Day-Night Cross-domain Vehicle Re-identification

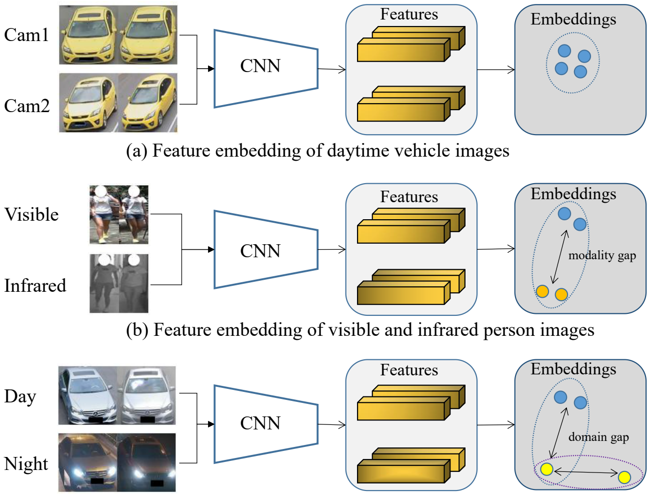
Hongchao Li[1], Jingong Chen[1], Aihua Zheng[2], Yong Wu[1*], Yonglong Luo[1]
[1]Anhui Normal University, [2]Anhui University
{lhc950304, ahzheng214}@foxmail.com, {chenjingong6, wuyong_tj}@163.com, ylluo@ustc.edu.cn

Abstract

Previous advances in vehicle re-identification (ReID) are mostly reported under favorable lighting conditions, while cross-day-and-night performance is neglected, which greatly hinders the development of related traffic intelligence applications. This work instead develops a novel Day-Night Dual-domain Modulation (DNDM) vehicle re-identification framework for day-night cross-domain traffic scenarios. Specifically, a unique night-domain glare suppression module is provided to attenuate the headlight glare from raw nighttime vehicle images. To enhance vehicle features under low-light environments, we propose a dual-domain structure enhancement module in the feature extractor, which enhances geometric structures between appearance features. To alleviate day-night domain discrepancies, we develop a cross-domain class awareness module that facilitates the interaction between appearance and structure features in both domains. In this work, we address the Day-Night cross-domain ReID (DN-ReID) problem and provide a new cross-domain dataset named DN-Wild, including day and night images of 2,286 identities, giving in total 85,945 daytime images and 54,952 nighttime images. Furthermore, we also take into account the matter of balance between day and night samples, and provide a dataset called DN-348. Exhaustive experiments demonstrate the robustness of the proposed framework in the DN-ReID problem. The code and benchmark are released at https://github.com/chenjingong/DN-ReID.

1. Introduction

Vehicle re-identification (ReID) is an important and advanced research topic in computer vision. It has great potential in multiple applications, such as intelligent transportation, urban surveillance, and social security. This is mainly because vehicles play a vital role as the primary objects of interest in urban surveillance scenarios. ReID aims to identify vehicle images of interest from the gallery captured by non-overlapping surveillance cameras. The main solution for ReID entails training a Convolutional Neural Network (CNN) [9, 10]. This allows samples of the same identity from different cameras to learn a consistent feature representation, as illustrated in Fig. 1 (a). Driven by large-scale datasets [17, 20, 22, 33] and the blossom of CNN, emerging re-identification methods [14, 21, 26, 42] keep setting state-of-the-art (SOTA) results in recent years.



(a) Feature embedding of daytime vehicle images

(b) Feature embedding of visible and infrared person images

(c) Feature embedding of daytime and nighttime vehicle images

Figure 1. Comparison of different ReID frameworks: (1) the most common single-modality visible domain ReID framework; (2) the cross-modality visible-infrared ReID framework; (3) the day-night cross-domain ReID framework.

Despite the advancements, the majority of vehicle re-identification (ReID) benchmarks and methods mainly concentrate on day-to-day matching, which is the common single-modality visible domain Re-ID problem. In numerous nighttime surveillance and low-light settings, thermal (near)-infrared cameras have the ability to capture the visual appearances of targets. This raises significant cross-modality visible-infrared ReID problems, such as visible-infrared person re-identification (VI-ReID) [27, 34]. The VI-ReID framework is typically formulated by learning modality-shared or invariant features to bridge the modal-

*Corresponding author.

Table 1. Publicly available benchmark datasets for person/vehicle re-identification (ReID). The term 'D-N proportion' refers to the proportion of identities with both day and night samples. 'D-N balance' signifies the equal distribution of training samples between day and night.

	Benchmark	ID	images	D-N proportion	D-N balance
person	VIPER [7]	632	1,264	0%	no
	iLIDS [44]	119	476	0%	no
	CUHK01 [16]	972	1,942	0%	no
	Market1501 [43]	1,501	32,668	0%	no
	DukeMTMC-ReID [29]	1,404	36,411	0%	no
	RegDB [27]	412	8,240	0%	no
	SYSU-MM01 [34]	491	303,357	0%	no
vehicle	VeRi-776 [20]	776	49,357	0%	no
	CityFlow-ReID [33]	666	229,680	0%	no
	VERI-Wild [22]	30,671	416,314	1.8%	no
	VERI-Wild 2.0 [1]	42,790	825,042	31.6%	no
	DNWild	2,286	140,897	100%	no
	DN348	348	34,077	100%	yes

ity gap, as shown in Fig. 1 (b). However, thermal (near)-infrared cameras have not been extensively utilized in vehicle ReID scenarios. There are several main reasons: 1. Thermal-infrared cameras are costly and possess limited resolution, which makes their utilization challenging in traffic situations. 2. Near-infrared cameras rely on emitted near-infrared light and can be easily influenced by vehicle headlights, street lights, and building lights in traffic situations. Given the reasons mentioned above, researchers cannot convert the day-to-night matching problem to the visible-to-infrared matching problem, as is done in person ReID. To facilitate the retrieval of nighttime images, we propose to utilize the setting of day-night cross-domain vehicle re-identification (DN-ReID). DN-ReID aims to identify visible images from the nighttime (daytime) gallery that belong to the same identity as the given daytime (nighttime) probe, as shown in Fig. 1 (c). Compared to the common single-modality visible domain ReID, DN-ReID encounters challenges arising from headlight glare and low-light environments. Moreover, viewpoint change, camera change, and occlusion problems which lead to large intra-class discrepancies in common ReID also bring difficulties to DN-ReID. Despite the practical significance of DN-ReID in real-world applications, there is currently a lack of research on this topic, making it an open issue that requires further exploration.

To support further research on DN-ReID problem, we provide a new dataset named DN-Wild. DN-Wild is collected from the testing set of VERI-Wild 2.0 dataset [1]. The DN-Wild training set consists of 70,981 daytime images and 35,384 nighttime images distributed over 1574 identities, while the query and gallery set is composed of 14,964 daytime images and 19,568 nighttime images from 712 identities for evaluation, respectively. Compared to other commonly used ReID datasets, as shown in Table 1, DN-Wild includes day-night vehicle image pairs for each identity. This allows users to evaluate the performance of day-night cross-domain scenarios. However, DN-Wild exhibits more noticeable sample imbalances, as depicted in Fig. 2 (a). The majority of classes have only a restricted number of nighttime (daytime) samples, whereas there is an abundance of daytime (nighttime) samples. To this phenomenon, we propose a new sample-balanced dataset, named DN-348. DN-348 is captured across four full months (from May to September) with a city-scale surveillance camera system covering an area of more than 400 km^2. The DN-348 training set comprises 200 identities with 9,962 daytime images and 10,022 nighttime images, while the query and gallery set is composed of 10,121 daytime images and 3972 nighttime images from 148 identities for evaluation, respectively. There are approximately 50 daytime images and 50 nighttime images per vehicle in the DN-348 training set, as shown in Fig. 2 (b). In general, DN-Wild and DN-348 exhibit different data sizes and distributions, thus offering the diverse evaluation for future DN-ReID methods.

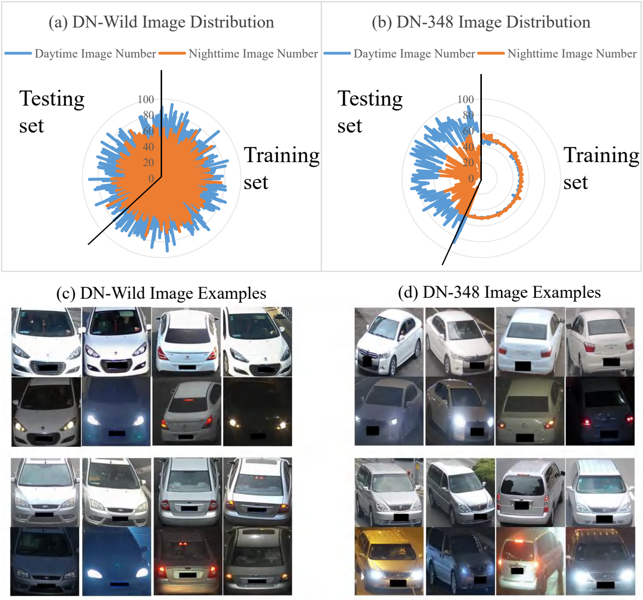


Figure 2. Image distribution and image examples in the DN-Wild and DN-348 datasets.

To provide a robust baseline algorithm, we propose a Day-Night Dual-domain Modulation (DNDM) network to consider the difficulties posed by headlight glare, low-light environments, and domain discrepancies. Inspired by the visual prompts [2, 19], we construct a night-domain glare suppression module and employ the highlighted area as visual prompts to effectively reduce headlight glare. Then, we present a dual-domain structure enhancement module that utilizes local feature gradients to learn geometric structures, which are resistant to interference from low-light environments. Additionally, we introduce a cross-domain class awareness module that utilizes the class activation maps to interact with the day-night cross-domain features. To minimize disparities in appearance and structural features between the two domains, we expand the dual-domain structure enhancement module and the cross-domain class awareness module to multiple stages within the backbone network.

Overall, The contributions of this paper can be summarized as follows.

- We provide two standardized benchmark datasets, DN-Wild and DN-348, to facilitate the study of DN-ReID. These benchmark datasets will be freely accessible to the public for academic research.
- We propose the Day-Night Dual-domain Modulation (DNDM) framework, which integrates the training of glare suppression, structure enhancement, and class awareness to dynamically modulate day-night cross-domain vehicle features.
- Comprehensive experiments conducted on our challenging benchmark datasets, DN-348 and DN-Wild, validate the superior performance and potential of our DNDM for day-night cross-domain vehicle ReID problem.

2. Related Work

We briefly review the related work in the following two folds, *i.e.*, vehicle re-identification and visible-infrared cross-modality person ReID (VI-ReID).

2.1. Vehicle Re-identification

The task of vehicle ReID has gained significant attention in recent years due to its widespread application in video surveillance and social security [10, 12, 42]. Liu *et al.* [17] introduce the VeRi-776 benchmark dataset and propose a deep relative distance learning approach for the vehicle ReID task. Lou *et al.* [22] introduce the VERI-Wild dataset for the vehicle ReID community in the wild, and design a feature distance adversary scheme to generate hard negative samples. Bai *et al.* [1] extend the VERI-Wild dataset and introduce the VERI-Wild 2.0 dataset. However, the majority of day-night sample pairs exist in the testing set of the VERI-Wild 2.0 dataset [1], making it unsuitable for addressing the day-night cross-domain problem.

In vehicle ReID, Zhou *et al.* [47] utilize a viewpoint-aware attentive multi-view inference model to acquire multi-view vehicle features. Lou *et al.* [23] incorporate an adversarial learning network into the vehicle ReID task to address the challenge of hard negative cross-view and same-view images. To learn part-based features for vehicle ReID, He *et al.* [8] integrate a detection branch to learn part-regularized features. Zhao *et al.* [42] propose a heterogeneous relational complement network that utilizes cross-level features and region-specific features as complements to enhance high-level features. Li *et al.* [15] propose a framework for vehicle ReID that utilizes knowledge vectors to guide the training of a transformer model. Shen *et al.* [30] propose a graph interactive transformer method to explore the interaction between local features and global features for vehicle ReID. While these methods have made significant advancements in solving the ReID problem, they fail to consider the potential relationships between vehicle images taken during the day and night.

2.2. Visible-Infrared Cross-modality Person Re-ID

In recent years, there has been a rise in the popularity of visible-infrared cross-modality person ReID (VI-ReID) [6, 35, 41]. Nguyen *et al.* [27] propose the RegDB (visible *vs* thermal-infrared) dataset, which consists of visible-infrared image pairs. This dataset is widely used for VI-ReID, but it is important to note that it was captured by only one visible-infrared camera. Wu *et al.* [34] introduce the SYSU-MM01 (visible *vs* near-infrared) dataset and propose a deep zero-padding network for matching visible-infrared images. Zhang *et al.* [40] propose a low-light visible-infrared cross-modality LLCM (visible *vs* near-infrared) dataset to aid in the research of VI-ReID for practical applications. While current VI-ReID datasets have started to address practical applications in low-light environments, they are not suitable for the vehicle ReID task.

In VI-ReID, Ye *et al.* [38] design an attention generalized mean pooling with weighted triplet loss method for VI-ReID. Liu *et al.* [18] propose a memory-augmented unidirectional metric learning method to learn cross-modality metrics. Sun *et al.* [31] propose a dense contrastive learning framework that promotes pixel-to-pixel dense alignment for visible-infrared image pairs. Jiang *et al.* [11] design a cross-modality transformer that aims to jointly explore a modality-level alignment module and an instance-level module for VI-ReID. Lu *et al.* [24] propose a named progressive modality-shared transformer to mitigate the adverse impact of modality gap. Yu *et al.* [39] design a modality unifying network, which can dynamically model the modality-specific and modality-shared representations to alleviate both cross-modality and intra-modality variations. Although these methods consider the intra-modality variations and cross-modality variations, they fail to address the challenges presented by headlight glare and low-light environments in nighttime vehicle images.

3. Method

3.1. Model Architecture

Fig. 3 provides an overview of the proposed Day-Night Dual-domain Modulation (DNDM) framework, which utilizes the ResNet-50 [9] as its backbone network. The backbone network is used to extract vehicle appearance feature tensors from both daytime and nighttime images at multi-

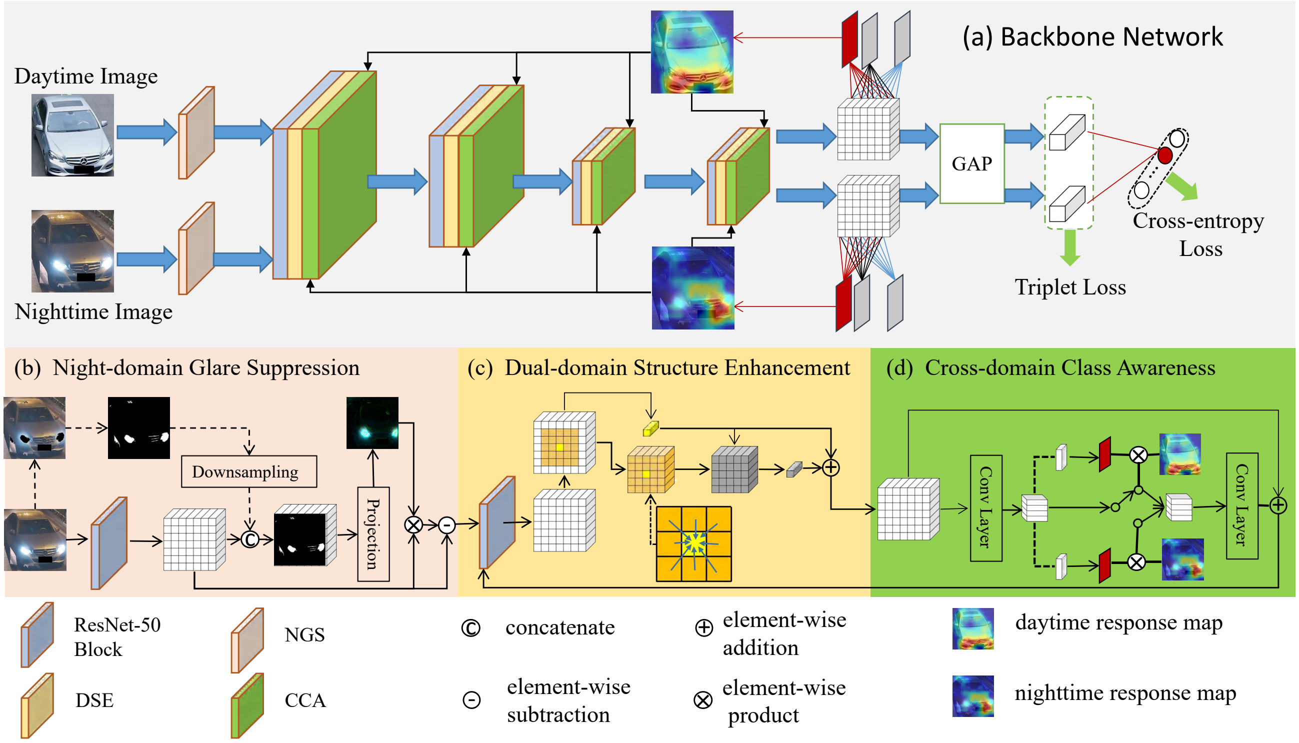


Figure 3. Pipeline of Day-Night Dual-domain Modulation (DNDM) framework. The whole network consists of a backbone network, a Night-domain Glare Suppression (NGS) module, a Dual-domain Structure Enhancement module (DSE), and a Cross-domain Class Awareness (CCA) module. The backbone network specifically employs ResNet-50 for vehicle appearance representation, using both cross-entropy loss and triplet loss for supervision. The NGS module uses a glare mask and a trainable projection vector to extract nighttime features while suppressing glare. The DSE module learns the structural representation by utilizing feature gradients within local windows. The CCA module improves the visual and structural representation by exchanging the class activation maps from the day and night domains.

ple stages. To enhance the learning of normal regions and mitigate glare in nighttime image, we propose the Night-domain Glare Suppression (NGS) module. Subsequently, the Dual-domain Structure Enhancement (DSE) module is introduced to aggregate gradients from local windows and capture diverse structural representations. To facilitate interaction between day-night cross-domain features under the same identity, we design a Cross-domain Class Awareness (CCA) module that follows the DSE module. The collaboration between the DSE and CCA modules aims to enhance appearance and structural representation, enabling their effective utilization across various stages of the backbone network.

3.2. Baseline

Day-night cross-domain vehicle re-identification (DN-ReID) aims to retrieve vehicles of interest in both daytime and nighttime environments. Given a pair of vehicle images $I = \{(I^{Day}, I^{Night}), y\}$, where I^{Day} and I^{Night} are the input daytime and nighttime vehicle images respectively, and y is the associated vehicle identity label. The corresponding multi-stage feature tensors encoded by the backbone network are denoted as $T_s^m \in \mathbb{R}^{H^s \times W^s \times C^s}$, $s \in \{0, 1, 2, 3, 4\}$, $m \in \{Day, Night\}$. Following the ResNet-50 [9] backbone as shown in Fig. 3 (a), we use a global average pooling (GAP) layer to obtain the corresponding feature vector $f^m = GAP(T_4^m)$. The network is then optimized with respect to a cross-entropy loss $\mathcal{L}_{ce}$ and a triplet loss $\mathcal{L}_{tri}$. The cross-entropy loss is formulated as:

$$\mathcal{L}_{ce} = -ylog(Softmax(FC_{class}(f^m))), \tag{1}$$

where FC_{class} denotes a fully connected layer that predicts the result of classification, $Softmax$ denotes the softmax function that gets the normalized probability. It is worth noting that f^{Day} and f^{Night} share the same FC_{class} layer. The triplet loss is formulated as:

$$\mathcal{L}_{tri} = max(0, d_{ij}^p + margin - d_{ik}^n), \tag{2}$$

where (i, j, k) represents a hard triplet within each training batch. For daytime anchor sample i, j is from the corresponding nighttime positive set, and k is from the daytime negative set. For nighttime anchor sample i, j is from corresponding daytime positive set, and k is from the nighttime negative set. d_{ij}^p/d_{ik}^n represents the pairwise distance of a positive/negative sample pair, and $margin = 0.3$ denotes the triplet distance margin.

Although the above backbone network can extract vehicle features, it does not effectively tackle the challenges presented by headlight glare, low-light environments, and domain discrepancies. To alleviate the challenges existing in DN-ReID, we introduce the Night-domain Glare Suppression (NGS) module, Dual-domain Structure Enhancement module (DSE), and Cross-domain Class Awareness (CCA) module in the following subsections.

3.3. Night-domain Glare Suppression (NGS)

Drawing upon the concept of visual prompts as highlighted in the studies [2, 19], the integration of visual cues can be crucial in identifying the exact details of a task. We propose a Night-domain Glare Suppression (NGS) module that utilizes glare prompts to guide attention towards glare-free regions and reduce the effects of headlight glare. Given a nighttime vehicle image I^{Night}, we initially adopt the convolutional block to obtain the feature tensor T_0^{Night}:

$$T_0^{Night} = Mxp_{2\times2}(ReLU(BN(conv_{7\times7}(I^{Night})))), \tag{3}$$

where $conv_{7\times7}$ represents a 7×7 convolutional operation, BN denotes the batch normalize operation, $ReLU$ denotes the rectified linear unit, and $Mxp_{2\times2}$ denotes a 2×2 max pooling operation.

Meanwhile, we convert the nighttime vehicle image I^{Night} to a grayscale image $I^G = rgb2gray(I^{Night})$. Then, we apply a brightness threshold of 220 to identify the highlighted pixels in the night image. After identifying the highlighted pixels, we combine the adjacent pixels into regions while discarding regions with fewer pixels. Finally, we obtain a binary mask $\mathbf{M}^G$ from the initial night image, where 1 represents the pixels affected by glare and 0 represents the pixels unaffected by glare. The binary mask is included as input in the glare suppression module and is modulated with the feature tensor T_0^{Night}. Specifically, we concatenate feature tensor $T_0^{Night} \in \mathbb{R}^{H^0\times W^0\times C^0}$ with the binary mask $\mathbf{M}^G \in \mathbb{R}^{H^0\times W^0\times 1}$, and then we feed them to a learnable projection vector $V \in \mathbb{R}^{(C^0+1)\times 1}$. Mathematically,

$$M_0^{Night} = Sigmoid(concat(T_0^{Night}, M^G)V), \tag{4}$$

where $M_0^{Night} \in \mathbb{R}^{H^0\times W^0\times 1}$ represents the learned domain suppression mask, and $Sigmoid$ refers to the sigmoid function.

In addition to the nighttime image, our NGS module is also guided by the daytime image. The basic idea is to utilize a dummy mask, $zeros(M^G)$, to prompt the difference between the glare-unaffected daytime feature and glare-affected nighttime feature:

$$M_0^{Day} = Sigmoid(concat(T_0^{Day}, zeros(M^G))V), \tag{5}$$

where $zeros(\cdot)$ represents the operation of setting all elements to 0. In the process of projection, one of the inputs is the concatenation of T_0^{Night} and $\mathbf{M}^G$. Another input is a glare-unaffected daytime feature tensor, T_0^{Night}, concatenated with a dummy all-zero mask $zeros(M^G)$. Based on the domain suppression mask $M_0^m, m \in \{Day, Night\}$, the final suppression procedure can be formulated as:

$$\overline{T}_0^m = T_0^m - \alpha M_0^m \odot T_0^m, \tag{6}$$

where $\overline{T}_0^m, m \in \{Day, Night\}$ denotes the daytime/ nighttime feature after the glare suppression operation, $\alpha = 0.5$ is a hyperparameter used to balance the original feature and the weakened feature. Fig. 3 (b) shows the result of our NGS module, illustrating the successful separation of glare regions. After performing the aforementioned operations, we feed $\overline{T}_m$ into the backbone network to acquire the corresponding feature tensors $T_s^m \in \mathbb{R}^{H^s\times W^s\times C^s}, s \in \{1, 2, 3, 4\}$.

3.4. Dual-domain Structure Enhancement (DSE)

The common vehicle re-identification network mainly focuses on extracting vehicle appearance features. However, the appearance features are easily influenced by low-light environments. To improve the feature consistency of day-night vehicle image pairs, we introduce a Dual-domain Structure Enhancement (DSE) module. The main idea of the DSE module is to extract the structural information from the appearance features through pixel-wise gradient. Specifically, the DSE module processes the intermediate feature map T_s^m and calculates the pixel-wise non-negative local gradient $G \in \mathbb{R}^{H^s\times W^s\times C^s\times N\times N}$ for each pixel position x and its surrounding region of size $N \times N$:

$$G(x, c, d) = \max(0, T_s^m(x + d, c) - T_s^m(x, c)) \tag{7}$$

where $c \in [1, C^s]$ represents the index of the channel dimension and $d \in [-d_n, d_n] \times [-d_n, d_n]$ denotes the neighbor position in the surrounding region of each pixel x. The size of the region is $N \times N$, with $d_n = (N - 1)/2$. Moreover, we propose a feature weighting operation that incorporates detailed local gradients into a concise structural descriptor, enabling the simultaneous learning of geometric structures from both domains:

$$\begin{aligned} S(x, c, d) &= \frac{G(x, c, d)}{1 + \sum_d G(x, c, d)} T_s^m(x + d, c), \\ S(x, c) &= \sum_d S(x, c, d), \end{aligned} \tag{8}$$

where $S \in \mathbb{R}^{H^s\times W^s\times C^s}$ has the same spatial and channel size as the original feature tensor T_s^m. The gradient-guided feature weighting operation aggregates the neighbor features into structural features, thereby reducing its spatial dimension from $N \times N$ to 1×1. This transformation converts the raw local gradient G into the structural descriptor S. In simpler terms, the structural descriptor is derived from the weighted appearance descriptor. Then, we utilize the structural descriptor as an additional input for the appearance descriptor:

$$\overline{T}_s^m = T_s^m + \beta S, \tag{9}$$

where $\overline{T}_s^m, m \in \{Day, Night\}$ denotes the daytime/ nighttime feature after the structure enhancement operation, the hyperparameter $\beta = 0.25$ is used to balance the original feature and the enhanced feature.

3.5. Cross-domain Class Awareness (CCA)

In our network, we first use ResNet-50 to extract appearance features from day-night cross-domain vehicle images. Then, we apply a glare suppression module to reduce the impact of glare in nighttime images. Furthermore, we introduce a structure enhancement module to improve the appearance features. However, these modules fail to consider the discrepancies between the day and night domains.

To tackle the day-night domain gap, we introduce a Cross-domain Class Awareness (CCA) module for the DN-ReID problem. Given day-night cross-domain vehicle features $\overline{T}_s^m \in \mathbb{R}^{H^s \times W^s \times C^s}, s \in \{1,2,3,4\}$. We convert it into its projection $P_s^m \in \mathbb{R}^{H^s \times W^s \times C^4}$ using a convolutional layer. This adjustment is to match the number of feature channels with the input size of the fully connected layer FC_{class} as described in formula 1. Mathematically,

$$P_s^m = BN(conv_{1\times1}(\overline{T}_s^m)), \tag{10}$$

where $conv_{1\times1}$ represents a 1×1 convolutional operation, and BN represents a batch normalization operation. Inspired by the class activation mapping (CAM) operation [46], which can highlight the class-specific discriminative regions. We introduce the projection P_s^m to calculate the class activation maps for daytime and nighttime images:

$$A_i^m|_{i=1}^{C'} = Sigmoid(FC_{class}(P_s^m)), \tag{11}$$

where C' represents the total number of classes in the training set. The class activation maps for the y-th class are denoted as A_y^{Day} and A_y^{Night}, where y represents the vehicle identity label. The sigmoid function is used to normalize the CAMs.

To facilitate the interaction between day-night cross-domain features, we suggest exchanging class awareness information between day and night samples:

$$\begin{aligned} \overline{P}_s^{Day} &= P_s^{Day} \odot A_y^{Night}, \\ \overline{P}_s^{Night} &= P_s^{Night} \odot A_y^{Day}, \end{aligned} \tag{12}$$

where $\odot$ represents element-wise multiplication. It is important to note that this exchange process only occurs during the training phase, and the class activation maps are not exchanged during the test phase. To ensure that $\overline{P}_s^m$ and $\overline{T}_s^m$ have the same number of channels, we incorporate the operation of $conv_{1\times1} + BN + ReLU$ into our CCA module. The resulting feature can be expressed as:

$$T^m = \overline{T}_s^m + ReLU(BN(conv_{1\times1}(\overline{P}_s^m))). \tag{13}$$

Overall loss. We utilize the widely adopted ResNet-50 as the backbone network. We integrate the proposed Night-domain Glare Suppression (NGS) module before the first block. Additionally, we incorporate the Dual-domain Structure Enhancement (DSE) module and the Cross-domain Class Awareness (CCA) module after each of the four convolution blocks (refer to Fig. 3). The entire network is trained in an end-to-end manner. The overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{tri}. \tag{14}$$

4. Experiments

We adopt the Cumulative Match Curve (CMC) with Rank-k matching accuracy and mean Average Precision (mAP) as the evaluation metrics. The red and blue respectively represent the first and second results.

4.1. Implementation details

In our experiments, we adopt ResNet-50 [9] pretrained on ImageNet [3] without the last spatial down-sampling layer as the backbone model followed by [32]. We use the Adam [13] optimizer with the initial learning rate of 1.0×10^{-2}. We apply a warmup [5] approach to initialize the network, gradually raising the learning rate from 1.0×10^{-2} to 1.0×10^{-1} over 10 epochs. Afterward, we maintain the learning rate at 1.0×10^{-1} from the 10-th to the 20-th epoch. The learning rate then further decays to 1.0×10^{-3} at the 20-th epoch and to 1.0×10^{-4} at the 50-th epoch, and this continues until a total of 80 epochs are completed. The training protocol follows the ReID strong baseline (BOT [25]) using random cropping and erasing [45] for data augmentation. In our implementation, all the input images are resized to 256×256. The size of the window region is $N \times N = 5 \times 5$ in the DSE module. The dimension of final features is 2048. We set eight IDs, and eight (four daytime + four nighttime) instances with the batch size of 64 in the training for the two datasets. We run our experiments on one NVIDIA GeForce RTX A6000 GPU with 48GB RAM.

4.2. Comparison to State-of-the-art Methods

Evaluation Results on DN-348. Table 2 reports the performance comparison of our DNDM against the state-of-the-art methods on the DN-348 dataset. From which we

Table 2. Comparison results of our method against the state-of-the-art methods on DN-348 and DN-Wild dataset.

Methods	DN-348						DN-Wild					
	Day-to-Night			Night-to-Day			Day-to-Night			Night-to-Day		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
BOT [25]	0.663	0.805	0.449	0.777	0.904	0.453	0.503	0.979	0.399	0.471	0.936	0.398
DDAG [36]	0.666	0.803	0.440	0.753	0.903	0.423	0.495	0.980	0.361	0.487	0.951	0.352
LbA [28]	0.680	0.828	0.449	0.766	0.897	0.429	0.472	0.936	0.263	0.417	0.871	0.254
AGW [38]	0.681	0.821	0.465	0.731	0.894	0.443	0.483	0.960	0.387	0.491	0.963	0.386
CAJ [37]	0.660	0.819	0.464	0.739	0.884	0.453	0.499	0.981	0.392	0.487	0.955	0.385
DCL [31]	0.675	0.813	0.443	0.789	0.920	0.428	0.499	0.979	0.348	0.472	0.943	0.342
PMT [24]	0.663	0.820	0.470	0.760	0.907	0.461	0.491	0.955	0.327	0.444	0.902	0.337
Baseline	0.674	0.826	0.456	0.723	0.889	0.439	0.492	0.965	0.395	0.485	0.951	0.349
DNDM	0.707	0.842	0.475	0.803	0.926	0.462	0.512	0.987	0.405	0.495	0.955	0.400

can see, the state-of-the-art VI-ReID methods have not achieved significant performance improvements compared to the ReID strong baseline BOT [25]. The reason is that these methods fail to address challenges presented by headlight glare and low-light environments in nighttime vehicle images. For the Day-to-Night setting, our approach significantly beats the VI-ReID methods as 70.7%, 84.2%, and 47.5% on the Rank-1, Rank-5, and mAP respectively. For the Night-to-Day setting, our approach significantly beats the VI-ReID methods as 80.3%, 92.6%, and 46.2% on the Rank-1, Rank-5, and mAP respectively. Compared with the baseline, our proposed DNDM significantly improves Rank-1, Rank-5 and mAP by 8.0%, 3.7%, and 2.3% respectively. This shows the promising achievement by employing night-domain glare suppression and dual-domain structure enhancement to improve the feature learning of nighttime vehicle images. In conclusion, these findings emphasize that DN-ReID is a challenging task with the potential to match vehicle images at night.

Evaluation Results on DN-Wild. Table 2 shows the comparison results on DN-Wild dataset on two different testing sets. In general, our proposed DNDM achieves promising performance compared to state-of-the-art methods. For the Day-to-Night setting, our approach significantly beats the ReID strong baseline BOT [25] by 51.2%, 98.7% and 40.5% on the Rank-1, Rank-5, and mAP respectively. For the Night-to-Day setting, compared with BOT [25], our proposed DNDM significantly improves Rank-1, Rank-5, and mAP by 2.4%, 1.9%, and 0.2% respectively. This shows the promising achievement by the training of glare suppression, structure enhancement, and class awareness to learn day-night cross-domain features. From Table 2, PMT [24] does not show the same competitiveness on the DN-Wild dataset as it does on the DN-348 dataset. It means that considering the domain gap is not sufficient for the DN-Wild, which suffers from drastic sample imbalance. Furthermore, it can be observed that our method outperforms the mAP of baseline model, which indicates the robust generalization ability of the proposed model in large-scale datasets.

4.3. Ablation Study

Settings			DN-348			
			Day-to-Night		Night-to-Day	
NGS	**DSE**	**CCA**	R-1	mAP	R-1	mAP
			0.674	0.456	0.723	0.439
✓			0.690	0.467	0.777	0.455
	✓		0.681	0.461	0.745	0.455
		✓	0.684	0.463	0.765	0.453
✓	✓		0.699	0.469	0.800	0.460
✓	✓	✓	**0.707**	**0.475**	**0.803**	**0.462**

Table 3. Ablation study on DN-348 dataset.

Effectiveness of each component. To verify the unique contributions of each module in our model, we implement the ablation study of several variants of our method on the DN-348 dataset. As shown in Table 3, the NGS, DSE, and CCA modules all show significant improvements. Specifically, on the DN-348 dataset, with only the NGS module activated, we observe Rank-1 performance of 69.0%, which further escalates to 69.9% when both the NGS and DSE modules are enabled. When the CCA module is combined with the NGS and DSE modules, the Rank-1 performance on DN-348 reaches 70.7%. These results validate the consistent performance improvement attributed to these three modules both individually and jointly.

The influence of which stage of ResNet-50 to plug the DSE module and CCA module. The DSE and CCA modules can be inserted into the backbone network at any stage. In our experiments, we utilize ResNet-50 as the backbone network, consisting of four stages. We analyze the impact of integrating the DSE and CCA modules at different stages of ResNet-50 in our experiments. In Table 4, it is evident that there is a clear improvement as our modules are integrated deeper into the stages of ResNet-

Methods	DN-348			
	Day-to-Night		Night-to-Day	
	R-1	mAP	R-1	mAP
Baseline	0.674	0.456	0.723	0.439
+ stage 1	0.696	0.470	0.784	0.454
+ stage (1+2)	0.696	0.473	0.785	0.459
+ stage (1+2+3)	0.699	0.475	0.799	0.459
+ stage (1+2+3+4)	**0.707**	**0.475**	**0.803**	**0.462**

Table 4. The influence of which stage of ResNet-50 to plug the DSE module and CCA module.

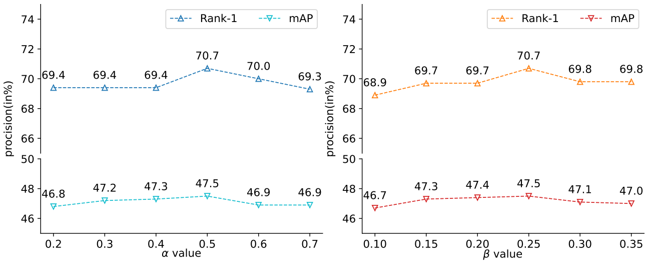


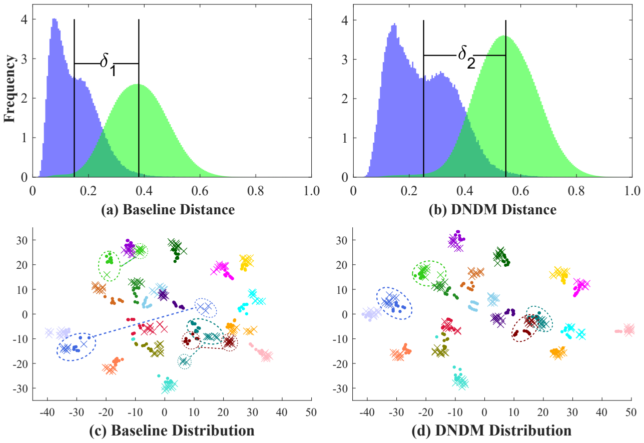
Figure 4. Parameter analysis (in %). The coefficients α and β are linked to the NGS and DSE modules, respectively.

50. In the DN-348 dataset, the Rank-1 score has increased steadily from 69.6% at the initial stage to 70.7% by the fourth stage. Simultaneously, the mAP performance has advanced from 47.0% following the initial stage to 47.5% after the fourth stage. It verifies the effectiveness of our day-night dual-domain modulation framework, which learns day-night cross-domain information on multiple stages to boost DN-ReID.

4.4. Other analysis

Parameter analysis. To evaluate the influence of the two hyperparameters, we give quantitative comparisons and report the results in Fig. 4. Different values of α and β significantly influence the performance of the NGS and DSE modules. As observed, the optimal performance is achieved with α and β values set to 0.5 and 0.25, respectively.

Visualization study. To further analyze the effectiveness of our DNDM, we conduct experiments on the DN-348 dataset to calculate the frequency of inter-identity and intra-identity distances. Fig. 5 (a, b) display the distance distributions acquired by the baseline and the proposed DNDM, respectively. Comparing Fig. 5 (b) with Fig. 5 (a), we can observe that $\delta_1 < \delta_2$. This indicates that the inter-identity and intra-identity distances are significantly separated using the proposed method. Moreover, we visualize the feature distribution of 20 vehicles in a 2D feature space using the T-SNE method [4]. In Fig. 5 (c, d), it is evident that the proposed DNDM markedly reduces the distances between day-night images of the same identity and successfully minimizes the domain discrepancy.



(a) Baseline Distance

(b) DNDM Distance

(c) Baseline Distribution

(d) DNDM Distribution

Figure 5. (a-b) The distributions of the two types of distances between the day-night cross-domain features. The intra-identity and inter-identity distances are represented by blue and green color, respectively. (c-d) The T-SNE distribution [4] shows 20 vehicles from the DN-348 testing set. Samples of the same color are from the same vehicle. The ”dot” and ”cross” markers represent images from the daytime and nighttime domains, respectively.

5. Conclusion

To our best knowledge, this is the first work to address the day-night cross-domain vehicle ReID (DN-ReID) problem. We have contributed two new DN-ReID datasets, along with an innovative DN-ReID approach. Compared to day-to-day vehicle ReID, DN-ReID faces challenges posed by headlight glare, low-light environments, and domain discrepancies. Therefore, we propose a Day-Night Dual-domain Modulation (DNDM) network that combines the learning of glare suppression, structure enhancement, and class awareness to dynamically modulate day-night cross-domain vehicle features. Extensive experiments demonstrate the promising performance of the proposed method. In addition, drawing from our research, we emphasize several crucial findings for DN-ReID. First, annotating vehicle images at night presents a challenge. Second, enhancing features in nighttime vehicle images is proven to be effective. Finally, it is worthwhile to consider the capability of day-night cross-domain data for identifying the same ID. In the future, we will enhance the aforementioned components to advance the state-of-the-art of DN-ReID and explore label-free DN-ReID.

Acknowledgement. This research is supported in part by the National Natural Science Foundation of China (Nos. 62302013, 62372003 and 62272006), and the Natural Science Foundation of Anhui Province (Nos. 2308085QF220, 2308085Y40).

References

[1] Yan Bai, Jun Liu, Yihang Lou, Ce Wang, and Ling-Yu Duan. Disentangled feature learning network and a comprehensive benchmark for vehicle re-identification. *IEEE TPAMI*, 44 (10):6854–6871, 2021. 2, 3

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Proceedings of the NeurIPS*, 35:25005–25017, 2022. 2, 5

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the CVPR*, pages 248–255, 2009. 6

[4] Laurens Van Der Maaten and Geoffrey E Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008. 8

[5] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Sphereid: Deep hypersphere manifold embedding for person re-identification. *JVCI*, 60:51–58, 2019. 6

[6] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the CVPR*, pages 22752–22761, 2023. 3

[7] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of the PETS Workshops*, pages 1–7, 2007. 2

[8] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the CVPR*, pages 3997–4005, 2019. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the CVPR*, pages 770–778, 2016. 1, 3, 4, 6

[10] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the ACM MM*, pages 9664–9667, 2023. 1, 3

[11] Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *Proceedings of the ECCV*, pages 480–496, 2022. 3

[12] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Juncheng Chen, and Rama Chellappa. A dual path model with adaptive attention for vehicle re-identification. In *Proceedings of the ICCV*, pages 6132–6141, 2019. 3

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Hongchao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. Attribute and state guided structural embedding network for vehicle re-identification. *IEEE TIP*, 31:5949–5962, 2022. 1

[15] Hongchao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. Mskat: Multi-scale knowledge-aware transformer for vehicle re-identification. *IEEE TITS*, 23(10):19557–19568, 2022. 3

[16] Wei Li, Rui Zhao, and Xiaogang Wang. Human re-identification with transferred metric learning. In *Proceedings of the ACCV*, pages 31–44, 2013. 2

[17] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the CVPR*, pages 2167–2175, 2016. 1, 3

[18] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the CVPR*, pages 19366–19375, 2022. 3

[19] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the CVPR*, pages 19434–19445, 2023. 2, 5

[20] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Proceedings of the ECCV*, pages 869–884, 2016. 1, 2

[21] Xinchen Liu, Wu Liu, Jinkai Zheng, Chenggang Yan, and Tao Mei. Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification. In *Proceedings of the ACM MM*, pages 907–915, 2020. 1

[22] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the CVPR*, pages 3235–3243, 2019. 1, 2, 3

[23] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Embedding adversarial learning for vehicle re-identification. *IEEE TIP*, 28(8):3794–3807, 2019. 3

[24] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI*, pages 1835–1843, 2023. 3, 7

[25] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE TMM*, 22(10):2597–2609, 2019. 6, 7

[26] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the CVPR*, pages 7103–7112, 2020. 1

[27] Dat Nguyen, Hyung Hong, Ki Kim, and Kang Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 1, 2, 3

[28] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the ICCV*, pages 12046–12055, 2021. 7

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the ECCV*, pages 17–35, 2016. 2

[30] Fei Shen, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng. Git: Graph interactive transformer for vehicle re-identification. *IEEE TIP*, 32:1039–1051, 2023. 3

[31] Hanzhe Sun, Jun Liu, Zhizhong Zhang, Chengjie Wang, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Not all pixels are

matched: Dense contrastive learning for cross-modality person re-identification. In *Proceedings of the ACM MM*, pages 5333–5341, 2022. 3, 7

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the ECCV*, pages 480–496, 2018. 6

[33] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the CVPR*, pages 8797–8806, 2019. 1, 2

[34] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the ICCV*, pages 5380–5389, 2017. 1, 2, 3

[35] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the CVPR*, pages 14308–14317, 2022. 3

[36] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of the ECCV*, pages 229–247, 2020. 7

[37] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the ICCV*, pages 13567–13576, 2021. 7

[38] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 44(6): 2872–2893, 2021. 3, 7

[39] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the ICCV*, pages 11185–11195, 2023. 3

[40] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the CVPR*, pages 2153–2162, 2023. 3

[41] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the ACM MM*, pages 788–796, 2021. 3

[42] Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, and Yonghong Tian. Heterogeneous relational complement for vehicle re-identification. In *Proceedings of the ICCV*, pages 205–214, 2021. 1, 3

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the ICCV*, pages 1116–1124, 2015. 2

[44] Wei Shi Zheng, Shao Gang Gong, and Tao Xiang. Associating groups of people. In *Proceedings of the BMVC*, pages 1–11, 2009. 2

[45] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI*, pages 13001–13008, 2020. 6

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the CVPR*, pages 2921–2929, 2016. 6

[47] Yi Zhou and Ling Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *Proceedings of the CVPR*, pages 6489–6498, 2018. 3