Towards Accurate Visual and Natural Language-Based Vehicle Retrieval Systems

Pirazh Khorramshahi *
Johns Hopkins University
pkhorra1@jhu.edu

Sai Saketh Rambhatla*
University of Maryland, College Park
rssaketh@umd.edu

Rama Chellappa
Johns Hopkins University
rchella4@jhu.edu

Abstract

In this work, we consider two tracks of the 2021 NVIDIA AI City Challenge, the City-Scale Multi-Camera Vehicle Re-identification and Natural language-based Vehicle Retrieval. For the vehicle re-identification task, we employ the state-of-art Excited Vehicle Re-Identification deep representation learning model coupled with best training practices and domain adaptation techniques to obtain robust embeddings. We further refine the re-identification results through a series of post-processing steps to remove camera and vehicle orientation bias that is inherent in the task of re-identification. We also take advantage of multiple observations of a vehicle using track-level information and finally obtain fine-grained retrieval results. For the task of Natural language-based vehicle retrieval we leverage the recently proposed Contrastive Language-Image Pre-training model and propose a simple yet effective text-based vehicle retrieval system. We compare our performance against the top submissions to the challenge and our systems are ranked 8^{th} in the public leaderboard for both tracks.

1. Introduction

Lately, there has been a great focus on the realization of automated and intelligent transportation systems at different scales. An autonomous vehicle can benefit from such automated systems to significantly reduce the risk of accidents, improve passenger experience and minimize travel distances. On the other hand, intelligent transportation systems can help large-scale traffic camera networks to dynamically learn traffic patterns, manage the flow of traffic, collect vehicle-level analytics such as speed, and retrieve a vehicle of interest using different attributes and descriptions. In recent years, the development of efficient and high performing deep neural networks has made such advancements quite possible. In addition, the NVIDIA AI City Challenge has facilitated the path to the realization of smart transportation systems during the past couple of years. In the 2021 version of this challenge, we participated in the tasks of City-Scale Multi-Camera Vehicle Re-Identification and Natural Language-Based Vehicle Retrieval.

Vehicle Re-Identification is the task of locating all instances of a particular vehicle identity in a gallery set consisting of a large volume of vehicle images which have been captured under diverse conditions using a network of traffic cameras. This is particularly challenging as vehicles with different identities can be of same manufacturer, model, year and color, resulting in very small inter-class variation. In addition, multiple images of a single vehicle identity can look significantly different under different view-points and angles, resulting in large intra-class variation. Therefore, learning highly discriminative and robust embeddings that are capable of handling both inter- and intra-class variations is critical. In this work, we employ the fast and accurate Excited Vehicle Re-Identification (EVER) [27] model which benefits from the Self-Supervised Residual Generation module [15] to excite the intermediate feature maps to learn robust embeddings. Moreover, we train EVER within the framework of FASTREID [11] using state-of-the art re-identification algorithm training techniques.

Natural Language-Based vehicle retrieval is a multi-modal task for retrieving single-camera tracks of vehicles that are consistent with a natural language query describing its visual and motion patterns. This is the first time AI City Challenge has introduced this task. Text-based image retrieval is inherently challenging due to the ambiguity of textual descriptions. Secondly, using single camera tracks might make the task of discerning the model of a vehicle hard resulting in poor retrieval results. Therefore, it is essential to train powerful multi-modal models that can effectively deal with such difficulties. In this work, we leverage a recently proposed multi-modal model CLIP (Contrastive Language-Image Pre-training) [29]. CLIP jointly trains a visual and text encoder by leveraging natural language descriptions as supervisors to learn powerful image representations and has demonstrated impressive zero shot performance in many image recognition tasks [29]. We design a

*The first two authors equally contributed to this work.

simple yet effective natural language-based vehicle retrieval system that given natural language queries, ranks tracks of vehicles using cosine similarity between visual and language features extracted from CLIP.

The paper is organized as follows. In Section 2, we briefly describe some of the recent works on vehicle re-identification and natural language-based retrieval. Then, we describe our method, experiments and results for the Vehicle Re-identification and Natural language-based vehicle retrieval tasks in Sections 3 and 4 respectively. Finally, in Section 5 we briefly summarize our efforts for the 2021 NVIDIA AI City Challenge and suggest a few directions for future research.

2. Related Works

Vehicle Re-Identification: Here we briefly review several recent and most relevant works in the area of vehicles re-identification. To learn discriminative vehicle embeddings, several large-scale re-id benchmarks have been proposed. VeRi [19], VehicleID [18], VERI-Wild [22] and Vehicle 1-M [7] have made it possible to learn robust visual features based on deep learning. Introduction of synthetic data [40] with diverse attributes has also been shown to contribute to the performance of re-identification models [41]. While learning global visual features of vehicles can be done in a straightforward fashion, learned embeddings are not robust to occlusion and changes in view-points [37]. In addition, the extracted features may usually fail to distinguish two similar looking vehicles that are of same make, model, color and year. Therefore, extracting local features from discriminating regions of vehicles plays a critical role. [14, 8] explored the idea of supervised attention in the form of vehicle key-points and vehicle parts location. In addition, the idea of image alignment based on local regions while extracting the features is shown to be effective [20]. Due to the scarcity of additional annotations to perform supervised attention, self-supervised models have been developed to overcome this bottleneck. [15, 27] by generating pseudo-saliency maps. As an alternative to convolutional neural networks (CNN), with the development of transformer models for visual domain [4], the idea of self-attention has been studied. In [12], the authors show that transformer-based models can yield competitive results to those based on CNNs.

Natural Language-Based Retrieval : Learning Cross-Modal (image-text) representations is fundamental to a wide range of vision-language (V+L) tasks, such as visual question answering, image-text retrieval, image captioning/grounding etc. Transformer-based [34] natural language models like BERT [3], have resulted in successful adaptation of similar architectures and training techniques to image and image-text representation learning. Lu et. al [23] proposed ViLBERT that extends BERT [3] to a multi-modal two stream architecture with novel Co-Attention transformer layers for learning task agnostic joint representations of image content and natural language which has shown impressive results on twelve different vision and language tasks [24]. Li et. al [17] argue that the self attention mechanism employed in contemporary Vision Language Pretraining (VLP) methods lack explicit alignment between image regions and text. To alleviate this issue, they propose OSCAR, a novel VLP method that leverages object tags detected in images as anchor points to facilitate efficient semantic alignment between image regions and text. Chen et. al [1] propose UNITER that uses conditional masking on pre-training tasks as opposed to the joint random masking of both modalities done in contemporary methods. UNITER is trained using four pre-training tasks namely Masked Language Modeling (MLM), Masked Region Modeling, Image-Text Matching (ITM), and Word-Region Alignment (WRA). While ITM helps achieve global image-text alignment, the proposed WRA leverages Optimal Transport (OT) to explicitly encourage fine-grained alignment between words and image regions during pre-training. While all the methods discussed above predict the exact word of the text using transformer based architectures, CLIP [29] is trained on a relatively easier task of matching the image to the right caption. CLIP is trained in a contrastive fashion using a symmetric cross entropy loss to assign a high similarity score to the correct (image, text) pair while simultaneously reducing the score for the incorrect pairings. CLIP has demonstrated impressive results in many image recognition tasks [29] and we employ it to solve the Natural language-based vehicle retrieval problem. Finally, datasets used for training and evaluating all of the text retrieval systems described above are generic, and in this paper we work with visual feed of vehicles. Such domain specific data presents its own set of challenges and have to be addressed appropriately.

3. Vehicle Re-Identification

In this section, we present our method for the City-Scale Multi-Camera Vehicle Re-Identification track of the 2021 NVIDIA AI City Challenge. Our approach has three distinct stages, namely Pre-Processing, Deep Feature Extraction, and Post-Processing. Figure 1 shows the overview of our proposed pipeline.

3.1. Pre-Processing

The 2021 edition of CityFlow Re-ID dataset [32] has 85058 images in total which are split among training, testing and query sets of size 52717, 31238 and 1103 respectively. The training data consists of 440 identities. To prepare the training data we performed margin removal and domain adaptation techniques as described in the following sections.

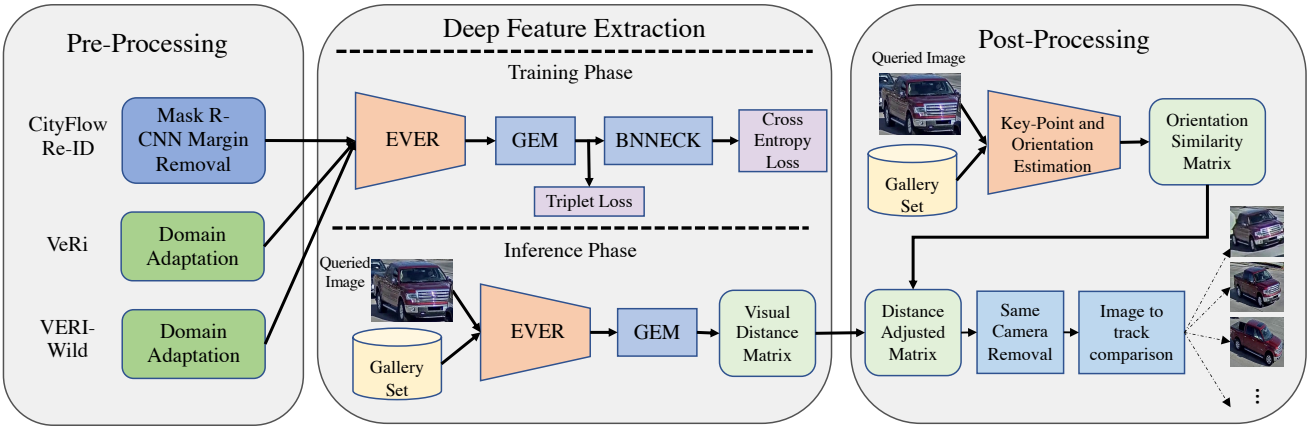

Figure 1: Our proposed approach for the task of City-Scale Multi-Camera Vehicle Re-Identification. The proposed pipeline consist of three distinct stages, namely pre-processing, deep feature extraction and post-processing.

3.1.1 Margin Removal (MR)

Cropped vehicle images in CityFlow Re-ID dataset often have a significantly large margin which can be viewed as background distractors. Therefore, following the practice of [16, 41], we use Mask R-CNN object detector [9] implemented in Detectron2 [39] to randomly tighten the bounding box of a vehicle. This process helps the representation learning model to better focus on the vehicle and its discriminative regions. Figure 2 demonstrates the impact of margin removal operation.

3.1.2 Domain Adaption (DA)

Typically high capacity deep learning models perform better with the introduction of additional training data with similar characteristics that resembles the original training data in terms of probability distribution of images. Hence, in an attempt to increase the size of training set, we use the two publicly available multi-view vehicle re-identification datasets, namely VeRi [19] and VERI-Wild [22]. However, the domain of these datasets is different from that of CityFlow Re-ID dataset. To compensate for this domain gap, we use CycleGAN [44] to perform the task of unpaired Image-to-Image translation. Therefore, we learn two mapping functions, G_1 and G_2 that can map images of VeRI and VERI-Wild datasets to the domain of CityFlow Re-ID dataset respectively. Figure 3 shows the transformation on a VeRi dataset image to the domain of CityFlow Re-ID dataset.

3.2. Deep Feature Extraction

To deal with the aforementioned inter- and intra-class similarities and variations that is prevalent in the vehicle re-identification task, accurate deep neural network models are required. These models should focus the attention to subtle details in the vehicle images and extract robust embeddings. Therefore, we choose to employ Excited Vehicle Re-Identification (EVER) model which was among the top performers of the City-Scale Multi-Camera Vehicle Re-Identification track of 2020 NVIDIA AI City Challenge.

EVER has a built-in self-supervised residual generation module inspired by [15] that can highlight the high-level details of vehicle corresponding to its identity and can help to distinguish the vehicle's identity from others. During the course of training, EVER excites its intermediate feature maps with the goal of attending to the discriminative regions within the vehicle image. However, as the training progresses the amount of excitation reduces so that once the model is fully trained, no more excitation is done. This can significantly improve the inference time as it only involves a single forward pass of the backbone ResNet [10] network in our case.

In addition, the global pooling layer of the backbone ResNet architecture that comes after the Res-5 block, is typically an average pooling layer. In our work, we replaced this with a learnable Generalized-Mean (GEM) pooling layer [28] due to the enhancements observed in FASTREID framework for re-identification tasks, with the following formulation:

$$f_c(x) = \left(\frac{1}{|W * H|} \sum_{i=1}^{W} \sum_{j=1}^{H} (x_{c,i,j})^p \right)^{1/p} \quad (1)$$

In Eq. 1, $f_c(x)$ is the c^{th} channel of the GEM layer's output for an input feature map x of shape $C * W * H$. Further, the pooling parameter p is trainable and is initiated with value of 1, *i.e.* the GEM layer performs average pooling operation initially. After training, the final value of p is 2.79.

3.2.1 Optimization Objective Functions

To train EVER, we employ Triplet [13] and Cross entropy loss functions. To ensure intra-class compactness while having larger inter-class distances, triplet loss tries to make the distance between an anchor and its positive pair smaller than the distance between the anchor and its negative pair by a distance margin. The integration of triplet loss with the batch hard sampling method is achieved by minimizing the loss function given below.

$$\mathcal{L}_t = \frac{1}{B}\sum_{i=1}^{B}\sum_{a\in b_i}\left[\gamma + \max_{p\in\mathcal{P}(a)}||x_a - x_p||_2 - \min_{n\in\mathcal{N}(a)}||x_a - x_n||_2\right]_+ \quad (2)$$

In Eq. 2, B, b_i, a, γ, $\mathcal{P}(a)$ and $\mathcal{N}(a)$ are the total number of batches, i^{th} batch, anchor sample, distance margin threshold, positive and negative sample sets corresponding to a given anchor respectively. Moreover, x_a, x_p, x_n are the extracted features for anchor, positive and negative samples. For the purpose of this loss, batches are constructed in a way that they have exactly 16 instances of each ID used.

In addition, the Cross entropy loss with label smoothing technique [31] to alleviate the issue of over-fitting is used. Note that to effectively apply both cross entropy and triplet losses to the extracted features, Batch Normalization Neck (BNNECK) [25] has been inserted into EVER. The Cross entropy loss is calculated as follows:

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left[y_j^i \log \hat{y}_j^i + (1-y_j^i)\log(1-\hat{y}_j^i)\right] \quad (3)$$

Where $\hat{y}_j^i = \log \frac{e^{(W_j^T x_i + b_j)}}{\left(\sum_{j=1}^{C} e^{W_j^T x_i + b_j}\right)}$ is the predicted logit corresponding to class j for the extracted feature x_i of the i^{th} training sample after applying the softmax layer. Furthermore, in Eq. 3, W_j, b_j are the classifier's weight vector and bias associated with j^{th} class respectively, and N and C represent the total number of samples and classes in the training dataset. Since we use label smoothing, $y_j^i = 1 - \frac{C-1}{C}\epsilon$ if $j = c$, otherwise $y_j^i = \frac{\epsilon}{C}$ where c is the true label of i^{th} sample.

3.3. Post-Processing

After we train EVER and extract visual embeddings for images in the gallery and query sets we perform a set of post-processing operations to enhance the accuracy of the retrieval process.

3.3.1 Same Camera Removal (SCR)

Images captured through a same camera share similarities in orientation, shape and background that can negatively impact the re-identification results by severely reducing the inter-class distance and lead to failure cases. In the 2021 version of CityFlow Re-ID dataset, camera labels are provided for the test set. Therefore, during the inference, we remove all gallery images with the same camera label as the query image. Given the recent improvements in the area of Multi-object single camera tracking [43, 36], especially on high quality data, the chance of occurring ID switches has reduced. Hence, this is a reasonable assumption. Note that this procedure contributes to the success of re-identification task considerably as discussed in section 3.4.3.

3.3.2 Orientation Bias Removal (OBR)

Although we remove images captured from identical cameras, there might be still query-gallery image pairs that are under similar view-points which in turn can impose a bias on visual similarity computed by EVER. Inspired by [45, 41], we use the key-point and orientation estimation model in [14] to extract orientation embeddings and adjust the distance of an image pair accordingly. To train the key-point and orientation estimation model, we use the domain adapted VeRi dataset, introduced in section 3.1.2, in which images are labeled with key-points and orientation annotations [37]. Afterwards, we extract orientation embeddings and adjust the distance of two given images I_q and I_g as the following:

$$d(I_q, I_g) = ||f(I_q) - f(I_g)||_2 + \lambda \frac{g(I_q).g(I_g)}{||g(I_q)||_2 ||g(I_g)||_2} \quad (4)$$

In Eq. 4, $d(.,.)$, $f(.)$ and $g(.)$ represent the distance of an image pair in L_2 norm, EVER deep feature extractor and key-point and orientation estimation model respectively. The intuition behind Eq. 4 is that images that have similar orientation have smaller visual distance, hence we increase the distance by adding a fraction, *i.e.* λ, of orientation similarity calculated based on cosine similarity.

3.3.3 Image to Track Comparison (ITC)

Relative to image to image comparison, image to track comparison is much more realistic as single camera tracking information is readily available and hence the chance of having similar images within the track to the query image from the perspective of the EVER model increases. CityFlow Re-ID dataset provides track-level information on the test set and we use this knowledge to rank the gallery by only considering the two samples in a track with the least distance to the query image.

3.3.4 Implementation details

We train the EVER model within the framework of FASTREID that employs state-of-the-art training tricks suited

(a) Original Image (b) Tightened

Figure 2: Margin removal via Mask R-CNN Object Detector

(a) Original Image (b) Mapped Image

Figure 3: Translation from VeRi to CityFlow Re-ID dataset

for the task of re-identification and thoroughly investigated in [25, 11].

In our work, all the images have been resized to $320*320$ pixels. Specifically during the training process the following steps are taken:

- As a data augmentation technique, images are randomly flipped with 0.5 probability.
- To address the issue of occlusion, Random Erasing Augmentation [42] step is employed with the goal of encouraging the network to learn more robust representations.
- Searched data augmentation policies from ImageNet dataset based on AutoAugment [2] is applied. The searched policy, involves translation, rotation, and shearing image processing operations. In addition, it carries information on the frequency and the degree to which these operations should be applied.
- To prevent the deep feature extraction model from over-fitting to the initial training batches, the Learning Rate Warm-up [5] technique is used.
- The backbone model is instantiated from ImageNet pre-trained weights. For the purpose of training, the classification layer of this model is replaced to meet the number of training labels and is initiated with randomly distributed weights. To improve this initialization, we used the backbone freeze technique in the first 5000 iterations and merely train the newly added classification layer. Subsequently, the entire model is trained.
- To optimize the deep network, Stochastic Gradient Descent (SGD) with momentum of 0.9 and Cosine Annealing learning rate scheduling [21] is utilized.

3.4. Results

In this part, first we describe the evaluation metric used to measure the performance and then present the evaluation results of the top ten performers of the City-Scale Multi-Camera Vehicle Re-Identification track in 2021 NVIDIA AI City Challenge. Lastly, we perform an ablation study to investigate the impact of each step in the proposed pipeline.

3.4.1 Evaluation Metric

To measure the performance, mean Average Precision (mAP) metric is considered. mAP shows how well a gallery set can be ranked based on a given query set and higher values of this metric shows the superiority of the performance. Note that in the city-scale multi-camera vehicle re-identification track in 2021 NVIDIA AI City Challenge, only the top 100 images in the ranked gallery participated in the mAP calculation.

3.4.2 Leaderboard Rankings

Table 1, shows the top ten performers of the city-scale multi-camera vehicle re-identification track in 2021 NVIDIA AI City Challenge. Our proposed pipeline, achieves the mAP of 62.16% and is ranked 8^{th} among submissions to the public leaderboard.

3.4.3 Ablation Study

Here we investigate the contribution of each modules in the pipeline to final performance. The baseline model is EVER trained with FASTREID framework on CityFlow Re-ID dataset without any further post-processing steps. Table 2 shows the result of this analysis. From Table 2 it can be seen that removing images with same camera label as the queried image significantly boosts the performance. In addition, we can appreciate the gain of margin removal (MR), domain adaptation (DA) on additional data, image to track comparison (ITC), and orientation bias removal to the overall performance.

4. Natural Language-Based Vehicle Retrieval

In this section we describe our method for the natural language based vehicle retrieval track of the 2021 NVIDIA

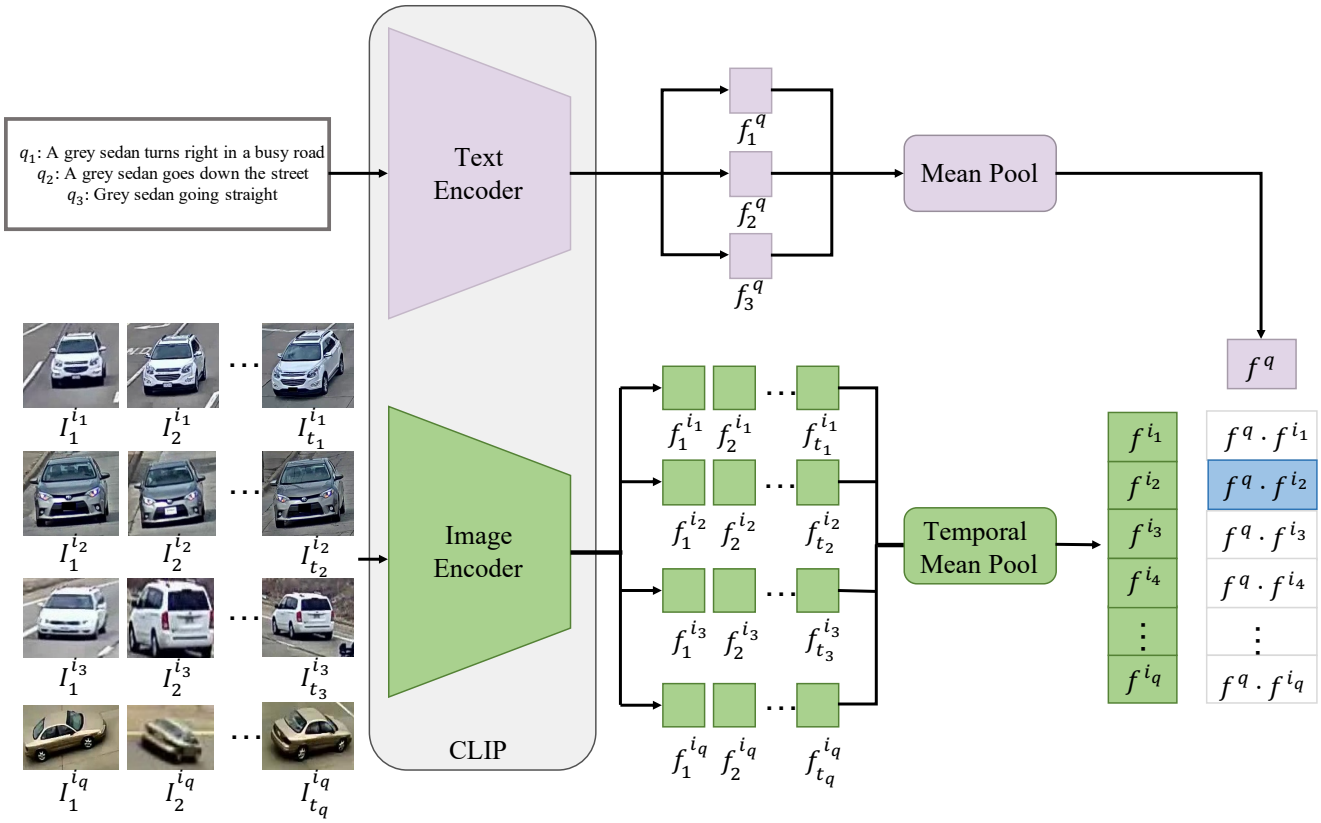

Figure 4: Our system pipeline for Natural Language-based vehicle retrieval. We leverage image and text embeddings from CLIP to compute the average text and track features which are then used to rank all the tracks in the gallery given a natural language query.

Table 1: Top 10 performers of city-scale multi-camera vehicle re-identification track in 2021 NVIDIA AI City Challenge

Rank	Team Name	Score (mAP)
1	DMT	0.7445
2	NewGeneration	0.7151
3	CyberHu	0.6650
4	For Azeroth	0.6555
5	IDo	0.6373
6	KeepMoving	0.6364
7	MegVideo	0.6252
8	**aiem2021 (Ours)**	**0.6216**
9	CyberCoreAI	0.6134
10	Janus Wars	0.6083

Table 2: Ablation Study Results of our proposed pipeline for the city-scale multi-camera vehicle re-identification track.

Model	Score (mAP)
Baseline	0.4019
Baseline + MR + ITC	0.4537
Baseline + MR + ITC + DA	0.4733
Baseline + MR + ITC + DA + SCR	0.6000
Baseline + MR + ITC + DA + SCR + OBR	**0.6216**

AI City Challenge. We first describe the approach followed by the implementation details, results and suggestions for future work.

4.1. Dataset and Metrics

The dataset for this track is built over the CityFlow-NL [6] dataset and consists of 2498 single camera tracks of vehicles for training. Each track is annotated with three natural language descriptions. Results for the challenge are reported on a separate set of 530 unique vehicle tracks to-

Table 3: Top 10 performers of Natural Language-based vehicle retrieval track in 2021 NVIDIA AI City Challenge.

Rank	Team Name	Score (MRR)
1	Alibaba-UTS	0.1869
2	TimeLab	0.1613
3	SBUK	0.1594
4	SNLP	0.1571
5	HUST	0.1564
6	HCMUS	0.1560
7	VCA	0.1548
8	**aiem2021 (Ours)**	**0.1364**
9	Enablers	0.1314
10	Modulabs	0.1195

gether with 530 natural language queries each with three descriptions form the test set.

All the submissions are evaluated on the test set using standard metrics for retrieval tasks, namely, Mean Reciprocal Rank (MRR) [35] and Recall @k ($k \in \{5, 10, 25\}$). **Mean Reciprocal Rank (MRR)** : Each individual query in the test set receives a score of the reciprocal of the rank at which the first correct response was returned. The value is zero if none of the five responses is the correct response. MRR is defined as the average of scores of all the queries.

Recall @k : Recall @k is the proportion of relevant items found in the top-k recommendations.

4.2. Pipeline

In this section, we describe our proposed pipeline for natural language-based vehicle retrieval in details. We adopt the recently proposed Contrastive Language-Image Pre-training (CLIP) [29] model for the task. We use CLIP to extract frame wise features. We average the frame-wise features of a track to obtain the track features. Similarly, we extract language features of all three captions and average them to get an average language descriptor. We then use cosine similarity between the language and track features to rank all the tracks in the gallery. In Section 4.2.1 we briefly describe CLIP and its training algorithm. Finally, in Section 4.2.2 we describe our pipeline. We present our vehicle retrieval pipeline in Figure 4.

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 5: Numpy style pseudo code for CLIP. Figure borrowed from [29]

4.2.1 Contrastive Language-Image Pre-training

Contrastive Language-Image Pre-training (CLIP) [29] leverages natural language descriptors of images as supervision to learn rich visual representations. In contrast to methods that predict the *exact* words of the text accompanying the image, CLIP solves an easier proxy task of predicting only which text as a whole is correctly paired with which image and not the exact words of the text. Next, we briefly describe the training methodology and implementation details.

Training: Given a batch consisting of B (image, text) pairs, CLIP is optimized to correctly predict the right (image, text) combination among the $B \times B$ possibilities. To achieve this, CLIP learns a multi-model embedding space by jointly training an image and text encoders to maximize the cosine similarity between the text and image embeddings of the B right pairs while simultaneously minimizing the $B^2 - B$ incorrect pairings. A symmetric cross entropy loss, similar to the ones proposed in [30, 33], is optimized to learn the text and image encoders. Refer to Fig.5 for a pseudo code for the core implementation of CLIP.

Implementation Details: Authors of [29] consider two architectures for the image encoder: a ResNet [10] (with several modifications) and a Visual Transformer backbones [38]. The text encoder is a 63M-parameter 12-layer 512-wide transformer [34] model with eight attention heads. For computational efficiency, the max sequence length was capped at 76. For more specific details about architectures, we refer the readers to [29]. All the CLIP models are trained for 32 epochs using ADAM optimizer with a very large

batch size of 32678 and a cosine schedule for learning rate decay. Mixed precision [26] was used to speed up training and save memory.

4.2.2 CLIP for Natural Language Vehicle Retrieval

In this section, we describe our method that leverages CLIP for natural language-based vehicle retrieval. We adopt the publicly available CLIP model that uses the Visual Transformer ViT-B/32 as the image encoder and a Text Transformer as the text encoder for our purpose. Given a track $\mathcal{T}_i = \{I_1^i, I_2^i, \ldots, I_{t_i}^i\}$ consisting of t_i crops centered around the vehicle, we use the Image encoder of CLIP to extract per frame features f_j^i $(j \in \{1, 2, \ldots, t_i\})$. We resize the crops to 256×256, take a center crop of 224×224 and normalize the crops to zero mean and unit standard deviation as a part of pre-processing. Each query consists of three natural language descriptions, q_i $(i \in \{1, 2, 3\})$. We use the text encoder of CLIP to extract text features f_i^q $(i \in \{1, 2, 3\})$. The per frame features are then average pooled to obtain the track feature $f^i = \frac{\sum_{j=1}^{t_i} f_j^i}{t_i}$. Similarly, the text features are averaged to obtain the average text feature $f^q = \frac{\sum_{j=1}^{3} f_j^q}{3}$. We normalize these features and use cosine similarity as the score for this pair. For each query description, this process is repeated over all tracks and the scores are sorted in descending order to get the retrieval output. Refer to Figure 4 for the retrieval pipeline.

4.3. Results

In Table 3 we show the top ten performers of the Natural Language-based vehicle retrieval track in 2021 NVIDIA AI City Challenge. Our method achieves a MRR score of 0.1364 and is ranked 8$^{\text{th}}$ among all the submissions to the public leaderboard.

4.4. Future Research

CLIP [29] has shown impressive zero shot performance in many tasks. It was shown that fine-tuning CLIP on the target data can further improve results. One direction we would like to explore is to fine-tune CLIP on the CityFlow-NL dataset. In our current approach, we naively average per-frame features to get the track features. However, not all frames are equally important due to occlusions, viewpoint variations etc and hence it is imperative to incorporate temporal modeling into the pipeline. To this end, we would like to explore self attention [34] as a way to intelligently aggregate visual information across time. Finally, without fine-tuning CLIP, it is unreasonable to expect the model to discriminate between various vehicle models. Discriminating between models is crucial as it helps with rejecting false positive recommendations due to color similarities.

5. Conclusion

In this paper, we summarizes our contributions in the 2021 NVIDIA AI City Challenge for City-Scale Multi-Camera Vehicle Re-Identification and Natural Language-Based Vehicle Retrieval tasks. We show how effective representation learning techniques in conjunction with post-processing steps and contrastive learning-based Language-Image pre-training can result in impressive real world vehicle retrieval systems. Both our proposed methods are ranked 8$^{\text{th}}$ on both the tasks.

6. Acknowledgement

This research is supported in part by the Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative, It is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] 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 *ECCV*, 2020. 2

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 5

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 60:51–58, 2019. 5

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

[7] Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, and Hanqing Lu. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[8] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 7

[11] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 6(7):8, 2020. 1, 5

[12] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021. 2

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4

[14] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6132–6141, 2019. 2, 4

[15] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *European Conference on Computer Vision*, pages 369–386. Springer, 2020. 1, 2, 3

[16] Pirazh Khorramshahi, Neehar Peri, Amit Kumar, Anshul Shah, and Rama Chellappa. Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding. 3

[17] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 2

[18] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016. 2

[19] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016. 2, 3

[20] Xinchen Liu, Wu Liu, Jinkai Zheng, Chenggang Yan, and Tao Mei. Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 907–915, 2020. 2

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[22] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019. 2, 3

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2

[24] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 4, 5

[26] P. Micikevicius, Sharan Narang, Jonah Alben, G. Diamos, Erich Elsen, D. García, Boris Ginsburg, Michael Houston, O. Kuchaiev, Ganesh Venkatesh, and H. Wu. Mixed precision training. *ArXiv*, abs/1710.03740, 2018. 8

[27] Neehar Peri, Pirazh Khorramshahi, Sai Saketh Rambhatla, Vineet Shenoy, Saumya Rawat, Jun-Cheng Chen, and Rama Chellappa. Towards real-time systems for vehicle re-identification, multi-camera tracking, and anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 622–623, 2020. 1, 2

[28] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 7, 8

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 7

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[32] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale

benchmark for multi-target multi-camera vehicle tracking and re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8797–8806, June 2019. 2

[33] Aäron van den Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 7

[34] 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. 2, 7, 8

[35] E. Voorhees. The trec-8 question answering track report. In *TREC*, 1999. 7

[36] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. *arXiv preprint arXiv:2104.03541*, 2021. 4

[37] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 4

[38] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 7

[39] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[40] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *ECCV*, 2020. 2

[41] Zhedong Zheng, Minyue Jiang, Zhigang Wang, Jian Wang, Zechen Bai, Xuanmeng Zhang, Xin Yu, Xiao Tan, Yi Yang, Shilei Wen, et al. Going beyond real data: A robust visual representation for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 598–599, 2020. 2, 3, 4

[42] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 5

[43] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 4

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3

[45] Xiangyu Zhu, Zhenbo Luo, Pei Fu, and Xiang Ji. Voc-reid: Vehicle re-identification based on vehicle-orientation-camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 602–603, 2020. 4