	rear license plate

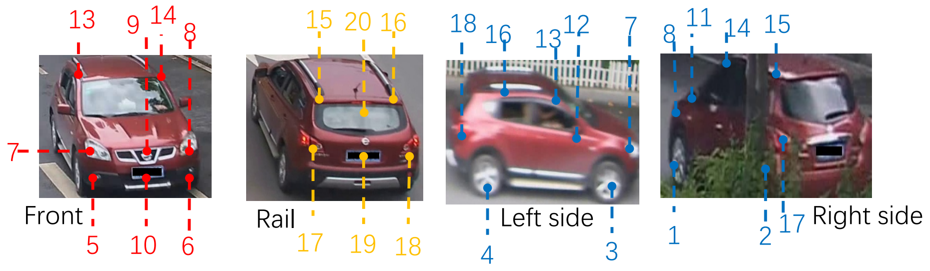


Figure 6. Illustrates four vehicle's poses categories and the keypoint annotations at the vehicle. The numbers indicate the keypoints.

The global feature extractor is adopted to obtain the global features. The network architecture consists of two ResNet Blocks [49]. The input images are resized to 256 × 256 and convolved by two ResNet Blocks. The size of the output feature map f_0^1 is 64 × 64 and the number of output channel is 512. A region mask R_I has the same size as a feature map f_0^1. For each local branch, f_0^1 is element-wisely multiplied to obtain each preliminary pose local feature maps $f_I^1 = f_0^1 \cdot R_I, (i = 1,2,3,4)$.

The pose feature extractor is applied to generate pose intermediate features. The network architecture is the features extractor of AlexNet [50]. For each local branch, the preliminary pose local feature map f_I^1 is convolved and max pooled four times. The output is the intermediate pose local feature f_I^2. The size of f_I^2 is 7 × 7, and the number of the output channel is 256. Each intermediate pose local feature is flattened into 12,544 dimensional feature vectors. These four feature vectors are concatenated together to obtain the final pose feature $f_P = [f_1^2, f_2^2, f_3^2, f_4^2]$. The total number of dimensions is 1024.

The pose classifier module yields the pose category. The network architecture is the classifier of AlexNet [50]. The final pose features f_P are fed to two fully connected (FC) layers, and the output is one of four pose categories.

3.4. Fusion of Visual Features and Pose Features

Because of the variation of vehicle's poses, the appearance features of vehicles vary significantly. High intra-class variance renders the use of only less discriminative visually based features. Hence, we propose to integrate both pose features and ID features to obtain pose-invariant feature f. A Keypoint-based Pose features Embedding Visual model (KPEV) is therefore proposed to extract pose-invariant features. As Figure 7 illustrates, KPEV consists of KPC, an ID features extractor and feature fusion layers. The network architecture of the ID features extractor is a ResNet18 [49] network with the last average-pooling layer and the FC layer removed. The ID features extractor takes the vehicle images as input and yields ID feature f_I with dimensions of 512 × 8 × 8. The ID feature f_I is flattened into 32,768 dimensional feature vectors. Feature fusion layers consist of two FC layers and batch normalization (BN) layers. The feature is generated by the concatenation of pose feature f_P and ID feature f_I, and it is fed into the first BN layer (BN1) and the FC layer (FC1) to obtain 2048-dimensional feature vectors. The FC1 layer is employed to project the second BN layer (BN2) and FC layer (FC2), taking 2048 dimensional feature as input and predicting the vehicle's ID. Pose-invariant feature f is the output of the FC1 layer in feature fusion layers.

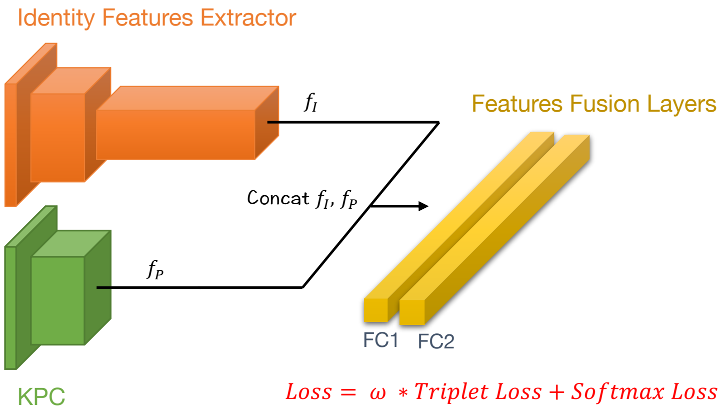

Identity Features Extractor
f_I
Features Fusion Layers
Concat f_I, f_P
f_P
FC1
FC2
KPC
$Loss = \omega * Triplet\ Loss + Softmax\ Loss$

Figure 7. Network architecture of the KPEV model.

To optimize the feature distances between inter-class and intra-class samples, a multiple loss function is designed for metric learning. It is formulated as:

$$\mathcal{L} = \omega \mathcal{L}_{Triplet} + (1-\omega)\mathcal{L}_{Softmax} \tag{2}$$

where we use hyperparameter ω to balance two types of loss.

Feature f_n^i is the pose-invariant feature of vehicle O_n^i. The visual-based match probability of O_n^i and O_m^j can be regarded as a similarity between features. Specifically, the similarity between f_n^i and f_m^j is measured by cosine distance, which is defined as:

$$P(\mathrm{Y}(O_n^i) = \mathrm{Y}(O_m^j) \mid f_n^i, f_m^j) = \frac{f_n^i \cdot f_m^j}{\|f_n^i\|\|f_m^j\|} \tag{3}$$

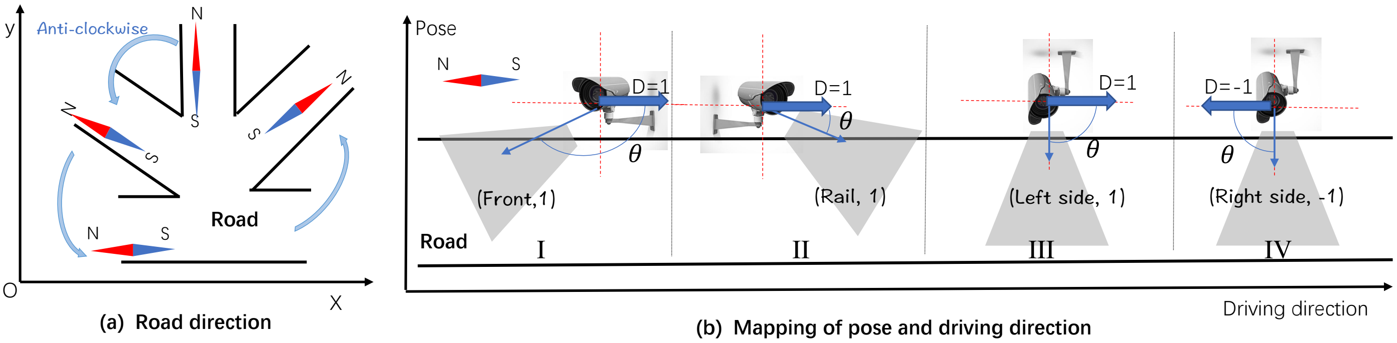
3.5. Pose-Guided Spatio-Temporal Model

In the urban surveillance system, camera shooting angles and camera topography are readily obtained. Combining a vehicle's captured poses and camera shooting angles, we can estimate the vehicle's relative driving direction. In a vehicle's complete trajectory, the vehicle must have the same relative driving direction at every location in the trajectory. In addition, the driving direction of the vehicle establishes the order in the spatio-temporal sequence. Hence, we develop an algorithm for the pose-guided spatio-temporal model. It contains three steps: estimating the vehicle's relative driving direction by combining the vehicle's captured poses and camera shooting directions, constructing the relative geographic relationship between camera pairs and computing the spatio-temporal match probability.

3.5.1. Estimating the Vehicle's Relative Driving Direction

A vehicle's poses and pose confidence can be estimated by the pose classifier. Pose confidence is denoted by α and is the maximum output probability score of the classifier. As shown in Figure 8a, we set the relative direction to be along the tangent of the road.

The north of the road is always the upstream of the road. Figure 8b illustrates the mapping of pose and driving direction. Different driving directions result in various captured poses. Angle θ denotes the inclination of shooting direction and the vehicle's driving direction. Based on observation, we examine the relationship between angle θ and poses, as Table 2 illustrates. The mapping of driving directions and a vehicle's poses is constructed based on the above assumption. If a vehicle's pose is known, the vehicle's relative driving direction D_n^i can be estimated by combining the cameras' shooting direction and the vehicle's poses. Specifically, the vehicle's driving direction from the north to the south of the road is represented as $D_n^i = 1$, and it is otherwise represented as $D_n^i = -1$.



(a) Road direction

(b) Mapping of pose and driving direction

Figure 8. (**a**) Illustrates the definition of relative direction of the road. (**b**) Illustrates the mapping of vehicle's poses and driving directions.

Table 2. The relationship between θ and poses.

Angle (θ)	$[0°, 60°]$	$[60°, 120°]$	$[120°, 180°]$
Pose	Front	Left or Right side	Rear

3.5.2. Constructing the Relative Geographic Relationship between Camera Pairs

To clearly describe the geographical relationship between camera pairs, we establish a 2D coordinate system for the camera network topology. As shown in Figure 9, the relative position between camera pair $C_{n,m}$ can be represented by a vector. For example, the position of C_n is represented by $G_n = (x_n, y_n)$, the $\overrightarrow{G_{n,m}}$ denotes the geographical position of C_n relative to that of C_m, which is formulated as $\overrightarrow{G_{n,m}} = G_m - G_n$, and its positive and negative signs represent the direction.

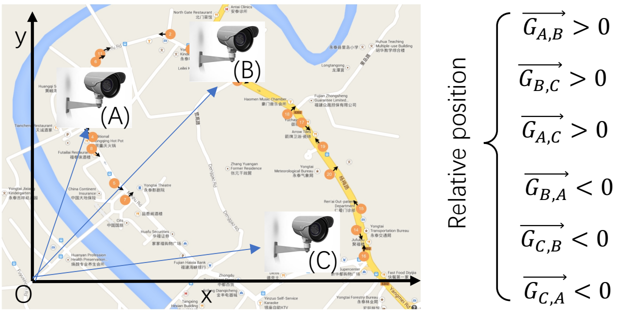

y
(A)
(B)
(C)
O
X
Relative position
$\overrightarrow{G_{A,B}} > 0$
$\overrightarrow{G_{B,C}} > 0$
$\overrightarrow{G_{A,C}} > 0$
$\overrightarrow{G_{B,A}} < 0$
$\overrightarrow{G_{C,B}} < 0$
$\overrightarrow{G_{C,A}} < 0$

Figure 9. Relative geographic relationship between camera pairs.

3.5.3. Computing the Pose Guide Spatio-Temporal Match Probability

To explore the effect of spatio-temporal information for vehicle Re-ID, we select vehicles with 576 IDs from 20 camera pairs in an urban surveillance and analyze all positive samples, and among them, the spatio-temporal translations time of each camera pair. The histogram in Figure 10 shows the time intervals of translations between camera pairs. We select several probability curves to fit the trend and observe that the lognormal-like probability model [44] performs best. Based on this assumption, the spatio-temporal match probability P_{st} is computed by the spatio-temporal model $f_{st}(\cdot)$. The time interval of O_n^i and O_m^j is represented by $\Delta t = |t_n^i - t_m^j|$ and $P(t_n^i, t_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j)) = f_{st}(\Delta t)$.

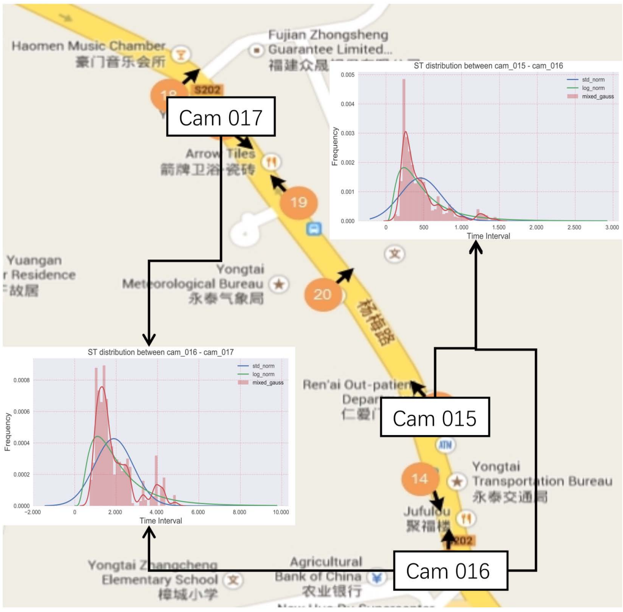


Figure 10. Spatio-temporal statistics for camera pairs.

Vehicle directions determine the direction of time flow and the relative transition direction. The probability $P(D_n^i, D_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j))$ is estimated by Algorithm 1.

Algorithm 1 Pose-Guided Spatio-Temporal Probability

```
Input: Probe vehicle image, $O_n^i$, and its occurrence moment, $t_n^i$; candidate vehicle's image, $O_m^j$, and its occurrence moment, $t_m^j$; geographic coordinates $G_n$ of camera $C_n$; geographic coordinates $G_m$ of camera $C_m$; mapping $M$ of pose and driving direction; spatio-temporal model $f_{st}(\cdot)$; direction-guidance hyperparameter $\lambda$;
Output: Pose-guided spatio-temporal match probability $P_{kpgst}$;
Input $O_n^i$ and $O_m^j$ into the KPC model to estimate the probe's pose $C_n^i$, the probe's pose confidence $\alpha_n^i$, the candidate's pose $C_m^j$ and the candidate's pose confidence $\alpha_m^j$;
Map $C_n^i$ and $C_m^j$ in $M$ to obtain the vehicles' relative driving direction $D_n^i, D_m^j \in \{-1, 1\}$;
Compute the relative distance vector between cameras $C_n, C_m$, $\overrightarrow{G_{n,m}} = G_n - G_m$;
if $D_n^i = D_m^j = 1$ and $(t_n^i - t_m^j) \cdot \overrightarrow{G_{n,m}} \leq 0$ or $D_n^i = D_m^j = -1$ and $(t_n^i - t_m^j) \cdot \overrightarrow{G_{n,m}} \geq 0$ then
    $P(D_n^i, D_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j)) = \alpha_n^i \cdot \alpha_m^j$
else
    $P(D_n^i, D_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j)) = 1 - \alpha_n^i \cdot \alpha_m^j$
end if
```

The direction and appearance time of the vehicle can be assumed to be independent. The pose guide spatio-temporal match probability P_{pgst} can be formulated as,

$$\begin{aligned} &P(t_n^i, t_m^j, D_n^i, D_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j)) \\ =&P(t_n^i, t_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j))P(D_n^i, D_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j)) \end{aligned} \tag{4}$$

3.6. Joint Metric of Visual Probability and Spatio-Temporal Probability

The data type of visual information is different from that of spatio-temporal information. In terms of the vehicle Re-ID problem, a vehicle's visual-based feature distribution and spatio-temporal feature distribution are independent of each other. Hence, we can integrate these two types of data based on a Bayesian probability model. The match probability of the probe O_n^i and the candidate O_n^i can be formulated as,

$$
\begin{aligned}
&P(\Upsilon(O_n^i) = \Upsilon(O_m^j) \mid t_n^i, t_m^j, D_n^i, D_m^j, f_n^i, f_m^j) \\
=&\frac{P(f_n^i, f_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j))P(t_n^i, t_m^j, D_n^i, D_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j))}{P(t_n^i, t_m^j, D_n^i, D_m^j)}
\end{aligned}
\tag{5}
$$

The prior probability $P(t_n^i, t_m^j, D_n^i, D_m^j)$ can be assumed to be equal for all data. Then, it is formulated as,

$$
\begin{aligned}
&P(\Upsilon(O_n^i) = \Upsilon(O_m^j) \mid t_n^i, t_m^j, D_n^i, D_m^j, f_n^i, f_m^j) \\
=&P(f_n^i, f_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j))P(t_n^i, t_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j))P(D_n^i, D_m^j \mid \Upsilon(O_n^i) = \Upsilon(O_m^j))
\end{aligned}
\tag{6}
$$

4. Experiments

4.1. Dataset and Evaluation Metric

To evaluate the effectiveness of our vehicle Re-ID framework, we mainly implement experiments on the VeRi-776 [18] dataset that contains spatio-temporal information on vehicles' movements and camera topology. Experiments on the VeRi-776 dataset show that the proposed approach effectively improves the performance of vehicle Re-ID performed without considering license plates.

- VeRi-776 [18]: The VeRi-776 dataset is the only one containing spatio-temporal information and topology of cameras. This dataset contains images collected with 20 cameras in a real-world traffic surveillance environment. The cameras are installed in arbitrary positions and directions. For each vehicle in the dataset, there are images captured from multiple viewpoints. The dataset contains approximately 50,000 images and 9000 tracks of 776 vehicles. It is split into a training set containing 37,781 images of 576 vehicles and a testing set with 11,579 images of 200 vehicles. A subset of 1678 query images is used to retrieve the corresponding images from the remainder of the set.
- VehicleID [39]: The VehicleID dataset contains data captured during the daytime by multiple real-world surveillance cameras distributed in a small city in China. There are 26,267 vehicles of 221,763 images in total in the entire dataset. Each image is attached with an ID label corresponding to its ID in the real world.

We follow the evaluation protocol proposed in [45]. The R-1, R-5, and R-20 accuracy as well as the mean Average Precision (mAP) are also adopted to evaluate the accuracy of the methods.

4.2. Effect of Estimating Vehicles' Poses Keypoint Regressor

The primary contribution that needs to be investigated is the effectiveness of extracting the vehicle's poses feature. Inspired by OIFE [44], the keypoint regressor [48] is an orientation-based regional proposal module. The keypoints of a vehicle in a visible orientation can be located by the keypoint regressor. The OIFE work [44] manually specifies 20 keypoints' coordinates and orientation class (such as front, rear, right side, left side) annotations for each vehicle in the VeRi-776 dataset. We adopt these orientation annotations to train and test the keypoint regressor. All input images are resized to 256 × 256 pixels. The ground truth heat map of size 64 × 64 consists of a 2D Gaussian (with standard deviation of 1 px) centered on the keypoint location. The Mean Squared Error (MSE) loss is computed to compare the predicted heat map with ground truth heat map. The RMSprop [51] optimization with a learning rate of 1.5×10^{-4} is used to train the network. The final prediction of the network is the maximum activating location of the heat map for a given keypoint. Parameter r_0 denotes the allowed error threshold of distance between the ground truth and a prediction value. As Table 3 shows, our trained keypoint regressor attains 93.87% accuracy (r_0 = 5) on the testing images and exceeds by 1.37% the result of OIFE [44].

Table 3. Comparisons (%) of the prediction accuracy of keypoint regressor on VeRi-776 dataset. The best results are shown in bold.

Models	*r* = 5	*r* = 3
OIFE [44]	92.05	**88.80**
Our keypoint regressor	**93.87**	80.5

4.3. Effect of Performance on Vehicle's Poses Classifier

The proposed KPC method classifies vehicle poses into four categories (front, rear, left side and right side). The AlexNet mode [50] is regarded as our basic pose classifier. We use two ResNet [49] blocks to obtain the vehicle global features. Hence, we also compare our KPC approach with ResNet18.

Our KPC training strategy is as follows. The learning rate is 2.5 × 10^{-3}. The loss function is the softmax loss. The size of a minibatch is set to 64 and the models are trained for 80 epochs. We first fix the parameters of the trained pose keypoint regressor in KPC and load the AlexNet blocks and the ResNet blocks in KPC. The AlexNet is pre-trained on ImageNet. The ResNet block is pre-trained on VeRi-776 and VehicleID by classifying each vehicle's ID. Note that there are two poses (front and rear) for each vehicle in VehicleID. We replace the four branches with two branches and output two categories.

Table 4 displays the number of training sets and testing sets on VeRi-776 and VehicleID datasets. Table 5 shows the performance of different methods on different VeRi-776 and VehicleID datasets. We use the precision and recall to evaluate each category separately. These two indicators are calculated by following Equation (7). The TP_I denotes the True Positive rate of category i. The FP_I denotes the False Positive rate of category i. The FN_I denotes the False Negative rate of category i.

$$\begin{aligned} precision_I &= \frac{TP_I}{TP_I + FP_I} \\ recall_I &= \frac{TP_I}{TP_I + FN_I} \end{aligned} \tag{7}$$

Table 4. The number of pose annotations on VeRi-776 and VehicleID datasets.

Dataset	Type	Front	Rear	Left Side	Right Side	Total
VeRi-776	train	501	501	504	504	2010
	test	251	250	238	252	991
VehicleID	train	5600	5600	-	-	11,200
	test	4575	1425	-	-	6000

Table 5. Comparisons (%) of the pose classifier on VeRi-776 and VehicleID datasets. The best results are shown in bold.

Dataset	Models	Front		Rear		Left Side		Right Side	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
VeRi-776	AlexNet [50]	93.80	85.12	73.20	91.04	92.04	74.55	82.80	**95.83**
	ResNet18 [49]	95.60	87.71	**84.00**	95.02	93.47	**81.06**	83.40	95.21
	KPC	**96.40**	**97.37**	81.40	**95.99**	**93.88**	79.72	**90.00**	91.09
VehicleID	AlexNet [50]	**98.58**	99.85	99.51	**95.62**	-	-	-	-
	ResNet18 [49]	98.40	99.78	99.30	95.09	-	-	-	-
	KPC	98.01	**99.89**	**99.65**	93.98	-	-	-	-

The performance of the pose classifier reflects the confidence degree of the pose category output. It also directly affects the performance of the pose-guided spatio-temporal model. It can be seen from the results that when the vehicle pose category increases, the effect of the pose classifier based on keypoint positioning gradually becomes obvious, and the average accuracy rate on each category is as high as 90.0% or more.

4.4. Effect of Extracting Vehicle's Visual Features

We select the final feature fusion layer in KPEV as the pose-invariant feature representation for the Re-ID task. The ResNet18 model [49] is regarded as our base network structure. To optimize the KPEV model, we vary hyper parameter ω, which is used to adjust the ratio between the softmax loss and the triplet loss. If ω equals 0, the final loss degenerates into the softmax loss, and in the other extreme of ω equal to 1, the final loss turns out to be the triplet loss. By multiple sets of comparative experiments on parameter ω, we find that the KPEV performs best on the VeRi-776 dataset when ω = 0.25, and on the VehicleID dataset when ω = 0.5. The learning rate for the VeRi-776 and VehicleID datasets starts from 0.001 and 0.01, respectively, and it decays every 150 rounds. The size of a minibatch is set to 64, and the models are trained for 500 epochs.

Our KPEV training strategy includes two steps. (1) We pre-train the base ResNet18 network with the softmax loss. (2) After parameters of the trained KPC have been fixed, we train the backbone of the feature fusion layer in KPEV and fine-tune the parameters of the ID feature extractor. Specifically, for the VehicleID dataset, if 13,161 IDs in the entire training set are used, it is difficult to achieve convergence of the KPEV model with the softmax loss. Hence, we randomly extract vehicle images of 1000 IDs from the entire training set and use them as a training subset for training the KPEV model.

Table 6 presents performance comparisons between KPEV and the state-of-the-art. For the VeRi-776 dataset, it attains approximately 9.94% mAP and 3.42% R-1 gains over KPEV over our baseline, which is ResNet18. For different test sizes of the VehicleID dataset, it achieves improvements of 4.98%, 4.60% and 3.89% in R-1 accuracy over ResNet18, and 3.19%, 2.53%, and 1.61% over LCSR [52].

Table 6. Comparison (%) of vehicle Re-ID performance of visual models on the VehicleID dataset. The best results are shown in bold.

Methods	VehicleID									VeRi-776		
	Test Size = 800			Test Size = 1600			Test Size = 2400			Query = 1678		
	R-1	R-5	R-20	R-1	R-5	R-20	R-1	R-5	R-20	R-1	R-5	R-20
LOMO [53]	19.76	32.01	45.04	18.85	29.18	39.87	15.32	25.29	35.99	9.78	23.87	39.14
GoogleNet [10]	47.88	67.18	78.46	43.40	63.86	74.99	38.27	59.39	72.08	17.81	52.12	66.79
FACT [18]	49.53	68.07	75.54	44.59	63.86	74.99	39.92	60.32	72.92	18.75	52.21	72.17
MixedDiff [19]	48.93	75.65	88.47	45.05	68.85	79.88	41.05	63.38	76.62	-	-	-
XVGAN [54]	52.87	80.83	91.86	49.55	71.39	81.73	44.89	86.65	78.04	24.65	60.20	77.03
OIFE [44]	-	-	-	-	-	-	-	-	-	48.00	65.92	87.66
C2F [55]	61.10	81.70	-	56.20	76.20	-	51.60	72.20	-	-	-	-
VAMI [13]	63.12	83.25	92.40	52.87	75.12	83.49	47.34	70.29	79.95	50.13	77.03	90.82
FDA-Net [40]	-	-	-	59.84	77.09	-	55.53	74.65	-	-	-	-
MLL + MLSR [56]	65.78	78.09	-	64.24	73.11	-	60.05	70.81	-	-	-	-
View-EALN [57]	67.19	78.20	-	63.23	77.12	-	59.98	74.20	-	50.32	81.34	90.88
SCFCL [58]	-	-	-	-	-	-	53.10	69.70	81.50	-	-	-
MALN [59]	67.71	**87.90**	-	61.50	82.77	-	54.51	77.29	-	45.06	72.05	88.86
MCD [60]	64.54	72.12	-	62.07	69.65	-	60.24	68.48	-	-	-	-
LCSR [52]	69.04	84.44	-	66.40	80.41	-	62.31	75.89	-	-	-	-
Baseline [49]	67.25	82.08	90.34	64.33	78.05	87.49	60.03	74.53	84.37	44.81	79.26	89.92
KPEV	**72.23**	87.41	**92.59**	**68.93**	**84.52**	**91.31**	**63.92**	**78.29**	**86.21**	**54.75**	**82.68**	**91.34**

XVGAN [54] and VAMI [13] both identify the vehicle's ID under different poses by generating the characteristics of the vehicle under different perspectives. Compared with them, our KPEV proposes to extract the pose's invariant feature, the mAP results over KPEV on the VeRi-776 dataset achieve 30.10% and 4.62% improvements, respectively. Moreover, the R-1 results over KPEV on the VehicleID dataset achieve approximately 19.00% and 9.00% improvements, respectively. Our visual framework is inspired by OIFE [44], but differently, we use four detailed branches to mine vehicle pose features and adopt triplet loss to fuse local features and global features. The result over KPEV on the VeRi-776 dataset attains 6.75% mAP gains over OIFE [44].

4.5. Comparisons of Spatio-Temporal Models

To investigate the effectiveness of pose-guided spatio-temporal models for vehicle Re-ID, we consider three spatio-temporal pattern comparisons as follows: (1) the standard normal spatio-temporal model (StdNorm ST); (2) the lognormal spatio-temporal model (LogNorm ST); (3) spatio-temporal histograms based on Parzen window (STHist).

- StdNorm ST: The probability of a vehicle's transition interval Δt observed by the camera pair is assumed to follow the standard normal distribution and can be computed using Equation (8). Parameters $\sigma_{n,m}$ and $\mu_{n,m}$ are the variance and the mean value (based on the training set) of the transition interval observed by camera pair (n, m).

$$P_{st} = \frac{1}{\sigma_{n,m}\sqrt{2\pi}} \exp\big(-\frac{(\Delta t - \mu_{n,m})^2}{2\sigma_{n,m}^2}\big) \tag{8}$$

- LogNorm ST: Inspired by the OIFE [44], we can estimate the spatio-temporal probability of transition interval Δt using the lognormal distribution. The respective probability is calculated using Equation (9). The two parameters $\mu_{n,m}$ and $\sigma_{n,m}$ are the mean and standard deviation of the variable's natural logarithm based on the training set.

$$P_{st} = \frac{1}{\Delta t \sigma_{n,m}\sqrt{2\pi}} \exp\big(-\frac{(\ln \Delta t - \mu_{n,m})^2}{2\sigma_{n,m}^2}\big) \tag{9}$$

- STHist: Inspired by the work [61] of spatio-temporal person Re-ID, we apply it to vehicle Re-ID using the histogram with the Parzen Window approach to estimate the spatio-temporal distribution. The spatio-temporal probability is computed by Equation (10). All the transition interval positive training samples between the camera pair (n, m) are placed into several bins. The width of each bin is denoted by d. Variable $N_{n,m}^k$ represents the total number of vehicles in the k-th bin. If the Δt is a transition interval of estimation, $k = \frac{\Delta t}{d}$ is calculated firstly, next, $N_{n,m}^k$ is estimated, and finally, we can obtain P_{st} as follows:

$$P_{st} = \frac{N_{n,m}^k}{\sum\limits_{l} N_{n,m}^l} \tag{10}$$

We apply our PGST algorithm to the above spatio-temporal models. All spatio-temporal methods are applied to the VeRi-776 dataset. Table 7 presents a comparison of performance of spatio-temporal models for Re-ID. It shows that the LogNorm spatio-temporal method performs best, which verifies that this log-norm distribution is more suitable for the spatio-temporal pattern of vehicles. Figure 11a shows that the PGST method generally attains more than 10% mAP gains over the spatio-temporal model without pose guidance. Figure 11b shows a comparison of CMC curves. We observe that PGST has clear advantages according to R-1 to R-50 match rates. The result demonstrates that the pose-guided spatio-temporal model method can effectively reduce Re-ID errors.

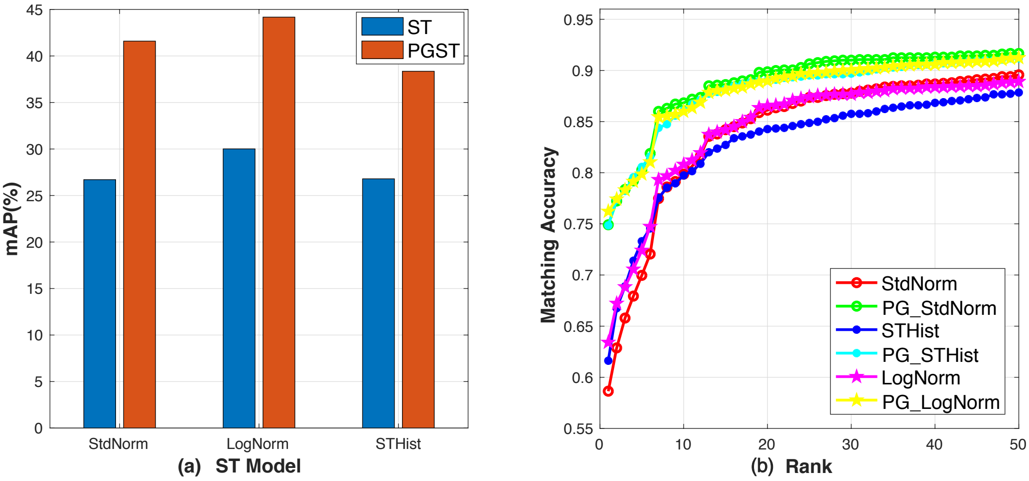


Figure 11. CMC results of comparisons of spatio-temporal model on the VeRi-776 dataset.

Table 7. Comparisons (%) of the performance of spatio-temporal model on the VeRi-776 dataset. The best results are shown in bold.

Models	Pose-Guided ST	mAP	R-1	R-5
StdNorm	×	26.70	58.64	69.96
	✓	40.25	73.89	79.92
LogNorm	×	30.01	63.40	72.40
	✓	**43.10**	**75.32**	**80.57**
STHist	×	26.79	61.62	73.30
	✓	37.51	73.78	79.91

4.6. Comparisons over Fusion Model with Visual and Spatio-Temporal

To extensively investigate the performance of the fusion model with visual and spatio-temporal features on the Re-ID task, we perform detailed experiments with various fusion methods. Our method of fusion with visual and spatio-temporal features is built on the Bayesian merging method. Table 8 presents a comparison of all of our fusion models on the VeRi-776 dataset. Figure 12 shows the CMC curves of fusion models on the VeRi-776 dataset. Compared to fusion models without pose-guided spatio-temporal features, fusion models with such KPGST features attain over 4.00% improvements of mAP. Moreover, we observe that the KPGST model of LogNorm PGST has the best performance, and it can be deemed to be the optimal model for the entire algorithm.

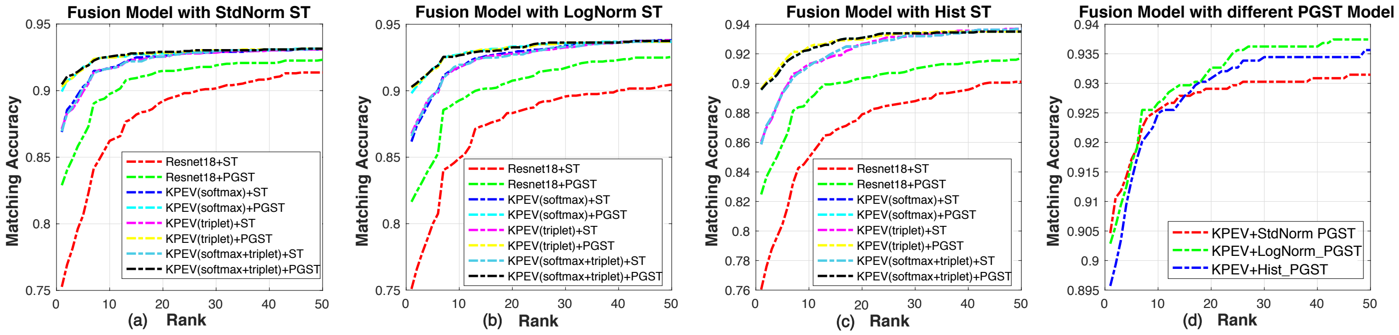


Figure 12. CMC results of comparisons of our fusion model on the VeRi-776 dataset.

Table 8. Comparisons (%) of performance of our KPGST approach on the VeRi-776 dataset. The best results are shown in bold.

Methods	Poses Guided	mAP	R-1	R-5
StdNorm + KPEV	×	60.14	89.01	91.43
	✓	65.38	92.30	93.41
LogNorm + KPEV	×	63.61	88.69	90.96
	✓	**68.73**	**92.35**	**93.92**
STHist + KPEV	×	56.02	87.94	90.24
	✓	59.82	91.78	92.74

Table 9 shows a comparison of our KPGST approach and the state-of-the-art fusion methods. OIFE + STR [44] proposes a lognormal spatio-temporal model for Re-ID. The VAMI + STR [13] and Siamese CNN + STR [45] models utilize the product of the time difference and physical distance as the final spatio-temporal matching score. The approaches of [13,44] both fuse models by adding weighted visual matching scores and visual matching scores. The Siamese CNN + PathLSTM [45] method uses a multi-layer perception (MLP) with two layers to process visual-spatio-temporal path proposals, and it subsequently matches candidate paths by an LSTM network. The KPGST method obtains 4.36%, 6.79% and 21.98% R-1 gains over VAMI + STR [13], Siamese CNN + PathLSTM [45] and OIFE + STR [44], respectively. Such significant improvements can be ascribed to two aspects. Firstly, an effective visual extractor of vehicle features is used that outperforms the methods of [44,45]. Moreover, we apply the pose-guided spatio-temporal algorithm to make the best use of spatio-temporal constraints. The improvements suggest that KPGST achieves better performance by considering spatio-temporal relationships in the vehicle Re-ID task.

Table 9. Comparisons (%) with our KPGST and state-of-the-art Re-ID methods on the VeRi-776 dataset. The best results are shown in bold.

Methods	mAP	R-1	R-5
FACT + STR [18]	27.77	61.44	78.78
OIFE + STR [44]	51.42	68.30	89.70
Siamese CNN + STR [45]	40.26	54.23	74.97
Siamese CNN + PathLSTM [45]	58.27	83.49	90.04
VAMI + STR [13]	61.32	85.92	91.84
KPGST (KPEV + LogNorm PGST)	**68.73**	**92.35**	**93.92**

5. Conclusions

Investigating practical scenarios, we consider the specific characteristics of vehicles and propose a simple but effective model called the KPGST model to solve the vehicle Re-ID problem in urban surveillance systems. The main idea in this paper is to exploit a two-branch framework that includes appearance features' fusion of global and pose-view and spatio-temporal constraints that are both guided by vehicles' poses. During the processing of pose features merged into global appearance features, we exploit keypoints to enhance the accuracy of pose recognition. Furthermore, we explore the mechanism of how the poses guide the spatio-temporal model based on a Bayesian network. Unlike the existing state-of-the-art methods that only utilize appearance features or some methods that analyze the vehicle Re-ID problem with the help of spatio-temporal information, we optimize the appearance features' framework with both global and regional features and develop a more accurate spatio-temporal constraint model. A performance analysis shows that our method for the practical vehicle Re-ID problem is reasonable and insightful.

Author Contributions: Conceptualization, W.H.; methodology, W.H. and X.Z.; software, M.F.; resources, X.Z.; writing—original draft preparation, W.H., X.J.; writing—review and editing, X.Z., W.L., Z.W. and S.S.; funding acquisition, X.Z. and W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Department of Science and Technology, Hubei Provincial People's Government under Grant 2021CFB281 and 2021CFB513.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the anonymous reviewers and the editor for their careful reviews and constructive suggestions to help us improve the quality of this paper.

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

References

1. He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; Feng, J. Foreground-Aware Pyramid Reconstruction for Alignment-Free Occluded Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8449–8458.
2. Wang, X.; Liang, C.; Chen, C.; Chen, J.; Wang, Z.; Han, Z.; Xiao, C. S3D: Scalable Pedestrian Detection via Score Scale Surface Discrimination. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 3332–3344.
3. Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Wang, Z.; Wang, X.; Jiang, J.; Lin, C. Rain-Free and Residue Hand-in-Hand: A Progressive Coupled Network for Real-Time Image Deraining. *IEEE Trans. Image Process.* **2021**, *30*, 7404–7418.
4. Stefanic, P.; Cigale, M.; Jones, A.C.; Knight, L.; Taylor, I.J.; Istrate, C.; Suciu, G.; Ulisses, A.; Stankovski, V.; Taherizadeh, S.; et al. SWITCH workbench: A novel approach for the development and deployment of time-critical microservice-based cloud-native applications. *Future Gener. Comput. Syst.* **2019**, *99*, 197–212.
5. Xu, Z.; Shah, H.S.; Ramachandran, U. Coral-Pie: A Geo-Distributed Edge-compute Solution for Space-Time Vehicle Tracking. In Proceedings of the 21st International Middleware Conference, Delft, The Netherlands 7–11 December 2020; pp. 400–414.
6. Wang, Z.; Jiang, J.; Yu, Y.; Satoh, S. Incremental Re-Identification by Cross-Direction and Cross-Ranking Adaption. *IEEE Trans. Multimed.* **2019**, *21*, 2376–2386.
7. Qian, X.; Fu, Y.; Xiang, T.; Jiang, Y.; Xue, X. Leader-Based Multi-Scale Attention Deep Architecture for Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 371–385.
8. Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; Yang, Y. Salience-Guided Cascaded Suppression Network for Person Re-Identification. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Vision, Seattle, WA, USA, 13–19 June 2020; pp. 3297–3307.
9. Li, H.; Wu, G.; Zheng, W. Combined Depth Space Based Architecture Search for Person Re-Identification. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Vision, Virtual, 19–25 June 2021; pp. 6729–6738.
10. Yang, L.; Luo, P.; Loy, C.C.; Tang, X. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3973–3981.
11. Peri, N.; Khorramshahi, P.; Rambhatla, S.S.; Shenoy, V.; Rawat, S.; Chen, J.; Chellappa, R. Towards Real-Time Systems for Vehicle Re-Identification, Multi-Camera Tracking, and Anomaly Detection. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2648–2657.
12. Zhong, X.; Gu, C.; Ye, M.; Huang, W.; Lin, C. Graph Complemented Latent Representation for Few-shot Image Classification. *IEEE Trans. Multimed.* **2022**. [CrossRef]
13. Zhou, Y.; Shao, L. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6489–6498.
14. Zhu, X.; Luo, Z.; Fu, P.; Ji, X. VOC-RelD: Vehicle Re-identification based on Vehicle-Orientation-Camera. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 2566–2573.
15. Jin, Y.; Li, C.; Li, Y.; Peng, P.; Giannopoulos, G.A. Model Latent Views With Multi-Center Metric Learning for Vehicle Re-Identification. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1919–1931.
16. Teng, S.; Zhang, S.; Huang, Q.; Sebe, N. Multi-View Spatial Attention Embedding for Vehicle Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 816–827.
17. Zhong, X.; Feng, M.; Huang, W.; Wang, Z.; Satoh, S. Poses Guide Spatiotemporal Model for Vehicle Re-identification. In Proceedings of the Springer International Conference on Multimedia Modeling, Thessaloniki, Greece, 8–11 January 2019; pp. 426–439.
18. Liu, X.; Liu, W.; Ma, H.; Fu, H. Large-scale vehicle re-identification in urban surveillance videos. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo, Seattle, WA, USA, 11–15 July 2016; pp. 1–6.

19. Liu, H.; Tian, Y.; Yang, Y.; Pang, L.; Huang, T. Deep relative distance learning: Tell the difference between similar vehicles. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2167–2175.
20. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 501–518.
21. Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; Yang, Y. Pose-Guided Feature Alignment for Occluded Person Re-Identification. In Proceedings of the IEEE/CVF International Conference Compution Vision, Seoul, Korea, 27 October–2 November 2019; pp. 542–551.
22. Ye, M.; Lan, X.; Leng, Q. Modality-aware Collaborative Learning for Visible Thermal Person Re-Identification. In Proceedings of the ACM International Conference Multimedia, Nice, France, 21–25 October 2019; pp. 347–355.
23. Wang, Z.; Jiang, J.; Wu, Y.; Ye, M.; Bai, X.; Satoh, S. Learning Sparse and Identity-Preserved Hidden Attributes for Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 2013–2025.
24. Jia, X.; Zhong, X.; Ye, M.; Liu, W.; Huang, W.; Zhao, S. Patching Your Clothes: Semantic-aware Learning for Cloth-Changed Person Re-Identification. In Proceedings of the International Conference MultiMedia Modeling, Qui Nhon, Vietnam, 5–8 April 2022; pp. 121–133.
25. Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; Lin, C. Grayscale Enhancement Colorization Network for Visible-infrared Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1418–1430.
26. Huang, Z.; Wang, Z.; Tsai, C.; Satoh, S.; Lin, C. DotSCN: Group Re-Identification via Domain-Transferred Single and Couple Representation Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2739–2750. [CrossRef]
27. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021.** [CrossRef]
28. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3960–3969.
29. Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; Hu, J. Pose Transferrable Person Re-Identification. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4099–4108.
30. Zheng, L.; Huang, Y.; Lu, H.; Yang, Y. Pose invariant embedding for deep person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 4500–4509.
31. Zheng, K.; Lan, C.; Zeng, W.; Liu, J.; Zhang, Z.; Zha, Z.J. Pose-Guided Feature Learning with Knowledge Distillation for Occluded Person Re-Identification. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, China, 20–24 October 2021; pp. 4537–4545.
32. Huang, W.; Hu, R.; Liang, C.; Yu, Y.; Wang, Z.; Zhong, X.; Zhang, C. Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations. In Proceedings of the International Conference on Multimedia Modeling, Miami, FL, USA, 4–6 January 2016; pp. 174–186.
33. Lv, J.; Chen, W.; Li, Q.; Yang, C. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7948–7956.
34. Jiang, N.; Bai, S.; Xu, Y.; Xing, C.; Zhou, Z.; Wu, W. Online inter-camera trajectory association exploiting person re-identification and camera topology. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 1457–1465.
35. Feris, R.S.; Siddiquie, B.; Petterson, J.; Zhai, Y.; Datta, A.; Brown, L.M.; Pankanti, S. Large-scale vehicle detection, indexing, and search in urban surveillance videos. *IEEE Trans. Multimed.* **2011**, *14*, 28–42. [CrossRef]
36. 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/CVF International Conference Compution Vision, Venice, Italy, 22–29 October 2017; pp. 562–570.
37. Chen, X.; Sui, H.; Fang, J.; Feng, W.; Zhou, M. Vehicle Re-Identification Using Distance-Based Global and Partial Multi-Regional Feature Learning. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1276–1286. [CrossRef]
38. Zheng, Z.; Ruan, T.; Wei, Y.; Yang, Y.; Mei, T. VehicleNet: Learning Robust Visual Representation for Vehicle Re-Identification. *IEEE Trans. Multimed.* **2021**, *23*, 2683–2693. [CrossRef]
39. Liu, X.; Liu, W.; Mei, T.; Ma, H. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 869–884.
40. Lou, Y.; Bai, Y.; Liu, J.; Wang, S.; Duan, L. VERI-Wild: A Large Dataset and a New Method for Vehicle Re-Identification in the Wild. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3235–3243.
41. Zapletal, D.; Herout, A. Vehicle re-identification for automatic video traffic surveillance. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition Workshops, Las Vegas, NV, USA, 27–30 June 2016; pp. 1568–1574.
42. Sochor, J.; Herout, A.; Havel, J. Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3006–3015.
43. Zhou, Y.; Liu, L.; Shao, L. Vehicle re-identification by deep hidden multi-view inference. *IEEE Trans. Image Process.* **2018**, *27*, 3275–3287. [CrossRef]

44. 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/CVF International Conference Compution Vision, Venice, Italy, 22–29 October 2017; pp. 379–387.
45. 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/CVF International Conference Compution Vision, Venice, Italy, 22–29 October 2017; pp. 1900–1909.
46. Liu, X.; Liu, W.; Mei, T.; Ma, H. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Trans. Multimed.* **2017**, *20*, 645–658. [CrossRef]
47. Li, P.; Li, G.; Yan, Z.; Li, Y.; Lu, M.; Xu, P.; Gu, Y.; Bai, B.; Zhang, Y.; Chuxing, D. Spatio-temporal Consistency and Hierarchical Matching for Multi-Target Multi-Camera Vehicle Tracking. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 222–230.
48. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference Compution Vision Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
51. Tieleman, T.; Hinton, G. Divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* **2012**, *4*, 26–31.
52. Zheng, A.; Dong, J.; Lin, X.; Liu, L.; Jiang, B.; Luo, B. Visual Cognition-Inspired Multi-View Vehicle Re-Identification via Laplacian-Regularized Correlative Sparse Ranking. *Cogn. Comput.* **2021**, *13*, 859–872. [CrossRef]
53. 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/CVF Conference Compution Vision Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
54. Zhou, Y.; Shao, L. Cross-view gan based vehicle generation for re-identification. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017.
55. Guo, H.; Zhao, C.; Liu, Z.; Wang, J.; Lu, H. Learning Coarse-to-Fine Structured Feature Embedding for Vehicle Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 6853–6860.
56. Hou, J.; Zeng, H.; Cai, L.; Zhu, J.; Chen, J.; Ma, K. Multi-label learning with multi-label smoothing regularization for vehicle re-identification. *Neurocomputing* **2019**, *345*, 15–22. [CrossRef]
57. Lou, Y.; Bai, Y.; Liu, J.; Wang, S.; Duan, L. Embedding Adversarial Learning for Vehicle Re-Identification. *IEEE Trans. Image Process.* **2019**, *28*, 3794–3807. [CrossRef] [PubMed]
58. Zhu, R.; Fang, J.; Li, S.; Wang, Q.; Xu, H.; Xue, J.; Yu, H. Vehicle re-identification in tunnel scenes via synergistically cascade forests. *Neurocomputing* **2020**, *381*, 227–239. [CrossRef]
59. Tumrani, S.; Deng, Z.; Lin, H.; Shao, J. Partial attention and multi-attribute learning for vehicle re-identification. *Pattern Recognit. Lett.* **2020**, *138*, 290–297. [CrossRef]
60. Roman-Jimenez, G.; Guyot, P.; Malon, T.; Chambon, S.; Charvillat, V.; Crouzil, A.; Péninou, A.; Pinquier, J.; Sèdes, F.; Sénac, C. Improving vehicle re-identification using CNN latent spaces: Metrics comparison and track-to-track extension. *IET Comput. Vis.* **2021**, *15*, 85–98. [CrossRef]
61. Wang, G.; Lai, J.; Huang, P.; Xie, X. Spatial-Temporal Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8933–8940.