sensors

MDPI

Article

Traffic Intersection Re-Identification Using Monocular Camera Sensors

Lu Xiong, Zhenwen Deng, Yuyao Huang, Weixin Du, Xiaolong Zhao, Chengyu Lu and Wei Tian *

Institute of Intelligent Vehicles, School of Automotive Studies, Tongji University, Shanghai 201804, China; xiong_lu@tongji.edu.cn (L.X.); dengzhenwen@tongji.edu.cn (Z.D.); huangyuyao@tongji.edu.cn (Y.H.); 1733345@tongji.edu.cn (W.D.); 1651860@tongji.edu.cn (X.Z.); luchengyu@tongji.edu.cn (C.L.)

* Correspondence: tian_wei@tongji.edu.cn

Received: 14 October 2020; Accepted: 11 November 2020; Published: 14 November 2020

Abstract: Perception of road structures especially the traffic intersections by visual sensors is an essential task for automated driving. However, compared with intersection detection or visual place recognition, intersection re-identification (intersection re-ID) strongly affects driving behavior decisions with given routes, yet has long been neglected by researchers. This paper strives to explore intersection re-ID by a monocular camera sensor. We propose a Hybrid Double-Level re-identification approach which exploits two branches of Deep Convolutional Neural Network to accomplish multi-task including classification of intersection and its fine attributes, and global localization in topological maps. Furthermore, we propose a mixed loss training for the network to learn the similarity of two intersection images. As no public datasets are available for the intersection re-ID task, based on the work of RobotCar, we propose a new dataset with carefully-labeled intersection attributes, which is called "RobotCar Intersection" and covers more than 30,000 images of eight intersections in different seasons and day time. Additionally, we provide another dataset, called "Campus Intersection" consisting of panoramic images of eight intersections in a university campus to verify our updating strategy of topology map. Experimental results demonstrate that our proposed approach can achieve promising results in re-ID of both coarse road intersections and its global pose, and is well suited for updating and completion of topological maps.

Keywords: monocular camera sensor; deep learning; intersection dataset; intersection re-identification; image matching

1. Introduction

Road intersection re-identification is an essential ability for human drivers. To arrive at the destination, a global driving route has to be planned. By re-identifying intersections on the planned route, drivers can decide the proper behaviors at different phases when traversing intersections. For instance, before entering the intersection, the vehicle should reduce the speed and select the correct lane according to the planned driving direction. Within the intersection, vehicles should take a correct motion and be alerted to environmental objects. When leaving the intersection, the vehicle recovers the instructed speed and continues to drive along the current road.

The same behavior decision at intersection also applies for automated driving. With the recurrence of deep learning, the road intersection re-ID task can be accomplished by leveraging image features (e.g., trees, infrastructure elements, road markings, traffic signs) and deep convolutional neural networks

(DCNNs). By introducing the prior information of a digital map, the intersection re-ID can also provide a rough positioning of the vehicle when the intersection is detected and identified, as shown in Figure 1. Compared to simultaneous localization and mapping (SLAM), visual place re-identification [1] is unobligated to save a vast amount of point clouds or visual feature points, and thus is suitable for sparse localization. (In this work, we refer to the localization w.r.t. the topological map, which focuses on the pose relative to the intersection and the driving direction of vehicle.) In this paper, we argue that the required information in driving behavior decision for intelligent vehicles (IVs) such as fine intersection attributes and sparse positioning w.r.t. the intersection topological map can be achieved under a rational road intersection re-ID approach.

Due to the importance of road intersections, recent studies have been focused on representation of intersections by exploiting recurrent sequences [2], producing HD maps [3], analysis of driving behavior at intersections by tracking strategies [4], or planning horizons [5]. However, this research either barely considers the intersection into larger designated areas or insufficiently uses vehicular sensors to re-identify the intersection in conjunction with the topological map.

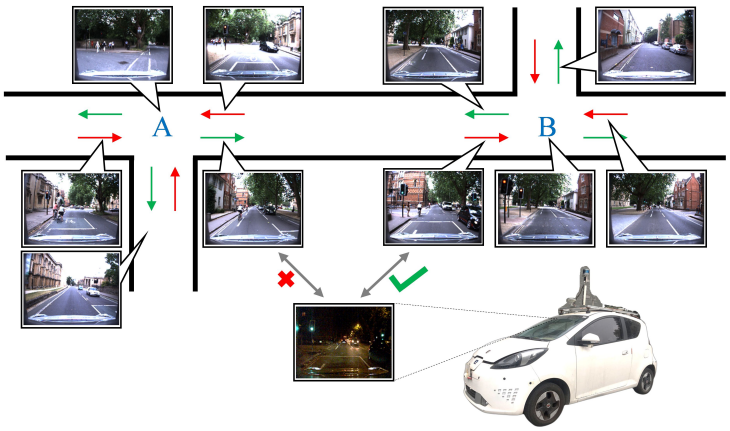

Figure 1. Traffic intersection re-identification for sparse localization of IV. Camera images captured at different positions in each intersection area are mapped and used to extract gallery features, which match the query image to achieve the sparse localization. The gallery images associated with the locations can make up a topological map. Image samples are from [6].

Monocular cameras are often considered as sensors for perception task solutions in the automotive industry. Compared with other vehicle-mounted sensors, such as LiDAR and radar, cameras can provide images with rich environmental semantic information and have a relative low production cost and high installation flexibility. Meanwhile, the training data and deep neural network structure (open source) from academic communities further improve the performance of image processing. Although machine learning based image processing has achieved significant progress in recent years, using a monocular camera sensor for intersection re-ID is fundamentally challenging for modern deep neural networks. As a sensor deficiency, the monocular camera has no depth information, resulting in the fact that the intersection geometry cannot be fully used. The complex traffic flow at the intersection can also interfere with the extraction of static environmental features. Currently, there is no framework designed for automated driving to complete multiple tasks including intersection classification, intersection recognition, and intersection re-ID with a monocular camera sensor. To the best of our knowledge, there are neither

public datasets available that address the evaluation of road intersection re-identification in the traffic environment, which has further slowed down the related research.

To address the issues above, in this paper, we firstly introduce the intersection re-ID task formulation, which is ready to associate proper driving behavior decision under a given intersection topological structure. Secondly, we design the Hybrid Double-Level traffic intersection re-ID network as an end-to-end multi-task framework for the classification of intersection and its fine attributes, and the global localization in topological maps by monocular camera images. Thirdly, we present two intersection datasets: the RobotCar Intersection dataset, which is based on the prior work of [6] and covers more than 30,000 finely labeled intersection images in different time dimensions, and the Campus Intersection dataset, which contains 3340 panoramic images captured at eight intersections in the Tongji University campus. Finally, we validate the performance of our approach in the re-identification of road intersections and its global pose, and the performance of updating strategy for topology map on proposed datasets. The fully labeled datasets and corresponding PyTorch code of our framework will be online available (https://github.com/tjiiv-cprg/Road-Intersection-Re-ID).

2. Related Works

Image-based traffic intersection re-ID is the re-identification of intersections by establishing correspondence between intersection images captured by camera sensors. In a closed area, as shown in Figure 1, the intelligent vehicle firstly traverses all intersections, and saves intersection images with pose labels in the gallery. When arriving at the intersection again, the global location of vehicle and its route can be determined by query-gallery image matching. For example, in Autonomous Valet Parking [7], driverless cars will automatically drive to the corresponding intersection for pick-up service by image based intersection re-ID. For driving in an open area, the intersection re-ID approach is inherently able to recognize new intersections (e.g., by a missing match). Thus, the stored topological map can be further extended with a rational updating strategy.

Despite the importance, intersections are considered as the most complex part of the traffic road, mainly because vehicles from different driving directions will gather there and drive to their respective destinations, thus forming a complex traffic flow. As the hubs of transportation networks, intersections are very valuable for research. However, there are few studies on intersection re-identification based on vehicular onboard visual sensors, and related datasets are also scarce. Thus, in this section, we mainly summarize research from aspects of intersection detection, visual place recognition, and visual place re-ID datasets, which are most relevant to our topic.

2.1. Intersection Detection

Detection of intersections has been an interest point for scholars since the last decade. Kushner et al. [8] firstly presented road intersection detection methods respectively based on monocular camera and range scanner. They used road boundaries and height profiles to determine the best match by a minimum goodness-of-fit measure. As the superiority of deep neural networks emerges, Bhatt et al. [2] proposed an end-to-end Long-term Recursive Convolutional Network (LRCN) and considered intersection detection as a binary classification task on the frame sequence. In the work of Bhattacharyya et al. [9], a method of spatial-temporal analysis of traffic conditions at urban intersections based on stereo vision and 3D digital maps was introduced. The depth cues of each pixel effectively provide more accurate intersection detection. Habermann et al. [10] proposed to detect road intersections based on 3D point clouds. Three classifiers, including support vector machine (SVM), AdaBoost, and artificial neural network (ANN), are used for classification of intersections.

In non-visual sensor based intersection detection, Xie et al. [11] detected intersections indirectly from common sub-tracks shared by different global navigation satellite system (GNSS) traces. Local distance matrices, image processing techniques, and Kernel Density Estimations are used to identify the intersection. Based on remote radar images, Cheng et al. [12] detected different types of road intersection in two stages: global area detection and local shape recognition. However, intersection detection by non-visual sensor typically takes a long observation time.

2.2. Visual Place Recognition

A cognitive map [13] interprets a distinctive place according to the sensory information such as place signature and place description. To distinguish between images, hand-crafted local feature descriptors such as scale-invariant feature transform (SIFT) [14], speed-up robust features (SURF) [15], and features from accelerated segment test (FAST) [16] have been used in image matching. With hundreds of local features extracted from images, the bag-of-words model [17] collects all local features into a vocabulary and the frequency of each word can be used as distinctiveness criterion for similarity evaluation between two images. Global place descriptors based on color histograms [18] and Principal Component Analysis (PCA) [19] are also used in early localization systems. Presently, the Gist [20], as a 512-dimensional feature, is extracted from an image using Gabor filters at different orientations and different frequencies. This approach is widely used in global place recognition [21].

In addition to hand-crafted approaches, the neural network is also introduced to solve signature verification as an image classification or matching problem. Chen et al. [22] trained a simple and lightweight ConvNet to classify each place image into place labels. The Siamese network [23], consisting of twin networks to search matched image pair, is optimized for place recognition in [24]. As the weights of the twin networks are shared, output features of two distinct images are computed in the same metric which is essential to similarity calculation. Such a structure is widely used in re-identification of persons [25] and vehicles [26]. However, differing from common approaches for pedestrian or vehicle re-ID, the task for intersection re-ID focuses on the complex road intersections, which involve information from a wide traffic scene. The intersection re-identification network needs to be able to automatically extract static background features, instead of focusing on the foreground targets, e.g., pedestrians and vehicles. These existing methods for pedestrian or vehicle re-ID cannot be directly applied to intersection re-identification.

2.3. Visual Place Re-ID Datasets

One of the related datasets is the Alderley Day/Night Dataset [27], which was captured along a fixed route in two different conditions: sunny daytime and rainy nighttime. The Ford Campus Vision and Lidar Dataset [28] recorded both time-registered laser points and images, with the vehicle driving along several large and small-scale loops. The images in the FABMAP Dataset [29] were collected from cameras mounted on both sides of a robot pan-tilt. The image collection was triggered every 1.5 m (on the basis of odometry). The St Lucia Multiple Times of Day [30] collected visual data with a forward facing webcam attached to the roof of a car. In total, ten subdatasets were captured at five different periods of daytime in a two-week interval. The CMU Dataset [31] recorded images in sequences during one year from the urban, suburban, and park by two front-facing cameras, pointing to the left/right side at approximately 45 degrees. However, the mentioned datasets are collected along trajectories which did not contain traffic intersections, or their intersection images lack for environmental diversity.

With an improved diversity, the Oxford RobotCar Dataset [6] contains 20 million images collected from six vehicular cameras along a fixed route (including more than 25 obvious intersections) during a period of more than one year (per two weeks on average). All weather conditions, including sunny,

overcast, rainy, or snowy weather in day and night, enable researchers to investigate long-term place re-identification in real-world and dynamic urban environments. However, this dataset can not be directly used in intersection re-ID due to the lack of detailed labels of traffic intersection.

Although there are many research and datasets related to intersection detection and visual place recognition, due to the essential difference, there is little academic work to address the re-identification of intersection and its attributes. In order to bridge the gap, we present the Hybrid Double-Level network and two corresponding datasets for this task.

3. Traffic Intersection Re-Identification Approach

3.1. Task Formulation

The visual sensor based intersection re-ID task in this paper is defined as the multi-task including classification of intersection and its fine attributes, and the determination of global vehicle pose. While the intersection identity can be used in coarse localization, its attribute identity denotes the interaction between the ego-car and the intersection, and the global identity determines the global pose in the map (shown in Figure 2). Furthermore, new intersections should also be recognized when driving in an open area. These are the main characteristics different from conventional tasks of intersection detection and visual place recognition.

Mathematically, we define the set of gallery images from a closed area as $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$. For each gallery image (actually the image feature) $\mathbf{x}_i$ in this set, we map it to an ID tuple by a function f interpreted as

$$f : \mathbf{x}_i \rightarrow (u_i, v_i, w_i) \in \mathbb{U} \times \mathbb{V} \times \mathbb{W}, \tag{1}$$

where (u_i, v_i, w_i) represents the tuple of intersection ID, attribute ID and global ID of $\mathbf{x}_i$. In addition, $\mathbb{U} \times \mathbb{V} \times \mathbb{W}$ represents the space of all valid intersection ID tuples. For a query image $\tilde{\mathbf{x}}$, it can also be mapped to an ID tuple $(\tilde{u}, \tilde{v}, \tilde{w})$ by function f. Thus, the intersection re-ID task in a closed area can be formulated as searching the optimal mapping function f by minimizing the classification error $\epsilon((\tilde{u}, \tilde{v}, \tilde{w}), (u_i, v_i, w_i))$ with the assumption of groundtruth ID tuple as (u_i, v_i, w_i). For an open area, if the query image $\tilde{\mathbf{x}}$ represents a new intersection, the function f should be able to recognize it and assign the image with a new ID tuple $(u_\perp, v_i, w_\perp)$, interpreted as:

$$f(\tilde{\mathbf{x}}) = \begin{cases} (u_i, v_i, w_i), & if\ \tilde{\mathbf{x}} = \mathbf{x}_i \\ (u_\perp, v_i, w_\perp), & otherwise \end{cases}. \tag{2}$$

Note that the attribute space $\mathbb{V}$ is unchanged in the task. For updating the topological map, the new intersection image along with its ID tuple should be added to the database, respectively yielding the extended gallery set $\mathbf{X}_{ex} = \tilde{\mathbf{x}} \cup \mathbf{X}$ and extended ID tuple space $\mathbb{U}_{ex} \times \mathbb{V} \times \mathbb{W}_{ex} = (u_\perp, v_i, w_\perp) \cup (\mathbb{U} \times \mathbb{V} \times \mathbb{W})$.

In this work, we interpret the function f with a DCNN architecture, which is introduced in Section 3.2. The mixed loss training of our network is explained in Section 3.3. In addition, the updating strategy of topology map is proposed in Section 3.4.

3.2. HDL Network Architecture

We propose a Hybrid Double-Level (HDL) network for traffic intersection re-identification to relieve the heavy burden from saving a vast amount of data in normal image matching. The proposed framework aims to learn a DCNN that can determine coarse classes like the intersection ID and attribute ID. The global ID is taken as a fine class, which is predicted by the similarity score between the query image feature and gallery image features.

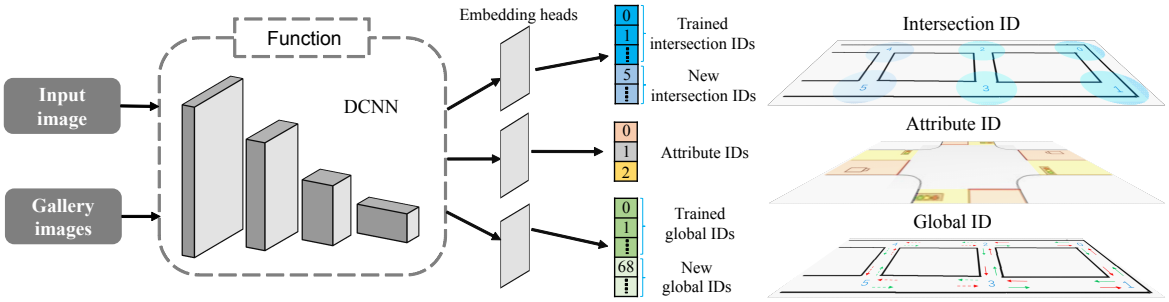

Figure 2. Multi-task of traffic intersection re-identification, which consists of determining the intersection ID, attribute ID, and global ID. The attribute ID denotes the interaction between ego-car and intersection, with 0 indicating at the entrance of the intersection, 1 for inside the intersection, and 2 for at the intersection exit. The intersection IDs and global IDs respectively represent different intersection areas and global poses (assigned with information of location and driving direction). Both of them can be partitioned into existing gallery IDs and new IDs.

For the HDL network structure, inspired by the work [26], we exploit two DCNN branches to accomplish different tasks in traffic intersection re-identification, as depicted in Figure 3. In the proposed network, we first map the input image of three color channels to a canonical dimension using bicubic sampling and normalization. This normalizes the image scale while maintaining its aspect-ratio. The pre-processed image is represented as $I \in \mathbb{R}^{H \times W \times 3}$ with the height H and width W. The proposed network is composed of one block of shared layers and two subsequent branches. The shared layers totally consist of seven convolutional filters of size 3×3 with ReLU activation layer, and three MaxPooling layers. The input image is firstly fed into shared layers to generate features for following two branches. The two DCNN branches are mainly for feature extraction in different scales: coarse and fine features. Each feature extractor contains convolutional blocks and fully connected blocks which are integrated with ReLU and dropout operation. We impose supervision on the coarse features to produce two embedding heads [32] for the classification of intersection and its attributes. The fine features with the same size are derived from the fine feature extractor and concatenated with coarse features as the input of a block of fusion layers. Note that the weights of both feature extractors are not the same. The fusion layers aim to fuse image information of different scales to represent global features of the intersection using a fully connected network. We perform the final embedding head to determine the global pose. The generation of coarse feature and global feature are respectively interpreted as

$$\mathbf{F}_c(I) = \sigma_c(\mathrm{Conv}_c(\sigma_s(\mathrm{Conv}_s(I)))), \tag{3}$$

$$\mathbf{F}_f(I) = \sigma_f(\mathrm{Conv}_f(\sigma_s(\mathrm{Conv}_s(I)))), \tag{4}$$

$$\mathbf{F}_g(I) = \sigma_g(\mathrm{Conv}_g(\mathbf{F}_c(I) \oplus \mathbf{F}_f(I))), \tag{5}$$

where $\mathbf{F}_{\{c,f,g\}}$ denotes generated feature from corresponding branches. The subscript c indicates coarse feature, f is for fine feature, and g denotes global feature. $\mathrm{Conv}_{\{c,f,g\}}$ represents the corresponding convolutional layers and $\sigma_{\{c,f,g\}}$ represents corresponding activation function at each stage. $\oplus$ is the element-wise concatenation operation.

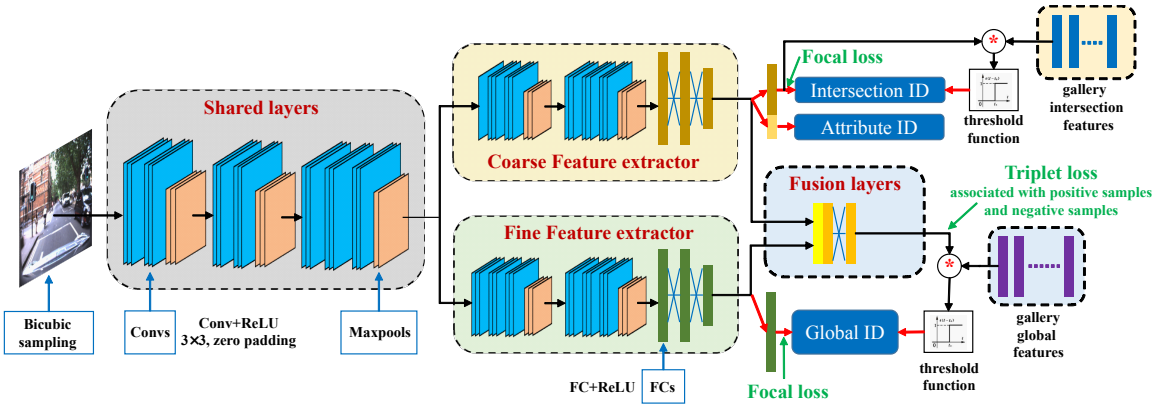

Figure 3. The proposed hybrid double-level network consists of two branches. The coarse features are concatenated into the mainstream to provide different level information for the generation of final feature. The similarity between the query feature and gallery features determines the global pose by threshold function. A coupled cluster losses help to train the network with three steps of forward propagation.

3.3. Mixed Loss Training

Since multiple tasks should be finished in our network, we strive to extract similar static background features from images of the same class as well as to distinguish the features between subclass images within one super-class. Such a task to recognize the subtle differences between highly similar images is called as the fine-grained image recognition. Moreover, to address the imbalance of training samples, we fuse a couple of loss functions to accelerate the training convergence and realize the multi-task.

In our network, we firstly introduce the Focal loss [33] based on the cross-entropy loss function of binary classification, which makes the network optimization less influenced by the imbalance of training samples. Both loss functions are interpreted as below:

$$\ell_{cross-entropy} = \begin{cases} -\log \tilde{y}, & if \quad y = 1 \\ -\log(1-\tilde{y}), & if \quad y = 0 \end{cases}, \tag{6}$$

$$\ell_{focal} = \begin{cases} -\alpha(1-\tilde{y})^{\gamma} \log \tilde{y}, & if \quad y = 1 \\ -(1-\alpha)\tilde{y}^{\gamma} \log(1-\tilde{y}), & if \quad y = 0 \end{cases}, \tag{7}$$

where $\tilde{y}$ is the output of activation function. y is ground truth label for binary classification. α and γ are hyperparameters for adjusting the imbalance of training samples. Compared with the standard cross entropy loss, the focal loss introduces an influence factor $\gamma > 0$ to reduce the loss proportion of easy samples, as shown in Equations (6) and (7), so that the network pays more attention to samples which are difficult to be classified. Here, we use the focal loss in each branch to evaluate the identity between the instances represented by both images.

Additionally, we employ the triplet loss [32] in our network training. Different from the classification losses, the triplet loss enables the network to project raw images into an embedding space in which images of different categories can be separated by a certain margin. By this loss, the dimension of network output layer can be smaller than the number of categories which results in output dimension reduction. The triplet loss can also be applied to the problem of fine-grained identification, especially in intersection re-identification. To better train our proposed network, we use the triplet loss to evaluate the subtle difference among intersection images of different global IDs. Moreover, we use an online training strategy to acquire negative samples from the training batch. Different from the traditional L1 or L2 distance, we use the cosine distance to measure the embedding distance, as shown below:

$$\ell_{triplet} = \sum_{n} \max\left\{f_{sim}(I^q, I^n) + M - f_{sim}(I^q, I^p),\, 0\right\}, \tag{8}$$

$$f_{sim}(I^q, I^n) = \frac{\mathbf{F}_g(I^q) * \mathbf{F}_g(I^n)}{\left\|\mathbf{F}_g(I^q)\right\| \left\|\mathbf{F}_g(I^n)\right\|}, \tag{9}$$

where f_{sim} is similarity function. I^q is the input tensor of query image. I^p is a positive sample with the same global ID. I^n is a negative sample with a different global ID. $*$ is the inner product. M represents the expected margin between inter-classes images and n represents the image number of a training batch. $\|\cdot\|$ denotes the L2 norm of feature vector.

3.4. Updating Strategy of Topology Map

When driving in an open area, vehicles will often run into new intersections. How to add information of new intersections into the preset topological map is essential for autonomous vehicles driving in a real traffic environment. In this regard, we propose an updating strategy for the intersection topology map, as shown in Figure 4. This strategy is decomposed into two steps. In the first step, camera images are processed by conventional intersection detection approaches (as introduced in Section 2) to determine whether it represents an intersection or not. If it is a straight road, the vehicle continues to drive along the planned path. When an intersection is detected, the system determines whether it is currently at a new intersection or at an existed intersection by comparing the similarity of query image feature with gallery image features.

For an image of new intersection, a new intersection ID and a new global ID will be assigned to this image. For an existed intersection, only the global ID will be updated if the current lane or direction is not matched with the gallery. Furthermore, the new image along with its labels will be added into the database. Note that, since the interaction between intersection and vehicle does not depend on the intersection recognition result, only the label space of intersection ID and global ID can be changed.

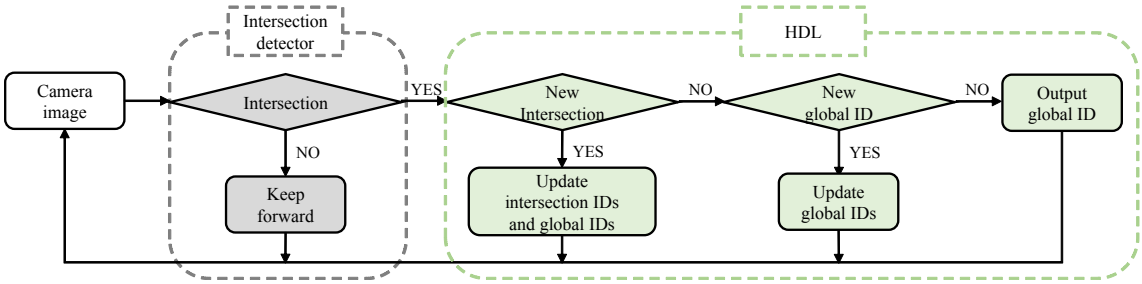

Figure 4. Updating strategy of intersection topological map. The system first perceives if the vehicle encounters an intersection. For a positive detection, the system checks if the image represents a new intersection or not. For a new intersection, the image will be assigned with new intersection and global ID. Both the image and its ID labels will be added into the database.

4. Proposed Intersection Datasets

4.1. RobotCar Intersection Dataset

Our proposed RobotCar Intersection dataset is based on the prior work RobotCar [6], which is captured in the central area of Oxford by driving vehicle several times along a fixed route. In our dataset, to simulate the habits of human drivers, we choose images of the front-view camera (Bumblebee XB3, CCD, only 66° Horizontal Field of View,) for intersection re-identification. Images of eight road intersections in a

closed region (over 30,000 images with a resolution of 1280 × 960) are selected to build our dataset, as shown in Figure 5a. Inheriting the characteristics of the original RobotCar dataset, the selected images are with environment conditions of different seasons (from May 2014 to November 2015), time (day/night/dusk), weather (sunny/overcast/cloudy/rainy/snowy), etc. Detailed statistics about these images are shown in Figure 5b,c.

Figure 5. Dataset information. (**a**) the selected intersections (marked with yellow number) on the map [6]. (**b**) data distribution at different time. (**c**) Data distribution under different weather.

To add refined attribute labels of intersection to the new dataset, firstly, we determine different traversing phases w.r.t. the relative location between the vehicle and the intersection. The entering phase is determined when the vehicle is located in front of the diversion/stop line at the entrance road of the intersection; the crossing phase is defined when the vehicle is located within the intersection, i.e., between the stop lines of the entrance and the exit road; The leaving phase is determined when the vehicle enters the exit road and completely leaves the intersection. These phases and their associated moving directions are considered as intersection attributes, which are along with the intersection ID assigned to corresponding images, as shown in Figure 6.

Furthermore, we assign each image with an additional global traversing pose ID for better recording the traversed route of the vehicle. Thus, each intersection image is labeled with three kinds of IDs. As shown in Figure 6, the intersection ID is only to identify the intersection location (0–7). The attribute ID of the intersection indicates the interactive relationship between the vehicle and the intersection: the label for entering the intersection is 0, the intersection crossing is labeled as 1, and the intersection exiting is denoted as 2. The global ID represents that the vehicle is at the designated location with designated driving direction. The extracted images are stored according to recording dates, and marked with the weather information. Corresponding labels are also stored in a similar way.

From the examples in Figure 6, it can be seen that the image texture information is the most abundant in cloudy days. The images in sunny weather are easily affected by exposure and halo. The earth ground in images are obviously different in rain and snow weather. The texture information of night images is less than that in the day. All of these facts make the intersection re-ID task on this new dataset challenging.

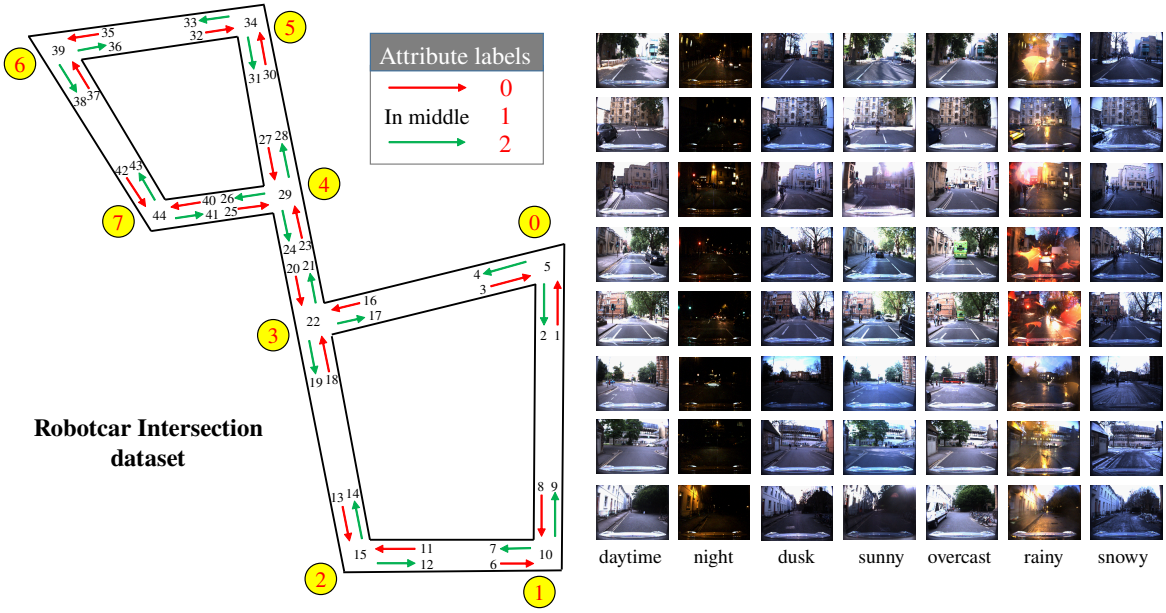

Figure 6. Examples of intersection labels. We label each image with an intersection ID, an attribute ID and a global ID. The ID arrangement is determined by the designated driving route in clockwise order. Image samples in different weathers are from [6].

4.2. Campus Intersection Dataset

The panoramic image has a wide perception area in the surrounding environment, which can also be used for traffic intersection re-identification. In order to explore the performance of our proposed framework on different camera models, we provide the "Campus Intersection" dataset which consists of 3340 spherical panoramic images from eight intersections in the Jiading campus of Tongji University, as shown in Figure 7. The images are captured by a panoramic camera setup introduced in work [34] with a resolution of 1800 × 900. The panoramic camera is formed by two fisheye cameras (S-YUE-U801, CMOS, 210° Horizontal Field of View), which are installed back-to-back in this case.

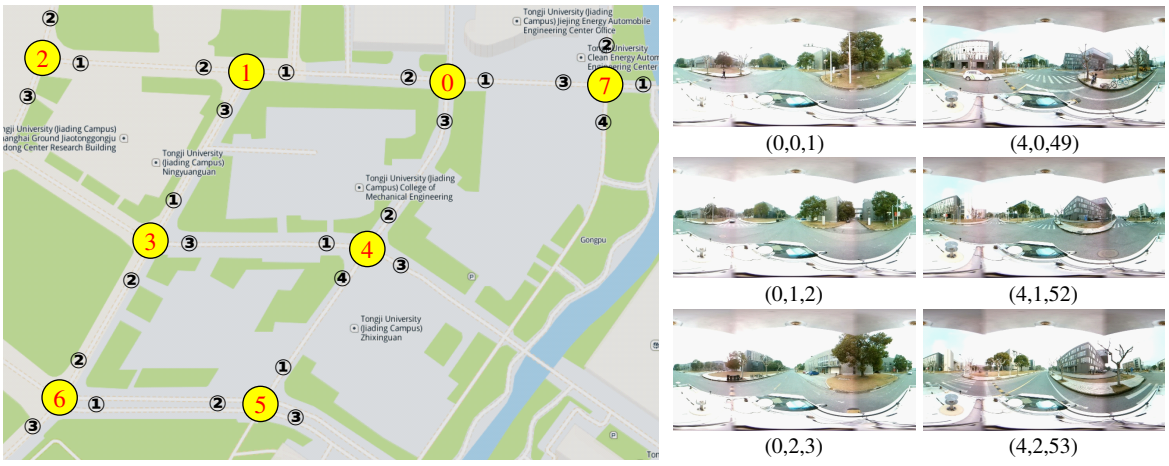

Figure 7. Selected campus intersections and labeled examples. We label intersection IDs according to the driving route and the global IDs in designed order (e.g., clockwise). Each image is labeled with a tuple of (intersection ID, attribute ID, global ID).

We label the Campus Intersection images with the same method as previously introduced. Intersection IDs are determined according to the driving route, and the global IDs are labeled with designed order (e.g., in clockwise order). Note that the intersection ID and attribute ID cannot determine the global position of the vehicle on the topological map, while only the global ID corresponds to the global pose of the vehicle.

5. Experiments

In this section, we evaluate the performance of proposed HDL network architecture on both the RobotCar Intersection and the Campus Intersection datasets, with the motivation of exploring its robustness against different camera models. The experimental results are evaluated by using the prevailing metric of accuracy, margin, precision and recall value, as shown below:

$$\text{Accuracy} = \frac{\sum_N(t_i)}{N}, \quad t_i = \begin{cases} 1, & if\ f_{sim}(I_i^q, I_i^p) > M + f_{sim}(I_i^q, I_i^n) \\ 0, & otherwise \end{cases}, \tag{10}$$

$$\text{Margin} = \frac{\sum_N(f_{sim}(I_i^q, I_i^p) - f_{sim}(I_i^q, I_i^n))}{N}, \tag{11}$$

$$\text{Precision} = TP/(TP + FP), \tag{12}$$

$$\text{Recall} = TP/(TP + FN), \tag{13}$$

where TP, FP, and FN respectively denote the number of true positives, false positives, and false negatives. N represents the number of query images. If the similarity between query image and gallery images is larger than a pre-defined threshold, the global ID of the most similar gallery image is assigned to the query image.

5.1. Training Configuration

For training, the weights of our proposed network are initialized by the pre-trained layers of a VGG16 [35] network. In detail, we use the first seven convolutional layers of the pre-trained VGG16 network to initialize the shared layers. The initial weights of each feature extractor are from the remaining layers of the pre-trained network. Weights of the fusion layer are initialized by a uniform distribution $\mathcal{U}(-\sqrt{1/k}, \sqrt{1/k})$, where k is the input tensor size. Additionally, our proposed model is trained with respect to a combination of detection loss and embedding loss, as follows:

$$\mathcal{L}_{Total} = \lambda_c \ell_c^F + \lambda_f \ell_f^F + \lambda_g^F \ell_g^F + \lambda_g^T \ell_g^T, \tag{14}$$

where ℓ_c^F, ℓ_f^F and ℓ_g^F respectively denote the Focal loss of coarse, fine, and global features. ℓ_g^T denotes the triplet loss of global features. The weights λ_c, λ_f, λ_g^F and λ_g^T for each loss are empirically set to 0.5, 0.5, 0.5, 1, respectively.

In the training process we use an online triplet mining method to select the most dissimilar positive sample and the most similar negative sample in the embedding space from the training batch. All networks are trained using the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001, a weight decay of 1×10^{-8} and a momentum of 0.9. The Focal loss is adopted with parameters $\alpha = 0.25$ and $\gamma = 2$ (under such parameters the focal loss can work best [33]). The batch size is set to 4, which means we need to feed $4 \times 5 = 20$ images in each training iteration (which is a compromise between network convergence speed and GPU computing load). All networks in our experiments are implemented by PyTorch. The evaluation environment is with an Nvidia GTX 1080Ti GPU, an Intel Xeon E5-2667 CPU of 3.2 GHz and a memory of 16 GB.

5.2. Data Augmentation

To enhance the generalization ability of trained networks especially on the illumination change, vehicle sway, camera extrinsic and intrinsic parameter change, we deploy data augmentation methods including random illumination, affine transform, and perspective transform. Since vehicles may also drive in extreme meteorological environments such as the rain or haze, data augmentation by adding rain and fog in images will improve the network performance under the extreme weather. Here, we introduce a nonlinear raindrop model [36], as shown below

$$I_r = \beta_r I_M + (1 - I_M) \odot I_c, \tag{15}$$

where I_r is a rainy image. I_c is the clean background image. I_M is the rain map and β_r is the brightness of raindrops. $\odot$ represents channel-wise operation between I_M and I_c. In order to simulate natural raindrops, we randomly set the raindrop falling direction within the range of $\pm 30°$, and the transparency of raindrops from 0.7 to 1.0, as shown in the first row of Figure 8.

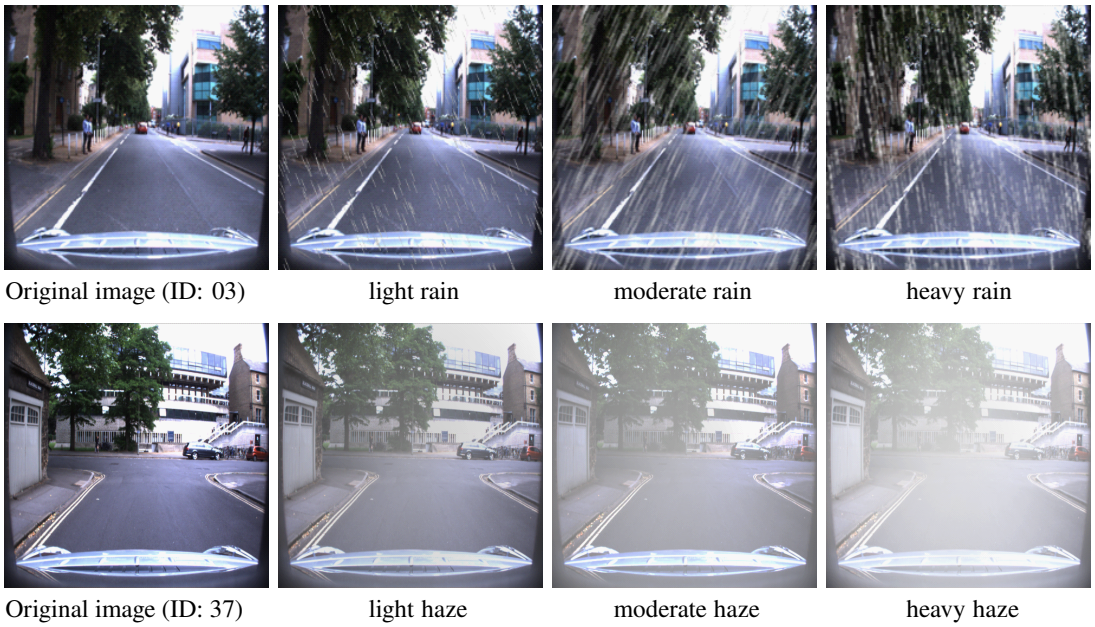

Figure 8. Intersection image with data augmentation by a nonlinear raining model (**first row**) and haze model (**second row**). Images samples are from [6].

For haze simulation, without the depth information, it is tricky to generate haze with only a single image. Here, we follow the classical haze generation [37] of atmospheric scattering model as

$$I(x) = J(x)t(x) + \beta_h A(1 - t(x)), \tag{16}$$

where $I(x)$ is the observed hazy image value at pixel x. $J(x)$ is the haze-free scene radiance to be recovered. β_h is the brightness of the haze. A denotes the white atmospheric light. $t(x)$ is the transmission matrix which is set with nonlinear proportion of pixel distance between each pixel and vanishing point. The density and brightness of haze is adjusted smoothly from 0.7 to 1.0, as shown in in the second row of Figure 8.

5.3. Evaluation on RobotCar Intersection

We divide the RobotCar Intersection dataset into two subsets according to the recording date. The first 60% of the total dataset, containing 21,968 images, are used for training. The remaining 40%, containing 14,620 images, are used for testing. The training data covers exactly a full year, allowing the network to fully learn the characteristics of data in different seasons and weather. The testing data are from a time period completely different from the training data.

In this experiment, we evaluate different network structures and augmentation methods. As mentioned in Section 3.2, we keep the first three blocks as the shared layers and the last three blocks as a feature extractor to build two branch networks (2B). Additionally, we integrate the triplet loss (TRI) instead of the single focal loss (FOC) (only for estimating global ID) to reinterpret the image classification problem as an image re-identification problem. Furthermore, image generation by rain and haze model which forms data augmentation (AUG), are also verified in our experiments. Totally, networks with five different configurations (see Table 1) are trained. Moreover, we also compare the results of the one branch networks (1B) which cancels the coarse feature extractor from the proposed HDL.

Table 1. Five different configurations for comparative experiments.

	1B	**2B**	**FOC**	**TRI**	**AUG**
1B + TRI + AUG	✓			✓	✓
2B + FOC		✓	✓		
2B + FOC + AUG		✓	✓		✓
2B + TRI		✓		✓	
2B + TRI + AUG		✓		✓	✓

Curves of loss, accuracy, and margin w.r.t. the epoch number are shown in Figure 9. It is obvious that the utilization of two branch networks can make the loss converge faster than those with only one branch, and the accuracy and margin curves behave similarly. This means that the coarse features indeed help the network to learn the fine features. It can also be seen that employing the triplet loss can accelerate the network learning, which demonstrates the effectiveness of triplet loss. In contrast, training with the data augmentation has larger loss value but less accuracy and margin before the network converges. This can be attributed to the fact that both data augmentation approaches increase the diversity of training data and make the learning process more complicated.

In experiments, the precision–recall (PR) curves of compared methods are depicted in Figure 10. From the curves, following points can be observed: (1) The method proposed in the paper can achieve high precision on intersection re-identification. (2) The performance of hybrid double-level network is better than that of the single-branch network, which is mainly because the HDL network can integrate coarse-grained features into the mainstream and provide semantic information of different scales. (3) The data augmentation also brings a minor performance gain, which demonstrates the positive effect by improving the diversity of training data.

The quantitative experiments are designed with three additional baseline methods, i.e., the bag-of-words (BOW) [17], which is frequently used in visual SLAM to determine if a vehicle revisits the starting position in a loop closure detection, ConvNet [22] and SiameseNet [24], which are widely used in image classification or image matching for visual place recognition. For a better impression of network performance, we report the precision of evaluated methods w.r.t. different environmental conditions in Table 2. Normally, because of the low-light, image features of intersections strongly degrade at night compared with the daytime. In addition, the road texture in rainy and snowy weather has lower contrast, and the light scattering of raindrops also forms light spots on the image. All of these phenomena mentioned

have severely affected the performance of the traditional bag-of-words method. However, the precision of deep learning based methods fluctuates much less. We owe it to the fact that the neural network can automatically learn more distinctive and robust intersection information from the image feature embedding space than the hand-crafted features used in traditional methods.

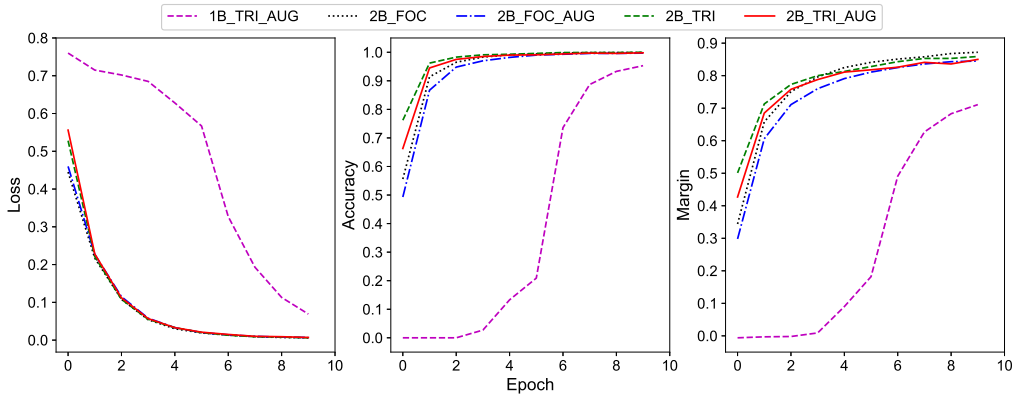

Figure 9. Loss, accuracy, and margin w.r.t. epoch number in training for different network structures and augmentation methods.

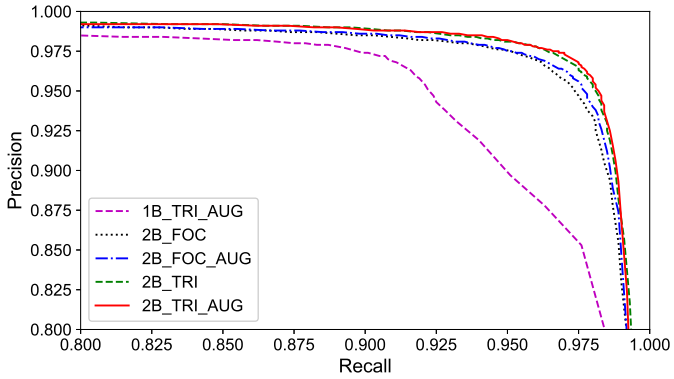

Figure 10. The PR curve.

Table 2. Precision of compared methods with baselines under different environmental conditions.

Methods	Total	Day	Night	Sunny	Overcast	Rainy
BOW [17]	0.689	0.736	0.311	0.685	0.746	0.544
ConvNet [22]	0.797	0.798	0.788	0.787	0.808	0.769
SiameseNet [24]	0.844	0.849	0.798	0.834	0.867	0.775
1B + TRI + AUG	0.964	0.963	0.972	0.956	0.968	0.893
2B + TRI	***0.971***	0.970	***0.989***	0.970	***0.973***	0.948
2B + TRI + AUG	***0.971***	***0.971***	***0.989***	***0.972***	0.972	***0.956***

From the results in Table 2, our proposed methods surpass the three baseline methods in the test on RobotCar Intersection data. This depends on the reasonable network structure and mixed loss. It can also be seen that the training with data augmentation fused with rain and haze model can further improve the precision of intersection re-identification. This is because the training and test dataset contains a small number of images captured during light rain or on a wet road. Training with augmented image data from

rain and haze model alleviates the imbalance of training samples. This also shows that the utilized rain and haze model can well simulate the image noise by rain and haze in the real traffic environment. For a further analysis, we also add generated rain and haze images (which are different from the images used in training) to the test data and re-evaluate each network. The dataset becomes more difficult, which can be seen from the decreased precision results shown in Table 3. However, the networks trained on augmented data still obtain better re-identification results on the augmented test data.

Table 3. Precision of compared methods on the RobotCar Intersection data fused with images generated by rain and haze models.

Methods	1B + TRI + AUG	2B + FOC	2B + FOC + AUG	2B + TRI	2B + TRI + AUG
Precision	0.933	0.906	0.958	0.901	***0.963***

Since dynamic objects can also interfere with the captured intersection scene by camera sensor, we conduct further experiments to evaluate the performance of proposed approach in different traffic flows. First, we assign intersection images of test set with labels which indicate different traffic flow: 0 for less surrounding dynamic objects, and 1 for more dynamic objects. The precision of compared methods are shown in Tables 4 and 5.

Table 4. Precision of compared methods in low traffic flow.

Methods	Total	Day	Night	Sunny	Overcast	Rainy
BOW [17]	0.728	0.743	0.349	0.652	0.777	0.356
ConvNet [22]	0.853	0.853	0.796	0.878	0.846	0.812
SiameseNet [24]	0.887	0.887	0.854	0.881	0.896	0.803
1B + TRI + AUG	0.975	0.975	0.968	0.983	0.976	0.931
2B + TRI	***0.982***	***0.982***	0.987	0.982	***0.984***	0.940
2B + TRI + AUG	***0.982***	***0.982***	***0.989***	***0.986***	0.983	***0.941***

Table 5. Precision of compared methods in high traffic flow.

Methods	Total	Day	Night	Sunny	Overcast	Rainy
BOW [17]	0.664	0.639	0.105	0.644	0.691	0.307
ConvNet [22]	0.738	0.738	0.745	0.707	0.763	0.713
SiameseNet [24]	0.861	0.861	0.910	0.887	0.887	0.690
1B + TRI + AUG	0.955	0.955	***0.980***	0.958	0.959	0.927
2B + TRI	***0.958***	0.957	***0.980***	***0.970***	***0.967***	0.911
2B + TRI + AUG	***0.958***	***0.958***	***0.980***	0.961	0.961	***0.928***

Comparing results in Tables 4 and 5, it is obvious that the accuracy of all intersection re-ID methods is reduced in high traffic flow with more dynamic traffic participants. This is mainly because the surrounding dynamic targets occlude part of the image of static environmental background, which makes some landmark signs, e.g., road markings, traffic facilities, surrounding buildings, not well recognized, and thus causes the mis-identification of intersection images. However, as seen from the comparison results, our proposed approach only exhibits a small accuracy decline and still outperforms other compared approaches.

In addition, the test results of BOW [17] show that the images captured on overcast days can be better recognized in different traffic flow. This indicates that the intersection images in overcast condition have richer textures than in other conditions, since the training of BOW [17] is to extract all point features in images. Moreover, the network by the training with rain-haze augmentation (2B + TRI + AUG),

in some extreme environmental conditions, such as in the rain, can obtain better re-identification results. This is mainly because the rain-haze augmentation imposes noise disturbance (which resembles the real scene) on the network. This enables the network to extract more robust features from the static environmental background.

5.4. Results on Campus Intersection

In this chapter, we use the Campus Intersection dataset to accomplish two main tasks: (1) verifying the effectiveness of the proposed method under panoramic camera modality; and (2) verifying if the double-level network could be used to detect new intersection images and expand the topological map. Hence, we divide the Campus Intersection dataset into two subsets according to the intersection ID. The 60% data of the first five intersections, containing 1456 images, are used to train the network. The remaining 40% images of the first five intersections, containing 995 images, are used to test the network on panoramic intersection images for existed intersection re-ID. The images of the last three intersections, containing 897 images, build the other subset of new intersection images and are used for the re-ID task of new intersections.

Due to minor changes in weather condition, in this experiment, we solely employ the double-level network with triplet loss and without rain-haze augmentation (2B + TRI). Similarly, images of each global ID are randomly selected from training images to form gallery images. Each testing image then is compared with each gallery image by the similarity function, and we calculate the accuracy of re-identification. The high precision results from the first row of the Table 6 show that the proposed network can well detect existed intersections with their intersection ID and attribute ID in the panorama image modality; meanwhile, it can re-identify the intersection with the global ID.

Table 6. Precision of proposed method under panoramic image modality in dealing with existed and new intersection images, which corresponds to the two main tasks in this chapter.

Task	Intersection ID	Attribute ID	Global ID
existed intersection re-ID (task 1)	0.985	0.987	0.998
new intersection detection (task 2)	-	0.855	0.990
new intersection re-ID (task 2)	-	0.855	0.991
all intersection re-ID (task 2)	-	0.924	0.996

Moreover, according to the expansion strategy of intersection topological map, the proposed network structure must be able to detect new intersection images, which can be categorized into two types: the first type of image is from an existed intersection but at a new lane or in a new direction. Thus, it is with a high intersection ID similarity but with a low global ID similarity. The second image type is from an entire new intersection, and thus with low similarities for both kinds of IDs. In this part of experiment, we use the global intersection ID to detect new intersection images. To determine the optimal threshold of similarity function, we first obtain the histograms by summarizing the similarity values between all query images (existed IDs and new IDs) and gallery images. Then, we set the trough of the histograms as the threshold, which is 0.87 for best detect precision of global ID. By adopting the optimal parameter settings in our experiment, the detection precision of attribute ID is 0.855 and for global ID, it is 0.990, as shown in the second row of Table 6.

For images with the same new global ID, we randomly select one of them and add it to the gallery. Then, we run the re-ID test with the new gallery again. The results are shown in the third row of Table 6 and demonstrate that our proposed intersection re-ID network can successfully handle images of new IDs. Since the attribute ID is from the classification of coarse features, it will be consistent with the result of

the second row. Now, the gallery is updated by the images of new IDs, then forming larger closed-paths. All the query images in the larger area will be tested and shown in the forth row of Table 6. All the results in Table 6 show that our proposed network are suitable for road intersection re-identification.

In the intelligent driving system, real-time performance is usually an important performance parameter of the designed network model. In the experiments, our approach can achieve a fast image processing speed of 16 frames per second (FPS). It can be applied in real-time autonomous driving tasks.

6. Conclusions and Future Work

In this paper, we propose a Hybrid Double-level network to solve an important but not well explored problem: traffic road intersection re-identification. We exploit a two-branch DCNN to accomplish multi-task including classification of intersection and its fine attributes, and global localization in topological maps. Compared with the traditional methods, our deep learning-based method avoids storing vast amount of feature points and has higher accuracy of re-identification. Experimental results demonstrate that our proposed network can complete the proposed multi-task of intersection re-identification mainly due to our elaborately designed network structure. We compared our network with state-of-the-art place recognition algorithms and the results demonstrated its superior performance, more than 12% recognition precision gain. For generalization and robustness in traffic environment, we consider the rain-haze model for data augmentation. It shows that the data augmentation brings a part of performance gain in re-ID task, which demonstrates the positive effect by improving the diversity of training data. For the lack of intersection re-identification datasets, we propose a carefully-labeled large-scale intersection dataset, called "RobotCar Intersection", which includes 36,588 images of eight intersections in different seasons and different day time. We also captured 3340 images of eight campus intersections to verify our re-ID approach on panorama modality and our expansion strategy of topology map. The related results show that our proposed method can detect and re-identify new intersection images.

In the future work, we plan to use attention mechanism, temporal characteristics, uncertainty estimation, fusion with GPS signals, etc., to improve the performance of our framework on intersection re-identification. Moreover, we plan to increase the intersection data by further recording real scene images as well as by generating virtual intersection scenes from simulation platforms and improved data augmentation models. To form a complete updating mechanism for intersection topology map, improved intersection detection is considered to fuse into our method.

Author Contributions: Conceptualization: L.X., Z.D., Y.H., W.T.; Supervision: L.X., Y.H., W.T.; Writing—review: Z.D., Y.H., X.Z., C.L., W.T.; Experiments: Z.D., W.D., X.Z., C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 52002285), the Shanghai Pujang Program (Grant No. 2020PJD075) and the Shenzhen Future Intelligent Network Transportation System Industry Innovation Center (Grant No. 17092530321).

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

References

1. Lowry, S.; Sunderhauf, N.; Newman, P.; Leonard, J.; Cox, D.; Corke, P.; Milford, M. Visual place recognition: A survey. *IEEE Trans. Robot.* **2016**, *32*, 1–19. [CrossRef]
2. Bhatt, D.; Sodhi, D.; Pal, A.; Balasubramanian, V.; Krishna, M. Have I reached the intersection: A deep learning-based approach for intersection detection from monocular cameras. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017; pp. 4495–4500.

3. Poggenhans, F.; Pauls, J.; Janosovits, J.; Orf, S.; Naumann, M.; Kuhnt, F.; Mayr, M. Lanelet2: A high-definition map framework for the future of automated driving. In Proceedings of the IEEE International Conference on Intelligent Transportation Systems, Maui, HI, USA, 4–7 November 2018; pp. 4495–4500.
4. Datondji, S.; Dupuis, Y.; Subirats, P.; Vasseur, P. A survey of vision-based traffic monitoring of road intersections. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2681–2698. [CrossRef]
5. Yu, C.; Feng, Y.; Liu, X.; Ma, W.; Yang, X. Integrated optimization of traffic signals and vehicle trajectories at isolated urban intersections. *Transp. Res. Part B-Methodol.* **2018**, *112*, 89–112. [CrossRef]
6. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 Year, 1000 km: The Oxford RobotCar dataset. *Int. J. Robot. Res.* **2016**, *36*, 3–15. [CrossRef]
7. Zhao, J.; Huang, Y.; He, X.; Zhang, S.; Ye, C.; Feng, T.; Xiong, L. Visual semantic landmark-based robust mapping and localization for autonomous indoor parking. *Sensors* **2019**, *19*, 161. [CrossRef] [PubMed]
8. Kushner, T.; Puri, S. Progress in road intersection detection for autonomous vehicle navigation. *Proc. SPIE* **1987**, *852*, 19–25.
9. Bhattacharyya, P.; Gu, Y.; Bao, J.; Liu, X.; Kamijo, S. 3D scene understanding at urban intersection using stereo vision and digital map. In Proceedings of the IEEE International Conference on Vehicular Technology, Sydney, Australia, 4–7 June 2017; pp. 1–5.
10. Habermann, D.; Vido, C.; Osorio, F. Road junction detection from 3D point clouds. In Proceedings of the IEEE International Conference on Neural Networks, Vancouver, BC, Canada, 24–29 July 2016; pp. 4934–4940.
11. Xie, X.; Philips, W. Road intersection detection through finding common sub-tracks between pairwise gnss traces. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 311. [CrossRef]
12. Cheng, J.; Gao, G.; Ku, X.; Sun, J. A novel method for detecting and identifying road junctions from high resolution SAR images. *J. Radars* **2012**, *1*, 100–109.
13. Kuipers, B.; Byun, Y. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Robot. Auton. Syst.* **1991**, *8*, 47–63. [CrossRef]
14. Lowe, D. Object recognition from local scale-invariant features. In Proceedings of the IEEE International Conference on Computer Vision, Liege, Belgium, 13–15 October 2009; pp. 1150–1157.
15. Bay, H.; Tuytelaars, T.; Van, L. SURF: Speeded up robust features. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
16. Rosten, E.; Drummond, T. Machine learning for high-speed corner detection. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 430–443.
17. Sivic, Z. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the IEEE International Conference on Computer Vision, Nice, France, 14–17 October 2003; pp. 1470–1477.
18. Ulrich, I.; Nourbakhsh, I. Appearance-based place recognition for topological localization. In Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco, CA, USA, 24–28 April 2000; pp. 1023–1029.
19. Krose, B.; Vlassis, N.; Bunschoten, R.; Motomura, Y. A probabilistic model for appearance-based robot localization. *Image Vis. Comput.* **2001**, *19*, 381–391. [CrossRef]
20. Oliva, A.; Torralba, A. Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res.* **2006**, *115*, 23–36.
21. Liu, Y.; Zhang, H. Visual loop closure detection with a compact image descriptor. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Algarve, Portugal, 7–12 October 2012; pp. 1051–1056.
22. Chen, Z.; Jacobson, A.; Sunderhauf, N. Deep learning features at scale for visual place recognition. In Proceedings of the IEEE International Conference on Robotics and Automation, Singapore, 29 May–3 June 2017; pp. 3223–3230.
23. Bromley, J.; Guyon, I.; Lecun, Y.; Sackinger, E.; Shah, R. Signature Verification using a "Siamese" Time Delay Neural Network. In Proceedings of the NIPS Conference on Neural Information Processing Systems, Denver, CO, USA, 26–28 May 1993; pp. 737–744.
24. Olid, D.; Facil, J.M.; Civera, J. Single-view place recognition under seasonal changes. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018; pp. 1–6.

25. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S. Deep learning for person re-identification: A survey and outlook. *arXiv* **2020**, arXiv:2001.04193.
26. Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; Huang, T. Deep relative distance learning: Tell the difference between similar vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–29 June 2016; pp. 737–744.
27. Milford, M.; Wyeth, G. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In Proceedings of the IEEE Conference on Robotics and Automation, Saint Paul, MN, USA, 15–18 May 2012; pp. 1643–1649.
28. Pandey, G.; Mcbride, J.; Eustice, R. Ford Campus vision and lidar data set. *Int. J. Robot. Res.* **2011**, *30*, 1543–1552. [CrossRef]
29. Cummins, M.; Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665. [CrossRef]
30. Glover, A.; Maddern, W.; Milford, M.; Wyeth, G. FAB-MAP+RatSLAM: Appearance-based SLAM for multiple times of day. In Proceedings of the IEEE Conference on Robotics and Automation, Anchorage, Alaska, 3–8 May 2010; pp. 3507–3512.
31. Badino, H.; Huber, D.; Kanade, T. Visual topometric localization. In Proceedings of the IEEE Conference on Intelligent Vehicles Symposium, Baden-Baden, Germany, 5–9 June 2011; pp. 794–799.
32. 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.
33. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 318–327.
34. Xiong, L.; Deng, Z.; Zhang, S.; Du, W.; Shan, F. Panoramic image mosaics assisted by lidar in vehicle system. In Proceedings of the IEEE Conference on Advanced Robotics and Mechatronics, Shenzhen, China, 18–21 December 2020; pp. 1–6.
35. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
36. Li, P.; Tian, J.; Tang, Y.; Wang, G.; Wu, C. Model-based deep network for single image deraining. *IEEE Access* **2020**, *8*, 14036–14047. [CrossRef]
37. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.* **2019**, *28*, 492–505. [CrossRef]