Semantic Feature-oriented Multi-Scale Dual Attention Network for Truck Re-Identification

Zhen Zeng and Jiangtao Ren*
School of Computer Science and Engineering
Sun Yat-sen University
Guangzhou, China
zengzh76@alumni.sysu.edu.cn, issrjt@mail.sysu.edu.cn

***Abstract*—Truck re-identification (Re-ID), a crucial subtask of vehicle Re-ID in intelligent transportation systems, faces unique challenges like significant front-carriage appearance differences, diverse carriage types, and cargo variations. Existing methods often overlook these structural differences and struggle to suppress variable visual features. To address these issues, we propose a Semantic Feature-oriented Multi-Scale Dual Attention (SF-MSDA) network, which includes: 1) a Segmentation module that enables truck front and carriage feature extraction at two granularity levels; 2) a Multi-Scale Dual Attention (MSDA) module that reconstructs front features by integrating spatial and channel dependencies, maximizing spatial distribution differences of each pair of features for fine-grained representation; and 3) a Part-Attention module that combines part features and global context to adaptively learn invariant features while suppressing cargo variations. Experiments on the Truck-ID dataset demonstrate that SF-MSDA outperforms both general vehicle and specialized truck Re-ID methods.**

***Index Terms*—Attention Mechanism, Feature Fusion, Semantic Segmentation, Truck Re-Identification, Multi-Scale Network.**

I. Introduction

Vehicle re-identification (Re-ID) refers to retrieving vehicles identical to a given query vehicle from a large gallery captured by different surveillance cameras. Recently, vehicle Re-ID has attracted increasing attention due to its importance in intelligent transportation applications. Due to the interference of factors such as illumination, viewpoint, occlusion, the intra-class differences of the same vehicle are large, while the inter-class similarities of different vehicles are strong. However, traditional vehicle Re-ID methods struggle to address these challenges.

In recent years, Deep Convolutional Neural Networks (DCNNs) have made significant breakthroughs in various computer vision tasks, and now dominate the Re-ID task. Current vehicle Re-ID methods include global-based, part-based and transformer-based methods. Global-based learning methods [1], [2] use typical feature extractors for end-to-end learning, such as VGG [3], ResNet [4], GoogleNet [5], etc. Compared with manual feature extractors, they more automatically learn discriminative and generalizable feature representations, but lack spatial perception of different local parts. Part-based methods [6], [7] enhance model robustness by extracting local features from specific parts of the vehicle, effectively addressing issues like viewpoint changes and occlusion. However, most of these methods directly cascade global feature and local features without fully integrating the global and local information. Transformer-based methods [8], [9] capture key information while reducing noise by dynamically adjusting attention to different regions of the image. DRA-Net [10] proposed a Dual Relational Attention Module (DRAM) to capture global feature dependencies in both spatial and channel dimensions, while SOFCT [11] used a patch feature-weighted transformer for the region view after semantic segmentation to enhance or attenuate each patch of different semantic features, thereby processing local features more finely. In addition, other methods are also helpful for Re-ID task. For example, multi-scale feature extraction [12], [13], capturing both local details and global context and learning positional and semantic features at different levels, has also improved model generalization.

Despite the effective progress in vehicle Re-ID, it still faces many challenges in specific scenarios, especially in the logistics field. Truck Re-ID, a special subtask of vehicle Re-ID, has its own unique importance and challenges. Compared with small cars, trucks have more complex and diverse appearances and structures, which results in limited effectiveness when directly applying vehicle Re-ID methods to truck. In addition, trucks have different types of fronts, carriages, and axles, as well as differences in cargo loading and axle configurations (Fig.1 shows the diversity of truck characteristics). These differences increase the difficulty of the truck Re-ID task. Although a few studies have attempted to address these issues, they still face certain limitations. For example, DGN [14] considered the structural characteristics of trucks and used a multi-branch network to jointly construct contextual information of global and local parts. However, it only divided the abstract global feature equally into front feature and body feature, which is a rough treatment of the truck parts. Moreover, it attaches equal importance to these two parts, failing to fully explore the unique appearance characteristics of trucks. Therefore, solving the truck Re-ID problem requires in-depth analysis of its particularity and proposing corresponding solutions.

Based on the diversity analysis of trucks in Fig.1, the special challenges of truck Re-ID include different front and carriage types, the number of axles, the axle configurations, fake tires, and loaded cargo, etc. Combining the two common challenges of vehicle Re-ID , namely, large intra-class differences and

Fig. 1: Diversity in Truck Characteristics. (a) Diversity of fronts: different colors, models and manufacturers. (b) Diversity of carriages: flatbed, palletized, tipper, barn-rail, box, tank, respectively. (c) Diversity of axles: different axle numbers and axle configurations.

small inter-class differences, we summarize the challenges of truck Re-ID as follows: 1) The front and carriage features of the same truck are inconsistent and the carriage appearance is unstable, 2) The carriage proportions vary significantly under different shooting angles, and 3) Different trucks of the same model and color have little difference. Combining the existing Re-ID ideas and specific structure of trucks, we further analyze the solutions to the three challenges of truck Re-ID. For the first challenge, a segmentation network can be introduced to divide the truck into front and carriage for differential treatment. For the second challenge, a part-aware attention mechanism is introduced in the feature fusion stage to assign different weights according to the local area ratio to alleviate the influence of viewpoint changes. For the third challenge, we focus on extracting fine-grained front feature and utilize a multi-scale branch network to capture multiple front features with different semantics for fine-grained recognition of the front.

To address above challenges, We propose a novel framework called Semantic Feature-oriented Multi-Scale Dual Attention (SF-MSDA), consisting of three key components: Semantic Segmentation, Multi-Scale Dual-Attention (MSDA) module, and Part-Attention module. The Semantic Segmentation network divides the truck into front and carriage,facilitating fine-grained front recognition and coarse-grained carriage recognition. The MSDA module utilizes multiple scale networks with varying depths and numbers of channels to obtain multi-level and multi-semantic features. By Maximizing Spatial Distribution Difference (MSDD) of each pair of multi-scale features, each scale network can learn complementary front features in different regions. Meanwhile, the Dual Attention mechanism is introduced to emphasize the important spatial and channel information of each front feature, and fuse spatial and channel features of multi-scale network to obtain a richer and more accurate front feature. The Part Attention module attempts to simulate the behavior of human observing and understanding an object, that is, from observing the whole to paying attention to the parts. And give different attention to these parts according to their location, function and size. Eventually, global and local features are fused to complete the truck Re-ID task.

In a nutshell, the mainly contributions of this article can be summarized as follows:

- We propose a novel semantic feature-oriented multi-scale attention truck re-identification network, which solves the problem of inconsistent features between the front and carriage of the truck by extracting features of the front and carriage parts at multi-level granularity.
- A multi-scale attention module is designed to mine more subtle discriminative features of the truck front. At the same time, a mechanism to maximize the spatial distribution difference is introduced to enable each branch to learn complementary information from different front regions.
- A Part Attention module is introduced to adaptively learn the importance of both the front and the carriage.
- Experimental validation is performed on the Truck-ID dataset and compared with SOTAs. The experimental results show that our model has significant performance advantages.

The rest of this paper is organized as follows: Section II introduces related work on vehicle Re-ID and truck Re-ID. Section III presents the proposed framework and details each module. Section IV provides ablation studies and baseline comparisons in a real dataset. Finally, Section V concludes the paper.

II. Related Work

In this section, we review recent related works on four methods for Re-ID.

A. Transformer-based Re-ID

While CNN-based Re-ID methods have achieved significant success, they process only one local neighborhood at a time, often leading to information loss due to convolution and downsampling operations. In contrast, transformers offer advantages for vehicle Re-ID by enabling the model to focus on identity-relevant information while ignoring distractions like background. He et al. [8] proposed a pure transformer-based Re-ID framework using multi-head attention to capture long-range dependencies and enhanced feature robustness via patch-based operations. Zheng et al. [10] introduced a dual relational attention module (DRAM) and added a non-similarity constraint, allowing different branches to focus on distinct image regions. Zhu et al. [9] highlighted DCNN's issues, such as poor feature alignment of vehicles from different views and lack of spatial perception of positioning, proposing a dual self-attention module to learn different regional dependencies.

B. Region/part learning-based Re-ID

In addition to global features, recent work utilizes local features to enhance feature characterization capabilities. EIA-Net [15] used Regional Proposal Networks (RPNs) to locate key vehicle parts, combining local and global features. Meng

et al. [6] developed a parsing-based view-aware embedding network (PVEN) for view-aware feature alignment, with co-visible attention designed to emphasize common visible views and suppress irrelevant views between vehicles. Guo et [16] designed the hard attention mechanism that Spatial Transformation Network (STN) to localize salient windshield and car-head parts, while the soft attention mechanism provides additional attentional refinement to focus on the unique features of each part. PGAN [17] proposed a component attention module to adaptively localize the most discriminative regions with highly attentive weights and to suppress the interference of irrelevant parts with lower weights.

C. Multi-branch network-based Re-ID

Many Re-ID methods focus on designing network architectures to extract more effective features, and a series of multi-branch networks have emerged, including multi-scale extraction and multi-resolution search, which provide better generalization performance for Re-ID. MSDeep [12] fused multiple levels of features at different scales to construct the final representation and introduced multi-scale attention to dynamically emphasize each scale's important channels for stronger global feature. LCSR [2] exploited the correlation and complementarity of deep-level feature generated by multiple baseline networks by Laplace regularized correlation sparse ranking, and utilized a re-ranking technique to further improve the performance of Re-ID. Most multi-branch network schemes that share low-level feature, the higher levels of the network ignore the interactions between branches. Therefore, Tang et al. [18] proposed a novel Harmonious Multi-branch Network(HMBN) to alleviate such problems. Combined with an attention module, it enables all branches to interact by modeling spatial contextual dependencies across branches.

D. Metric learning-based Re-ID

ReID based on metric learning obtains a more divisible feature space by learning the degree of similarity of feature embeddings of vehicle images. Schroff et al. [19] proposed a classic triple loss to shorten the distance in the feature space of positive samples and widen the distance in the feature space of negative samples. Most Re-ID work takes the reduction of cross-entropy loss and triplet loss as training goals. However, Luo et [1] proved experimentally that the objectives of these two losses are inconsistent in the embedding space. Therefore, they proposed the BNNeck module to solve the problem of loss oscillation and disappearance when the two are optimized at the same time, and greatly improve the performance of Re-ID. Wang et al. proposed a Mahalanobis-based Triplet loss with a Transformation Relation Matrix to encode the changing factors into the learning process and enhance the model robustness. Yan et al. [20] divided the vehicle dataset into multi-grains based on image relationships, and proposed two multi-grain ranking constraints, Generalized Pairwise Ranking and Multi-Grain based List Ranking, to alleviate the precise vehicle search problem.

III. Methodology

For trucks with complex appearances and structures, we propose a Semantic Feature-oriented Multi-scale Dual Attention (SF-MSDA) framework, illustrated in Fig.2. This method improves the performance of the truck Re-ID model by expanding the high-response regions for front feature, adaptively identifying invariant carriage feature, and mitigating the impact of variable cargo.

A Multi-Scale Dual Attention (MSDA) module is utilized to extract rich front information, in which a multi-scale network is utilized to capture features with multi-levels and multi-semantics, a maximization of the spatial distribution difference (MSDD) of each pair of these features is introduced to enable each scale network to learn complementary feature from different regions, and a dual attention mechanism is adopted to enable each branch to have the ability to distinguish different trucks. In the carriage feature extractor, the carriage mask is directly mapped to global feature, forming the final carriage representation. Since the carriage occupies most of the truck's area and has stable characteristics in color, shape, and texture, high-level feature can adequately distinguish invariant carriage structures. For variable information like cargo and tarpaulin, suppression is preferred over amplification. Finally, the Part Attention module integrates position, function, and size information of both components in the image, adaptively fusing global, front, and carriage features to generate final Re-ID feature for truck similarity calculation.

In this section, we detail the proposed model from four key perspectives.

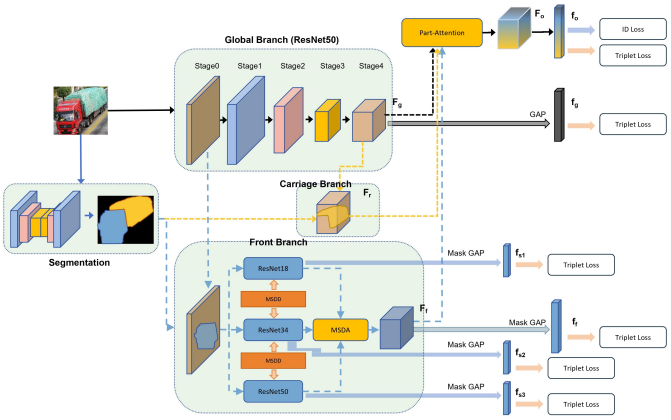

Fig. 2: SF-MSDA pipeline. The front branch extracts fine-grained front features by MSDA module and MSDD constraint, while the carriage branch extracts coarse-grained carriage feature. The Part Attention module adaptively weights the global, front, carriage features to output the final Re-ID feature.

A. Multi-Scale Dual-Attention Module

Among the common strategies for fine-grained feature extraction, multi-scale feature extraction captures complementary feature across scales, reducing the risk of overfitting

but potentially introducing feature redundancy or attention to irrelevant areas. In contrast, methods based on attention mechanisms can dynamically focus on critical regions, enhancing the interpretability of the model. Therefore, we propose the Multi-Scale Dual-Attention (MSDA) mechanism, as shown in Fig.3, combining the strengths of both, which dynamically focuses on key regions at each scale, minimizing redundancy and improving feature effectiveness.

This module mainly consists of two parts: single-scale dual attention network and multi-scale feature fusion. Before MSDA module, multiple front features are extracted by the multi-scale feature extractor which consists of several different resnet networks with different depths and channels. Shallow networks mainly focus on local detail features, while deep networks can extract more abstract high-level semantic information. Therefore, networks of multiple scales can obtain rich feature information from different perspectives. In the single-scale dual attention network, we introduce dual attention mechanisms on space and channels for each individual scale network. The spatial attention mechanism can focus on key areas according to the importance of spatial positions; the channel attention mechanism adjusts the weights of each channel according to the importance of different channels. Finally, the features of multiple scale networks are reconstructed and fused to integrate the different levels of information captured by each scale network. By fusing these spatial features and position features, a richer and more accurate representation of the truck front is finally obtained.

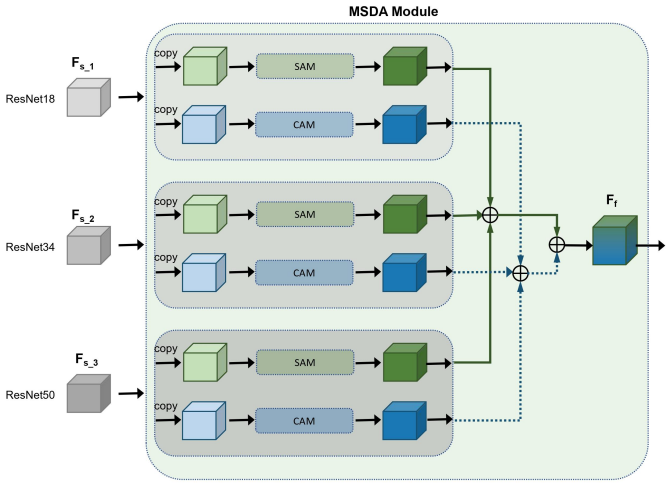

Fig. 3: Multi-Scale Dual-Attention Module.

Specifically, let F_s denote the feature map from scale s (where $s = 1, 2, 3$). Each feature map is processed through spatial attention module (SAM) and channel attention module (CAM) as proposed in DANet [21] to capture relevant dependencies. The overall feature representation is expressed as:

$$F_f = \text{Concat}\left(\sum_{i=1}^{3} \mathcal{S}(F_{s\_i}), \sum_{i=1}^{3} \mathcal{C}(F_{s\_i})\right) \tag{1}$$

Where $\mathcal{S}(\cdot)$ represents SAM, while $\mathcal{C}(\cdot)$ represents CAM. This process combines spatial and channel features across multiple scales, yielding a robust and discriminative representation for fine-grained recognition task.

B. Maximizing Spatial Distribution Difference

In the front branch, we introduce the Maximizing Spatial Distribution Difference (MSDD) mechanism , alongside triplet loss optimization for each scale branch's feature space. The MSDD encourages high-response regions of front features across different scales to exhibit maximum spatial inconsistency, thus broadening the front discrimination area.

Specifically, let $F_s \in \mathbb{R}^{C_s \times H_s \times W_s}$ denote the output feature map of scale s. A 1×1 convolution reduces the channel dimension, yielding the transformed feature map F'_s, and define the spatial probability distribution P_s for each branch as:

$$F'_s = \text{Conv}(F_s) \tag{2}$$

$$P_s(i,j) = \frac{\exp(F'_s(i,j))}{\sum_{i,j} \exp(F'_s(i,j))} \tag{3}$$

To quantify the spatial distribution difference between two scales s_1 and s_2, we employ Kullback-Leibler divergence:

$$\text{KL}(P_{s_1} \parallel P_{s_2}) = \sum_{i,j} P_{s_1}(i,j) \log \frac{P_{s_1}(i,j)}{P_{s_2}(i,j)} \tag{4}$$

The objective of the MSDD is to maximize divergence across all scale combinations:

$$\mathcal{L}_{\text{SD}} = -\sum_{s_1 \neq s_2} \text{KL}(P_{s_1} \parallel P_{s_2}) \tag{5}$$

Minimizing $\mathcal{L}_{\text{SD}}$ encourages spatial distribution inconsistency among the output feature, expanding the model's perceptual capacity for the front detection area.

C. Part Attention Module

Inspired by the human recognition process that shifts attention from global appearance to discriminative parts, we propose a Part Attention module. It takes global feature and part features as input, and uses an attention mechanism to dynamically adjust the attention weights according to the importance of each part in the global context.

The part attention mechanism is similar to the image-based self-attention calculation method. We redefine the query vector Q, key vector K, and value vector V. Specifically, we calculate the query vector Q_g for the global feature F_g, and calculate the corresponding key vector K_p and value vector V_p for each part feature F_p (where $F_p = \{F_f, F_r\}$, representing the features of the front and carriage branches respectively). The calculation formula is as follows:

$$Q_g = W_Q F_g, \quad K_p = W_K F_p, \quad V_p = W_V F_p \tag{6}$$

where W_Q, W_K and W_V are learned projection matrices for the query, key and value transformations respectively.

The attention weight between the global feature and each part feature is computed by using the query from the global

feature and the key from each part feature. Additionally, the part area ratio α_p, which indicates the proportion of each part relative to the image, is incorporated as a second weight to modulate the attention:

$$A_p = \alpha_p \cdot \frac{\exp\left(\frac{Q_g K_p}{\sqrt{d_k}}\right)}{\sum_{j=1}^{n} \exp\left(\frac{Q_g K_p}{\sqrt{d_k}}\right)} \tag{7}$$

where d_k is the dimension of the key vector. This attention weight A_p reflects the importance of each part feature relative to the global context.

The final part feature F_p' is reconstructed using the weighted attention for adjusts the part feature based on its relevance in the global context:

$$F_p' = A_p V_p \tag{8}$$

where the reconstructed front and carriage features are represented as F_f' and F_r' respectively.

Finally, the reconstructed part features are fused with the global feature to form the final enhanced feature representation F_o:

$$F_o = \text{Concat}(F_g, F_f', F_r') \tag{9}$$

D. Loss Function

During the training phase, we use cross-entropy loss, triplet loss, and maximizing spatial distribution difference for joint optimization. In the multi-scale branch, each branch is optimized with triplet loss $\mathcal{L}_{tri}^{s\_i}$ to ensure its capacity to distinguish truck front. The MSDD constraint $\mathcal{L}_{\text{SD}}$ is applied to each branch pair, encouraging distinct spatial focus and complementary feature learning. In summary, the total loss function can be written as:

$$\mathcal{L} = \mathcal{L}_{id}^{o} + \mathcal{L}_{tri}^{o} + \lambda_1 \mathcal{L}_{tri}^{g} + \lambda_2 \mathcal{L}_{tri}^{f} + \lambda_3 \sum_{i}^{3} \mathcal{L}_{tri}^{s\_i} + \lambda_4 \mathcal{L}_{\text{SD}} \tag{10}$$

where $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ is the loss weight to weigh the effect of the global triplet loss, the front triplet loss, and maximizing the spatial distribution difference loss function.

IV. Experiments

A. Dataset and Evaluation Metrics

Dataset. We select the Truck-ID dataset, which contains 32,353 truck images from seven traffic surveillance sites in Anhui Province of China, including 13,137 license plate IDs. The gallery of Truck-ID dataset is further divided into three sub-datasets based on the difficulty of truck Re-ID in order to evaluate the quality of different truck Re-ID models more comprehensively. The Truck-ID dataset not only contains the common challenges of vehicle Re-ID, but also covers the specific fine-grained challenges of truck Re-ID, such as the number of axles on a truck, the type of axle set locations, dummy tires, and the loaded cargo. Fig.1 reflect the diversity and complexity of the dataset.

Evaluation Metrics. We use two standard Re-ID metrics: Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP).

B. Implementation Details

We use ResNet-50 as the backbone to extract global feature, adjusting the last layer's stride to 1 for higher feature granularity, as suggested by [1]. The semantic segmentation extractor is implemented with YOLOv8, trained on a dataset containing 917 labels. A warm-up strategy is applied to guide training. The learning rate starts at $3.5e^{-4}$ for the first 10 epochs and is reduced to $3.5e^{-5}$ and $3.5e^{-6}$ at the 30th and 50th epochs, respectively, until training completes.

C. Experimental Evaluation

1) Quantitative analysis: We compare SF-MSDA with recent state-of-the-art models using mAP and CMC. These methods can be broadly categorized into three groups: (1) global-feature-based approaches, such as RK [1], AGW [22], and FPB [23]; (2) part-based methods using fixed spatial partitioning, including MGN [24] and DGN [14]; and (3) methods leveraging detectors or segmenters to obtain local regions, such as PVEN [6], PGAN [17], and TransReID [8].

Tables I-II compare SF-MSDA with these SOTA models on the Truck-ID dataset. As shown in Table I, our method outperforms the SOTAs on the Gallery-Easy dataset. Compared to PVEN, which also uses a segmentation network, our method improves mAP by 1.9%, demonstrating that segmenting into two parts is more effective than using four views for Truck Re-ID. Additionally, our method surpasses DGN, another truck Re-ID model, with a 0.8% improvement in mAP due to multi-scale feature fusion and part-based attention.

TABLE I: Results on Gallery-Easy Dataset

Method	mAP	R1	R5
RK	83.7	84.9	97.3
AGW	86.8	87.3	97.3
FPB	89.2	88.8	97.9
MGN	87.9	85.3	96.2
DGN	90.2	90.2	98.2
PVEN	89.1	89.0	97.8
PGAN	90.3	90.3	98.7
TransReID	89.7	89.3	98.1
FS-MSDA(Ours)	**91.0**	**90.9**	**99.1**

TABLE II: Results On Gallery-Medium And Gallery-Hard Dataset

Method	Gallery-Medium			Gallery-Hard		
	mAP	R1	R5	mAP	R1	R5
RK	74.4	72.5	91.8	70.1	59.0	84.6
AGW	81.4	78.6	95.2	74.7	64.0	89.5
FPB	84.9	80.8	96.3	77.1	67.5	90.2
MGN	82.9	75.2	94.4	72.9	61.4	89.1
DGN	85.7	82.8	97.2	80.1	71.8	92.1
PVEN	84.7	81.5	96.2	77.0	69.0	89.2
PGAN	85.9	82.9	97.3	80.1	71.9	92.1
TransReID	85.0	82.3	97.1	78.5	69.8	90.7
FS-MSDA(Ours)	**86.4**	**83.1**	**97.5**	**80.3**	**72.1**	**92.0**

2) Qualitative analysis: Fig.4 displays the Top-10 results on Gallery-Easy dataset. PVEN tends to identify trucks from similar viewpoints, while TransReID, with the added edge information, ranks the same truck from different viewpoints higher. However, both of them are difficult to consider two images of the same vehicle, loaded and unloaded cargo, as the same ID, but our model can do it with a higher recognition rate. This is due to the fact that we add MSDA module to refine the front features and Part Attention module to adaptively fuse the front, carriage and global features.

Immediately afterwards, we randomly select some truck images and display the corresponding activation maps, As shown in Fig.5. The first column of Fig.5 represents original images, the second column visualizes the output feature of stage 4 of the backbone network, the third column visualizes the front feature, the fourth column visualizes the carriage feature, and the last column visualizes the final fusion feature. It can be seen that our model learns the relationship between global and local features, adaptively adjusts the weights of the front, carriage and global features, and finally extracts more discriminative feature of the truck Re-ID.

D. Ablation Experiments

We conduct ablation experiments to evaluate the contributions of different components to Re-ID performance.

1) Effectiveness of components: To validate the effectiveness of each component, we perform the following comparisons:

- Global Only (GO): Model includes only the global branch (baseline).
- Global + Local (GL): Model includes global and two local branches.
- Multiscale (MS): Adds a Multi-Scale Feature Extractor (without Dual-Attention) and MSDD in the front branch.
- Multi-Scale Dual-Attention (MSDA): Incorporates all proposed strategies for the front branch.

The results in Table III indicate the following conclusions: First, this locally enhanced model (GL) distinguishes trucks with subtle appearance differences, showing that the feature from the two stages are well complementary. Second, adding a multi-scale feature extractor in the front branch and constraining each extractor to focus on different areas of the front can achieve better performance. Lastly, MSDA uses channel and spatial attention to highlight key areas across scales, merging them to improve attention to front features, yielding a 0.6% improvement in mAP, reaching 91.6%.

TABLE III: Results On Different Components

Method	mAP	R1	R5
GO	87.7	87.9	97.6
GL	89.1	89.3	97.9
MS	90.4	90.3	98.7
MSDA	**91.0**	**90.9**	**99.1**

2) The influence of multi-scale branch numbers: In this paper, we propose the MSDA module to extract features from multiple front regions, constrained by MSDD and triplet loss. More branches theoretically capture richer features. Based on this assumption, we test with 2-4 scale branches on the front branch. The results of experiments are reported in Table IV, where there is a small performance improvement with 4 branches, but is also a concomitant sharp increase in the number of parameters. To balance the efficiency and performance, we finally choose three scale branches.

TABLE IV: Results On Different Number Of Multi-Scale Branches

Method	mAP	R1	R5	parameters
2	90.6	90.4	98.6	142.18M
3	91.0	90.9	99.1	143.49M
4	91.1	91.0	99.1	151.13M

3) The influence of feature fusion: In this subsection, we design four fusion schemes:.

A. Channel concatenation: Three features are concatenated in the channel dimension directly.
B. Numerical addition and subtraction: Three features are fused through numerical addition.
C. Self-Attention: Reconstruct each part's feature using self-attention mechanism, then assign weights and sum them.
D. Part Attention: Using the Part Attention module as the proposed framework.

Experimental results in Table V reveal that scheme B performs worst, indicating that simply enhancing or weakening the part features as a whole is not feasible, and the F_g learned from the baseline is likely to contain both discriminative and suppressive global feature. Scheme C outperforms A, implying that appropriately focusing on the critical regions of the parts and reconstructing the feature can attenuate the suppressive feature and enhance the influence of the discriminative feature, thus effectively reducing semantic gaps in the fusion stage. Scheme D achieves the best results, demonstrating that fine-tuning local feature by the association of global and local features can effectively highlights the discriminative regions of local feature.

TABLE V: Results On Different Feature Fusion Schemes

Method	mAP	R1	R5
A	90.4	90.3	98.3
B	87.4	86.6	97.3
C	90.8	90.8	98.9
D	**91.0**	**90.9**	**99.1**

4) The influence of loss function: We explore the effect of different loss functions, as shown in Table VI, L_{GFI} refers to $L_{id}^{g} + L_{id}^{f}$, L_{GFT} refers to $L_{tri}^{g} + L_{tri}^{f}$, L_{MS} refers to $L_{SD} + L_{tri}^{s\_i}$. Performance is lowest when only cross-entropy and triplet losses are applied to the fused feature f_o. Adding cross-entropy loss for global feature f_g and front feature f_f

Fig. 4: Top-10 query results from different methods. Green IDs match the probe ID, while red IDs differ.

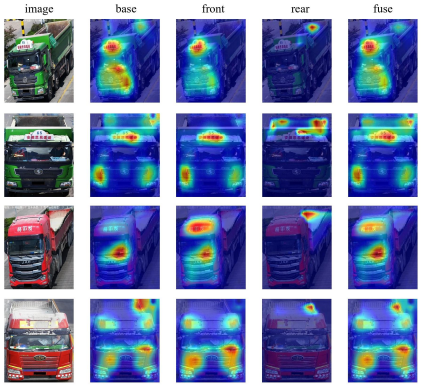

Fig. 5: Feature visualization in different images and different viewpoints.

TABLE VI: Results On Loss Function

L_{id}^{o}	L_{tri}^{o}	L_{GFI}	L_{GFT}	L_{MS}	mAP	R1	R5
Y	Y	N	N	N	83.7	84.2	96.8
Y	Y	N	Y	N	90.8	90.6	98.7
Y	Y	Y	Y	N	90.1	90.1	98.6
Y	**Y**	**N**	**Y**	**Y**	**91.0**	**90.9**	**99.1**

improves performance slightly, but combining both the cross-entropy loss and the triplet for f_g and f_f leads to a decrease. which suggests that f_g and f_f have difficulty in recognizing truck identities, but help to increase inter-class distance and reduce intra-class variance. When constraining the maximum spatial distribution difference L_{SD} for each pair of multi-scale front features and the triple loss $L_{tri}^{s\_i}$ for each multi-scale front features, the discriminative region of the front can be enlarged to capture more valid information about the front and obtain the higher performance.

To further demonstrate the effectiveness of L_{SD} + $L_{tri}^{s\_i}$, we visualize the activation maps of the three front multi-scale features in Fig.6. Without L_{SD} + $L_{tri}^{s\_i}$ constraints, the multi-scale feature extractor captures overlapping regions. In contrast, with these constraints, interest regions become distinct for each scale branch. For example, in the green truck on the right, the first scale branch focuses on the manufacturer's text on the top of the front, the second branch detects the windshield, the

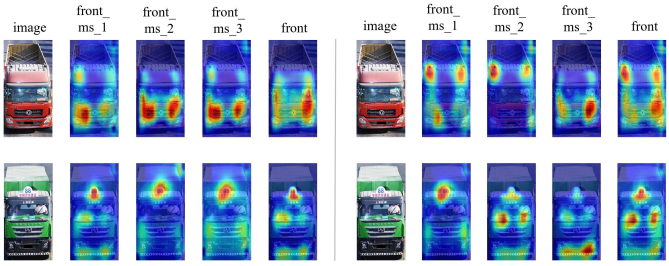

Fig. 6: Visualizations with and without L_{SD} + $L_{tri}^{s\_i}$ in SF-MSDA. The left figure does not apply L_{SD} + $L_{tri}^{s\_i}$ to multi-scale branch, while the right figure does the opposite. The first column is the original image, the second to fourth columns are the three multi-scale front features extracted from the multi-scale branch, and the fifth column is the front feature after fusion by the MSDA module.

third branch locates the bumper and the fused heatmap covers a wider area, indicating that the L_{SD} + $L_{tri}^{s\_i}$ constraints help learn important discriminative features in separate regions.

Additionally, we visualize intra-class and inter-class distances of truck features using t-SNE in Fig.7. Some of these selected trucks have very small differences in manufacturer and color, while others have significant differences in carriage type and whether or not load cargo. By comparing the figures, it can be observed that our proposed method can better distinguish the identities that are hard to distinguish in the baseline.

5) The influence of hypermeters: In equation 10, the learnable parameter $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ is used to adjust the weights of the loss function. We evaluate its influence on Re-ID through experiments on Gallery-Easy with varying parameter sets, as shown in Table VII.

When the weight of f_g matches or exceeds f_f, the model performs worse. This may be due to the global feature being more sensitive to viewpoint changes, leading to larger differences between front and side views. Conversely, the refined front feature contains more discriminative information, but variability in the carriage can cause global feature discrepancies. When λ is set to $\{0.1, 0.5, 1.0, 1.0\}$, our model achieves the best balance between global and local features, resulting

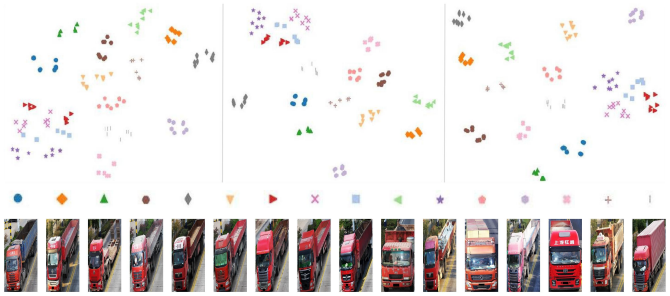

Fig. 7: T-SNE visualization on some trucks. The left figure shows the results of baseline, the middle figure shows the results of SF-MSDA without L_{SD} + $L_{tri}^{s\_i}$ constraints, and the right figure shows the results of SF-MSDA with L_{SD} + $L_{tri}^{s\_i}$ constraints.

TABLE VII: Results On Different Hyperparameters In Gallery-Easy

λ_1	λ_2	λ_3	λ_4	mAP	R1	R5
1.0	1.0	1.0	1.0	89.7	89.2	98.2
0.5	0.5	1.0	1.0	89.5	88.0	98.0
0.1	**0.5**	**1.0**	**1.0**	**91.0**	**90.9**	**99.1**
0.1	0.5	0.5	1.0	90.4	90.2	98.8
0.5	0.1	1.0	0.5	85.4	85.6	97.2

in the highest performance.

V. Conclusion

In this paper, we propose the SF-MSDA framework to address three challenges in truck Re-ID: unstable carriage appearance in the same truck, significant carriage differences across viewpoints, and subtle differences among trucks from the same manufacturer and color. The MSDA module integrates complementary information at different levels to capture rich front feature, while the part attention module adaptively learns the significance of the front and carriage in recognition. To avoid feature imbalance due to viewpoint changes, we incorporate the part area ratio into the part attention module. Finally, global and local features are combined to enhance truck Re-ID performance. Extensive experiments and ablation studies demonstrate the effectiveness of SF-MSDA.

References

[1] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[2] D. Sun, L. Liu, A. Zheng, B. Jiang, and B. Luo, "Visual cognition inspired vehicle re-identification via correlative sparse ranking with multi-view deep features," in *Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings 9*. Springer, 2018, pp. 54–63.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[6] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z.-J. Zha, X. Gao, S. Wang, and Q. Huang, "Parsing-based view-aware embedding network for vehicle re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7103–7112.

[7] X. Chen, H. Yu, F. Zhao, Y. Hu, and Z. Li, "Global–local discriminative representation learning network for viewpoint-aware vehicle re-identification in intelligent transportation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[8] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 013–15 022.

[9] W. Zhu, Z. Wang, X. Wang, R. Hu, H. Liu, C. Liu, C. Wang, and D. Li, "A dual self-attention mechanism for vehicle re-identification," *Pattern Recognition*, vol. 137, p. 109258, 2023.

[10] Y. Zheng, X. Pang, G. Jiang, X. Tian, and Q. Meng, "Dual-relational attention network for vehicle re-identification," *Applied Intelligence*, vol. 53, no. 7, pp. 7776–7787, 2023.

[11] Z. Yu, Z. Huang, J. Pei, L. Tahsin, and D. Sun, "Semantic-oriented feature coupling transformer for vehicle re-identification in intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 3, pp. 2803–2813, 2023.

[12] Y. Cheng, C. Zhang, K. Gu, L. Qi, Z. Gan, and W. Zhang, "Multi-scale deep feature fusion for vehicle re-identification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1928–1932.

[13] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, and J. Zhao, "Msinet: Twins contrastive search of multi-scale interaction for object reid," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 243–19 253.

[14] S.-B. Chen, Z.-H. Lin, C. H. Ding, and B. Luo, "Fine-grained truck re-identification: A challenge," *Cognitive Computation*, vol. 15, no. 6, pp. 1947–1960, 2023.

[15] L. Liang, C. Lang, Z. Li, J. Zhao, T. Wang, and S. Feng, "Seeing crucial parts: Vehicle model verification via a discriminative representation model," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1s, pp. 1–22, 2022.

[16] H. Guo, K. Zhu, M. Tang, and J. Wang, "Two-level attention network with multi-grain ranking loss for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4328–4338, 2019.

[17] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen, "Part-guided attention learning for vehicle instance retrieval," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3048–3060, 2020.

[18] Z. Tang and J. Huang, "Harmonious multi-branch network for person re-identification with harder triplet loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 4, pp. 1–21, 2022.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[20] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 562–570.

[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[22] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[23] S. Zhang, Z. Yin, X. Wu, K. Wang, Q. Zhou, and B. Kang, "Fpb: feature pyramid branch for person re-identification," *arXiv preprint arXiv:2108.01901*, 2021.

[24] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.