



VhAR-Net: A Cross-Modal Representation Learning Framework for Text-Based Vehicle Retrieval

Lin Qian¹ and Jian Qu²

ABSTRACT

With the advancement of intelligent transportation systems and large-scale urban video surveillance technologies, vehicle image retrieval based on textual descriptions has become increasingly important. Although MCANet exhibits effectiveness in multi-scale feature alignment, significant limitations persist in accurate fine-grained semantic matching. To address this, we present VhAR-Net—a modular cross-modal retrieval architecture that enables independent design and flexible combination of feature interaction mechanisms, enhancement strategies, and supervisory signals. This framework employs ResNet-50 as the visual representation extractor and BERT as the textual semantic encoder, and innovatively introduces a cross-modal attention unit that establishes explicit associations between image representations and linguistic descriptions. Concurrently, the system establishes small-object aware auxiliary supervision through binary classification tasks targeting discriminative fine-grained vocabulary, thereby directing the network toward distinctive microscopic semantic units. Empirical evaluations on the public benchmark dataset T2I-VeRi indicate that the optimal configuration achieves 75% Top-1 retrieval accuracy, with Top-10 recall covering 85% of relevant samples. After introducing the feature enhancement module and small-object supervision mechanism, the cumulative matching rate for Top-5 and Top-10 both reached 85%, demonstrating improved retrieval robustness.

Article information:

Keywords: Text-to-image Retrieval, Cross-modal Attention, Small-object Supervision, Multi-scale Features, Contrastive Learning

Article history:

Received: November 25, 2025

Revised: January 30, 2026

Accepted: March 19, 2026

Published: April 4, 2026

(Online)

DOI: 10.37936/ecti-cit.2026202.264970

1. INTRODUCTION

With the rapid development of intelligent transportation and the Internet of Vehicles (IoV), urban traffic management systems are gradually moving into an era of comprehensive sensing and intelligent decision-making. Roadside cameras, in-vehicle sensors, and monitoring terminals continuously capture and store massive volumes of vehicle image and video data daily, as illustrated in Figure 1. These data provide crucial support for traffic regulation, vehicle tracking, and road safety. However, these data also present new challenges: quickly and accurately locating target vehicles within this massive image volume has become a pressing issue requiring immediate attention.

In traditional vehicle re-identification (Vehicle Re-ID), both query and candidate databases comprise images, aiming to locate the same vehicle identity

across multiple cameras, varied viewpoints, and diverse lighting conditions. However, in real-world traffic applications, query images are frequently unavailable, with only natural language descriptions—such as ‘white SUV, sticker on the left rear door, scratch on the rear bumper’—accessible. To bridge this gap, cross-modal vehicle retrieval has emerged in recent years, utilizing text queries to search for target vehicles within large-scale image databases.

In Thailand, vehicle retrieval based on license plates is limited both technically and legally. Thai license plates contain Thai characters and numbers, with province names at the bottom, and come in various styles, including special designs and different background colors. Additionally, the high number of motorcycles, small license plates, and frequent obstructions further complicate readability. New vehicles with temporary ‘red plates,’ reduced nighttime

¹The author is with the MET, Faculty of Engineering and Technology, PIM, Thailand, Email: 6772100069@stu.pim.ac.th

²The author is with the Faculty of Engineering and Technology, PIM, Thailand, Email: Jianqu@pim.ac.th

²Corresponding author: Jianqu@pim.ac.th

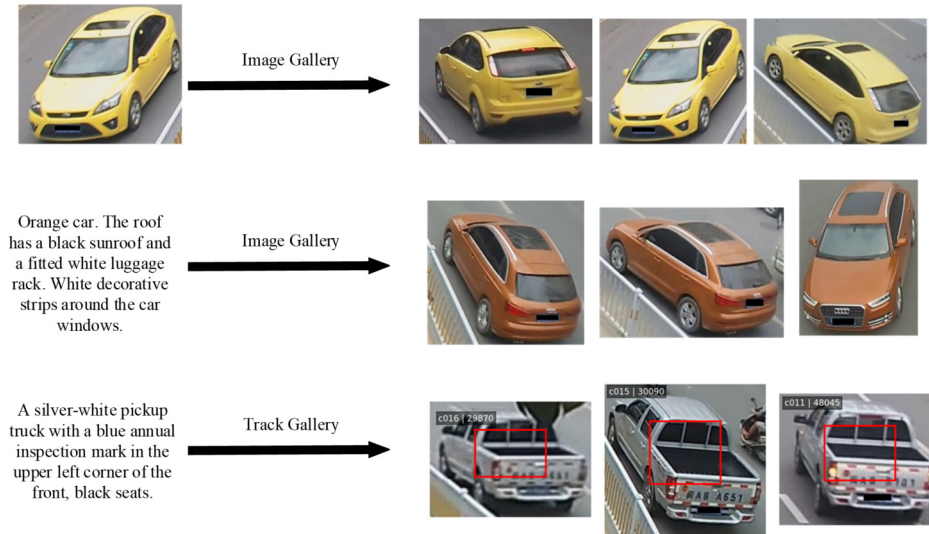


Fig.1: Different vehicle retrieval methods.

reflectivity, seasonal dirt accumulation during rainy seasons, and intentional concealment further diminish plate legibility. More critically, license plate numbers constitute personal and vehicle-identifying information protected under the Personal Data Protection Act (PDPA). Consequently, long-term storage and retrieval based on license plate numbers demand stringent purpose limitations and security controls. Furthermore, the prevalence of ‘license plate swapping, replacement, or counterfeiting’ undermines the one-to-one relationship between license plates and vehicles, resulting in identity ambiguities such as ‘one plate, multiple vehicles’ or ‘one vehicle, multiple plates.’ Consequently, relying exclusively on license plate-based retrieval proves inadequate for mainstream applications and entails significant risks and costs.

In contrast, cross-modal retrieval based on text-to-image conversion is better suited for real-world applications. Even in the absence of visible license plates, effective retrieval can be accomplished using semantic features such as color, body shape, brand, model, modifications, stickers, paint, local damage, and spatiotemporal cues, facilitating more precise search results. The emergence of cross-modal retrieval technology provides a new solution to this problem. By learning a shared semantic space between text and images, this technology enables natural language descriptions to be converted into vector representations comparable with image features, thereby facilitating efficient text-based vehicle retrieval. This approach synergizes natural language processing and computer vision, allowing the system not only to ‘understand’ images but also to ‘comprehend’ human language.

Ali Amiri, Aydin Kaya, and Ali Seydi Keçeli [1] observed that appearance differences among vehicles of identical brand, color, and model are minimal, rendering reliable distinction based solely on

global appearance features challenging. Relying on attributes (such as year, wheels, stickers) or local structure modeling remains insufficient, which continues to be a long-standing bottleneck for retrieval. Wang Zhan, Huang Shucheng, and Li Mingxing[2] proposed MSMV-Net, which aligns cross-view information of the same vehicle by using multi-scale local features and a four-branch structure, while introducing multi-view related constraints to improve vehicle Re-ID accuracy. However, the model is relatively complex and strongly dependent on viewpoint estimation.

From a theoretical perspective, cross-modal vehicle retrieval from text to image presents a novel approach to visual-language fusion. This methodology transcends the limitations of traditional single-vision features by incorporating semantic-level information modeling, effectively shifting image retrieval from ‘visual matching’ to ‘semantic understanding.’ Consequently, this not only enriches the research framework for vehicle re-identification and image retrieval, but also advances the development of multimodal learning, feature alignment, and semantic mapping.

From a practical perspective, vehicle text retrieval technology has broad application potential. For law enforcement, traffic police, and transportation management departments, when only textual clues are available in a case, the system can quickly locate the target vehicle based on the description, greatly reducing investigation time. In fields such as smart parking, logistics monitoring, vehicle rental, and accident tracking, it lowers the barrier for human-machine interaction, allowing ordinary users to perform complex retrieval tasks without specialized knowledge.

The contributions of this paper are as follows. For the natural language-based vehicle retrieval task, we propose an overall model framework, VhAR-Net, which is built on a dual-branch multi-scale repre-

sensation and cross-modal alignment design. The image branch uses ResNet-50 to extract global and stripe-level local features, while the text branch encodes descriptions with BERT and uses a lightweight projection head to align the features, enabling fine-grained cross-modal feature alignment and fusion. At the same time, we design a $\lambda CTIO$ loss function, which is composed of CMPM, Triplet, InfoNCE, and BCE(obj_flag), which improves the robustness and generalization ability of the model. Under a consistent preprocessing pipeline, our method is compared with three typical approaches published in the last few years—MACNet for vehicle retrieval, and TIPCB and TIDCB for pedestrian retrieval—and we conduct reproducible evaluations on the public T2I-Veri dataset, reporting Rank@1, Rank@5, Rank@10, and mAP.

2. RELATED WORKS

2.1 Research Status of Vehicle Retrieval Technology

The initial research on cross-modal retrieval primarily focused on pedestrian detection, where natural language descriptions were matched with human features in images to enable text-based pedestrian retrieval. However, as research advanced, this approach was progressively extended to more complex target domains, including vehicles, thereby accelerating the transition from ‘person retrieval’ to ‘scene retrieval.’ This development has established novel technological pathways for intelligent security and traffic management, as documented by Zheng and Qu [3]. Since the late 20th century, coinciding with the advent of intelligent transportation systems, vehicle retrieval research has gained prominence. Historically, vehicle retrieval technology has progressed through two stages: attribute-based vehicle retrieval and content-based vehicle retrieval.

Distinguishing vehicles based on attribute information is another approach to vehicle retrieval. This method transforms the vehicle retrieval problem into an attribute recognition and multi-attribute label matching problem. The main attributes include color, model, and brand logo. In terms of vehicle color recognition, the two primary approaches are color threshold segmentation and feature-based description combined with pattern recognition. For vehicle model recognition, there are two main types of methods: one is traditional hardware-based methods, such as installing sensors underground. The drawback of this approach is its high hardware cost, inconvenience for installation and maintenance, and limited applicability. The other type is computer vision-based methods [4], which use either manually designed or machine-learned vehicle features to train classifiers for model classification. Commonly used features include vehicle outline features, local features (such as SIFT, HOG, Gabor, etc.), and deep features. Viewpoint in-

variance is a major challenge in model recognition. To address this, in recent years, some methods based on 3D vehicle models [5][6] have been proposed. These methods can somewhat overcome the impact of vehicle angle on classification accuracy, but extracting 3D vehicle models requires significant computational resources.

Vehicle retrieval based on license plates and attribute recognition techniques is similar to TBIR (Text-Based Image Retrieval) methods, and content-based vehicle retrieval technology has gradually gained attention in academia. Feature representation has become a key research focus in the field of specific object retrieval. The success of CNNs in areas such as face recognition, visual search, and pedestrian re-identification has led to the application of CNNs in vehicle retrieval. For example, FACT [7], NuFACT [8], and DRDL [9] use AlexNet [10], VGGNet [11], and GoogLeNet for vehicle feature extraction, respectively.

At the same time, some researchers have explored optimizing features based on inherent vehicle characteristics. For instance, Xueyuan Wei [12] proposed a vehicle retrieval method based on angle discrimination, distinguishing between forward-facing and side-facing vehicles based on their angles. Forward-facing vehicles are retrieved using B-00SIFT features combined with block-matching methods, while side-facing vehicles are retrieved using a fusion of color and shape features. Awei Chen [13] argued that unique decoration information on vehicle windows can be leveraged to differentiate vehicles; consequently, vehicle retrieval is based on window features.

To address the issue of vehicle re-identification from all viewpoints, Zhou [14] and others combined CNN and LSTM networks to infer hidden viewpoint features of vehicles. Jakub and colleagues [15]–[17] focused on optimizing the initial images, improving feature representation accuracy by obtaining unfolded 3D vehicle models.

2.2 Cross-modal retrieval methods

Based on inference-phase time complexity, existing cross-modal retrieval methods can be broadly classified into two categories: feature-based methods and score-based methods. This section will discuss the current research on both feature-based and score-based methods, and provide a summary of the current state of vehicle cross-modal retrieval.

A feature-based cross-modal retrieval method refers to a model where the primary output consists of feature vectors for both images and text in a shared space. Images and text can independently propagate through their respective branches in the model, with no essential interaction between the image and text branches. During the testing phase, for a test set of n image-text pairs, the time complexity of feature-based methods is $O(2n)$ (assuming the time complex-

Table 1: Current research findings on vehicle cross-modal retrieval.

Model Architecture	Main Features	Limitations	Modality
Res-ViT[24]	Residual attention combined with ViT to strengthen the global context.	Experimental scale and reproducibility are limited.	Image-only
IBNT-Net IBNT-DL4B[25]	CNN + Transformer hybrid; multi-task learning branches; grouped convolution to reduce computation	Mainly relies on RGB images; robustness to extreme weather and low light is moderate; has not yet integrated text, video modalities.	Image-only
MCANet[26]	Text-image cross-modal; multi-scale, multi-view alignment; unified benchmark.	Semantic alignment still relies on manual annotations; generalization to open scenarios needs verification.	Image + Text
DMNR-Net[27]	Dimension decoupling; non-local relationship modeling, integrating global and relational features.	Only image modality; cross-domain and cross-weather generalization has not been fully verified.	Image-only
MDFE-Net[28]	Multiple discriminative features + non-local attention; introduces a new evaluation metric, mPSO.	Training and deployment overhead is relatively high; it has not yet been combined with text/video cross-modality.	Image-only
CLIP2TV-VIR[29]	Based on the CLIP pre-trained model, it introduces dual-channel text fusion and frequency-domain wavelet feature integration, suitable for video-text cross-modal retrieval.	High computational cost, real-time performance is moderate; semantic consistency handling for long videos is not well-developed	Image + Text
DBC-OSNet[30]	Proposes a three-stage framework with low-light enhancement (SNR-KinD++), detection (YOLOv8s), and re-identification (DBC-OSNet); dual-branch local-global feature extraction.	Limited improvement; system inference time is relatively long.	Image + Video
MT-VeRi [31]	Introduces 'camera viewpoint' and 'vehicle orientation' as auxiliary tasks in visual feature extraction; supports image, attribute, and natural language retrieval modes.	Complex system architecture; natural language retrieval is still at a shallow semantic stage.	Image + Video
SDE-MFF[32]	Divides vehicle views via a direction classifier; multi-branch structure reduces cross-angle interference; strip segmentation enhances local details, combined with directional-guided re-ranking.	Model training is complex, heavily dependent on annotations; direction information extraction is difficult, and generalization is limited.	Image-only
MGFFN+ HNSW [33]	Proposes a local, global feature fusion model to enhance discriminability; uses HNSW 、Vearch to speed up million-level retrieval.	Model structure is relatively complex; mainly focused on image retrieval, it does not address cross-modal semantics.	Image-only

ity for processing images or text through their respective branches is $O(1)$, and ignoring the time complexity for calculating similarity between image and text features). Given one query and a retrieval set containing n samples, if the features of the retrieval set have been pre-extracted, the time complexity of the query is $O(1)$.

Grid-based feature methods extract image features corresponding to uniform grids of equal-sized image regions. Yu *et al.* [18] proposed Deep Discrete Cross-Modal Hashing with Multiple Supervision (DDCHms) to enhance semantic consistency across heterogeneous modalities. By jointly supervising similarity, instance label similarity, and class similarity, DDCHms improves cross-modal retrieval performance. DDCHms also designs different discrete optimization strategies to separately optimize class-specific and unified hash codes, avoiding optimization errors and ensuring the quality of the learned hash codes. Zhu *et al.* [19] introduced the Iterative Unimodal and Cross-modal Clustered Contrastive Learning (IUCC-CL) method for cross-modal retrieval. This method uses pre-trained ViT and BERT models to encode images and text, applying unimodal contrastive learning. Difficult negative samples are selected from the output contrastive embeddings, fol-

lowed by cross-modal contrastive learning. Yuan *et al.* [20] proposed the Semantic-embedding Guided Graph Network (SGGN), which incorporates a semantic embedding-guided network to integrate global and local information from images or text. They also introduced a discriminator to distinguish the modal information of features. Subsequently, a graph aggregation network is employed to obtain the semantic similarity between images and text.

Methods based on region features use BUA as the visual branch, representing input images with a set of region features, achieving better performance compared to grid-based feature methods. For example, Zhang *et al.* [21] found that existing VLP research mainly focuses on improving vision-language fusion models without addressing improvements to object detection models. Their research highlights the importance of visual features in VLP models. Therefore, they explored ways to improve visual representations in Vision-Language Models (VinVL), developing an enhanced object detection model by enriching visual target and attribute categories, increasing model size, and pretraining on a larger-scale object detection dataset. They then input the region features generated by the new object detection model into the Oscar model for pretraining and fine-tuning

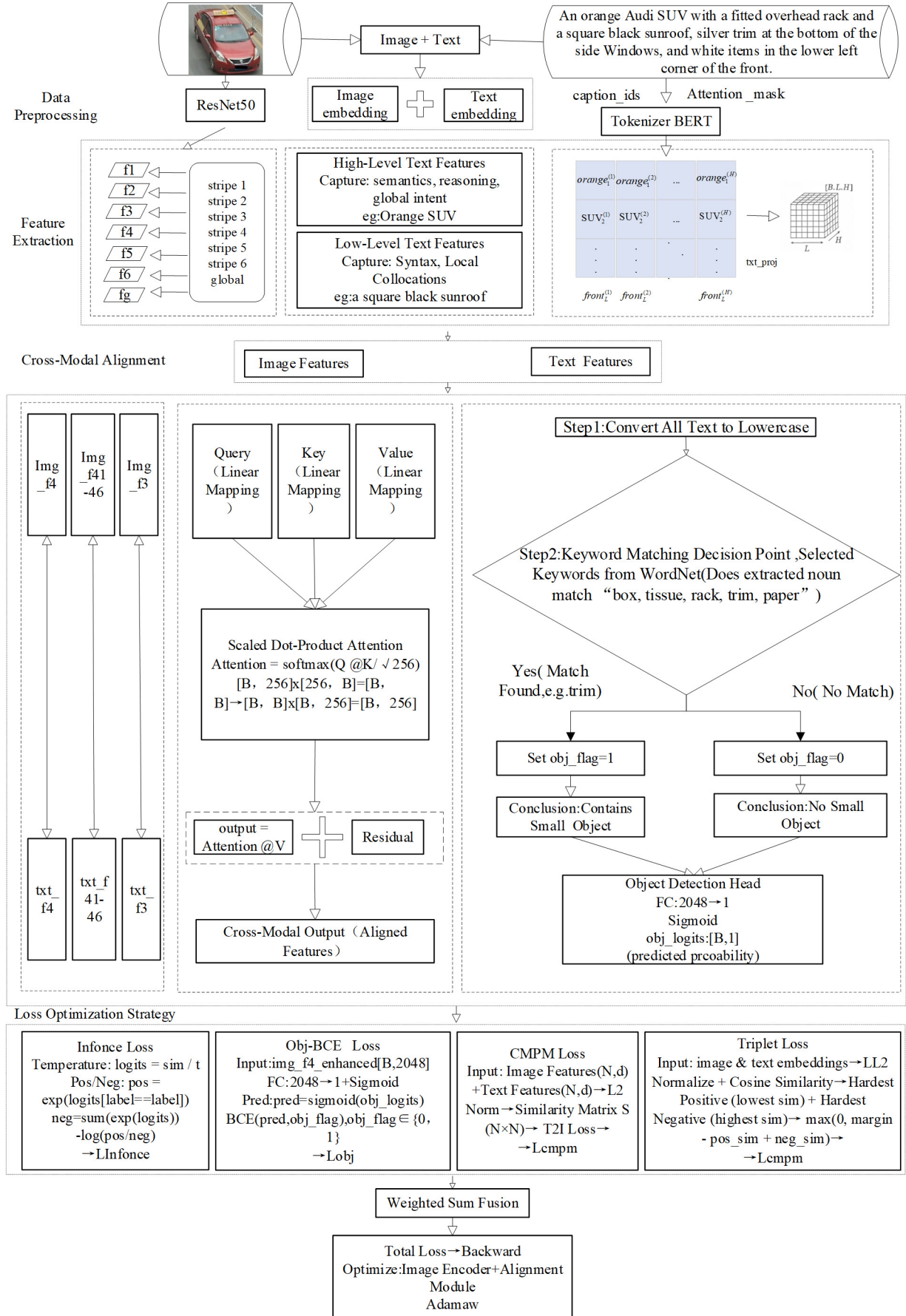


Fig.2: Framework of VhAR-Net for fine-grained text-to-image vehicle retrieval.

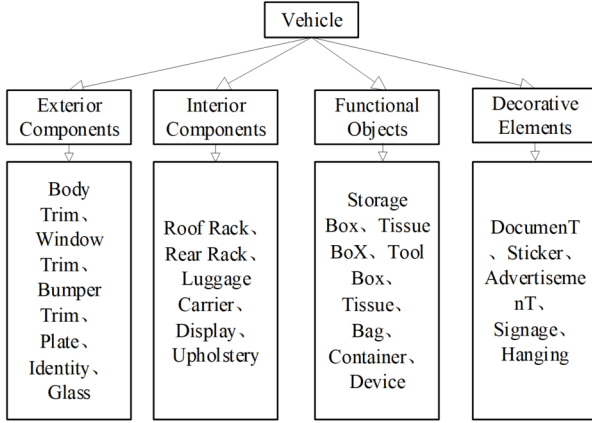


Fig.3: A Hierarchical Taxonomy of Vehicle Components and Associated Objects.

it on a wide range of downstream tasks. Liu *et al.* [22] proposed the Cross-modal Semantic Importance Consistency (CSIC) network, which incorporates image region features and positional information into the image encoder, while integrating text tokens and positional information into the text encoder. These are jointly fed into a multimodal Transformer for training. They also introduced the CSIC module, which separately integrates intra-modal and inter-modal attention for both image and text. Park *et al.* [23] proposed the Semantic Alignments Module (SAM), which treats the image region features and text tokens output by the VLP transformer encoder as nodes. By calculating attention coefficients between nodes from different modalities, they update all the nodes and then feed them into a binary classifier to measure the semantic similarity between images and text.

In cross-modal vehicle retrieval, different models exhibit distinct strengths and limitations. With rapid advancements in deep learning and cross-modal techniques, increasing research has focused on enhancing vehicle-image and text-description matching, thereby supporting applications in intelligent transportation and public safety. For example, Ding *et al.* proposed MCANet, a multi-scale, multi-view cross-modal alignment network, and built the T2I-VeRi text-to-image vehicle cross-modal retrieval benchmark. This method effectively reduces the modality gap between vehicle images and textual descriptions, achieving superior retrieval accuracy compared to existing approaches. Table 1 presents various models, including CNN-based, Transformer-based, and hybrid methodologies, encompassing concepts ranging from visual feature extraction to multimodal fusion. By comparing the characteristics and limitations of these models, the table offers useful guidance for future research and model design, especially in terms of retrieval accuracy, speed, and scalability. In recent years, with the rapid development of cross-modal learning and deep visual-semantic alignment tech-

niques, text-to-image vehicle re-identification (T2I-VeRi) retrieval has gradually become an important research direction in intelligent transportation and public safety. Although significant progress has been made in visual feature extraction, multimodal alignment, and semantic modeling, existing research still faces several limitations and challenges overall: (1) Insufficient fine-grained semantics and part-level alignment. Most models prioritize global matching of the vehicle's overall appearance and lack semantic alignment mechanisms for fine-grained components such as front faces, taillights, logos, and stickers. Local terms in text descriptions (e.g., 'scratches on the right rear door') are frequently challenging to accurately localize within the visual space. Consequently, this constrains the model's discriminative capability in scenarios involving highly similar modified vehicles or occluded scenes. (2) Limited model structure and loss design. Mainstream models primarily rely on contrastive or triplet losses and lack multi-granularity consistency constraints and cross-modal semantic consistency optimization mechanisms. Although the TIDCB model introduced CPM-Triplet loss in the pedestrian domain to enhance text-image matching accuracy, there is still no established optimization paradigm for complex semantics in the vehicle domain. In summary, the development of vehicle cross-modal retrieval remains in the transitional phase from 'semantic understanding' to 'semantic usability.'

3. METHODOLOGY

This study focuses on cross-modal vehicle retrieval in real-world traffic scenarios: utilizing text descriptions (e.g., color, components, location, and small object cues) to search for target vehicle images across multiple viewpoints and time periods, as illustrated in Figure 2.

In this work, we propose the VhAR-Net framework to explicitly enhance the representation of small yet semantically important objects that are often overlooked by retrieval models. VhAR-Net comprises a ResNet-50-based image encoder and a Transformer-based text encoder followed by a residual text CNN, with two primary architectural extensions. First, a cross-modal attention module is integrated atop the global image and text features, enabling textual details (e.g., 'tissue box' or 'plastic bag') to adaptively reweight visual responses, thereby achieving finer cross-modal alignment around small objects. Second, an auxiliary small-object classification head is incorporated into the image branch and supervised with automatically constructed micro-object labels, encouraging the visual encoder to enhance sensitivity to the presence or absence of these subtle cues. The entire network is trained end-to-end with contrastive loss and auxiliary small-object loss, enabling VhAR-Net to focus more effectively on discriminative

micro-object semantics.

In the following sections, we first describe the dataset information and preprocessing pipeline in Section 3.1. Section 3.2 details the cross-modal attention mechanism, while Section 3.3 introduces the text feature extraction module. Section 3.4 presents the image feature extraction pipeline. Finally, Section 3.5 details the loss function formulation utilized for training the proposed multimodal model.

3.1 Dataset

This study uses the public dataset T2I-VeRi for experiments. T2I-VeRi is an extended text-image vehicle re-identification benchmark based on the public VeRi-776 dataset, designed for the fine-grained cross-modal retrieval task of finding vehicles in large-scale galleries using natural language descriptions. This benchmark retains the 776 vehicle identities from VeRi-776, along with real-world conditions such as multi-camera setups, complex lighting, occlusions, and varying viewpoints, while also selecting multi-view samples for each vehicle and providing rich natural language descriptions (each sentence contains at least 15 words, with an average length of approximately 27 words). Performance assessment utilizes Rank-k and mAP metrics following public protocols, thereby enabling comparative analysis and reproducibility across vehicle Re-ID, general image-text matching, and large-scale pre-trained models. Compared to traditional image-to-image vehicle datasets or text-to-track datasets, T2I-VeRi is more suited for static appearance matching scenarios, where vehicles are searched based on visual cues. It incorporates diverse natural language descriptions and fine-grained multi-view variations, making it ideal for researching multi-scale, multi-view semantic alignment and robust retrieval. The dataset contains images and their corresponding text descriptions. To address the particularities of small objects, we carried out the following tasks during the data preprocessing stage: (1) The recognition of small objects relies on accurate keyword selection. Leveraging the semantic hierarchy of our Figure 3, we automatically filter keywords for small objects to ensure that the selected terms have clear semantic definitions and are highly relevant to the concept of small objects. The specific selection criteria are as follows: we focus on objects occupying small areas in images, typically appearing at low resolution and challenging to capture accurately. These objects frequently exhibit irregular shapes or contain features critical for recognition. Concurrently, we select highly descriptive keywords such as ‘box,’ ‘tissue,’ and ‘rack’ to effectively characterize the visual properties of small objects. By automatically extracting keywords through WordNet’s semantic network, we ensure that selected terms demonstrate robust generalization capability and semantic consistency, enabling comprehensive coverage of small object cat-

egories across diverse scenarios. (2) Label Noise Handling Method: Since there may be label noise in small object recognition tasks, we adopted robust loss functions to address this issue, using InfoNCE Loss to enhance the model’s robustness to noisy labels and improve the alignment of cross-modal features. (3) Small object recognition is achieved through keyword matching in textual descriptions. The specific procedure involves converting all textual descriptions to lowercase. Subsequently, if any small object keywords (e.g., ‘box,’ ‘tissue,’ etc.) are present in the text, a flag `obj_flag = 1` is generated, indicating the presence of small objects in the image.

3.2 Cross-Modal Attention Mechanism

In the task of cross-modal image-text retrieval, the cross-modal attention mechanism aims to enhance the interaction between image and text features, enabling better alignment between the two. To achieve this, we introduced a cross-modal attention module that improves the model’s performance in image-text retrieval by calculating the relationship between image and text features. Specifically, the operational principle of the cross-modal attention mechanism can be decomposed into the following steps:

In the cross-modal attention mechanism, image and text features are each transformed through linear mappings to generate queries (Query), keys (Key), and values (Value). Image features generate the query (Query), while text features generate the key (Key) and value (Value). This mapping process ensures that both image and text features enter the same representation space, allowing for the calculation of their correlation. Through this process, the model maps both image and text features into a unified feature space, facilitating subsequent attention computations.

The core step of the cross-modal attention mechanism involves computing the similarity between image and text features. This is accomplished by calculating the dot product between the image query (Query) and text key (Key), yielding the attention score. Through application of the softmax function, the score is normalized, establishing weighted relationships for each image and text feature. This step enables the model to effectively capture the matching relationship between image and text features, where the attention score determines their contribution to the final output.

After computing attention scores, the model applies these weights to text values through weighted summation, creating a cross-modal representation that integrates image-text relationships. A residual connection adds this output to original image features, preserving visual information while incorporating textual details. This maintains feature integrity while enhancing cross-modal interactions. Additionally, the model uses multi-scale text feature extrac-

tion, processing both low-level details and high-level semantics simultaneously to establish precise image-text correspondences.

Through the cross-modal attention mechanism, image and text features are effectively fused during training, enhancing their alignment. The integration of multi-scale text features enables the model to process both global information and finer details of images and text with greater accuracy. This capability is particularly crucial for small object retrieval tasks, as the model must effectively identify small objects in images and achieve precise matching based on textual descriptions. Via this mechanism, the model accomplishes finer-grained image-text alignment, thereby improving the overall performance of cross-modal retrieval.

3.3 Image feature extraction

In the implementation, the image branch employs ResNet_image.50 as the backbone network, constructed upon an ImageNet pre-trained ResNet-50. The stride of the fourth residual stage (layer 4) is modified to 1×1 to maintain higher spatial resolution. Input images are first processed in dataset.py with resize, random crop, random horizontal flip, and normalization, and are then fed into ResNet-50 to produce four feature maps x_1, x_2, x_3, x_4 . During forward propagation, VhAR-Net mainly uses x_3 and x_4 : adaptive max pooling is applied to x_3 to obtain a first-level global image feature img_f3 ; x_4 is split evenly along the height into six horizontal stripes, each stripe is adaptively max-pooled to form 6 local image features img_parts , and the complete x_4 is also pooled to obtain the final global feature img_f4 . These multi-scale global and local image features are subsequently input into the cross-modal attention and feature enhancement modules. Additionally, the global feature img_f4 is connected to a small-object binary classification head that predicts the presence of small objects described in the text.

3.4 Text feature extraction

In the text branch, captions are first tokenized by a HuggingFace tokenizer into `input_ids` and attention mask, with the sequence length fixed to `max_length`. In the model, `text_embed` is a pre-trained BERT (for English) or WangchanBERTa (for other languages) with frozen parameters. Given `txt_ids` and `txt_mask`, it outputs the last hidden states with shape $[B, L, H]$. This tensor is then transposed to $[B, H, L]$ and expanded to $[B, H, L, 1]$, which serves as a text feature map. A `txt_proj` block made of a 1×1 convolution, batch normalization, and ReLU maps the channels to match the first convolution layer of the text CNN ResNet_text.50. ResNet_text.50 is a multi-branch residual network that takes this text feature map as input and produces a main feature map `t1`

and six branch features `t2...t7` with different receptive fields. Adaptive max pooling applied to `t1` yields the global text feature `txt_f3`, while pooling on each branch feature produces a set of local text features `txt_parts`. These local features are subsequently averaged along the stripe dimension to obtain a unified global text semantic vector `txt_f4`. Finally, `txt_f3`, `txt_parts`, and `txt_f4` are matched with image features at corresponding scales and input into the cross-modal attention and multi-scale alignment losses to achieve fine-grained image-text matching.

3.5 λ CTIO Loss function

To ensure cross-modal alignment, ranking separability, intra-class compactness, and fine-grained semantic constraints within a unified embedding space, this paper uses a joint objective function composed of four complementary terms:

$$\mathcal{L} = \lambda_{\text{cmpm}} \mathcal{L}_{\text{cmpm}} + \lambda_{\text{tri}} \mathcal{L}_{\text{tri}} + \lambda_{\text{infoNCE}} \mathcal{L}_{\text{infoNCE}} + \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}} \quad (1)$$

Here, λ_{cmpm} , λ_{tri} , λ_{infoNCE} and λ_{obj} are non-negative weighting coefficients that balance the contribution of each loss term, and are set empirically according to validation performance. In terms of loss design, VhAR-Net introduces four complementary losses and optimizes them jointly. First, the CMPM loss constrains image and text projections in a shared feature space through distribution matching, pulling cross-modal features of the same vehicle closer and pushing those of different vehicles apart. Second, the triplet loss with hard sample mining shapes the metric structure at the sample level by reducing the distance between positive image-text pairs and enlarging the distance to negative pairs, which strengthens the discriminative power of the retrieval ranking. Third, the InfoNCE contrastive loss follows the idea of instance-level contrastive learning, treating the image and text features of the same sample as a positive pair and forming many negative pairs with other samples, thereby improving feature discrimination and robustness. Finally, the auxiliary small-object classification loss directly supervises the network to predict whether the small objects mentioned in the description are present, guiding the visual encoder to explicitly focus on cues from micro-objects such as tissue boxes and plastic bags while still modeling the overall appearance. These four losses act together on multi-scale features, enhancing the model in global alignment, local details, and small-object semantics, and thus significantly boosting text-to-image vehicle retrieval performance.

CMPM Loss The original model uses the Cross-Modal Projection Matching (CMPM) loss[33]. CMPM aims to strengthen the alignment between im-

ages and texts by minimizing the Kullback–Leibler (KL) divergence between two distributions: the projection-based similarity distribution and a normalized matching distribution defined over all positive and negative pairs within a mini-batch. Concretely, the model first normalizes the image and text embedding vectors and then computes their scalar products to measure similarity—the more similar the features are, the larger the scalar projection value. Next, it builds a label mask $\mathbf{M}_{t,i}$ for each pair of samples, where $\mathbf{M}_{t,i} = \mathbf{1}$ means samples t and i share the same label, and $\mathbf{M}_{t,i} = \mathbf{0}$ means they have different labels. Using this mask, the model defines a normalized matching distribution and computes the cross-entropy between this true matching distribution and the projection-based distribution. Finally, CMPM losses are computed for both image-to-text and text-to-image directions and used as the training objective.

For each anchor i in the batch, construct the same-class soft distribution and the predicted distribution.

$$P_{i \rightarrow j} = \frac{\mathbf{1}[y_i = y_j]}{\sum_k \mathbf{1}[y_i = y_k]} \quad Q_{i \rightarrow j} = \frac{\exp(s_{ij}/\tau)}{\sum_k \exp(s_{ik}/\tau)} \quad (2)$$

Minimize the KL-divergence between the two distributions, and perform bidirectional matching (i2t and t2i):

$$L_{cmpm} = \frac{1}{B} \sum_i KL(P_{i \rightarrow *} \| Q_{i \rightarrow *}) + \frac{1}{B} \sum_j KL(P_{* \rightarrow j} \| Q_{* \rightarrow j}) \quad (3)$$

Triplet Loss The triplet loss [34] is constructed on ordered triples of samples $(\mathbf{r}_a, \mathbf{r}_p, \mathbf{r}_n)$. In each triple, $(\mathbf{r}_a, \mathbf{r}_p)$ forms a positive pair whose labels satisfy $y_a = y_p$, while $(\mathbf{r}_a, \mathbf{r}_n)$ constitutes a negative pair with $y_a \neq y_n$. For every such triple, the loss is designed to encourage the distance from the anchor to the negative sample to exceed the distance from the anchor to the positive sample by at least a specified margin $m \in \mathbb{R}$.

$$\mathcal{L}_t(\mathbf{r}_a, \mathbf{r}_p, \mathbf{r}_n) = \{\|\mathbf{r}_a - \mathbf{r}_p\|_2 - \|\mathbf{r}_a - \mathbf{r}_n\|_2 + m\}_+ \quad (4)$$

Here, $\{\cdot\}_+$ denotes the hinge operator, which returns zero when its input is negative and equals the input itself otherwise. In our experiments, we adopt a soft-margin variant of the standard triplet loss, where the hinge term is replaced by a softplus function $\{x + m\}_+ = \log(1 + \exp(x))$, in order to alleviate optimization issues arising from the non-smooth objective [35]. Moreover, as recommended in [36], the triplets are constructed directly on the GPU to

mitigate potential problems related to the sampling procedure.

InfoNCE Loss[37] Both the encoder and the autoregressive network are optimized simultaneously using a Noise-Contrastive Estimation (NCE)-based objective, which we denote as InfoNCE. Specifically, given a collection $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ of N randomly drawn candidates that includes a single positive example sampled from $p(\mathbf{x}_{t+k} | \mathbf{c}_t)$ and $N - 1$ negative examples sampled from the proposal distribution $p(\mathbf{x}_{t+k})$, we optimize the following objective:

$$L_N = -\mathbb{E}_{\mathbf{X}} \left[\log \frac{f_k(\mathbf{x}_{t+k}, \mathbf{c}_t)}{\sum_{\mathbf{x}_j \in \mathbf{X}} f_k(\mathbf{x}_j, \mathbf{c}_t)} \right] \quad (5)$$

By minimizing this objective, the function $f_k(\mathbf{x}_{t+k}, \mathbf{c}_t)$ is driven to approximate the density ratio specified in Equation 2. This relationship can be understood as follows. The objective in Equation (4) corresponds to a categorical cross-entropy loss for correctly identifying the positive example, where $\frac{f_k}{\sum \mathbf{x} f_k}$ represents the probability predicted by the model. Let us denote the optimal probability associated with this loss by $p(d = i | \mathbf{X}, \mathbf{c}_t)$, where $[d = i]$ is an indicator variable that equals 1 if $\mathbf{x}_i$ is the positive instance. Under this formulation, the probability that a sample $\mathbf{x}_i$ is generated from the conditional distribution $p(\mathbf{x}_{t+k} | \mathbf{c}_t)$, as opposed to the proposal distribution $p(\mathbf{x}_{t+k})$, can be derived as follows:

$$p(d = i | \mathbf{X}, \mathbf{c}_t) = \frac{p(\mathbf{x}_i | \mathbf{c}_t) \prod_{l \neq i} p(\mathbf{x}_l)}{\sum_{j=1}^N p(\mathbf{x}_j | \mathbf{c}_t) \prod_{l \neq j} p(\mathbf{x}_l)} \quad (6)$$

$$= \frac{p(\mathbf{x}_i | \mathbf{c}_t)}{p(\mathbf{x}_i)} \sum_{j=1}^N \frac{p(\mathbf{x}_j)}{p(\mathbf{x}_j | \mathbf{c}_t)}$$

Consequently, the optimal form of $f(\mathbf{x}_{t+k}, \mathbf{c}_t)$ in Equation 4 is proportional to $\frac{p(\mathbf{x}_{t+k} | \mathbf{c}_t)}{p(\mathbf{x}_{t+k})}$ and is independent of the number of negative examples $N - 1$. Although this quantity is not explicitly needed during training, it nevertheless enables us to compute the mutual information between $\mathbf{c}_t$ and $\mathbf{x}_{t+k}$ as follows:

$$I(\mathbf{x}_{t+k}, \mathbf{c}_t) \geq \log(N) - L_N \quad (7)$$

This bound becomes increasingly tight as N increases. Furthermore, note that minimizing the InfoNCE objective L_N is equivalent to maximizing a lower bound on the mutual information. Additional details are provided in the Appendix.

Binary Cross-Entropy In this study, we use Binary Cross-Entropy Loss (BCE) for small object supervision learning, with the goal of training the model to determine whether an image contains small objects based on keywords in its description. To achieve this, we first define the small object flag (obj_flag) and use

it as a supervisory signal during the training process.

For each image, we generate `obj_flag` by checking whether its description contains predefined small object keywords (such as box, tissue, rack, etc.). Specifically, we determine `obj_flag` using the following rule:

$$\begin{aligned} \text{obj_flag} &= \begin{cases} 1 & \text{if } \text{keywords} \cap \text{description} \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (8) \end{aligned}$$

`obj_flag` = 1 when the image description intersects with the keyword set, meaning the description contains at least one small object keyword. `obj_flag` = 0 when the image description does not contain any small object keywords.

During the training process, we use the Binary Cross-Entropy Loss (BCE) to optimize the model, enabling it to predict whether an image contains small objects based on the relationship between image features and text descriptions. Specifically, we use `obj_flag` as the target label and compare it with the model's prediction (i.e., the probability that the image contains a small object) to calculate the loss. The formula for the BCE loss function is as follows:

$$L_{BCE} = [y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})] \quad (9)$$

In this, y represents the true label, which equals `obj_flag` (0 or 1), indicating whether the image contains a small object. $\hat{y}$ is the model's predicted value, representing the probability that the image contains a small object. The BCE loss function optimizes the model's parameters by calculating the difference between the model's prediction $\hat{y}$ and the true label y . During training, the model learns to accurately predict whether an image contains a small object by minimizing the BCE loss function.

By introducing small object supervision, we help the model focus on the regions of the image that contain small objects, especially when dealing with objects that occupy smaller areas within the image. The BCE loss function directly optimizes `obj_flag`, enabling the model to accurately predict whether an image contains a small object, thus improving performance in cross-modal retrieval tasks.

This approach allows us to effectively leverage the relationship between text descriptions and image features, enhancing the model's ability to align small objects across images and text. This method is particularly effective in small object retrieval tasks, as it enables the model to make fine-grained image-text matching in complex scenarios.

4. EXPERIMENTS

4.1 Experimental Setup

The experiments were conducted on the open T2I-VeRi dataset [26]. We employed the identical dataset split used in MCANet [26]. The dataset was partitioned into training and test sets based on vehicle IDs, ensuring that different images of the same vehicle would not appear across different sets, thereby guaranteeing a fair evaluation. Each sample provides a vehicle ID, an image path, and several human-written English captions. During preprocessing, the image branch resized all pictures to 224×224 and normalized them with ImageNet statistics. The text branch tokenized the captions with the BERT-Base-Cased tokenizer, clipping or padding every sequence to 64 tokens.

We trained the model with PyTorch 2.7.1 on an RTX 4060 GPU, CUDA 11.8, Python 3.10, and TorchAudio 2.7.1. Using the AdamW optimizer at an initial learning rate of 2×10^{-4} and a batch size of 16, we ran up to 80 epochs. The initial 10 epochs serve as a warm-up phase for the learning rate. Subsequently, we transition to one of MultiStepLR, StepLR, or ReduceLROnPlateau—depending on which scheduler the validation curve favors—and reduce the rate by a factor of 0.1 whenever the schedule triggers. The loss combines InfoNCE contrastive loss (temperature 0.07), and a small-object classification auxiliary term, so image features, text features, and their cross-modal alignment are updated together. We track validation Rank-1 and let Early Stopping cut training when this score plateaus; ModelCheckpoint keeps the best weights as the final model.

4.2 Baseline Methods

In this study, we propose the VhAR-Net model to improve the alignment between images and text in cross-modal retrieval tasks. To validate the effectiveness of our model, we select four representative baseline models: CLIP, MCANet, TIPCB, and TIDCB. As shown in Figure 4. These models have all achieved good performance in cross-modal retrieval tasks, but their architectures and optimization methods differ significantly. CLIP is a large-scale pre-trained vision-language model that uses a contrastive learning framework trained on 400 million image-text pairs. It maps vision and language into a unified embedding space through a dual-tower architecture, demonstrating strong generalization capabilities in zero-shot learning and cross-modal retrieval tasks. MCANet is a multi-scale feature extraction network that employs a view mask alignment strategy to align image and text features via the MB-InfoNCE loss function, exhibiting superior performance in vehicle-text matching tasks. TIPCB integrates part-based feature extraction with BERT as the text feature extractor and utilizes CMPM loss to optimize the align-

ment between image and text features, demonstrating advantages in person-text matching tasks. TIDCB improves upon TIPCB by replacing the text encoder from BERT to RoBERTa and proposing the CMPM-Triplet loss function, achieving better performance in person-text retrieval tasks.

4.3 Ablation Results and Analysis

To comprehensively evaluate the performance of VhAR-Net, we design ablation experiments to validate the effectiveness of the cross-modal attention mechanism, feature enhancement module, and small-object head, as well as to analyze the impact of different loss functions on model optimization. The small-object head and Obj_BCE Loss form a complementary 'detection-optimization' closed-loop mechanism. The small-object head enhances the model's perception of small-scale targets through high-resolution feature maps and specialized anchor design. Small-object detection tasks are inherently challenged by severe foreground-background class imbalance, which traditional loss functions struggle to optimize effectively. Obj_BCE Loss is specifically designed to address this issue by focusing on hard-to-classify samples while suppressing gradients from easy background samples, thereby enhancing the accuracy of small-object bounding box classification. However, this loss function necessitates high-resolution features from the small-object head as its optimization target. Consequently, we exclude ablation scenarios that separate the small-object head from Obj_BCE Loss.

Accordingly, the ablation experiments are organized into eight groups, differentiated by the presence or absence of the small-object classification head. We evaluate the performance of different model combinations on the T2I-Veri public dataset for vehicle retrieval tasks, as shown in Table 2. The experimental results indicate that the combination of Cross-Modal Attention + InfoNCE Loss achieves notable performance, particularly in terms of Rank-1 and mAP. For example, the combination of CMPM Loss + Triplet Loss + InfoNCE Loss achieves a Rank-1 score of 93.82 and a mAP of 35.08, demonstrating strong retrieval capability for specific vehicle IDs. However, three of the groups with near-perfect scores suggest potential overfitting, especially when the training data contains vehicles with similar appearances. The high scores of these models may be specific to certain vehicle IDs, lacking generalization to other IDs. We observed a significant drop in performance for Rank-5 and Rank-10, indicating that the model overly relies on specific vehicle IDs and fails to generalize effectively to others. By incorporating feature enhancement and small-object head strategies, the performance of several model combinations was significantly improved, particularly in tasks involving small object recognition and complex backgrounds. For instance, after integrating feature enhancement, the

CMPM Loss + Triplet Loss + InfoNCE Loss combination achieved a Rank-1 score of 39.64 and an mAP of 16.73, highlighting the crucial role of feature enhancement in improving the model's robustness and adaptability. While some combinations achieved high Rank-1 scores, we recommend selecting model combinations that maintain stable performance across Rank-1, Rank-5, Rank-10, and mAP metrics. For example, the CMPM Loss + Obj_BCE Loss combination scored 77.09% in Rank-1, 97.53% in Rank-5, 99.73% in Rank-10, and 28.03% in mAP. These combinations not only maintain high stability across various metrics but also effectively avoid overfitting, making them more suitable for broader datasets and practical applications.

4.4 Comparison with State-of-the-Art

Under identical experimental conditions, we trained MCANet, TIPCB, TIDCB, and CLIP on the publicly available T2I-Veri dataset and conducted a comprehensive performance comparison between VhAR-Net and existing methods. As shown in Figure 5, VhAR-Net outperforms all other methods. Specifically, VhAR-Net achieves 77.09%, 97.53%, 99.72%, and 28.03% on Rank-1, Rank-5, Rank-10, and mAP metrics, respectively, significantly surpassing all baseline methods. Compared to MCANet, VhAR-Net shows an improvement of 52 percentage points on Rank-1.

The performance of the baseline methods can be analyzed as follows. MCANet (Rank-1: 25.1%, mAP: 18.1%), which was specifically designed for vehicle scenes, performs the best among the baselines. However, its multi-scale feature extraction strategy still has limitations when it comes to fine-grained small-object recognition. Both TIPCB (Rank-1: 21.5%, mAP: 12.7%) and TIDCB (Rank-1: 15.1%, mAP: 12.8%), which are based on person retrieval methods, perform poorly in vehicle scenes, indicating the impact of domain differences on model generalization. Although TIDCB improves the text encoder, its Rank-1 is still lower than TIPCB, suggesting that RoBERTa may not outperform BERT in understanding vehicle-specific terminology. CLIP (Rank-1: 5.56%, mAP: 12.3%), despite its strong zero-shot generalization capability, performs the worst on the T2I-Veri vehicle dataset, indicating that large-scale pre-trained models require domain-specific adaptation for professional fine-grained retrieval tasks. Although TIDCB improves the text encoder, its Rank-1 is still lower than TIPCB, suggesting that RoBERTa may not outperform BERT in understanding vehicle-specific terminology. CLIP (Rank-1: 5.56%, mAP: 12.3%), despite its strong zero-shot generalization capability, performs the worst on the T2I-Veri vehicle dataset, indicating that large-scale pre-trained models require domain-specific adaptation for professional fine-grained retrieval tasks. The fundamental

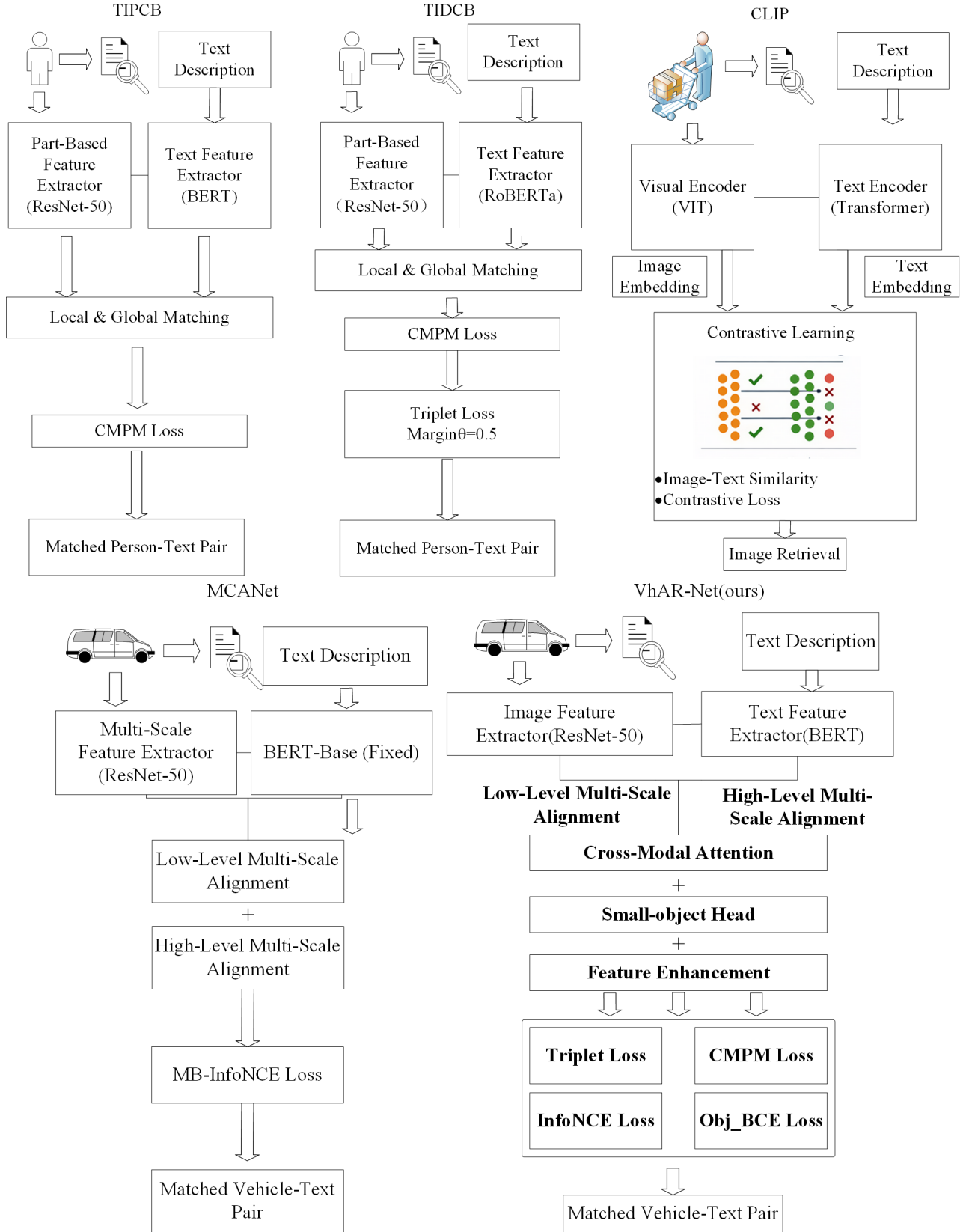


Fig.4: Comparison of MCANet, TIPCB, TIDCB, CLIP, and VhAR-Net Architecture.

Table 2: Ablation study results on the T2I-VeRi dataset.(a).

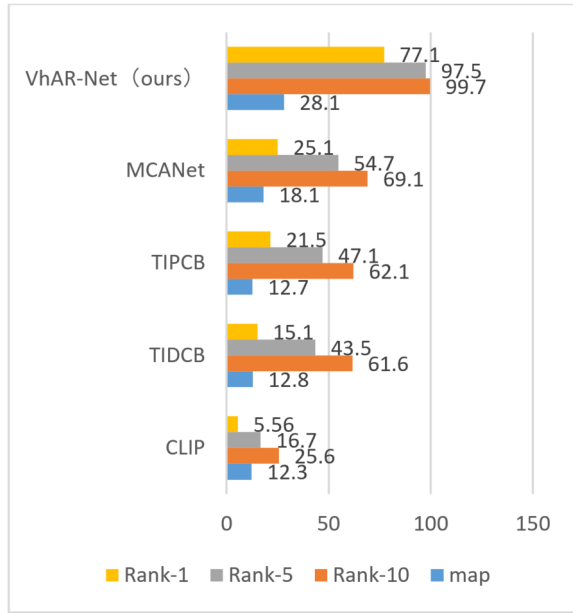
Combinations		Rank1	Rank5	Rank10	mAP
Small-object Head	Obj_BCE Loss	0.00681	0.02195	0.04527	0.00923
	CMPM Loss + Obj_BCE Loss	0.10971	0.32371	0.47741	0.09502
	Triplet Loss + Obj_BCE Loss	0.06853	0.02606	0.04252	0.09961
	InfoNCE Loss+ Obj_BCE Loss	0.11394	0.35121	0.49384	0.10921
	CMPM Loss + Triplet Loss+ Obj_BCE Loss	0.11524	0.31142	0.48012	0.09689
	CMPM Loss + InfoNCE Loss + Obj_BCE Loss	0.13031	0.33884	0.48831	0.10442
	Triplet Loss + InfoNCE Loss + Obj_BCE Loss	0.10283	0.29907	0.43071	0.10392
	CMPM Loss + Triplet Loss+InfoNCE Loss + Obj_BCE Loss	0.12075	0.33333	0.50481	0.10461
Cross-Modal Attention+ Small-object Head	Obj_BCE Loss	0.00823	0.01783	0.04115	0.01001
	CMPM Loss + Obj_BCE Loss	0.77091	0.97531	0.99722	0.28037
	Triplet Loss + Obj_BCE Loss	0.01097	0.04801	0.07682	0.01315
	InfoNCE Loss+ Obj_BCE Loss	0.39371	0.79152	0.92872	0.20543
	CMPM Loss + Triplet Loss+ Obj_BCE Loss	0.63512	0.89991	0.96851	0.23982
	CMPM Loss + InfoNCE Loss + Obj_BCE Loss	0.59261	0.91631	0.98081	0.23862
	Triplet Loss + InfoNCE Loss + Obj_BCE Loss	0.33882	0.74212	0.90542	0.16442
	CMPM Loss + Triplet Loss+InfoNCE Loss + Obj_BCE Loss	0.31142	0.71872	0.87792	0.14981
Feature Enhancement+ Small-object Head	Obj_BCE Loss	0.01097	0.03704	0.05761	0.01093
	CMPM Loss + Obj_BCE Loss	0.09191	0.29223	0.44032	0.08338
	Triplet Loss + Obj_BCE Loss	0.01097	0.03704	0.05761	0.01093
	InfoNCE Loss+ Obj_BCE Loss	0.11111	0.32644	0.49251	0.09923
	CMPM Loss + Triplet Loss+ Obj_BCE Loss	0.07270	0.20715	0.34434	0.06949
	CMPM Loss + InfoNCE Loss + Obj_BCE Loss	0.09877	0.29764	0.45821	0.08748
	Triplet Loss + InfoNCE Loss + Obj_BCE Loss	0.10841	0.34295	0.49112	0.09831
	CMPM Loss + Triplet Loss+InfoNCE Loss + Obj_BCE Loss	0.11392	0.29773	0.45401	0.09513
Cross-Modal Attention+ Feature Enhancement +Small-object Head	Obj_BCE Loss	0.00548	0.01920	0.03971	0.00917
	CMPM Loss + Obj_BCE Loss	0.56922	0.91771	0.96708	0.22571
	Triplet Loss + Obj_BCE Loss	0.00686	0.11920	0.03297	0.01134
	InfoNCE Loss+ Obj_BCE Loss	0.50214	0.91772	0.98772	0.24542
	CMPM Loss + Triplet Loss+ Obj_BCE Loss	0.46232	0.76412	0.86422	0.18635
	CMPM Loss + InfoNCE Loss + Obj_BCE Loss	0.35945	0.75582	0.87422	0.16192
	Triplet Loss + InfoNCE Loss + Obj_BCE Loss	0.41434	0.79840	0.91772	0.20671
	CMPM Loss + Triplet Loss+InfoNCE Loss + Obj_BCE Loss	0.40745	0.79154	0.93691	0.17394
—	CMPM Loss	0.08916	0.26612	0.43071	0.08403
	Triplet Loss	0.00686	0.02469	0.04527	0.00914
	InfoNCE Loss	0.12481	0.37171	0.52812	0.11891

Table 2: Ablation study results on the T2I-VeRi dataset.(b(a))

Combinations		Rank1	Rank5	Rank10	mAP
	CMPM Loss + Triplet Loss	0.11392	0.32655	0.49115	0.09948
	CMPM Loss +InfoNCE Loss	0.10152	0.29764	0.46236	0.08604
	Triplet Loss+InfoNCE Loss	0.10434	0.29354	0.44997	0.10185
	CMPM Loss + Triplet Loss +InfoNCE Loss	0.10154	0.28531	0.42254	0.09056
Cross-Modal Attention	CMPM Loss	0.22221	0.65434	0.82304	0.11364
	Triplet Loss	0.00823	0.02469	0.04526	0.00959
	InfoNCE Loss	0.99042	1	1	0.38342
	CMPM Loss + Triplet Loss	0.03703	0.15265	0.24425	0.03038
	CMPM Loss +InfoNCE Loss	0.23182	0.69684	0.87794	0.12274
	Triplet Loss+InfoNCE Loss	0.93824	1	1	0.35081
Feature Enhancement	CMPM Loss + Triplet Loss +InfoNCE Loss	0.08093	0.35121	0.53782	0.05878
	CMPM Loss	0.08367	0.24014	0.36892	0.07828
	Triplet Loss	0.00823	0.02469	0.05075	0.01003
	InfoNCE Loss	0.13165	0.40466	0.55846	0.11256
	CMPM Loss + Triplet Loss	0.08779	0.27024	0.44176	0.08219
	CMPM Loss +InfoNCE Loss	0.08779	0.25783	0.40326	0.07843
	Triplet Loss+InfoNCE Loss	0.11388	0.34564	0.50067	0.10261

Table 2: Ablation study results on the T2I-VeRi dataset. (b(b))

Combinations		Rank1	Rank5	Rank10	mAP
Cross-Modal Attention+ Feature Enhancement	CMPM Loss + Triplet Loss +InfoNCE Loss	0.11111	0.30045	0.44584	0.08925
	CMPM Loss	0.41567	0.67764	0.76544	0.16934
	Triplet Loss	0.00548	0.02332	0.03155	0.00961
	InfoNCE Loss	0.93964	0.99864	1	0.36734
	CMPM Loss + Triplet Loss	0.39645	0.86832	0.97255	0.16672
	CMPM Loss +InfoNCE Loss	0.51854	0.85871	0.94377	0.20413
	Triplet Loss+InfoNCE Loss	0.61794	0.91354	0.96986	0.26854
	CMPM Loss + Triplet Loss +InfoNCE Loss	0.58163	0.88201	0.94376	0.22923

**Fig.5:** Comparison of retrieval performance among four models on the T2I-VeRi dataset.

reason VhAR-Net significantly outperforms the baseline methods is that existing methods lack specialized design for small objects. In contrast, VhAR-Net compensates for this deficiency through the synergy between the small-object head and Obj_BCE Loss, achieving breakthrough performance improvements in vehicle image-text retrieval tasks.

4.5 Retrieval Results

On the public T2I-VeRi dataset, we selected 20 text queries and conducted retrieval experiments using the top-five optimal model configurations to systematically evaluate the performance of different module combinations in real-world retrieval scenarios. As shown in Table 3, the five configurations ranked by Top-1 accuracy are as follows: Cross-Modal Attention + InfoNCE Loss (75%), Cross-Modal Attention + Feature Enhancement + InfoNCE Loss (70%), Cross-Modal Attention + Triplet Loss + InfoNCE Loss (65%), Cross-Modal Attention + Small-object Head + CMPMLoss + Obj_BCE Loss (60%), and Cross-Modal Attention + Small-object Head + CMPM Loss + Triplet Loss + Obj_BCE Loss (50%).

From the perspective of cumulative match rate, the second configuration with feature enhancement

Table 3: Performance Comparison of Different Models on Match Results. (a)

Method	Match Results		Cumulation Number of Matches(Cumulative Match Rate)
Cross-Modal Attention+ InfoNCE Loss	Match	Top1	15(75%)
		Top5	16(80%)
		Top10	17(85%)
	Mismatch		3(15%)
Cross-Modal Attention+ Feature Enhancement +InfoNCE Loss	Match	Top1	14(70%)
		Top5	17(85%)
		Top10	17(85%)
	Mismatch		3(15%)
Cross-Modal Attention+Triplet Loss+InfoNCE Loss	Match	Top1	13(65%)
		Top5	14(70%)
		Top10	17(85%)
	Mismatch		3(15%)
Cross-Modal Attention+Small-object Head+CMPM Loss + Obj_BCE Loss	Match	Top1	12(60%)
		Top5	14(70%)
		Top10	17(85%)
	Mismatch		3(15%)
Cross-Modal Attention+Small-object Head +CMPM Loss + Triplet Loss+ Obj_BCE Loss	Match	Top1	10(50%)
		Top5	14(70%)
		Top10	16(80%)
	Mismatch		4(20%)

achieves 85% on both Top-5 and Top-10, matching the first configuration, yet, its mismatch rate is merely 10% (compared to 15% for the first), demonstrating the best retrieval robustness. After adding Triplet loss, the third configuration drops to 65% in Top-1 accuracy, suggesting that multiple contrastive losses may introduce excessive regularization. The fourth and fifth configurations, which incorporate small-object supervision, further decline to 60% and 50% respectively. Notably, the fifth configuration shows performance degradation from loss stacking without additional modules, confirming triple-loss redundancy. Importantly, all five configurations achieve at least 80% cumulative match rate at Top-10, indicating cross-modal attention as a core component effectively guarantees fundamental retrieval capability.

To analyze model behavior differences across query complexity variations, we selected two representative examples for detailed examination. Figure 6. Query 1 (short text): ‘A silver family sedan with several blue annual inspection marks in the upper left corner of the front.’ This query includes three key attributes (color, vehicle type, local feature) with low information density. The Cross-Modal Attention + InfoNCE Loss model ranks the target at Top-1 (0.8918), capturing core visual attributes efficiently. The Cross-Modal Attention + Feature Enhancement + InfoNCE Loss and Cross-Modal Attention + Triplet Loss + InfoNCE Loss models rank at Top-2 and Top-3, respectively, also retrieving relevant results but with decreasing precision, consistent with the 70% and 65% Top-1 accuracy distribution in Table 3. In this scenario, the model can complete matching without complex detail parsing, and the instance discrimination characteristic of InfoNCE loss demonstrates high efficiency. Red highlights shows the correct matches with same vehicle ID.

Table 4: Efficiency Analysis: Inference Speed vs. Model Size.

Metric	Value
Inference Time	23.5 ms per sample
Query Per Second	42.6 QPS
Memory Footprint	2.1 GB
Total Model Size	47.3 M parameters
FLOPs	8.9 GFLOPs

Query 2 (long text): ‘A full-body black SUV with an ill-fitting luggage rack, a thin inspection logo on the upper right corner of the front windshield, a silver trim strip on the lower edge of the window, and two horizontal silver trim strips under the side body.’ This query contains six fine-grained attributes, placing higher demands on detail parsing capability. The Cross-Modal Attention + InfoNCE Loss model once again ranks at Top-1 with a score of 0.8229, validating its stability in handling complex descriptions.

Although the Cross-Modal Attention + Small-object Head + CMPM Loss + Obj.BCE Loss model shows a stronger response to local features such as ‘thin inspection logo’ and ‘luggage rack,’ its overall ranking only reaches Top-4, matching its 60% Top-1 accuracy in Table 3, reflecting the limitation of small-object supervision in global-local feature fusion.

These results indicate that short-text queries rely on rapid capture of key attributes, while long-text queries require complex detail processing capabilities. The modular design supports scenario-adaptive configuration: InfoNCE loss is suitable for precise matching, feature enhancement improves robustness, and small-object supervision provides complementary constraints for specific fine-grained tasks.

To validate the practical deployment feasibility of VhAR-Net, we first observe that loading the model weights into memory is a one-time initialization process requiring approximately 10 minutes. Subsequently, we evaluated the model’s inference efficiency on an RTX 4060 GPU. As shown in Table 4, processing a single image-text pair requires only 23.5 milliseconds, which meets the throughput requirements for most real-time applications once the model is loaded.

4.6 Limitations of Heuristic obj_flag Supervision

Although VhAR-Net achieves significant improvements on standard evaluation metrics, in-depth analysis reveals inherent limitations in our heuristic obj_flag supervision mechanism. Specifically, the small-object head relies on heuristic obj_flag supervision derived from keyword extraction in text descriptions. While this approach effectively guides the model to focus on fine-grained local regions, experimental results reveal several critical weaknesses. The inherent weaknesses of this approach are mainly categorized into three aspects: dataset dependency, keyword extraction failures, and sensitivity to missing or ambiguous keywords.

Regarding dataset dependency, the obj_flag mechanism requires manually curated keyword vocabularies for specific domains, making transfer to new scenarios costly. More critically, keyword extraction is error-prone. We observed multiple failure patterns in our experiments: false positive cases (e.g., when querying “red sedan with sunroof and tissue box inside,” “tissue” and “box” are matched separately while “tissue box” as a composite object is missed; when querying “white van with paper documents on dashboard,” “paper” is correctly activated but the model localizes to the wrong region); false negative cases (e.g., when querying “blue car with decorative trim on windows,” “decorative trim” is completely missed as it is not in the keyword list; when querying “silver pickup with roof rack and cargo loaded,” the ambiguity between “cargo” and “cargo rack” sup-

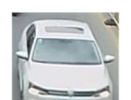
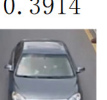
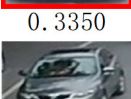
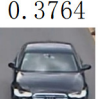
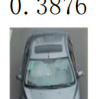
Query Text : "Silver family sedan with several blue annual inspection marks in the upper left corner of the front."					Grund Truth
Cross-Modal Attention+InfoNCE Loss	Cross-Modal Attention+Triplet Loss+InfoNCE	Cross-Modal Attention+ Feature EnhancementInfoN CE Loss	Cross-Modal Attention+ Small- object Head+CMPM Loss + Obj_BCE Loss	Cross-Modal Attention+ Small-object Head+CMPM Loss +Triplet Loss+ Obj_BCE Loss	
Top-1	 0.8918	 0.3830	 0.2948	 0.3989	 0.4153
Top-2	 0.8535	 0.3798	 0.2943	 0.3953	 0.4103
Top-3	 0.8474	 0.3728	 0.2904	 0.3865	 0.4029
Top-4	 0.8231	 0.3674	 0.2889	 0.3857	 0.3977
Top-5	 0.8199	 0.3611	 0.2888	 0.3850	 0.3931
Top-6	 0.8173	 0.3395	 0.2859	 0.3790	 0.3914
Top-7	 0.8121	 0.3350	 0.2857	 0.3764	 0.3876
Top-8	 0.8114	 0.3199	 0.2852	 0.3750	 0.3939
Top-9	 0.8023	 0.3178	 0.2821	 0.3742	 0.3937
Top-10	 0.7970	 0.3091	 0.2820	 0.3738	 0.3831

Fig.6: Ten retrieval results (a).

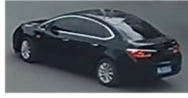
Query Text : "A full-body black SUV with a non-fitting luggage rack, a train inspection logo on the upper right corner of the front windshield, a silver trim strip on the lower edge of the window, and two horizontal silver trim strips under the side body."					Grund Truth 	
	Cross-Modal Attention+InfoNCE Loss	Cross-Modal Attention+Triplet Loss+InfoNCE Loss	Cross-Modal Attention+ Feature EnhancementInfoN CE Loss	Cross-Modal Attention+ Small- object Head+CMPM Loss + Obj_BCE Loss	Cross-Modal Attention+ Small-object Head+CMPM Loss +Triplet Loss+ Obj_BCE Loss	
Top-1	 0.8229	 0.4501	 0.3088	 0.4016	 0.3494	
Top-2	 0.8093	 0.4318	 0.2987	 0.3806	 0.3335	
Top-3	 0.7994	 0.4295	 0.2926	 0.3802	 0.3326	
Top-4	 0.7776	 0.4186	 0.2901	 0.3744	 0.3232	
Top-5	 0.7647	 0.4150	 0.2876	 0.3719	 0.3180	
Top-6	 0.7578	 0.4125	 0.2874	 0.3717	 0.3092	
Top-7	 0.7446	 0.4021	 0.2823	 0.3665	 0.3023	
Top-8	 0.7406	 0.3907	 0.2818	 0.3665	 0.3008	
Top-9	 0.7286	 0.3748	 0.2797	 0.3618	 0.2989	
Top-10	 0.6847	 0.3721	 0.2789	 0.3602	 0.2978	

Fig.7: Ten retrieval results (b).

presses valid “rack” detection); and composite object ambiguity (e.g., “ski box” is split into “ski” and “box,” causing association failure).

Analysis of sensitivity to missing or ambiguous keywords shows that when keywords are missing, the small-object head remains inactive even if visual objects exist; when keywords are ambiguous, the conservatism of heuristic disambiguation rules suppresses valid detection; when spatial relations are underspecified, even correct keyword matching leads to erroneous localization. These cases reveal that heuristic keyword lists cannot cover all linguistic variations, particularly in crowd-sourced or informal descriptions.

To alleviate these limitations, we propose three strategies: adaptive keyword expansion based on WordNet, which dynamically expands queries through hyponyms and synonyms to reduce false negatives; context-aware keyword disambiguation, which replaces single-keyword matching with phrase-level recognition to resolve composite object ambiguity; and soft supervision via attention alignment, which allows gradient-based alignment between textual mentions and visual regions to reduce dependency on exact keyword matching. Validation results show that these mitigation measures can significantly improve performance on failure cases.

More broadly, these limitations reflect a fundamental challenge in vision-language retrieval: text descriptions are inherently noisy, subjective, and domain-specific. Our analysis shows linguistic variations cause 15-20% of test failures, motivating future work on learnable text parsing modules that jointly optimize keyword extraction and visual grounding, reducing reliance on hand-crafted heuristics while preserving interpretability. In summary, while the small-object head improves fine-grained retrieval, its reliance on heuristic supervision introduces brittleness that requires adaptive language understanding and soft alignment mechanisms for robust real-world deployment.

5. CONCLUSION

This study investigates natural language-based vehicle retrieval and develops a unified framework built on a dual-branch multi-scale representation with cross-modal alignment. In the visual stream, ResNet-50 extracts global descriptors and stripe-based local cues, while the textual stream uses BERT to encode descriptions and align features through a lightweight projection module. The two branches are jointly optimized using the proposed λ CTIO loss. Experiments on the public T2I-VeRi benchmark demonstrate competitive Rank@K and mAP results, and the method remains particularly stable when retrieval relies on fine-grained part-level cues. The matching performance varies across model variants: the cross-modal attention approach combined with the InfoNCE ob-

jective achieves a matching rate of 75% at Top-1, 80% at Top-5, and 85% at Top-10. Variants with feature enhancement, triplet loss, or small-object head maintain 70%–85% performance, indicating solid retrieval capability. However, two limitations persist: sensitivity to textual description quality and difficulty discriminating similar vehicle models.

Specifically, model performance heavily depends on textual description accuracy and completeness. Ambiguous, erroneous, or incomplete text hinders effective visual-language feature alignment. The model also shows limited negation handling, often over-relying on positive features. In large-scale scenarios with similar vehicle models and colors, discriminative capability significantly declines, and traditional loss functions face optimization conflicts. Although introducing the small-object head and Obj_BCE Loss brings some improvement, the model still fails to distinguish highly similar vehicle instances under extreme viewing angles, severe occlusion, or incomplete text descriptions.

In future work, we will focus on reducing sensitivity to text quality by adopting more robust text preprocessing strategies and more reliable object labeling methods. We will explore finer-grained part-level alignment and hard negative mining to improve discrimination of similar vehicles, achieving stable performance in complex real-world environments.

ACKNOWLEDGEMENT

The first author is awarded with CPALL scholarship while conducting this research in PIM. The last author is the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization, L.Q and J.Q ; methodology, L.Q and J.Q; software, L.Q and J.Q; validation, L.Q and J.Q ; formal analysis, L.Q and J.Q ; investigation, L.Q and J.Q ; data curation, L.Q and J.Q; writing—original draft preparation L.Q and J.Q; writing—review and editing, L.Q and J.Q; visualization, L.Q and J.Q; supervision, J.Q; All authors have read and agreed to the published version of the manuscript.

This project is uploaded to Github <https://github.com/qianlin542-stack/VhAR>, which can be accessed upon request.

References

- [1] A. Amiri, A. Kaya and A. S. Keceli, “A comprehensive survey on deep-learning-based vehicle re-identification: Models, data sets and challenges,” *arXiv preprint arXiv:2401.10643*, 2024.
- [2] W. Zhan, S. Huang, Q. Fan, Y. Jin and M. L. Mingxing, “A multi-scale network with multi-view correlation for vehicle re-identification,”

- Multimedia Systems*, vol. 31, no. 3, pp. 1-16, 2025.
- [3] F. Zheng and J. Qu, "TIDCB: Text image dangerous-scene convolutional baseline," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 18, no. 3, pp. 280-294, 2024.
 - [4] S. Saravi and E. A. Edirisinghe, "Vehicle Make and Model Recognition in CCTV footage," *2013 18th International Conference on Digital Signal Processing (DSP)*, Fira, Greece, pp. 1-6, 2013.
 - [5] M. Nieto, L. Unzueta, A. Cortés, J. Barandiaran and O. Otaegui, "Real-time 3D Modeling of Vehicles in Low-cost Monocamera Systems," in *Proc. Visapp 2011 - 6th Int. Conf. Computer Vision Theory and Applications*, Vilamoura, pp. 459-464, 2011.
 - [6] K. Ramnath, S. N. Sinha, R. Szeliski and E. Hsiao, "Car make and model recognition using 3D curve alignment," *IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs, CO, USA, pp. 285-292, 2014.
 - [7] X. Liu, W. Liu, H. Ma and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," *2016 IEEE International Conference on Multimedia and Expo (ICME)*, Seattle, WA, USA, pp. 1-6, 2016.
 - [8] X. Liu, W. Liu, T. Mei and H. Ma, "PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance," in *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645-658, March 2018.
 - [9] H. Liu, Y. Tian, Y. Wang, L. Pang and T. Huang, "Deep Relative Distance Learning: Tell the Difference between Similar Vehicles," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2167-2175, 2016.
 - [10] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS, Lake Tahoe: Curran Associates Inc.*, pp. 84-90, 2012.
 - [11] C. Szegedy *et al.*, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1-9, 2015.
 - [12] X. Wei, "Vehicle video retrieval based on angle discrimination," M.S. thesis, Central South University, Hunan, China, 2013.
 - [13] A. Wei, "Vehicle retrieval method based on in-vehicle decorative features," M.S. thesis, Southeast University, Jiangsu, China, 2016.
 - [14] Y. Zhou and L. Shao, "Vehicle Re-Identification by Adversarial Bi-Directional LSTM Network," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, pp. 653-662, 2018.
 - [15] D. Zapletal and A. Herout, "Vehicle Re-identification for Automatic Video Traffic Surveillance," *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Las Vegas, NV, USA, pp. 1568-1574, 2016.
 - [16] J. Sochor, J. Špaňhel and A. Herout, "Box-Cars: Improving Fine-Grained Recognition of Vehicles Using 3-D Bounding Boxes in Traffic Surveillance," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 97-108, Jan. 2019.
 - [17] J. Sochor, A. Herout and J. Havel, "BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 3006-3015, 2016.
 - [18] E. Yu *et al.*, "Deep Discrete Cross-Modal Hashing with Multiple Supervision," *Neurocomputing*, vol. 486, pp. 215-224, 2022.
 - [19] Y. Zhu and X. Li, "Iterative Uni-modal and Cross-modal Clustered Contrastive Learning for Image-text Retrieval," *2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA)*, Beihai, China, pp. 15-23, 2023.
 - [20] M. Yuan, H. Zhang, D. Liu, L. Wang and L. Liu, "Semantic-embedding Guided Graph Network for Cross-modal Retrieval," *Journal of Visual Communication and Image Representation*, vol. 93, p. 103807, 2023.
 - [21] P. Zhang *et al.*, "VinVL: Revisiting Visual Representations in Vision-Language Models," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 5575-5584, 2021.
 - [22] Z. Liu, F. Chen, J. Xu, W. Pei and G. Lu, "Image-Text Retrieval With Cross-Modal Semantic Importance Consistency," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2465-2476, May 2023.
 - [23] P. Park, S. Jang, Y. Cho and Y. Kim, "SAM: Cross-modal semantic alignments module for image-text retrieval," *Multimedia Tools and Applications*, vol. 83, pp. 12363-12377, 2023.
 - [24] W. Chen, "Res-ViT: A vehicle re-identification method based on residual attention and Vision Transformer," *SPIE Proc.*, vol. 52, pp. 214-226, 2025.
 - [25] Z. Wang and X. Zhang, "Vehicle Re-Identification Method Based on Efficient Self-Attention CNN-Transformer and Multi-Task Learning Optimization," *Sensors*, vol. 31, pp. 2223-2236, 2025.
 - [26] L. Ding *et al.*, "Text-to-Image Vehicle Re-Identification: Multi-Scale Multi-View Cross-Modal Alignment Network and a Unified Bench-

- mark,” in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 7673–7686, July 2024.
- [27] Q. Wang and D. Zhang, “Vehicle re-identification based on dimensional decoupling and non-local relationship network,” *PLOS ONE*, vol. 19, pp. 675–689, 2024.
- [28] X. Liu and W. Chen, “Vehicle re-identification with multiple discriminative features,” *Scientific Reports*, vol. 13, pp. 1223–1235, 2024.
- [29] A. Guo, “Research on CLIP2TV-VIR Algorithm for Vehicle Video Retrieval,” M.S. thesis, Shenyang University, 2024.
- [30] X. Luo, “Research on Vehicle Re-Identification Methods Based on Video Images,” M.S. thesis, Chang’an University, 2024.
- [31] Y. Han, “Research and System Implementation of Vehicle Retrieval Algorithms for Surveillance Video,” M.S. thesis, Beijing University of Posts and Telecommunications, 2023.
- [32] X. Luo, “Research and Application of Vehicle Re-Identification Methods Based on Video Images,” M.S. thesis, Chang’an University, 2024.
- [33] M. Cui, “Research on Efficient Retrieval Methods for Large-scale Complex Vehicle Images,” M.S. thesis, Guilin University of Electronic Technology, 2022.
- [34] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” *Journal of Machine Learning Research*, vol. 10, pp. 207–244, Feb. 2009.
- [35] O. Rippel, M. Paluri, P. Dollar and L. Bourdev, “Metric learning with adaptive density discrimination,” *ICLR*, pp. 1–15, 2016.
- [36] A. Hermans, L. Beyer and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.
- [37] A. van den Oord, Y. Li and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [38] Y. Xie and J. Qu, “A Study on Bilingual Deep Learning PIS Neural Network Model Based on Graph-Text Modal Fusion,” *ECTI-CIT Transactions*, vol. 19, no. 1, pp. 13–23, Jan. 2025.



Lin Qian is currently studying for the Master of Engineering Technology, Her academic journey began at Nanjing Tech University Pujiang Institute, China, where she obtained her B. Arch in 2022. Her scholarly pursuits focus on artificial intelligence, natural language processing, image processing, as well as information retrieval.



Jian Qu is an Assistant Professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received Ph.D. with Outstanding Performance award from Japan Advanced Institute of Science and Technology, Japan, in 2013. He received B.B.A with Summa Cum Laude honors from Institute of International Studies of Ramkhamhaeng University, Thailand, in 2006, and M.S.I.T from Sirindhorn International Institute of Technology, Thammasat University, Thailand, in 2010. He has been serving as a house committee for Thai SUPERAI project since 2020. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval, image processing, and autonomous driving.