Highlights

Flare-Aware Cross-modal Enhancement Network for Multi-spectral Vehicle Re-identification

Aihua Zheng,Zhiqi Ma,Yongqi Sun,Zi Wang,Chenglong Li,Jin Tang

- A mutual flare mask prediction module that predicts the flare mask through flare-affected RGB and NI modalities.
- A flare-aware enhancement module that enhances masked RGB and NI features using flare-immunized TI information.
- A multi-modal consistency loss to enhance semantic consistency in multi-spectral vehicle under intense flare.
- A more realistic and comprehensive large-scale Wild Multi-spectral Vehicle Re-ID dataset WMVeID863.

Flare-Aware Cross-modal Enhancement Network for Multi-spectral Vehicle Re-identification

Aihua Zheng[a], Zhiqi Ma[b], Yongqi Sun[b], Zi Wang[b], Chenglong Li[a] and Jin Tang[b]

[a]*Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University , Hefei, 230601, China*

[b]*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology , Anhui University, Hefei, 230601, China*

ARTICLE INFO

Keywords:
Multi-spectral
Vehicle Re-identification
Cross-modal Enhancement
Intense Illumination

ABSTRACT

Multi-spectral vehicle Re-identification (Re-ID) aims to incorporate complementary visible and infrared information to tackle the challenge of re-identifying vehicles in complex lighting conditions. However, in harsh environments, the discriminative cues in RGB (visible) and NI (near infrared) modalities are significantly lost by the strong flare from vehicle lamps or the sunlight. To handle this problem, we propose a Flare-Aware Cross-modal Enhancement Network (FACENet) to adaptively restore the flare-corrupted RGB and NI features with the guidance from the flare-immunized TI (thermal infrared) spectra. First, to reduce the influence of locally degraded appearance by the intense flare, we propose a Mutual Flare Mask Prediction (MFMP) module to jointly obtain the flare-corrupted masks in RGB and NI modalities in a self-supervised manner. Second, to utilize the flare-immunized TI information to enhance the masked RGB and NI, we propose a Flare-aware Cross-modal Enhancement module (FCE) to adaptively guide feature extraction of masked RGB and NI spectra with the prior flare-immunized knowledge from the TI spectra. Third, to mine the common semantic information of RGB and NI, and alleviate the severe semantic loss in the NI spectra using TI, we propose a Multi-modality Consistency (MC) loss to enhance the semantic consistency among the three modalities. Finally, to evaluate the proposed FACENet while handling the intense flare problem, we contribute a new multi-spectral vehicle Re-ID dataset, named WMVEID863 with additional challenges, such as motion blur, huge background changes, and especially intense flare degradation. Comprehensive experiments on both the newly collected dataset and public benchmark multi-spectral vehicle Re-ID datasets verify the superior performance of the proposed FACENet compared to the state-of-the-art methods, especially in handling the strong flares. The codes and dataset will be released at this link.

1. Introduction

Multi-spectral vehicle Re-ID Li et al. (2020b); Zheng et al. (2023) endeavours to address the challenge of re-identifying vehicles in intricate lighting conditions by combining complementary RGB, near-infrared (NI) and thermal infrared (TI) data.

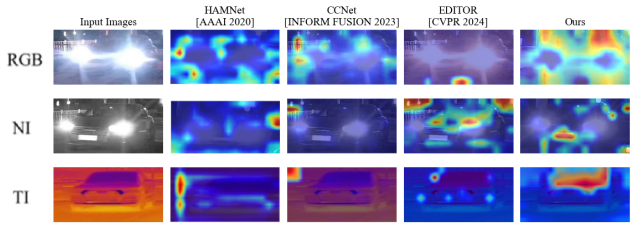


Figure 1: Feature map visualization of how the intense flares affect existing multi-modal Re-ID methods.

Despite the progress in both methodologies and datasets, the ubiquitous intense flare problem in real-life complex transportation systems is neglected, since both RGB and NI modalities are sensitive to intense flares such as vehicle lamps or strong sunlight. As a result, discriminative cues in RGB and NI images are partially lost, which significantly affects the modality fusion in existing multi-modal methods, as shown in Fig. 1. The most straightforward solution is to employ image restoration strategies Zamir et al. (2020); Singh et al. (2020) to restore the local region of the light-corrupted images. However, modelling a projection between light-corrupted and clean images requires a large amount of paired data, which is tedious at best and impossible at worst.

As a new challenge task in multi-spectral vehicle Re-ID, we observe that there are three crucial issues concerning the intense flare problem. **First**, to eliminate the impact of the local information degradation from the intense flare, it is straightforward to locate the flare-corrupted region in the images. Meanwhile, different modalities, such as RGB and NI, may produce discrepant degradation under the intense flare. Therefore, how to automatically and jointly predict the flare-corrupted local region in both RGB and NI modalities is essential for this task. **Second**, we observe that the thermal-infrared (TI) spectra is generally flare-immunized, which can provide critical information, especially for the

*Corresponding Author: Chenglong Li. Zhiqi Ma and Yongqi Sun are with equal contributions.

ahzheng214@foxmail.com (Aihua Zheng); doermzq0398@foxmail.com (Zhiqi Ma); yong-qi-sun@foxmail.com (Yongqi Sun); ziwang1121@foxmail.com (Zi Wang); lcl1314@foxmail.com (Chenglong Li); tangjin@ahu.edu.cn (Jin Tang)

ORCID(s): 0000-0002-7233-2739 (Chenglong Li)

flare-corrupted region in RGB and NI modalities. Therefore, how to utilize the flare-immunized TI information to guide the feature learning of the masked RGB and near-infrared (NI) spectra is crucial. **Third**, after the cross-modal enhancement, the guidance of the TI spectra may introduce unexpected modality-specific information to RGB and NI modalities, leading to a biased feature distribution. Therefore, intuitively enforcing the three modalities such as 3M loss Wang et al. (2022b) and CdC loss Zheng et al. (2023) may result in suboptimal performance. How to align the semantic information of enhanced RGB and NI features is essential for multi-spectral vehicle Re-ID.

To solve the above problems, we propose Flare-Aware Cross-modal Enhancement Net (FACENet) to adaptively restore the flare-corrupted RGB and NI features with the guidance of the flare-immunized TI spectra. **First**, to acquire regions that are partially impaired due to strong light interference, we propose a Mutual Flare Mask Prediction (MFMP) module. This module is designed to extract the areas affected by strong light in both RGB and NI spectra in a self-supervised manner. It then effectively combines them based on the extent to which they are influenced by the lighting conditions. **Second**, although we can relieve the influence of the intense flare from the masked images, the crucial local information is simultaneously missing with the masks. The TI spectra contains a wealth of information invariant to strong lighting. Given the premise of coarse alignment in multi-spectral image data, the TI spectra can serve as a guide when RGB and NI information are locally disturbed. Therefore, we propose a Flare-aware Cross-modal Enhancement (FCE) module to adaptively guide the feature extraction of the masked RGB and NI spectra with the prior flare-immunized knowledge from the TI spectra in a conditional learning manner.

Third, we observe that compared to the TI spectra generally producing the global information of the vehicles, the RGB and NI modalities contain more detailed semantic information that is beneficial to the Re-ID task. Meanwhile, during the inference process of the baseline, the NI spectra may negatively impact the inference results due to the loss of detailed semantic information caused by intense flare occlusion. Inspired by Wang et al. (2022b), we tend to apply consistency constraints to multi-spectra features to fully utilize the shared representative information within the multi-spectra. Unlike ordinary multi-modal vehicle Re-ID, the semantic loss in RGB and NI modalities caused by flares makes directly using the 3M loss for all modality information misleading for network optimization. Specifically, we propose a multi-modal consistency loss (MC Loss) to enhance semantic feature extraction, constrain the consistency between multi-modal spectra, and alleviate the severe semantic loss in the NI spectra. It is worth noting that when there is no intense flare, we can deactivate the FCE module, as shown in Fig. 2. In this way, the features of RGB and NI will remain unchanged.

In addition, existing multi-spectral vehicle Re-ID datasets such as Zheng et al. (2023); Li et al. (2020b) mainly focus on low illumination scenarios with still vehicles in a limited number of identities while without intense flares. Therefore, we contribute a more realistic large-scale **W**ild **M**ulti-spectral **Ve**hicle Re-**ID** dataset WMVeID863 in a complex environment. WMVeID863 is captured with vehicles in motion with more challenges, such as motion blur, huge background changes, and especially intense flare degradation from car lamps, and sunlight, with a large quantity of identities. We conclude our main contributions as follows:

- We are the first to launch the intense flare issue in vehicle Re-ID and propose the Flare-Aware Cross-modal Enhancement Network (FACENet) to adaptively restore the flare-corrupted RGB and NI features with guidance from the flare-immunized TI spectra.
- To reduce the influence of locally degraded appearance by the intense flare, We propose the Mutual Flare Mask Prediction (MFMP) module, which enables the interaction between the flare-affected features in RGB and NI modalities in a self-supervised manner to predict the flare mask.
- To enhance the masked RGB and NI features by the flare-immunized TI information, We propose the Flare-aware Cross-model enhancement (FCE) module to guide local feature learning of RGB and NI branches with prior information from the TI branch.
- To explore the semantic consistency in multi-spectral vehicle re-identification under intense flare, we propose enhancing multi-modal consistency (MC) loss between multiple modalities by minimizing the KL divergence.
- We contribute a more realistic and comprehensive large-scale Wild Multi-spectral Vehicle Re-ID dataset WMVeID863 with more challenges including intense flare to evaluate the effectiveness of FACENet.

2. Related Work

2.1. Multi-spectral Re-ID

Recent methods targeting the visible Re-ID task have become well-established, typically categorized into attribute-based and feature-based approaches Zhou et al. (2020); Chen et al. (2020); Zhou et al. (2020); Li et al. (2022a); Pang et al. (2023); Shen et al. (2023b,c,a), viewpoint-based methods Wu et al. (2021); He et al. (2021); Zheng et al. (2022); Pang et al. (2022), and distance metric learning strategies Liu et al. (2016); Li et al. (2022b); Teng et al. (2023). Multi-spectral re-identification leverages data from various spectra, including RGB, NI, and TI, to improve feature fusion and enhance performance in re-identification tasks, particularly under challenging lighting conditions. Multi-spectral Re-ID has been advanced primarily through CNN-based Li et al. (2020b); Zheng et al. (2021); Wang et al. (2022b); Guo et al. (2022); Zheng et al. (2023);

Kamenou et al. (2023); He et al. (2023); Pang et al. (2024) and transformer-based Wang et al. (2024); Zhang et al. (2024) approaches, which leverage the complementary information of different spectra to improve robustness in varying environments. However, the high computational resource demands remain a significant challenge, particularly with the increased training costs for multi-spectral ReID tasks. Some work focuses on optimizing the computational efficiency of transformers to mitigate these issues Guo et al. (2019). Li et al. (2020b) construct the first multi-spectra vehicle Re-ID dataset RGBN300 (visible and near-infrared) and RGBNT100 (visible, near-infrared, and thermal infrared) to solve this problem and propose a baseline method HAMNet to effectively fuse multi-spectra information through CAM (Class Activation Map) Zhou et al. (2016). Guo et al. (2022) introduce the Generative and Attentive Fusion Network (GAFNet), which incorporates generated transitional modality images and an attentive feature fusion module, addressing the limitations of existing methods in handling the disparity between different modalities. Pang et al. (2024) propose a novel cross-modal similarity learning framework that emphasizes shape information while also encouraging the learning of modality-invariant and identity-related features. Wang et al. (2024) propose a recurrent token permutation framework, further leveraging multi-spectral features extracted by Transformer to fully exploit the local details of different spectra, generating more discriminative multi-spectral features. Zhang et al. (2024) construct a token selection framework that adaptively selects object-centered tokens based on spatial and frequency information, and further reduces the influence of background through hierarchical frequency filtering. Therefore, it is ineffective to directly employ the existing multi-spectral Re-ID methods for the intense flare issue.

2.2. Cross-modal Enhancement

Cross-modal enhancement is committed to boosting one modality from other modalities. Wang et al. (2023a) propose to enhance text representations by integrating visual and acoustic information into a language model. Wang et al. (2022b) propose to exchange the information between modalities to absorb complementary information from other modalities while simultaneously maintaining the modality-specific information. Fang et al. (2022) present the Spatial-Spectral Enhancement Module designed for cross-modal information interaction in deep neural networks, specifically enhancing the spatial and spectral representations of hyper-spectral and LiDAR data. Cheng et al. (2020) propose a self-supervised framework with a co-attention mechanism to learn generic cross-modal representations to solve the pretext task of audiovisual synchronization. Mercea et al. (2022) proposes to learn multi-modal representations from audiovisual data using cross-modal attention for the alignment between the audio and visual modalities. Wang et al. (2023b) develop the Cross-modal Enhancement Network (CENet) model for multimodal sentiment analysis which significantly improves text representations by integrating visual and acoustic information into a language model, and effectively aligns and fuses verbal and nonverbal modalities. Wang et al. (2023d) incorporate cross-modal information to utilize the complementary information of multi-source data and propose RSRNet. Wang et al. (2023c) introduce the Context-Aware Proposal–Boundary (CAPB) network for audiovisual event localization in videos, addressing the limitations of previous methods by incorporating proposal-level feature encoding and a local-global context encoder. Jiang et al. (2023) introduce the Random Online Hashing (ROH) method, effectively resolving the scalability and semantic information preservation challenges in supervised cross-modal hashing for large-scale multimedia databases. Zheng et al. (2021) integrate the multi-modal features at part level to capture the complementary local information among modalities. However, existing cross-modal enhancement methods are limited while facing flare problems on certain spectra, since the intense flare introduces large influence and diversity in different modalities. Therefore, directly performing cross-modal enhancement will bring noise and reduce the representation ability of RGB and NI spectra due to multi-spectra heterogeneity.

2.3. Image Flare Removal

Flare is an optical phenomenon that typically occurs when strong light is scattered or reflected within an optical system, resulting in radiant bright areas and light spots in images. This effect is particularly noticeable during nighttime photography. Flare often blurs details around light sources, significantly affecting visual quality and performance. While professional lenses can somewhat mitigate flare, they cannot fully eliminate the issue. Feng et al. (2023) primarily focus on removing flare artifacts caused by under-display cameras. Qiao et al. (2021) propose a deep framework with light-source-aware guidance for single-image flare removal (SIFR). To guide the removal of flare artifacts based on the detected light source and flare regions, effectively eliminating various types of flares from the image. Dai et al. (2023) propose a new end-to-end pipeline to preserve the light source while removing lens flares. In vehicle Re-ID tasks, flare is often present on the vehicle itself, such as reflections on the car body and headlights, which severely degrade the vehicle's detail in images, thereby impacting the accuracy of Re-ID tasks. Therefore, developing effective flare removal methods is crucial.

3. Flare-Aware Cross-modal Enhancement Network

To make full use of the flare-immunized information in the TI spectra and the residual effective information in the flare-corrupted RGB and NI spectra, we propose the **F**lare-**A**ware **C**ross-modal **E**nhancement **N**etwork (FACENet) for multi-spectral vehicle Re-ID, as shown in Fig. 2. First of all, we propose the Mutual Flare-aware Mask Prediction

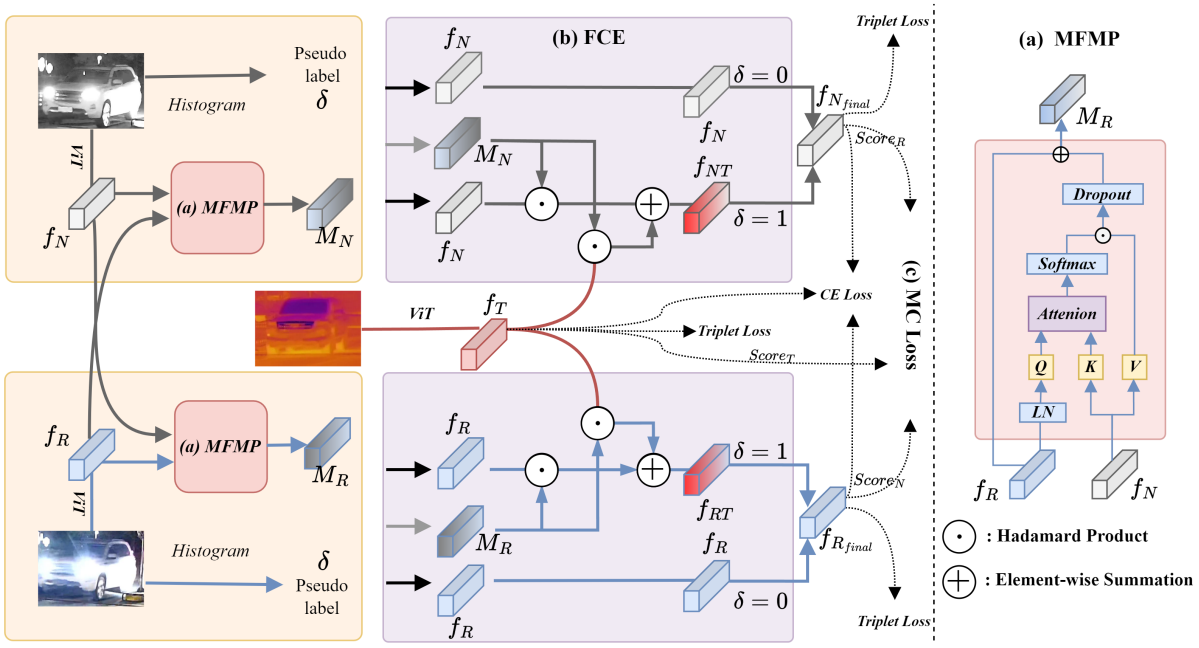


Figure 2: The overall structure of the proposed method FACENet. First, we obtain the pseudo-label according to the histogram. We extract features of multi-spectra vehicle images by three individual ViTs to obtain features f_R, f_N, f_T of each spectra. f_R and f_N are fed into the MFMP to obtain the flare masks for each spectra. Finally, the flare-immunized f_T is fed into a flare-aware cross-modal enhancement module to guide f_R and f_N with the aid of the predicted flare masks and the pseudo-labels. The overall training stage for Re-ID is supervised by Cross-Entropy loss and Triplet loss on three individual spectra, along with the proposed MC loss on RGB and NI spectra. In the testing stage, we concatenate the multi-branch features for final distance measuring.

(MFMP) module, which extracts regions heavily affected by the flare through inter-modal self-supervised mask prediction. Then, we propose the **F**lare-aware **C**ross-modal **E**nhancement (FCE) module to use the local feature in TI spectra as a condition to guide the local feature recovery in RGB and NI spectra. Moreover, to minimize the similarity between predictions of RGB and NI branch, and alleviate the severe semantic loss in the NI spectra. we design **M**ulti-Modality **C**onsistency Loss (MC Loss) for training. We will elaborate on the main components below.

3.1. MFMP: Mutual Flare Mask Prediction Module

We observe that both RGB and NI modalities are susceptible to flare, introducing flare noise into the results extracted by the model. To reduce the influence of intense flare on local appearance, we first propose the Mutual Flare Mask Prediction (MFMP) module to jointly obtain the flare-corrupted masks in RGB and NI modalities in a self-supervised manner. The MFMP aims to predict the regions affected by flare through a self-supervised approach by integrating features from both RGB and NI to produce a more effective flare-aware mask.

While a flare mask can be derived from a feature map of a single spectra, this approach overlooks the fact that flares affect both RGB and NI spectra. Therefore, training the flare mask separately is suboptimal. To address this limitation, we propose integrating features from both RGB and NI to identify the local areas affected by flares accurately. As shown in Fig. 2, given features from RGB and NI branches f_R and f_N with the shape of $B \times (N+1) \times D$, we split f_R and f_N into class tokens and patch tokens, f_R^{Class} and $f_N^{Class} \in \mathbf{R}^{B\times 1\times D}$, f_R^{Patch} and $f_N^{Patch} \in \mathbf{R}^{B\times N\times D}$. To better capture the flare-corrupted regions in the spectra, we fuse the patch tokens that contain more detailed information from both modalities, to generate a flare-aware mask. Notably, to ensure the accuracy of the mask positions, we generate the flare-aware mask M_i for each of the RGB and NI modalities separately, where $f_i, i \in \{R, N\}$ indicates mask from RGB or NI spectra. The following is an example of the flare-aware mask for the RGB spectra.

We first perform matrix operations on the obtained f_R^{Patch} and f_N^{Patch} separately to generate query matric Q from f_R^{Patch} and key matric K and values matric V from f_N^{Patch}, as shown in Eq. (1).

$$Q_R = W_Q \odot f_R^{Patch}, K_N = W_K \odot f_N^{Patch}, V_N = W_V \odot f_N^{Patch}, \tag{1}$$

where W denotes the weight matrix that equalizes the feature dimensions. We utilize the excellent attention mechanism to obtain a more comprehensive flare region. First, we

compute the similarity between the query matrix Q and the key matrix K, then normalize it with softmax and multiply it by the value matrix V. Finally, the resulting attention output is added to the RGB, yielding the RGB flare mask $M_R \in \mathbf{R}^{B\times N\times D}$. In the same manner, we can further obtain the NI flare mask M_N, as shown in Eq. (2).

$$\begin{aligned} M_R &= \text{softmax}\left(\frac{Q_R K_N^T}{\sqrt{d_C}}\right) V_N \oplus f_R^{Patch}, \\ M_N &= \text{softmax}\left(\frac{Q_N K_R^T}{\sqrt{d_C}}\right) V_R \oplus f_N^{Patch}, \end{aligned} \tag{2}$$

where d_C denotes the feature dimension and $\oplus$ stands for matrix addition. Based on the high correlation of flare information between the two modalities, we accurately capture the flare regions in both RGB and NI modalities using the cross-attention mechanism. Considering the differences in information loss between different modalities, we generate representative flare masks M_R and M_N for the RGB and NI modalities, respectively.

3.2. FCE: Flare-aware Cross-modal Enhancement Module

Given the flare mask obtained from the MFMP module, a crucial issue for multi-spectra vehicle Re-ID under intense flare conditions is how to recover the flare-corrupted features in RGB and NI spectra. Inspired by conditional learning Li et al. (2020a); Wang et al. (2018), which incorporates additional information such as labels, context, or prior knowledge, into the learning process to help the model better understand the input data and the specific task requirement, we propose a Flare-aware Cross-modal Enhancement (FCE) module to use the flare-immunized TI spectra to guide the feature learning of the flare-corrupted RGB and NI spectra.

To supervise the FCE module in more accurately repairing flare-corrupted regions, we derive a pseudo-label that indicates whether a particular image is affected by flares based on the histogram of the image. Specifically, we calculate the percentage of pixels in the image with pixel values between 250 and 255 relative to the total number of pixels in the image. To obtain the pseudo-label, we manually select a bar 5%. If the percentage is greater than the bar, which means there are more than 5% amount of pixels with high value in an image, the image is considered a flare-corrupted sample. The pseudo-label δ is also utilized in the FCE module to exclude samples that are not significantly affected by flares, as shown in Eq. (3).

$$\begin{aligned} &\text{if } \frac{\text{count}(250 < pv < 256)}{H \times W} > 5\%, \\ &\quad \text{then } \delta = 1, \\ &\quad \text{else } \delta = 0, \end{aligned} \tag{3}$$

where pv indicates pixel value, and $H * W$ indicates the amount of pixels in an image.

The FCE module aims to guide the flare-susceptible RGB and NI spectra with the flare-immunized information in the TI spectra. Therefore, we consider the local features masked by flare-mask prediction from MFMP in the TI spectra as the condition. And the masked features in RGB and NI as the input of conditional learning. Note that FCE is performed respectively on RGB and NI spectra. Taking RGB as an example of FCE:

$$f_{RT} = C(f_R^{Class}, f_T^{Patch} \odot M_R \oplus (f_R^{Patch} \odot \tilde{M}_R)), \tag{4}$$

where C is the operation used in PyTorch to concatenate tensors, M_R is the flare mask provided by MFMP module, and $\tilde{M}_R$ is the negation of M_R.

Fig. 2 illustrates the objective of FCE, which aims to guide the learning of flare-corrupted f_i from flare-immunized features f_T, where $f_i, i \in \{R, N\}$ indicates features from RGB or NI spectra. Specifically, the TI branch provides flare-immunized knowledge in the form of f_T, and FCE seeks to restore flare-corrupted information in f_i by utilizing the prior knowledge. After obtaining the mask M_i from MFMP, we perform element-wise production on feature $f_i^{Patch} \odot \tilde{M}_i$ to obtain the local features that are not affected by flare while performing $f_T^{Patch} \odot M_i$ to obtain the corresponding flare-immunized local feature in the TI spectra. At last, the two local features are element-wise summed as the final feature representation for conditional learning of the i-th branch. **Note that** the FCE module is designed for the flare problem, for certain samples without severe flare degradation, we utilize pseudo-label to skip from FCE, and use the feature f_R, f_N from backbone.

3.3. MC Loss: Multi-modality Consistency Loss

To fully utilize the complementary information of RGB and NI spectra and avoid the negative impact on the final result due to severe semantic loss in the NI spectra, we propose a method to align the semantic consistency of RGB and NI, as well as NI and TI simultaneously. Inspired by Wang et al. (2022b), we propose to enforce the semantic consistency of prediction scores between RGB, NI and TI spectra.

Specifically, we propose to use KL-divergence to pairwise enhance the feature distribution between specific modalities. This aims to minimize the similarity between the classification results of RGB and NI and compensate for the excessive semantic degradation of NI. Phuong and Lampert (2019) verify that implying consistency loss on the later prediction layer of deep networks works better. The classifiers from later stages have larger capacity and predicted scores can more accurately reflect feature information. Given classification scores of three modalities $\{S_R, S_N, S_T\}$ from corresponding branches, we employ KL-divergence Kullback and Leibler (1951) to compute the distance.

We obtain the smaller value from the KL distances between RGB, NI and TI, as the proposed multi-modality

consistency loss:

$$KL(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \tag{5}$$

$$\begin{aligned}\mathcal{L}_{MC} = &\min\Big(KL(S_R \parallel S_N), KL(S_N \parallel S_R)\Big) \\ &+ \min\Big(KL(S_N \parallel S_T), KL(S_T \parallel S_N)\Big),\end{aligned} \tag{6}$$

where S_i represents the final classifier prediction score for modality i, $i \in \{R, N, T\}$.

Note that, by constraining the consistency of the distribution results between RGB and NI modalities, the two branches can enhance the information pertinent to classification through unified prediction outcomes. This information may be complementary or analogous, which does not contradict the notion of complementarity at the feature level. By constraining the consistency of the distribution between NI and TI modalities, the network can further mine the common semantic information of NI and TI, while compensating for the semantic degradation of the NI spectra. By enforcing the MC loss, the network can optimize the extraction of richer semantic information, thereby achieving robust multi-modal representation.

3.4. Overall Loss Function

In the training stage, our method is under the joint supervision of identify (ID) loss, Triplet Loss Hermans et al. (2017), and the proposed MC loss.

To supervise the optimization of the MFMP module, we also ensure the consistency of the mask prediction regions for the RGB and NI modalities by calculating the KL divergence of the distribution results of flare masks n and r. The loss is defined as:

$$\mathcal{L}_{KL} = \min((KL(S_{M_R} \parallel S_{M_N}), KL(S_{M_N} \parallel S_{M_R})). \tag{7}$$

The ID loss is calculated as the cross entropy between the predicting probability and is denoted as $\mathcal{L}_{id}$.

$$\mathcal{L}_{id}(y) = -\sum_{i=1}^{N}\sum_{j=1}^{C} \hat{y}^n \log p(y_j^n), \tag{8}$$

where $\hat{y}^n$ is a one-hot matrix indicates the identity label of the n-th sample, where $\hat{y}_i^n = 0, i \in \{0, 1, ..., C\}$ except $\hat{y}_c^n = 1$.

To perform hard sample mining in a batch, we adopt Triplet loss Hermans et al. (2017) denoted as $\mathcal{L}_{Tri}(f)$, where f indicated the input feature:

$$\begin{aligned}\mathcal{L}_{Tri}(f) = \sum_{i=1}^{P}\sum_{a=1}^{K}[m + &\overbrace{\max_{p=1,...,K} D(f_a^i, f_p^i)}^{hardest\ positive} \\ &- \overbrace{\min_{n=1,...,K} D(f_a^i, f_n^i)}^{hardest\ negative}.\end{aligned} \tag{9}$$

We train the multiple branches with MC Loss, KL loss, ID loss and Triplet loss, the overall loss is defined as:

$$\begin{aligned}\mathcal{L}_{all} = &\mathcal{L}_{id}(Score_R) + \mathcal{L}_{id}(Score_N) + \mathcal{L}_{id}(Score_T) \\ &+\mathcal{L}_{Tri}(f_{R_{final}}) + \mathcal{L}_{Tri}(f_{N_{final}}) + \mathcal{L}_{Tri}(f_{T_{final}}) \\ &+\mathcal{L}_{KL} + \mathcal{L}_{MC}.\end{aligned} \tag{10}$$

The ID loss in FACENet is effective in distinguishing between identities, while the Triplet loss optimizes the inner-class and intra-class distances. In addition, our proposed MC loss aligns the distribution of classifier score results and adjusts the modality distance, resulting in a more robust feature representation for Re-ID.

4. WMVeID863: Wild Multi-spectral Vehicle Re-identification Dataset

To evaluate the proposed method while handling the intense flare issue, we contribute a Wild Multi-spectral Vehicle re-IDentification Dataset, named WMVeID863.

4.1. Data Acquisition and Processing

4.1.1. Data Acquisition

To ensure the diversity of vehicle data and environments, we selected two places with vehicles passing through on campus to record the original videos, with each place containing 8 viewpoints. WMVeID863 dataset is collected on campus by triplicated cameras to simultaneously record RGB, NI, and TI video data in both day and night with a two-month time span. The dataset is captured in four different weather conditions including sunny, cloudy, windy, and hot days in both morning and night respectively in videos. The collection period for the data was varied, encompassing morning, afternoon, and evening times to ensure a diverse and extensive dataset for research in vehicle Re-Identification. The raw data contributes to 37 hours of videos in total, with RGB, NI, and TI, respectively.

4.1.2. Data Processing

Following the acquisition of the raw video footage, we refined and adapted this data to align with the specific needs of the multi-modal Re-Identification task. The RGB and NI images are captured by the paired 360 D866 cameras with a resolution of 2560×1440 in 15 fps, and the TI images are captured by a DALI thermal telescope and a HIKVISION DS-7800HQH-K1 recording device with a resolution of 1280×720 in 15 fps. Firstly, we manually adjusted and synchronized the resolution of the RGB and NI images to correspond with that of the TI images for consistent quality. Then, vehicle video clips with nearly frontal viewpoints were chosen for identity labeling, based on their license plate numbers. Subsequently, we select the corresponding multi-spectral video clips based on their ID and obtain image data by extracting frames at specific intervals. Last, we crop each sample to obtain the bounding boxes of the vehicles for Re-ID.

4.2. Data Description

WMVeID863 contains 4709 image triplets of 863 IDs of 8 camera views, the number of image triplets for each vehicle varies from 1 to 39. The dataset distribution of the number of scenes across the number of identities is shown in Fig. 3. We randomly select 603 IDs with 3482 image triplets for training and 260 IDs with 1226 image triplets for testing. The split ratio between the training and testing sets is approximately 3:1. The gallery samples consist of 260 IDs with 1226 image triplets. The query samples are randomly selected from the gallery samples, consisting of 210 IDs with 959 image triplets.

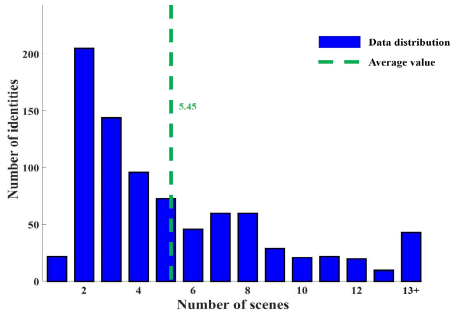


Figure 3: Distribution for the number of scenes across the number of identities in WMVeID863.

4.3. Special Challenges

The WMVeID863 dataset introduces one distinct and challenging element absent in current vehicle ReID datasets. **Intense illumination.** In real-world scenarios, vehicles appear in various settings. A common yet often overlooked challenge arises in certain specific scenes where the camera is aligned with the vehicle lights or reflects off certain light sources. As a result, the captured images contain limited visible information about the vehicles. WMVeID863 dataset includes scenarios where daylight reflects off the car's paint, nighttime street lights cause reflections on the vehicle, and interference from strong light emitted by car headlights. Strong light interference can lead to the degradation of image information and a decline in image quality. This challenge is difficult to address for common re-identification systems that rely on visible light images. However, multi-spectral vehicle re-identification systems have a natural advantage due to the strong light invariance of the thermal infrared modality.

4.4. Comparison with Existing Datasets

We compare our WMVeID863 dataset with existing multi-spectral vehicle ReID datasets, as shown in Fig. 4, in addition to the common challenges like view changes (VC), partly occlusion (PO), and various resolutions (VR), WMVeID863 introduces new challenges like intense light (IL). RGBN300 dataset consists of 300 IDs and a substantial total of 100,250 images, captured across 2004 different scenes. It, however, does not specify the presence of Vehicle-in-Motion ViM or intense light IL, which are crucial

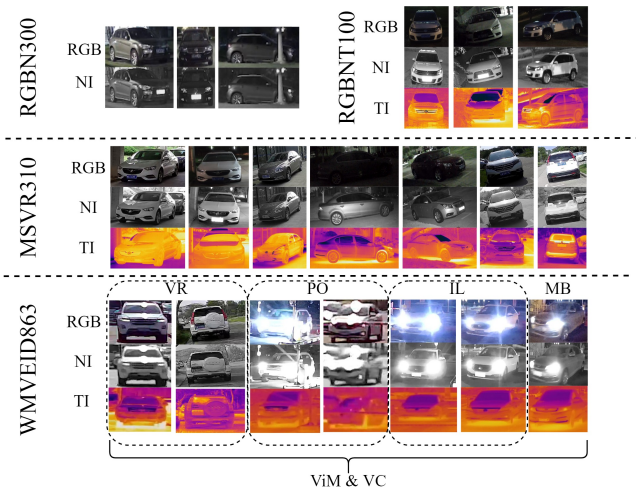


Figure 4: Challenge comparison between the proposed WMVeID863 and existing multi-spectral datasets. WMVeID863 presents a more challenging and realistic multi-spectral vehicle Re-ID scenario.

Table 1
Comparison with existing multi-spectra vehicle Re-ID datasets.

Benchmarks	IDs	Images	Scenes	ViM	IL
RGBN300	300	100250	2004	-	-
RGBNT100	100	17250	690	-	-
MSVR310	310	6261	2061	-	-
WMVeID863	863	14127	4709	✓	✓

factors in vehicle re-identification scenarios. RGBNT100 dataset is smaller in scale and it does not provide data on ViM and IL. MSVR310 offers a moderate-sized dataset. However, similar to the above datasets, it lacks information on ViM and IL. Different from the existing datasets that capture the vehicles in still, we capture the vehicles in motion, thereby bringing the additional challenge of motion blurring (MB). Our dataset, WMVeID863, significantly expands on the existing datasets by including 863 IDs and 14,127 images, spread over 4,709 scenes. What sets WMVeID863 apart is its inclusion of Vehicle-in-Motion and Illumination Variation. These features make WMVeID863 a more comprehensive and challenging dataset, better suited for evaluating vehicle re-identification algorithms under various real-world conditions. The WMVeID863 dataset stands out in its comprehensive coverage of scenarios, number of identities, and especially in its incorporation of dynamic factors like ViM and IL. This makes it a valuable resource for advancing research in the field of multi-spectra vehicle re-identification, providing a more realistic and challenging platform for the development and evaluation of state-of-the-art algorithms.

Comparing with existing multi-spectra Re-ID datasets as shown in Table 1, WMVeID863 has two major advantages:

Table 2
Comparison with State-of-the-Art methods for multi-spectra Re-ID on WMVeID863 (in %). The best three scores are marked in red, blue and green (in %).

	Method	Ref	Bac	mAP	R-1	R-5	R-10
Single-Modal	DenseNet	CVPR'17	CNN	42.9	47.9	61.9	68.7
	ShuffleNet	CVPR'18	CNN	34.2	37.2	52.3	58.9
	MLFN	CVPR'18	CNN	43.7	47.0	61.7	69.8
	HACNN	CVPR'18	CNN	46.9	48.9	66.9	73.8
	StrongBaseline	CVPR'19	CNN	51.1	55.7	69.8	74.7
	OSNet	ICCV'19	CNN	42.9	46.8	61.9	69.4
	AGW	TPAMI'21	CNN	30.3	35.3	43.3	46.5
	ViT	Arxiv'20	ViT	66.5	73.7	80.6	82.6
	TransReID	ICCV'21	ViT	67.0	74.7	79.5	82.4
	PFD	AAAI'22	ViT	50.2	55.3	69.8	75.3
Multi-Modal	HAMNet	AAAI'20	CNN	45.6	48.5	63.1	68.8
	PFNet	AAAI'21	CNN	50.1	55.9	68.7	75.1
	IEEE	AAAI'22	CNN	45.9	48.6	64.3	67.9
	CCNet	INFFUS'23	CNN	50.3	52.7	69.6	75.1
	TOP-ReID	AAAI'24	ViT	67.7	75.3	80.8	83.5
	EDITOR	CVPR'24	ViT	65.6	73.8	80.0	82.3
	FACENet	Ours	ViT	69.8	77.0	81.0	84.2

- WMVeID863 is the largest multi-spectral vehicle Re-ID dataset so far with more realistic challenges compared to MSVR310 Zheng et al. (2023), RGBN300 Li et al. (2020b) and RNBNT100 Li et al. (2020b).
- It consists of more background changes and realistic challenges like intense light and motion blur, which are significantly ignored in existing multi-spectral vehicle Re-ID datasets.

5. Experiments

Implementation Details.

We employ ViT Dosovitskiy et al. (2020) pretrained on ImageNet-21K Deng et al. (2009) as our backbone. The implementation platform is PyTorch 1.10.1 with one NVIDIA RTX 4090 GPU. We use SGD Singarimbun et al. (2019) optimizer to optimize the network with the initial learning rate as 3×10^{-3} with a weight decays of 1×10^{-4} at total 120 epochs. The images are resized to 256×128 as the input of the network. In the training phase, the features of multiple spectra are trained separately without parameter sharing and jointly supervised by Cross-Entropy loss and Triplet Loss, while R and N spectra pairs, and N and T spectra pairs are trained by an additional proposed MC loss. In evaluation, we concatenate the features extracted from three parallel branches as the final representation for a sample, for the feature matching.

Evaluation Protocols.

Following the protocols in Zheng et al. (2015), we employ the commonly used metrics in Re-ID task mean Average Precision (mAP) and Matching Characteristic curve (CMC) to evaluate our method, where R-*n* indicates the first *n* closest samples to the query sample with the same ID from different cameras.

5.1. Comparison with State-of-the-Art Methods

We evaluate the performance of our proposed FACENet method against a variety of state-of-the-art single-modal and multi-modal methods for multi-spectra re-identification on the WMVeID863 dataset as shown in Table 2. For single-modal methods, we simply extend them to multi-modality versions for comparison by expanding the single-branch network into multiple branches and adding a corresponding loss function to each branch.

Most existing single-modal methods Huang et al. (2017); Zhang et al. (2018); Chang et al. (2018); Li et al. (2018); Luo et al. (2019); Ye et al. (2021); Wang et al. (2022a) fail to fully utilize the complementary information of multi-modal data, even with strong feature extraction capabilities in a single modality, especially in overcoming the effects of strong illumination. Note that the transformer-based methods He et al. (2021); Dosovitskiy et al. (2020) significantly outperform CNN-based methods, the reason is that the patch-based transformers discover both global and local parts of an image, thus can better model the connection between the flare-corrupted region and flare-immunized region to suppress the influence of intense flare. Our FACENet significantly outperforms the second-best method TOP-ReID Wang et al. (2024) for 2.2% in mAP and 1.7% in R-1. This demonstrates the superiority of FACENet in dealing with multi-spectra scenarios. In addition, it not only surpasses the current best multi-modal method, but also the strongest single-modal methods like TransReID He et al. (2021) and ViT Dosovitskiy et al. (2020).

Existing multi-modal methods aim to fuse complementary features, however, cannot recover the information corrupted by the intense flare. By contrast, FACENet focuses on utilizing the flare-immunized information to guide the feature learning of flare-susceptible spectra, thus leading to the best overall feature representation.

Table 3
Ablation study of FACENet on WMVeID863 (in %).

Method	mAP	R-1	R-5	R-10
Baseline	66.5	73.7	80.6	82.6
+ FCE	68.2	75.4	80.6	83.3
+ FCE + $\mathcal{L}_{MC}$	69.0	76.3	80.8	83.1
+ MFMP + FCE	69.2	76.3	81.3	83.6
+ MFMP + FCE + $\mathcal{L}_{MC}$	69.8	77.0	81.0	84.2

5.2. Ablation Study

We evaluate the contribution of the three main components, mutual flare mask prediction module (MFMP), flare-aware cross-modal enhancement module (FCE), and multi-modality consistency loss (MC loss), as shown in Table 3.

The baseline is evaluated as ViT with ID loss and Triplet loss. First, we introduce the flare-aware cross-modal enhancement module (FCE) into the baseline. At this stage, we directly use f_R and f_N, output from the RGB and NI backbone branches respectively, as flare masks input to FCE. However, this does not lead to significant improvement. The primary reason is that the flare masks predicted solely by RGB and NI features capture only a small portion of the flare-corrupted regions. Then, we introduce our proposed multi-modality consistency loss (MC loss $\mathcal{L}_{MC}$) into the flare-aware cross-modal enhancement module (FCE), which leads to a significant improvement. By introducing the proposed mutual flare mask prediction module (MFMP), the overall mAP and R-1 scores improve, demonstrating the effectiveness of MFMP in predicting the overall flare regions. Finally, by using our proposed three main components on the backbone network, the overall experimental results achieve optimal performance.

5.3. Evaluation on MFMP and FCE

To demonstrate the effectiveness of our proposed MFMP and FCE modules in handling intense flare problems, we compare our method with two flare removal methods: flare source removal and flare halo removal Dai et al. (2023). Note that flare source removal is an image-level operation that directly sets flare pixels to zero to suppress flares, while flare halo removal Dai et al. (2023) is an image restoration method to synthesize de-flared images. We replace the original images with the ones generated by the model for training and testing. The comparison results are shown in Table 4. For a fair comparison, we remove the MC loss in FACENet. Although flare source removal and flare halo removal Dai et al. (2023) slightly improve the accuracy of some content, they reduce flare corruption by suppressing background noise and restoring information based on randomly added flares, leading to limited performance improvement. By jointly predicting flare masks and locally guiding flare-immune information to RGB and NI, our FACENet significantly outperforms both flare source removal and flare halo removal methods.

Table 4
Comparison with the flare removal methods to decrease the influence of intense extra light (in %).

	mAP	R-1	R-5	R-10
Baseline	66.5	73.7	80.6	82.6
+ Flare Source Removal	66.3	74.7	79.7	82.6
+ Flare Halo Removal	66.7	75.1	80.2	82.4
+ MFMP + FCE	**69.2**	**76.3**	**81.3**	**83.6**

Table 5
Comparison with existing modality relationship loss functions (in %).

	mAP	R-1	R-5	R-10
FACENet w/o $\mathcal{L}_{MC}$	69.2	76.3	**81.3**	83.6
+ $\mathcal{L}_{3M}$	68.1	75.8	80.1	81.7
+ $\mathcal{L}_{HC}$	68.4	75.3	80.8	82.7
+ $\mathcal{L}_{CdC}$	68.5	75.4	80.8	83.5
+ $\mathcal{L}_{MC}$ (Ours)	**69.8**	**77.0**	81.0	**84.2**

5.4. Evaluation on MC Loss: $\mathcal{L}_{MC}$

5.4.1. Comparing with the state-of-the-art loss functions.

We further compare our MC loss with the state-of-the-art loss functions as shown in Table 5.

3M loss Wang et al. (2022b) enforces TI differently from the TI-guided RGB and NI features, which suppresses the effectiveness of MFMP and FCE modules. CdC loss Zheng et al. (2023) enforces RGB and NI more similar to TI features, thus losing their original discriminative power. HC loss Zhu et al. (2020) is designed for cross-modal Re-ID by constraining the consistency of RGB and NI spectra. Implying HC loss in our task explores the common effective features in RGB and NI, therefore outperforming 3M loss and CdC loss. However, the performance is still behindhand, since HC loss aims to maximize the similarity of modality centers, which may excessively focus on the features that are enhanced by the TI spectra. By contrast, the proposed MC loss explores the semantic similarity according to prediction scores of RGB and NI spectra, thus leading to a significant performance improvement.

5.4.2. Hyper-parameter Analysis

Following Multi-Stream Network (MSN) in HAM-Net Li et al. (2020b), we employ Triplet loss and Cross Entropy loss for each stream and set the weight of each loss to 1.0. For the proposed MC loss, we empirically select the weight by combining baseline and MC loss, as shown in Fig. 5.

Fig. 5 presents an insightful analysis of the impact of different weight settings for the MC loss on the performance of the vehicle re-identification system. The results are shown in terms of mean Average Precision (mAP) and Rank-1 (R-1) accuracy, with weights varying from 0.2 to 2.0. A key observation from this analysis is the optimal performance achieved at a weight setting of 1.0 for the MC loss. In this

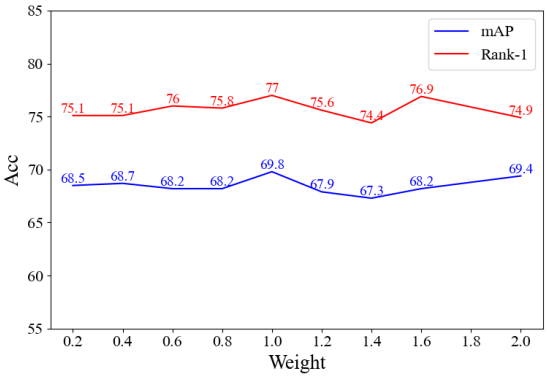


Figure 5: Analysis of weight settings of MC loss (in %).

Table 6
Experimental results of MC loss performing on features $\mathcal{L}_{MC}(F)$ and prediction results $\mathcal{L}_{MC}(P)$ (in %).

	mAP	R-1	R-5	R-10
FACENet w/o $\mathcal{L}_{MC}$	69.2	76.3	81.3	84.2
+ $\mathcal{L}_{MC}(F)$	69.1	76.2	80.1	82.7
+ $\mathcal{L}_{MC}(P)$	69.8	77.0	81.0	84.2

setting, the system attains the highest mAP of 69.8% and the highest R-1 accuracy of 77.0%.

5.4.3. *Plugging into Feature & Score*

We propose the MC loss to ensure that the predictions of the three modality branches are accurate and consistent. By constraining the consistency of the distribution results among the modalities, these three branches enhance the information relevant to classification through unified prediction outcomes. We perform MC loss on feature and classification scores to compare the performance. Experimental results, as presented in Table 6, suggest that enforcing feature-level consistency can inadvertently compromise the integrity of complementary information, culminating in a decline in performance. In contrast, imposing consistency on the level of prediction results sustains the complementary nature of feature-level information.

5.4.4. *Plugging to state-of-the-art multi-modal Re-ID methods.*

To verify the effectiveness of the proposed MC Loss $\mathcal{L}_{MC}$ in enhancing vehicle re-identification performance, we conduct comprehensive experiments on the WMVEID863 dataset. These experiments involve integrating $\mathcal{L}_{MC}$ with existing multi-modal Re-ID methods and assessing their performance improvements. A key observation from the results in Table 7 is the consistent improvement across all metrics when $\mathcal{L}_{MC}$ is added to existing methods. Notably, the integration of $\mathcal{L}_{MC}$ with CCNet exhibits the most significant enhancement, with the mAP increasing from 50.3% to 60.6%, and R-1 accuracy improving from 52.7% to 68.5%. This represents the highest improvement in both mAP and R-1 metrics across all tested methods.

In addition to the highest overall improvements, it is also worth noting the exceptional individual value improvements, such as the improvement in R-5 and R-10 accuracies with CCNet Zheng et al. (2023), reaching up to 79.2% and 82.4%, respectively. These results validate the capability of our proposed $\mathcal{L}_{MC}$ in significantly enhancing the performance of multi-modal vehicle re-identification systems.

Table 7
Experimental results of plugging the proposed MC Loss $\mathcal{L}_{MC}$ into existing multi-modal Re-ID methods on WMVEID863 (in %).

	mAP	R-1	R-5	R-10
HAMNet	45.6	48.5	63.1	68.8
+ $\mathcal{L}_{MC}$	48.8$_{+3.2}$	52.1$_{+3.6}$	67.8$_{+4.7}$	72.1$_{+3.3}$
PFNet	50.1	55.9	68.7	75.1
+ $\mathcal{L}_{MC}$	54.0$_{+3.9}$	58.5$_{+2.6}$	72.1$_{+3.4}$	76.9$_{+1.8}$
IEEE	45.9	48.6	64.3	67.9
+ $\mathcal{L}_{MC}$	51.4$_{+5.5}$	59.3$_{+10.7}$	72.6$_{+8.3}$	75.1$_{+7.2}$
CCNet	50.3	52.7	69.7	75.3
+ $\mathcal{L}_{MC}$	60.6$_{+10.3}$	68.5$_{+15.8}$	79.2$_{+9.5}$	82.4$_{+7.1}$
EDITOR	60.5	63.8	80.0	82.3
+ $\mathcal{L}_{MC}$	66.0$_{+5.5}$	74.2$_{+10.4}$	81.0$_{+1.0}$	82.9$_{+0.6}$
TOP-ReID	67.7	75.3	80.8	83.5
+ $\mathcal{L}_{MC}$	68.5$_{+0.8}$	75.6$_{+0.3}$	81.0$_{+0.2}$	83.5$_{+0.0}$

Table 8
Evaluation of the proposed method on different backbones, where Ours = MFMP + FCE + $\mathcal{L}_{MC}$ (in %).

Method	mAP	R-1	R-5	R-10
ResNet50	53.7	60.3	69.1	74.4
+ Ours	62.1$_{+8.4}$	67.3$_{+7.0}$	76.7$_{+7.6}$	80.1$_{+5.7}$
TransReID	67.0	74.7	79.5	82.4
+ Ours	69.2$_{+2.2}$	76.0$_{+1.3}$	80.2$_{+0.7}$	83.5$_{+1.1}$
ViT	66.5	73.7	80.6	82.6
+ Ours	69.8$_{+3.3}$	77.0$_{+3.3}$	81.0$_{+0.4}$	84.2$_{+1.6}$

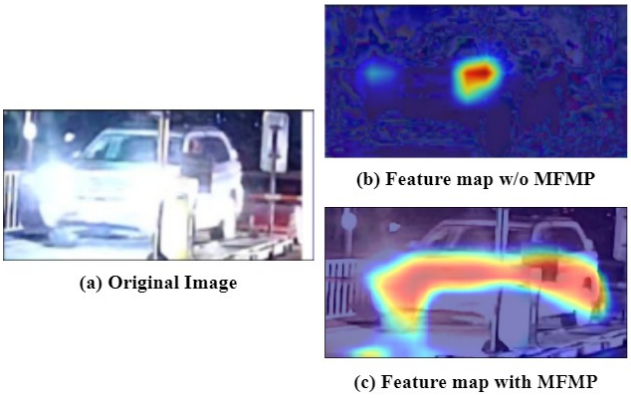
5.5. Evaluation on Different Baselines

To verify the effectiveness and generalization ability of our proposed method, we conduct experiments using different backbone architectures: ResNet50 He et al. (2016), TransReID He et al. (2021) and ViT Dosovitskiy et al. (2020), as shown in Table 8. The results demonstrate that our method consistently enhances the performance across all metrics, regardless of the backbone used. Notably, when applied to the ViT Dosovitskiy et al. (2020) backbone, our method yields a significant improvement in mAP, escalating from 66.5% to 69.8%. Similarly, R-1 accuracy improves from 73.7% to 77.0%.

In summary, the integration of our method with various backbone architectures validates its adaptability and robustness. It demonstrates the capability of our proposed combination of MFMP, FCE, and $\mathcal{L}_{MC}$ in significantly enhancing the performance metrics of vehicle re-identification

Table 9
Comparison with state-of-the-art methods for multi-spectra Re-ID on RGBNT100 Li et al. (2020b) and MSVR310 Zheng et al. (2023) datasets. The best three scores are marked in red, blue and green (in %).

Methods	Ref	Bac	RGBNT100		MSVR310	
			mAP	R-1	mAP	R-1
PCB	ECCV'18	CNN	57.2	83.5	23.2	42.9
MGN	ACMMM'18	CNN	58.1	83.1	26.2	44.3
OSNet	ICCV'19	CNN	75.0	95.6	28.7	44.8
StrongBaseline	CVPR'19	CNN	78.0	95.1	23.5	38.4
AGW	TPAMI'21	CNN	73.1	92.7	28.9	46.9
HAMNet	AAAI'20	CNN	74.5	93.3	27.1	42.3
PFNet	AAAI'21	CNN	68.1	94.1	23.5	37.4
CCNet	INFUS'23	CNN	77.2	96.3	36.4	55.2
TransReID	ICCV'21	ViT	75.6	92.9	18.4	29.6
TOP-ReID	CVPR'24	ViT	81.2	96.4	35.9	44.6
EDITOR	CVPR'24	ViT	82.1	96.4	39.0	49.3
FACENet	Ours	ViT	81.5	96.9	36.2	54.1



(a) Original Image

(b) Feature map w/o MFMP

(c) Feature map with MFMP

Figure 6: Visualization of feature map of MFMP. Mask prediction without MFMP concentrates on the small region near the car lamp. With the aid of MFMP, the mask prediction focuses on a much broader flare-corrupted area.

systems, irrespective of the underlying backbone architecture.

5.6. Results on the benchmark RGBNT100 and MSVR310 datasets

To validate the generalization ability of our method, We further evaluate it on MSVR310 Zheng et al. (2023) and RGBNT100 Li et al. (2020b) in Table 9. Note that FACENet is particularly designed to handle intense flare issues in vehicle ReID. Even though, it still reaches state-of-the-art levels on both RGBNT100 and MSVR310 datasets without intense flare issue. Although the RGBNT100 Li et al. (2020b) and MSVR310 Zheng et al. (2023) datasets are not explicitly degraded by flare, some samples under different lighting conditions would benefit from our FCE module and MC loss function.

Table 10
Evaluate the complexity of different methods on WMVeID863. T_{train}, $T_{inference}$ and $Memory$ represent training times, inference times and memory usage for each epoch of the method respectively.

	Bac	mAP	R-1	T_{train}	$T_{inference}$	$Memory$
HAMNet	CNN	45.6	48.5	27s	24s	9408M
CCNet	CNN	50.3	52.7	24s	21s	11804M
ViT	ViT	66.5	73.7	17s	15s	9590M
TOP-ReID	ViT	67.7	75.3	22s	19s	11196M
FACENet	ViT	69.8	77.0	20s	18s	11868M

Table 11
Analysis of the ID loss, Triplet loss, KL loss and MC loss on WMVeID863 (in %).

ID	Triplet	KL	MC	mAP	R-1	R-5	R-10
✓	×	×	×	61.6	67.6	76.5	79.7
✓	✓	×	×	68.3	75.8	79.5	82.9
✓	✓	✓	×	69.2	76.3	**81.3**	83.6
✓	✓	✓	✓	**69.8**	**77.0**	81.0	**84.2**

5.7. Complexity Analysis

To evaluate the computational complexity of FACENet, we measure the training time, inference time, and memory usage for each epoch on the WMVeID863, as shown in Table 10. The results indicate that our method requires less training and inference time and has comparable memory usage to the currently proposed methods. While it only slightly improves resource consumption during training and inference compared to the baseline ViT Dosovitskiy et al. (2020), our method achieves superior performance over existing state-of-the-art methods.

5.8. Evaluation on Different Losses

Our method primarily utilizes four loss functions: ID loss, Triplet loss, KL loss, and MC loss. We evaluate the effectiveness of KL loss and MC loss in Table 11. The four losses both contribute positively to the network. When jointly utilizing the four losses, our method achieves the best performance on the WMVeID863 dataset.

5.9. Visualization

To demonstrate the effectiveness of our proposed Mutual Flare Mask Prediction Module (MFMP) on WMVeID863 dataset, we utilize Grad-CAM Selvaraju et al. (2020) to visualize the class activation maps (CAMs). This allows us to observe the network's attention to flares before and after applying the MFMP module, with the resulting CAMs superimposed on the original RGB images, as shown in Fig. 6.

To further demonstrate the effectiveness of the proposed method, we visualize the top-10 retrieval results on the different datasets and compare them with the latest state-of-the-art methods, as shown in Fig. 7. Fig. 7 (a) shows that our method can effectively utilize the flare immunity

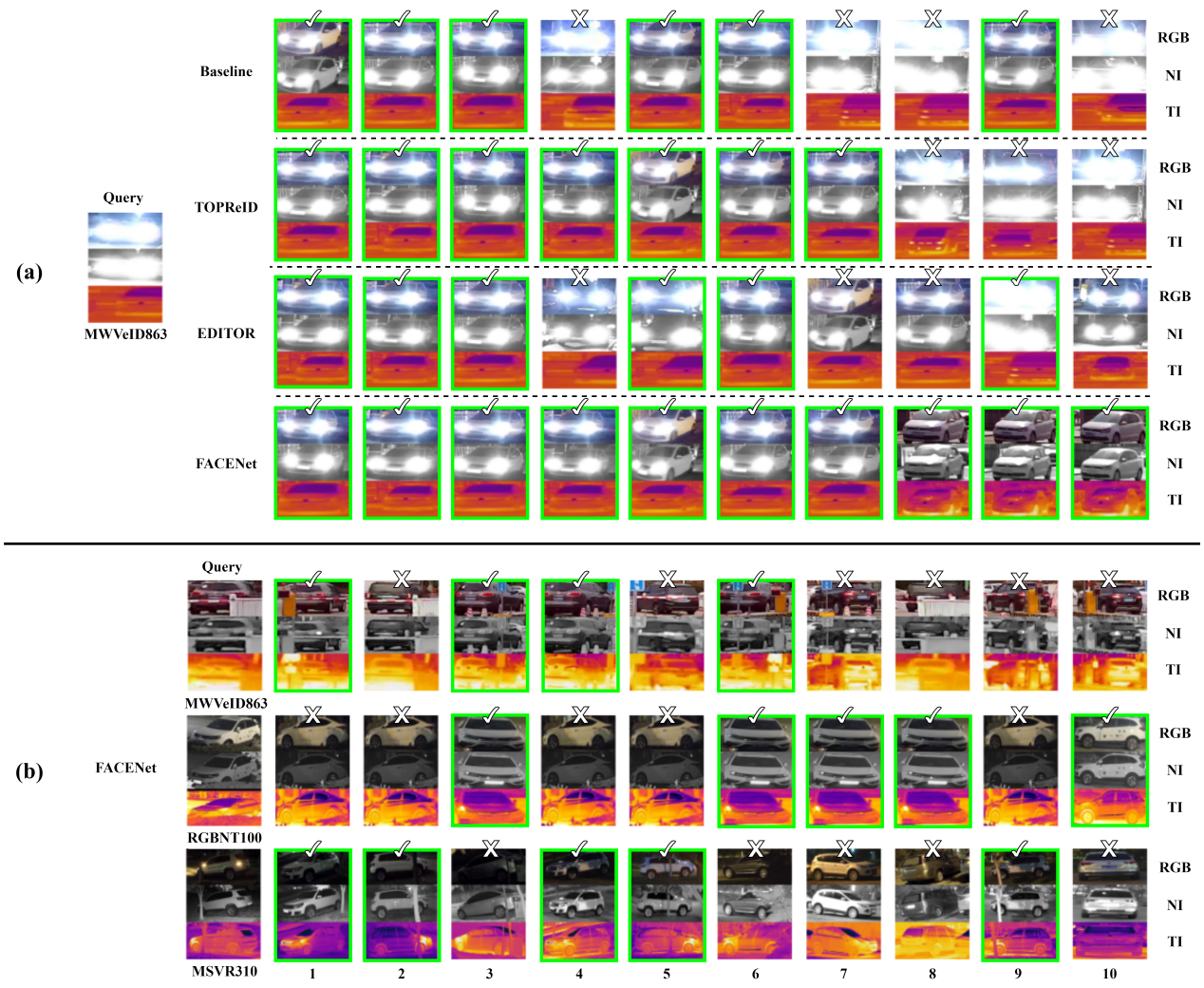


Figure 7: (a) Ranking results of a sample query under different methods on MWVeID863, and (b) Results of FACENet's ranking on different datasets. The green boxes indicate the correct matchings.

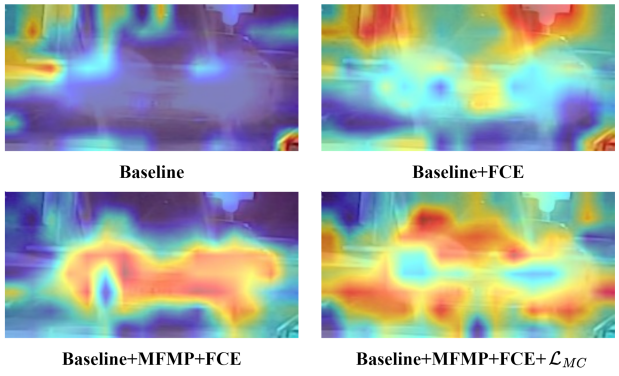


Figure 8: Effectiveness visualization of the proposed modules.

information of the TI spectra, and thus effectively mitigates the effects of nighttime flares on the RGB and NI spectra. However, as Fig. 7 (b) shows, it still faces challenges with significant impact such as perspective distortion and occlusion. In our future work, we will carefully consider these challenges to propose more generalizable and robust methods to address these scenarios.

Finally, to verify the effectiveness of the proposed modules, we apply Grad-CAM Selvaraju et al. (2020) visualization on the final results of different modules, as shown in Fig. 8. It shows that the proposed modules use the flare-immune TI spectra to supplement the missing information in the RGB spectra, which is affected by flare interference, thereby minimizing the impact of nighttime flare on the RGB and NI spectra.

6. Conclusion

In this paper, we first investigate the intense flare challenge in vehicle Re-ID, which results in a significant drop in image quality in certain spectra. To address the problem, we propose a novel Flare-Aware Cross-modal Enhancement

Network (FACENet), which accurately localizes the flare-corrupted area in a self-supervised manner and enhances the flare-susceptible spectra with the guidance of the flare-immuned spectra in a conditional learning manner. In addition, we propose a simple yet effective multi-modal consistency loss to align the semantic information of flare-susceptible spectra and mitigate the semantic degradation caused by flares, thereby further utilizing useful information under intense flare conditions. Moreover, we contribute a new dataset WMVeID863 with more realistic challenges such as intense flare and motion blur for multi-spectral vehicle Re-ID and related communities. Comprehensive experimental results on the proposed WMVeID863 demonstrate the promising performance of the proposed method FACENet, especially while handling intense flares.

Data Availability Statement

The data used to support the findings of this study are included in the paper.

Declaration of Competing Interest

No potential conflict of interest was reported by the authors.

Author Statement

Aihua Zheng: Conceptualization of this study and Methodology. Zhiqi Ma: Investigation and Writing - Original Draft. Yongqi Sun: Validation and Visualization. Zi Wang: Writing Review and Editing. Chenglong Li: Formal Analysis and Data Curation. Jin Tang: Resources and interpretation of data.

Acknowledgements

This research is partly supported by the National Natural Science Foundation of China (No. 62372003), the Natural Science Foundation of Anhui Province (No.2308085Y40 and No. 2208085J18) and the University Synergy Innovation Program of Anhui Province (No. GXXT-2022-036).

References

Chang, X., Hospedales, T.M., Xiang, T., 2018. Multi-level factorisation net for person re-identification, in: CVPR, pp. 2109–2118.

Chen, T.S., Liu, C.T., Wu, C.W., Chien, S.Y., 2020. Orientation-aware vehicle re-identification with semantics-guided part attention network, in: ECCV, pp. 330–346.

Cheng, Y., Wang, R., Pan, Z., Feng, R., Zhang, Y., 2020. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning, in: ACM MM, pp. 3884–3892.

Dai, Y., Li, C., Zhou, S., Feng, R., Luo, Y., Loy, C.C., 2023. Flare7k++: Mixing synthetic and real datasets for nighttime flare removal and beyond. arXiv preprint arXiv:2306.04236 .

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: CVPR, pp. 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Fang, S., Li, K., Li, Z., 2022. S²enet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and lidar data. IEEE GRSL 19, 1–5. doi:10.1109/LGRS.2021.3121028.

Feng, R., Li, C., Chen, H., Li, S., Gu, J., Loy, C.C., 2023. Generating aligned pseudo-supervision from non-aligned data for image restoration in under-display camera, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5013–5022.

Guo, J., Zhang, X., Liu, Z., Wang, Y., 2022. Generative and attentive fusion for multi-spectral vehicle re-identification, in: ICSP, pp. 1565–1572. doi:10.1109/ICSP54964.2022.9778769.

Guo, Y., Zheng, Y., Tan, M., Chen, Q., Chen, J., Zhao, P., Junzhou, H., 2019. Nat: Neural architecture transformer for accurate and compact architectures, in: NeurIPS, pp. 735–747.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR, pp. 770–778.

He, Q., Lu, Z., Wang, Z., Hu, H., 2023. Graph-based progressive fusion network for multi-modality vehicle re-identification. IEEE TITS 24, 12431–12447. doi:10.1109/TITS.2023.3285758.

He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W., 2021. Transreid: Transformer-based object re-identification, in: ICCV, pp. 15013–15022.

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 .

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: CVPR, pp. 4700–4708.

Jiang, K., Wong, W.K., Fang, X., Li, J., Qin, J., Xie, S., 2023. Random online hashing for cross-modal retrieval. IEEE TNNLS , 1–15doi:10.1109/TNNLS.2023.3330975.

Kamenou, E., del Rincón, J.M., Miller, P., Devlin-Hill, P., 2023. A meta-learning approach for domain generalisation across visual modalities in vehicle re-identification, in: CVPR Workshops, pp. 385–393.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. IMS Ann Stat 22, 79–86.

Li, C., Xia, W., Yan, Y., Luo, B., Tang, J., 2020a. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. IEEE TNNLS 32, 3069–3082.

Li, H., Li, C., Zheng, A., Tang, J., Luo, B., 2022a. Attribute and state guided structural embedding network for vehicle re-identification. IEEE TIP 31, 5949–5962.

Li, H., Li, C., Zhu, X., Zheng, A., Luo, B., 2020b. Multi-spectral vehicle re-identification: A challenge, in: AAAI, pp. 11345–11353.

Li, K., Ding, Z., Li, K., Zhang, Y., Fu, Y., 2022b. Vehicle and person re-identification with support neighbor loss. IEEE TNNLS 33, 826–838. doi:10.1109/TNNLS.2020.3029299.

Li, W., Zhu, X., Gong, S., 2018. Harmonious attention network for person re-identification, in: CVPR, pp. 2285–2294.

Liu, H., Tian, Y., Yang, Y., Pang, L., Huang, T., 2016. Deep relative distance learning: Tell the difference between similar vehicles, in: CVPR, pp. 2167–2175.

Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification, in: CVPR Workshops.

Mercea, O.B., Riesch, L., Koepke, A., Akata, Z., 2022. Audio-visual generalised zero-shot learning with cross-modal attention and language, in: CVPR, pp. 10553–10563.

Pang, Z., Wang, C., Zhao, L., Liu, Y., Sharma, G., 2023. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. IEEE TCSVT .

Pang, Z., Zhao, L., Liu, Q., Wang, C., 2022. Camera invariant feature learning for unsupervised person re-identification. IEEE TMM 25, 6171–6182.

Pang, Z., Zhao, L., Liu, Y., Sharma, G., Wang, C., 2024. Inter-modality similarity learning for unsupervised multi-modality person re-identification. IEEE TCSVT .

Phuong, M., Lampert, C.H., 2019. Distillation-based training for multi-exit architectures, in: ICCV.

Qiao, X., Hancke, G.P., Lau, R.W.H., 2021. Light source guided single-image flare removal from unpaired data, in: ICCV, pp. 4157–4165. doi:10.1109/ICCV48922.2021.00414.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-cam: Visual explanations from deep networks via gradient-based localization. IJCV , 336–359doi:10.1007/s11263-019-01228-7.

Shen, F., Du, X., Zhang, L., Shu, X., Tang, J., 2023a. Triplet contrastive representation learning for unsupervised vehicle re-identification. arXiv preprint arXiv:2301.09498 .

Shen, F., Shu, X., Du, X., Tang, J., 2023b. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval, in: ACM MM, pp. 8922–8931.

Shen, F., Xie, Y., Zhu, J., Zhu, X., Zeng, H., 2023c. Git: Graph interactive transformer for vehicle re-identification. IEEE TIP 32, 1039–1051.

Singarimbun, R.N., Nababan, E.B., Sitompul, O.S., 2019. Adaptive moment estimation to minimize square error in backpropagation algorithm, in: ICoSNIKOM, pp. 1–7. doi:10.1109/ICoSNIKOM48755.2019.9111563.

Singh, A., Bhave, A., Prasad, D.K., Singh, A., 2020. Single image dehazing for a variety of haze scenarios using back projected pyramid network, in: ECCV, pp. 166–181.

Teng, X., Lan, L., Zhao, J., Li, X., Tang, Y., 2023. Highly efficient active learning with tracklet-aware co-cooperative annotators for person re-identification. IEEE TNNLS , 1–14doi:10.1109/TNNLS.2023.3289178.

Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., Gao, X., 2023a. Cross-modal enhancement network for multimodal sentiment analysis. IEEE TMM 25, 4909–4921. doi:10.1109/TMM.2022.3183830.

Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., Gao, X., 2023b. Cross-modal enhancement network for multimodal sentiment analysis. IEEE TMM 25, 4909–4921. doi:10.1109/TMM.2022.3183830.

Wang, H., Zha, Z.J., Li, L., Chen, X., Luo, J., 2023c. Context-aware proposal–boundary network with structural consistency for audiovisual event localization. IEEE TNNLS , 1–11doi:10.1109/TNNLS.2023.3290083.

Wang, J., Li, W., Wang, Y., Tao, R., Du, Q., 2023d. Representation-enhanced status replay network for multisource remote-sensing image classification. IEEE TNNLS , 1–13doi:10.1109/TNNLS.2023.3286422.

Wang, T., Liu, H., Song, P., Guo, T., Shi, W., 2022a. Pose-guided feature disentangling for occluded person re-identification based on transformer, in: AAAI, pp. 2540–2549.

Wang, X., Yu, K., Dong, C., Loy, C.C., 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform, in: CVPR, pp. 606–615.

Wang, Y., Liu, X., Zhang, P., Lu, H., Tu, Z., Lu, H., 2024. Top-reid: Multi-spectral object re-identification with token permutation, in: AAAI.

Wang, Z., Li, C., Zheng, A., He, R., Tang, J., 2022b. Interact, embed, and enlarge (ieee): Boosting modality-specific representations for multi-modal person re-identification, in: AAAI, pp. 2633–2641.

Wu, L., Wang, Y., Gao, J., Wang, M., Zha, Z.J., Tao, D., 2021. Deep coattention-based comparator for relative representation learning in person re-identification. IEEE TNNLS 32, 722–735. doi:10.1109/TNNLS.2020.2979190.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C., 2021. Deep learning for person re-identification: A survey and outlook. IEEE TPAMI 44, 2872–2893.

Zamir, S.W., Arora, A., Khan, S.H., Hayat, M., Khan, F.S., Yang, M., Shao, L., 2020. Learning enriched features for real image restoration and enhancement, in: ECCV, pp. 492–511.

Zhang, P., Wang, Y., Liu, Y., Tu, Z., Lu, H., 2024. Magic tokens: Select diverse tokens for multi-modal object re-identification, in: CVPR.

Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: CVPR, pp. 6848–6856.

Zheng, A., Wang, Z., Chen, Z., Li, C., Tang, J., 2021. Robust multi-modality person re-identification, in: AAAI, pp. 3529–3537.

Zheng, A., Zhu, X., Ma, Z., Li, C., Tang, J., Ma, J., 2023. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. Information Fusion 100, 101901.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015. Scalable person re-identification: A benchmark, in: ICCV, pp. 1116–1124.

Zheng, Z., Wang, X., Zheng, N., Yang, Y., 2022. Parameter-efficient person re-identification in the 3d space. IEEE TNNLS , 1–14doi:10.1109/TNNLS.2022.3214834.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: ICCV, pp. 2921–2929.

Zhou, R., Chang, X., Shi, L., Shen, Y.D., Yang, Y., Nie, F., 2020. Person reidentification via multi-feature fusion with adaptive graph learning. IEEE TNNLS 31, 1592–1601. doi:10.1109/TNNLS.2019.2920905.

Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D., 2020. Hetero-center loss for cross-modality person re-identification. Neurocomputing 386, 97–109.

Aihua Zheng received B.Eng. degrees and finished Master-Docter combined program in Computer Science and Technology from Anhui University of China in 2006 and 2008, respectively. And received Ph.D. degree in computer science from University of Greenwich of UK in 2012. She visited University of Stirling and Texas State University during June to September in 2013 and during September 2019 to August 2020 respectively. She is currently an Associate Professor and PhD supervisor at the School of Artificial Intelligence, Anhui University. Her main research interests include vision based artificial intelligence and pattern recognition. Especially on person/vehicle re-identification, audio visual computing, and multi-modal intelligence.

Zhiqi Ma received his B.Eng. degree in 2021 and is currently pursuing the M.Eng degree in the School of Computer Science and Technology, Anhui University, Hefei, China. His research interests include Computer Vision, Multi-modal Intelligence and Vehicle Re-identification.

Yongqi Sun received his B.Eng. degree in 2022 and is currently pursuing the M.Eng degree in the School of Artificial Intelligence, Anhui University, Hefei, China. His research interests include Computer Vision, Multi-modal Intelligence and Vehicle Re-identification.

Zi Wang received his B.Eng. degree in 2021 and is currently pursuing the M.Eng degree in the School of Computer Science and Technology, Anhui University, Hefei, China. His research interests include Computer Vision, Multi-modal Intelligence and Vehicle Re-identification.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he worked as a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a postdoctoral research fellow at the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an Associate Professor and PhD supervisor at the School of Artificial Intelligence, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in Computer Science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor and PhD supervisor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.