Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MFFP-Net: Multi-directional Feature Fusion and Position-Aware Network

Yazhong Si

siyazhong123@163.com

Wuxi University

Jingyu Chen

Nanjing University of Information Science and Technology

Hongxu Li

Wuxi University

Chen Li

Wuxi University

Research Article

Keywords: Vehicle re-identification, Multi-directional feature fusion, Position awareness, Depthwise separable convolution

Posted Date: March 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8783703/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MFFP-Net: Multi-directional Feature Fusion and Position-Aware Network

Yazhong Si[1*], Jingyu Chen[2], Hongxu Li[3], Chen Li[4]

[1*]School of Electronic and Information Engineering, Wuxi University, 333 Xishan Avenue, Xishan District, Wuxi, 214105, Jiangsu, China.
[2]School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, 219 Ningliu Road, Nanjing, 210044, Jiangsu, China.
[3]Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence, Wuxi University, 333 Xishan Avenue, Xishan District, Wuxi, 610101, Jiangsu, China.
[4]School of Cybersecurity and Informatization, Wuxi University, 333 Xishan Avenue, Xishan District, Wuxi, 610101, Jiangsu, China.

*Corresponding author(s). E-mail(s): siyazhong123@163.com;
Contributing authors: 1498332213@qq.com; hongxuli@cwxu.edu.cn;
papertg2018@163.com;

Abstract

Vehicle re-identification faces significant challenges in feature modeling due to viewpoint variations, illumination changes, occlusions, and intra-class similarity. To this end, we propose a VeRi algorithm based on multi-directional feature fusion and position awareness, which synergistically integrates multi-dimensional features and positional information to suppress background interference and improve feature discriminability. Specifically, we design multi-directional depthwise separable convolution kernels which square kernels for local details, horizontal strip kernels for long-range dependencies, and vertical strip kernels for spatial distribution to capture comprehensive directional features; the DOConv module fuses depthwise convolution and conventional convolution to balance fine-grained texture extraction and global structural integration without additional sub-networks; and a position encoding module with horizontal-vertical tensor

initialization enhances the model's perception of key components' relative positions. Experimental results on the VeRi-776 and Veri-Wild datasets demonstrate the superior performance of the proposed algorithm.

Keywords: Vehicle re-identification, Multi-directional feature fusion, Position awareness, Depthwise separable convolution

1 Introduction

With the acceleration of urbanization and the in-depth deployment of Intelligent Transportation Systems (ITS), vehicle identity matching and trajectory tracking across non-overlapping cameras have become core requirements for improving traffic supervision efficiency and ensuring public safety [1–3]. License plate recognition (LPR) as a traditional vehicle identification method often encounters performance degradation in extreme scenarios such as license plate forgery, partial occlusion and low-light-induced, blurring thus rendering VeRi an indispensable alternative technology. As a key supporting technology, VeRi can realize cross-spatiotemporal retrieval of target vehicles through visual features without relying on license plate recognition, and has been widely applied in fields like traffic violation traceability, hit-and-run vehicle investigation, smart parking lot management, and urban security deployment.

It was not until Liu et al. [4] the VeRi-776 dataset based on real urban surveillance scenarios that deep model training supported by large-scale annotated data was first realized, promoting VeRi research into a stage of rapid development. Subsequently, large-scale datasets such as Veri-Wild [5] and VehicleNet [6] have further advanced the field: Veri-Wild, as a large-scale wild dataset, captures diverse vehicle images under unconstrained real environments, while VehicleNet integrates four public datasets to provide more comprehensive training samples [6]. Meanwhile, various advanced networks and frameworks have emerged, including the Strip-based and Attribute-aware Network (SAN) [7], the Domain Adaptive Vehicle Re-identification (DAVR) framework [8], and the Multi-view Spatial Attention Embedding network [9], enriching the technical landscape of VeRi but also highlighting the complexity of the task.

Nevertheless, complex factors in real surveillance environments pose severe challenges to VeRi technology: (1) Significant intra-class inconsistency caused by viewpoint shifts, illumination fluctuations, and background interference [10]; (2) Prominent inter-class similarity among vehicles of the same brand and model [4]; (3) Poor cross-camera feature consistency due to camera parameter differences and vehicle motion blur [11]; (4) Low resolution and lack of local features in long-distance capture scenarios [12]. Additionally, the lack of fine-grained annotation in some datasets and cross-domain discrepancies further degrade model performance [13, 14].

To address these challenges, academia has formed five parallel research directions: global-local feature synergy, attention mechanism enhancement, Transformer-based feature decomposition, memory units with contrastive learning, and generative methods with attribute learning. However, these directions still face inherent limitations: global-local feature methods often disrupt spatial correlation or rely on complex

sub-networks; attention-based models fail to adapt to dynamic viewpoint changes; Transformer architectures suffer from poor generalization in few-shot scenarios; memory unit methods lack directional feature guidance; generative and attribute learning approaches struggle to balance fine-grained and global features. Collectively, existing methods have four core shortcomings: (1) Single dimension of feature capture [15, 16]; (2) Insufficient utilization of positional information [17, 18]; (3) Limited robustness to extreme scenarios [12]; (4) Conflict between model complexity and generalization [17, 19]. To tackle these core issues, we proposes a vehicle re-identification algorithm (MFFP-Net) based on multi-directional feature fusion and position awareness. The algorithm achieves multi-dimensional feature interaction optimization through three synergistic core designs:

(1) A multi-directional feature extraction module (MFE) with customized convolution kernels to capture full-dimensional vehicle features;

(2) A DOConv module that fuses depthwise and traditional convolution to balance local and global features without additional sub-networks;

(3) A positional encoding module (PE) with horizontal-vertical tensor initialization to enhance positional information utilization.

These components collaboratively suppress background interference and improve robustness to complex scenarios. Experimental results on VeRi-776 and Veri-Wild datasets verify the superior performance of the proposed algorithm.

The collaborative work of these three components can suppress background noise interference and enhance robustness to illumination variations, occlusions, extreme viewpoints, and similar vehicle models. Experimental validation demonstrates that the proposed algorithm outperforms mainstream baselines in terms of Rank-1 accuracy and mAP metrics on the VeRi-776 and Veri-Wild datasets, providing a new solution for multi-dimensional challenges in VeRi.

2 Related Work

Existing vehicle re-identification research primarily centers on robust feature extraction and discriminative metric learning, evolving into five mainstream technical paths: global-local feature synergy, attention mechanism enhancement, Transformer-based feature decomposition, memory units with contrastive learning, and generative methods with attribute learning.

2.1 Multi-directional Feature Capture

Most existing methods suffer from single-dimensional feature capture, failing to comprehensively model the horizontal, vertical, and local detail features of vehicles. In the direction of global-local feature synergy, Xu et al. [15] designed a horizontal four-partition local branch network, but rigid segmentation disrupts the spatial correlation of vehicle components; Zhang et al. [16] proposed a dual-attention granularity network, which only focuses on coarse and fine-grained feature differentiation without directional feature modeling. In attention mechanism research, the SE block [20] only calibrates channel weights and lacks the ability to capture directional dependencies; Jiang et al. [21]'s global reference attention network ignores the dynamic impact of

viewpoint changes on directional feature significance. Khorramshahi et al. [22] propose a dual-path model with adaptive attention to enhance feature adaptability, yet it still fails to fully capture the multi-directional structural features of vehicles. Generative methods such as XVGAN [23] and MV-GAN [24] alleviate view imbalance but cannot break through the limitations of single-dimensional feature representation. Lou et al. [25] adopt embedding adversarial learning to improve feature robustness against interference, yet this method still lacks targeted modeling of horizontal and vertical directional features. Tumrani et al. [26]'s view-aware attribute-guided network emphasizes the impact of viewpoint on attribute expression, but it still cannot comprehensively integrate directional structural features. To address these issues, the proposed MFE module designs three customized depthwise separable convolution kernels, which synergistically capture local details, horizontal long-range dependencies, and vertical spatial distribution, making up for the deficiency of incomplete feature representation in existing methods.

2.2 Local-Global Feature Balance

Traditional methods face a conflict between local-global feature modeling and model complexity. In global-local feature fusion, Sun et al. [17]'s TBE-Net relies on additional object detection/segmentation sub-networks, leading to a 60% increase in parameters and reduced inference speed; PGAN [18] struggles to suppress background redundancy while integrating local and global features. Sun et al. [27] propose a multi-feature learning model with enhanced local attention, which strengthens the extraction of local component features but weakens the modeling of global structural continuity. Transformer-based methods such as TransReID [28] and Git [19] enhance structural feature capture but bring high computational complexity and poor generalization in few-shot scenarios. Almeida et al. [29] adopt multi-branch representation learning to enrich feature diversity, yet they still face the dilemma of balancing local detail extraction and global structure integration. Traditional convolution layers use symmetric square kernels, which cannot balance the modeling needs of horizontal/vertical directional features and local/global structures. The DOConv module proposed in this paper fuses depthwise convolution for capturing fine-grained textures and traditional convolution for modeling global structures, achieving efficient local-global feature synergy without additional sub-networks while maintaining the inference efficiency of traditional convolution.

2.3 Positional Information Utilization

Existing methods generally underutilize the positional correlation of key vehicle components, weakening the model's discriminative ability. In global-local feature synergy, Jin et al. [30]'s multi-center metric learning method and Zhang et al. [11]'s part-guided attention network do not explicitly model the relative positions of components such as headlights and wheel hubs. Tumrani et al. [31] integrate partial attention and multi-attribute learning to enhance feature discriminability, but they ignore the positional correlation between different attributes. In memory unit and contrastive

learning research, MoCo [32] and MSCL [33] focus on feature distribution optimization but lack positional feature guidance. Cheng et al. [34] extend memory-based contrastive learning to cross-modality person re-identification, which provides a reference for feature embedding optimization but does not involve the modeling of vehicle component positional correlations. Transformer-based methods such as Git [35] use complex graph convolution to model component correlations, increasing computational overhead. Chu et al. [36] propose viewpoint-aware metric learning to optimize inter-class distance, but it does not explicitly encode the horizontal and vertical positional information of vehicles. The proposed PE module adopts a horizontal-vertical tensor initialization strategy, which efficiently models the relative positional relationships of key components through truncated normal distribution initialization and dimension expansion, addressing the issue of insufficient positional information utilization in existing methods while avoiding complex structural design.

3 Proposed method

The proposed network is designed based on the ResNet50-IBN architecture. Input images are first processed by the Position Encoding Module; they are then fed into the backbone network, where feature extraction is conducted sequentially through layer1, layer2, and layer3, followed by the extraction of output features from layer1 and layer2 respectively. These two layers of features are further subjected to convolution (Conv) and Global Average Pooling (GAP) operations to adjust their spatial dimensions, after which they are jointly fed into the multi-directional feature capture module for processing together with the backbone features. Finally, the processed features enter the last layer of the network (layer4) to facilitate the processing of subsequent tasks. The network structure is shown in Figure 1.

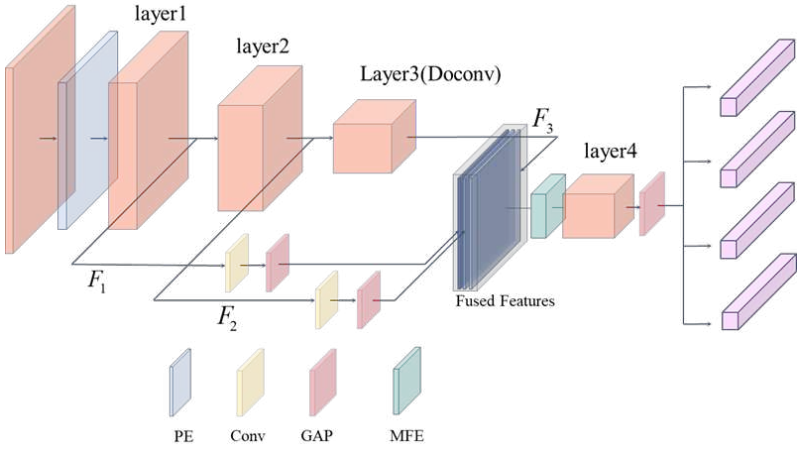


Fig. 1: Overall architecture of the proposed framework.

F1 demonstrates the output feature of layer1 in the backbone network, F2 is the feature obtained after the backbone network sequentially outputs through layer1 and layer2 and then adjusts the spatial dimension and channel dimension through convolution and global average pooling operations, and F3 is the output feature of layer3 in the backbone network; the three are fused to form Fused Features.

4 Algorithm Implementation

4.1 Multi-directional Feature Extraction Module

Due to the significant viewpoint differences, pose variations, and intra-class similarity in vehicle appearances, how to accurately capture discriminative multi-dimensional features has always been a core challenge for improving model performance. In vehicle images, horizontal vehicle body contours, vertical structural proportions, and local regional detail features together constitute the key information for distinguishing different vehicles. However, traditional feature extraction methods often fail to simultaneously balance multi-directional feature modeling of global structures and local details. To address this issue, this paper proposes a Multi-directional Feature Extraction Module (MFE). Through specially designed convolution kernel structures, this module achieves efficient capture of vehicles' horizontal long-range dependencies, vertical spatial distribution, and local detail features, aiming to provide a more comprehensive and discriminative feature representation for vehicle re-identification tasks, thereby improving the accuracy and robustness of vehicle matching in cross-scenario settings.

This module takes depthwise separable convolution as its core architecture, and achieves accurate extraction and efficient encoding of multi-dimensional features from vehicle images through custom-designed convolution kernel structures and feature fusion mechanisms. Its core design logic is as follows: Targeting the global structural characteristics of vehicle features in horizontal and vertical directions and the discriminative details of local regions, differentiated convolution kernels are adopted for directional capture; additionally, through feature fusion and dimension regularization, a comprehensive feature representation integrating both global and local information is formed. The module structure is illustrated in Figure 2:

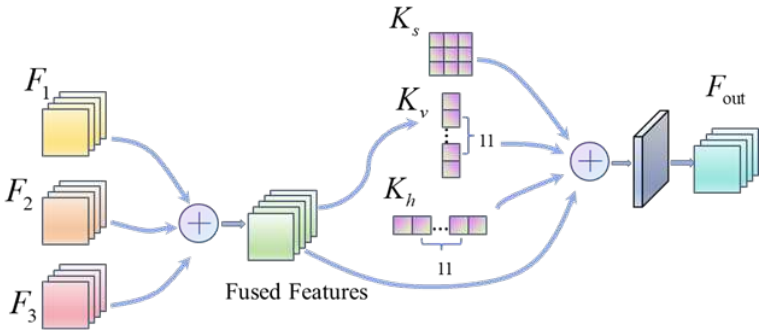


Fig. 2: Illustration of the proposed Multi-directional Feature Extraction Module.

Input features F_1, F_2, F_3 are channel-wise concatenated to form the intermediate Fused Features. This Fused Features is fed into three parallel convolutional branches with unique kernels K_s, K_ν, K_h for complementary multi-directional discriminative feature extraction. The resulting features from the three branches are further channel-wise concatenated with the initial Fused Features followed by a pooling operation to refine feature scale and produce F_{out}.

From the perspective of mathematical modeling for feature extraction, let the input vehicle feature map be $X \in R^{C\times H\times W}$, where C denotes the number of channels, and H, W represent the height and width of the feature map, respectively. The module comprises three parallel depthwise separable convolution branches, and each branch extracts directional features through the "main channel independent convolution" mechanism. A square convolution kernel $K_{\mathcal{S}} \in R^{\mathrm{k_z}\times \bar{k}_z}$ is adopted for square regions to capture local detail features via depthwise convolution, as shown in Equation (1):

$$F_s(i,j) = \sum_{m=0}^{k_z-1}\sum_{n=0}^{k_z-1} X_c(i+m-\lfloor k_s/2\rfloor, j+n-\lfloor k_s/2\rfloor)\cdot K_{s,c}(m,n) \tag{1}$$

where $F_{\mathrm{s}} \in R^{C\times H\times W}$ is the square feature output; c$\in \{1,2,\ldots,C\}$ denotes the channel index; $\lfloor\cdot\rfloor$ represents the floor operation, and $K_{s,c}$ is the square convolution kernel for the c-th channel (since the number of groups g=C, channel-wise independent convolution is achieved); the padding parameter $\mathfrak{p}_s = [k_s/2]$ ensures the output size is consistent with the input size, and k_{s} is the size of the square convolution kernel; i is the index of the feature map in the vertical dimension, j corresponds to the index in the width direction (horizontal dimension), m and n represent the local indices inside the square convolution kernel $K_{s,c}$,where m corresponds to the vertical dimension of the convolution kernel and n corresponds to the horizontal dimension.

For the horizontal direction, a horizontal strip convolution kernel $K_{\mathrm{h}} \in R^{1\times \mathrm{k}_b}$ is adopted,where $K_{\mathrm{h},c}$ is the size in the horizontal direction. It captures horizontal longrange dependencies via depthwise convolution, as shown in Equation (2):

$$F_{\mathrm{h}}(i,j) = \sum_{n=0}^{k_b-1} X_c(i, j+n-\lfloor k_b/2\rfloor)\cdot K_{h,c}(1,n) \tag{2}$$

where $F_{\mathrm{h}} \in R^{\mathbb{C}\times H\times W}$ is the horizontal output, $K_{\mathrm{h.c}}$ is the horizontal convolutior kernel for the c-th channel; the padding parameter $\mathfrak{p}_h = [k_h/2]$ ensures scale invariance in the horizontal direction. For the vertical direction, a vertical strip convolution kernel $K_{\mathrm{v}} \in R^{k_b\times 1}$ is adopted, where $k_{_c}$ is the size in the vertical direction, and its depthwise convolution calculation is shown in Equation (3):

$$F_\nu(i,j) = \sum_{n=0}^{k_b-1} X_c(i, j+n-\lfloor k_b/2\rfloor, j)\cdot K_{\nu,c}(m,1) \tag{3}$$

where $F_\nu \in \mathbb{R}^{C\times H\times W}$ is the vertical feature output, $K_{v,c}$ is the vertical convolution kernel for the c-th channel, and the padding parameter $\mathfrak{p}_\nu = [k_\nu/2]$ ensures scale invariance in the vertical direction. In the feature fusion stage, the original input and the features from the three branches are concatenated along the channel dimension to obtain the fused feature, as shown in Equation (4). The fused feature is represented as:

$$F_{\mathrm{merge}} = \mathrm{Concat}(X, F_s, F_h, F_\nu) \in \mathbb{R}^{4C\times H\times W} \tag{4}$$

where $\mathrm{Concat}(\cdot)$ denotes the channel-dimension concatenation operation. By integrating the original features and directional features, F_{merge} retains completeness across multiple scales and directions.

To achieve dimension regularization and feature reorganization, a 1x1 convolution layer is adopted for channel compression, as shown in Equation (5):

$$F_{\mathrm{out}} = \mathrm{RELU}(BN(\sum_{c'=1}^{4C} W_{c,c'}\cdot F_{\mathrm{merge},c'} + b_c)) \tag{5}$$

where $W \in R^{C\times 4C\times 1\times 1}$ is the weight of the 1x1 convolution kernel; c' is the channel index of the fused feature map $F_{\mathrm{merge},c'}$, $b \in R^C$ is the bias; $BN(\cdot)$ denotes the batch normalization operation; the output feature $F_{\mathrm{out}} \in R^{C\times H\times W}$ has the same dimension as the input, ensuring the module can be embedded into existing networks.

The local neighborhood receptive field of the square convolution kernel K_s can accurately capture the detail features of local vehicle components, such as the edge morphology and internal light source distribution of headlights, as well as the spoke texture and hollow structure of wheel hubs. The horizontal strip receptive field of the horizontal convolution kernel K_h can effectively cover the extensional features of horizontally continuous components on the vehicle body, such as the horizontal contours of the upper and lower edges of windows, and the horizontal arrangement and overall coherence of taillights. The vertical strip receptive field of the vertical convolution kernel K_v is adapted to the distribution features of the vehicle body's longitudinal structure, such as the vertical tilt angle and connection morphology of pillars, as well as the position of vertical dividing lines and longitudinal distribution features of doors. Through their respective spatial receptive fields, the three kernels focus on local details and global structural features of the vehicle respectively, forming a comprehensive feature encoding for key components such as headlights, wheel hubs, windows, taillights, pillars, and doors.

4.2 DOConv Convolution Module

Vehicle appearance exhibits significant differences due to factors such as viewpoint, illumination, and occlusion, and it is difficult to distinguish fine-grained features (e.g., headlight morphology, wheel hub texture) of the same vehicle type. Traditional convolution layers struggle to balance the multi-dimensional modeling requirements of global structure and local details during feature extraction, resulting in limited adaptability of models to complex scenarios.

Traditional convolution layers handle channel and spatial dimensions simultaneously with a single convolution kernel, which has significant limitations in vehicle re-identification tasks. Since traditional convolutions mostly use square convolution kernels with symmetrically distributed spatial receptive fields, their multi-directional feature capture capability is insufficient, making it impossible to adapt to the directional differences of vehicle features in a targeted manner. The continuous contours of the vehicle body in the horizontal direction require long-range dependency modeling, while the structural dependencies in the vertical direction rely on longitudinally continuous features. The balanced effect of square convolution kernels cannot simultaneously enhance these two types of features, leading to the loss of discriminative information in the horizontal/vertical directions.

Meanwhile, there is a contradiction between parameter redundancy and expressive capability. The parameter dimension of traditional convolution kernels is $C_{\text{out}} \times C_{in} \times M \times N$, and the parameter scale grows linearly with the number of channels and kernel size. A large number of parameters are used for redundant channel-spatial cross-mapping rather than effective feature learning. Many parameters are not optimized for local details or global structures of vehicles, resulting in low feature expression efficiency.

In addition, the receptive field of traditional convolution kernels is fixed. If the kernel size is small, it is difficult to cover the global structure of the vehicle body; if the kernel size is large, it will dilute local detail features. This local global trade-off characteristic makes the model perform poorly in distinguishing similar vehicle types, especially in viewpoint-varying scenarios, where feature matching accuracy drops significantly, as shown in Figure 3.

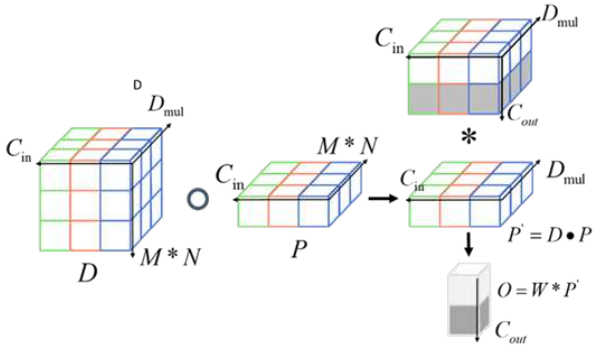


Fig. 3: Illustration of DOConv Module.

The size of the input patch P is MxN, the trainable kernel dimension D, the number of input channels , and the number of output channels; different input patches can generate corresponding output channels (the cubes in the figure). The DOConv module achieves local-global feature fusion through the combination of the depthwise convolution operator and the conventional convolution operator $*$.

DOConv addresses the defects of traditional convolution through the mechanisms of depthwise convolution over-parameterization and kernel combination. Its core design is to capture local details with depthwise convolution and integrate global structures with traditional convolution, providing sufficient feature learning space through over-parameterization while maintaining inference efficiency.Let the input vehicle feature map be $X \in R^{C \times H \times W}$, and its calculation process is as follows.

Depthwise convolution captures local fine-grained features and adopts channel-independent convolution, where each input channel corresponds to a dedicated convolution kernel, focusing on the features of local vehicle components, as shown in Equation (6):

$$P'_{c,d} = \sum_{i=1}^{M \times N} D_{c,i,d} \otimes P_{c,j} \tag{6}$$

where $P'_{c,d}$ is the intermediate feature output by depthwise convolution, c denotes the input channel index, d denotes the depth multiplier index(corresponding to D_{mul}, the over-parameterized depth multiplier); $D_{c,i,d}$ denotes the depthwise convolution kernel, i is the spatial position index of the convolution kernel, with dimension $R^{C_{in} \times (M \times N) \times D_{\text{mul}}}$, C_{in} is the number of channels of the feature map input to the current layer, M and N are the spatial sizes of the convolution kernel; $P_{c,j}$ is the feature value at the $j-th$ spatial position of channel c in the input feature; $\otimes$ denotes the spatial convolution operation.

Traditional convolution integrates global structural features, performing cross-channel weighted fusion on the output of depthwise convolution to model the global structural dependencies of vehicles, as shown in Equation (7):

$$O_o = \sum_{d=1}^{D_{\text{mul}}} \sum_{c=1}^{C_{in}} W_{o,d,c} \cdot P'_{c,d} \tag{7}$$

where O_o is the feature output by traditional convolution, o denotes the output channel index (corresponding to C_{out}); $W_{o,d,c}$ denotes the traditional convolution kernel, with dimension $R^{C_{out} \times D_{mi} \times C_{in}}$, $\cdot$ denotes the weights summation operation. Finally, for feature fusion and dimension regularization, the original features and enhanced features are fused. The original features and directionally enhanced features are integrated through channel-dimension concatenation, as shown in Equation (8):

$$F_c = Concat(X, P_{c,d}, O_o) \tag{8}$$

where F_c is the fused feature; X is the original input feature map; $Concat(\cdot)$ is the channel-dimension concatenation operation, integrating original features, depthwise convolution features, and traditional convolution features.

Channel compression is performed via a 1x1 convolution to ensure the output dimension is consistent with the input, enabling seamless embedding into existing networks, as shown in Equation (9):

$$Y = BN(W_{1\times1} \cdot F + b) \tag{9}$$

where Y is the final output feature of the module; $W_{1\times1}$ is the 1×1 convolution kernel; BN denotes the batch normalization operation; and b is the bias term.

DOConv enables depthwise convolution to dynamically learn directional weights through a parameterization mechanism, flexibly capturing horizontal/vertical features of vehicle bodies without the need for manual design of strip kernels. Through per-channel depthwise convolution and cross-channel traditional convolution, parameter utilization is more efficient, allowing for precise encoding of fine-grained differences such as wheel hubs and headlights.

Meanwhile, depthwise convolution focuses on local components, while traditional convolution associates global structures, forming complementary local and global features to enhance the distinguishability of similar vehicle models, especially improving the recognition of vehicles without license plates. During training, it converges faster, is more robust to occlusion and illumination changes, and is compatible with existing networks. Its inference efficiency is consistent with that of traditional convolution, making it suitable for real-time scenarios.

4.3 Positional Encoding Module

The positional information of vehicles in images is of great significance for distinguishing different vehicles. It can help models better understand the spatial layout of vehicles and the relative positions of key components, thereby enhancing the discriminative ability of models. Traditional Convolutional Neural Networks (CNNs) mainly rely on convolution operations to extract local features and reduce the spatial resolution of feature maps through pooling operations. Although this operation can effectively capture appearance features, its utilization of positional information is limited, which weakens the model's perception of vehicle positions to a certain extent. In addition, different vehicles vary in size in the horizontal and vertical directions; some are longer horizontally, while others are taller vertically. Therefore, models need to be able to flexibly handle feature information in different directions. Accurately grasping the positional features of vehicles has become the key to improving recognition accuracy.

To address this deficiency, this paper designs a Positional Encoding Module (PE) to improve the model's perception of positional information, and its structure is shown in Figure 4. The Positional Encoding Module adds positional information for the horizontal and vertical directions respectively by initializing positional encoding tensors.

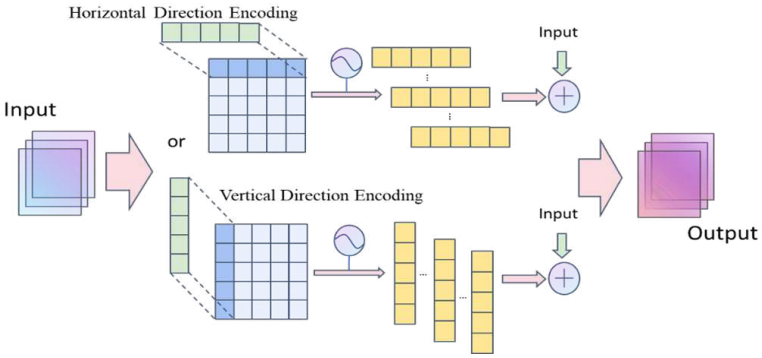


Fig. 4: Illustration of Positional Encoding Module.

Figure 4 is configured with dual horizontal/vertical directional encoding branches, first initializing positional encoding parameters via a truncated normal distribution, then expanding the encoding dimensionality to match the spatial resolution and channel dimensions of the input feature maps, and finally generating output features through element-wise addition between the positional encodings and the original input feature maps.

For the horizontal positional encoding, a tensor with the initial shape $P_H \in R^{1\times C\times W\times 1}$ is initialized; for the vertical encoding, a tensor with the initial shape $P_V \in R^{1\times C\times 1\times W}$ is initialized. Here, C denotes the number of channels of the feature map, and W is the global window size. Both the horizontal and vertical positional encoding tensors are initialized using a truncated normal distribution, as shown in Equation (10):

$$P_H, P_V \sim N_{\text{trunc}}(\mu = 0, \sigma = 0.02) \tag{10}$$

where P_H is the horizontal positional encoding tensor; P_V is the vertical positional encoding tensor; $N_{\text{trunc}}(\bullet)$ represents the Truncated Normal Distribution; μ is the mean, and σ is the standard deviation (set to 0.02).

The positional encoding tensors are then matched to the size of the input feature map via an expansion operation, resulting in an expanded shape of $1 \times C \times S_h \times S_w$, which allows alignment with each position of the feature map. Mathematically, this process can be described in detail as shown in Equations (11) and (12):

$$P_{H_exp} = \text{Expand}(P_H, \text{target_shape} = (1, C, S_h, S_w)) \tag{11}$$

$$P_{V_exp} = \text{Expand}(P_V, \text{target_shape} = (1, C, S_h, S_w)) \tag{12}$$

where $Expand(\cdot)$ is the expansion operation, $P_{H\ exp}$ is the horizontally expanded tensor, and $P_{V\ exp}$ is the vertically expanded tensor; S_h is the height dimension of the feature map; S_w is the width dimension of the feature map. The expanded positional encoding is added element-wise to the input feature map $F \in R^{1\times C\times S_h\times S_w}$, as shown

in Equation (13):

$$F_{out} = F + P_{H_exp} + P_{V_exp}$$

where F_{out} is the output feature map fused with positional information. This feature map contains the positional information of the vehicle in both horizontal and vertical directions, helping the model capture the spatial features of key vehicle parts more accurately.

The PE module enables the model to learn the relative positions of vehicles in images, especially for vehicles with significant dimensions in the horizontal or vertical direction. This targeted feature enhancement helps improve the model's ability to recognize different types of vehicles. With the feature map enhanced by positional encoding, the model can more precisely capture key vehicle parts, whose positional information plays a critical role in vehicle identity recognition. In addition, the positional encoding module not only enhances the model's sensitivity to positions but also guides feature learning, allowing the network to utilize spatial information more effectively. This mechanism helps the model accurately recognize vehicles even when dealing with complex scenarios such as occlusion, viewpoint variations, and illumination changes.

4.4 Loss Function

To enhance the learning capability of the proposed network structure and accelerate network convergence, we combines classification loss and metric loss with a multi-branch architecture to improve the discriminative ability and generalization ability of the model.

4.4.1 Classification Loss

Classification loss is used to improve the model's performance in identity recognition tasks. Specifically, classification loss is implemented via Cross-Entropy Loss, as shown in Equation (14):

$$L_{cls} = -\sum_{c=1}^{C} y_{gt}^{j} \log y_i^p$$

where L_{cls} denotes the cross-entropy loss value of the classification branch, y_{gt}^{j} represents the true label distribution of the i-th category, y_i^p denotes the model's predicted probability for the i-th category, and C is the number of categories in training. The label smoothing strategy adjusts the distribution of true labels by introducing a smoothing parameter ε, as shown in Equation (15):

$$y_{gt}^{i} = \begin{cases} 1 - \frac{C-1}{C}\varepsilon & i \neq class \\ \varepsilon & i = class \end{cases} \tag{13}$$

where ε is the smoothing parameter, and this parameter is set to 0.1 in this paper.

4.4.2 Metric Loss

Metric loss is used to enhance the discriminative ability of feature embedding, especially for vehicle features under different viewpoints. This paper adopts Batch-Hard triplet loss [37] to optimize feature embedding, as shown in Equation (16):

$$L_{tri} = \frac{1}{PK}\sum_{i=1}^{P}\sum_{k=1}^{K}\max[0, m + \max_{p\in P(\mathrm{a})} d(x_a, x_p) - \min_{n\in N(a)} d(x_a, x_n)] \tag{14}$$

where P denotes the number of identities, K is the number of samples per identity, m is the distance threshold, $P(a)$ represents the positive sample set for each anchor, $N(a)$ denotes the negative sample set for each anchor, x_a is the anchor, x_p is the extracted embedding of the positive sample, x_n is the extracted embedding of the negative sample, and $d()$ stands for Euclidean distance.

4.4.3 Overall Loss Function

The overall loss function is implemented via the weighted linear combination of classification loss and metric loss, as shown in Equation (17):

$$L = \sum_{i=1}^{N} \omega_i^{cls} L_i^{cls} + \omega_i^{tri} L_i^{tri} \tag{17}$$ (15)

where ω_i^{cls} and ω_i^{tri} are the weights of classification loss and metric loss respectively; L_i^{cls} and L_i^{tri} are the classification loss and metric loss of the i-th branch respectively.

5 EXP

5.1 Network Parameter Settings

Experiments are conducted on two mainstream datasets: VeRi-776, a 2016-proposed urban surveillance dataset with 776 vehicle identities and 49,357 images; and Veri-Wild, a 2019-released large-scale wild dataset containing 40,671 identities and 416,314 images covering unconstrained real environments. We use an Intel(R) Core(TM) i7-10700 CPU, 32GB memory, NVIDIA GeForce RTX 3070 (8GB) GPU, and the PyTorch framework; the input image size is 256x256x3, with data augmentation including edge padding, random cropping, and horizontal flipping. Training batches are constructed via the PK sampling strategy (P=6 identities per batch, K=8 samples per identity, total batch size=48); the Adam optimizer is adopted with an initial learning rate of 0.0001 (decayed by 0.1 at epoch 45, 90, and 130), and the stride of the last convolutional layer is adjusted to 1 before training.

5.2 Comparison with other methods

This paper compares the proposed algorithm with current mainstream algorithms on the VeRi-776 and Veri-Wild datasets, and the results are presented in Table 1 and Table 2.

On the VeRi-776 dataset, the experimental results in Table 1 show a clear evolutionary gradient in the performance of existing methods. Early basic models including GoogleNet and FACT have mAP values below 20% and CMC1 values less than 60%, as they only capture basic features and fail to handle intra-class variations caused by viewpoint shifts and illumination fluctuations. Even the generative-based XVGAN only increases mAP to 24.65%, which still cannot break through the limitations of single-feature dimension. Later methods including MSCL, GSTE and VAAG optimize via metric learning or attention mechanisms, raising mAP to 45.90-67.58% and CMC1 to 81.20-92.21%. However, their performance still faces bottlenecks due to single-dimensional feature capture or insufficient utilization of positional information. In contrast, our method achieves optimal performance on the dataset: it attains 77.89% mAP, 96.66% CMC1, and 98.33 CMC5. These superior results stem from the synergy of three core components: multi-directional feature extraction, the DOConv module, and positional encoding.

Table 1: Performance Comparison on the VeRi-776 Dataset

Method	mAP	CMC1	CMC5
GoogleNet[38]	17.81	52.12	66.78
FACT[4]	18.73	51.85	67.16
XVGAN[23]	24.65	60.20	77.03
MSCL[33]	45.90	81.20	–
GSTE[39]	59.43	92.21	96.25
VAAG[26]	65.23	91.35	94.76
NVSL[40]	67.58	91.52	94.83
TCPM[14]	74.6	94.0	97.1
SN++[13]	75.7	95.1	98.1
Ours	**77.69**	**96.56**	**98.33**

On the large-scale wild dataset Veri-Wild (experimental results are shown in Table 2), complex scenarios such as low resolution, severe occlusion, and cross-camera settings pose severe challenges to existing methods. Early Triplet-based methods only achieve an mAP of 15.69 and a CMC1 of 44.67 (3000-scale), which struggle to adapt to feature distribution shifts. Methods like GoogleNet and GSTE show improvements, but their maximum mAP is only 31.42 under the 3000-scale setting. Even the relatively new CTCAL, which raises the 3000-scale mAP to 70.35% and CMC1 to 83.64% via a feature calibration strategy, still fails to fully balance fine-grained texture and global structural features.

Our algorithm demonstrates stronger robustness: under the 3000-scale setting, it achieves 82.34% mAP, 94.35% CMC1, and 98.56% CMC5; under the 5000-scale setting, it still maintains excellent performance with 78.94% mAP and 93.43% CMC1. Its core advantages lie in multi-directional features covering full-dimensional information, the DOConv module balancing local and global features, and positional encoding

enhancing the correlation of key components—all of which effectively adapt to feature variations in wild scenarios.

Table 2: Performance Comparison on the Veri-Wild Dataset

Method	3000			5000		
	mAP	CMC1	CMC5	mAP	CMC1	CMC5
GoogleNet	24.27	57.61	75.13	24.15	53.16	71.13
Triplet[41]	15.69	44.67	63.33	13.34	40.34	58.98
GSTE	31.42	60.48	80.13	26.18	52.13	74.98
AAVER[22]	62.35	75.86	92.76	53.56	68.23	88.79
CTCAL[42]	70.35	83.64	92.63	65.73	80.31	90.75
Ours	**82.34**	**94.35**	**98.56**	**78.54**	**93.10**	**97.61**

Vehicle query matching results are shown in Figure 5. Results are sorted in descending order of computed similarity, with the top 5 matches returned. Here, green boxes indicate results consistent with the query vehicle's identity, while red boxes denote mismatches. The visualization shows that MFFP Net achieves higher retrieval accuracy than the baseline.

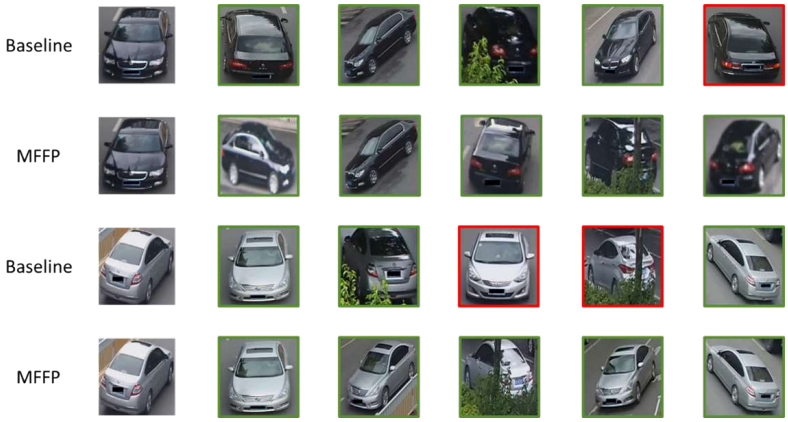


Fig. 5: Illustration of presentation of matching results of the baseline network and MFFP Net on the VeRi-776 dataset.

6 Ablation Experiments

To verify the role of each module in the network, ablation experiments are conducted on the VeRi-776 and Veri-Wild datasets to analyze the effectiveness of each module. The metrics mAP, CMC1, and CMC5 are adopted to determine the optimal architecture of the Multi-directional Feature and Position-Aware Fusion Network. The ablation experiments of the model on VeRi-776 are shown in Table 3.

Table 3: Quantitative Results of Comparative Schemes on the VeRi-776 Dataset

	Baseline	MFE	DOConv	PE	mAP	CMC1	CMC5
1	✓				74.40	95.46	97.81
2	✓	✓			75.97	96.24	98.33
3	✓	✓	✓		77.16	96.42	98.46
4	✓	✓	✓	✓	**77.69**	**96.66**	**98.53**

The ablation experiments of the model on VeRi-776 are shown in Table 3. The mAP, CMC1, and CMC5 of the Baseline are 75.82%, 96.40%, and 98.33%, respectively. After adding the MFE module, the mAP increased slightly by 0.41%; when the DOConv module was added, the mAP increased by 1.16%, indicating that the DOConv module has a more prominent optimization effect on balancing local and global features. After introducing the complete model, the performance reached the optimal level, with mAP of 77.89%, CMC1 of 96.66%, and CMC5 of 98.33%. All metrics are significantly improved compared with the Baseline, which confirms that the synergy of the three modules can effectively enhance feature discriminability and suppress background interference.

Figure 6 shows the training trends of the baseline model (represented by the red line) and the MFFP model (represented by the blue line) on the VeRi-776 dataset. The left panel is the CMC curve, where both curves rise rapidly and stabilize during iteration; the right panel is the mAP curve, where the MFFP model's curve remains superior, reflecting its better convergence and feature discriminability on this dataset.

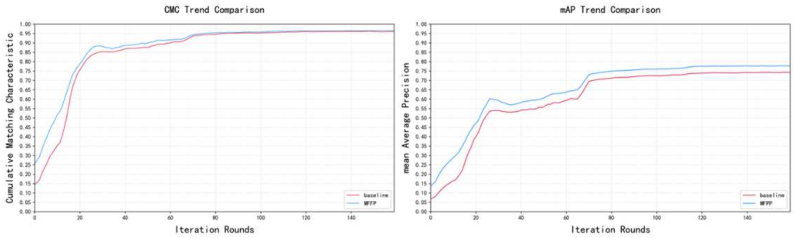


Fig. 6: Line Chart of Evaluation Metrics Based on the VeRi-776 Dataset

The ablation experiments on Veri-Wild are shown in Table 4. As a large-scale wild dataset, its data is more complex including low-resolution and severely occluded samples. Figure 7 presents the metric changes of the baseline model (represented by the red line) and the MFFP model (represented by the blue line) on the large-scale Veri-Wild dataset. Both the left CMC curve and right mAP curve indicate that the MFFP model outperforms the baseline in convergence speed and final metrics, verifying its robustness in complex unconstrained scenarios. Under the 3000-scale evaluation, the mAP, CMC1, CMC5 of the Baseline are 81.41%, 93.81%, and 97.96% respectively; all metrics improved after adding a single module, and the complete model further

Table 4: Quantitative Results of Comparative Schemes on the Veri-Wild Dataset

Schemes	Baseline	MFE	DOConv	PE	3000			5000		
					mAP	CMC1	CMC5	mAP	CMC1	CMC5
1	✓				80.41	93.21	97.83	75.43	92.21	97.15
2	✓	✓			80.85	93.54	98.09	76.26	92.41	97.31
3	✓	✓	✓		81.28	94.01	98.33	76.97	92.73	97.49
4	✓	✓	✓	✓	**82.31**	**94.64**	**98.49**	**77.94**	**93.43**	**97.79**

boosted them to 83.31%, 94.64%, and 98.49%. Under the 5000-scale setting, which features a larger sample size and more complex scenarios, the mAP of the baseline dropped to 76.56%, yet the complete model still achieved a stable improvement with an mAP of 77.88% and a CMC1 of 93.21%. This demonstrates that the module synergy mechanism can still optimize performance in extreme scenarios, reflecting the strong generalization ability of the model.

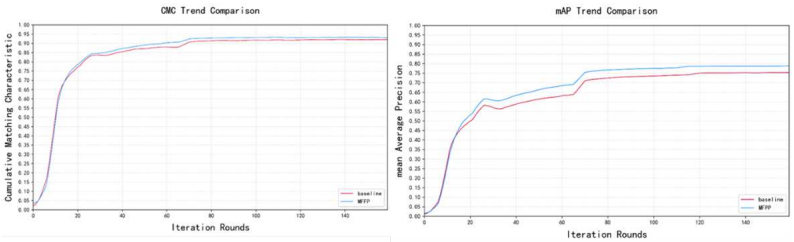


Fig. 7: Line Chart of Evaluation Metrics Based on the Veri-Wild Dataset

6.1 Conclusion

Vehicle re-identification faces critical challenges including viewpoint variations, illumination changes, occlusions, intra-class inconsistency, and inter-class similarity, while existing methods suffer from single-dimensional feature capture, insufficient positional information utilization, limited robustness, and complexity-generalization conflicts. To address these issues, this paper proposes MFFP-Net, a vehicle re-identification algorithm built on the ResNet50-IBN backbone, integrating three synergistic core modules: the Multi-directional Feature Extraction Module (MFE) with customized depthwise separable convolution kernels for comprehensive directional feature capture, the DOConv module fusing depthwise and traditional convolution to balance local-global features without additional sub-networks, and the Positional Encoding Module (PE) with horizontal-vertical tensor initialization to enhance positional perception, along with a combined loss function of cross-entropy and Batch-Hard Triplet Loss. Quantitative experiments show MFFP-Net achieves state-of-the-art performance: 77.89% mAP and 96.66% CMC1 on VeRi-776, 83.31% mAP and 94.64% CMC1 (3000-scale) as well as 77.88% mAP and 93.21% CMC1 (5000-scale) on Veri-Wild. Qualitative results

confirm its higher retrieval accuracy and stronger robustness in complex scenarios, while ablation experiments verify the synergistic effectiveness of the three modules, providing an efficient and robust solution for real-world vehicle re-identification.

References

[1] Wang, Q., Min, W., Han, Q., *et al.*: Inter-domain adaptation label for data augmentation in vehicle re-identification. IEEE Transactions on Multimedia **24**, 1031–1041 (2021)

[2] Tang, Z., Naphade, M., Liu, M.-Y., *et al.*: Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8797–8806 (2019)

[3] He, Q., Lu, Z., Wang, Z., *et al.*: Graph-based progressive fusion network for multi-modality vehicle re-identification. IEEE Transactions on Intelligent Transportation Systems **24**(11), 12431–12447 (2023)

[4] Liu, X., Liu, W., Ma, H., *et al.*: Large-scale vehicle re-identification in urban surveillance videos. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE, ??? (2016)

[5] Lou, Y., Bai, Y., Liu, J., *et al.*: Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3235–3243 (2019)

[6] Zheng, Z., Ruan, T., Wei, Y., *et al.*: Vehiclenet: Learning robust visual representation for vehicle re-identification. IEEE Transactions on Multimedia **23**, 2683–2693 (2021)

[7] Qian, J., Jiang, W., Luo, H., *et al.*: Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. Journal of Physics: Conference Series **1515**(9), 095401 (2020)

[8] Peng, J., Wang, H., Xu, F., *et al.*: Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification. Neurocomputing **401**, 133–144 (2020)

[9] Teng, S., Zhang, S., Huang, Q., *et al.*: Multi-view spatial attention embedding for vehicle re-identification. IEEE Transactions on Circuits and Systems for Video Technology **31**, 816–827 (2020)

[10] Gao, F., *et al.*: Incorporating intra-class variance to fine-grained visual recognition. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 1452–1457. IEEE, ??? (2017)

[11] Wang, Z., Tang, L., Liu, X., *et al.*: Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: IEEE International Conference on Computer Vision (ICCV), pp. 379–387 (2017)

[12] Cormier, M., Sommer, L.W., Teutsch, M.: Low resolution vehicle re-identification based on appearance features for wide area motion imagery. In: 2016 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 1–7. IEEE, ??? (2016)

[13] Li, K., Ding, Z., Li, K., *et al.*: Vehicle and person re-identification with support neighbor loss. IEEE Transactions on Neural Networks and Learning Systems **33**(2), 826–838 (2022)

[14] Wang, H., Peng, J., Jiang, G., *et al.*: Discriminative feature and dictionary learning with part-aware model for vehicle re-identification. Neurocomputing **438**, 55–62 (2021)

[15] Xu, Y., Rong, L., Zhou, X., *et al.*: Joint multiple fine-grained feature for vehicle re-identification. Array **14**, 100152 (2022)

[16] Zhang, J., Almeida Echen, J., Cao, J., *et al.*: Dual attention granularity network for vehicle re-identification. Neural Computing and Applications **34**(4), 2953–2964 (2022)

[17] Sun, W., Dai, G., Zhang, X., *et al.*: TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification. IEEE Transactions on Intelligent Transportation Systems **23**(9), 14557–14569 (2021)

[18] Zhang, X., Zhang, R., Cao, J., et al.: Part-guided attention learning for vehicle re-identification. arXiv preprint (2019). https://doi.org/10.48550/arXiv.1909.06023

[19] He, S., Luo, H., Wang, P., *et al.*: TransReID: Transformer-based object re-identification. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 103–107 (2021)

[20] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)

[21] Jiang, G., Pang, X., Tian, X., *et al.*: Global reference attention network for vehicle re-identification. Applied Intelligence **53**(9), 11328–11343 (2023)

[22] Khorramshahi, P., Kumar, A., Peri, N., *et al.*: A dual-path model with adaptive attention for vehicle re-identification. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6132–6141 (2019)

[23] Zhou, Y., Shao, L.: Cross-view GAN based vehicle generation for re-identification.

In: British Machine Vision Conference (BMVC), pp. 1–12 (2017)

[24] Zhang, F., Ma, Y., Yuan, G., *et al.*: Multiview image generation for vehicle re-identification. Applied Intelligence **51**(8), 5665–5682 (2021)

[25] Lou, Y., Bai, Y., Liu, J., *et al.*: Embedding adversarial learning for vehicle re-identification. IEEE Transactions on Image Processing **28**(8), 3794–3807 (2019)

[26] Tumrani, S., Ali, W., Kumar, R., *et al.*: View-aware attribute-guided network for vehicle re-identification. Multimedia Systems **29**(4), 1853–1863 (2023)

[27] Sun, W., Chen, X., Zhang, X., *et al.*: A multi-feature learning model with enhanced local attention for vehicle re-identification. Computers, Materials & Continua **69**(3), 3550–3561 (2021)

[28] Wei, R., Gu, J., He, S., *et al.*: Transformer-based domain-specific representation for unsupervised domain adaptive vehicle re-identification. IEEE Transactions on Intelligent Transportation Systems **24**(3), 2935–2946 (2023)

[29] Almeida, E., Silva, B., Batista, J.: Strength in diversity: Multi-branch representation learning for vehicle re-identification. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). IEEE, ??? (2023)

[30] Jin, Y., Li, C., Li, Y., *et al.*: Model latent views with multi-center metric learning for vehicle re-identification. IEEE Transactions on Intelligent Transportation Systems **22**(3), 1919–1931 (2021)

[31] Tumrani, S., Deng, Z., Lin, H., *et al.*: Partial attention and multi-attribute learning for vehicle re-identification. Pattern Recognition Letters **138**, 290–297 (2020)

[32] He, K., Fan, H., Wu, Y., *et al.*: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738 (2020)

[33] Wang, Y., Wei, Y., Ma, R., *et al.*: Unsupervised vehicle re-identification based on mixed sample contrastive learning. Signal, Image and Video Processing **16**(8), 2083–2091 (2022)

[34] Cheng, D., Wang, X., Wang, N., *et al.*: Cross-modality person re-identification with memory-based contrastive embedding. In: AAAI Conference on Artificial Intelligence, pp. 425–432 (2023)

[35] Shen, F., Xie, Y., Zhu, J., *et al.*: GIT: Graph interactive transformer for vehicle re-identification. IEEE Transactions on Image Processing **32**, 1039–1051 (2023)

[36] Chu, R., Sun, Y., Li, Y., *et al.*: Vehicle re-identification with viewpoint-aware metric learning. In: IEEE/CVF International Conference on Computer Vision

(ICCV), pp. 8282–8291 (2019)

[37] Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint (2017)

[38] Yang, L., Luo, P., Change Loy, C., *et al.*: A large-scale car dataset for fine-grained categorization and verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3973–3981 (2015)

[39] Bai, Y., Lou, Y., Gao, F., *et al.*: Group-sensitive triplet embedding for vehicle re-identification. IEEE Transactions on Multimedia **20**(9), 2385–2399 (2018)

[40] Qiao, W., Ren, W., Zhao, L.: Vehicle re-identification in aerial imagery based on normalized virtual Softmax loss. Applied Sciences **12**(9), 4731 (2022)

[41] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)

[42] Yu, J., Kim, J., Kim, M., *et al.*: Camera-tracklet-aware contrastive learning for unsupervised vehicle re-identification. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), pp. 905–911 (2022)