University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

Enhanced Deep Feature Embedding for Vehicle Re-Identification using Hybrid CNN-Transformer Architecture

Akram Jabbar Mohaisen, Informatics Institute for Postgraduate Studies, University of Information Technology and Communications, Baghdad, Iraq (ms202420025@uoitc.edu.iq).

Huda kadhim tayyeh, University of Information Technology and Communications, Baghdad, Iraq (haljobori@uoitc.edu.iq).

Abstract

Vehicle re-identification (Re-ID) plays a central role in modern intelligent surveillance systems, yet its performance remains highly sensitive to real-world imaging conditions, particularly variations in illumination and contrast. Although recent hybrid architectures have advanced representation learning by combining convolutional and transformer-based models, the influence of input quality on the resulting feature embeddings is often underestimated. To address this limitation, this paper introduces the Preprocessing-Enhanced Hybrid Embedding Network (PHE-Net), an end-to-end framework that explicitly integrates input enhancement with hybrid feature learning. Rather than introducing a completely new preprocessing operator or a new backbone family, the contribution of this work lies in a surveillance-oriented integration strategy that explicitly couples input-quality normalization with hybrid local-global embedding learning for vehicle re-identification. The preprocessing stage is designed to mitigate common visual degradations through Contrast Limited Adaptive Histogram Equalization (CLAHE) and gamma correction, while maintaining vehicle geometry using aspect-ratio-aware resizing. By stabilizing the visual appearance of surveillance images prior to feature extraction, this stage provides a more reliable foundation for downstream embedding learning. For representation learning, PHE-Net combines the strong local inductive biases of ConvNeXt with the global context modeling capability of Swin Transformer blocks. This hybrid design enables the network to jointly capture fine-grained texture details and long-range structural relationships, resulting in a more expressive and discriminative vehicle representation. The model is trained using PK sampling and a joint optimization objective that integrates identity classification loss with triplet loss, encouraging embeddings that are both class-discriminative and retrieval-friendly. Extensive experiments on the VeRi-776 benchmark validate the effectiveness of the proposed framework. PHE-Net achieves 98.20% Rank-1 accuracy and 82.60% mAP, demonstrating that explicitly coupling input enhancement with hybrid CNN–Transformer feature learning leads to more robust and reliable vehicle re-identification under challenging environmental conditions.

Keywords : Vehicle Re-Identification, Hybrid Architecture, ConvNeXt, Swin Transformer, PHE-Net

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

1. Introduction

Vehicle Re-ID is a task whereby images of the same vehicle are retrieved by the different time and non-overlapping cameras by learning an identity-relevant embedding space. It is also popular in smart transportation and smart security system whereby a query vehicle is required to be tracked through huge camera systems and extremely huge galleries [1]. Vehicle Re-ID is challenging due to the fact that vehicles with different identities may appear nearly identical (high inter-class similarity) and the same vehicle may also appear different across cameras (high intra-class variation) due to a change in viewpoint, occlusion, and clutter in the background [2]. In practice, night low contrast, shadows, under/over-exposures, compression artifacts, and motion blur make the challenge worse by eliminating fine-grained cues required to make a discrimination [1]. A number of methods thus extend appearance-based matching with the supplementary context, or spatio-temporal information to regularize the matching of appearance in case appearance itself is not sufficient [3]. Vehicle Re-ID is now dominated by deep learning techniques, typically trained on the goal of classification and metric learning on benchmarks like VeRi-776 [2]. Most recent research studies CNN-Transformer hybrids, the primary objective of which is to represent both locality and long-range dependencies, and has demonstrated high results on vehicle Re-ID tasks [4] [21]. On the backbone level, ConvNeXt offers a more modernized convolutional architecture that has high representational capability [5], whereas Swin Transformer offers effective hierarchical self-attention and multi-scale capabilities [6]. Such developments drive hybrid architectures that mix both contemporary CNN and Transformer blocks that result in powerful vehicle embeddings capable of crossing-camera generalization [21].

Nevertheless, input enhancement is still treated as a secondary step in many vehicle Re-ID pipelines, even though surveillance images often suffer from severe illumination variation, low contrast, and other degradations that can directly reduce feature discriminability and retrieval reliability [1], [4]. Classical enhancement techniques such as CLAHE can improve local contrast without excessively amplifying noise, thereby helping recover fine-grained visual details important for identity recognition [7]. However, the combined effect of explicitly integrating input-quality normalization with hybrid CNN–Transformer representation learning for vehicle Re-ID has not been sufficiently investigated.

Motivated by this gap, we propose the Preprocessing-Enhanced Hybrid Embedding Network (PHE-Net), a vehicle re-identification framework designed for challenging surveillance conditions. The novelty of this work does not lie in introducing a completely new preprocessing operator or a new backbone family in isolation. Rather, its scientific contribution lies in a task-oriented integration strategy that explicitly couples input-quality enhancement with hybrid local-global embedding learning for vehicle re-identification. Specifically, the proposed framework treats preprocessing as a representation-support stage to reduce illumination and contrast degradation before feature extraction, while combining ConvNeXt-based local structural modeling with Swin Transformer-based global context modeling in a unified retrieval-oriented embedding pipeline. In addition, the

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

study provides empirical evidence on the VeRi-776 benchmark that this integrated design improves retrieval effectiveness over single-backbone counterparts. The main contributions of this paper are summarized as follows: (1) we formulate input enhancement as an explicit component of representation learning rather than a generic standalone preprocessing step; (2) we design a hybrid embedding framework that integrates local discriminative cues and global contextual dependencies for vehicle Re-ID; and (3) we validate the effectiveness of this integrated design through comparative and ablation experiments on VeRi-776, where PHE-Net achieves 98.20% Rank-1 accuracy and 82.60% mAP.

The rest of the paper is structured in the following way. Section 2 is a literature review. The proposed method is described in section 3. Experiments and comparisons are provided in section 4. Section 5 is a conclusion of the paper and it talks about the future directions.

2. Related Work

Recent works on Re-ID have begun to focus more on the viewpoint-induced appearance changes, appearance background clutter, and fine-grained inter-class similarity through stronger attention mechanisms, exploiting semantic attributes, and utilizing transformer-style modeling. Lee et al. suggest the Multi-Attention-based Soft Partition (MUSP) network, that learns multiple soft spatial attention mask, in addition to channel-wise attention, and explicitly learns to reduce noisy attention by aggregating insignificant/background areas into a separate map that does not form part of the final representation, thus enhancing discrimination without the need to access metadata annotations [9]. Rong et al. solve background interference and incomplete cues with a superior multi-branch feature fusion system that combines global local feature fusion, channels attention introduced to emphasize identity-relevant channels, and weighted local features to regulate background/noise effects, with notable gains on mainstream benchmarks, including VeRi-776, VRIC and VehicleID [10]. Going outside of controlled conditions, Wang et al. investigate unsupervised vehicle Re-ID and alleviate identity-irrelevant background variation by training both SAM-based masked auto-encoders on pre-training and background-informed meta-learning to learn transferable representations and minimise background interference [11]. Semantically speaking, Tumrani et al. propose the view-sensitive attribute-conditioned architecture that simultaneously utilizes view data and vehicle features (e.g., color/type hint) to produce more consistent embeddings when changing camera views and problematic imaging scenarios [12]. Li et al. also enhance feature discriminability through DMNR-Net, a combination of global feature extraction with non-local relationship capture module (learning saliency in spatial and channel dimensions) and dimensional decoupling strategy that separates the spatial/channel subspaces in order to learn complementary coarse-to-fine cues [13]. Transformer-oriented designs Lian et al. introduce a Transformer-Based Attention Network (TAN) that learns identity-relevant vehicle representations using transformer-based attention modeling (with no keypoint labels) to pay attention to both image-level structure and fine-grained details to perform robust matching[14]. Semantic attributes are explicitly modeled by Sun, HSV color and vehicle category by extracting the information with the

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

help of the multi-attribute dense linking network and a distance control module enlarging the inter-class distance and shrinking intra-class variation is introduced by Sun to increase the reliability of retrieval [15]. Zhu and Sang combine frequency-aware improvement by developing a wavelet feature enhancement (Multi-scale decomposition to retain edges/textures) that combines both global and local cues through differential attention and the use of attribute (color / type) to enhance discrimination across view point and resolution variations [16]. And lastly, Qian et al. propose V2ReID, an outlook-attention network constructed on the Vision Outlooker backbone that learns the finer-level visual cues and has a high performance when an individual only has visual information and no auxiliary side metadata [17].

In summary, prior vehicle Re-ID studies can be broadly grouped into four methodological directions: attention-based and background-suppression approaches, attribute or view-guided methods, transformer-oriented representation models, and enhancement-assisted feature learning strategies. Attention- and fusion-based methods improve discrimination mainly by refining salient local regions or combining complementary feature branches [22], while attribute and view-aware methods rely on semantic side information to reduce viewpoint ambiguity. Transformer-based approaches strengthen global context modeling and long-range dependency learning [20] [23], whereas enhancement-oriented methods attempt to preserve informative texture and edge cues under challenging imaging conditions. Despite these advances, most existing studies emphasize either architectural design or auxiliary semantic modeling, while giving comparatively less attention to explicitly integrating input-quality normalization with hybrid embedding learning in a unified retrieval framework. In contrast, the proposed PHE-Net combines lightweight preprocessing for illumination and contrast stabilization with a hybrid ConvNeXt–Swin feature extractor, thereby linking input enhancement and local-global representation learning within a single vehicle Re-ID pipeline. This distinction clarifies the methodological contribution of the present work relative to prior studies.

3.The Proposed Method

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

In this section, we introduce PHE-Net, which describes its general structure, main parts, and training plan. The approach is aimed at learning robust embeddings that can be identity-relevant in the presence of difficult real-life scenarios including viewpoint or background clutter, as shown in Fig. 1.

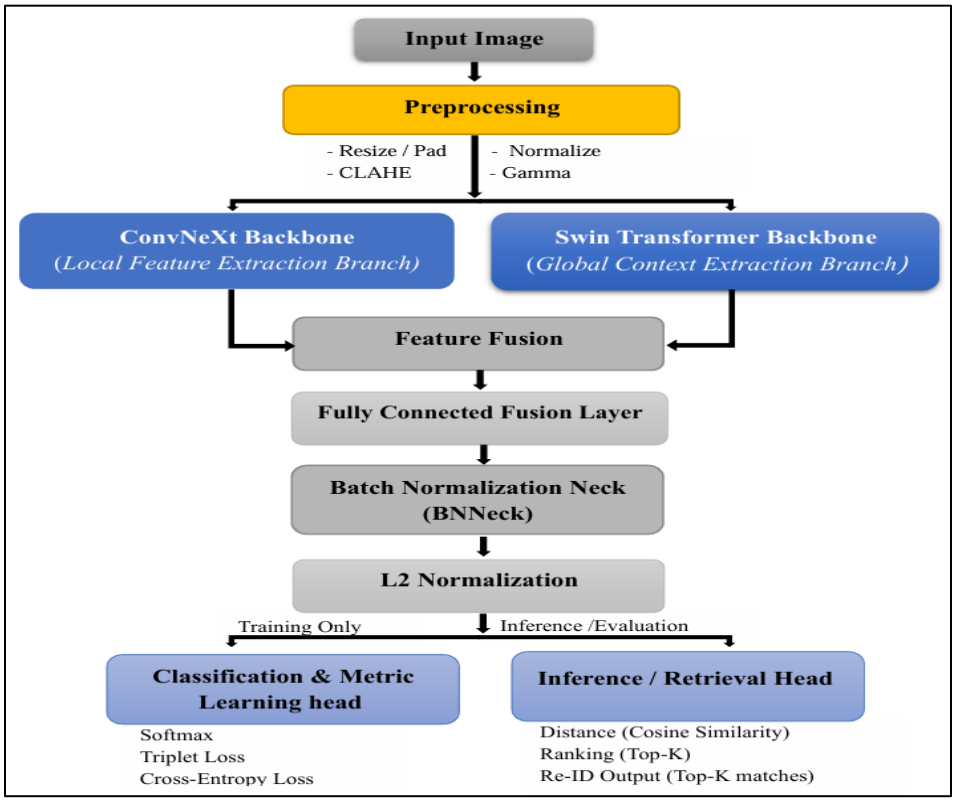

Fig. 1: Architectural pipeline of PHE-Net.

3.1. Framework Overview

We present (PHE-Net) an end-to-end vehicle Re-ID architecture that acquires robust and identity-sensitive embeddings in conditions of real-world appearance changes. The framework involves three closely-linked components: (i) it contains a preprocessing step to diminish illumination and geometric inconsistencies, (ii) a hybrid CNN-Transformer backbone to capture jointly, fine-grained, local appearance cues, and global contextual relationships, and (iii) it contains a training strategy that is based on metric-learning to explicitly organize the embedding space to conduct a retrieval. The design specifically addresses the major issues of vehicle Re-ID, namely, viewpoint variation, background clutter, and high inter-class similarity, and allows Top-K matching to be performed reliably in large-scale galleries.

3.1.1 Datasets

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

PHE-Net is tested on VeRi-776 where large-scale vehicle re-identification is performed on a large-scale dataset of images gathered in real-world urban surveillance settings and first presented by Liu et al. [18]. The data has 51,035 images of 776 vehicle identities which were captured using 20 non-overlapping cameras and there are high differences in point of view, light, background clutter and image resolution. Training is done to 576 identities (37,778 images), and the rest of the identities (200) are considered as a test set, which is further grouped into a query set of 1,678 images and a gallery of 11,579 images. The overall statistics and divisions are summarized in Table 1. The data can be found on the website [19].

Table 1: VeRi-776 Dataset.

Item	Value
Cameras	20
Total identities	776
Total images	51,035
Training set	576 IDs / 37,778 images
Test set (gallery)	200 IDs / 11,579 images
Query set	1,678 images (from the test IDs)

3.1.2 Preprocessing Module

The images taken of vehicles under the surveillance of multi-accounts of cameras are usually affected by uneven light distribution, insufficient contrast, and scale alterations arising due to different perspectives. In order to reduce these effects before feature extraction, we use a small but efficient preprocessing module. CLAHE is first applied to improve local contrast, but it does not exaggerate noise. This is succeeded by gamma correction so that the world brightness variations in different scenes are normalized.

In order to retain shape-sensitive information, we use a padding-sensitive resizing methodology that does not change the aspect ratio, but instead pads the shorter dimension of the image with a constant amount of padding (with a neutral gray value). The method avoids distorting the geometry of the vehicle and has minimal sharp transitions of intensity at the edges of the vehicle as compared to standard zero-padding. Padded image is then brought to geometric consistency at fixed scale of 224×224 which makes sure that the image is consistent and can be viewed at varied angles, and still works with other standard backbone architectures as shown in Fig. 2.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

Fig. 2: Comparison between original, preprocessed, and augmented vehicle samples.

3.1.3 Hybrid Feature Embedding Model

In order to address the shortcoming of either that of convolutional or transformer-based representations, we use a hybrid feature embedding model that capitalizes on its advantages. Convolutional backbone A Convolutional backbone is created in the style of ConvNeXt to capture fine-grained local structures like textures, contours and subtle structural details that play a critical role in vehicle identity discrimination. Simultaneously, a Swin Transformer-based branch model captures the global context and long-range dependencies of the model using window-based self-attention, which is resistant to changes in perspective and variability in the background. The two branch outputs are combined to create a single representation which collectively represents both local fine-grained cues and global contextual information leading to richer feature encoding.

3.1.4 Feature Aggregation and Embedding Generation

We obtain the fused representation by combining the outputs of the two backbone branches, followed by Generalized Mean (GeM) pooling to generate a compact retrieval descriptor [24]. In the proposed feature design, the two branches provide complementary representations of the same vehicle image. The ConvNeXt

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

branch emphasizes fine-grained local patterns such as texture, contour, and subtle appearance details, while the Swin Transformer branch captures broader structural layout and long-range contextual dependencies. Their fused representation is therefore intended to preserve both discriminative local cues and global relational information. GeM pooling is applied to retain salient activations more effectively than standard average pooling while producing a compact descriptor. The pooled vector is then passed through a BNNeck layer to stabilize feature statistics during training. Finally, L2 normalization is applied to obtain the final embedding vector z, which lies in a unit-length space suitable for cosine-similarity-based retrieval.

3.1.5 Training Strategy

A retrieval-oriented embedding space cannot be learned effectively using only a conventional supervised classification objective. Therefore, we adopt a training strategy that combines efficient mini-batch construction with joint optimization. PK sampling is used to construct each mini-batch by randomly selecting P = 12 vehicle identities and K = 6 images per identity. This strategy increases the presence of informative positive and negative samples within each batch, which is important for metric learning. The network is optimized using a combined objective consisting of identity classification loss (cross-entropy) and metric learning loss (triplet loss). The classification loss encourages inter-class separability by learning class-discriminative decision boundaries, whereas the triplet loss directly shapes the embedding space by pulling samples of the same identity closer and pushing samples of different identities farther apart with a margin constraint. This combination enables the model to benefit from both classification-oriented supervision and retrieval-oriented metric learning, thereby improving optimization effectiveness for vehicle re-identification.

3.1.6 Inference and Matching

In the inference, the query and gallery images go through a similar preprocessing and backbone pipeline to generate single embedding vectors. Embedding is L2-normalized and cosine similarity is used to compute similarity between query and gallery features. Images on the gallery are then ranked based on their similarity scores and the Top-K best similar samples are provided as the final vehicle Re-ID results.

4. Experiments and Results

It is a section that test-drives the effectiveness of PHE-Net by conducting large-scale experiments on a public benchmark. We compare the quantitative performance, ablation results, and qualitative retrieval behavior to illustrate the effect of the design component.

4.1 Evaluation Protocol

In order to achieve a just and thorough evaluation of the proposed car re-identification system, a set of usual evaluation metrics of retrieval is used. These measures are commonly used in the literature of person and vehicle

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

re-identification, as they combine the top-match accuracy, ranking robustness, and general retrieval quality. The evaluation metrics applied in this research and its practical interpretation are summarized in Table 2.

Table 2: Evaluation Metrics Used in Vehicle Re-Identification

Metric	What it reports	How to interpret higher values
Rank-1 (CMC)	Percentage of queries where the correct ID appears at rank 1 in the gallery	Better top-match retrieval
Rank-5 (CMC)	Percentage of queries where the correct ID appears within the top 5 results	Better top-K reliability
mAP	Mean Average Precision over all queries (accounts for all true matches and their ranking positions)	Better overall ranking quality, not just the first hit

Rank-1 accuracy is a result of the Cumulative Matching Characteristic (CMC) curve and is used to measure how well the system can correctly identify the first hit of a vehicle identity that is important in real-time or fully automated surveillance systems. Rank-5 accuracy generalizes this view to evaluate the presence of the correct identity in the top 5 retrieved identities that offer a more tolerant perspective that adjusts to semi-automated inspection or human-in-the-loop conditions. Mean Average Precision (mAP) is a more comprehensive measure that uses all the correct matches of a particular query and the relative position of the match in the gallery. In contrast to Rank-k metrics, mAP also does not reward consistent quality in ranking when incorrect ordering is present and punishes uniform quality in ranking throughout the result set. Therefore, mAP has found special application especially where vehicle re-identification activities are required and several pictures of the same vehicle might exist under different view-points, light conditions and occlusions. These metrics, in combination, offer a balanced assessment tool that considers both instant retrieval performance and global ranking functionality, so that the results reported can be considered to be the true deployment conditions in the real-world.

4.2. Implementation

The experiments were conducted on Google Colab using an NVIDIA Tesla T4 GPU with 16 GB VRAM. The proposed preprocessing module, consisting of CLAHE, gamma correction, and padding-aware resizing, was applied to the input images before feature extraction. Padding-aware resizing was used to adapt all vehicle images to a fixed input resolution of 224 × 224 while preserving the original aspect ratio. During training, PK sampling was adopted with P = 12 identities per batch and K = 6 images per identity, resulting in a batch size of 72. This sampling strategy ensures that each mini-batch contains informative positive and negative examples for metric learning. For clarity and reproducibility, the main implementation and training settings used in the reported experiments are summarized in Table 3.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

Table 3: Implementation and Training Settings of PHE-Net

Item	**Reported setting**
Hardware	Google Colab, NVIDIA Tesla T4 (16 GB VRAM)
Input size	224 × 224
Batch construction	PK sampling, P = 12 and K = 6 (batch size = 72)
Losses	Cross-entropy + Triplet loss
Inference similarity	Cosine similarity on L2-normalized embeddings
Optimizer	AdamW
Initial learning rate	0.0003
Number of epochs	100
Triplet margin α	0.3
Embedding dimension	512
Pretrained initialization	ImageNet-pretrained weights

The settings listed in Table 3 define the optimization and embedding configuration used throughout the reported experiments. Both backbone branches were initialized with ImageNet-pretrained weights prior to task-specific fine-tuning. The network was trained using a joint objective that combines identity classification loss and metric learning loss:

$$L_{total} = \lambda_{id} L_{id} + \lambda_{tri} L_{tri} \quad \text{..... (1)}$$

where L_{id} denotes the cross-entropy identity loss and L_{tri} denotes the triplet loss. The triplet loss was computed using the learned embeddings z as follows:

$$L_{tri} = \sum_{i=1}^{N} max\left(0, d\left(z_i^a, z_i^p\right) - d(z_i^a, z_i^n) + \alpha\right) \quad \text{..... (2)}$$

where z_i^a, z_i^p and z_i^n denote the anchor, positive, and negative embeddings, respectively; α is the margin; d (.,.) is a distance function.

At test time, each image produces a single embedding vector. We apply L2 normalization and perform retrieval using cosine similarity, ranking gallery embeddings for each query.

4.3. Ablation Study

To more rigorously quantify the contribution of the proposed design, we conduct two complementary ablation analyses on the VeRi-776 dataset. First, we isolate the effect of the preprocessing module by evaluating the same hybrid backbone under different preprocessing settings, including no preprocessing, CLAHE only, gamma correction only, their combined use, and the full proposed preprocessing pipeline. Second, we examine the

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

contribution of the backbone design itself by comparing single-branch and hybrid feature learning configurations. This two-level analysis allows the effect of preprocessing to be distinguished from the effect of architectural strength and feature fusion.

Table 4: Isolated Effect of Preprocessing Components on VeRi-776

Model Setting	CLAHE	Gamma Correction	Resizing Strategy	mAP (%)	Rank-1 (%)
PHE-Net without preprocessing	✗	✗	Standard resize	77.50%	95.80%
PHE-Net + CLAHE only	✓	✗	Standard resize	79.80%	97.10%
PHE-Net + Gamma correction only	✗	✓	Standard resize	78.90%	96.50%
PHE-Net + CLAHE + Gamma correction	✓	✓	Standard resize	81.10%	97.60%
PHE-Net + Full preprocessing module (proposed)	✓	✓	Padding-aware resize	82.60%	98.20%

Table 4 clearly isolates the contribution of the preprocessing module while keeping the hybrid backbone unchanged. Compared with the no-preprocessing baseline, applying CLAHE alone improves performance from 95.80% to 97.10% in Rank-1 and from 77.50% to 79.80% in mAP, indicating that local contrast enhancement contributes positively to discriminative feature learning. Gamma correction alone also improves performance to 96.50% Rank-1 and 78.90% mAP, confirming that brightness normalization provides an additional benefit, although its effect is slightly smaller than that of CLAHE in this setting. When CLAHE and gamma correction are combined, performance further increases to 97.60% Rank-1 and 81.10% mAP, suggesting a complementary effect between local contrast enhancement and global brightness adjustment. The best results are obtained with the full proposed preprocessing pipeline, which additionally includes padding-aware resizing, reaching 98.20% Rank-1 and 82.60% mAP. These results confirm that the observed performance gain cannot be attributed to backbone strength alone, but is also significantly supported by the proposed preprocessing design. For clearer visual interpretation, the preprocessing ablation results are also illustrated in Fig. 3.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

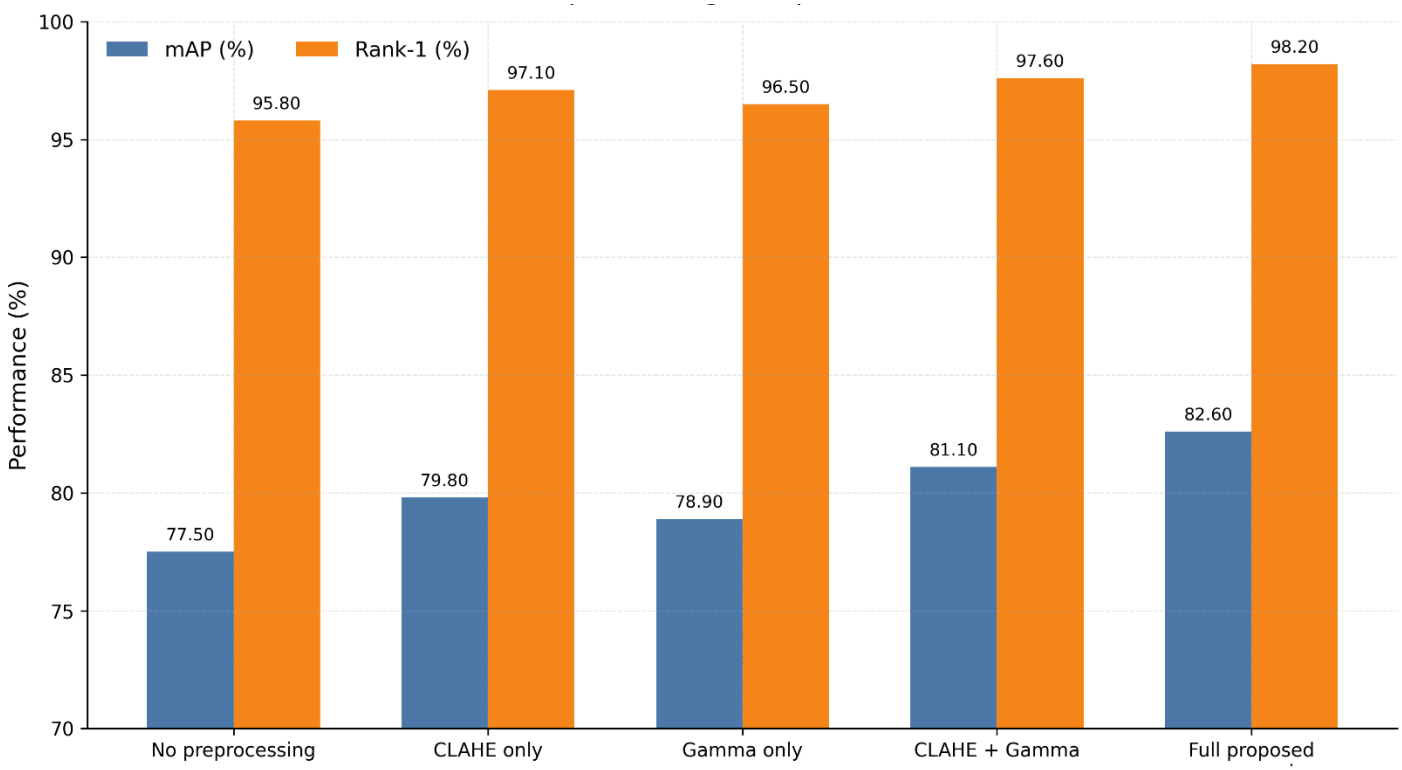

Fig. 3. Effect of preprocessing components on VeRi-776 in terms of mAP and Rank-1.

As shown in Fig. 3, performance improves progressively as preprocessing components are added. CLAHE and gamma correction each provide measurable gains over the no-preprocessing baseline, while their combination yields a stronger improvement. The best performance is achieved when padding-aware resizing is added to the combined enhancement pipeline, confirming the effectiveness of the full proposed preprocessing module.

The backbone-level ablation results are summarized in Table 5, where the contribution of single-branch and hybrid feature learning is analyzed separately from the preprocessing component study.

Table 5: Backbone-Level Ablation on VeRi-776

Model variant	Hybrid backbone	mAP (%)	Rank-1 (%)
ConvNeXt	✗	76.85	95.10
Swin Transformer	✗	79.40	97.80

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

Model variant	Hybrid backbone	mAP (%)	Rank-1 (%)
PHE-Net	✓	82.60	98.20

Table 5 presents the backbone-level ablation under the same full preprocessing setting. Using only the ConvNeXt branch yields 95.10% Rank-1 and 76.85% mAP, showing that convolution-based local feature extraction provides a solid baseline for vehicle discrimination. Replacing this with a Swin Transformer backbone increases performance to 97.80% Rank-1 and 79.40% mAP, highlighting the benefit of global context modeling and long-range dependency learning. The full PHE-Net configuration, which combines both branches, achieves the best overall performance with 98.20% Rank-1 and 82.60% mAP. This confirms that the hybrid local-global design provides complementary gains beyond those obtained by either single-branch backbone alone. For clearer visual comparison, the backbone-level ablation results are also illustrated in Fig. 4.

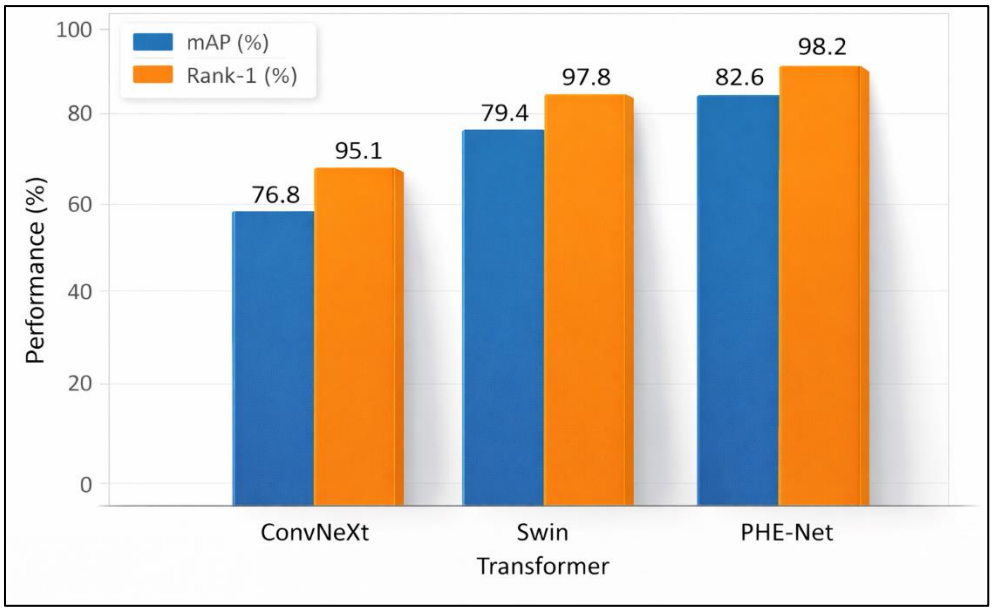

Fig. 4. Backbone-level ablation results on VeRi-776 in terms of mAP and Rank-1.

As shown in Fig. 4, the performance improves consistently from ConvNeXt to Swin Transformer and reaches its highest values with the full PHE-Net configuration. This trend confirms that combining convolution-based local feature extraction with transformer-based global context modeling provides complementary advantages over using either backbone individually.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

4.4. Comparison with Recent Methods

We benchmark PHE-Net against a set of representative current vehicle Re-ID methods on the VeRi-776 benchmark on the basis of conventional metrics of retrieval (mAP and CMC Rank-k). On the whole, the findings indicate that the integration of input improvement and hybrid CNN-Transformer feature learning can give credible retrieval results in viewpoint variability and background clutter.

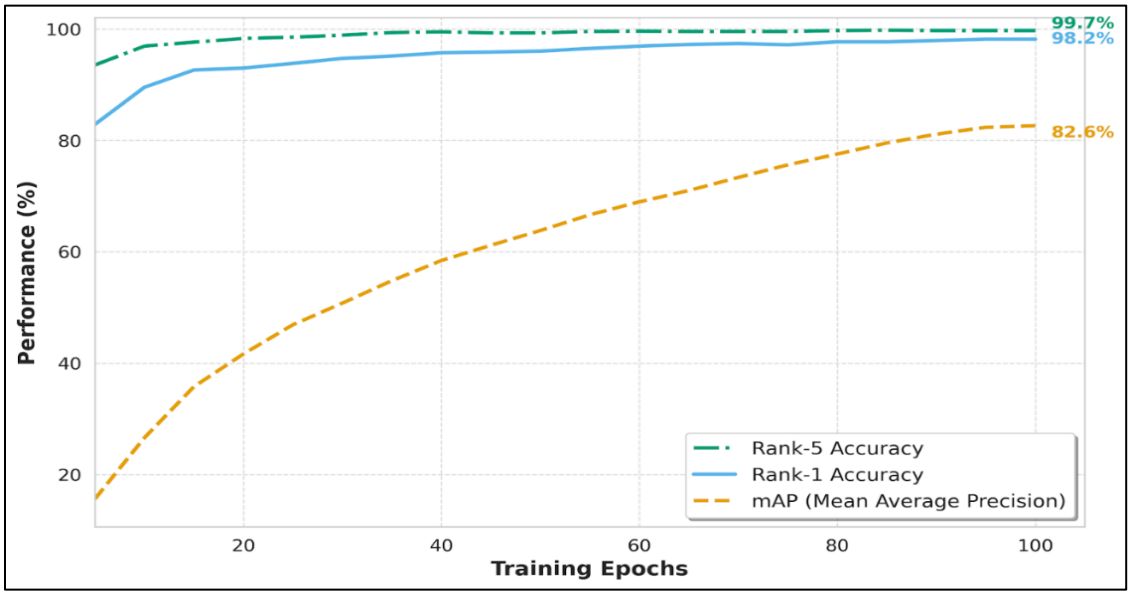

Fig. 5**:** mAP and CMC (Rank-1/Rank-5) trends on VeRi-776.

According to Fig. 5, Rank-1 and Rank-5 accuracies grow fast in the early epochs and hence stabilize, meaning that the model swiftly acquires fixed identity-relevant cues that facilitate steady Top-k matching. By contrast, mAP gets better over time during training, just as a result of overall global ranking quality refinement. This is a normal phenomenon of retrieval-optimization, where the mAP of all correct matches (mAP) can be increased only through long-term learning to achieve a more pronounced separation of hard negatives and narrowing of intra-identity cluster of features. Table 6 provides a comparative discussion of some popular vehicle re-identification techniques comparing their results on the VeRi-776 dataset, as well as recently published works and those published relatively close to the one. The comparison reveals the differences in architectural design, feature learning strategies, and reported performance, which gives a definite perspective of placing the proposed approach in relation to the current work.

Table 6: Comparative Analysis of Vehicle Re-Identification Methods on VeRi-776

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

Ref.	Year	Method	Core Idea	Backbone Family	Hybrid CNN–Transformer	Attribute / View Modeling	Explicit Image or Background Enhancement	mAP (%)	Rank-1 (%)
[11]	2024	SAM-MAE + Meta	SAM-based masking with MAE pretraining and background-aware meta-learning	Transformer (MAE / ViT)	No	No	Yes	38.40%	86.80%
[12]	2022	VAAG	View-aware and attribute-guided feature learning	CNN (ResNet-50)	No	Yes	No	63.01%	92.20%
[15]	2024	MADLN + DC	Dense multi-attribute linking with distance control	CNN (DenseNet-style, multi-branch)	No	Yes	No	68.83%	92.94%
[10]	2021	MBFF (Improved)	Multi-branch global–local fusion with channel attention	CNN (ResNet-50)	No	No	No	77.12%	96.30%
[9]	2023	MUSP	Multi-attention soft partition (spatial and channel attention)	CNN	No	No	No	78.00%	95.60%
[16]	2025	Wavelet-FE + GLDAF	Frequency-domain enhancement with global–local attention fusion	Transformer (Swin)	No	Yes	Yes	79.20%	97.00%
[17]	2022	V²ReID	Transformer-based outlook attention modeling	Transformer (VOLO)	No	No	No	80.31%	97.13%
[14]	2022	TAN	CNN backbone with transformer-based attention branch	ResNet-50 + Transformer	Yes	Yes	No	80.50%	95.40%

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

[13]	2024	DMNR-Net	Dimensional decoupling with non-local relational modeling	CNN (ResNet-50, three-branch)	No	No	No	82.00%	95.80%
(Ours)	2026	**PHE-Net**	Preprocessing-aware hybrid embedding with local–global fusion	ConvNeXt + Swin	**Yes**	No	**Yes**	**82.60%**	**98.20%**

A systematic comparison of recent and recent vehicle re-identification approaches to the VeRi-776 dataset is provided in Table 4, mapping the performance reported in Rank-1 and mAP to the major architectural and methodological decisions of the approaches. Specifically, the table emphasizes a method having a hybrid CNN Transformer architecture, using attribute or view based semantic cues or including an explicit image/background enhancement phase to enhance the quality of input under challenging surveillance conditions. A number of trends are evident as a result of this comparison. Approaches based on attribute or view supervision, like those in [12] and [15], have a higher stability in matching when compared to more perspective-independent approaches, but have significantly lower mAP scores. This implies that semantic cues are useful in the process of aligning coarse visual properties, but in general are not effective enough to resolve finer-grained ambiguities when ranking large and visually similar galleries.

Conversely, strategies with more focus on the representation learning with a stronger emphasis on the attention mechanisms, feature fusion, or relational modeling, such as [9], [10], and [13] have higher mAP always. These gains suggest better overall ranking quality than the better scoring of the top retrieved match. Transformer-based designs like [14] and [17] further improve the context modeling of the globe as well as provide competitive performance that supports the importance of long-range dependency modeling in vehicle re-identification. Likewise, frequency-sensitive refinement techniques, which are studied in [16] can indicate that texture and edge preservation can be a profitable idea, but these techniques do not always provide the best mAP among approaches. Combined, the table shows a sensible gap in the literature. The majority of current methods focus on backbone architecture and feature-learning methods, whereas comparatively few directly consider the issue of input degradation due to changes in illumination, contrast loss and background clutter- aspects which are characteristic of real-world surveillance data. The given limitation is directly addressed in PHE-Net where preprocessing is the central design factor and it is paired with a hybrid ConvNeXtSwin embedding framework. The combination gives the best overall performance that is recorded in the table with a Score of 98.20% Rank-1 and 82.60% mAP. Lastly, even the unsupervised framework in [11] is significantly inferior to supervised ones, which highlights the fact that empowering robustness and decreasing the use of labeled data is an open and difficult avenue of research in the future of vehicle re-identification.

5. Conclusion and Future Work

This paper introduced PHE-Net, an end-to-end vehicle re-identification framework specifically designed to maintain reliable performance under challenging real-world surveillance conditions, including illumination variations, low contrast, background clutter, and viewpoint changes. The proposed framework integrates three complementary components. First, a lightweight preprocessing module using CLAHE, gamma correction, and padding-aware resizing stabilizes visual appearance and preserves vehicle geometry prior to feature extraction. Second, a hybrid CNN–Transformer embedding backbone jointly exploits fine-grained texture and shape cues alongside global contextual structure, enabling more expressive and discriminative representations. Third, a metric-learning–oriented training strategy encourages identity-relevant embeddings that are well suited for retrieval tasks. Comprehensive experiments on the VeRi-776 benchmark demonstrate the effectiveness of the proposed design, with PHE-Net achieving 98.20% Rank-1 accuracy and 82.60% mAP, reflecting both strong top-hit retrieval and improved overall ranking quality across large galleries. Ablation studies further verify that each component contributes meaningfully to the final performance and that combining preprocessing with hybrid feature learning yields complementary gains beyond those achievable with a single backbone architecture. In summary, these findings support the central conclusion that explicitly addressing input quality, together with hybrid representation learning, provides a practical and effective direction for robust vehicle re-identification in large-scale camera networks.

Looking forward, future work will extend PHE-Net along three main directions. First, explainable re-identification mechanisms will be incorporated to improve the interpretability of matching decisions. Second, robustness under domain shifts will be strengthened through lightweight masking or segmentation strategies and enhanced cross-camera generalization, particularly in low-light scenarios. Third, efficiency-oriented optimizations will be explored to accelerate embedding extraction and large-gallery retrieval, facilitating real-time deployment in operational surveillance systems.

6. References

[1] S. D. Khan and H. Ullah, “A survey of advances in vision-based vehicle re-identification,” *Comput. Vis. Image Underst.*, vol. 182, pp. 50–63, 2019, doi: 10.1016/j.cviu.2019.03.001.

[2] X. Liu, W. Liu, T. Mei, and H. Ma, “A deep learning-based approach to progressive vehicle re-identification for urban surveillance,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 869–884. doi: 10.1007/978-3-319-46475-6_53.

[3] W. Shen, “Learning Deep Neural Networks for Vehicle Re-ID With Visual-Spatio-Temporal Path Proposals,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1904–1912.

[4] Y. Wang, R. Li, and Y. Shao, “Vehicle Re-Identification Method Based on Efficient Self-Attention CNN–Transformer and Multi-Task Learning Optimization,” *Sensors*, vol. 25, no. 10, 2025.

[5] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11976–11986.

[6] Z. Liu *et al.*, “Swin Transformer: Hierarchical vision Transformer using shifted windows,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022. doi: 10.1109/ICCV48922.2021.00986.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

[7] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed., San Diego, CA, USA: Academic Press, 1994, pp. 474–485. doi: 10.5555/180895.180940.

[8] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017. [Online]. Available: https://arxiv.org/abs/1703.07737

[9] S. Lee, T. Woo, and S. H. Lee, "Multi-attention-based soft partition network for vehicle re-identification," *J. Comput. Des. Eng.*, vol. 10, no. 2, pp. 488–502, 2023.

[10] L. Rong, "A Vehicle Re-Identification Framework Based on the Improved Multi-Branch Feature Fusion Network," *Sci. Rep.*, vol. 11, p. 20210, 2021.

[11] D. Wang *et al.*, "SAM-driven MAE pre-training and background-aware meta-learning for unsupervised vehicle re-identification," *Comput. Vis. Media*, vol. 10, no. 4, pp. 771–789, 2024, doi: 10.1007/s41095-024-0424-2.

[12] S. Tumrani, W. Ali, R. Kumar, A. A. Khan, and F. A. Dharejo, "View-aware attribute guided network for vehicle re-identification," *Multimed. Syst.*, vol. 29, pp. 1853–1863, 2022.

[13] X. Li, X. P. Id, and Q. Meng, "Vehicle re-identification based on dimensional decoupling strategy and non-local relations," pp. 1–18, 2024, doi: 10.1371/journal.pone.0291047.

[14] J. Lian, D. Wang, S. Zhu, Y. Wu, and C. Li, "Transformer-Based Attention Network for Vehicle Re-Identification," pp. 1–17, 2022.

[15] X. Sun *et al.*, "Vehicle re-identification method based on multi-attribute dense linking network combined with distance control module," *Front. Neurorobot.*, vol. 17, p. 1294211, 2024.

[16] B. Zhu and H. Sang, "Vehicle Re-Identification Based on Wavelet Feature Enhancement and Global-Local Differential Attention Fusion," *J. Comput. Sci. Artif. Intell.*, vol. 2, no. 1, pp. 53–60, 2025.

[17] Y. Qian, J. Barthelemy, U. Iqbal, and P. Perez, "V2ReID: Vision-Outlooker-Based Vehicle Re-Identification," *Sensors*, vol. 22, no. 22, p. 8651, 2022.

[18] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *IEEE International Conference on Multimedia & Expo (ICME)*, 2016, pp. 1–6. doi: 10.1109/ICME.2016.7553002.

[19] "VeRi Vehicle Re-identification Dataset," 2026. [Online]. Available: https://www.kaggle.com/datasets/abhyudaya12/veri-vehicle-re-identification-dataset

[20] A. Dosovitskiy *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[21] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 22–31.

[22] Y. Guo, G. Yuan, W. Li, and H. Li, "DAFF-Net: A Dual-Branch Attention-Guided Feature Fusion Network for Vehicle Re-Identification," Algorithms, vol. 18, no. 11, Art. no. 690, 2025, doi: 10.3390/a18110690.

[23] M. Zhu and Q. Feng, "Transformer-based vehicle re-identification with view information," Scientific Reports, vol. 15, Art. no. 40576, 2025, doi: 10.1038/s41598-025-24392-y.

[24] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, Jul.2019, doi: 10.1109/TPAMI.2018.2846566