Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 1, 2025

Vehicle Re-Identification Based on Wavelet Feature Enhancement and Global-Local Differential Attention Fusion

Bochi Zhu *, Haifeng Sang

School of Information Science and Engineering, Shenyang University of Technology, Shenyang, Liaoning 110870, China
* **Corresponding author**: Bochi Zhu (Email: Sut_22@smail.sut.edu.cn)

Abstract: In the continuous evolution of intelligent transportation systems, vehicle re-identification technology faces numerous technical challenges, including variations in perspective and equipment resolution. These factors lead to significant intra-class discrepancies in the performance of identical vehicles under varying conditions, as well as inter-class confusion among vehicles with similar appearances. To address these challenges, we integrate vehicle color and type attribute information, enhancing the model's ability to capture semantic features and improve its discriminative performance. Additionally, we propose a wavelet feature enhancement module that employs wavelet transform to decompose images at multiple scales, effectively capturing fine-grained features such as edges and textures. This enables the model to better represent intricate visual details. Finally, we introduce a differential attention mechanism that combines global and local features, strengthening contextual understanding through interactive feature modeling. Experimental results demonstrate the effectiveness of our approach, achieving a Rank-1 accuracy of 97.0% on the VeRi-776 dataset and 85.2% on the VehicleID dataset, outperforming existing methods and highlighting the efficacy of our proposed framework.

Keywords: Attribute aggregation, Swin transformer, Vehicle re-identification, Wavelet transform.

1. Introduction

In real-world traffic surveillance scenarios, the images of vehicles captured are inevitably influenced by various factors, including changes in viewing angles and the differing resolutions of camera equipment. These factors can result in intra-class variations for the same vehicle observed from different perspectives, as well as inter-class similarities among different vehicles of the same model, thereby presenting significant challenges for researchers involved in vehicle re-identification tasks. Earlier methodologies predominantly relied on convolutional neural networks to extract global features from vehicle images, frequently utilizing classical backbone architectures such as ResNet in conjunction with global pooling layers. He et al. [1] pioneered the integration of Vision Transformers (ViT) [2] into the field of vehicle re-identification. ViT utilizes a multi-head self-attention mechanism to effectively capture global contextual information. However, its reliance on fixed-sized patches and lack of hierarchical representation limits its ability to capture fine-grained local features, which are crucial for distinguishing visually similar vehicles. Liu et al. [3] further advanced this field by implementing the Swin Transformer, which applies a sliding window approach to divide images into localized regions and performs self-attention computations within each region, enhancing the extraction of detailed local features. However, while Swin Transformer's sliding window approach prioritizes local feature extraction, it offers limited interaction among these local features. Consequently, there is an over-reliance on localized details, which undermines the model's capacity to recognize relationships among features, thereby negatively impacting its performance in distinguishing similar vehicles during re-identification tasks. To address this limitation, Qian et al. [4] proposed a strategy that categorizes input features into global and local types. While this approach partially alleviates the issues associated with the independent processing of these feature types, it still fails to establish effective interactions between them and neglects potential correlations across different local regions. In response to these challenges, we have developed a global-local differential attention mechanism designed to enhance interaction between global and local features while maintaining their correlations. This mechanism aims to reduce excessive dependence on localized attributes and improve the overall quality of feature representation. In recent years, Jaderberg et al. [5] introduced random spatial transformations to image datasets as a means of augmenting diversity and enhancing model robustness. However, these random alterations may lead to the loss of critical image features, particularly when excessive zooming results in the obliteration of important details. Subsequently, Phan et al. [6] proposed a method that involves random cropping of input images, aimed at improving model resilience against local occlusions. Nevertheless, the removal of even small portions of an image can significantly compromise essential details, such as license plates or vehicle tags, thereby adversely affecting identification accuracy. In order to address this issue, we propose the implementation of a Wavelet feature extraction module, which incorporates wavelet transform techniques to mitigate the challenges of feature loss and feature inconsistency. Furthermore, the attribute information associated with vehicles is crucial for re-identification tasks. Yu et al. [7] utilized two-dimensional convolutions in conjunction with channel attention mechanisms to aggregate attributes such as color, vehicle type, and perspective. While capturing dependencies among attributes is advantageous, this straightforward concatenation approach inadequately addresses the hierarchical structures and fine-grained variations inherent within the attribute data itself. Additionally, traditional convolution operations

primarily focus on spatial features, often neglecting frequency-based insights (e.g., texture and detail), which may result in the loss of valuable information due to overly simplistic spatial-domain manipulations. To address this limitation, we propose an innovative fusion methodology that combines channel attention mechanism with spatial attention mechanism. By integrating coded attribute representations with visual imagery, this technique avoids the pitfalls of traditional splicing, ensuring the retention of finer-grained characteristic distinctions and thereby enhancing the expressive capabilities related to vehicular attributes.

The contributions of our paper are as follows:

(1) The Global-Local Differential Attention Mechanism (GDAT) facilitates interactive modeling between global and local features, addressing the limitations of independently processing these features. By enabling comprehensive feature interaction, it improves the model's capacity to distinguish vehicles with high inter-class similarity.

(2) The Wavelet Feature Enhancement Module (WFE) employs wavelet transform and convolution to perform multi-scale decomposition of vehicle images, enabling the extraction of edge and texture details across various frequency bands. This approach resolves the problem of feature detail loss and markedly enhances the ability to discern fine-grained appearance details of vehicles.

(3) We propose an Attribute Information Fusion Mechanism (ATF) that integrates vehicle attributes, such as color and type, with image features. By incorporating Channel and Spatial Attention Module (CBAM), which leverages both channel and spatial attention mechanisms, this method effectively addresses the issue of information loss and significantly improves re-identification accuracy in complex scenarios.

2. Related Work

In the field of vehicle re-identification, early methodologies predominantly utilized convolutional neural networks to extract global features from images. The models developed by Simonyan K et al. utilizing VGGNet [8], Szegedy C et al. employing GoogLeNet [9], and He K et al. leveraging ResNet [10] are grounded in classical convolutional neural network architectures, which serve as the foundational backbone structures. However, although these convolutional networks promoted the development of vehicle re-identification technology at the initial stage, with the increase of task complexity, the extraction of global features only showed limitations when dealing with scenes such as cross-camera perspectives and complex angular variations. To overcome these limitations, Transformer [11] have been introduced for visual tasks. ViT was first proposed, which divides images into multiple patches and processes them using tokenization strategies similar to NLP tasks. This methodology facilitates effective feature extraction on a global scale and establishes a theoretical foundation for subsequent visual tasks. Following this, Li et al. [12] designed an efficient multi-perspective Transformer. By optimizing local and global feature alignment under different perspectives, the feature expression consistency of vehicles in multi-perspective environments can be ensured, so that the model can maintain excellent identification effect across perspectives. Lian et al. [13] use Transformer to enhance the spatial attention mechanism of the model, enabling it to concentrate on both global and local discriminative features across different camera perspectives. Nevertheless, despite the introduction of local segmentation and global extraction strategies to capture global and local features of vehicles, these models exhibit deficiencies in the deep interaction and fusion of local and global features, particularly in scenarios involving similar vehicles across different regions. To enhance the visual representation of models, recent studies have explored the integration of frequency transformation techniques with deep learning architectures. Li et al. [14] proposed SFPFusion, which employs a hyper-feature attention mechanism and wavelet-guided pooling to enhance the extraction of both global and local image features, demonstrating significant improvements in image fusion tasks. Wang et al. [15] introduced a residual attention network based on wavelet decomposition, where images are separated into high- and low-frequency components via a two-dimensional wavelet transform, effectively capturing subtle variations and improving super-resolution reconstruction outcomes. More recently, Yao et al. [16] presented the Wavelet Vision Transformer, which combines reversible downsampling and wavelet decomposition within a self-attention framework, enabling lossless downsampling through inverse transformation and promoting multi-scale feature fusion. While these methods leverage frequency decomposition to enhance edge and detail features of images to varying degrees, most lack comprehensive fusion of diverse frequency features and fail to fully integrate attention mechanisms to enhance the selectivity of detailed features. As a result, their ability to mine and identify fine-grained vehicle texture features remains limited in vehicle re-identification tasks. In the vehicle re-identification task, in addition to the progress of feature extraction technology, alongside the use of vehicle attribute information also significantly enhanced the accuracy of vehicle re-identification under cross-camera and multi-perspective. Jiang et al. [17] proposed an attribute fusion network, which combines multiple vehicle attribute features with global features in series to enhance the overall description ability of the model. Yu et al. [7] proposed the Vehicle Attribute Transformer by embedding attributes such as color and vehicle type into a Transformer structure to capture intricate interactions between attribute features and global vehicle features through self-attention mechanism, and finally improve the recognition effect and robustness of the network. Despite the promising results achieved through these methods in vehicle attribute fusion, most existing studies employ simplistic concatenation techniques to integrate attributes like color and vehicle type, and lacks in-depth modeling of complex relationships between attributes and both local and global features. This limitation hinders the deep application of attribute information in intricate traffic scenarios, thereby adversely affecting the representation and matching accuracy of fine-grained vehicle features.

3. Organization of the Text

We propose a vehicle re-identification model that utilizes the Swin Transformer as the backbone, which combines a global feature modeling module, a local feature modeling module, and a wavelet feature-enhanced convolution module, with the objective of effectively capturing both global and local features of vehicle images. The overall architecture of the model is shown in Figure 1.

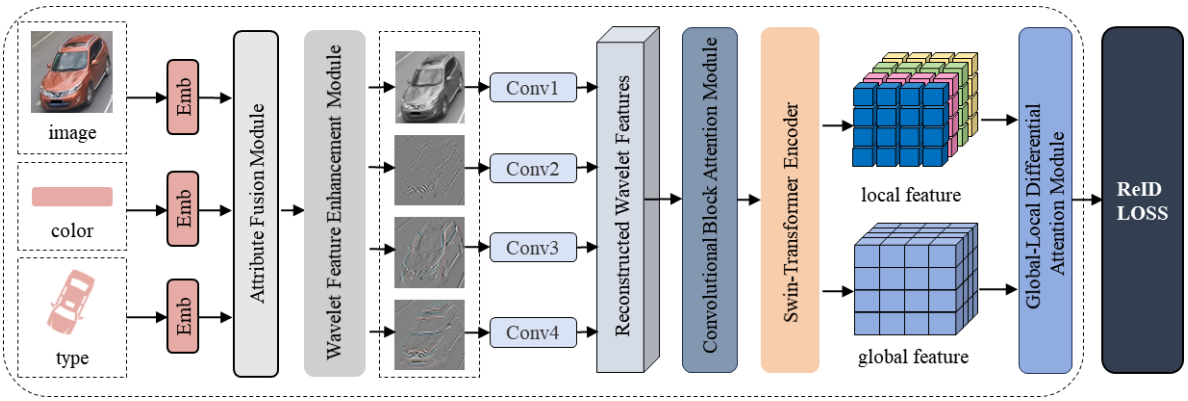

Figure 1. The proposed comprehensive model for vehicle re-identification

Fig. 1 illustrates the framework of the proposed model, which is composed of several interconnected modules to achieve effective vehicle re-identification. First, vehicle images and attribute information are independently encoded. These representations are fused through the proposed ATF, resulting in integrated features. The fused features are then processed by the WFE to extract multi-scale wavelet features. Subsequently, these features undergo convolution operations and are reconstructed via the reconstructed wavelet module, producing feature maps that are further refined using the CBAM attention mechanism to match the dimensionality of the input. The resulting features are fed into the Swin Transformer, which extracts global and local features. Finally, the features are passed through the proposed Global-Local Differential Attention Mechanism to enhance feature interaction and discrimination, followed by a classification module for the re-identification task.

3.1. Attribute Aggregation Module

Initially, the visual representation of the vehicle, denoted as $F_{visual} \in R^{H\times W\times C}$, serves as the input for the model, where H and W correspond to the height and width of the image, respectively, and the value of C indicates the RGB color channel. The input image contains the visual information of the vehicle, which forms the foundation for subsequent processing. Building upon this foundation, we introduce an ATF that focuses on vehicle color and type to augment the model's comprehension of vehicle attributes. To ensure that the attribute information operates synergistically with the visual features of the image, we define embedding functions, referred to as $\phi_{embedding}$, to map the features of color and vehicle type into a high-dimensional feature space. The specific formulation is as follows:

$$F_{color\_embedding} = \phi_{embedding}(A_c) \quad (1)$$

$$F_{type\_embedding} = \phi_{embedding}(A_t) \quad (2)$$

Where A_c and A_t denote the attribute information of color and vehicle type, respectively. The embedded color feature, denoted as $F_{color\_embedding} \in R^{H\times W\times d_c}$, and the vehicle type feature, represented as $F_{type\_embedding} \in R^{H\times W\times d_t}$. Additionally, d_c and d_t are the embedding dimensions for color and vehicle type, respectively. Subsequently, the embedded features for color and vehicle type are integrated with the visual feature F_{visual} derived from the original image, resulting in an enhanced joint feature representation:

$$F = F_{visual} + F_{color\_embedding} + F_{type\_embedding}$$

This feature fusion strategy not only enriches the model's feature representation but also significantly enhances its ability to comprehend semantic details. Through collaborative mapping within the semantic space, image features and attribute information are closely combined in multiple dimensions, enabling the model to capture the details of vehicle appearance and to perceive the semantic information pertaining to the vehicle more comprehensively.

3.2. Wavelet Feature Enhancement Module

After the attribute fusion module, the features are further transferred to the WFE. The core design of the WFE is to use discrete wavelet transform to decompose the spatial frequency information of the input image and extract the detailed features of different scales, so as to enhance the perception ability of the model. The input tensor is characterized by the shape $F \in R^{H\times W\times C}$. Initially, the input feature F is processed using a two-dimensional Discrete Wavelet Transform, which decomposes it into four frequency subbands:

$$W = DWT(S) \quad (3)$$

Where, DWT (·) denotes a two-dimensional discrete wavelet transform. $W = \{W_{LL}, W_{LH}, W_{HL}, W_{HH}\}$. W_{LL} is the low-frequency subband, which mainly contains the global information of the image, and W_{LH}, W_{HL}, W_{HH} are the high-frequency subbands, which capture local edges and details within the image. The low-frequency component preserves the overall structure of the image, while the high-frequency component enhances the capture of the image edges and textures. To further augment the feature representation capabilities, we apply a convolution operation to each subband following decomposition, thereby strengthening the features of different frequency components. For each subband W_i, where $i \in \{LL, LH, HL, HH\}$, the convolution operation is expressed as follows:

$$W_i' = Conv(W_i) \quad (4)$$

The convolution operation is specifically designed to enhance the extraction of both global and local features, particularly in capturing the appearance details of the vehicle such as shape and texture. The convolutional subbands are reconstructed via Reconstructed Wavelet Transform (RDWT) to restore the original input features:

$$S' = \mathrm{RDWT}\left(W_{LL}', W_{LH}', W_{HL}', W_{HH}'\right) \quad (5)$$

RDWT serves as the inverse wavelet transform, effectively reconstructing an image from its subbands after wavelet decomposition and convolution operations. In this process, both low-frequency and high-frequency information are comprehensively integrated, thereby preserving the global information of the image while simultaneously enhancing the capture of intricate details. Subsequently, the fused features are passed to the CBAM, which employs an adaptive mechanism to selectively amplify significant features and diminish redundant information. CBAM is comprised of two components: channel attention and spatial attention. The channel attention mechanism assesses the importance of features along the channel dimension through global average pooling and max pooling operations, whereas the latter selectively emphasizes key parts in the spatial dimension through the global information present in the feature map:

$$F_{cbam} = f\left(S'\right) \quad (6)$$

Here, $f(\cdot)$ denotes the application of the CBAM attention mechanism. Ultimately, the dimensions of the output tensor remain consistent with those of the input tensor. By integrating wavelet transforms with convolutional operations, the model can make full use of the frequency information in the image. Notably, with the adaptive enhancement mechanism of the CBAM module, the model can more precisely capture both global information and intricate features of the vehicle. This capability significantly enhances the accuracy and robustness of vehicle re-identification.

3.3. Global-Local Differential Attention Feature Modeling

3.3.1. Global Feature Modeling

Initially, the model extracts global features utilizing the Swin Transformer. The output features from the Swin Transformer are denoted as F. To further aggregate global information, the model introduces a learnable global cls token, Q_{global}, into the extracted global features, which serves to aggregate global context within the self-attention mechanism. During this process, the key matrix K_{global} and the value matrix V_{global} are generated from features extracted by the Swin Transformer.

The fundamental aspect of global feature modeling is the differential attention mechanism. In particular, two sets of features, Q'_{global}, K'_{global}, V'_{global} and Q''_{global}, K''_{global}, V''_{global}, are derived from the dimensional segmentation of Q_{global}, K_{global}, and V_{global}. Subsequently, the two sets of features are computed independently for attention, resulting in two outputs are obtained:

$$out_1 = \mathrm{Attention}\left(Q'_{global}, K'_{global}, V'_{global}\right) \quad (7)$$

$$out_2 = \mathrm{Attention}\left(Q''_{global}, K''_{global}, V''_{global}\right) \quad (8)$$

Then, the learnable parameters λ_{q1}, λ_{k1}, λ_{q2}, λ_{k2} are introduced to regulate the attention contributions from various components. The dot product summation of λ_{q1} and λ_{k1}, λ_{q2} and λ_{k2} element by element:

$$\lambda_{q1k1} = \lambda_{q1}[i] \cdot \lambda_{k1}[i], \lambda_{q2k2} = \lambda_{q2}[i] \cdot \lambda_{k2}[i] \quad (9)$$

The two results are then exponentially added together to give the final global weight λ_{global}:

$$\lambda_{global} = e^{\lambda_{q1k1}} + e^{\lambda_{q2k2}} + \lambda_{init} \quad (10)$$

Where λ_{init} is a constant, out_1 and out_2 are weighted to get the final global output:

$$out_{global} = out_1 + \lambda_{global}\, out_2 \quad (11)$$

Through this mechanism, the interaction between cls token and the global feature can get the global feature representation, thereby effectively capturing the overall appearance information of the vehicle.

3.3.2. Global Feature Modeling

This partial model divides the output feature F of the Swin Transformer into four distinct local regions, with the features of each local region denoted as $F_{locali,}$ i=1, 2, 3, 4. In order to capture local features, the model introduces an independent learnable local cls token for each local region, whose query matrix is Q_{locali}. This local cls token participates in the construction of the value matrix V_{locali}. In local feature modeling, the query matrix Q_{global} and the key matrix K_{global} are derived from the global features, thereby ensuring that the model can capture global context information, and the value matrix V_{locali} is derived from the features of this local region. Similar to global modeling, we divide Q_{global}, K_{global}, and V_{locali} to obtain two sets of features, Q'_{global}, K'_{global}, V'_{locali} and Q''_{global}, K''_{global}, V''_{locali} then perform attention calculations on these two sets of features to get out3 and out4.

Next, similar to global modeling, we define the learnable parameter λ_{q3}, λ_{k3}, λ_{q4}, λ_{k4}, use the same calculations to get λ_{q3k3} and λ_{q4k4}, and by applying exponential weighting, we obtain the final local weight:

$$\lambda_{local_i} = e^{\lambda_{q3k3}} + e^{\lambda_{q4k4}} + \lambda_{init} \quad (12)$$

Finally, the sum of out_3 and out_4 is weighted by λ_{locali} to get the local output out_{locali}:

$$out_{local_i} = out_3 + \lambda_{local_i}\, out_4 \quad (13)$$

Through this attention mechanism, local features are capable of embedding global contextual information while simultaneously preserving local detail representation to obtain local feature representation.

3.3.3. Global Feature Modeling

To effectively integrate global and local features, the model first performs global average pooling operations on the global features F and the local features F_{locali} (where i=1, 2, 3, 4) extracted by the backbone. Next, the features obtained through the Global-Local Differential Attention mechanism and the pooled features are added to obtain F^*_{global} and F^*_{locali}. Finally, the local features are concatenated and added to the global features to produce the final comprehensive feature representation:

$$F_{final} = F^*_{global} + \mathrm{Concat}\left(F^*_{local_1}, F^*_{local_2}, F^*_{local_3}, F^*_{local_4}\right) \quad (14)$$

Through the integration of global-local features, the model is capable of preserving the overall contour information of the vehicle while simultaneously capturing the nuanced differences among various components of the vehicle. Figure 2 provides an illustration of the Global-Local Differential Attention Mechanism.

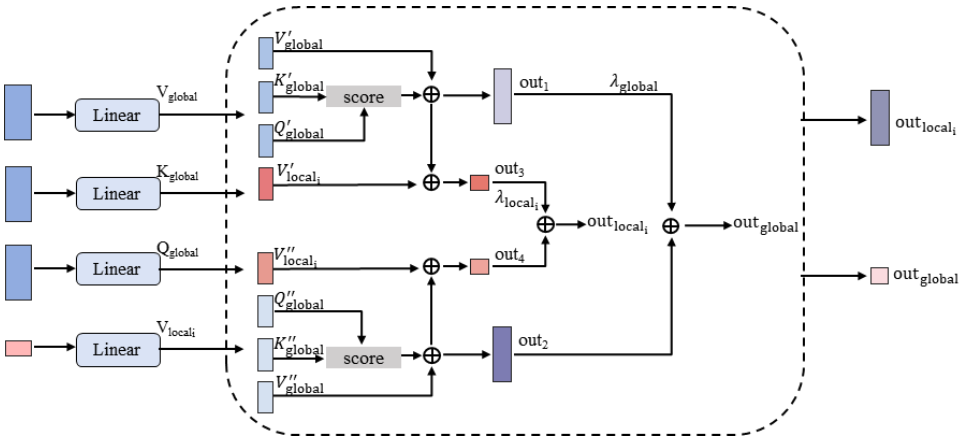

Figure 2. The proposed global-local differential attention mechanism

3.4. Loss Functions

In the task of vehicle re-identification, we adopt a comprehensive loss function that integrates cross-entropy loss and triplet loss to effectively optimize model performance. This approach is designed to enhance the model's discriminative capability and optimize inter-class distances. The cross-entropy loss function is employed in the classification task of vehicle ID to ensure the accuracy of the model on vehicle ID classification. By calculating the difference between the predicted class probability distribution $\hat{y}$ and the true label y, the model is driven to optimize the feature representation of the vehicle. The goal of the cross-entropy loss function is to maximize the logarithmic probability of the true label, as follows:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log\left(\hat{y}_i\right) \tag{15}$$

Where C is the number of categories, y_i is the one-hot encoding of the true label, and $\hat{y}_i$ is the class probability output generated by the model. Within the vehicle re-identification task, this loss function serves to effectively direct the model in acquiring the capability to differentiate between various vehicle identities, thereby establishing more distinct class boundaries within the feature space.

The triplet loss function is used to optimize both the intra-class and inter-class distances within the feature space. The input to the triplet loss comprises an anchor sample A, a positive sample P (the same vehicle), and a negative sample N (different vehicles). The objective is to minimize the distance between the anchor and the positive sample while simultaneously maximizing the distance between the anchor and the negative sample. The mathematical formulation of the triplet loss is presented as follows:

$$L_{Triplet} = \max\left(d(A,P) - d(A,N) + a, 0\right) \tag{16}$$

Where $d(\cdot)$ is the distance metric between samples, and α is a predetermined interval. This loss ensures that vehicle image representations within the feature space are more discriminative by directly optimizing within-class compactness versus between-class separability. Ultimately, the integration of cross-entropy loss and triplet loss results in a comprehensive loss function that not only preserves the accuracy of vehicle ID classification, but also enhances the separation ability between classes. The mathematical representation of the combined loss function is as follows:

$$L = L_{CE} + \lambda L_{triplet} \tag{17}$$

Here, λ is the weight coefficient that equilibrates the contributions of the two terms of loss. Through this design, the model can not only accurately distinguish different vehicles, but also construct a global-local feature distribution with strong discrimination within the feature space, thus achieving excellent performance in complex vehicle re-identification scenarios.

4. Experiment and Results

4.1. Evaluation Metrix

In the context of vehicle re-identification, we adopted the Mean Average Precision (mAP) and Rank values to conduct a comprehensive evaluation of the model's performance. These metrics provide a multifaceted assessment of the model's effectiveness and accuracy in retrieval tasks.

4.2. Datasets

Commonly utilized publicly available datasets for vehicle re-identification tasks include VeRi-776 and VehicleID, which provide test benchmarks for assessing vehicle re-identification across cameras.

(1) VeRi-776 dataset

The VeRi-776 dataset is a widely recognized dataset utilized in vehicle re-identification tasks, comprising over 50,000 images of 776 distinct vehicles. These images are captured from various camera perspectives and across different temporal contexts, with vehicles appearing in different environments and lighting conditions. The images of each car cover different angles, distances, and partial occlusions. The dataset also provides the ID annotation of each vehicle, camera ID, timestamp, and other information, which is suitable for assessing cross-camera vehicle re-identification tasks.

(2) VehicleID dataset

The VehicleID dataset is another important vehicle re-identification dataset, containing 221,736 images of 26,267 vehicles. This dataset focuses more on the front side and back images of vehicles, with high picture resolution and rich types of vehicles.

The dataset of VehicleID is assigned to different training and test sets according to the vehicle ID, where the test set is divided into various scales to evaluate the recognition performance of the model at different scales.

4.3. Datasets

All experiments were conducted on a single workstation equipped with an Intel® Core i5-12700KF CPU operating at 3.58 GHz, featuring 10 cores, and an Nvidia RTX 3080 Ti GPU. The implementation of the Transformer architecture utilized 4 attention heads and an embedding dimension of 768, with input tokens having n=196 and a feature dimension (dim) of 768. During the training phase, the Adam optimizer was employed with a batch size of 64. A decaying learning rate policy was adopted, utilizing a stochastic optimization approach, with an initial learning rate set at 0.01.

4.4. Experimental Analysis

In this section, we compare our approach with state-of-the-art methods applied to the VeRi-776 and VehicleID datasets, utilizing evaluation metrics including mAP and Rank-1. Our method demonstrates superior performance compared to all baseline models. This shows the model's capacity to effectively extract and represent features across varying scales and levels of feature importance, thereby improving its interpretative capabilities regarding the images. In tasks such as vehicle re-identification, our method significantly enhances the model's ability to recognize intricate vehicle details, increasing both the accuracy and robustness of the recognition process.

Table 1. Comparison with state-of-the-art methods

	VeRi-776		VehicleID	
Method	mAP	Rank-1	mAP	Rank-1
AAVER [18]	58.5%	88.7%	-	72.5%
PRM [19]	70.2%	92.2%	-	78.4%
FDA-Ne [20]	56.0%	84.3%	-	59.8%
VANet [21]	66.3%	89.8%	-	83.3%
SAN [22]	72.5%	93.3%	-	75.6%
UMTS [23]	75.9%	95.8%	87.0%	80.9%
MVAN [24]	72.5%	92.6%	-	-
MV-GAN [25]	63.2%	91.1%	-	74.5%
LABNet-50 [26]	79.5%	95.7%	87.5%	81.2%
TCPM [27]	75.0%	94.0%	85.1%	82.0%
TransReID [1]	78.0%	96.1%	82.9%	-
DSN [28]	76.3%	94.8%	81.7%	80.6%
URRNet [29]	72.2%	93.0%	-	76.5%
CFSA [30]	78.5%	94.3%	-	-
SSRNet [31]	78.3%	96.4%	89.1%	83.1%
Baseline	78.2%	96.5%	88.1%	82.9%
Baseline+WFE	78.4%	96.5%	88.7%	83.9%
Baseline+ATTR+WFE	78.4%	96.6%	88.9%	84.2%
our	79.2%	97.0%	89.4%	85.2%

In Table 1, Baseline refers to the configuration where the model solely utilizes our backbone, the Swin Transformer, for feature extraction. Our method refers to the complete framework, incorporating the ATTR module, WFE module, and GDAT module as the final integrated design.

In the experimental results presented in Table 1, our method is compared with several state-of-the-art models. In the VeRi-776 dataset, our model achieves a mAP of 79.2% and a Rank-1 metric of 97.0%. Additionally, it achieves 89.4% mAP and 85.2% Rank-1 in the VehicleID dataset. This significant performance improvement demonstrates the capability of our model to capture multi-level features and details of vehicle images and further verifies the effectiveness of our design strategy. The results presented in Table I indicate that our model demonstrates a significant improvement in performance when compared to the Baseline and other methodologies listed in the table, including AAVER, FDA-Ne, UMTS, and LABNet-50, among others. This performance benefits from the synergy of the multi-module design.

In comparison to the Baseline, both Baseline+ATF+WFE and Baseline+WFE demonstrate a notable enhancement in performance. For instance, on the VehicleID dataset, Baseline+WFE yields an improvement in mAP of 0.6%, resulting in a value of 88.7%. Additionally, the Rank-1 index rises to 83.9%. This improvement can be attributed to the frequency domain feature module's ability to capture multi-scale information from the image, thereby enhancing the model's effectiveness in identifying detailed features within complex scenes. Furthermore, Baseline+ATF+WFE significantly enhances the model's fine-grained discrimination capability, with the mAP and Rank-1 indexes increasing to 88.9% and 84.2%, respectively. These findings indicate that the various modules exert independent influences on detail capture and global feature representation, with the performance gains being particularly pronounced in complex scenes.

Module synergy compared with AAVER, FDA-Ne, UMTS, and other methods, our model significantly improves the recognition accuracy on the VeRi-776 and VehicleID datasets, showing stronger robustness. For example, UMTS achieves only 80.9% Rank-1 on the VehicleID dataset, while our model achieves 85.2%. This improvement enhances the expression ability of details through multi-module collaborative design. The introduction of the GDAT strengthens the interaction between global and local features, enabling the model to achieve more refined recognition ability by focusing attention on key parts. The wavelet transform provides multi-scale analysis, capturing richer vehicle details at different spatial scales. The attribute features further complement the model's semantic expression by embedding high-dimensional information. Therefore, through the collaboration between modules, the stronger adaptability and accuracy of our model are demonstrated.

4.5. Ablation Experiments

To evaluate the effectiveness of our model, ablation experiments are conducted on two datasets for several schemes proposed in this paper. In the ablation experiments, we maintained the other components constant while examining the various modules introduced.

(1) Validation of the Attribute Aggregation Module

To assess the efficacy of our proposed ATF in the vehicle re-identification task, we conducted experiments on the VeRi-776 and VehicleID datasets, as presented in Table 2. The experimental results show that the implementation of the ATF module results in an increase in the mAP and Rank-1 accuracy for the VeRi-776 dataset, achieving values of 78.3% and 96.3%, respectively. These results represent an enhancement of 0.4% for both metrics when compared to the baseline. Similarly, for the VehicleID dataset, the mAP and Rank-1 accuracy reached 88.6% and 83.5%, respectively, reflecting improvements of 0.5% and 0.6% over the baseline. The ATF module enhances the expression ability of detailed attributes by encoding the features of vehicle color and model attributes, and introduces the CBAM attention mechanism to further

optimize the selective expression of features, which significantly improves the fine-grained identification capabilities of vehicles.

Table 2. Ablation experiments of the ATF modules

	VeRi-776		VehicleID	
Method	mAP	Rank-1	mAP	Rank-1
baseline	77.9%	95.9%	88.1%	82.9%
baseline+color	78.1%	96.0%	88.4%	83.3%
Baseline+type	78.2%	96.1%	88.3%	83.1%
Baseline+ATF(CAT)1	76.5%	95.1%	87.3%	82.3%
baseline+ATF	78.3%	96.3%	88.6%	83.5%

In Table 2, an ablation study of the ATF module is conducted on the VeRi-776 and VehicleID datasets. 1 Denotes the concatenation of attribute information and image features.

(2) Validation of the Wavelet Processing Module

Ablation experiments were conducted using the WFE on the VeRi-776 and VehicleID datasets to validate its effectiveness in capturing detailed information, as shown in Table 3. Experiments show that the incorporation of the WFE module leads to an increase in the mAP and Rank-1 accuracy for the VeRi-776 dataset, achieving values of 78.4% and 96.5%, respectively. Similarly, for the VehicleID dataset, the mAP and Rank-1 accuracy improved to 88.7% and 83.9%, respectively. The WFE module utilizes wavelet transforms to extract multi-scale features across various frequencies, thereby enhancing the model's ability to identify details in complex backgrounds and multi-view scenarios. This approach provides a more nuanced representation of the key appearance features of vehicles.

Table 3. Ablation Experiments of the WFE Modules

	VeRi-776		VehicleID	
Method	mAP	Rank-1	mAP	Rank-1
baseline	77.9%	95.9%	88.1%	82.9%
baseline+wavelet1	78.2%	96.3%	88.5%	83.6%
baseline+WFE	78.4%	96.5%	88.7%	83.9%

In Table 3, An ablation study of the WFE module conducted on the VeRi-776 and VehicleID datasets.1 refers to the configuration where wavelet transform is applied to the image without incorporating the CBAM attention mechanism.

Table 4. Ablation Experiments of the WFE Modules

	VeRi-776		VehicleID	
Method	mAP	Rank-1	mAP	Rank-1
baseline	77.9%	95.9%	88.1%	82.9%
baseline+GDAT1	78.1%	96.2%	88.4%	83.9%
baseline+GDAT*	78.5%	96.5%	88.9%	84.1%
baseline+GDAT	78.8%	96.8%	89.1%	84.5%

In Table 4, An ablation study of the GDAT module conducted on the VeRi-776 and VehicleID datasets.1 refers to the process within the Global-Local Differential Attention Mechanism where global and local features are subjected to attention calculations independently, without any interaction between them. GDAT* indicates that Global-Local Differential Attention Mechanism has not undergone differential attention operations.

(3) Validation of the Global-Local Differential Attention Module

To evaluate the impact of the GDAT on model performance, we conducted ablation experiments utilizing two datasets, as presented in Table 4. The introduction of the GDAT module resulted in a mAP of 78.8% and a Rank-1 accuracy of 96.8% for the VeRi-776 dataset. In contrast, the mAP and Rank-1 accuracy for the VehicleID dataset increased to 88.9% and 84.1%, respectively. By facilitating the interactive modeling of global and local features, the GDAT module achieves a comprehensive integration of these information types, addressing the limitations of conventional methodologies that process global and local features separately. This enhancement contributes to the improved robustness and discriminative capability of the model in complex traffic scenarios.

5. Conclusion

In this paper, we introduce a novel vehicle re-identification model that significantly enhances performance through the integration of three distinct modules: the ATF, the WFE, and the GDAT modules. The ATF module improves the model's sensitivity to detailed vehicle features by encoding attribute information, such as color and type, and integrating it with the image features. The WFE module employs wavelet transform to conduct multi-scale decomposition of the image in the frequency domain, thereby capturing more nuanced global and local information, which enhances the model's ability to discern subtle differences. Lastly, the GDAT module enhances feature representation adaptively through a Global-Local Differential Attention mechanism, thereby augmenting the model's ability to capture essential information while minimizing the influence of irrelevant data. Comprehensive experiments conducted on the VeRi-776 and VehicleID datasets substantiate the efficacy of the three proposed modules. Ablation studies reveal that the ATF module, the WFE module, and the GDAT module significantly improve the mAP and Rank-1 accuracy of the model, thereby highlighting their contributions to the vehicle re-identification task. In summary, the model introduced in this study not only substantially enhances the accuracy of vehicle re-identification but also emphasizes the potential applications of attribute information, wavelet processing, and attention mechanisms in fine-grained identification tasks. Future research will focus on further optimizing the feature fusion strategy, particularly in achieving efficient recognition of cross-view detailed features of vehicles within complex and dynamic environments, to meet the diverse requirements of transportation applications.

Acknowledgment

We would like to express our sincere gratitude to Haifeng Sang for his valuable assistance. This work has greatly benefited from his support.

References

[1] He S, Luo H, Wang P, Wang F, Li H, Jiang W. Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 15013–15022.

[2] Alexey D. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.

[3] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 10012–10022.

[4] Qian W, Luo H, Peng S, Wang F, Chen C, Li H. Unstructured feature decoupling for vehicle re-identification. In: European Conference on Computer Vision. Springer; 2022. p. 336–353.

[5] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks. Advances in neural information processing systems. 2015;28.

[6] Phan N, Huy TD, Duong ST, Hoang NT, Tran S, Hung DH, et al. Logovit: Local-global vision transformer for object re-identification. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2023. p. 1–5.

[7] Yu Z, Pei J, Zhu M, Zhang J, Li J. Multi-attribute adaptive aggregation transformer for vehicle re-identification. Information Processing & Management. 2022;59(2):102868.

[8] Simonyan K. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014.

[9] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

[11] Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems. 2017.

[12] Li Z, Zhang X, Tian C, Gao X, Gong Y, Wu J, et al. Tvg-reid: Transformer-based vehicle-graph re-identification. IEEE Transactions on Intelligent Vehicles. 2023.

[13] Lian J, Wang D, Zhu S, Wu Y, Li C. Transformer-based attention network for vehicle re-identification. Electronics. 2022;11(7):1016.

[14] Li H, Xiao Y, Cheng C, Song X. SFPFusion: An improved vision transformer combining super feature attention and wavelet-guided pooling for infrared and visible images fusion. Sensors. 2023;23(18):7870.

[15] Wang C, Wu J, Fang A, Zhu Z, Wang P, Chen H. An efficient frequency domain fusion network of infrared and visible images. Engineering Applications of Artificial Intelligence. 2024;133:108013.

[16] Yao T, PanY, Li Y, Ngo CW, Mei T. Wave-vit: Unifying wavelet and transformers for visual representation learning. In: European Conference on Computer Vision. Springer; 2022. p. 328–345.

[17] Jiang Y, Liu Q, Liu MT. Attribute Feature Fusion Network for Pedestrian Detection and Re-Identification. In: 2023 5th International Conference on Robotics and Computer Vision (ICRCV). IEEE; 2023. p. 36–40.

[18] Khorramshahi P, Kumar A, Peri N, Rambhatla SS, Chen JC, Chellappa R. A dual-path model with adaptive attention for vehicle re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 6132–6141.

[19] He B, Li J, Zhao Y, Tian Y. Part-regularized near-duplicate vehicle re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 3997–4005.

[20] Lou Y, Bai Y, Liu J, Wang S, Duan L. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 3235–3243.

[21] Chu R, Sun Y, Li Y, Liu Z, Zhang C, Wei Y. Vehicle re-identification with viewpoint-aware metric learning. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 8282–8291.

[22] Qian J, Jiang W, Luo H, Yu H. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. Measurement Science and Technology. 2020;31(9):095401.

[23] Jin X, Lan C, Zeng W, Chen Z. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34; 2020. p. 11165–11172.

[24] Teng S, Zhang S, Huang Q, Sebe N. Multi-view spatial attention embedding for vehicle re-identification. IEEE Transactions on Circuits and Systems for Video Technology. 2020;31(2):816–827.

[25] Zhang F, Ma Y, Yuan G, Zhang H, Ren J. Multiview image generation for vehicle re-identification. Applied Intelligence. 2021;51(8):5665–5682.

[26] Taufique AMN, Savakis A. LABNet: Local graph aggregation network with class balanced loss for vehicle re-identification. Neurocomputing. 2021;463:122–132.

[27] Wang H, Peng J, Jiang G, Xu F, Fu X. Discriminative feature and dictionary learning with part-aware model for vehicle re-identification. Neurocomputing. 2021;438:55–62.

[28] Zhu W, Wang Z, Wang X, Hu R, Liu H, Liu C, et al. A dat self-attention mechanism for vehicle re-identification. Pattern Recognition. 2023;137:109258.

[29] Qian J, Pan M, Tong W, Law R, Wu EQ. URRNet: A Unified Relational Reasoning Network for Vehicle Re-Identification. IEEE Transactions on Vehicular Technology. 2023;72(9):11156–11168.

[30] Huang F, Lv X, Zhang L. Coarse-to-fine sparse self-attention for vehicle re-identification. Knowledge-Based Systems. 2023;270:110526.

[31] Xu Z, Wei L, Lang C, Feng S, Wang T, Bors AG, et al. SSR-Net: A Spatial Structural Relation Network for Vehicle Re-identification. ACM Transactions on Multimedia Computing, Communications and Applications. 2023;19(6):1–22.