IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 February 2023, accepted 12 March 2023, date of publication 21 March 2023, date of current version 30 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260149

APPLIED RESEARCH

DAKRS: Domain Adaptive Knowledge-Based Retrieval System for Natural Language-Based Vehicle Retrieval

SYNH VIET-UYEN HA, (Senior Member, IEEE), HUY DINH-ANH LE, (Student Member, IEEE), QUANG QUI-VINH NGUYEN, AND NHAT MINH CHUNG

Vietnam National University—Ho Chi Minh City International University (VNU-HCMIU), Ho Chi Minh City 700000, Vietnam
Vietnam National University, Ho Chi Minh City 700000, Vietnam

Corresponding author: Synh Viet-Uyen Ha (hvusynh@hcmiu.edu.vn)

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2022-28-04.

ABSTRACT Given Natural Language (NL) text descriptions, NL-based vehicle retrieval aims to extract target vehicles from a multi-view multi-camera traffic video pool. Solutions to the problem have been challenged by not only inherent distinctions between textual and visual domains, but also by the complexities of the high-dimensionality of visual data, the diverse range of textual descriptions, a major lack of high-volume datasets in this relatively new field, alongside prominently large domain gaps between training and test sets. To deal with these issues, existing approaches have advocated computationally expensive models to separately extract the subspaces of language and vision before blending them into the same shared representation space. Through our proposed Domain Adaptive Knowledge-based Retrieval System (DAKRS), we show that by taking advantage of multi-modal information in a pretrained model, we can better focus on training robust representations in the shared space of limited labels, rather than on robust extraction of uni-modal representations that comes with increased computational burdens. Our contributions are threefold: (i) An efficient extension of Contrastive Language-Image Pre-training (CLIP)'s transfer learning into a baseline text-to-image multi-modular vehicle retrieval framework; (ii) A data enhancement method to create pseudo-vehicle tracks from the traffic video pool by leveraging the robustness of baseline retrieval model combined with background subtraction; and (iii) A Semi-Supervised Domain Adaptation (SSDA) scheme to engineer pseudo-labels for adapting model parameters to the target domain. Experimental results are benchmarked on Cityflow-NL to obtain 63.20% MRR with 150.0 M of parameters, illustrating our competitive effectiveness and efficiency against state-of-the-arts, without ensembling.

INDEX TERMS Contrastive representation learning, text-to-image retrieval, vehicle retrieval, semi-supervised learning, domain adaptation, background subtraction.

I. INTRODUCTION

Vehicle retrieval, which refers to extracting target vehicles in multiple traffic videos recorded from different views and cameras via a visual or textual input query, is essential in developing Intelligent Traffic Systems (ITS) in smart cities. Most existing vehicle retrieval systems are built using image-to-image matching solutions from the vehicle Re-Identification (ReID) task, which aims to use an image query to retrieve the target vehicle from a pool of vehicle images. Recently, with the rise of many large language models with promising results, image-text retrieval has become more prominent. In particular, being able to query a specific vehicle of interest from a pool of large databases using only intuitive, natural language descriptions is a powerful capability, as it is not only cost-effective but also alleviates the problems in many data-intensive applications where query with image modality is not available.

However, NL-text vehicle retrieval is a very challenging task due to inherent semantic gaps between the features of

The associate editor coordinating the review of this manuscript and approving it for publication was Sergio Consoli.

FIGURE 1. (a) Multi-view, (b) Multi-camera properties in CityFlow-NL where the multi-camera multi-view properties of the dataset, the intra-class variations (i.e. vehicles and scenes) are drastically enlarged and cause domain bias.

images and texts, along with each domain's distinctive data properties. On the language side, the corresponding textual descriptions of each target vehicle track (tracking video of the target vehicle) in the query set can be very diverse in terms of linguistic style but might be grammatically poor, semantically generic, unclear, or even convey conflicting information. These data samples have been known to add noise to the training and assessment of a retrieval model. On the vision side, cars with distinct identities may exhibit little inter-class differences and major intra-class variations due to changes in perspectives and visual resolutions. They often have the same static qualities (e.g., color and type) or dynamic properties (e.g., motion patterns), making cross-modal matching more challenging. Furthermore, because the field is relatively new, there is a lack of high-volume datasets that comprehensively include video descriptions for vehicle tracks and synonymous textual descriptions of vehicle queries. Notably, the CityFlow-NL dataset [1] is one of the first publicly available datasets in the field. Still, the small amount of annotated language-vehicle pairings being supplied in the training set is limited in terms of how much it can address the domain gap from the test sets. Furthermore, as shown in FIGURE 1. due to the multi-camera multi-view properties of the dataset, the intra-class variations are drastically enlarged and cause domain bias that leads to smaller inter-class differences between vehicles in the same scenario.

To our knowledge, existing methods [2], [3], [4] have dealt with the aforementioned issues under the following limitations,

- Existing approaches mainly focus on exploiting specialized knowledge from single-modal pre-training (i.e., using the pre-trained vision encoder or the language encoder) before blending them to facilitate multi-modal learning. In particular, by leveraging the robust performances of model-driven approaches, they separately pre-trained the vision/language knowledge using different specialized models, then transfer-learn the embeddings towards the same embedding subspace. By neglecting the multi-modal corresponding information, models can be good at the visual or textual tasks but would require significant efforts and data to adapt to the generalized theme of multi-modal learning. As a consequence, it typically leads to correspondingly higher computational expenses and more data.
- In addition to real-world scenarios, the existing approaches of text-to-vehicle retrieval have yet actively addressed the distribution gap (termed as domain shift, or subpopulation shift) between the data used for training the model (source domain) and the test data (target domain), especially in the setting of limited data. Despite of the fact that recent studies on Domain Adaptation (Self-Supervised, Semi-Supervised, etc.) have proposed various strategies for models to adapt and overcome poor generalization with data of different distributions, the lack of utilization of those techniques has resulted in performance limitations as models overfit to the training domain.

In this research, we aim to simultaneously address the inherent problems of NL-based vehicle retrieval and alleviate the issues of existing approaches. Intuitively, we believe that it is possible to take advantage of multi-modal information and generalization capabilities of a pretrained model, so that we can instead focus on training robust representations in the shared space of limited labels, rather than on bridging uni-modal representations from scratch which apparently have entailed intensive computational costs and training burdens. Our proposed domain adaptive knowledge retrieval system shall be described in three aspects, all of which also correspond to our scientific contributions:

1) In order to work with limited data without unnecessarily increasing computational complexities, we propose an extension of the CLIP [5], a large neural network study on unlabeled vision-language multi-modular pre-training. By adopting CLIP, we could circumvent the shortage of labelled data and better attain the full potential of network learning when large-scale labelled data are not available.
2) To alleviate the issues of limited high-quality labels, we propose a data enhancement method that generates augmented data. In fact, by visually extracting unlabelled vehicle tracks from the traffic video data with background subtraction, and textually presenting a simplified version for text descriptions, we introduce the method to expand the data pool and address the linguistic ambiguity residing in the language query set.
3) To address the domain gap between the training and testing sets and to further amplify the retrieval performance of the baseline model, we propose SSDA to deal with problems from the domain shift by adapting the knowledge of the test set through the pseudo-labels. Our SSDA approach for retrieval combines a small number of labelled samples from the target domain with the remaining unlabelled target data, significantly

different from current works focusing on supervised training settings.

A preliminary version of this work has been published as a technical paper [6]. In this work, we focus on more scientific aspects and key modifications in the data enhancement method to further support the SSDA training method, which allows us to enhance model performance in retrieval results significantly. From experimental results on the CityFlow-NL dataset, our proposed method has been validated to achieve competitive results while having less computational cost than any state-of-the-art methods without using any ensemble and post-processing methods.

II. RELATED WORKS

A. NATURAL LANGUAGE-BASED VEHICLE-BASED VIDEO RETRIEVAL

In recent years, natural language-based video retrieval tasks have garnered significant interest. Early research focuses on extracting representative features from video and text data to establish a link between text and video. To encode the language, these efforts employ a textual feature extractor such as ERNIE [7], Word2Vec [8], or LSTM [9], and a powerful vision network such as ViT [10] to extract visual features. CLIP4Clip [11] investigates a method to transfer the knowledge of the vision-language pre-training model to video-language retrieval problems. Come Recently, large-scale pre-trained vision-language models have shown excellent performance in video retrieval tasks. CLIPBERT [12] uses sparse sampling to make end-to-end learning for video language challenges inexpensive. In addition, [13] proposes a novel framework that employs multiple queries as inputs to provide more accurate results, as opposed to combining the similarity outputs of multiple queries from the previous single query-trained models. In contrast to the typical video retrieval job, the vehicle retrieval task is primarily an instance-level retrieval problem requiring a model to comprehend traffic scenes and vehicle properties better. To accelerate research in the field, the 5th and 6th AI City Challenge [14], [15] hosted by NVIDIA has specifically organised the natural language-based vehicle retrieval challenge to encourage text-to-image vehicle retrieval systems development to advance research in the field.

In the 5th NVIDIA AI City Challenge, the majority of teams [2], [16], [17], [18], [19], [20] chose to extract sentence embeddings of the queries, whereas two teams [21], [22] processed the NL queries using conventional NLP techniques. For cross-modality learning, certain teams [2], [20] used ReID models with the adoption of vision models pre-trained on visual ReID data and language models pre-trained on the given queries from the dataset. The motion of vehicles is an integral component of the NL descriptions. Consequently, a number of teams [2], [18], [22] have developed specific methods for measuring and representing vehicle motion patterns.

As for the 6th NVIDIA AI City Challenge, all participating teams utilised InfoNCE losses [23] to train for the text-to-image retrieval task. In addition, to represent the NL descriptions, most teams used some pre-trained phrase embedding models, such as BERT [24], to represent the NL descriptions. Using an NL parser, the [4] team obtained tracked vehicles' color, kind, and movement. In addition to the ReID-based strategy, these attributes were employed to post-process the retrieval results. Consequently, several groups [3], [25], [26] utilised the global motion image provided by Bai et al. [2] to generate a vehicle motion stream. The [3] team developed an enhanced motion image utilising the inter-frame IoU of the tracked targets.

However, among the state-of-the-art solutions, every team mostly focused on combining separate pre-training vision and language models while neglecting the research on using vision-language pre-training models. In tackling that gap, our preliminary version leveraged CLIP vision-language pre-training model, proposed an SSDA training process and performed motion analysis and post-processing with the pruning of retrieval results.

B. VEHICLE RETRIEVAL DATASETS

The vehicle retrieval task is derived from Vehicle ReID, an important task in computer vision that matches vehicle images taken from different cameras and viewpoints to a unique vehicle identity. Over the years, various vehicle ReID datasets have been created and used as the benchmark for evaluating and improving vehicle ReID models. Some of the most notable datasets are VeRi-776 [27], a large-scale dataset that contains over 50,000 images of 776 vehicles captured by 20 cameras from various viewpoints and weather conditions; VehicleID [28] dataset consisting of 2211,567 images of 26,328 vehicles; CompCar [29] dataset is designed for fine-grained car recognition and contains 136,726 images of 1,716 car models; CityFlow [30] dataset is among the most massive vehicle re-id datasets. There are 666 vehicle identifiers labelled with bounding boxes. Realistically, each image is captured from 40 cameras with 36,935 training images of 333 vehicles and 19,342 testing images of other 333 vehicles.

Despite the enlargement and constant development in image-to-image vehicle datasets, text-to-image vehicle datasets still lack large-scale quality realistic datasets. Thus, a benchmark proposed by NVIDIA AI City Challenge called CityFlow-NL is the first publicly available dataset intended to allow study at the intersection of multi-object tracking, retrieval by natural language specification, and temporal localisation of occurrences. The benchmark is drawn from CityFlow, a public benchmark that has been the focal point of several previous NVIDIA AI City Challenge workshop challenges centred on Multi-Target Multi-Camera (MTMC) tracking and ReID.

C. VISION-LANGUAGE PRE-TRAINING

Vision-Language pre-training has made significant strides recently and obtained outstanding results on various

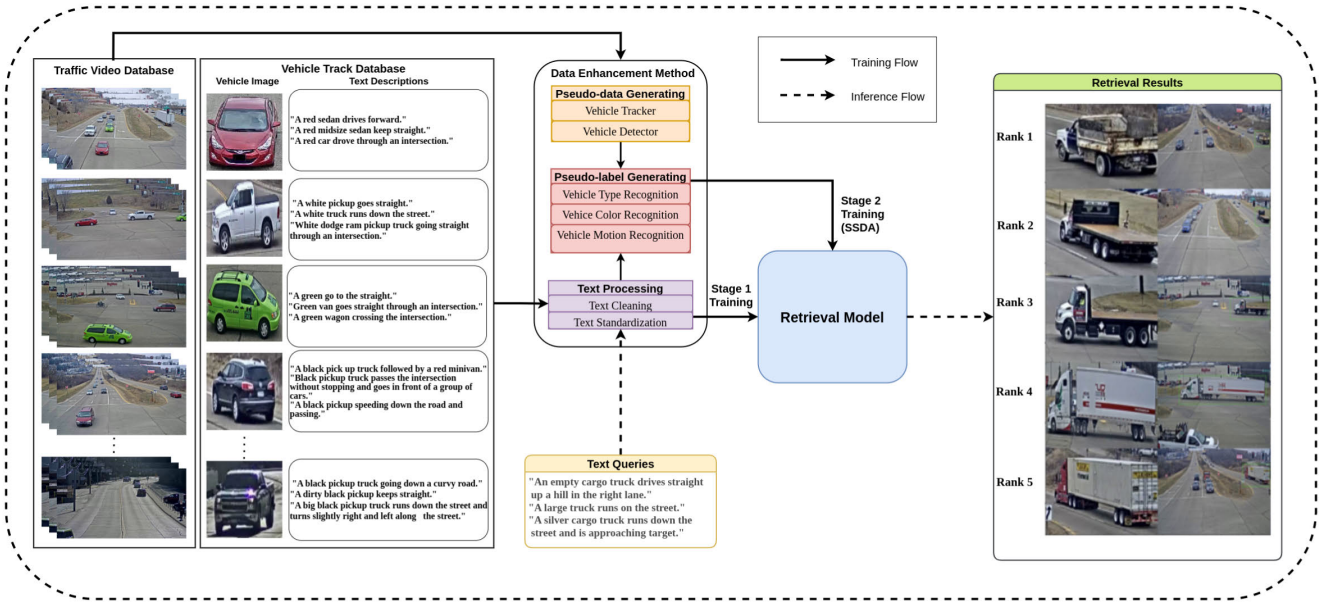


FIGURE 2. The overall framework of DAKRS for vehicle retrieval through natural language descriptions includes the data enhancement method. Given a Traffic Video Database, we want to match vehicle data to the corresponding textual data as a Vehicle Track Database. For enlarging the training dataset, the Traffic Video Database is used in our Data Enhancement Method, whose pseudo-data, pseudo-labels, and pre-processed texts are used to train the Retrieval Model that leverages the multi-modal capabilities of CLIP.

multi-modal downstream tasks. Several ways were developed using semantic supervision from large-scale image data to learn visual representations from textual representations. MIL-NCE [31] primarily investigated utilizing noisy, large-scale Howto100M [32] instructional films to learn a superior video encoder in an end-to-end fashion. SimVLM [33] lowered the complexity of training by utilizing large-scale weak supervision and adopting a single prefix language modelling objective end-to-end method. CLIP, a recent method, has demonstrated remarkable success matching two modalities' representations in the embedding space by utilizing internet-collected large-scale image and text pairs. CLIP implemented contrastive learning with high-capacity language models and visual feature encoders to identify appealing visual concepts for zero-shot picture categorization.

Inspired by recent encouraging results of transferrirng the knowledge of CLIP models to downstream tasks, such as video captioning, video-text retrieval, and image synthesis, we propose to exploit the generalization capabilities of the CLIP approach to extend it to a robust and efficient baseline in NL-based vehicle retrieval.

III. THE METHODOLOGY

In this section, we introduce our proposed domain adaptive knowledge-based retrieval system in FIGURE 2. which leverages and extends the multi-modal capabilities of CLIP to refer to vehicle retrieval in the format of texts and videos. Furthermore, to address the gap between source and target domains, we discuss our proposed data enhancement method and our SSDA strategy, where augmented data and pseudo-labels are generated from raw videos to adapt model parameters towards the target domain.

A. PROBLEM FORMULATION OF THE TEXT-TO-TRACK VEHICLE RETRIEVAL

Given a set of n traffic video clips $\boldsymbol{V} = \{v_1, v_2, \ldots, v_n\}$ and a corresponding NL-text query database $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_n\}$, we seek to learn a function $\mathbf{s}(v_i, q_j)$ such that $\boldsymbol{q_j} = \left\{q_j^1, q_j^2, \ldots, q_j^m\right\}$ is the set of m synonymous text descriptions, and each clip v_i is annotated with the bounding box coordinates of the tracked-vehicle as $\boldsymbol{B}(v_i) = \left\{b_1, b_2, \ldots, b_{|v_i|}\right\}$ over the video length $|v_i|$. In particular, suppose that each set of NL-description comprises 3 descriptions corresponding to each vehicle track v_i. The main objective of this problem is to successfully retrieve video v_i from $\boldsymbol{V}$ based on $\boldsymbol{q_j} = \left\{q_j^1, q_j^2, q_j^3\right\}$ from $\boldsymbol{Q}$. Thus, a solution of interest needs 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 $\boldsymbol{Q}$, $q_k \neq q_j$.

Furthermore, with two domains of interest: the source domain set and target domain set, we denote the source domain set as $\boldsymbol{D_s} = \{(\boldsymbol{V}, \boldsymbol{Q})\}$, where $\boldsymbol{V}$ refers to the source domain's vehicle track database and $\boldsymbol{Q}$ is referred to as their hand-labelled labels. Thus, the target domain set is similarly denoted as $\boldsymbol{D_t} = \left\{(\boldsymbol{V}', \boldsymbol{Q}')\right\}$ where $\boldsymbol{V}' = \left\{v_1', v_2', \ldots, v_p'\right\}$ is target domain vehicle track database, and $\boldsymbol{Q}' = \left\{q_1', q_2', \ldots, q_p'\right\}$ is the corresponding, unseen query set of the target domain that needs to be matched pair-wise with

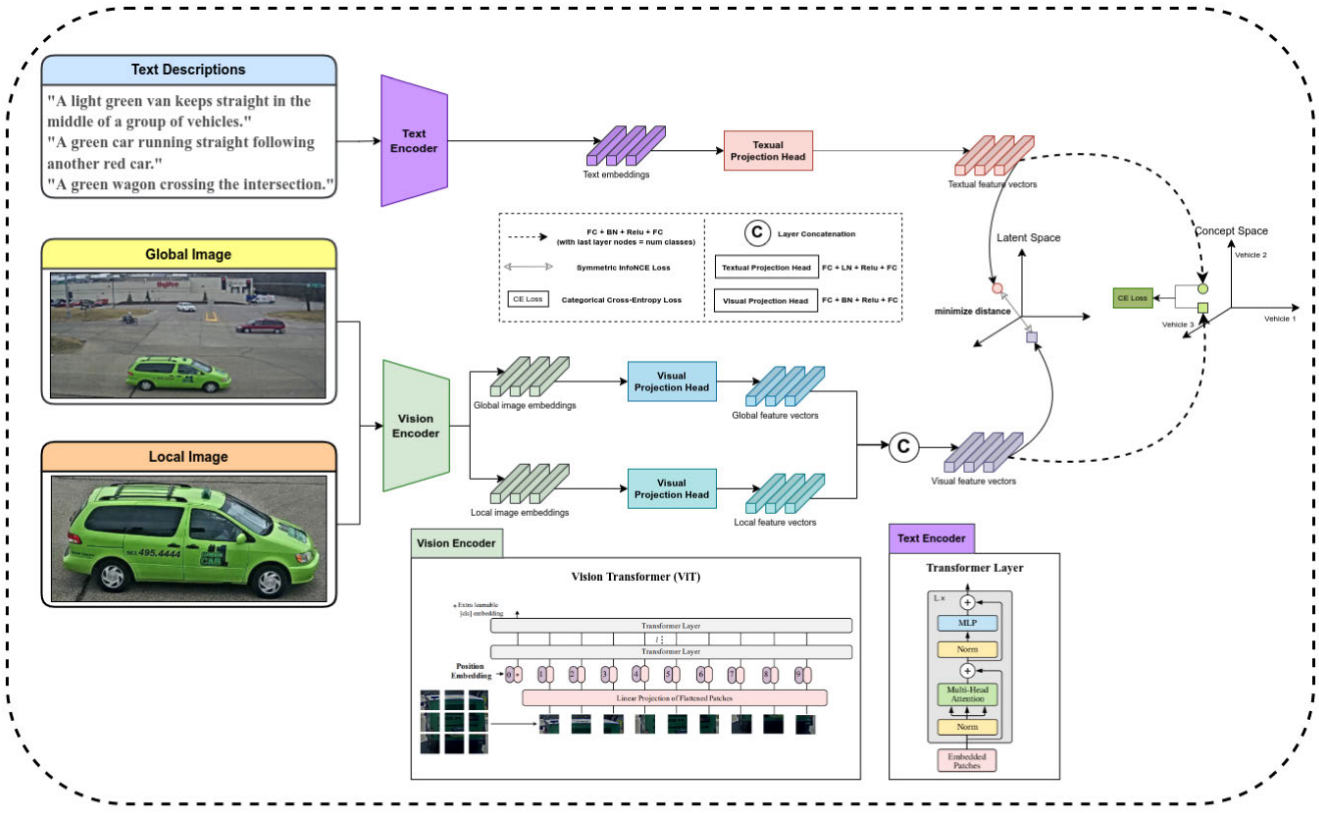


FIGURE 3. Our baseline retrieval model leverages CLIP that includes visual and textual feature extractors, which respectively are Transformer as the Vision Encoder and Text Transformer as the Text Encoder. Textual Descriptions are encoded to textual vectors via the Text Encoder, and vehicle scene (Global Image) along with its close-up appearance (Local Image) are encoded to visual vectors by the Vision Encoder. Via hybrid space learning of both latent space and concept space learning, text and image representations mutally adapt to each other's variations in the shared domain, thereby creating generalizable embeddings for text-image retrieval task.

elements in $\boldsymbol{V}'$. Hence, by learning the similarity function $\mathbf{s}_\theta(\cdot)$ parameterized by θ on $\boldsymbol{D_s}$, we seek to maximize the general objective,

$$\mathbf{S} = \sum_{i=1}^{n} \mathbf{M}\left(i, \ \arg\max_j \left[\mathbf{s}_\theta(v_i', q_j')\right]\right)$$
$$\mathbf{M}(i,j) = \begin{cases} 1 & if \quad i = j \\ 0 & if \quad otherwise \end{cases} \tag{1}$$

where ideally, and without seeing $\boldsymbol{Q}'$,

B. THE PROPOSED DOMAIN-ADAPTIVE TEXT-TO-VEHICLE RETRIEVAL FRAMEWORK

Parallel with related works in the field [2], [16], [21], in FIGURE 3. we constructed a framework that leverages the multi-modal ''zero-shot'' representation capabilities of a pre-trained network, such that by having pre-trained it with an abundantly available source of supervision of unfiltered, highly varied, and highly noisy data (the internet), there would be less need to scale its parameterization size. Furthermore, choosing a suitable backbone as the feature extractor is vital for obtaining robust embeddings since they contain abstract features that are disentangled from varying degrees of inessential variations, making them more generalizable for text-image retrieval tasks.

With critical insights into leveraging NL-text and images as flexible prediction spaces to enable generalization and transfer, we propose an architecture that extends the CLIP model for the text-to-vehicle retrieval problem, thanks to its extensive knowledge in constructing strong representations for visual-textual feature extraction tasks. CLIP uses pre-trained models Vision Transformer as the Image Encoder $f_i(\cdot)$, and a Text Transformer [34] as the Text Encoder $f_t(\cdot)$. We show our proposed extensions in FIGURE 3. with a dual-stream visual branch and a single-stream textual branch for each set of 3 textual inputs.

1) DUAL STREAM

We utilize a dual-stream processing component to align the multi-granularity information from text descriptions. Our insights are to construct a visual representation to enhance the knowledge the model can capture in each vehicle track regarding global and local visual views.

As defined, each vehicle track is represented by a video clip v_i and its corresponding set of bounding box coordinates $\boldsymbol{B}(v_i)$. Hence, each vehicle track has its set of global and local

images corresponding to the global view of the video clip and the cropped view of the vehicle in that video clip. We further denote the j^{th} global image in v_i as $\boldsymbol{I}_{\boldsymbol{g}}^{\boldsymbol{j}} \in \mathbb{R}^{H\times W\times 3}$, which is the original frame, and the local image $\boldsymbol{I}_{\boldsymbol{l}}^{\boldsymbol{j}} \in \mathbb{R}^{H\times W\times 3}$ that is cropped on a global image by a bounding box b_j and resized to the original frame size. Both inputs are then simultaneously encoded via a shared-weight image encoder $\mathbf{f_i}(\cdot)$ in a parallel fashion to obtain:

$$\begin{aligned}\mathbf{h}_{\mathbf{g}}^{\mathbf{j}} &= \mathbf{f_i}(\mathbf{I}_{\mathbf{g}}^{\mathbf{j}})\\ \mathbf{h}_{\mathbf{l}}^{\mathbf{j}} &= \mathbf{f_i}(\mathbf{I}_{\mathbf{l}}^{\mathbf{j}})\end{aligned} \tag{2}$$

where $\mathbf{h}_{\mathbf{g}}^{\mathbf{j}} \in \mathbb{R}^{B\times 512}$, $\mathbf{h}_{\mathbf{l}}^{\mathbf{j}} \in \mathbb{R}^{B\times 512}$ are global and local feature embeddings, respectively. Note that B is the batch size, H and W are the height and width of the image.

Finally, for the tracked-vehicle representation of v_i, we combine both local and global views across the sampled j^{th} frames to obtain the visual representation as,

$$\begin{aligned}\mathbf{h_v} &= \mathbb{E}\left[\mathbf{h}_{\mathbf{v}}^{\mathbf{j}}\right]\\ \mathbf{h}_{\mathbf{v}}^{\mathbf{j}} &= \left[\mathbf{h}_{\mathbf{g}}^{\mathbf{j}}||\mathbf{h}_{\mathbf{l}}^{\mathbf{j}}\right]\end{aligned} \tag{3}$$

where $\mathbf{h}_{\mathbf{v}}^{\mathbf{j}} \in \mathbb{R}^{B\times 1024}$ and channel-wise concatenation operation is defined as $||$, This way, we learn the vehicles' expected motion poses and appearances to match them with directional trajectory and appearance-wise textual descriptions over the video clip.

2) SEMANTIC EXTRACTION OF TEXT REPRESENTATION

Due to poor textual quality in grammar and spelling between sentences in each description set, text-processing steps are performed to ensure consistency across different description sets. We identify each misspelt word and calculate the Levenshtein distance to replace it with the corresponding correct word in the prepared dictionary where the distance between two words is the smallest. After that, we use the SRL tool [35] to extract verbs and perform text stemming that converts verbs into their base form.

Nevertheless, regarding variations of style and context in NL descriptions, we propose to disentangle the factors of variations of m synonymous descriptions in terms of style, while context differences are still taken into account by direct learning of each representation with respect to the visual description. By previously denoting a query description for v_i as q_i in $\boldsymbol{Q}$, we randomly sample one sentence among m synonymous sentences as an example and encode it with a text encoder to obtain the textual representation as,

$$\mathbf{h_t} = \mathbb{E}\left[\mathbf{h}_{\mathbf{t}}^{\mathbf{j}}\right] \tag{4}$$

where $\mathbf{h}_{\mathbf{t}}^{\mathbf{j}} = \mathbf{f_t}(q_i^j) \in \mathbb{R}^{B\times 1024}$ is text feature embedding. Note that B is the batch size.

3) PROJECTION HEADS

Inspired by [36], we then feed text and image representation respectively into each separated projection heads $\mathbf{g_v}(\cdot)$ and $\mathbf{g_t}(\cdot)$, with the intention of mapping 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 $\mathbf{z_t}$ can be represented as:

$$\begin{aligned}\mathbf{z_v} = \mathbf{g_v}\left(\mathbf{h_v}\right) &= \mathbf{W^{(2)}}\sigma\left(\mathbf{BN}\left(\mathbf{W^{(1)}h_v}\right)\right)\\ \mathbf{z_t} = \mathbf{g_t}\left(\mathbf{h_t}\right) &= \mathbf{W^{(2)}}\sigma\left(\mathbf{LN}\left(\mathbf{W^{(1)}h_t}\right)\right)\end{aligned} \tag{5}$$

where the projection head is a small Multi-Layer Perceptron (MLP) with one fully-connected **(FC)** layer that contains $\mathbf{W^{(1)}}$ and $\mathbf{W^{(2)}}$ as the parameter of the fully-connected on each side, respectively, the MLP is using a non-linear activation function σ ReLU and a Normalization Layer with Batch Normalization (**BN**) for visual representation and Layer Normalization (**LN**) for text representation. An additional image projection head is leveraged as the concatenation features between dual-stream visual feature vectors for the concept space learning task in section III-B4.b. All the vectors are then normalized to be unit vectors. Additionally, a special projection head, also known as classification head $\mathbf{g_c}(\cdot)$, is employed to map visual and textual feature vectors into a classification space where each pair of visual and textual feature vectors corresponds to a vehicle track id.

4) MULTI-MODAL DOMAIN-ADAPTIVE LEARNING

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 in the source domain $\boldsymbol{D_s}$. We use cosine similarity as the parameterized measurement:

$$\mathbf{s}_\theta\left(v_i, q_j\right) = \frac{\mathbf{z}_{\mathbf{v}}^{(\mathbf{i})} \cdot \mathbf{z}_{\mathbf{t}}^{(\mathbf{j})}}{\left\|\mathbf{z}_{\mathbf{v}}^{(\mathbf{i})}\right\| \left\|\mathbf{z}_{\mathbf{t}}^{(\mathbf{j})}\right\|} \tag{6}$$

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

a: 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 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:

$$\mathbf{L}_{v\to q} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp\left(\mathbf{s}_\theta\left(v_i, q_i\right)\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{s}_\theta\left(v_i, q_j\right)\right)} \tag{7}$$

- Text-to-Image Loss:

$$\mathbf{L}_{q\to v} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp\left(\mathbf{s}_\theta\left(v_i, q_i\right)\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{s}_\theta\left(v_j, q_i\right)\right)} \tag{8}$$

Finally, the InfoNCE Loss is formulated as follows:

$$\mathbf{L}_{InfoNCE} = \mathbf{L}_{v\to q} + \mathbf{L}_{q\to v} \tag{9}$$

b: 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:

$$\mathbf{x} = \mathbf{g_c}\left(\mathbf{z}\right) = \mathbf{W^{(2)}}\sigma\left(\mathbf{BN}\left(\mathbf{W^{(1)}z}\right)\right) \tag{10}$$

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 as follows:

$$\mathbf{L}_{concept} = -\frac{1}{C}\sum_{i}^{C}\log\frac{\exp\left(\mathbf{x_i}\right)}{\sum_{j=1}^{C}\exp\left(\mathbf{x_j}\right)} \tag{11}$$

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

$$\mathbf{L}_{final} = \mathbf{L}_{InfoNCE} + \mathbf{L}_{concept} \tag{12}$$

C. THE PROPOSED DATA ENHANCEMENT METHOD

To actively alleviate the issues of limited high-quality labels in the source domain $\boldsymbol{D_s}$, we introduce the data enhancement method that further generates augmented data and pseudo-labels on the target domain $\boldsymbol{D_t}$ by visually extracting unlabelled vehicle tracks from the traffic video data, and textually presenting a standardized version for text descriptions to address the linguistic ambiguity residing in the language query set on both $\boldsymbol{D_s}$ and $\boldsymbol{D_t}$. All of this serves as a way to fully exploit the vast amount of unlabeled data since the training set only reflects a small number of vehicle tracks inside each video. Therefore, we propose two new components: Pseudo-label Generating and Augmented-data Generating, to complement the shortage of datasets in the source domain and gradually bridge the discrepancy between the two domains.

a: VEHICLE TEXT GRAMMAR

As mentioned in section III-B2, due to the vast diversity in each description, creating pseudo-labels near that content level required many resources and effort. Hence, we denote the original text format as $t_{original}$ and we propose a standardized text grammar for vehicle retrieval and formulated as follows:

$$t_{standardized} = a_c + a_t + a_m + a_r \tag{13}$$

where:

- a_c denotes the attribute vehicle's color as Red, Blue, Yellow, Black, etc.
- a_t denotes the attribute vehicle's type as Car (generic), Sedan, SUV, Truck, or Large Truck, etc.
- a_m denotes the attribute vehicle's motion as Straight Ahead, Turn Left, Turn Right, Turn Around, etc.
- a_r denotes the attribute vehicle's other surrounding information, such as ''Behind a Red Sedan'', etc.

This format enforces each description into the same content level and benefits the proposed pseudo-label technique. The SRL tool extracts verbs from the sentence and helps identify vehicle color and type along with any relative descriptions on sub-target in the query. In particular, each vehicle track corresponds with m synonymous text descriptions and can cause variation in the color and type of vehicle track. Thus, we select the most prevalent vehicle color and type based on the number of instances that appear statistically the most and replace it for all descriptions in each query.

b: AUGMENTED-DATA GENERATION

Due to the relatively new field, limitations from the lack of a large-scale dataset are concerning. We propose an augmentation strategy based on data observation to tackle that issue when working with existing datasets. As seen in the global image of FIGURE 3. the target of interest is only one vehicle in the scene, while there can be other separately moving vehicles in the scene.

Hence, given a vehicle track database $\boldsymbol{V}$, whether from the source or the target domain, we propose to use **Background Subtraction** [37] to extract all separately moving vehicle coordinates, then we employ **Vehicle Tracking** on the vehicles based on their speed and momentum such as in [38]. By doing so, we can obtain an augmented vehicle database $\boldsymbol{V^a} = \left\{v_1^a, v_2^a, \ldots, v_b^a\right\}$, with b as a chosen number of vehicles. The algorithm is shown in Algorithm 1.

c: PSEUDO-LABEL GENERATION

In a setting where labelled data on the target domain is unavailable, such as for the augmented set $\boldsymbol{V^a}$ and the target set $\boldsymbol{V'}$, Unsupervised Domain Adaptation (UDA) is a common approach where pseudo-labels on the target domain are generated to refine the model which aims at transferring knowledge from a strong label source domain to unlabeled target domain. Lin et al. [39] propose utilizing the characteristic to facilitate the transmission of information and eliminate the pseudo-label noise. Zhang et al. [40] and Yang et al. [41] provide an asymmetric co-teaching framework in which two distinct modules create pseudo-label for each other. However, a commonly used strategy for generating pseudo-labels in UDA (e.g. Feature clustering on the target domain) is prone to significant errors in this problem with limited data since the domain gap between textual and visual features is significant. Thus, we introduce a simple but effective pseudo-label strategy where information from the source domain and knowledge of the baseline model is leveraged by **Vehicle Attribute Recognition** to create pseudo-label for target domain.

In particular, for a given vehicle track v_i, we fine-tuned CLIP's Vision Transformer to develop classification models

Algorithm 1 Augmented-Data Generation

```
M ← Background Subtraction (BGS) Model
T ← Vehicle Tracking Model
V^a ← {}                                  ▷ Dictionary for Augmented Data
procedure Augment(V)
    i ← 1
    while i ≠ |V| do                      ▷ Background Initialization
        M.update(I_i)                     ▷ Background Update
        i ← i + 1
    end while

    i ← 1
    while i ≠ |V| do                      ▷ Data Generation
        O ← M.subtract(I_i)               ▷ BGS to Vehicles
        for o in O do                     ▷ For each Vehicle Object
            v ← T.track(o)
            if v.id not in V^a then
                V^a = [ ]
            end if

            V^a [v.id] .append({I_i, o})  ▷ Augment

        end for
        i ← i + 1
    end while
end procedure
```

for vehicle color π_c and vehicle type π_t based on the training dataset. This is meant to use CLIP's general learning on data from the large-scale vision-language datasets and specific attributes of the vehicle retrieval domain. Finally, different from our previous version using various heuristics to calculate the vehicle's trajectory and extract the vehicle's direction, which leads to many errors due to different views in cameras. Thus, in this version, to further enhance the accuracy in generating pseudo-label, we leverage the training videos and corresponding text queries to train an addition classifier to predict the vehicle's motion direction as π_d. Finally, the pseudo-labelling, also known as the textual query with the format $t_{standardized}$, can be defined as the concatenation of:

$$\hat{q}_i = \pi_c(v_i) \,||\, \pi_t(v_i) \,||\, \pi_d(v_i) \tag{14}$$

Through this approach, we can produce pseudo-labels for any vehicle track attribute based on three attribute classification modules: color, type, and movement. For color and type classifications, we discovered that the baseline model could retrieve the vehicle track that closely matches the text descriptions with the same vehicle color and type.

D. SEMI-SUPERVISED DOMAIN ADAPTATION (SSDA)

Due to the limited data with strong labels, training the model using only the samples in the source domain can easily lead to overfitting because of the domain gap between the source and target domains. In place of strong labels, we propose to additionally fine-tunne the **general multi-modal model** (i.e. our CLIP-extended framework) with pseudo-label generated by **specialized classification models** on the target domain. By doing so, our model is guided with information about the subpopulation shift of data in the test domain, and thus adapt its parameters for increased accuracy.

- **General multi-modal model:** to extract generalizable representations, our CLIP-extended framework is capable of mixing the visual and textual subspaces to obtain generalizable results via Vision Transformer and the Text Transformer. To mix with the Text Transformer, the Vision Transformer needs to encapsulate visual-linguistic context into its embeddings.
- **Specialized classification models:** By teaching the Vision Transformer to perform a narrow set of visual concepts for classification, it can be good at one specific task of generating our classifications of interest, thereby generating strong pseudo-label that are approximately close to the ground truths.

The limitation of training data is a setting where knowledge bias from learning on the source domain can be alleviated by the pseudo-labels on the target domain. After learning, it is possible to align source-trained model parameters towards the target domain's data distribution. Intuitively, our insight is that the pseudo-label generated by specialized models shall provide labels closer to the ground truth than a general multi-modal model. Hence we could obtain higher accuracy while tolerating certain degrees of label errors.

In particular, suppose $\hat{\boldsymbol{Q}}' = \left\{\hat{q}'_1, \hat{q}'_2, \ldots, \hat{q}'_p\right\}$ are the pseudo-label query set generated for the target vehicle set $\boldsymbol{V}'$, we employ the text representation extractor $\mathbf{f_t}(\cdot)$ on $\hat{\boldsymbol{Q}}'$ and by Equation (12) such that the embeddings, similarities and losses are generated as if the set $\hat{\boldsymbol{Q}}'$ comprises of the strong corresponding labels for $\boldsymbol{V}'$.

IV. EXPERIMENTS

A. DATASET

CityFlow-NL Dataset: The CityFlow-NL benchmark includes 666 target vehicles in 3598 single-view vehicle tracks captured by 46 calibrated cameras with 5 different scenarios and 6784 distinct NL descriptions. At least three crowdsourcing employees supplied NL descriptions for each target to capture better true differences and ambiguities expected in real-world application domains. The NL descriptions include details about the vehicle's color, movement, traffic scene, and relationships with other vehicles. The dataset is originally built upon the CityFlow benchmark, which contains 3.58 hours (215.03 minutes) of footage from 46 cameras spanning 16 intersections in a mid-sized US metropolis with a distance of 4 km between the two furthest existing benchmarks.

B. EVALUATION METRICS

The Vehicle Retrieval by NL descriptions task is evaluated using standard metrics for retrieval tasks. The Mean

TABLE 1. Comparison between the preliminary work and DAKRS.

Method	MRR	Recall@5	Recall@10
preliminary version [6]	47.73%	66.30%	80.43%
DAKRS (ours)	**63.20%**	**79.35%**	**93.48%**

Reciprocal Rank (MRR) is used and formulated as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}, \tag{15}$$

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

C. IMPLEMENTATION DETAILS

1) BASELINE TRAINING STAGE

We choose CLIP's ViT-B/32 and Text Transfomer as as the backbone of the image encoder and the text encoder, respectively. The size of embedding visual and textual vectors' dimensions are 512 and 1024, respectively. All the training images are resized to 224 × 224 as the input for the visual encoder. We train the model with 180 epochs, with the size of each mini-batch set to 64. During both training stages, we use AdamW [42] as the optimizer with the initial learning rate set to $1e^{-2}$ and decay $1e^{-1}$ every 30 epochs for better convergence. We train 2 variations of the baseline models with different text formats, where **Type 1 is (original format)** and **Type 2 is (standardized format)** respectively, to evaluate the effectiveness of each text format.

2) SSDA TRAINING STAGE

All the training settings are the same as the baseline model. We then fine-tune the baseline model for 90 epochs with the pseudo-label generated from section III-C All experiments are conducted with Pytorch 1.13 on 1 GPU NVIDIA Quadro RTX 6000.

3) INFERENCE STAGE

During the inference stage, we use the text format similar to the format in which the model is trained/fine-tuned to maximize each text format's retrieval performance.

4) CLASSIFICATION MODELS

All the training settings are the same as the baseline model, except the batch size is changed to 128. Each classification model is fine-tuned from the baseline retrieval model vision branch without using the global images as inputs. We train each model for 210 epochs and use it to automatically generate type, color and direction for each vehicle track by inference through all frames of each track and then choose the class with the highest occurrence.

D. EXPERIMENTAL RESULTS

1) QUANTITATIVE RESULTS

As shown in FIGURE 4. we plot the variation in the performance of DAKRS between the Baseline retrieval model and

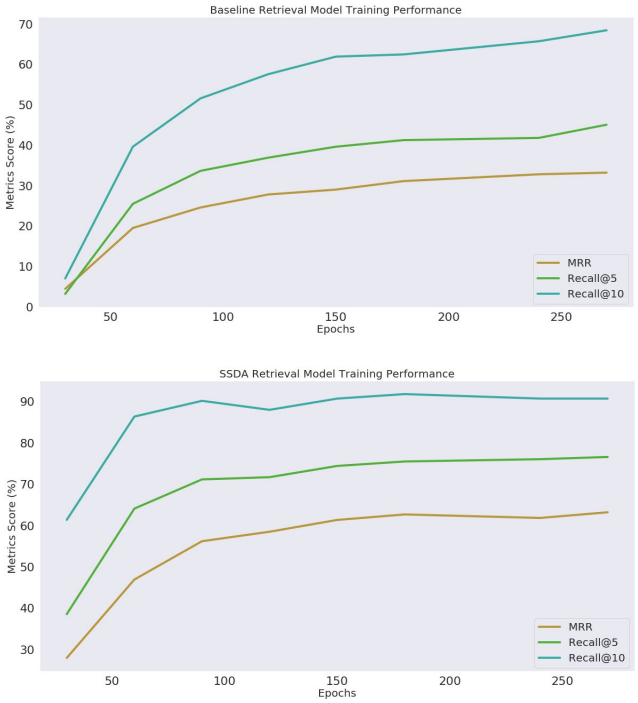


FIGURE 4. Training performance of the Baseline and SSDA retrieval model on CityFlow-NL.

TABLE 2. Summary of datasets for training on CityFlow-NL where "Data Enhancement Method" refers to "DEM" and "w/" refers to "with."

Training Dataset	Total ids	Total queries	Total images
Training set	2,155	6,465	185,199
Training set w/ DEM	3,032	7,342	275,517

TABLE 3. Performance (MRR) comparisons of retrieval model's results between state-of-the-art methods on CityFlow-NL. * In MRR column, $Red_{(1)}$ is for the best MRR, $Blue_{(2)}$ is for the second best MRR. * In Parameters column, $Red_{(1)}$ is for the smallest parameters, $Blue_{(2)}$ is for the second smallest parameters.

Method	MRR	Parameters
Baidu-SYSU [4]	$37.05\%_{(1)}$	300.18 M
Megvii [3]	32.52%	391.17 M
HCMUS-UDayton [43]	31.37%	$151.80\ M_{(2)}$
BUPT-ChinaMobile [25]	30.12%	333.43 M
Terminus-CQUPT [26]	20.05%	420.98 M
DAKRS (ours)	$33.58\%_{(2)}$	$150.02\ M_{(1)}$

SSDA retrieval model on the CityFlow-NL dataset during 2 training stages, and in TABLE 2. we show the amount of training data used for the baseline retrieval model and SSDA retrieval model. Then, we compare the results of the proposed model with other existing methods on the CityFlow-NL dataset. Finally, we thoroughly compare the performance of each stage of our proposed DAKRS retrieval system with other methods participating in the challenge, including state-of-the-art solutions on CityFlow-NL dataset such as Baidu-SYSU [4], Megvii [3], HCMUS-UDayton [43], BUPT-ChinaMobile [25], Terminus-CQUPT [26].

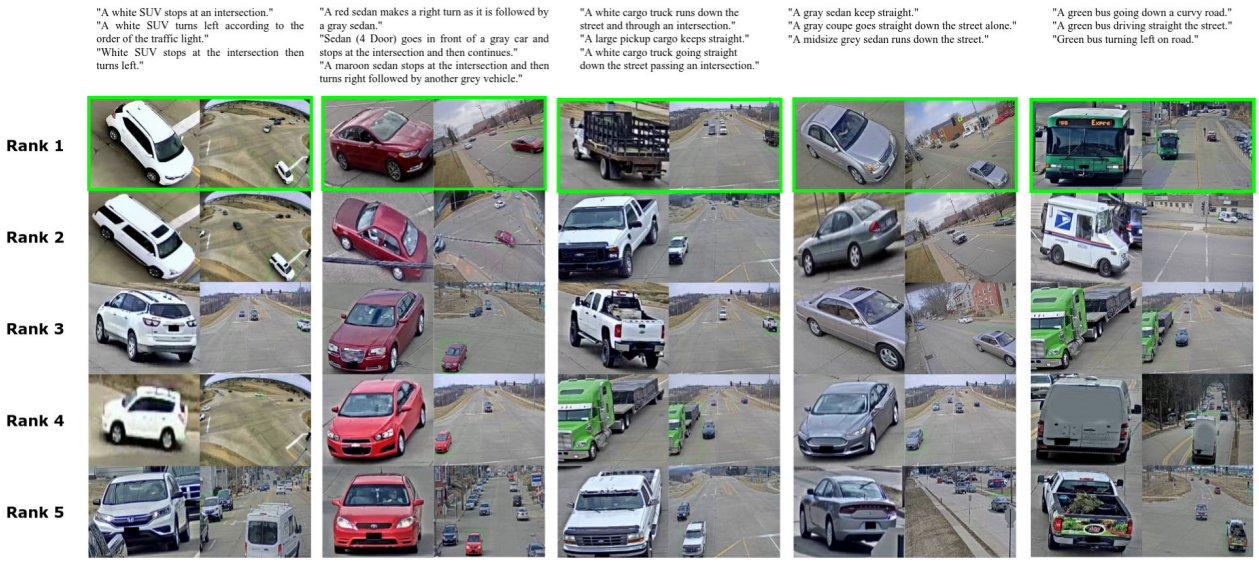


FIGURE 5. Qualitative results on our baseline retrieval model. Despite having lower accuracy, the baseline model produces several reasonable matches. Green vehicle images indicate the true top-1 images match the text descriptions, where the baseline retrieval model achieves top-1 results.

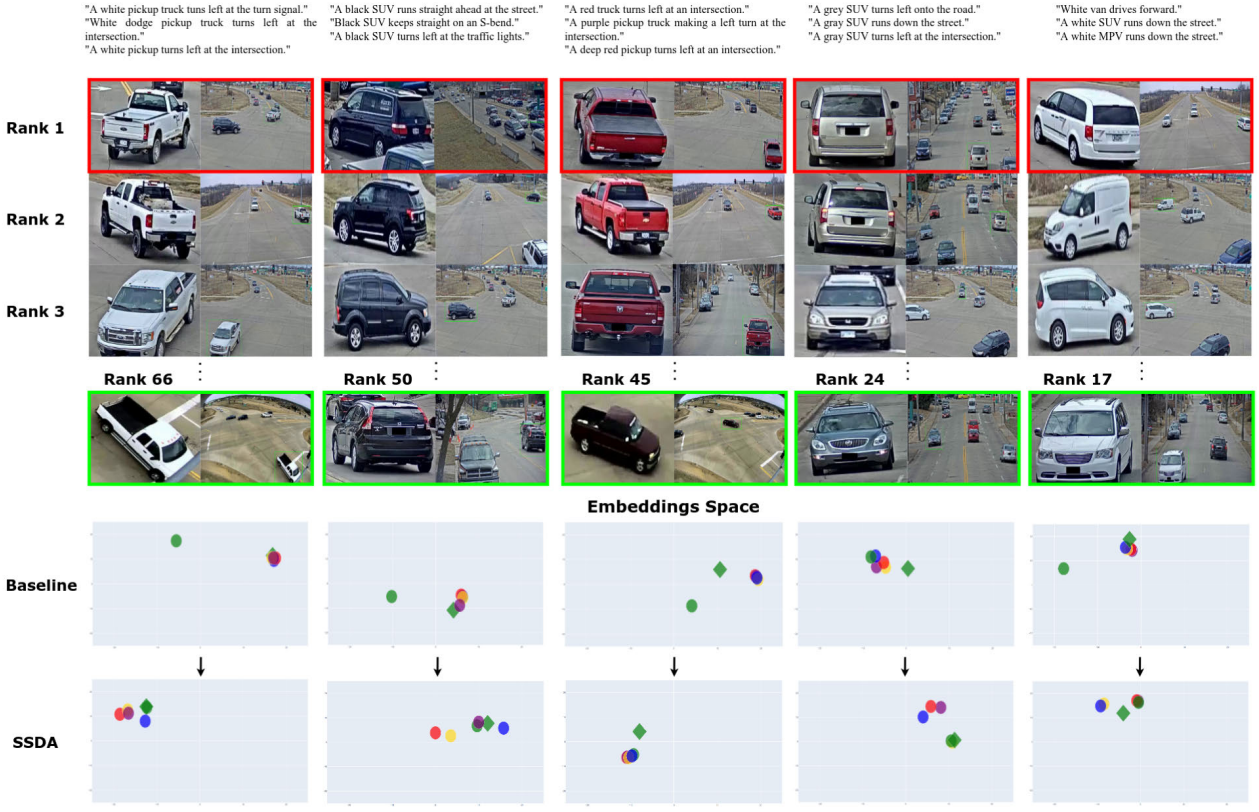


FIGURE 6. Qualitative results for SSDA training stage. These are example cases where the baseline's embeddings are not reasonably close, but after SSDA the embeddings are closer to the ground truth. (Upper Figure) Red vehicle images indicate the false top-1 images match with the text descriptions, where the baseline retrieval model achieves low ranking. Green vehicle images indicate the true top-1 images match with the text descriptions, after using SSDA our retrieval model produces top-1. (Lower Figure) Green icons indicate the target embedding pairs, other colors denote unmatched predicted embeddings. The circle icon refers to image embedding, and the diamond icon refers to text embedding.

The experimental results on CityFlow-NL of baseline retrieval models and whole retrieval systems are detailed in TABLE 3. and TABLE 4., respectively. TABLE 3. shows that our method achieves the second-best results with 33.44% on MRR metrics while requiring the least computational cost where the DAKRS baseline model's parameters are half

TABLE 4. Overall framework performance (MRR) comparisons with other methods on CityFlow-NL. * In each column, *Red*$_{(1)}$ is for the best MRR, *Blue*$_{(2)}$ is for the second best MRR. † Methods that have not published papers and disclosed their methods to validate their result with NVIDIA, whose MRR results are taken from the public leaderboard available at https://eval.aicitychallenge.org/aicity2022.

Method	MRR	Ensemble	Post-Process
Baidu-SYSU [4]	66.06%$_{(1)}$	Yes	Yes
Thursday †	52.51%	—	—
HCMIU-CVIP [6]	47.73%	Yes	Yes
Megvii [3]	43.92%	Yes	Yes
HCMUS-UDayton [43]	36.11%	Yes	Yes
P & L †	33.38%	—	—
Terminus-CQUPT [26]	33.20%	Yes	**No**
MARS_WHU †	32.05%	—	—
BUPT-ChinaMobile [25]	30.12%	Yes	**No**
folklore †	28.32%	—	—
HYFL †	28.04%	—	—
alpha †	28.02%	—	—
SEEE-HUST †	23.33%	—	—
ETRI_AIA †	03.89%	—	—
DAKRS (ours)	63.20%$_{(2)}$	**No**	**No**

compared to the top-1 team Baidu-SYSU model's parameters, while the MRR is less than 3.47%. Moreover, our proposed baseline achieves better MRR than the top-3 solution from the Megvii team with a difference of 1.06% MRR while requiring significantly fewer computational costs, which are less than half the parameters required by Megvii's model. Our proposed baseline model shows a promising balance between effectiveness and efficiency compared to other approaches, where the model can achieve robust retrieval results while maintaining the least computational costs.

In TABLE 4. we report the performance of DAKRS retrieval results on the CityFlow-NL benchmark compared to other approaches in NVIDIA AI City Challenge 2022. Our retrieval system obtains 63.20% MRR using only a single two-stage retrieval model without any further post-processing methods on retrieval results or ensemble methods with multiple models. Compared with other methods, our Domain-Adaptive Knowledge Retrieval System (DAKRS) achieves the second-best performance in MRR metrics with a small gap of 2.86% MRR compared to the top-1 team while maintaining notable results compared to other teams where our result is 15.47% and 27.09% higher than [3] and [43], respectively. Interestingly, our proposed system improves up to 15.47% MRR compared to our preliminary version.

2) QUALITATIVE RESULTS

We visualize the ranking of retrieval results as the qualitative results where the difference in the performance of the baseline model and after SSDA is shown in FIGURE 5. and FIGURE 6. respectively. Although three descriptions contain different fine-grained level information about the vehicle in each query, the proposed method can still find the right matches with the highest similarity to achieve top-1 in the ranking. Furthermore, all the top-5 tracks are relevant to the query descriptions and, without proper observations, may be indistinguishable even to humans. Thus, validating the effectiveness of the proposed baseline model.

In FIGURE 6., we visualize retrieval cases in which the baseline model ranks remarkably wrong in the target video, which is placed at a very low ranking. We can observe that the top-3 results retrieved from the baseline model are identical to the vehicle's appearance query descriptions. The difference between the target and top-3 vehicles is dynamic attributes such as trajectories. However, with different camera viewpoints, this information is hard to distinguish and can lead to confusion even for humans. Also, we notice that the baseline model ranks vehicles with scenarios that appear in the train set higher than unseen scenarios which causes domain bias between seen and unseen scenarios and leads to poor performance in cases where target vehicles are in unseen scenarios. Thus to tackle these problems, the SSDA approach is alleviated. To get an intuitive view of how SSDA works, we use the t-SNE algorithm to transform the multi-modal representation into a pair of 2-dimension feature points. We provide the visualization of the embeddings space of those unseen scenarios cases between baseline models and after SSDA, where we visualize the target pair and top-4 result image embeddings for comparison. In the baseline model, the distance between the text embedding and image embedding of the target vehicle is very far apart, while top-4 image embeddings are very close, which verifies the domain bias problem of the baseline model. But, after the SSDA stage, the target vehicle pair of embeddings are pulled closer to each other and push other image embeddings far away from text embeddings. We can infer that the SSDA approach helps create a pseudo-data point that acts as an intermediary to pull the target pair closer to each other and simultaneously moves the text embedding far away from the wrong image embeddings. Therefore, illustrate the effectiveness of the SSDA in resolving domain bias and model confusion due to ambiguous information.

Finally, as illustrated by FIGURE 7. we observe that the embeddings appear clustered into large clusters in the baseline retrieval model's representation embedding space. This shows that the difference between inter-class embeddings is quite small, where a text embedding can be surrounded by multiple image embeddings with similar static properties (e.g., color and type) and mostly appear in the same scenario. This leads to a mismatch in the retrieval process and drastically decreases the MRR score. While in SSDA retrieval model's embedding space, we can see that each cluster contains fewer embeddings while the distance between each cluster is farther and more sparse than the one in the baseline. Hence, it suggested that SSDA resolve the domain bias problems with large intra-class variation and small inter-class differences by creating pseudo-embedding near the real embedding and pulling each true pair closer while pushing the false pair away, reducing the false pair with high similarity during the matching embeddings process.

FIGURE 7. Embedding representation space between the Baseline and SSDA retrieval model on CityFlow-NL. We can qualitatively observe the density and sparsity between NL-Track matches. Circle icons denote visual embeddings, and diamond icons denote textual embeddings.

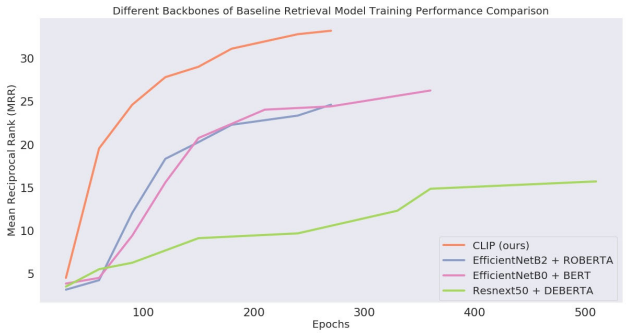


FIGURE 8. Training performance of the baseline retrieval model using CLIP and different pre-trained models as backbones on CityFlow-NL.

TABLE 5. Experiments to verify the robustness of CLIP when using as the backbone to extract text-image features on CityFlow-NL compared to other backbones.

Method	MRR	Recall@5	Recall@10	Params
EfficientNetB2 + ROBERTA	24.65%	31.52%	54.35%	398.16 M
EfficientNetB0 + BERT	26.30%	35.33%	51.63%	147.31 M
Resnext50 + DEBERTA	15.72%	21.20%	29.89%	247.23 M
CLIP-based Baseline (ours)	**33.58%**	**46.20%**	**67.39%**	**150.02 M**

E. ABLATION STUDY

To validate the effectiveness of each component in DAKRS and find an optimal structure, we conduct extensive experiments on the CityFlow-NL dataset to obtain TABLE 5. and TABLE 6.

1) THE VALIDATION OF CLIP

In TABLE 5. we conduct several experiments on our baseline model by replacing CLIP with different textual and visual feature extractors to validate the robustness and effectiveness of CLIP. Through thorough observations, the results show that CLIP outperforms the experiments conducted by three other backbones combinations in terms of effectiveness and efficiency, with the convergence time faster than other backbones as in FIGURE 8. illustrated. In detail, for visual feature extractors, we experiment on three models with two EfficientNet [44] variants B0, B2 and Resnext50 [45]. As for textual feature extractors, we tested on three variants of BERT models, namely BERT [24], ROBERTA [46] and DEBERTA [47].

TABLE 6. Experiments to verify "Concept Space Learning" refers to "CSL" and "Semi-Supervised Domain Adaptation" refers to "SSDA" effectiveness on CityFlow-NL. "w/o" refers to "without", "w/" refers to "with."

Baseline	MRR	Recall@5	Recall@10
w/ Original text format	33.58%	46.20%	67.39%
w/o CSL	28.21%	40.22%	59.24%
w/ SSDA	55.21%	67.39%	83.70%
w/ Standardized text format	30.14%	39.13%	65.22%
w/o CSL	28.41%	39.67%	58.15%
w/ SSDA	**63.20%**	**79.35%**	**93.48%**

When experimenting with EfficientNetB2 and ROBERTA we observe that the model converged at epoch 270th with 24.65% MRR, smaller than CLIP 8.93%, while the model's parameters are more than double CLIP's. For EfficientNetB0 and BERT, the model converged much slower than CLIP at epoch 330th while the number of parameters is smaller than CLIP with nearly 3 M params, but the performance is significantly less than with a 7.28% MRR difference. Finally, with Resnext50 and DEBERTA, the model converged at epoch 510th however the performance is notably lower compared to CLIP with a 17.86% MRR difference while the number of parameters is more than half of CLIP. This implies the robustness of CLIP in constructing strong representations for textual-visual tasks, and further supports the idea of multi-modal robust embeddings that are disentangled from varying degrees of inessential variations.

2) THE EFFECTIVENESS OF SSDA

In our two-stage retrieval where the second stage uses SSDA to resolve the domain gaps and align the distributions between the train set (source domain) and test set (target domain). However, to tackle the ambiguity of the queries, we propose a standardized version of the queries for creating pseudo-labels for the Domain Adaptation approach. We thoroughly validate the performance of standardized text formats with the baseline model and SSDA approach on the CityFlow-NL dataset. From TABLE 6. we can observe that standardized text formats lead to a decrease of 3.44% in MRR, 7.07% in Recall@5 and 2.17% in Recall@10. However, using the SSDA approach on the baseline with standard text format achieves higher results than the original text format and brings a difference of 7.99% in MRR, 11.96% in Recall@5 and 9.78% in Recall@10. This indicates that standardized text format confuses the baseline model in discriminating different text-image pairs due to the omission of unique information about the vehicle in each text description and making some descriptions nearly identical. However, the increase in performance after the SSDA approach shows that due to pseudo-label of the SSDA approach in standardized text format and the inconsistency between the new and original format can lead to a decline in performance. Thus, maintaining the similarity of data between the two training stages

is crucial for the performance of the SSDA strategy. Furthermore, the SSDA results between our preliminary version and current works have a difference of 15.47% in MRR, 13.05% in Recall@5, and 13.05% in Recall@10 validate our improvement of the data enhancement method and further emphasize the essence of data quality and data-driven properties of our SSDA approach.

3) THE EFFECTIVENESS OF CSL

Our two-stage retrieval model consists of two objective learning functions: Latent space learning, also referred to as contrastive learning, as a common approach to retrieval tasks, and Concept Space Learning or CSL, which enforces the model to learn at the instance level for local feature alignment. We separately validate the effectiveness of CSL on the CityFlow-NL dataset in TABLE 6. We have the following observations: CSL brings an improvement of 5.37% and 1.73% in MRR, 5.98% and 0.54% in Recall@5, 8.15% and 7.07% in Recall@10 with baseline model on two different text formats which are original text format and standardized text format, respectively. This demonstrates the effectiveness of leveraging local feature information and indicates that instance-level information can boost the performance of contrastive learning in discriminating different text-image pairs.

V. CONCLUSION AND FUTURE WORK

In this work, we have looked at the usage of multi-modal pre-trained models in the vehicle retrieval task through text descriptions and further enhance the retrieval performance of a single retrieval model without any further ensemble or post-processing method through a simple but effective SSDA method using the generated pseudo-data and pseudo-label from our proposed data enhancement method. Undoubtedly, our proposed CLIP-driven retrieval model shows the effectiveness of the CityFlow-NL dataset with the second-best result while requiring the least amount of computational cost compared to the top-1 performing team's. In addition, to ensure the retrieval model performance can achieve the state-of-the-art result without any further ensemble or post-processing, the SSDA approach shows its capability in minimising the domain shift in representation space between the training and testing sets. And with the support of the data enhancement method, the SSDA approach has shown its exceptional capability in adapting knowledge using pseudo-label and pseudo-data after only a little fine-tuning.

Due to the lack of data for in-depth domain generalization research in this area, we are more inclined to improve our approach in terms of domain adaptation for text-image retrieval tasks. We observe that the representation embedding space of the retrieval model drastically changes after the SSDA fine-tuning, with the distance between embeddings of the image-text true pair being pulled closer while pushing the embeddings of the false pair far away despite having the identical attributes and contexts. In particular, our next goals are to fully leverage the method's capability with meta-data constraints (e.g. traffic scenario, vehicular appearance, etc.) with the hope of reducing the impact of noisy pseudo-data, and we also aim to make the adaptation approach fully self-supervised with self-structured textual data.

ACKNOWLEDGMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2022-28-04. The authors would like to thank International University-Vietnam National University Ho Chi Minh City for facilitating this work and also would like to thank their colleagues for their support, which significantly improved the manuscript.

(Synh Viet-Uyen Ha, Huy Dinh-Anh Le, and Quang Qui-Vinh Nguyen are co-first authors.)

REFERENCES

[1] Q. Feng, V. Ablavsky, and S. Sclaroff, "CityFlow-NL: Tracking and retrieval of vehicles at city scale by natural language descriptions," 2021, *arXiv:2101.04741*.

[2] S. Bai, Z. Zheng, X. Wang, J. Lin, Z. Zhang, C. Zhou, H. Yang, and Y. Yang, "Connecting language and vision for natural language-based vehicle retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4029–4038.

[3] C. Zhao, H. Chen, W. Zhang, J. Chen, S. Zhang, Y. Li, and B. Li, "Symmetric network with spatial relationship modeling for natural language-based vehicle retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3226–3233.

[4] J. Zhang, X. Lin, M. Jiang, Y. Yu, C. Gong, W. Zhang, X. Tan, Y. Li, E. Ding, and G. Li, "A multi-granularity retrieval system for natural language-based vehicle retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3215–3224.

[5] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[6] H. D. Le, Q. Q. Nguyen, V. A. Nguyen, T. D. Nguyen, N. M. Chung, T.-T. Thái, and S. V. Ha, "Tracked-vehicle retrieval by natural language descriptions with domain adaptive knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3299–3308.

[7] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1441–1451. [Online]. Available: https://aclanthology.org/P19-1139

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Red Hook, NY, USA: Curran Associates, 2013.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[11] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval," 2021, *arXiv:2104.08860*.

[12] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: CLIPBERT for video-and-language learning via sparse sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7327–7337.

[13] Z. Wang, Y. Wu, K. Narasimhan, and O. Russakovsky, "Multi-query video retrieval," 2022, *arXiv:2201.03639*.

[14] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, X. Yang, Y. Yao, L. Zheng, P. Chakraborty, A. Sharma, Q. Feng, V. Ablavsky, and S. Sclaroff, "The 5th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4258–4268.

[15] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. Shaiqur Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa, "The 6th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3347–3356.

[16] Z. Sun, X. Liu, X. Bi, X. Nie, and Y. Yin, "DUN: Dual-path temporal matching network for natural language-based vehicle retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4056–4062.

[17] T. M. Nguyen, Q. H. Pham, L. B. Doan, H. V. Trinh, V.-A. Nguyen, and V.-H. Phan, "Contrastive learning for natural language-based vehicle retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4240–4247.

[18] S. Lee, T. Woo, and S. H. Lee, "SBNet: Segmentation-based network for natural language-based vehicle search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4049–4055.

[19] P. Khorramshahi, S. S. Rambhatla, and R. Chellappa, "Towards accurate visual and natural language-based vehicle retrieval systems," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4178–4187.

[20] C. Sebastian, R. Imbriaco, P. Meletis, G. Dubbelman, E. Bondarev, and P. H. N. de With, "TIED: A cycle consistent encoder–decoder model for text-to-image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4133–4141.

[21] T.-P. Nguyen, B.-T. Tran-Le, X.-D. Thai, T. V. Nguyen, M. N. Do, and M.-T. Tran, "Traffic video event retrieval via text query using vehicle appearance and motion attributes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4160–4167.

[22] E.-J. Park, H. Kim, S. Jeong, B. Kang, and Y. Kwon, "Keyword-based vehicle retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4215–4222.

[23] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[25] Y. Du, B. Zhang, X. Ruan, F. Su, Z. Zhao, and H. Chen, "OMG: Observe multiple granularities for natural language-based vehicle retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3123–3132.

[26] B. Xu, Y. Xiong, R. Zhang, Y. Feng, and H. Wu, "Natural language-based vehicle retrieval with explicit cross-modal representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3141–3148.

[27] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 869–884.

[28] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[29] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," 2015, *arXiv:1506.08959*.

[30] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," 2019, *arXiv:1903.09254*.

[31] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9876–9886.

[32] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.

[33] Z. Wang, J. Yu, A. Wei Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," 2021, *arXiv:2108.10904*.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[35] N. Omar and S. S. Hasbullah, "SRL TOOL: Heuristics-based semantic role labeling through natural language processing," in *Proc. Int. Symp. Inf. Technol.*, 2008, pp. 1–7.

[36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.

[37] S. V.-U. Ha, N. M. Chung, H. N. Phan, and C. T. Nguyen, "TensorMoG: A tensor-driven Gaussian mixture model with dynamic scene adaptation for background modelling," *Sensors*, vol. 20, no. 23, p. 6973, Dec. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/23/6973

[38] S. V. Ha, N. M. Chung, T.-C. Nguyen, and H. N. Phan, "Tiny-PIRATE: A tiny model with parallelized intelligence for real-time analysis as a traffic countEr," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4114–4123.

[39] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[40] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8221–8230.

[41] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X.-W. Guo, F. Huang, R. Ji, and S. Li, "Asymmetric co-teaching for unsupervised cross domain person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1–15.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[43] T.-L. Nguyen-Ho, M.-K. Pham, T.-P. Nguyen, H.-D. Nguyen, M. N. Do, T. V. Nguyen, and M.-T. Tran, "Text query based traffic video event retrieval with global-local fusion embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3133–3140.

[44] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

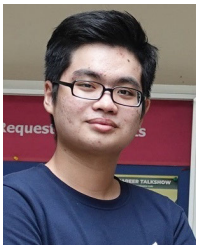
[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016, *arXiv:1611.05431*.

[46] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[47] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arXiv:2006.03654*.

SYNH VIET-UYEN HA (Senior Member, IEEE) received the B.Eng. degree in information technology from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, in 1996, the M.Sc. degree in computer science from the Ho Chi Minh City University of Science (HCMUS), Vietnam, in 1999, and the Ph.D. degree in electrical and computer engineering from Sungkyunkwan University, South Korea, in 2010. After working as a Postdoctoral Scholar with the Automation Laboratory, Sungkyunkwan University, he joined Vietnam National University—Ho Chi Minh City International University (VNU-HCMIU) as a Lecturer, where he is currently the Head of the Computer Vision and Image Processing Laboratory. His research interests include signal processing, computer vision, machine learning, and deep learning, especially with applications in smart traffic systems.

HUY DINH-ANH LE (Student Member, IEEE) is currently pursuing the bachelor's degree in computer science with Vietnam National University—Ho Chi Minh City International University (VNU-HCMIU), Ho Chi Minh City, Vietnam. He is also a member of the Olympiad Team in Informatics, VNU-HCMIU, where he has been doing research related to video-based surveillance systems under the supervision of Dr. Synh Viet-Uyen Ha with the Computer Vision and Image Processing Laboratory, since 2021, to investigate the topic of contrastive representation learning, and generative models in image restoration. His current research interests include image processing, computer vision, and artificial intelligence.



QUANG QUI-VINH NGUYEN is currently pursuing the bachelor's degree in computer science with Vietnam National University—Ho Chi Minh City International University (VNU-HCMIU), Ho Chi Minh City, Vietnam. He is currently pursuing the B.Eng. degree in computer science. He is also a co-coach and a member of the Olympiad Team in Informatics. He has been doing research related to video-based surveillance, and been working under the supervision of Dr. Synh Viet-Uyen Ha with the Computer Vision and Image Processing Laboratory, VNU-HCMIU, since 2020, to investigate the topic of contrastive representation learning of multi-modal retrieval systems. His current research interests include image processing, computer vision, and artificial intelligence.

NHAT MINH CHUNG received the B.Eng. degree in computer science from Vietnam National University—Ho Chi Minh City International University (VNU-HCMIU), Vietnam, in 2021, where he is currently pursuing the M.Eng. degree in computer science. He has been working under the supervision of Dr. Synh Viet-Uyen Ha with the Computer Vision and Image Processing Laboratory, VNU-HCMIU, since 2019, to investigate the topic of background modeling and change detection in vision-based surveillance systems. His current research interests include computer vision, machine learning, representation learning, and semantic segmentation.