Natural Language-Based Vehicle Retrieval with Explicit Cross-Modal Representation Learning

Bocheng Xu[1*] Yihua Xiong[1*] Rui Zhang[1] Yanyi Feng[1]
Haifeng Wu[2]
[1]Department of AI R&D, Terminus Technologies, China
[2]Chongqing University of Posts and Telecommunications, China
{xu.bocheng, xiong.yihua, zhang.rui_sh, feng.yanyi}@tslsmart.com, S190231138@stu.cqupt.edu.cn

Abstract

On the account of the explosive growth in the large-scale transportation videos, vehicle retrieval plays an important role in the public transportation security and the intelligent transport system recently. Most vehicle retrieval algorithms are vision-based and consist of vehicle re-identification and vehicle tracking. However, the performance of vision-based vehicle retrieval algorithms is constrained as the limited information provided by traffic video streams. In this paper, we propose a contrastive cross-modal vehicle retrieval solution, maximizing the value of the complementation between natural language representation and vision representation. The framework of the proposed solution includes: (1) Preprocess a source video in four ways for generating local motional semantics and global motional semantics; (2) Correspondingly, preprocess relevant description sentences in two ways, including Textual Local Instance Semantics Extraction (TLISE) and Textual Local Motional Semantics Extraction (TLMSE); (3) Use a two-stream architecture model with four visual encoders and four text encoders to extract visual features and textual embeddings; (4) Fuse visual features and textual embeddings respectively by concatenating them along the feature channel in the order of importance, and use them for retrieval. By using the proposed solution, we achieved MRR score of 33.20%, ranking the 7th place in the AI City Challenge 2022 Track 2. The code is publicly available at https://github.com/Katherinaxxx/2022AICITY_T2.

1. Introduction

Vehicle retrieval is an important and active research domain in the last decade, and it has a wide range of applications in the industry. Most vehicle retrieval systems are empowered by vision-based re-identification (ReID) [1, 2, 3]. The vision-based ReID models enable to extract vehicle characteristics (e.g., appearance features) from an image or a video stream, helping the systems retrieve the top-k similar instances from a large-scale gallery by calculating the similarity among those vehicle characteristics. In practice, if traffic managers or police officers want to locate a suspicious target, they probably only have some language description about the suspicious target rather than an image or a video. To resolve this situation, a cross-modal vehicle retrieval paradigm is proposed, which focuses on how to leverage heterogeneous data from different signal domains like Computer Vision (CV) and Natural Language (NL). The paradigm is shown as Fig. 1.

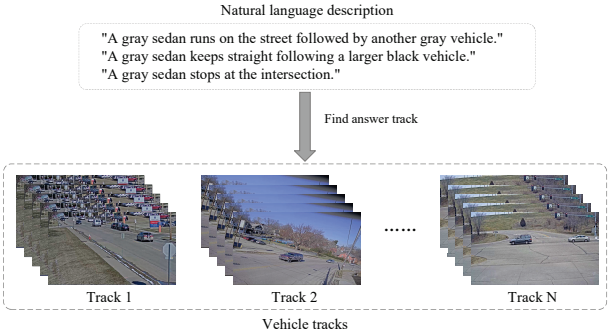

Fig. 1. Problem definition.

For the CNN-based contrastive representation methods, some works [4, 5, 6] have been successfully applied to the natural language-based vehicle retrieval. Firstly, they use a CNN-based visual encoder and a CNN-based text encoder to obtain high-level semantic visual representations and text representations respectively. Secondly, two representations are embedded into the same manifold space. Finally, two embedded representations sharing the same space are used to calculate the similarity matrix, and instances from the vehicle gallery will be ranked based on the similarity matrix.

*These authors contributed equally to this work

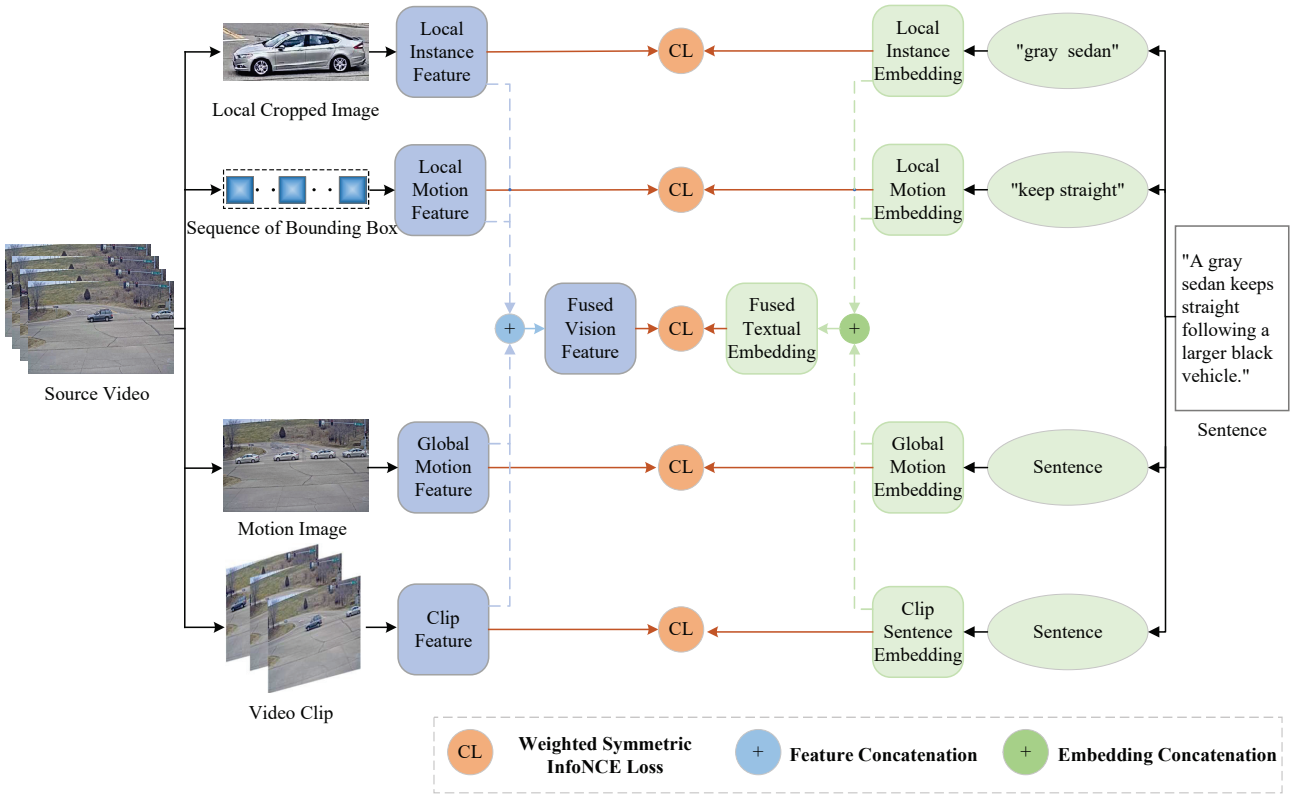

Fig. 2. The framework of our vehicle retrieval solution.

Although the CNN-based contrastive representation methods obtained acceptable results, most of them either blindly increase the number of convolution layers to pursue a slight performance improvement at the cost of increasing the computation burden. Moreover, most of them didn't fully consider the characteristics of the vehicle retrieval task in designing the network structure, which also leads to the limited improvement of vehicle retrieval performance due to the lack of context information about traffic flow and words in natural language.

Another works [6, 7, 8] break through the above bottleneck by combining other information like the images, video clips or box coordinates, to learn more explicit mutual information between NL and CV, which confirms that obtaining enhanced high-level vehicle semantics through with limited data is crucial for cross-modal contrastive learning.

Inspiring by these works, in this paper, we propose a natural language-based vehicle retrieval method based on cross-modal contrastive learning. The framework of our vehicle retrieval solution is shown as Fig. 2.

Firstly, for learning more mutual information between NL and CV, we proposed a set of effective data augmentation strategies. For local explicit information, we crop vehicles from images and track vehicles to generate single-camera trajectories. Correspondingly, we implement dependency parsing to extract the attributes and the motion descriptions of vehicles that correspond to vehicle cropped images and trajectories. For global information, we generate motion images for each vehicle by pasting cropped images of the relevant vehicle trajectory into one image. Meanwhile, we cut video streams into a set of video clips with 32 frames. Considering the context continuity of sentences, we directly use complete sentences to extract global textual features.

Secondly, we adopt a mature two-stream architecture framework, which consists of a visual encoding module and a text encoding module. Both contain four encoders respectively, and they are applied for extracting local instance features, local motional features, global motional features and clip features. Specifically, the visual encoding module is used to extract spatial-temporal visual features from images and video clips, while the text encoding module is used to extract high-level fine-grained semantic textual features from extracted attributes and a complete sentence. However, the visual features and the textual features are learned from different domains which causes difficulties to model training based on the contrastive loss without alignment. Thus, we utilize MLPs as projection heads for mapping each embedding into the sharing space.

In the end, we introduce a symmetric weighted infoNCE

loss for training, which includes text-to-vision infoNCE loss [9] and vision-to-text infoNCE loss at the same time. Besides, we add a weight discount factor on the symmetric infoNEC loss to make the global embeddings with higher weights.

The contributions of this paper are summarized as follows:

- We propose a set of effective data augmentation strategies which can build an explicit connection between NL and CV.
- For enhancing the high-level vehicle semantics, we introduce four types of features for CV and NL respectively, including local instance features, local motional features, global motional features, and clip features.
- We introduce a symmetric weighted infoNCE loss for the case where multiple embedding exists, which helps converge to the global optimum.

2. Related Work

2.1. Text-Video Retrieval

The prototypical approach to text-video retrieval is to integrate textual and visual streams through a combination of a pre-trained linguistic model and a video model typically pre-trained for various tasks and modalities, followed by late fusion. Depending on whether multi-branch networks are used for video representation, we categorize existing methods into two groups, i.e., single-branch [10] and multiple-branch methods [6]. A common implementation of single-branch methods is to first extract visual features from video frames by pre-trained CNN models, and subsequently aggregate the frame-level features into a video-level feature. For multiple-branch methods, multiple parallel video encoding branches are jointly used to represent videos. One simple way is to utilize multiple independent encoding branches with different video features as inputs [6]. MoEE [11], CE[6], MMT [12], MDMMT [5] and TeachText [13] are all such efforts. Some works benefit from pre-training their models with text images or pre-training them on large-scale text-video datasets, such as ActBERT [14] and ClipBERT [8] are single-stream models jointly embedded in text-video pairs via a BERT-like architecture for early cross-modal fusion.

Recent work by CLIP4Clip [15] and StraightCLIP [16] uses as a backbone a joint language-visual model pre-trained by CLIP [17] on a large-scale text-image dataset. Even using CLIP in a zero-beat manner exceeds most of the recent works mentioned above [16], and it is particularly noteworthy that CLIP's rich joint text-image understanding can be extended to video. CLIP4Clip [15] proposes several video aggregation schemes, including mean pooling, self-attention, and multimodal transformers, yet none of them allow text to be directly matched with its most relevant video subregions to be directly matched.

2.2. Data Augmentation in NLP

With the development of deep learning technology, the requirement for better neural network models for large-scale data has also gradually increased. For classification tasks, if the amount of data is very limited or the amount of data differs greatly between different categories, the model will be overfitted. Text augmentation is a very common method to enhance the robustness of models, and common text augmentation methods can be divided into lexical-level methods, sentence-level methods, and model-based methods.

As for lexical-level methods, synonym replacement with WordNet and word embedding substitution are very common techniques, which are widely used for text augmentation, especially for semantic similarity tasks; Back-translation is the most commonly used sentence-level method, which can sometimes change the syntactic structure, and retain semantic information compared to replacement words. And back-translation tends to increase the diversity of text data. However, the data produced by back-translation depends on the quality of the translation methods, and some of which may not be as accurate as we assume.

Model-based methods can be classified as semi-supervised and unsupervised methods. Semi-supervised and unsupervised learning methods were proposed to make better use of unlabeled data and alleviate the dependence on large-scale labeled datasets, and have proven to be a powerful learning paradigm. For example, MixMatch [18] mixes unlabeled data with labeled data by approximating the low entropy labels of unlabeled samples generated by the means of MixUp. UDA [19] filters the augmented data similar to the original data by training classifiers.

3. Method

In the Natural Language-based vehicle retrieval system, it is important to build a connection between visual knowledge and textual knowledge. Bai et al. [20] proposed a framework that directly used global textual embedding and fused visual features on cross-modal representation learning. However, we consider that the local textual description like "gray sedan" is equally important. Furthermore, building a mutual connection between local visual features and local textual embeddings is a more fine-grained method for improving model performance. Based on the ideas mentioned, we develop a set of data augmentation strategies and construct a dual-stream network that can learn cross-modal knowledge effectively.

Sentence	Type and Color	Motion
A gray sedan stopped at the intersection.	gray sedan	stop at intersection
A red wagon drives straight through the intersection.	red suv	drive straight
A wine-red SUV runs down the street not yielding to a pedestrian.	red suv	run down the street

Table 1. A example of text attributes extraction.

3.1. Data Augmentation

Although there are many kinds of vehicles on the road in reality, we can mainly describe a target from three aspects: appearance characteristics, motion trajectories, and surrounding environment. For the first aspect, most people usually use colors and types of vehicles to describe them, such as "a gray sedan". For the second aspect, when describing a vehicle's motion, people tend to use terms like "turn left". For the last aspect, making a detailed description with the help of the surrounding environment is an efficient way of locating the target rapidly. For example, we can easily locate the target from traffic flow with the sentence "A gray sedan keeps straight following a larger black vehicle" on the account of the given "larger black" reference. According to the three aspects mentioned, we develop some data augmentation strategies for sentences and source videos.

3.1.1 Textual Data Augmentation

Sentences contain higher-level semantics compared with images, and more intuitive information e.g. the characteristics and motion of vehicles. However, it is still difficult to learn the representation of target vehicles and retrieve their trajectories correctly with limited sentences. The situation will be worse when the description sentences are ambiguous, inaccurate, or extremely similar.

According to the problems mentioned above, we develop a three-stage textual augmentation strategy:

The first stage is spelling correction. We find that there exists some spelling errors in both training and test set. They will be recognized as "UNK" causing a negative influence on the model. To ensure the accuracy of description sentences, we correct some spelling errors. For example, we change the wrong word "mint" to "mini", "SVU" to "SUV".

The second stage is information extraction shown as Table 1, including Textual Local Instance Semantics Extraction (TLISE) and Textual Local Motional Semantics Extraction (TLMSE). Specifically, we find that the description sentences are mostly of short length and without complex syntactic structure. Furthermore, the target vehicle is the subject of most description sentences, while the motion of the target vehicle is the verb. Therefore, we propose the TLISE and the TLMSE based on dependency parsing. For the TLISE, we use "nltk" to extract the subject and the adjective modifying the subject in each sentence as the type and color of the target vehicle, respectively. For the TLMSE, we extract the verb and its complement as the motion.

The third stage is disambiguation. In this dataset, we observe that there are some descriptive differences and diversity in the sentences when describing the same trajectory. For example, some people use the present tense, but others tend to use the present progressive tense. Besides, as we know, there are many words to describe colors. When it comes to a red car, some people may use "wine-red" instead of "red". To reduce the impact of these differences and diversity, we further normalize the extracted attributes. First, we summarize the types and colors of the extracted vehicles into mapping tables separately. We use these tables to normalize the extracted type and color of vehicles, for example, "suv", "wagon", "mpv" are mapped to "suv", "red", "maroon", "reddish" are mapped to "red". Then we change all the verbs of the motion to the original verb form and change all the nouns to lowercase.

In addition, we find that most of the sentences in the training set are repeated. If we only use "nl" sentences to train the model, the small quantity of training data will easily lead to overfitting. Therefore, we use both "nl" and "nl_other_view" sentences for data augmentation and training, in which we use the reciprocal of the number of occurrences of each sentence as the sampling weight.

3.1.2 Visual Data Augmentation

For source videos, we mainly divide the augmentation into four parts:

Firstly, we crop instances from random frames according to their bounding boxes. The cropped images mainly contain appearance information of instances, which correspond to the short appearance description extracted by TLISE. Thus, the subset of cropped images can be used to train the Local Instance Encoder from the visual encoding module, which focuses on extracting instance features.

Secondly, we track vehicle instances and generate the corresponding single-camera trajectories. The trajectories are composed of sequences of bounding boxes. Similarly, the trajectory sequences correspond to the short motion description extracted by TLMSE. The sequences of bounding boxes are used to train the Local Motion Encoder from the sequences of bounding boxes, which focus on extracting in-

stance motion features.

Thirdly, in order to learn semantics from global motion and the surrounding environment, we generate motion images for each vehicle by pasting cropped images of the relevant vehicle trajectory into the same background image.

Finally, in order to learn special-temporal information, we cut video streams into multiple video clips. For each video stream, the first and last frames of the source video are retained, and the remaining frames are evenly sampled to obtain 32 frames of video clips. The clips lacking enough frames are padded by all-zero frames.

3.2. Explicit Cross-Modal Knowledge Learning

Cross-modal feature extraction plays a key role in the whole NL-based vehicle retrieval framework. In order to learn fine-grained knowledge from cross-modal data generated by our data augmentation, we construct a dual-stream architecture model which consists of a visual encoding module and a textual encoding module. Each encoding module has two local encoders and two global encoders. Moreover, we utilize MLPs as projection heads to map each module's output into the sharing space.

3.2.1 Fused Visual Features

For visual features, we construct a visual encoding module with Local Instance Encoder, Local Motion Encoder, Global Motion Encoder, and Clip Encoder.

The Local Instance Encoder and Global Motion Encoder are CNN-based models, on account of CNN-based models being able to provide more robust spatial features. We use SE-ResNeXt 101 model [21] or EfficientNet B3 model [22] pretrained on ImageNet [23].

RNN-based models and Transformer are good at processing temporal information. Considering the temporal characteristic of motion information, we use Bi-directional Long Short-Term Memory (BiLSTM) [24] for Local Motion Encoder, and use Video Swin Transformer [25] or video encoder in VideoClip [26] for Clip Encoder. The Local Motion Encoder can well model the context information of vehicle trajectories based on the sequences of bounding boxes, and the Clip Encoder can provide rich spatial-temporal features that are benefited from the Transformer architecture.

In addition, for each four visual encoders, we utilize projection heads for mapping each encoder's output into the spaces of contrastive representation learning. The projection head uses a MLP can be expressed as:

$$e_{vis} = g_i(h_i) = W_i\sigma(BN(W_i h_i)), \tag{1}$$

where $BN(\cdot)$ is a Batch Normalization (BN) layer, σ is a ReLU layer. W_i is a fully connected layer and the output dimension is 512. h_i is the visual features extracted by the visual encoders, which correspond to local feature, local motion feature, global motion feature and clips feature.

3.2.2 Fused Textual Embeddings

As for textual embeddings, we routinely use pretrained transformer-based models, i.e., RoBERTa [27] or DeBERTaV3 [28], as the Textual Local Instance Encoder, the Textual Local Motion Encoder, and the Global Motion Encoder. They focus on extracting textual embeddings from appearance description, motion description and the complete sentence description, respectively. Meanwhile, we use a pretrained encoder from VideoClip to encode description sentences as one of global text encoders. Equally, we map all textual embeddings into the sharing space by the projection heads. The projection head uses a MLP can be expressed as:

$$e_t = g_t(h_t) = W_t\sigma(LN(W_t h_t)), \tag{2}$$

which is textual embedding, where h_t is last hidden state extracted by the text encoder and $LN(\cdot)$ is a Layer Normalization (LN) layer. W_t is another fully connected layer and the output dimension is 512.

3.2.3 Representation Learning on Fused Cross-modal Features

After visual feature extracting and textual embedding extracting, we can obtain a set of cross-modal features pairs S_{cross}:

$$S_{cross} = \begin{bmatrix} \{E_{vi}^{local}, E_{ti}^{local}\} \\ \{E_{vm}^{local}, E_{tm}^{local}\} \\ \{E_{vm}^{global}, E_{tm}^{global}\} \\ \{E_{vc}^{global}, E_{tc}^{local}\} \end{bmatrix} \tag{3}$$

where $\{E_{vi}^{local}, E_{ti}^{local}\}$ is a pair of local instance features from visual encoding module and textual encoding module. Equally, $\{E_{vm}^{local}, E_{tm}^{local}\}$ is a pair of local motion features, $\{E_{vm}^{global}, E_{tm}^{global}\}$ is a pair of global motion features, and $\{E_{vc}^{global}, E_{tc}^{local}\}$ is a pair of clip features.

Intuitively, the global features are supposed to contain more information than the local features. Therefore, we fuse all visual features and textual embeddings respectively in the order from local to global, to obtain the fused visual features and the fused textual embeddings.

For representation learning, in Fig. 2, we not only calculate the contrastive losses between features in each cross-modal features pairs, but also the contrastive loss between fused visual feature and fused textual embedding. Particularly, the contrastive losses of local instance cross-modal feature pairs and local motion cross-modal feature pairs, provide an explicit supervised signal to the models, as there is almost a one-to-one correlation between cross-modal information in each pair.

3.3. Symmetric Weighted InfoNCE Loss

To learn the representation of visual information and textual descriptions, we perform symmetric weighted InfoNCE loss to achieve well-aligned between cross-modal features mentioned above.

Given a batch of N text-vision embedding pairs, which is consist of M embeddings, i.e., $e_{t,j,\gamma}, e_{vis,i,\gamma}, (i,j \in N, \gamma \in M)$.

Therefore, the batch consists of NxN sample pairs, in which only one pair is positive and others are negative. To maximize the cosine similarity of the text and vision embeddings, we first utilize symmetric infoNCE loss, which is consist of two parts, text-to-vision and vision-to-text. Suppose τ denotes a temperature learnable parameter initialized with 1 and cos denotes the cosine similarity.

The text-to-vision infoNCE loss is formulated as:

$$L_{t2vis,\gamma} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(e_{vis,i,\gamma}, e_{t,i,\gamma})/\tau)}{\sum_{j=1}^{N}\exp(\cos(e_{vis,i,\gamma}, e_{t,i,\gamma})/\tau)} \tag{4}$$

The vision-to-text infoNCE loss is formulated as:

$$L_{vis2t,\gamma} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(e_{t,i,\gamma}, e_{vis,i,\gamma})/\tau)}{\sum_{j=1}^{N}\exp(\cos(e_{t,i,\gamma}, e_{vis,i,\gamma})/\tau)} \tag{5}$$

The symmetric InfoNCE loss is formulated as:

$$L_{S,\gamma} = \frac{L_{t2vis,\gamma} + L_{vis2t,\gamma}}{2} \tag{6}$$

Since the embeddings of each pair are of different importance, we optimize the symmetric InfoNCE loss by adding an increasing weight discount factor, which makes the latter embeddings with greater weights. The weighted symmetric InfoNCE loss is formulated as:

$$L_{SWNCE} = \sum_{\gamma=1}^{M}\alpha^{M-\gamma}L_{S,\gamma} \tag{7}$$

where α is the weight discount factor, we set $\alpha = 0.5$. As mentioned above, we have five kinds of pairs of visual features and textual embeddings, we set $M = 5$.

4. Results

4.1. Dataset

The CityFlow-NL dataset consists of a training set and a test set, containing 2155 and 184 trajectories respectively, and each trajectory was annotated with three natural language descriptions tagged as "nl". It is worth noting that the training set also provides several additional descriptions for each trajectory, tagged as "nl_other_views". As we have mentioned in subsection 3.1.1, there exists noise in both training and test set. To avoid the influence of noise data, we use the mean textual embeddings to find the most similar samples.

During the training process, we deploy 10-fold cross validation on the training set to reduce overfitting.

4.2. Evaluation Metics

In the leaderboard, the Natural Language-Based Vehicle Retrieval task uses the mean reciprocal rank (MRR) [29] as the main evaluation metric. MRR is formulated as

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{8}$$

where $rank_i$ refers to the rank position of the correct trajectory for the i-th text descriptions, and Q is the set of description sentences. Recall@5, Recall@10 are also evaluated for all submissions

In addition, we also consider Recall@1 to choose the best version during training process.

4.2.1 Comparsion with Other Teams

As shown in Table 2, the proposed method currently rank 7th on the private test published by the organizers, with a MRR score of 0.3320. Moreover, the consistent performance on all Test datasets demonstrates the effectiveness and robustness of the proposed method.

Rank	Team ID	Team Name	Score
1	176	Must Win	0.6606
2	6	Thursday	0.5251
3	4	HCMIU-CVIP	0.4773
4	183	MegVideo	0.4392
5	91	HCMUS	0.3611
6	44	P & L	0.3338
7	**10**	**Terminus-AI**	**0.3320**
8	41	MARS_WHU	0.3205
9	24	BUPT_MCPRL_T2	0.3012
10	56	folklore	0.2832

Table 2. The private test result.

4.2.2 Abation Study

As illustrated in Table 3, we conduct ablation studies with different modules of our proposed method. "Baseline" represents the use of ResNet50 as local instance encoder and global motion encoder, and RoBERTa as text encoders. "Local motion & TLMSE & TLISE" donates the three-stream architecture with local cropped image, sequence of bounding boxes and motion image. "Symmetric Weighted InfoNCE Loss" represents the use of Symmetric Weighted InfoNCE Loss as objective function. "Video clip" provides the four-stream architecture with local cropped image, sequence of bounding boxes, motion image and video clip. The introduction of Video clip has achieved a relative MRR

Method	Performance						
Baseline	✓	✓	✓	✓	✓	✓	✓
Local motion & TLMSE & TLISE		✓	✓	✓	✓	✓	✓
Symmetric Weighted InfoNCE Loss			✓	✓	✓	✓	✓
Video clip				✓	✓	✓	✓
Large Size & Model					✓	✓	✓
10-fold cross validation						✓	✓
Ensemble							✓
MRR(%)	20.05	21.09	21.81	24.07	27.97	31.76	33.20

Table 3. Ablation Study on TestA in the online evaluation system.

improvement of 22.6%, which indicates that the global motion feature obtained by Video clip plays a vital role in natural language-based vehicle retrieval. "Large Size&Model" means using a larger pretrained model, such as DeBERTa as text encoders and SE-ResNeXt 101 as local instance encoder and global motion encoder, the input size of the image is increased from 288×288 to 320×320, which proves the great importance of the input size and parameter quantity. "10-fold cross validation" shows that MRR is significantly improved after training with cross validation, indicating that it can reduce overfitting caused by data noise and the small amount of data. Finally, through model ensemble, we improve the baseline from 21.09% to 33.20% mAP MRR on the test set.

5. Conclusion

In this paper, using cross-modal contrastive learning, we propose a robust two-stream architecture framework to learn the vehicle and textual representations for the text-vehicle retrieval task. To establish an explicit connection between CV and NL, we propose an effective data augmentation pipeline. Further, we design a two-stream architecture model with four visual encoders and four text encoders to efficiently extract visual features and textual embeddings, and apply the four types of features for CV and NL respectively to enhance the high-level vehicle semantics and model robustness. Finally, we competed in AICity Challenge 2022 and achieves 33.20% MRR accuracy and reaches the 7th place in the private test.

Future work will continually explore the efficient text-video retrieval methods for the intelligent transportation system. Besides, more investigation will be made to discover powerful and stable optimization objectives, Additionally, we will utilize more elaborate multi-modal fusion architectures to enhance the learning of text-video representation.

References

[1] Yan Tian, Tao Chen, Guohua Cheng, Shihao Yu, Xi Li, Jianyuan Li, and Bailin Yang. Global context assisted structure-aware vehicle retrieval. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1

[2] 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, 2020. 1

[3] Qiantong Xu, Ke Yan, and Yonghong Tian. Learning a repression network for precise vehicle search. *arXiv preprint arXiv:1708.02386*, 2017. 1

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1

[5] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021. 1, 3

[6] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 1, 2, 3

[7] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[8] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 2, 3

[9] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 3

[10] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 3

[11] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 3

[12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 3

[13] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 3

[14] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 3

[15] 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. *arXiv preprint arXiv:2104.08860*, 2021. 3

[16] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021. 3

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[18] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[19] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 3

[20] Shuai Bai, Zhedong Zheng, Xiaohan Wang, Junyang Lin, Zhu Zhang, Chang Zhou, Hongxia Yang, and Yi Yang. Connecting language and vision for natural language-based vehicle retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4034–4043, 2021. 3

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[22] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[24] Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012. 5

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021. 5

[26] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 5

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[28] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. 5

[29] Ellen M Voorhees et al. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82, 1999. 6