Tracked-Vehicle Retrieval by Natural Language Descriptions With Multi-Contextual Adaptive Knowledge

Huy Dinh-Anh Le[1,2], Quang Qui-Vinh Nguyen[1,2], Duc Trung Luu[1,2],
Truc Thi-Thanh Chau[1,2], Nhat Minh Chung[1,2], Synh Viet-Uyen Ha[1,2,*]

[1] School of Computer Science and Engineering, International University, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper introduces our solution for Track 2 in AI City Challenge 2023. The task is tracked-vehicle retrieval by natural language descriptions with a real-world dataset of various scenarios and cameras. Our solution mainly focuses on four points: (1) To address the linguistic ambiguity in the language query, we leverage our proposed standardized version for text descriptions for the domain-adaptive training and post-processing stage. (2) Our baseline vehicle retrieval model utilizes CLIP to extract robust visual and textual feature representations to learn the unified cross-modal representations between textual and visual features. (3) Our proposed semi-supervised domain adaptive (SSDA) training method is leveraged to address the domain gap between the train and test set. (4) Finally, we propose a multi-contextual post-processing technique that prunes out the wrong results based on multi-contextual attributes information that effectively boosts the final retrieval results. Our proposed framework has yielded a competitive performance of 82.63% MRR accuracy on the test set, achieving 1st place in the competition. Codes will be available at https://github.com/zef1611/AIC23_NLRetrieval_HCMIU_CVIP

1. Introduction

In the era of data-driven technology, the concept of a smart city has garnered significant attention as a means to improve the quality of life for citizens. One crucial element of a smart city is vehicle retrieval systems, which utilize data from surveillance cameras to address various transportation-related issues such as reducing traffic congestion, improving traffic flow, and creating a sustainable and efficient transportation system. While vision-based vehicle re-identification has been a significant advancement in this field, text-based systems are also gaining popularity due to their ability to provide semantically concise and clear information such as speed, direction, location, branch, color, size, etc. Text-image retrieval is a cross-modal retrieval task that involves learning modality representations and their shared embedding space to obtain latent features. Therefore, most studies in the language-vision retrieval task focus on improving the learning of embedding feature vectors to achieve higher accuracy in representation matching. However, recent advancements in deep learning frameworks have shifted the focus towards discovering semantic concepts in the language and vision input, rather than solely relying on matching and ranking algorithms.

In order to advance research of text-to-image retrieval systems, the 7th AI City Challenge [16] has established a challenging research track to promote active participation. Despite promising results in the past, major challenges remain. Firstly, natural textual data is highly diverse and presents significant challenges to machines. While text data is intuitive for humans, it is difficult for machines to distinguish different descriptions of the same vehicle, such as "A vehicle is moving straight" and "The vehicle is heading forward". The limited amount of training data further exacerbates this issue in learned models. Secondly, there is a significant scarcity of high-quality training data for text-to-image vehicle retrieval, as it is a relatively new domain. Unlike established datasets such as ImageNet [20] and COCO [13], which contain millions of samples for feature training, manual annotations for this domain are limited. Effective models must therefore leverage pre-trained parameters as much as possible and use the few available labels for fine-tuning. Finally, while existing state-of-the-art methods are useful, their performances are largely probabilistic in the high-dimensional text-to-image domain. Consequently, prediction outputs may lack proper constraints to accurately match a query with its true video. Addressing these technical challenges will require continued innovation

*Corresponding author. Email: hvusynh@hcmiu.edu.vn

and collaboration in the field.

Therefore, the main contributions of our paper are stated as follows:

- Our study introduces an enhanced version of our baseline vehicle retrieval model using CLIP [19], which incorporates both symmetric InfoNCE Loss [22] and Circle Loss [21] to interconnect the text and image modalities. By projecting both modalities into a shared representation space, our proposed approach ensures that they are aligned and linked together.
- Building upon our previous research [12], we present an enhanced version of the Semi-Supervised Domain Adaptive (SSDA) training strategy. Our approach aims to mitigate the domain gap that arises between the training and testing sets, and additionally addresses the distribution shift problem that had previously occurred in prior research.
- We have developed a unique approach called multi-contextual pruning that enhances retrieval results by utilizing multiple contextual pieces of information as stringent constraints. This approach refines the final results to differentiate between vehicle tracks that appear highly visually identical.
- Experiments show that our system achieves 1st place on the testing set of the challenge.

2. Related Works

2.1. Video Retrieval by Natural Language

The text-to-video retrieval task involves mapping text and video representations into common embedding spaces and measuring cross-modal similarity. Text can be encoded using conventional text encoders or transformer-based architectures, such as BERT [4], which have demonstrated superior performance. Visual features are extracted from video frames using pre-trained CNN models and combined into video-level features. The ViLBERT [15] and UNITER [3] models use a shared transformer for image-text joint representations, while a multi-modal transformer is proposed by Gabeur et al. [9] compute three matching text-video similarities using Object, Activity, and Place experts. Current research mainly focuses on learning cross-modal similarity between text and video encodings.

2.2. Contrastive Representation Learning

Contrastive representation learning has been a popular research topic in the field of machine learning and computer vision in recent years. A number of works have been proposed that explore different approaches to improving the performance of contrastive learning methods. One notable approach is SimCLR [2], which introduced a simple yet effective framework for contrastive learning. It leverages data augmentation techniques and a large batch size to improve the quality of learned representations. Another approach is MoCo [11], which uses a momentum-based update rule to improve the stability and performance of contrastive learning. It also introduces a dynamic dictionary to improve the quality of negative samples used during training. BYOL [10] is a recent method that focuses on learning representations in a self-supervised manner without negative samples. It uses a target network to generate predictions for augmented views of the input, and optimizes the model to match the target network's predictions. Contrastive Language-Image Pre-training (CLIP) [19] is a multi-modal approach that learns representations for both images and text. It uses a contrastive loss such as InfoNCE [22] to encourage the model to learn representations that are predictive of both modalities. Regarding metric optimization criteria, Circle Loss [21] is popularly seen as a general unification of the pairwise similarity optimization.

Overall, these works highlight the importance of contrastive representation learning in improving the quality of learned representations and enabling effective transfer learning. In our works, we propose to adopt Circle Loss onto the CLIP multi-modal framework.

2.3. Tracked Vehicle Retrieval by Natural Language

In the 6th AI City Challenge [17], various approaches based on ReID approaches were introduced by teams to learn representative motion and visual appearance features. In particular, all teams involved in the text-to-image retrieval task utilized InfoNCE losses during training. They also used pre-trained sentence embedding models, such as BERT [4], CLIP [19], to represent NL descriptions. A team proposed a multi-granularity loss function [7] that formulated the ReID problem by using a pair-wise InfoNCE loss between NL streams and visual streams.

Meanwhile, post-processing of retrieval results were based on the keywords of relations and motions in the NL descriptions by participating teams to further improve the retrieval results. One team [24] utilized an NL parser to extract information about the color, type, and motion of tracked-vehicles. Meanwhile, other teams (namely, [7], [26], [24]) used the global motion image originally presented by Bai et al. [1] to create a vehicle motion stream. Additionally, the Megvii team [24] developed a refined motion image that was based on the inter-frame intersection over union (IoU) of the tracked targets.

In our approach, we focused not only on representation learning of multimodality, but also on how to post-process and prune based on the keyword extractions effectively, which were the main difficulty for matching the vehicle gallery with the NL queries. Furthermore, we adopt the post-processing and pruning processes into the multimodal learnig process.

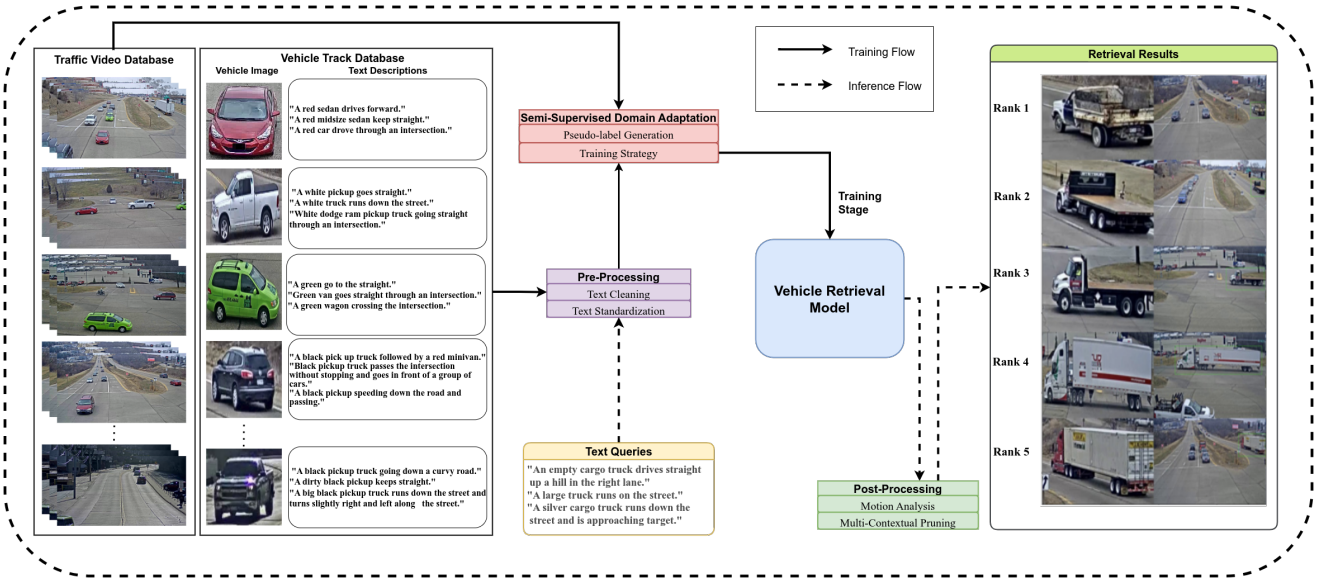

Figure 1. **Overview of our proposed system**

3. Methodology

3.1. Overview

In this section, we present our proposed multi-contextual adaptive knowledge-based retrieval system shown in Fig. 1 It is composed of four main modules, each addressing the task's challenges at varying levels of detail, and achieving remarkable performance on the benchmark dataset. The four modules are (1) Pre-processing, (2) Baseline vehicle retrieval model, (3) Semi-Supervised Domain-Adaptive (SSDA) training strategy, and (4) Multi-contextual post-processing.

3.2. Pre-processing

Due to the presence of different forms of words in a text, such as plurals, and tenses, different texts can have the same semantic meaning but provide a different wealth of information. However, the diversity in the queries can lead to degradation in the performance of the retrieval model due to a lack of consistency. To solve the aforementioned problems, we propose a text cleaning step to utilize Stemming and Lemmatization, common techniques in Natural Language Processing (NLP), to derive related word forms to a common base form to alleviate the generalization of the model's knowledge during the learning stage.

Text Cleaning. With each text description, we utilize stemming and lemmatization, common techniques in Natural Language Processing (NLP) where stemming involves cutting off the end of a word to obtain its root form, while lemmatization involves reducing a word to its base form using knowledge of the language and its grammar. Using those two techniques, we performed stop word cleaning, misspelt word correcting and converting all of them to their base forms to ensure identical text format structure.

Text Standardization. To tackle the linguistic ambiguity in the language query and reduce the variance in textual embeddings for the learning process, a consistent format is needed for every text description. Based on our observations, the text descriptions often have three contextual attributes about the vehicle, which are color, type and movement. Thus, we use English PropBank Semantic Role Labeling (SRL) [5] to extract all of the aforementioned information and propose a new standardized format $t_{standardized} = a_c + a_t + a_m$ for the natural language queries where a_c, a_t, a_m denotes the attribute vehicle's color, type, motion, respectively. Through data analysis, we discovered that many words have different synonyms but express the same semantic meaning. Thus, to minimize the diversity between different text descriptions, several clusters are created based on semantic similarity and replaced with their cluster name. i.e. synonyms for brown: brownish, bay, or beige; or synonyms for coupe: mini cooper, couple, or coup. The same goes for vehicle movements, which we categorize into four movements based on their trajectory: go straight, turn left, turn right and stop.

3.3. CLIP-based vehicle retrieval baseline model

Problem Formulation. Given a collection of n traffic video event clips $V = \{v_1, v_2, \ldots, v_n\}$ and a corresponding text descriptions database $Q = \{q_1, q_2, \ldots, q_n\}$, we aim to discover a function $\mathbf{s}(v_i, q_j)$ such that $q_j = \{q_j^1, q_j^2, \ldots, q_j^m\}$ is the set of m corresponding text de-

Original	A white truck drives straight through the intersection in the left lane.
Clean	white truck drive straight through the intersection in the left lane.
Standardized	white pickup-truck go straight.

Table 1. The example of the text standardization

scriptions, and each clip v_i is annotated with the bounding box coordinates of the tracked-vehicle as $B(v_i) = \{b_1, b_2, \ldots, b_{|v_i|}\}$ over the video length $|v_i|$. In particular, each tracked-vehicle v_i has a set comprising 3 NL-text descriptions. The primary goal of this problem is to retrieve video v_i from V based on $q_j = \{q_j^1, q_j^2, q_j^3\}$ from Q. Consequently, the main objective is to focus on maximizing the similarity $\mathbf{s}(v_i, q_j)$ between v_i and it is corresponding q_j while simultaneously minimizing the similarity of $\mathbf{s}(v_i, q_k)$ with v_i against all other queries in $Q, q_k \neq q_j$.

3.3.1 Model Architecture

Backbone. Choosing a proper backbone as the feature extractor is vital for obtaining robust embeddings because they contain abstract features disentangled from varying degrees of inessential variations, thereby making them more generalizable for text-image retrieval tasks. As a result, CLIP is used as our primary backbone to leverage its potent knowledge in constructing robust representations for feature extraction tasks with the pre-trained models Vision Transformer [6] as the Image Encoder $f_i(\cdot)$, and a Text Transformer [23] as the Text Encoder $f_t(\cdot)$.

Visual Embeddings. The visual input for each vehicle track v_i is represented by a randomly extracted j^{th} frame also known as global image and its corresponding set of bounding box coordinates $B(v_i)$. To enable dual-stream inputs, each track is constructed by incorporating a global image and a local image where we denote j^{th} global image in v_i as I_g^j, which is the original frame, and the local image I_l^j, which capture the visual global and local features, respectively. The original frame of the vehicle track serves as the global image. In contrast, the local image is obtained by cropping the corresponding global image using the ground truth bounding box of the primary vehicle. A shared-weights image encoder $\mathbf{f_i}(\cdot)$ encodes both streams of images to obtain are global and local visual feature embeddings, respectively.

Textual Embeddings. In section 3.2, pre-processed text descriptions are utilized as the textual input for the system. To obtain textual feature vectors, one of the three text descriptions q_i^j associated with each vehicle track is randomly selected and tokenized. A text encoder $\mathbf{f_t}(\cdot)$ then encodes the NL description to obtain textual feature embeddings.

Projection Head. To map each modality representation into the same space, each text and image representation is then fed into each separated projection head $\mathbf{g_v}(\cdot)$ and $\mathbf{g_t}(\cdot)$, to map each embedding from its domain space into a shared latent space where contrastive learning is applied. Visual feature vector $\mathbf{z_v}$ and textual feature vector $\mathbf{z_t}$.

3.3.2 Loss Functions

Contrastive Loss. Given a batch B pairs of video vehicle track v_i and text query q_i, we want to learn representations of v_i that adapt to variations in q_i and vice versa. In particular, there are $B \times B$ possible sample pairs, so our main objective is to maximize the similarity between vehicle track v_i and text query q_j. We use cosine similarity as the parameterized measurement:

$$\mathbf{s}_\theta\left(v_i, q_j\right) = \frac{\mathbf{z_v^{(i)}} \cdot \mathbf{z_t^{(j)}}}{\left\|\mathbf{z_v^{(i)}}\right\| \left\|\mathbf{z_t^{(j)}}\right\|} \tag{1}$$

where $\cdot$ denotes the dot product operation, and $\left\|\mathbf{z_v^{(i)}}\right\|$, $\left\|\mathbf{z_t^{(j)}}\right\|$ denote the **L2 norm** of the feature vectors.

Latent Space Learning. As the visual feature vector z_v and textual feature vector z_t are projected into a common latent space, an appropriate similarity function shall pull relevant video-sentence pairs close together and irrelevant pairs far apart in the latent space. Thus, We adopt the infoNCE Loss and Circle Loss to connect the representations of the two modalities of the text and the image, ensuring that they are projected into a unified representation space. The InfoNCE Loss is chosen due to its ability to alleviate the model to learn multi-modal embedding space by jointly training visual and text embedding to maximize the similarity between B positive pairs and minimize $B \times (B-1)$ opposing pairs simultaneously. The loss consists of two parts: Image-to-Text and Text-to-Image.

- Image-to-Text Loss:

$$\mathbf{L}_{v \to q} = -\frac{1}{B} \sum_{i}^{B} \log \frac{\exp\left(\mathbf{s}_\theta\left(v_i, q_i\right)\right)}{\sum_{j=1}^{B} \exp\left(\mathbf{s}_\theta\left(v_i, q_j\right)\right)} \tag{2}$$

- Text-to-Image Loss:

$$\mathbf{L}_{q \to v} = -\frac{1}{B} \sum_{i}^{B} \log \frac{\exp\left(\mathbf{s}_\theta\left(q_i, v_i\right)\right)}{\sum_{j=1}^{B} \exp\left(\mathbf{s}_\theta\left(q_j, v_i\right)\right)} \tag{3}$$

Finally, the InfoNCE Loss is formulated as follows:

$$\mathbf{L}_{InfoNCE} = \mathbf{L}_{v \to q} + \mathbf{L}_{q \to v} \tag{4}$$

Different from the InfoNCE loss, Circle loss minimize the similarity of all negative pairs. Denote the positive pair

and negative pairs as s_p and s_n, respectively. Circle loss is defined as follows:

$$\mathrm{L}_{circle} = \log\left[1 + \sum_{j=1}^{L} \exp\left(\gamma\alpha_n^j\left(s_n^j - \Delta_n\right)\right)\sum_{i=1}^{K} \exp\left(-\gamma\alpha_p^i\left(s_p^i - \Delta_p\right)\right)\right] \quad (5)$$

where the $K = 1$ and $L = 2(N-1), \Delta_p$ and Δ_n are the intra-class and inter-class margins, respectively. α_p and α_n are calculated as:

$$\begin{cases} \alpha_p^i = \left[O_p - s_p^i\right]_+ \\ \alpha_n^j = \left[s_n^j - O_n\right]_+ \end{cases} \quad (6)$$

in which O_p and O_n are optimums of s_p and s_n, respectively. These parameters are set $O_p = 1 + m, \Delta_p = 1 - m$, $O_n = -m$ and $\Delta_n = m$, respectively. The final latent space loss is formulated as follows:

$$\mathbf{L}_{latent} = \mathbf{L}_{InfoNCE} + \mathbf{L}_{circle} \quad (7)$$

Concept Space Learning. Aside from classifying each pair based on similarity, leveraging concept features such as vehicle id, where each id is unique, is crucial since learning at the instance level ensures local feature alignment. Hence, concept space learning can be naturally expressed as a multi-class classification task. Thus, we project visual feature vector $\mathbf{z_v}$ and textual feature vector $\mathbf{z_t}$ into a shared-weight classification head $\mathbf{g_c}(\cdot)$ to obtain:

$$\mathbf{x} = \mathbf{g_c}\left(\mathbf{z}\right) = \mathbf{W^{(2)}}\sigma\left(\mathbf{BN}\left(\mathbf{W^{(1)}z}\right)\right) \quad (8)$$

where $\mathbf{x}$ is the final linear classifier and $\mathbf{z}$ represents both $\mathbf{z_v}$ and $\mathbf{z_t}$. The final linear classifier is then used to calculate categorical cross-entropy loss [25] as follows:

$$\mathbf{L}_{concept} = -\frac{1}{C}\sum_{i}^{C}\log\frac{\exp\left(\mathbf{x_i}\right)}{\sum_{j=1}^{C}\exp\left(\mathbf{x_j}\right)} \quad (9)$$

with C denoting the number of vehicle tracks as each vehicle track is a unique id. Then, the **final loss** is formulated as:

$$\mathbf{L}_{final} = \mathbf{L}_{latent} + \mathbf{L}_{concept} \quad (10)$$

3.4. Semi-Supervised Domain Adaptation (SSDA)

Owing to the scarcity of data with accurate labels, training a model solely on the samples available in the training set can lead to overfitting due to the inherent domain gap between the training and testing sets. To overcome this challenge and address the potential domain bias between the two sets, which can result in unobserved scenarios during testing, we propose a Semi-Supervised Domain-Adaptive (SSDA) training strategy. The proposed approach comprises two primary components: the generation of pseudo-labels and the corresponding training strategy. By utilizing this approach, we aim to improve the model's ability to generalize across different domains and effectively handle the unseen scenarios encountered on the testing set with the intuition that incorporating pseudo-labels for the testing set can mitigate the knowledge bias that results from solely learning on the training set.

Pseudo-labels Generation. For a given vehicle track v_i, we leverage the baseline model trained Image Encoder $f_i(\cdot)$ on training set and fine-tune it to develop classification models for vehicle color π_c, vehicle type π_t based on the training dataset and label extracted from section 3.2. Different from our previous version, π_d is generated using various heuristics, which leads to many errors due to different views in cameras. we leverage the training videos and corresponding text queries to train an addition classifier to predict the vehicle's motion direction as π_d and use our new improved version of motion analysis in section 3.5 to refine the results, thus boost the accuracy of the vehicle's direction prediction. Finally, the pseudo-labelling $\hat{q}_i$, also known as the textual query with the format $t_{standardized}$, can be defined as the concatenation of:

$$\hat{q}_i = \pi_c(v_i) \mid\mid \pi_t(v_i) \mid\mid \pi_d(v_i) \quad (11)$$

Our proposed approach is based on the intuition that task specialized models generating pseudo-labels can provide more accurate labels, closer to the ground truth, than pre-trained classification models. By utilizing these specialized models, we can achieve higher accuracy while simultaneously allowing for a certain degree of label errors.

Training Strategy. In contrast to our prior research, we have found that training the model from scratch by combining the training set with standardized text and pseudo-labels can significantly enhance its overall learning performance. This is primarily due to the fact that the resulting mixed set covers the entire distribution of the two sets, thereby enabling the model to better distinguish between different scenarios during the learning stage. By leveraging this approach, we can effectively augment the training data and improve the model's ability to generalize to unseen scenarios encountered during testing. In comparison to fine-tuning the model solely with pseudo-labels, our approach yields superior performance by further enhancing the model's ability to differentiate between relevant information within the training data.

3.5. Post-processing

3.5.1 Motion Analysis

To tackle the problem that occurs due to different camera types and angles, we propose several heuristic-driven algorithms based on our previous work to analyze the right vehicle movement.

Turn Event Detector. To detect whether the vehicle is turning left or right, it is necessary to know the relative position of motion vectors, this can be obtained by utilizing the basic property of the cross-product, as demonstrated in the formula below:

$$\mathcal{D} = \overrightarrow{p_0 p_1} \times \overrightarrow{p_0 p_n} \tag{12}$$

where p_0, p_1, p_n are the positions of the vehicle in the first, second, and last frame, respectively. If the result of the cross-product is negative, we can conclude that the vehicle is turning *left*; otherwise, it is turning *right*.

This approach is based on the obvious assumption that the vehicle turns gradually in one direction while moving (i.e., there is no way the vehicle turns right and then turns left). Furthermore, this method has been proven to be simple and efficient, as it only requires the positions of the vehicle in the first, second, and last frames.

Stop Event Detector. Adopt from [18], to detect the stop event of a vehicle, we follow a procedure that involves calculating the **L2** magnitude of the velocity vectors. First, we obtain a list of velocity vectors $\vec{v_i} = \overrightarrow{p_i p_{i+1}}$, where p_i and p_{i+1} are the positions of the vehicle in frame i and $i+1$, respectively. Next, we calculate the **L2** magnitude of each velocity vector, which represents how far the vehicle moves after one frame. If there exists an interval of time $[L, R]$ such that the **L2** magnitude of the velocity vectors, $\|\vec{v}_{L..R}\|$, is less than a constant threshold eps (which approximates 0), and the duration of the interval is long enough, then we conclude that there exists a stop event.

3.6. Multi-contextual Pruning

Based on our observations, we have determined that imposing strict constraints on the contextual attributes of the vehicle can significantly improve the accuracy of the final results. Accordingly, we propose a novel multi-contextual pruning approach that eliminates tracks with attributes that differ from those specified in the description, such as color, type, and direction, from the top of the rank. In contrast to our prior approach, which relied solely on unidirectional information, our proposed approach leverages a new motion analysis module to prune tracks based on bidirectional information, which takes into account several directions of the vehicle. By integrating this module, we can more accurately identify and eliminate tracks that do not conform to the specified contextual attributes, thereby improving the overall accuracy of the model.

First Stage Pruning. In order to identify the correct vehicle track, we consider the contextual attributes of each vehicle in the current rank. The type π_t and color π_c attributes are checked in sequence for each vehicle, starting from the first and proceeding to the last in the rank. If these attributes match the corresponding attributes in the description, the track is retained in the priority list; otherwise, it is demoted to a lower priority. Once this step is completed, the remaining tracks in the priority list are expected to have the same type and color as the description. The priority list is then used in the next step, where the tracks are rearranged based on their primary direction.

Second Stage Pruning. The priority list is then used in the next step, where the tracks are rearranged based on their directions where determining the direction π_d attribute is a complex task as some descriptions may provide multiple directions for the target track. To fully leverage this additional information, we first predict all possible directions of the track. We then use the priority list obtained from the previous step to rearrange the tracks based on their predicted directions. In general, the pruning process is described in the Algorithm 1.

Algorithm 1 Multi-Contextual Pruning

function PRUNING($text$, $current\_rank$)
  $irrelevant\_list \leftarrow \{\}$ ▷ initialize empty lists
  $priority\_list \leftarrow \{\}$
  for u in $current\_rank$ **do** ▷ filter type, color
    if $get\_type(u) = get\_type(text)$ and $get\_color(u) = get\_color(text)$ **then**
      $priority\_list$.append(u)
    else
      $irrelevant\_list$.append(u)
    end if
  end for
  $highly\_relevant \leftarrow \{\}$
  $likely\_relevant \leftarrow \{\}$
  $directions \leftarrow get\_directions(text)$
  for u in $priority\_list$ **do** ▷ re-rank priority list
    $u\_directions \leftarrow get\_directions(u)$
    if $match\_all(u\_directions, directions)$ **then**
      $highly\_relevant$.append(u)
    else
      $likely\_relevant$.append(u)
    end if
  end for
  return $highly\_relevant + likely\_relevant + irrelevant$
end function

4. Experiments

4.1. Dataset

This work uses the CityFlow-NL [8] Benchmark dataset, consisting of 3.25 hours of footage from 40 cameras across 10 junctions in a mid-sized US metropolis. The dataset contains 2155 tracks of vehicles with three natural language de-

scriptions each, and 184 unique vehicle tracks were selected for this challenge.

4.2. Evaluation Metrics

The Vehicle Retrieval by NL descriptions task is evaluated using standard metrics for retrieval tasks. The Mean Reciprocal Rank (MRR) is used and formulated as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}, \tag{13}$$

where $rank_i$ refers to the rank position of the right track for the i_{th} text description, and Q is the set of text queries. In addition, Recall@5 and Recall@10 are also evaluated.

4.3. Implementation Details

All the training images are resized to 224 $\times$ 224 and normalized. In both training stages, we use AdamW [14] as the optimizer with the initial learning rate set to $1e^{-2}$. We train the model with 11 epochs with a batch size of 64. Each classification model settings are the same as the baseline model, except the batch size is changed to 128 and training epochs are 4. Each vehicle tracks' attributes are generated by inference through all frames of each track and then choose the class with the highest occurrence. All models are trained on one GPU RTX 6000.

4.4. Ablation Study

Ablation study between previous works and current works. Based on the results presented in Tab. 2, it is evident that our proposed approach, which is an enhanced version of our previous work, outperforms the prior work in terms of mean reciprocal rank (MRR) score. Specifically, we observe a significant increase in MRR score for both the baseline model and the semi-supervised domain adaptive (SSDA) training strategy. These results demonstrate the effectiveness of our proposed improvements in enhancing the performance of the retrieval model for tracked-vehicle retrieval by natural language descriptions, as compared to the prior work.

Ablation study of the proposed sytem. According to Tab. 3, our experimental results demonstrate that our baseline vehicle retrieval model, which employs the standardized version of text, achieves an impressive MRR score of 30.28% which shows the effectiveness of the CLIP architecture as the backbone of our model is thereby confirmed. Notably, the robustness of our model is demonstrated despite the use of the standardized text, which contains less information compared to the original text version. Subsequently, by employing the SSDA training strategy, we achieve a remarkable improvement of 24.12% MRR, resulting in a new MRR score of 54.40%. Our experimental findings demonstrate the effectiveness of the proposed SSDA approach for NL-based vehicle retrieval tasks. In addition, we conducted a multi-contextual post-processing strategy that comprised a two-stage pruning approach. Notably, the first stage of pruning led to a significant improvement in MRR, with a boost of 18.73% achieved, which boost the MRR to 73.13%. Our experimental findings highlight the efficacy of the proposed pseudo-label generation technique in producing highly accurate predictions that are closely aligned with the ground truth. This approach is further leveraged as a contextual constraint in a multi-stage pruning strategy to enhance the overall performance of the final results. Finally, in order to fully exploit the benefits of the pruning strategy, we propose a novel multi-contextual pruning approach that leverages the bidirectional attribute. This attribute serves as a strict constraint to effectively differentiate between vehicle tracks that exhibit highly similar visual contexts, which can potentially lead to confusion in the model's predictions. Our experimental results demonstrate the effectiveness of this approach in further improving the accuracy and reliability of our retrieval system, resulting in a notable improvement of 9.5% MRR, leading to a new state-of-the-art MRR score of 82.63%.

Methods	MRR	Recall@5	Recall@10
Baseline [12]	29.59%	40.22%	64.67%
+SSDA	47.73%	66.30%	80.43%
Baseline (ours)	30.28%	44.02%	67.39%
+SSDA	**54.40%**	**69.57%**	**90.22%**

Table 2. The ablation study between solely using InfoNCE (prior work) versus InfoNCE and CircleLoss

Baseline	SSDA	1st Prune	2nd Prune	MRR	Recall@5	Recall@ 10
✓				30.28%	44.02%	67.39%
✓	✓			54.40%	69.57%	90.22%
✓	✓	✓		73.13%	92.93%	100.00%
✓	✓	✓	✓	**82.63%**	**99.46%**	**100.00%**

Table 3. Ablation study on system components

4.5. Challenge Results

As shown in table 4, the final score of our team (Team ID 9) final mean reciprocal rank for the test set is 0.8263. We achieved rank #1 on Track 2 Natural Language-Based Vehicle Retrieval of AI City Challenge 2023.

5. Conclusions

In conclusion, this paper presents a solution for Track 2 in the AI City Challenge 2023 for tracked-vehicle retrieval by natural language descriptions. The proposed solution addresses linguistic ambiguity in the query, utilizes CLIP for feature extraction, uses Semi-Supervised Domain

Rank	Team ID	Team Name	MRR
1	**9**	**HCMIU-CVIP (ours)**	**82.63%**
2	28	IOV	81.79%
3	85	AIO-NLRetrieve_VGU	47.95%
4	151	AIO2022_VGU	46.59%
5	76	DUT_ReID	43.92%

Table 4. The overall ranking on MRR score of AI City Challenge 2023 - Track 2: The Natural language based vehicle retrieval

Adaptive training to overcome domain gap, and employs a post-processing technique to prune out wrong results based on multi-contextual attributes information. The proposed framework achieved a competitive performance of 82.63% MRR accuracy on the test set and won 1st place in the competition. The success of this solution demonstrates the effectiveness of the proposed approach in addressing real-world challenges in cross-modal retrieval tasks.

Acknowledgment

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2022-28-04. We would like to express our heartfelt appreciation to Ho Chi Minh City International University—Vietnam National University (HCMIU-VNU) for facilitating our efforts. Additionally, we would like to express our heartfelt appreciation to all of our colleagues for their contributions, which considerably aided in the revision of the manuscript.

References

[1] 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 (CVPR) Workshops*, pages 4034–4043, June 2021. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. 2

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2019. 2

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 2

[5] Cícero Nogueira dos Santos and Ruy Luiz Milidiú. Semantic role labeling. 2012. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 4

[7] Yunhao Du, Binyu Zhang, Xiangning Ruan, Fei Su, Zhicheng Zhao, and Hong Chen. Omg: Observe multiple granularities for natural language-based vehicle retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3124–3133, June 2022. 2

[8] Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. *ArXiv*, abs/2101.04741, 2021. 6

[9] Valentin Gabeur, Chen Sun, Alahari Karteek, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, 2020. 2

[10] Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733, 2020. 2

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019. 2

[12] Huy Dinh-Anh Le, Quang Qui-Vinh Nguyen, Vuong Ai Nguyen, Thong Duy Nguyen, Nhat Minh Chung, Tin Trung Thai, and Synh Viet-Uyen Ha. Tracked-vehicle retrieval by natural language descriptions with domain adaptive knowledge. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3299–3308, 2022. 2, 7

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 1

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 7

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*, 2019. 2

[16] 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. 1

[17] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mo-

hammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Alice Li, Shangru Li, and Rama Chellappa. The 6th ai city challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3347–3356, June 2022. 2

[18] Thang-Long Nguyen-Ho, Minh-Khoi Pham, Tien-Phat Nguyen, Hai-Dang Nguyen, Minh N. Do, Tam V. Nguyen, and Minh-Triet Tran. Text query based traffic video event retrieval with global-local fusion embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3134–3141, June 2022. 6

[19] 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 *International Conference on Machine Learning*, 2021. 2

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2014. 1

[21] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6397–6406, 2020. 2

[22] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 2

[23] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 4

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3216–3225, June 2022. 2

[25] Zhilu Zhang and Mert Rory Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *ArXiv*, abs/1805.07836, 2018. 5

[26] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3226–3233, June 2022. 2