Comprehensive Visual Features and Pseudo Labeling for Robust Natural Language-based Vehicle Retrieval

Bach Hoang Ngo[1†], Dat Thanh Nguyen[2†], Nhat-Tuong Do-Tran[1], Phuc Pham Huy Thien[3], Minh-Hung An[2], Tuan-Ngoc Nguyen[1], Loi Nguyen Hoang[1], Vinh Dinh Nguyen[4], Quang-Vinh Dinh[*]

[1] AI Vietnam Research, Vietnam
[2] FPT Telecom, Vietnam
[3] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology
[4] School of Computing and Information Technology, Eastern International University, Vietnam

Abstract

Vehicle retrieval has become crucial for public transportation and intelligent transportation systems due to the exponential development of large-scale transportation videos. Vehicle re-identification and vehicle tracking are the main components of most vision-based vehicle recovery systems. Unfortunately, the limited amount of information provided by traffic video feeds limits the vision-based vehicle retrieval algorithm's efficacy. Therefore, this article proposes a contrastive cross-modal vehicle retrieval approach to maximize the complementarity of natural language and visual representations. An efficient method to fuse multiple input image features is also proposed to extract comprehensive information from various vehicles along with pseudo labeling and efficient post-processing techniques to enhance retrieval accuracy. The proposed method achieved the 3rd ranking of Mean Reciprocal Rank (MRR) score of 0.4795 on the test set for the Challenge Track 2: Tracked-Vehicle Retrieval by Natural Language Descriptions 2023. Source code for the proposed approaches is openly accessible at https://github.com/anminhhung/AI-City-2023-Track2.

1. Introduction

Efficient traffic management in an AI city requires the capability to search for vehicles. Current methods for vehicle search mainly involve image matching, and hence include vehicle detection, re-identification, and tracking [11] [26]. However, the query image may not be accessible, for real-world scenarios, with only a general description of the target vehicle. Therefore, there is a pressing need for natural language (NL) based vehicle retrieval solutions. The objective for NL-based vehicle retrieval is to locate particular vehicles based on textual descriptions, a crucial aspect of smart transportation and urban surveillance systems. Current vehicle retrieval systems typically rely on matching images, i.e, performing vehicle re-identification by comparing a query image against images in a gallery. Real-world scenarios can often be restricted by the image being searched for not being accessible, and only providing vague explanations or descriptions for the target vehicle.

Obtaining and using natural language text descriptions is simpler than image queries, and provides greater flexibility. However, it also presents significant retrieval challenges since text and image have marked different modalities. Typical current vehicle retrieval methods using NL embed images and descriptions into a common feature space(s) and use cross-modal similarities to rank vehicle images. Dominant techniques [3] establish visual encoder and text encoder using fixed backbones and optimize a limited number of projection layers. DUN [21] proposed a dual-path network to align video and text embeddings for vehicles. However, they did not consider motion information, which can help differentiate between similar vehicles. In contrast, CLV [17] utilized symmetric InfoNCE loss to learn cross-modal representation and then subsequently generated global motion images for each track, which retained vehicle appearance information. They also enhanced subject descriptions in the text, improving vehicle appearance information impacts. However, motion maps for CLV lack vehicle appearance information, which significantly reduces model performance and effectiveness.

Zhao et al. [25] proposed an NL-based vehicle retrieval method, overcoming CLV limitations using spatial relationship modeling (SSM) to learn both visual and linguistic representations. Adopting a symmetric network means SSM can extract both internal and external vehicle characteris-

[*]Corresponding author; [†] equal contribution

tics, including type, color, shape, motion state, and surrounding environment. This was achieved using separate visual and text encoders to extract vehicle appearance features and text embeddings, respectively. In contrast to [4] and [5], SSM optimizes the extracted features using symmetric InfoNCE and pair-wise Circle losses. However, despite its effectiveness, SMM only uses a single frame as model input. Thus, SSM can not consider different vehicle angles, causing significant vulnerable to misrepresentation for local vehicle attributes and poor general features tracking. Considering these issues, current SMM model performance could possibly be increased by providing more features from vehicle angles. Therefore this paper proposes an efficient method to fuse multiple frames, allowing the model to learn various vehicle angles, and also includes pseudo-labeling and several post-processing modules to enhance retrieval results. The proposed method achieved a 3rd rank with Mean Reciprocal Rank(MRR) = 0.4795 on the Challenge Track 2 test set: Tracked-Vehicle Retrieval by Natural Language Descriptions 2023.

The remainder of this paper is structured as follows: Section 2 presents a survey of existing approaches for natural language-based vehicle retrieval systems. An efficient method to fuse multiple input image features is introduced to extract comprehensive information from various vehicles along with pseudo labeling and efficient post-processing techniques to enhance retrieval accuracy in Section 3. Section 4 discusses the performance of the proposed method and baseline method. Finally, Section 5 presents the concluding remarks.

2. Related Work

Video retrieval by NL has become a notable research area, with recent works focusing on mapping features from diverse spaces onto a common semantic space. Text queries are encoded using language feature extractors [22], and visual information is obtained using sparse frames and video segments [13]. Encoders, such as attention mechanisms and convolutional neural networks (CNNs), are commonly employed to learn global and local contexts for the video retrieval frameworks [8].

Metric learning techniques are applied while learning a function to minimize feature distances, hence enabling effective video-language understanding. For example, Bai et al. [1] employed InfoNCE loss to manage sample similarity and improve performance.

Recent studies have considered various approaches to achieve effective video retrieval using natural language. Luo et al. [8] conducted an empirical study on end-to-end video clip retrieval using the clip4clip method, whereas Wang et al. [22] focused on feature extraction and natural language processing for deep learning of the English language. Xiong et al. [13] considered sparse spatiotemporal representation with adaptive regularized dictionary learning for low-bit-rate video coding.

2.1. Text Feature Extraction

Numerous studies have demonstrated the usefulness of vector space in representing words. Early approaches such as Word2Vec [9] and LSTM [4] were used to encode NL for language representations, but transformer-family architectures have become more popular for word representation due to their effectiveness. Devlin et al. [2] highlight the significance of bidirectional pre-training for language representations, which can outperform other methods in both sentence-level and token-level tasks. Thus, bidirectional pre-training enables the model to understand language context and meaning, enabling better performance for natural language processing tasks. Findings from Devlin et al.'s provide insights into the most effective approach to word representation, which can then be applied in various domains such as machine translation, text classification, and question answering.

2.2. Image Feature Extraction

Convolutional neural networks are fundamental for extracting image features and finding widespread applications in various vehicle tracking tasks in urban environments, including vehicle classification [23], vehicle detection [20], and vehicle re-identification [7].

On the other hand, CNNs can also be used as a specific feature extractor for identifying vehicle color [5] and type [18]. Although these vehicle characteristics are essential for retrieving individual objects, global features from the video, such as environmental attributes, scenes, and related entities, can significantly influence candidate video ranking accuracy.

2.3. Contrastive Language-Image Pre-training

Retrieving a single object based on new and unseen inputs poses a challenging problem, particularly when relying solely on a supervised retrieval pipeline. Recent studies have shown unsupervised domain adaptation to be a promising solution for improving model generalization by leveraging knowledge from the training data to apply to unlabeled data [15]. Although this approach has demonstrated significant effectiveness in various domains, its application in text-video retrieval remains limited. Therefore, the current study proposes a domain adaptive method to bridge the gap between training and test sets. Specifically, the proposed approach utilizes unsupervised data augmentation for consistency in training to enhance the model's ability to handle new inputs that may not have been present in the training data. Thus, the proposed approach improves text-video retrieval system performance and enhances its effectiveness in real-world scenarios.

3. Proposed Method

3.1. Overview

The proposed method comprises two branches that simultaneously learn static information(colors, type) and dynamic information (direction, location, nearby vehicles) of that vehicle. The two branches are subsequently merged to generate global features between visual and text (as shown in Fig. 1). Symmetric InfoNCE [12] and Circle [19] are applied(similar to previous studies) to associate text and image representations, and the joint project these embedding features onto a unified representation space. The main contribution of this paper is to alter the input to show more information about vehicle characteristics. We then concatenate multiple frames to allow the model to learn various vehicle angles and subsequently employ pseudo-labeling and several post-processing modules to enhance retrieval results.

3.2. Multi-View Association

Previous studies considering the local vehicle image branch, [25] only employ a single frame as model input to identify local vehicle information. However, the model will be biased using this approach toward that frame for the vehicle itself on the whole road. This means the model can not consider different vehicle angles, making it vulnerable to misrepresentation of local vehicle attributes. We adjusted image information in this branch by extracting more frames, intending to also access vehicle trajectory information.

Specifically, we concatenate N frames in a trajectory onto a single input image. And then uniformly sample N frames from the track such that each frame has an approximately similar size and their positions are evenly distributed in the vehicle' trajectory. We selected N = 4 frames, and the 4 corner frames are pasted onto the image corner frames to create the concatenated image, as shown in Fig. 2.

3.3. Pseudo Labeling

The test set could be a potential unlabeled data source for training; hence we implemented pseudo-labeling [6] as semi-supervised learning to cultivate the training process. This step, combine pseudo data generated from the best results from the 50% test set and available training data, creating a larger training dataset. For each query, we select 3 tracks that have the highest similarity score and then create 3 new data samples. These new samples are subsequently merged into the original data to create a richer training dataset.

$$y_i^{'} = \begin{cases} 1 & \text{if } f_i(x) \text{ in top 3 of } f(x) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The utilized approach enhances the model's ability to generalize by utilizing unlabeled data [6]. It relies on the data samples being located in high-density regions which are more likely to have similar labels among their neighboring samples. Therefore, the method encourages the network's output to remain consistent despite any variations that may occur in low-dimensional manifold's directions.

3.4. Loss Functions

Consider a batch with N language-vision feature pairs $\{(f_i^{lang}, f_i^{vis})\}_{i=1}^{N}$, that includes total $N \times N$ sample pairs.We apply the symmetric infoNCE [12] and Circle [19] loss algorithms to simultaneously maximize similarity between N positive pairs and minimize $N \times (N-1)$ negative pairs.

3.4.1 Symmetric InfoNCE loss

The symmetric InfoNCE loss algorithm minimizes the cosine similarity for cross-model negative pairs and maximizes positive language-vision pairs. The formula consists of two parts vision-to-language and language-to-vision:

$$\mathcal{L}_{INCE} = \mathcal{L}_{vis \to lang} + \mathcal{L}_{lang \to vis} \tag{2}$$

Vision-to-language InfoNCE loss helps vision to distinguish between their corresponding language and other language features. Thus, It is formulated as:

$$\mathcal{L}_{vis \to lang} = \frac{1}{N} \sum_{n=1}^{N} -log \left(\frac{e^{cos(f_i^{vis}, f_i^{lang})/\tau}}{\sum_{j=1}^{N} e^{cos(f_i^{vis}, f_j^{lang})/\tau}} \right) \tag{3}$$

Similarly, we used language-to-vision infoNCE loss, to help language features distinguish between corresponding vision and other visual features.

$$\mathcal{L}_{lang \to vis} = \frac{1}{N} \sum_{n=1}^{N} -log \left(\frac{e^{cos(f_i^{lang}, f_i^{vis})/\tau}}{\sum_{j=1}^{N} e^{cos(f_i^{lang}, f_j^{vis})/\tau}} \right) \tag{4}$$

where τ in Equations 3 and 4 denotes a temperature learnable parameter, $cos(\cdot, \cdot)$,i.e, mean cosine similarity.

3.4.2 Circle Loss

Circle loss [19] is a deep metric learning loss function that learns where the inter-class variance is maximized for a feature embedding space where the intra-class variance is minimized. It achieves by penalizing the similarity of negative pairs and by encouraging the similarity of positive pairs. Therefore, Circle loss is expressed as:

$$\mathcal{L}_{circle} = \log[1 + \sum_{j=1}^{L} exp(\gamma \alpha_n^j (s_n^j - \Delta n)) \sum_{i=1}^{K} exp(-\gamma \alpha_p^i (s_p^i - \Delta p))] \tag{5}$$

where s_p and s_n are positive and negative pairs, respectively; and coefficients α_p and α_n:

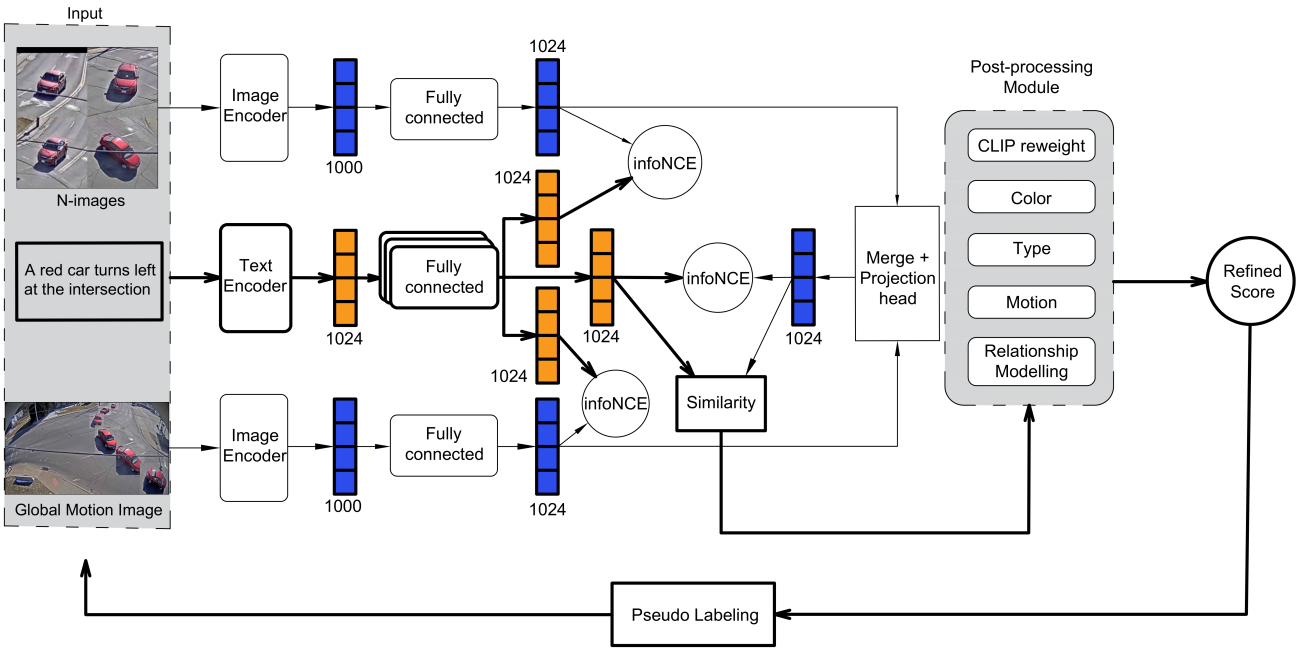

Figure 1: Proposed method architecture, comprising two branches to learn information from two aspects. The upper branch is dedicated to identifying static vehicular features, whereas the lower branch focuses on extracting dynamic features. These representations, in conjunction with the text description representation, are integrated to compute cosine similarity scores. Further refinement of this score is accomplished using a post-processing module that includes clip reweighting, attribute matching, and relationship matching; to produce the final refined score.

$$\begin{cases} \alpha_p^i = max(0, \mathcal{O}_p - s_p^i) \\ \alpha_n^i = max(0, s_n^j - \mathcal{O}_n) \end{cases} \tag{6}$$

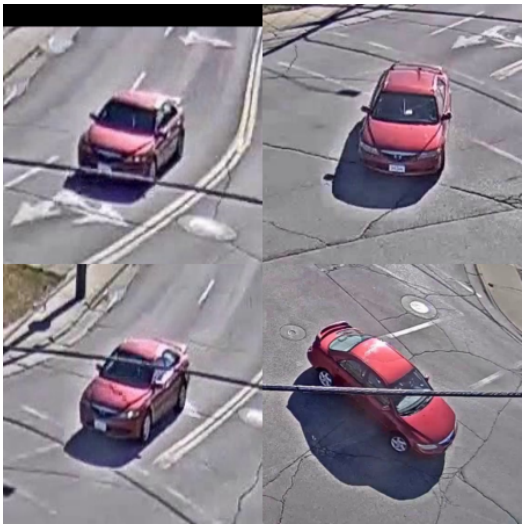

Figure 2: An example of multi-view association method. N-cropped vehicle frames are concatenated into a single representative image.

Margins are δp and Δn, where Δp is the intra-class margin and Δn is the inter-class margin. Reference values $\mathcal{O}_0$ and $\mathcal{O}_n$ can be expressed as :

$$\begin{cases} \mathcal{O}_p = 1 + m \\ \mathcal{O}_n = -m \end{cases} \tag{7}$$

and margin values were set to $\Delta p = 1 - m$, $\Delta n = m$. Values for K and L are determined by the number of classes N, K = 1 and L = 2(N - 1). These values are used to calculate the exponential values in the Circlelloss function.

3.5. Post-processing module

We apply several post-processing modules to further refine retrieval results from the model. Following [24] [25], we implement CNNs to classify attributes, along with a module to reconsider the relationships between the main vehicle and other objects in a track.

3.5.1 Fine-Grained Vehicle Feature Refinement

Following [24], the proposed approach leverages CNNs to learn both local and global vehicle attributes. In particular, we formulate three CNN-based neural networks to clas-

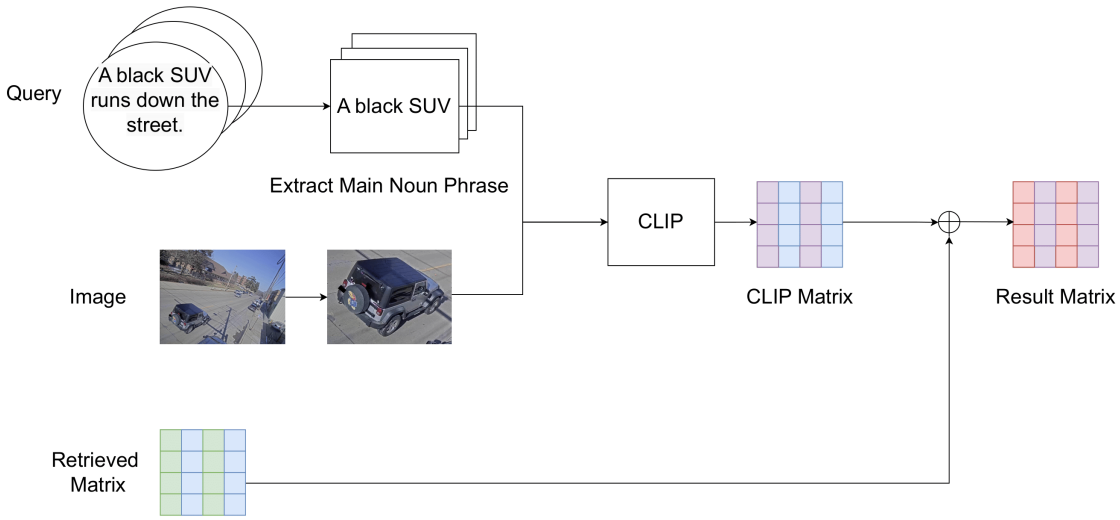

Figure 3: Proposed method using query sentences with the prefix "A photo of" to obtain representative images and embedding vectors. The similarity is subsequently calculated using cosine similarity and a hyper-parameter. This enhances local vehicle attributes and hence visual recognition performance.

sify vehicle color, type, and motion. Color and type classifiers are trained, using cropped vehicle images as input; and model, color, and type extracted from natural language queries as labels. For the motion classifier, we laid the vehicle trajectory onto the motion map and used the extracted motion to train the classifier.

We also employ a module to extract information about related vehicles using natural language and video frames. We first use an object detection module to identify potential related vehicle locations and filter irrelevant vehicles. The module calculates intersection over union (IOU) between each bounding box area and the target vehicle's trajectory mask to filter bounding boxes with IOU that is less than a threshold value. We can calculate the L2 distance as:

$$d\left(p,q\right)=\sqrt{\sum_{i=1}^{n}\left(q_i-p_i\right)^2} \tag{8}$$

where $d\left(p,q\right)$ is the distance between the main vehicle p and related vehicle q; find vehicles in front of and behind the target vehicle.

Given these relevant background details and attribute classifiers, we apply a re-weighting strategy to generate score matrices for front and back vehicles, including scores for color, type, and main vehicle motion. Score matrices are then added together (using a coefficient) to obtain the refined similarity matrix. Thus, we can refine retrieval results from the model.

3.5.2 CLIP feature re-weighting

We propose a strategy utilizing the contrastive language-image pre-Training (CLIP) model [16] to enhance local vehicle attributes using a re-weighting approach (as shown in Fig. 3). In particular, the proposed method extracts the main subject from each query sentence, then includes the prefix "A photo of" before the corresponding noun phrase. This creates the appropriate query sentence, such as "A photo of a red sedan" or "A photo of a brown pickup truck.". Representative cropped images are subsequently obtained for each vehicle track, and embedding vectors are derived for both images and queries through the CLIP model. The similarity between queries and tracks is then calculated using cosine similarity, and the resulting matrix is adjusted with a hyper-parameter coefficient before being added to the original similarity matrix obtained from the model. Leveraging this approach will effectively enhance local vehicle attributes and improve visual recognition performance.

3.5.3 Long and Short-Distance Relationship Modeling

The proposed approach uses long-distance relationship modeling [25] to model the relationships between text and images. This technique gives a higher similarity score for all identified intersections described in both text and image terms. Vehicle location is calculated in each frame using the bounding box, and vehicle movement is determined using frame coordinates. Vehicles are determined to be located at the intersection of vehicle movement distance in n frames are all zero. Visual and text location embeddings

are obtained, and their similarity is calculated using the dot product to obtain similarity matrix S_l.

We extract all noun phrases in the sentence and keep all phrases describing the vehicle to model relationships between multiple vehicles close together in the video frame. We then extract language embeddings for all the vehicle descriptions and use detection files to extract bounding boxes for all cars in the frame. This creates a matrix S_r, representing relationship similarity between each query and track.

$$S_{final} = S + \alpha S_l + \beta S_r \tag{9}$$

where S_{final} is the final similarity matrix; S_l is the location similarity matrix; S_r is the relationship similarity matrix; α and β are hyperparameters that set the importance of each similarity matrix and we set $\alpha = 1$, $\beta = 0.3$, as recommended in the paper [25].

4. Experiments

4.1. Dataset and Evaluation metrics

The proposed model was trained and tested using the CityFlow-NL dataset [14], which was already divided into distinct training and testing sets. The training set has 2,155 identical vehicle tracks. Each sample comprises three descriptive sentences, whereas the testing set consists of 184 unique vehicle tracks.

We evaluated model results for each track using the mean reciprocal rank (MRR). The reciprocal rank for a query response is the multiplicative inverse of the rank of the first correct answer, and MRR is the average reciprocal rank over the entire test set:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{10}$$

where N is the number of descriptive sentences and $rank_i$ is the rank position for the first correct answer for the i^{th} query. We also calculated MRR scores of R@5 (Recall@5) and R@10 (Recall@10).

In the ablation study section, we use Acc@5 to choose the best number of frames in the multi-view association experiments. Top-5 accuracy means that any of the model's top 5 highest probability predicted results must match the labels.

$$Acc@5 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{y_{label}^{(i)} \in \hat{y}_{top5}^{(i)}\} \tag{11}$$

where N is the number of descriptive sentences, $y_{label}^{(i)}$ is the correct answer and $\hat{y}_{top5}^{(i)}$ is the top 5 highest probability predicted results for the i^{th} query.

4.2. Implementation

Using the proposed method, we input images for each track at 5/3 fold actual bounding box, subsequently resized to 288×288 pixel. Input text was augmented by translating into French and then back to English to increase the number of sample queries. Pseudo-labels were used to create dummy labels for the test dataset. We use the ViT-B/16 [16] weight for the CLIP re-weighting model extracting media and object features, as described in the query. We set a batch size of 40 for all inputs, and models were trained over 400 epochs with the AdamW optimizer. The learning rate was initialized at 0.01, with a standard launch strategy for 40 epochs. A step delay scheduler was subsequently applied to reduce the learning rate every 80 epochs. Parameters $\Gamma = 48$ and $m = 35$ for circle loss; whereas we set $m = 0$ for triple loss. Defined models we trained under their different configurations and gathered for final similarity calculation.

4.3. Ablation Study

4.3.1 Multi-View Association Experiments

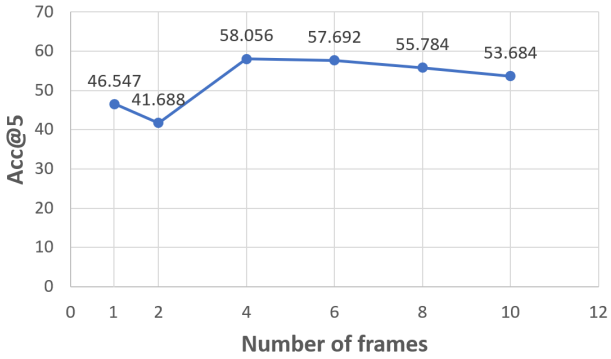

Figure 4: Multi-view association experiment results on the validation set. The proposed method (Acc@5) configured with four frames achieves the best performance.

In this experiment, we utilized three cameras: 'S05_c010', 'S05_c019', and 'S05_c020' for validation, and the remaining cameras from the CityFlow-NL dataset [14] for training. We analyzed the Acc@5 of the models on the validation set when changing N. N is the number of concatenation frames:

$$N = \begin{cases} 1 & \text{for} \quad n = 0 \\ 2n & \text{for} \quad n \geq 1 \end{cases} \text{ where n} \in \mathbb{N} \tag{12}$$

This experiment examines the model with $n \in [0, 5]$. The case when n = 0 indicates N = 1. In this case, the input is an image with only one frame, similar to the baseline [25]. In other cases, the number of association views must

be even because of the symmetric property. All concatenate frames must be equal in size and resolution. Fig. 4 shows that the optimal choice is $N_{best} = 4$ when n = 2.

4.3.2 Post-processing Module Experiments

In this section, we also use the validation set and the baseline like the experiments in Section 4.3.1

Table 1: Ablation experiments for Post-processing module. All Three Combined denote the proposed post-processing method, and "↑" denotes that larger numbers are better.

Method	MRR(↑)
Baseline	0.25564
Baseline + Feature Refinement	0.27094
Baseline + CLIP Reweighting	0.25757
Baseline + Relationship Modeling	0.26350
Baseline + All Three Combined	**0.28047**

The first row of Table 1 shows the baseline MRR score, which is 25.564%. The other rows of the table represent the MRR scores achieved when different methods were applied to the baseline system. Refining the vehicle features achieved the highest increase among the three single methods, with MRR = 27.094%. The baseline system is improved by applying CLIP reweighting, resulting in an MRR score of 25.757%. The MRR score of the baseline model combined with modeling long and short-distance relationships increased by 0.786% compared to the baseline. All three previous methods provided positive impacts on the model performance. The fifth row shows the MRR score achieved when all three methods are combined, resulting in the highest MRR score = 28.047%.

4.3.3 Model Experiments

In this section, we use the results from the evaluating system of AI City Challenge 2023 on CityFlow-NL dataset [14]. We investigate the effect of our proposed components including multi-view association (MA), pseudo labeling (PL), and post-processing (PP) in the whole proposed method.

Following the rule of the competition, the scores in Table 2 are computed on a 50% subset of the test data from 1/23/2023 to 02/20/2023. The ensemble model utilizing MA and PP achieved MRR = 0.4579. Adding PL to that model, increased MRR by 3.5%.

Table 3 shows the quantitative results of the proposed method and other testing methods using the benchmark that uses the full test set. We used [25] as the baseline model achieving MRR = 0.2931. Baseline+MA and Baseline+PL performances increased with the MRR values of 0.3480 and 0.3715, respectively. Similarly, the Baseline+MA and Baseline+PP performances overall increased using the R@5 and R@10 metrics.

Table 2: Ablation study for different models using on 50% test dataset. Ensemble+MA+PP+PL denotes the proposed method.

Method	MRR(↑)
Ensemble + MA + PP	0.4579
Ensemble + MA + PP + PL	**0.4929**

MA helps the model exploit more comprehension information from the multiple images, and PL utilizes unlabeled data and promotes consistency in the model's output. Therefore, using them gave a positive effect on Baseline.

Ensemble+MA+PP and the proposed method (Ensemble+MA+PP+PL) performed much better than Baseline, Baseline+MA, and Baseline+PL. Overall, Ensemble+MA+PP+PL had the best performance in our experiments.

Table 3: Model performance on the full test set. Ensemble+MA+PP+PL denotes the proposed method.

Method	MRR(↑)	R@5(↑)	R@10(↑)
Baseline	0.2931	0.4891	0.6793
Baseline + MA	0.3480	0.5598	0.6739
Baseline + PL	0.3715	0.5924	0.6957
Ensemble + MA + PP	**0.4795**	0.6467	**0.7826**
Ensemble + MA + PP + PL	0.4719	**0.6522**	**0.7826**

4.4. Evaluation Results

The proposed method achieved 3rd rank with MRR = 0.4795, as shown in Table 4.

Table 4: Evaluation results for various methods and other competitors using the full test set from the competition benchmark. The proposed method is named AIO-NLRetrieve in this table.

Rank	Team ID	Team name	MRR Score
1	9	HCMIU-CVIP	0.8263
2	28	IOV	0.8179
3	**85**	**AIO-NLRetrieve**	**0.4795**
4	151	AIO2022	0.4659
5	76	DUT_ReID	0.4392
6	42	Ctyun-AI	0.4286
7	101	Lighthouse	0.3544
8	210	CT_CORE	0.3444
9	197	cv_bird	0.3057
10	19	Cpay Penguins	0.0315

Fig. 5 shows the top three retrieval results of the proposed model on the CityFlow-NL dataset. The proposed

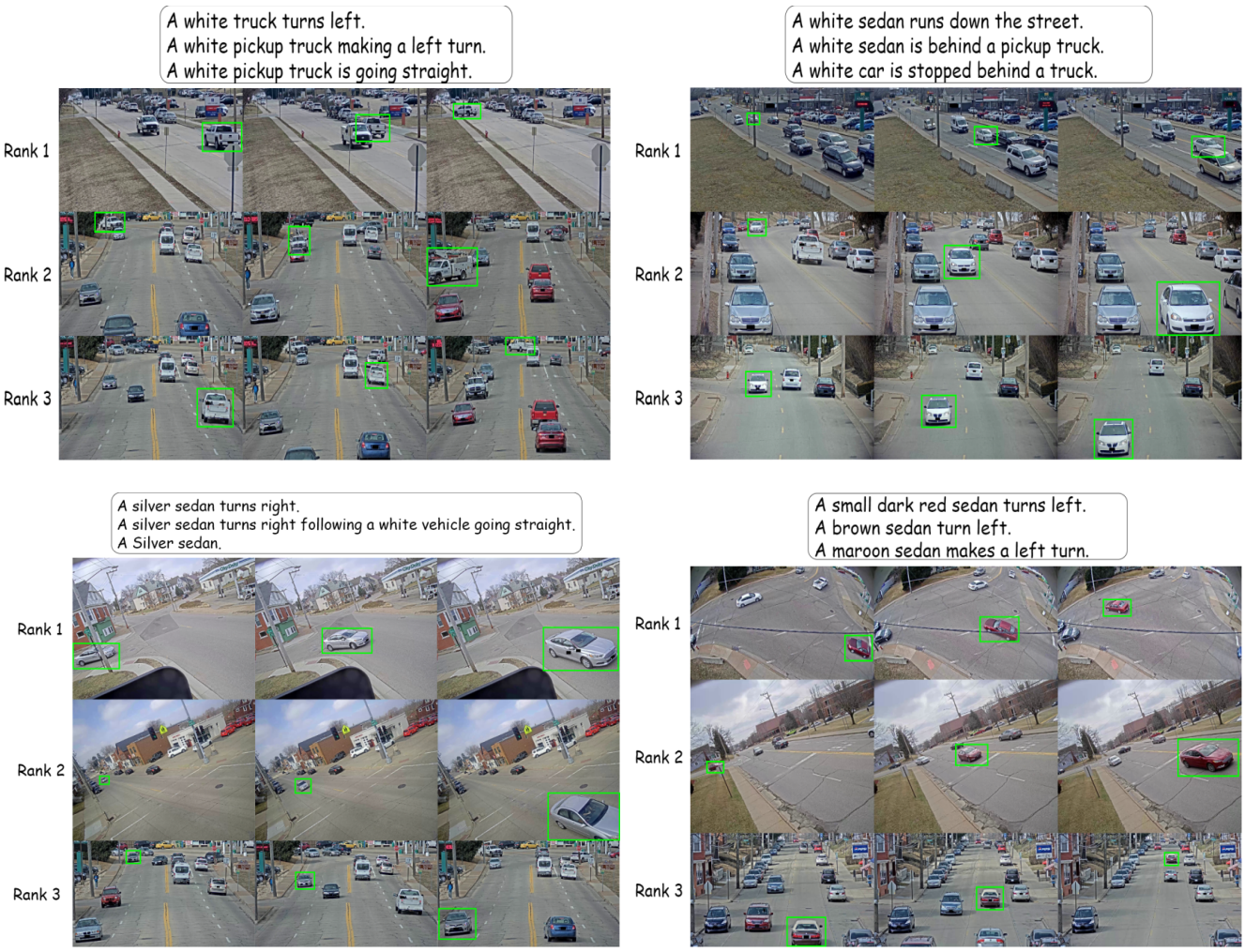

Figure 5: Visualization sample retrieval results of the proposed method on the test dataset of the CityFlow-NL dataset. Only the first 3 results of the query are listed and 3 frames were sampled for each orbit. Retrieved target vehicles are marked with green bounding boxes. The proposed model exhibits semantic understanding, focusing heavily on different aspects of the object even if it is occluded.

model can query the vehicle even though it passes behind an obstacle, i.e., becomes obscured, in the camera view. Unlike other methods that only use one image of the object as input, our proposed method Multi-view association (Section 3.2) uses 4 input images, so if the object passes through an obstacle, it will still get the comprehensive features at various angles because of benefits of multi-view association. Different degrees might eliminate the possibility that the obstacle obscures the characteristics of the vehicle. Because of the CLIP feature re-weighting (Section 3.5.2), the queried objects focus more on their external appearance, causing queries from the first 3 ranks to return correct object results which visual representations are described as in the query.

5. Conclusion

This study proposed an improved two-stream architectural framework for natural language-based vehicle retrieval. The proposed approach provides a successful pathway to increase visual input features and produced a comprehensive relationship by considering multiple input vehicle images. A pseudo-labeling step was introduced to effectively increase data diversity in training. Several post-processing techniques were applied using three efficient neural networks to classify vehicle attributes. We applied the proposed approach to the AI City Challenge 2023, achieving 3rd position in the private test with MRR = 47.95%. Works in this Challenge are summarized in the paper [10].

References

[1] Shuai Bai, Zhedong Zheng, Xiaohan Wang, Junyang Lin, Zhu Zhang, Chang Zhou, Yi Yang, and Hongxia Yang. Connecting language and vision for natural language-based vehicle retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, 05 2021. 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[3] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9338–9347, 2019. 1

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[5] Chuanping Hu, Xiang Bai, Li Qi, Pan Chen, Gengjian Xue, and Lin Mei. Vehicle color recognition with spatial pyramid deep learning. *IEEE Trans. Intell. Transp. Syst.*, 16(5):2925–2934, 2015. 2

[6] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013. 3

[7] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016. 2

[8] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2

[9] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013. 2

[10] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023. 8

[11] Vinh Dinh Nguyen, Hau Van Nguyen, Dinh Thi Tran, Sang Jun Lee, and Jae Wook Jeon. Learning framework for robust obstacle detection, recognition, and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1633–1646, 2017. 1

[12] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[13] Zhiming Pan and Hongkai Xiong. Sparse spatio-temporal representation with adaptive regularized dictionaries for super-resolution based video coding. In *2012 Data Compression Conference*, pages 139–148, 2012. 2

[14] Stan Sclaroff Qi Feng, Vitaly Ablavsky. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. *arXiv preprint arXiv:2101.04741*, 2021. 6, 7

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5, 6

[17] Xiaohan Wang Junyang Lin Zhu Zhang Chang Zhou Hongxia Yang Shuai Bai, Zhedong Zheng and Yi Yang. Connecting language and vision for natural language-based vehicle retrieval. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4034–4043, 2021. 1

[18] Pranjay Shyam and Kuk-Jin Yoon. Adversarially-trained hierarchical feature extractor for vehicle re-identification. pages 13400–13407, 05 2021. 2

[19] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020. 3

[20] Zehang Sun, George Bebis, and Ronald Miller. On-road vehicle detection: A review. *IEEE transactions on pattern analysis and machine intelligence*, 28:694–711, 06 2006. 2

[21] Ziruo Sun, Xinfang Liu, Xiaopeng Bi, Xiushan Nie, and Yilong Yin. Dun: Dual-path temporal matching network for natural language-based vehicle retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4056–4062, 2021. 1

[22] Dongyang Wang, Junli Su, and Hongbin Yu. Feature extraction and analysis of natural language processing for deep learning english language. *IEEE Access*, pages 46335–46345, 2020. 2

[23] Wei Wu, Zhang Qi-sen, and Wang Mingjun. A method of vehicle classification using models and neural networks. *IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings (Cat. No.01CH37202)*, 4:3022–3026 vol.4, 2001. 2

[24] Jiacheng Zhang, Xiangru Lin, Minyue Jiang, Yue Yu, Chenting Gong, Wei Zhang, Xiao Tan, Yingying Li, Errui Ding,

and Guanbin Li. A multi-granularity retrieval system for natural language-based vehicle retrieval. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3215–3224, 2022. 4

[25] Chuyang Zhao, Haobo Chen, Wenyuan Zhang, Junru Chen, Sipeng Zhang, Yadong Li, and Boxun Li. Symmetric network with spatial relationship modeling for natural language-based vehicle retrieval, 2022. 1, 3, 4, 5, 6, 7

[26] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: Learning robust visual representation for vehicle re-identification. *IEEE Transactions on Multimedia*, 23:2683–2693, 2021. 1