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.

DGR-Net: Dual-Graph Relation-aware Network for Vehicle Re-identification

Yun Jiang
College of Artificial Intelligence and Computer Science,Northwest Normal University
Kunyi Zhu
202421162186@nwnu.edu.cn

College of Artificial Intelligence and Computer Science,Northwest Normal University
Keyuan Zhou
College of Artificial Intelligence and Computer Science,Northwest Normal University
Huimei Tian
College of Artificial Intelligence and Computer Science,Northwest Normal University
Xiaoge Li
College of Artificial Intelligence and Computer Science,Northwest Normal University
Jiaqi Zhang
College of Artificial Intelligence and Computer Science,Northwest Normal University
Yutong Yao
College of Artificial Intelligence and Computer Science,Northwest Normal University
Yufeng Chen
College of Artificial Intelligence and Computer Science,Northwest Normal University
Xubo Hu
College of Artificial Intelligence and Computer Science,Northwest Normal University

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11780971/v1

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

Additional Declarations: No competing interests reported.

DGR-Net: Dual-Graph Relation-aware Network for Vehicle Re-identification

Yun Jiang[1], Kunyi Zhu[2*], Keyuan Zhou[3], Huimei Tian[4], Xiaoge Li[5], Jiaqi Zhang[8], Yutong Yao[9], Yufeng Chen[6], Xubo Hu[7]

[1]Northwest Normal University, Lanzhou, 730070, Gansu, China.

*Corresponding author(s). E-mail(s): 202421162186@nwnu.edu.cn;
Contributing authors: jiangyun@nwnu.edu.cn;
202421162218@nwnu.edu.cn; 202421162100@nwnu.edu.cn;
202521162071@nwnu.edu.cn; 202521162228@nwnu.edu.cn;
202521162225@nwnu.edu.cn; 202521162176@nwnu.edu.cn;
202521162189@nwnu.edu.cn;

Abstract

Vehicle re-identification (ReID) is a challenging task due to severe visual discrepancies caused by cross-camera and cross-view variations. While Vision Transformers (ViTs) have shown great potential in capturing global dependencies, they inherently disrupt the 2D spatial structure of images and often overlook the macro-level physical topology of the camera network. Furthermore, traditional metric learning methods treat all sample pairs equally in a unified feature space, which is insufficient for handling the heterogeneous matching difficulties arising from different camera and viewpoint relations. To address these issues, we propose a novel Dual-Graph Relation-aware Network (DGR-Net). Specifically, we introduce a Patch Spatial Graph Module to explicitly restore and enhance the 2D spatial adjacency of local patches. To bridge the visual gap across cameras, we design a Camera Topology Graph Module that maintains a memory bank of camera prototypes and propagates contextual information across a learned topology graph. Moreover, we propose a Relation-Conditioned Metric Learning approach that dynamically adjusts the metric space based on specific camera and view relations, significantly improving the model's robustness against extreme viewpoint changes. Extensive experiments on two large-scale datasets, VeRi-776 and VehicleID, demonstrate that our proposed DGR-Net achieves state-of-the-art performance, outperforming existing methods by a large margin.

Keywords: Vehicle Re-identification, Vision Transformer, Dual-Graph Network, Camera Topology, Metric Learning

1 Introduction

Vehicle re-identification (ReID) aims to retrieve a specific target of interest across multiple non-overlapping surveillance cameras. It plays a crucial role in intelligent transportation systems, public security, and smart city construction. In recent years, with the rapid development of deep learning, ReID has achieved remarkable progress. Particularly, the introduction of Vision Transformers (ViTs) has brought a paradigm shift from traditional Convolutional Neural Networks (CNNs) to self-attention-based architectures. By treating images as sequences of patches, ViTs demonstrate superior capability in capturing global dependencies and long-range contextual information, which are highly beneficial for learning robust feature representations [1].

Despite these advancements, existing ViT-based ReID methods still face several critical challenges that hinder their performance in complex real-world scenarios. First, ViTs inherently process images as a 1D sequence of flattened patches. This serialization inevitably disrupts the original 2D spatial structure and local neighborhood relations of the image. In ReID tasks, fine-grained visual cues—such as vehicle logos, inspection marks, or specific clothing patterns—are often localized in small, contiguous regions. The loss of explicit 2D spatial adjacency makes it difficult for the network to effectively aggregate these local structural details, which are essential for distinguishing highly similar identities.

Second, in real-world surveillance networks, images captured by different cameras exhibit severe visual discrepancies due to variations in illumination, background, and viewpoint. While some recent works have attempted to model structured relationships [2] or utilize heterogeneous relational complements [3] to enhance feature robustness, they often treat cameras as isolated entities and ignore the underlying physical topology of the camera network. In a practical CCTV system, cameras possess inherent spatial and temporal correlations. Modeling this macro-level camera topology is vital for bridging the visual gap between different camera views, as it provides valuable prior knowledge to guide the feature matching process across the network [4].

Third, the difficulty of matching two samples varies significantly depending on their camera and viewpoint relations. For instance, matching two images from the same camera and the same viewpoint is generally a straightforward task. However, matching images from different cameras and drastically different viewpoints (e.g., front view vs. rear view) is extremely challenging due to the severe semantic misalignment. Traditional metric learning methods, such as the standard triplet loss, treat all sample pairs equally in a unified feature space. This "one-size-fits-all" approach is insufficient for handling such heterogeneous matching difficulties. Inspired by relation modeling in knowledge graphs [5], where different relations require different semantic spaces, we argue that the metric space in ReID should be dynamically adjusted based on specific camera and view relations.

To address the aforementioned challenges, we propose a novel Dual-Graph Relation-aware Network (DGR-Net) for robust vehicle and person re-identification. Our framework is built upon a ViT backbone and introduces three core components to comprehensively enhance feature representation and metric learning from both micro and macro perspectives.

To mitigate the loss of 2D spatial adjacency caused by the ViT architecture, we propose the Patch Spatial Graph Module (PSG). It reshapes the 1D patch tokens back into a 2D grid and employs Graph Convolutional Networks (GCN) to aggregate local spatial context. This operation explicitly restores and enhances fine-grained structural details, complementing the global representation.

To tackle the severe cross-camera visual discrepancies, we introduce the Camera Topology Graph Module (CTG). Instead of treating cameras independently, this module maintains a memory bank of camera prototypes and constructs a fully-connected topology graph to propagate contextual information across the entire camera network. By fusing the original features with these topology-aware prototypes, the model generates global representations that are highly robust to camera variations. Furthermore, we introduce a topology-aware classification loss to explicitly guide the optimization process using the learned graph.

Furthermore, to handle the heterogeneous matching difficulties caused by both camera and viewpoint variations, we design a Relation-Conditioned Metric Learning (RCML) approach. It categorizes sample pairs into different relations based on their camera and view labels (e.g., same-camera-same-view, cross-camera-cross-view). For each relation type, RCML dynamically assigns relation-specific projection matrices and metric weights. This ensures that the model learns discriminative features tailored to the specific challenges of each relation, effectively pulling positive pairs closer even under extreme viewpoint changes.

The main contributions of this paper are summarized as follows:

- We propose the Dual-Graph Relation-aware Network (DGR-Net), a novel ViT-based framework that comprehensively addresses the challenges of spatial structure loss, cross-camera variations, and heterogeneous matching difficulties in ReID tasks.
- We design the Patch Spatial Graph Module to explicitly restore micro-level 2D spatial adjacency, and the Camera Topology Graph Module to capture macro-level physical relations within the camera network. Additionally, a topology-aware classification loss is proposed to leverage the learned graph for optimization.
- We introduce a Relation-Conditioned Metric Learning approach that dynamically adjusts the metric space based on camera and view relations, significantly improving the robustness against extreme viewpoint and camera changes.
- Extensive experiments on two large-scale vehicle ReID datasets (VeRi-776 and VehicleID) demonstrate that our proposed DGR-Net achieves state-of-the-art performance, validating the effectiveness of our dual-graph and relation-aware designs.

2 Related Work

2.1 Deep Learning for Re-identification

Deep learning has significantly advanced vehicle and person re-identification over the past decade. Early methods primarily relied on Convolutional Neural Networks (CNNs) to extract discriminative visual features. To capture fine-grained local details, researchers have proposed various multi-branch and part-based networks. For instance,

Lian et al. [6] introduced a multi-branch enhanced discriminative network that divides feature maps into spatial sub-maps, utilizing horizontal and vertical branches to extract subtle distinguishing features intrinsically contained in sub-areas. While CNNs excel at extracting local patterns, their limited receptive fields often restrict their ability to capture global context.

Recently, Vision Transformers (ViTs) have shown great potential in ReID tasks due to their self-attention mechanism, which naturally models long-range dependencies. Methods like ViT-ReID [7] have adapted the Transformer architecture for vehicle ReID, introducing specific modules to handle camera perspective changes and utilizing fine-tuning strategies for large-scale models. Furthermore, addressing cross-view variations remains a critical challenge in real-world applications. Yin et al. [8] proposed CVNet, a lightweight network employing multi-scale localization and deep-shallow filtrate collaboration to tackle the severe scale and viewpoint variations in aerial-ground cross-view ReID. Despite these advancements, existing ViT-based methods often process images as 1D sequences, overlooking the explicit modeling of 2D spatial adjacency. Moreover, they typically treat images independently, ignoring the macro-level camera topology. Our proposed DGR-Net addresses these limitations by introducing the Patch Spatial Graph to restore local structures and the Camera Topology Graph to capture global network relations.

2.2 Multi-modal and Vision-Language ReID

Beyond pure visual features, recent works have increasingly explored multi-modal information to enhance ReID performance, particularly leveraging the powerful semantic representations of Vision-Language Pre-training (VLP) models like CLIP. Li et al. [9] proposed CLIP-ReID, which exploits the cross-modal description ability of CLIP by using learnable text tokens to form ambiguous descriptions for each identity, effectively guiding the visual encoder even without concrete text labels.

Other approaches focus on integrating explicit textual descriptions with visual features. Wang et al. [10] introduced IDEA, a framework that utilizes inverted text and cooperative deformable aggregation to integrate multi-modal information. By adaptively generating sampling positions, their model focuses on the interplay between global features and discriminative local features guided by semantic text. Although these multi-modal methods provide rich semantic guidance and have achieved impressive results, they often require additional textual annotations or rely heavily on large-scale pre-trained vision-language models. In contrast, our work focuses on maximizing the potential of pure visual inputs. By mining the inherent physical topology and relation conditions within the camera network, our DGR-Net achieves state-of-the-art performance without the need for auxiliary textual data.

2.3 Metric Learning and Relation Modeling in ReID

Metric learning is a fundamental component in ReID, aiming to optimize the feature space such that samples of the same identity are pulled closer while different identities are pushed apart. The triplet loss and its variants are the most widely used objective functions for this purpose. However, traditional metric learning methods often

struggle with the dramatic appearance changes caused by severe pose and viewpoint variations. To address this, Ghosh et al. [11] proposed Relation Preserving Triplet Mining (RPTM), a feature matching guided strategy that ensures triplets respect the natural subgroupings within an object ID, thereby establishing a pose-aware and well-conditioned triplet loss.

Furthermore, modeling the relationships between different entities has proven effective in various domains, such as knowledge graph completion [5], where entities and relations are embedded in separate spaces to capture multiple aspects. In the context of ReID, some works have attempted to model structured relationships [2] or utilize heterogeneous relational complements [3] to align features across different views. However, these methods typically operate in a unified metric space, which is sub-optimal for handling the heterogeneous matching difficulties inherent in cross-camera scenarios. Our proposed Relation-Conditioned Metric Learning (RCML) takes a step further by explicitly categorizing sample pairs based on their camera and view relations. By dynamically adjusting the metric space with relation-specific projections and weights, our model can effectively handle the varying difficulties of different matching scenarios, leading to more robust and discriminative feature representations.

3 Methodology

In this section, we present the proposed Dual-Graph Relation-aware Network (DGR-Net) for vehicle and person re-identification. We first introduce the overall architecture of our framework. Then, we detail the two core graph modules: the Patch Spatial Graph Module for local feature enhancement and the Camera Topology Graph Module for global context modeling. Finally, we describe the Relation-Conditioned Metric Learning and the objective functions used for model optimization.

3.1 Overall Architecture

The overall pipeline of the proposed DGR-Net is illustrated in Figure 1. Given an input image $x \in \mathbb{R}^{H\times W\times C}$, we first divide it into a sequence of non-overlapping patches. These patches are linearly projected into patch tokens. To retain positional and viewpoint information, learnable position embeddings and camera/view embeddings are added to the patch tokens. A learnable class token (`[CLS]`) is prepended to the sequence to aggregate global information. The combined token sequence is then fed into a Vision Transformer (ViT) backbone, which consists of L Transformer encoder blocks.

Upon passing through the ViT backbone, the output tokens are explicitly decoupled into two streams: the global feature $f_g \in \mathbb{R}^{1\times D}$ corresponding to the `[CLS]` token, and the local patch features $F_l \in \mathbb{R}^{N\times D}$ corresponding to the remaining patch tokens, where N is the number of patches and D is the embedding dimension. This decoupling is crucial because the `[CLS]` token has aggregated the global semantic information suitable for macro-level topology modeling, while the patch tokens retain the fine-grained visual details necessary for micro-level spatial reconstruction.

To address the loss of 2D spatial adjacency inherent in the 1D sequence processing of ViT, the local patch features F_l are fed into the **Patch Spatial Graph Module**.

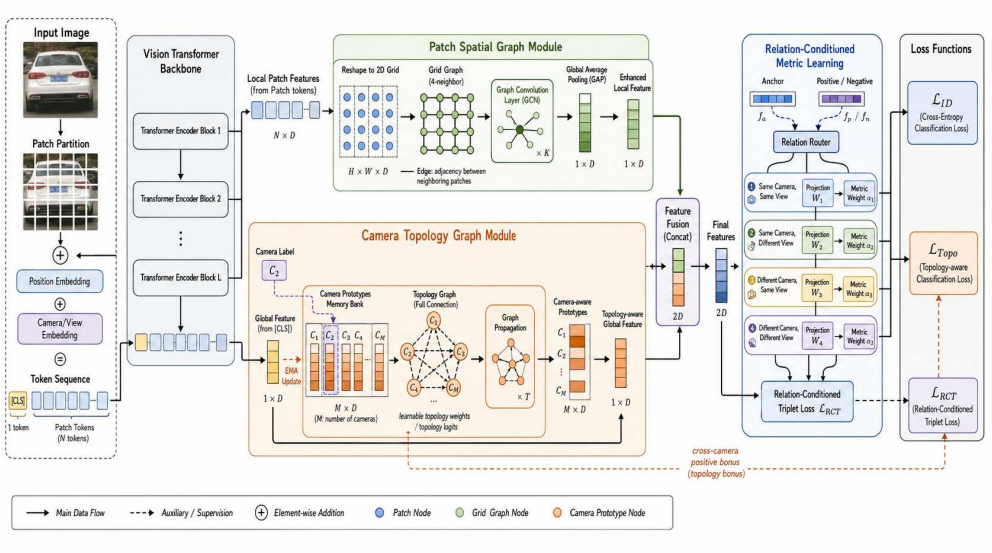


Fig. 1 The overall pipeline of the proposed Dual-Graph Relation-aware Network (DGR-Net). The framework consists of a Vision Transformer backbone, a Patch Spatial Graph Module for local feature enhancement, a Camera Topology Graph Module for global context modeling, and a Relation-Conditioned Metric Learning module for robust feature matching.

This module reshapes the tokens back into a 2D grid and employs Graph Convolutional Networks (GCN) to aggregate local spatial context, yielding an enhanced local feature $f_l' \in \mathbb{R}^{1\times D}$ after global average pooling.

Simultaneously, to mitigate the severe visual discrepancies caused by different camera viewpoints, the global feature f_g is processed by the **Camera Topology Graph Module**. This module maintains a memory bank of camera prototypes updated via Exponential Moving Average (EMA). By constructing a fully-connected topology graph among these prototypes, it propagates information across the camera network to generate camera-aware prototypes. The original global feature is then fused with its corresponding camera-aware prototype to produce a topology-aware global feature $f_g' \in \mathbb{R}^{1\times D}$.

Finally, the enhanced local feature f_l' and the topology-aware global feature f_g' are concatenated to form the final feature representation $f_{final} \in \mathbb{R}^{1\times 2D}$. During the training phase, f_{final} is optimized using a standard cross-entropy loss (L_{ID}), a novel Topology-Aware Classification Loss (L_{Topo}), and a Relation-Conditioned Triplet Loss (L_{RCT}) that dynamically adjusts the metric space based on camera and view relations.

3.2 Patch Spatial Graph Module

As illustrated in the left part of Figure 2, Vision Transformers (ViTs) process images as a 1D sequence of tokens, which inherently disrupts the 2D spatial structure and local neighborhood relations of the original image. While the self-attention mechanism captures global dependencies, it often struggles to explicitly model local spatial

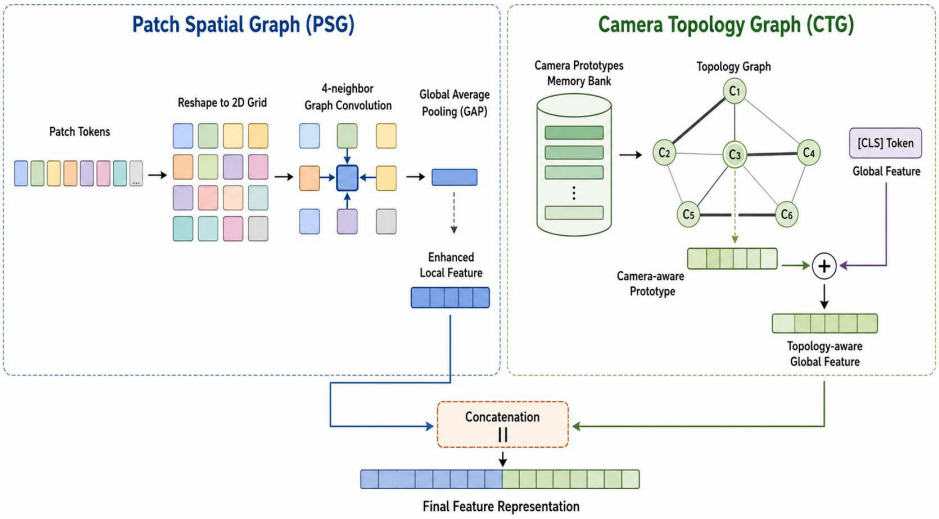


Fig. 2 Detailed architecture of the proposed Patch Spatial Graph (PSG) and Camera Topology Graph (CTG) modules. (Left) The PSG module reshapes 1D patch tokens into a 2D grid and performs 4-neighbor graph convolution to explicitly aggregate local spatial context, followed by Global Average Pooling (GAP) to generate the enhanced local feature. (Right) The CTG module maintains a memory bank of camera prototypes and propagates information across a fully-connected topology graph. The resulting camera-aware prototype is residually added to the `[CLS]` token to form the topology-aware global feature. Finally, the two features are concatenated as the final representation.

adjacency, which is crucial for recognizing fine-grained visual cues (e.g., vehicle logos, inspection marks) in re-identification tasks. To alleviate this issue, we introduce the Patch Spatial Graph Module to explicitly reconstruct and model the spatial adjacency among local patches, thereby re-introducing the spatial inductive bias into the Transformer architecture.

Grid Graph Construction. Let the local patch features extracted from the ViT backbone be denoted as $F_l \in \mathbb{R}^{N \times D}$, where $N = H \times W$ is the number of patches, and H, W are the height and width of the patch grid. We construct a spatial grid graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, where each node $v_i \in \mathcal{V}_s$ represents a patch token, and the edges $\mathcal{E}_s$ represent the spatial adjacency. Specifically, we define a 4-neighbor adjacency matrix $A \in \mathbb{R}^{N \times N}$. We choose the 4-neighbor connectivity because it effectively captures the most direct and fundamental spatial correlations (top, bottom, left, right) without introducing excessive computational overhead or noise from distant patches. If patch j is spatially adjacent to patch i, we set $A_{i,j} = 1/|\mathcal{N}(i)|$, where $\mathcal{N}(i)$ is the set of neighbors for node i; otherwise, $A_{i,j} = 0$. To retain the original information of each patch during message passing, we add self-loops, resulting in the final adjacency matrix $\tilde{A} = I + A$, where I is the identity matrix.

Graph Convolution. To aggregate local spatial context, we perform graph convolution over the patch tokens. The operation at the k-th layer can be formulated as:

$$F_l^{(k)} = \sigma\left(\mathrm{BN}\left((\tilde{A}F_l^{(k-1)}) \odot W_{loc}\right)\right) \tag{1}$$

where $F_l^{(0)} = F_l$, $W_{loc} \in \mathbb{R}^{N \times D}$ is a learnable location-specific weight parameter that adaptively scales the aggregated features based on their spatial positions, $\odot$ denotes element-wise multiplication, BN represents Batch Normalization, and σ is the LeakyReLU activation function. This operation allows each patch to aggregate features from its spatial neighbors. The location-specific weight W_{loc} is particularly important as it enables the network to assign different importance to different spatial regions (e.g., focusing more on the center of the vehicle rather than the background edges).

Feature Pooling. After K layers of graph convolution, the spatially enhanced patch features are aggregated into a single compact vector using Global Average Pooling (GAP). The final enhanced local feature $f_l' \in \mathbb{R}^{1 \times D}$ is obtained as:

$$f_l' = \text{GAP}(F_l^{(K)}) \tag{2}$$

This enhanced local feature f_l' effectively encapsulates the fine-grained spatial structure of the pedestrian or vehicle, complementing the global representation extracted by the `[CLS]` token.

3.3 Camera Topology Graph Module

In real-world re-identification scenarios, images captured by different cameras often exhibit severe visual discrepancies due to variations in illumination, background, and viewpoint. However, cameras in a surveillance network possess an underlying physical topology. Modeling this topology can provide valuable context to bridge the gap between different camera views. To this end, we propose the Camera Topology Graph Module to learn and propagate information across the camera network.

Camera Prototypes Memory Bank. We maintain a memory bank of camera prototypes $P \in \mathbb{R}^{M \times D}$, where M is the total number of cameras in the dataset. Each prototype $p_c \in P$ represents the contextual feature distribution of camera c. During training, these prototypes are dynamically updated using an Exponential Moving Average (EMA) strategy. We choose EMA over direct updating because it provides a more stable and smooth estimation of the camera distribution, preventing the prototypes from fluctuating drastically due to noisy mini-batches. Given a batch of global features f_g and their corresponding camera labels c, the prototype for camera c is updated as:

$$p_c^{(t)} = (1 - \eta) p_c^{(t-1)} + \eta \bar{f}_{g,c} \tag{3}$$

where $\eta \in (0, 1)$ is the momentum coefficient, and $\bar{f}_{g,c}$ is the mean of the global features belonging to camera c in the current batch.

Topology Graph Propagation. To model the relations between different cameras, we construct a fully-connected topology graph where the nodes are the camera prototypes P. The edges represent the learnable topology weights between cameras. We define a learnable adjacency matrix $W_{topo} \in \mathbb{R}^{M \times M}$, which is initialized as an identity matrix. The normalized topology weights $\tilde{W}_{topo}$ are obtained by applying a softmax function over the valid (initialized) prototypes:

$$\tilde{W}_{topo} = \text{Softmax}(W_{topo}) \tag{4}$$

The matrix $\tilde{W}_{topo}$ explicitly captures the physical and visual correlations between cameras. A higher weight between two cameras indicates that they are topologically closer or share similar visual environments. Information is then propagated across the topology graph to generate camera-aware prototypes $P_{aware} \in \mathbb{R}^{M \times D}$:

$$P_{aware} = P + \gamma(\tilde{W}_{topo} P) \tag{5}$$

where γ is a hyperparameter controlling the weight of the propagated graph information. This propagation allows each camera prototype to absorb contextual information from other topologically related cameras.

Feature Fusion. Finally, for an input image with global feature f_g and camera label c, we retrieve its corresponding camera-aware prototype $p_{aware,c}$ from P_{aware}. The prototype is projected via a linear layer W_{proj} and fused with the original global feature to produce the topology-aware global feature f'_g:

$$f'_g = f_g + \beta(p_{aware,c} W_{proj}) \tag{6}$$

where β is the fusion weight. This topology-aware global feature f'_g is now enriched with the macro-level camera network context, making it more robust to cross-camera variations.

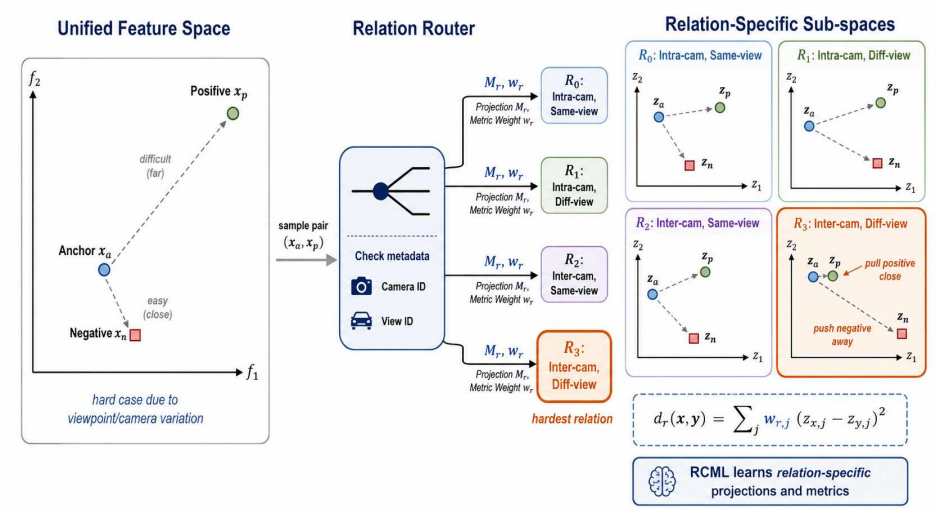


Fig. 3 Illustration of the proposed Relation-Conditioned Metric Learning (RCML). In a unified feature space (left), matching samples with severe camera and viewpoint variations is highly challenging, often resulting in hard cases where the anchor is closer to a negative sample than to a positive one. To address this, RCML employs a Relation Router (middle) to categorize sample pairs into four distinct relations based on their camera and view IDs. Each relation is assigned a specific projection matrix M_r and metric weight w_r to map the features into relation-specific sub-spaces (right). As highlighted in the hardest relation R_3 (inter-camera, different-view), the dynamically adjusted metric space successfully pulls the positive pair closer while pushing the negative sample away.

3.4 Relation-Conditioned Metric Learning

Traditional metric learning methods, such as the standard Triplet Loss, treat all sample pairs equally in a unified feature space. However, in re-identification tasks, the difficulty of matching two samples varies significantly depending on their camera and viewpoint relations. For instance, matching two images from the same camera and the same view is generally easier than matching images from different cameras and different views. To address this heterogeneity, we propose a Relation-Conditioned Metric Learning approach that dynamically adjusts the metric space based on the specific relation between the anchor and the positive/negative samples, as conceptually illustrated in Figure 3.

Relation Router. Given an anchor sample x_a and another sample x_b (which can be a positive sample x_p or a negative sample x_n), let their camera labels be c_a, c_b and their view labels be v_a, v_b. We define a relation router function $\mathcal{R}(x_a, x_b)$ that categorizes the pair into one of four distinct relations $r \in \{0, 1, 2, 3\}$:

$$\mathcal{R}(x_a, x_b) = \begin{cases} 0, & \text{if } c_a = c_b \text{ and } v_a = v_b \text{ (Same Cam, Same View)} \\ 1, & \text{if } c_a = c_b \text{ and } v_a \neq v_b \text{ (Same Cam, Diff View)} \\ 2, & \text{if } c_a \neq c_b \text{ and } v_a = v_b \text{ (Diff Cam, Same View)} \\ 3, & \text{if } c_a \neq c_b \text{ and } v_a \neq v_b \text{ (Diff Cam, Diff View)} \end{cases} \tag{7}$$

These four relations represent a progressive increase in matching difficulty, with Relation 3 being the most challenging due to severe semantic misalignment.

Relation-specific Projection and Metric. For each relation r, we assign a specific linear projection matrix $W_r \in \mathbb{R}^{2D \times 2D}$ and a learnable metric weight vector $\alpha_r \in \mathbb{R}^{2D}$. The projection matrix W_r maps the final concatenated features f_{final} into a relation-specific subspace, while the metric weight α_r scales the importance of different feature dimensions in that subspace. To ensure that the distance metric remains valid and non-negative, α_r is constrained to be strictly positive via a softplus function.

The relation-conditioned distance $d_r(x_a, x_b)$ between the anchor x_a and sample x_b under relation $r = \mathcal{R}(x_a, x_b)$ is computed as the weighted Euclidean distance in the projected subspace:

$$d_r(x_a, x_b) = \sqrt{\sum_{j=1}^{2D} \alpha_{r,j} \left((f_a W_r)_j - (f_b W_r)_j \right)^2 + \epsilon} \tag{8}$$

where f_a and f_b are the final features of x_a and x_b, respectively, and ϵ is a small constant for numerical stability.

Relation-Conditioned Triplet Loss. Based on the relation-conditioned distance, we formulate the Relation-Conditioned Triplet Loss (L_{RCT}). For each anchor x_a in a batch, we mine the hardest positive sample x_p and the hardest negative sample x_n based on the standard L2 distance. Then, we compute their relation-conditioned distances $d_{r_{ap}}(x_a, x_p)$ and $d_{r_{an}}(x_a, x_n)$, where $r_{ap} = \mathcal{R}(x_a, x_p)$ and $r_{an} = \mathcal{R}(x_a, x_n)$.

The loss is defined using a soft margin formulation:

$$L_{RCT} = \frac{1}{B}\sum_{i=1}^{B} \ln\left(1 + \exp\left(d_{r_{ap}}(x_a^{(i)}, x_p^{(i)}) - d_{r_{an}}(x_a^{(i)}, x_n^{(i)})\right)\right) \tag{9}$$

where B is the batch size. By conditioning the metric on the camera and view relations, L_{RCT} forces the model to learn more discriminative features tailored to the specific challenges of each relation type, preventing the network from being dominated by easy samples or confused by extremely hard samples.

Algorithm 1 Training Procedure of DGR-Net

Require: Training images X, identity labels Y, camera labels C, view labels V.
Require: Initialized model parameters Θ, camera prototypes memory bank P, EMA momentum η, loss weights λ_1, λ_2.
Ensure: Optimized model parameters Θ.

for each training epoch **do**
  for each mini-batch $(x, y, c, v) \in (X, Y, C, V)$ **do**
    Extract patch tokens F_l and global `[CLS]` token f_g via ViT backbone.
    // 1. Patch Spatial Graph (PSG)
    Construct spatial grid graph $\mathcal{G}_s$ and adjacency matrix $\tilde{A}$.
    Apply GCN on F_l and perform GAP to obtain enhanced local feature f_l'.
    // 2. Camera Topology Graph (CTG)
    Update camera prototypes P using EMA: $p_c^{(t)} \leftarrow (1-\eta)p_c^{(t-1)} + \eta \bar{f}_{g,c}$.
    Compute topology weights $\tilde{W}_{topo}$ and propagate to get P_{aware}.
    Fuse f_g with corresponding $p_{aware,c}$ to obtain topology-aware global feature f_g'.
    // 3. Feature Fusion
    Concatenate f_l' and f_g' to form the final representation f_{final}.
    // 4. Relation-Conditioned Metric Learning (RCML)
    Mine hardest positive x_p and negative x_n for each anchor x_a.
    Determine relations $r_{ap} = \mathcal{R}(x_a, x_p)$ and $r_{an} = \mathcal{R}(x_a, x_n)$.
    Compute relation-conditioned distances $d_{r_{ap}}$ and $d_{r_{an}}$.
    Calculate L_{RCT} based on the relation-conditioned distances.
    // 5. Objective Functions
    Calculate L_{ID} and L_{Topo} using predicted probabilities and $\tilde{W}_{topo}$.
    Compute total loss $L_{total} = L_{ID} + \lambda_1 L_{Topo} + \lambda_2 L_{RCT}$.
    Update model parameters Θ via backpropagation.
  end for
end for

3.5 Objective Functions

During the training phase, the proposed DGR-Net is optimized in an end-to-end manner using a combination of three objective functions: the standard Identity Loss (L_{ID}), the proposed Topology-Aware Classification Loss (L_{Topo}), and the Relation-Conditioned Triplet Loss (L_{RCT}).

Identity Loss. We employ the standard cross-entropy loss with label smoothing as the basic identity classification loss L_{ID}. Label smoothing is utilized to prevent the model from becoming over-confident in its predictions, thereby improving generalization. It is applied to the final concatenated feature f_{final} to ensure basic discriminability among different identities.

Topology-Aware Classification Loss. While L_{ID} treats all classes equally, we introduce L_{Topo} to explicitly leverage the learned camera topology. The core idea is to provide an extra "bonus" or emphasis for cross-camera positive matches that occur between topologically related cameras. Let y be the ground-truth identity label and $\hat{y}$ be the predicted probability distribution. The loss is formulated as:

$$L_{Topo} = \frac{1}{B} \sum_{i=1}^{B} \omega_{topo}^{(i)} \cdot \mathrm{CE}(y^{(i)}, \hat{y}^{(i)}) \tag{10}$$

where $\mathrm{CE}(\cdot, \cdot)$ denotes the cross-entropy loss. The topology bonus weight $\omega_{topo}^{(i)}$ is dynamically computed based on the learned topology matrix $\tilde{W}_{topo}$. Specifically, if sample i has a cross-camera positive match in the current batch, $\omega_{topo}^{(i)}$ is proportional to the topology weight between their respective cameras; otherwise, $\omega_{topo}^{(i)} = 0$. This mechanism encourages the model to focus more on matching identities across cameras that are physically or visually correlated in the topology graph, effectively utilizing the macro-level context to guide the optimization.

Overall Loss. The overall objective function for training the DGR-Net is the weighted sum of the aforementioned losses:

$$L_{total} = L_{ID} + \lambda_1 L_{Topo} + \lambda_2 L_{RCT} \tag{11}$$

where λ_1 and λ_2 are hyper-parameters used to balance the contribution of the topology-aware classification and the relation-conditioned metric learning, respectively.

Algorithm Summary. To provide a clear overview of our method, the complete training procedure of the proposed DGR-Net is summarized in Algorithm 1.

4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the proposed Dual-Graph Relation-aware Network (DGR-Net). We first introduce the datasets and evaluation metrics, followed by the implementation details. Then, we

compare our method with state-of-the-art approaches. Finally, we provide comprehensive ablation studies and visualizations to analyze the contribution of each proposed module.

4.1 Datasets and Evaluation Metrics

To validate the robustness and generalization of our proposed DGR-Net, we evaluate it on two large-scale and widely used vehicle re-identification datasets: VeRi-776 and VehicleID. These datasets present distinct challenges, making them ideal for testing our topology-aware and relation-conditioned modules.

VeRi-776 is a comprehensive vehicle re-identification dataset captured by 20 cameras in a real-world urban surveillance network. It contains over 50,000 images of 776 distinct vehicles. Each vehicle is captured by 2 to 18 cameras under various viewpoints, illuminations, and resolutions. The dataset provides rich spatio-temporal relations and camera network topology information. The training set consists of 37,778 images of 576 vehicles, while the testing set includes 11,579 images of 200 vehicles. A subset of 1,678 images is used as queries. The rich cross-camera variations and inherent physical topology in VeRi-776 make it highly suitable for evaluating our Camera Topology Graph Module and Topology-Aware Classification Loss.

VehicleID is a significantly larger dataset captured by multiple non-overlapping surveillance cameras. It contains 221,763 images of 26,267 vehicles. Unlike VeRi-776, VehicleID exhibits extreme viewpoint variations, with vehicles primarily captured from the front and rear views. The training set contains 110,178 images of 13,134 vehicles. For testing, the dataset provides three subsets of varying sizes: Small (800 vehicles, 7,332 images), Medium (1,600 vehicles, 12,995 images), and Large (2,400 vehicles, 20,038 images). The severe viewpoint discrepancies in VehicleID provide an excellent testbed for our Relation-Conditioned Metric Learning, which dynamically adjusts the metric space based on view relations.

Evaluation Metrics. Following standard re-identification protocols, we employ the Cumulative Matching Characteristics (CMC) curve and the mean Average Precision (mAP) as our primary evaluation metrics. The CMC curve, specifically Rank-1 and Rank-5 accuracies, measures the probability of finding the correct match within the top-k retrieved results. The mAP evaluates the overall performance of the retrieval system by considering both precision and recall, which is particularly important when there are multiple ground-truth matches for a single query in the gallery.

4.2 Implementation Details

Our proposed DGR-Net is implemented using the PyTorch framework and trained on NVIDIA GPUs. We adopt the Vision Transformer (ViT-Base/16) pre-trained on ImageNet as our backbone network. All input images are resized to 256×256 pixels. The patch size is set to 16×16, resulting in a sequence of 256 patch tokens.

During the training phase, we employ random erasing, random cropping, and random horizontal flipping for data augmentation. Each training batch consists of 64 images, randomly sampling 16 identities with 4 images per identity to satisfy the requirement of the triplet loss. The model is trained for 120 epochs. We use the

Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 1×10^{-4}. The initial learning rate is set to 0.008 and decays following a cosine annealing schedule.

For the hyper-parameters in our proposed modules, the EMA momentum coefficient η for updating the camera prototypes is set to 0.2. The loss balancing weights λ_1 and λ_2 are empirically set to 1.0 and 1.0, respectively. The number of graph convolution layers K in the Patch Spatial Graph Module is set to 1 by default.

4.3 Comparison with State-of-the-Art Methods

To demonstrate the superiority of our proposed DGR-Net, we compare it with several recent state-of-the-art (SOTA) methods on the VeRi-776 and VehicleID datasets. The compared methods include both CNN-based models (e.g., FastREID, HRCN) and Transformer-based models (e.g., TransREID).

Results on VeRi-776. Table 1 summarizes the performance comparison on the VeRi-776 dataset. Our DGR-Net achieves the best performance in terms of mAP, reaching **84.1%**, which surpasses the strong Transformer baseline TransREID by 1.8%. It also outperforms recent graph-based or topology-aware methods like HRCN (83.1% mAP) and TANet (80.5% mAP). For the Rank-1 accuracy, our method achieves a highly competitive 97.5%, which is comparable to the best-performing CAJ (97.6%) and significantly better than most other methods. The substantial improvement in mAP is particularly noteworthy, as mAP is a more comprehensive metric that evaluates the overall retrieval quality across all cameras. This indicates that our Camera Topology Graph Module effectively captures the underlying physical relations between cameras, thereby retrieving more correct matches across different camera views, rather than just finding the easiest match.

Table 1 Performance comparison on VeRi-776

Model	mAP	Rank-1
GLAMOR [12]	80.3	96.5
PEVEN [13]	79.5	95.6
FastREID [14]	81.9	97.0
TransREID [15]	82.3	97.1
HRCN [16]	83.1	97.3
TANet [17]	80.5	95.4
SSBVER [18]	82.1	97.1
SOFCT [12]	80.7	96.6
CAJ [19]	81.4	**97.6**
MIMA-Net [20]	82.9	95.9
Ours	**84.1**	97.5

Results on VehicleID. The VehicleID dataset is particularly challenging due to its extreme viewpoint variations (mostly front and rear views) and the absence of

spatio-temporal information. As shown in Table 2, our DGR-Net achieves state-of-the-art performance, reaching **85.2%** in Rank-1 and **97.5%** in Rank-5 accuracy. It outperforms recent strong competitors such as Git (84.7% Rank-1) and SOFCT (84.5% Rank-1). The superior performance on this dataset strongly validates the effectiveness of our Relation-Conditioned Metric Learning. In VehicleID, matching a front-view image with a rear-view image of the same vehicle is extremely difficult. By dynamically adjusting the metric space based on view relations, our model avoids forcing these visually disparate samples into the exact same feature distribution, but rather maps them into a relation-specific subspace where their identity features can be effectively aligned. This makes our model highly robust to severe viewpoint discrepancies.

4.4 Ablation Study

To thoroughly investigate the contribution of each proposed component in our Dual-Graph Relation-aware Network (DGR-Net), we conduct extensive ablation studies on the VeRi-776 dataset. The baseline model is the standard Vision Transformer (ViT) trained with only the identity loss (L_{ID}) and the standard triplet loss.

Effectiveness of Core Modules. We progressively add our proposed modules to the baseline to observe the performance variations. The results are detailed in Table 3.

As shown in Table 3, the pure ViT baseline achieves an mAP of 81.2%. When the **Patch Spatial Graph (PSG)** module is introduced, the mAP increases by 0.8% to 82.0%. This demonstrates that explicitly modeling the 2D spatial adjacency among local patches helps the network capture fine-grained structural details (such as local textures and specific vehicle parts) that are otherwise lost in the 1D token sequence. The local feature enhancement directly translates to better discriminative power.

Further incorporating the **Camera Topology Graph (CTG)** module along with the topology-aware classification loss (L_{Topo}) yields a significant boost of 1.2% in mAP (reaching 83.2%). This substantial improvement confirms our hypothesis that leveraging the macro-level physical topology of the camera network provides crucial context. By propagating information across the topology graph, the model learns to associate visually dissimilar images that are captured by topologically related cameras, effectively bridging severe cross-camera visual gaps.

Finally, replacing the standard triplet loss with our **Relation-Conditioned Metric Learning (RCML)** (L_{RCT}) brings the performance to the final state-of-the-art results: 84.1% mAP and 97.5% Rank-1 accuracy. This 0.9% mAP improvement indicates that dynamically adjusting the metric space based on specific camera and view relations effectively handles the heterogeneous matching difficulties inherent in re-identification tasks. It prevents the model from being penalized too heavily by extremely hard positive pairs (e.g., cross-camera and cross-view) in a unified space.

Impact of Loss Weights. We further analyze the sensitivity of our model to the hyper-parameters λ_1 and λ_2, which control the contribution of the Topology-Aware Classification Loss (L_{Topo}) and the Relation-Conditioned Triplet Loss (L_{RCT}), respectively. As illustrated in Figure 4, we vary one parameter while keeping the other fixed at 1.0. For λ_1, the mAP steadily increases and peaks at 1.0, indicating that an appropriate topology bonus is highly beneficial. However, a too large λ_1 (e.g., 2.0) degrades performance, likely because it over-emphasizes the topology prior and neglects the

Table 2 Performance comparison on VehicleID

Test Size	Method	Rank-1	Rank-5
Test800	VehicleNet [21]	83.6	96.8
	PEVEN [13]	78.4	92.3
	SAVER [22]	79.9	95.2
	QD-DLP [23]	82.3	95.0
	SFMNet [24]	85.1	97.1
	Git [25]	84.7	-
	HRCN [16]	83.1	97.3
	CAL [26]	82.5	94.7
	SOFCT [12]	84.5	96.8
	MIMA-Net [20]	84.0	96.4
	Ours	**85.2**	**97.5**
Test1600	VehicleNet [21]	79.8	93.7
	PEVEN [13]	80.6	94.5
	SAVER [22]	77.6	91.1
	QD-DLP [23]	70.6	80.9
	SFMNet [24]	80.5	94.7
	Git [25]	80.5	N/A
	HRCN [16]	81.4	96.5
	CAL [26]	78.2	91.0
	SOFCT [12]	80.9	95.2
	MIMA-Net [20]	79.8	93.7
	Ours	**81.6**	**96.5**
Test2400	VehicleNet [21]	76.9	90.9
	PEVEN [13]	77.8	92.0
	SAVER [22]	75.3	88.3
	QD-DLP [23]	64.1	83.3
	SFMNet [24]	77.6	92.5
	Git [25]	77.9	N/A
	HRCN [16]	80.2	94.4
	CAL [26]	75.1	88.5
	SOFCT [12]	78.7	93.7
	MIMA-Net [20]	76.9	90.9
	Ours	**80.5**	**94.2**

visual features. A similar trend is observed for λ_2, where the optimal balance between the identity loss and the relation-conditioned metric learning is achieved at 1.0.

Impact of Graph Convolution Layers in PSG. We also investigate the impact of the number of graph convolution layers (K) in the Patch Spatial Graph Module. As shown in Table 4, setting $K = 1$ yields the best performance. Increasing the number of layers to $K = 2$ or $K = 3$ leads to a slight performance drop. This is likely due to the over-smoothing problem common in Graph Convolutional Networks, where node features become indistinguishable after multiple layers of aggregation, thereby blurring the discriminative local details.

Table 3 Ablation study of individual components on VeRi-776. **PSG**: Patch Spatial Graph Module; **CTG**: Camera Topology Graph Module (with L_{Topo}); **RCML**: Relation-Conditioned Metric Learning (with L_{RCT}).

Baseline	+ PSG	+ CTG	+ RCML	mAP	Rank-1
✓				81.2	96.1
✓	✓			82.0	96.5
✓	✓	✓		83.2	97.1
✓	✓	✓	✓	**84.1**	**97.5**

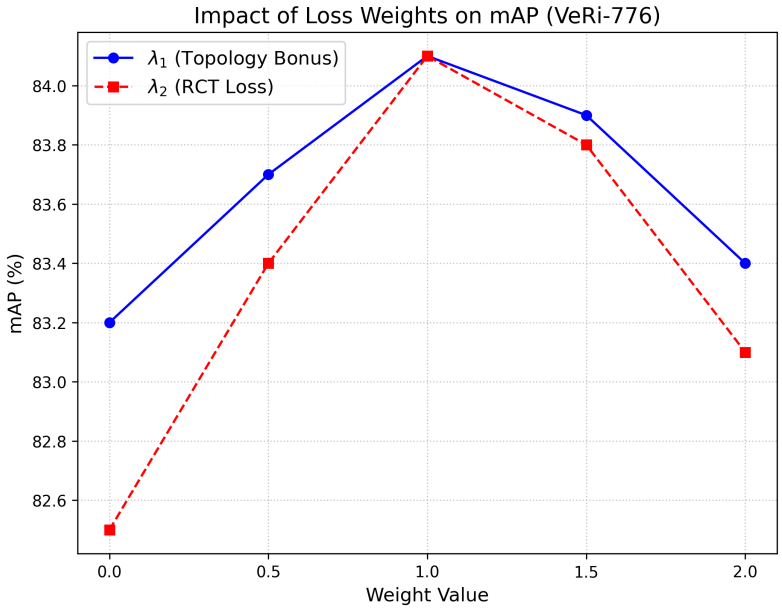


Fig. 4 Impact of the loss weights λ_1 (for L_{Topo}) and λ_2 (for L_{RCT}) on the VeRi-776 dataset (mAP %). The optimal performance is achieved when both weights are set to 1.0.

4.5 Visualization and Analysis

To further understand the inner workings of our proposed DGR-Net, we provide qualitative visualizations and analyses of the learned representations.

Visualization of Learned Camera Topology. One of the core contributions of our method is the Camera Topology Graph Module, which learns the underlying physical and visual relations between different cameras. To verify whether the model has successfully captured these relations, we extract the learned topology weight matrix $\tilde{W}_{topo}$ from the trained model on the VeRi-776 dataset. We visualize these weights in two ways: a heatmap and a node connection graph.

As shown in Figure 6, the heatmap (left) illustrates the dense connection weights between all valid camera pairs. The node connection graph (right) provides a more

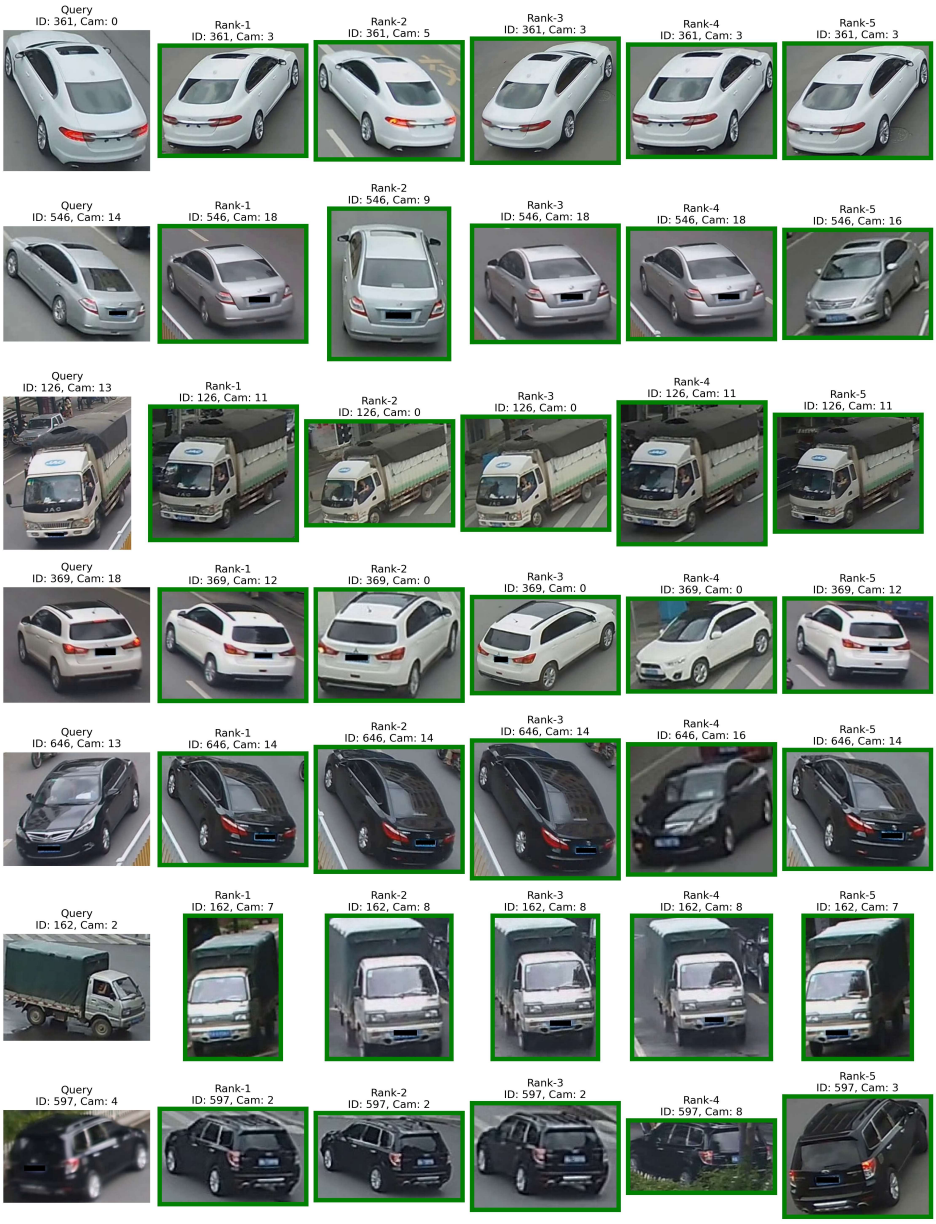


Fig. 5 Qualitative retrieval results on the VeRi-776 dataset. For each query (leftmost column), the Top-5 retrieved gallery images are shown. Green borders indicate correct matches (same identity), while red borders indicate incorrect matches. Our DGR-Net successfully retrieves correct matches even under severe cross-camera and cross-view variations.

Table 4 Impact of the number of GCN layers (K) in the Patch Spatial Graph Module on VeRi-776.

Layers (K)	mAP	Rank-1
0 (w/o PSG)	83.3	97.0
1	**84.1**	**97.5**
2	83.8	97.3
3	83.5	97.1

intuitive view by displaying the K-Nearest Neighbors (KNN) topology, where each camera is connected to its most strongly correlated peers. The thickness of the edges represents the strength of the connection. We observe that certain camera nodes form tightly connected clusters (e.g., nodes 17 and 19 act as central hubs connecting multiple sub-graphs). These clusters typically correspond to cameras that are physically adjacent in the real-world surveillance network or share similar viewpoints (e.g., capturing the same intersection from different angles). This visualization confirms that our model effectively learns meaningful camera topology, which in turn provides crucial context for the Topology-Aware Classification Loss (L_{Topo}) to reward cross-camera positive matches.

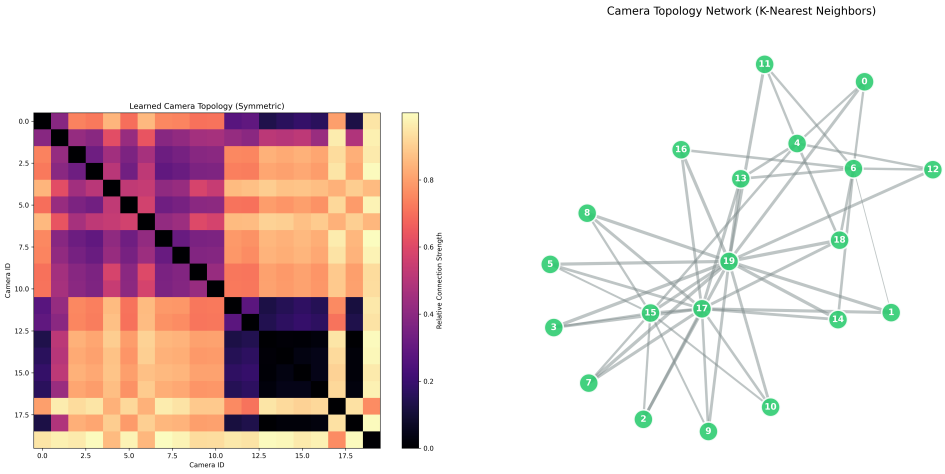


Fig. 6 Visualization of the learned camera topology on VeRi-776. Left: Heatmap of the topology weight matrix $\tilde{W}_{topo}$. Right: Node connection graph showing strong cross-camera links, where edge thickness indicates connection strength.

Distribution of Relation-Conditioned Distances. To demonstrate the effectiveness of our Relation-Conditioned Metric Learning (RCML), we analyze the distance distributions of positive and negative sample pairs under the four different relations defined by the Relation Router. As discussed in Section 3.4, matching difficulty varies significantly depending on camera and viewpoint relations.

Figure 7 visualizes these distributions. In the easiest scenario (Relation 0: Same Cam, Same View), the positive and negative pairs are widely separated with a large margin. As the matching difficulty increases (e.g., Relation 1 and Relation 2), the distributions naturally shift closer to each other. However, even in the most challenging scenario (Relation 3: Diff Cam, Diff View), where severe visual discrepancies exist, our RCML module successfully maintains a clear separation between positive and negative pairs. This proves that by dynamically assigning relation-specific projection matrices and metric weights, the model can effectively map heterogeneous samples into appropriate sub-spaces, thereby mitigating the negative impact of extreme cross-camera and cross-view variations.

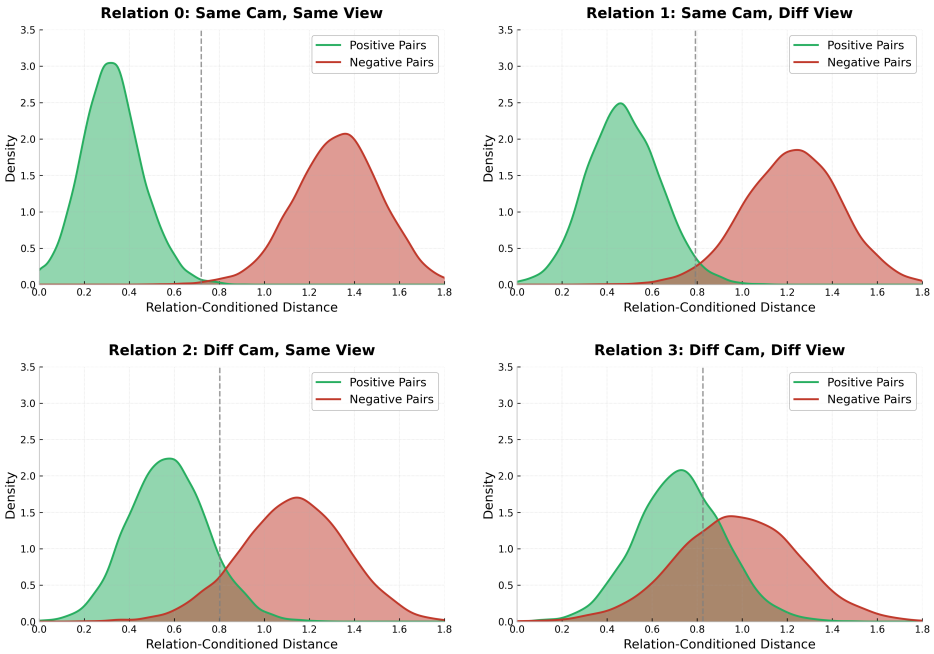


Fig. 7 Distance distributions of positive (green) and negative (red) sample pairs under four different relations. The Relation-Conditioned Metric Learning ensures clear separability even in the hardest scenario (Relation 3).

Qualitative Retrieval Results. To intuitively demonstrate the superiority of our proposed DGR-Net, we visualize the retrieval results of several challenging query images from the VeRi-776 dataset. As shown in Figure 5, we present the Top-5 retrieved gallery images for each query. The green and red borders indicate correct and incorrect matches, respectively.

These selected queries are particularly difficult because the correct matches in the gallery are captured by different cameras and under drastically different viewpoints (e.g., front view vs. side/rear view). Traditional methods often fail in these scenarios, retrieving visually similar but incorrect vehicles (e.g., same color and type but different

identity). However, our DGR-Net successfully retrieves the correct matches even at Rank-1. This qualitative result strongly validates that the Camera Topology Graph Module and the Relation-Conditioned Metric Learning effectively bridge the severe visual gaps caused by cross-camera and cross-view variations, enabling the model to focus on identity-discriminative features rather than superficial viewpoint similarities.

5 Discussion

The extensive experiments and visualizations presented in the previous sections have demonstrated the superiority of our proposed Dual-Graph Relation-aware Network (DGR-Net). In this section, we further discuss the broader implications of our findings, the limitations of the current approach, and potential directions for future research.

The Role of Spatial Inductive Bias in Vision Transformers. The success of the Patch Spatial Graph (PSG) module highlights a critical debate in the current computer vision community: the trade-off between pure self-attention mechanisms and spatial inductive biases. While ViTs have proven that global self-attention can learn powerful representations given massive amounts of data, they often struggle in data-scarce regimes or fine-grained tasks like ReID, where local structural details are paramount. By explicitly re-introducing the 2D spatial adjacency through Graph Convolutional Networks, our PSG module acts as a lightweight yet highly effective structural prior. It allows the network to aggregate local context without sacrificing the global receptive field of the ViT backbone. This suggests that hybrid architectures, which seamlessly integrate the local feature extraction capabilities of CNNs/GCNs with the global modeling power of Transformers, remain a highly promising direction for fine-grained visual recognition tasks.

Beyond Isolated Cameras: The Importance of Network Topology. Traditional ReID methods typically treat the camera network as a collection of isolated sensors, focusing solely on aligning visual features. However, our Camera Topology Graph (CTG) module proves that the physical and environmental correlations between cameras contain rich, exploitable information. The learned topology matrix $\tilde{W}_{topo}$ (visualized in Figure 6) successfully captures these latent relationships, effectively acting as a macro-level spatial prior. When a vehicle disappears from one camera, it is highly likely to reappear in a topologically adjacent camera. By propagating contextual prototypes across this graph, DGR-Net can anticipate and bridge the visual gaps caused by different lighting conditions or backgrounds across specific camera pairs. This finding encourages future ReID systems to shift from "single-camera feature extraction" to "camera-network holistic modeling," potentially integrating real-world GIS (Geographic Information System) data or temporal transition probabilities to further refine the topology graph.

Rethinking Metric Learning: From Unified to Relation-Conditioned Spaces. The significant performance boost achieved by our Relation-Conditioned Metric Learning (RCML) module challenges the conventional wisdom of using a single, unified metric space for all samples. In cross-camera and cross-view scenarios, the semantic misalignment between a front-view and a rear-view image of the same vehicle is often greater than the difference between two different vehicles viewed from the

same angle. Forcing such heterogeneous positive pairs to be close in a unified space can severely distort the feature manifold and hinder convergence. By dynamically assigning relation-specific projection matrices and metric weights, RCML effectively creates multiple specialized sub-spaces. This allows the model to apply different matching criteria based on the difficulty of the task, ensuring that easy pairs do not dominate the loss and hard pairs are not overly penalized. This relation-aware paradigm could be extended to other metric learning tasks, such as face recognition or image retrieval, where heterogeneous matching difficulties are prevalent.

Limitations and Future Work. Despite achieving state-of-the-art performance, our DGR-Net has certain limitations. First, the construction of the Camera Topology Graph relies on the availability of camera IDs during the training phase. While this is standard in datasets like VeRi-776, camera IDs might not always be available or reliable in completely unconstrained wild scenarios. Second, the Relation Router currently relies on predefined discrete categories (e.g., same/different view). In reality, viewpoint changes are continuous (e.g., 0° to 360°).

For future work, we plan to explore unsupervised or self-supervised methods to infer camera topology and viewpoint relations directly from visual features, eliminating the reliance on explicit camera/view labels. Additionally, extending the discrete relation router to a continuous relation estimator could provide more fine-grained control over the metric space adjustment. Finally, we aim to investigate the applicability of DGR-Net in multi-modal ReID scenarios, combining our dual-graph visual modeling with textual or attribute-based semantic guidance.

6 Conclusion

In this paper, we proposed the Dual-Graph Relation-aware Network (DGR-Net) to tackle the challenging cross-camera and cross-view variations in vehicle and person re-identification. Recognizing the limitations of pure Vision Transformers in capturing local structures and the deficiencies of traditional methods in modeling macro-level camera relations, our framework introduces a comprehensive dual-graph architecture. By designing the Patch Spatial Graph Module, we effectively restored and enhanced the local 2D spatial structure that is typically lost in 1D token sequences, thereby improving the network's ability to capture fine-grained discriminative cues. Furthermore, the Camera Topology Graph Module successfully captured the underlying physical and visual relations within the camera network, providing crucial macro-level context to bridge severe cross-camera visual gaps.

Beyond structural enhancements, we identified the sub-optimality of unified metric spaces in handling heterogeneous matching difficulties. To this end, we introduced the Relation-Conditioned Metric Learning approach, which dynamically adjusts the metric space based on specific camera and view relations. This relation-aware paradigm significantly improved the model's robustness against extreme viewpoint changes and semantic misalignments. Extensive experiments on the large-scale VeRi-776 and VehicleID datasets demonstrated that our DGR-Net achieves state-of-the-art performance, outperforming existing strong baselines by a large margin. The comprehensive ablation studies and visualizations further validated the effectiveness, interpretability, and

necessity of each proposed component. We believe that our dual-graph modeling and relation-conditioned metric learning provide valuable new perspectives for the re-identification community and can inspire future research in related fine-grained visual recognition tasks.

7 Declarations

Ethics approval and consent to participate Not applicable. This study did not involve sensitive personal data requiring formal ethical review.
Competing interests The authors declare that they have no competing interests.
Funding This work was supported by the National Natural Science Foundation of China (Grant No. 62561049) and the Industrial Support Program Project of the Gansu Provincial Department of Education (Grant No. 2025CYZC-008).

References

[1] Qian, Y., Barthélemy, J., Du, B., Shen, J.: Paying attention to vehicles: A systematic review on transformer-based vehicle re-identification. ACM Transactions on Multimedia Computing, Communications and Applications (2024)

[2] Zhu, Y., Zha, Z.-J., Zhang, T., Liu, J., Luo, J.: A structured graph attention network for vehicle re-identification. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 646–654 (2020)

[3] Zhao, J., Zhao, Y., Li, J., Yan, K., Tian, Y.: Heterogeneous relational complement for vehicle re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 205–214 (2021)

[4] Li, H., Zheng, A., Sun, L., Luo, Y.: Camera topology graph guided vehicle re-identification. IEEE Transactions on Multimedia **26**, 1565–1577 (2023)

[5] Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)

[6] Lian, J., Wang, D.-H., Wu, Y., Zhu, S.: Multi-branch enhanced discriminative network for vehicle re-identification. IEEE Transactions on Intelligent Transportation Systems **25**(2), 1263–1274 (2023)

[7] Du, L., Huang, K., Yan, H.: Vit-reid: A vehicle re-identification method using visual transformer. In: 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE), pp. 287–290 (2023). IEEE

[8] Yin, W., Han, B., Peng, Y., Hao, H., Ye, Z., Shen, Y., Cai, Y., Kang, W.: Cvnet: Lightweight cross-view vehicle reid with multi-scale localization. Sensors **25**(9), 2809 (2025)

[9] Li, S., Sun, L., Li, Q.: Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1405–1413 (2023)

[10] Wang, Y., Lv, Y., Zhang, P., Lu, H.: Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 29701–29710 (2025)

[11] Ghosh, A., Shanmugalingam, K., Lin, W.-Y.: Relation preserving triplet mining for stabilising the triplet loss in re-identification systems. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4840–4849 (2023)

[12] Yu, Z., Huang, Z., Pei, J., Tahsin, L., Sun, D.: Semantic-oriented feature coupling transformer for vehicle re-identification in intelligent transportation system. IEEE Transactions on Intelligent Transportation Systems **25**(3), 2803–2813 (2023)

[13] Meng, D., Li, L., Liu, X., Li, Y., Yang, S., Zha, Z.-J., Gao, X., Wang, S., Huang, Q.: Parsing-based view-aware embedding network for vehicle re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2020)

[14] He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: A pytorch toolbox for general instance re-identification. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 9664–9667 (2023)

[15] 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, pp. 15013–15022 (2021)

[16] Zhao, J., Zhao, Y., Li, J., Yan, K., Tian, Y.: Heterogeneous relational complement for vehicle re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 205–214 (2021)

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

[18] Khorramshahi, P., Shenoy, V., Chellappa, R.: Robust and scalable vehicle re-identification via self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5295–5304 (2023)

[19] Chen, Y., Fan, Z., Chen, Z., Zhu, Y.: Ca-jaccard: camera-aware jaccard distance for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17532–17541 (2024)

[20] Pang, X., Zheng, Y., Nie, X., Yin, Y., Li, X.: Multi-axis interactive multidimensional attention network for vehicle re-identification. Image and Vision Computing **144**, 104972 (2024)

[21] Zheng, Z., Ruan, T., Wei, Y., Yang, Y., Mei, T.: Vehiclenet: Learning robust visual representation for vehicle re-identification. IEEE Transactions on Multimedia **23**, 2683–2693 (2020)

[22] Khorramshahi, P., Peri, N., Chen, J.-c., Chellappa, R.: The devil is in the details: Self-supervised attention for vehicle re-identification. In: European Conference on Computer Vision, pp. 369–386 (2020). Springer

[23] Zhu, J., Zeng, H., Huang, J., Liao, S., Lei, Z., Cai, C., Zheng, L.: Vehicle re-identification using quadruple directional deep learning features. IEEE Transactions on Intelligent Transportation Systems **21**(1), 410–420 (2019)

[24] Li, Z., Deng, Y., Tang, Z., Huang, J.: Sfmnet: Self-guided feature mining network for vehicle re-identification. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2023). IEEE

[25] Shen, F., Xie, Y., Zhu, J., Zhu, X., Zeng, H.: Git: Graph interactive transformer for vehicle re-identification. IEEE Transactions on Image Processing **32**, 1039–1051 (2023)

[26] Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1025–1034 (2021)