TIED: A Cycle Consistent Encoder-Decoder Model for Text-to-Image Retrieval

Clint Sebastian[1,2], Raffaele Imbriaco[1], Panagiotis Meletis[1],
Gijs Dubbelman[1], Egor Bondarev[1], Peter H.N. de With[1,2]
[1]VCA Group, Eindhoven University of Technology
[2]Cyclomedia B.V
{c.sebastian, r.imbriaco, p.c.meletis}@tue.nl

Abstract

Retrieving specific vehicle tracks by Natural Language (NL)-based descriptions is a convenient way to monitor vehicle movement patterns and traffic-related events. NL-based image retrieval has several applications in smart cities, traffic control, etc. In this work, we propose TIED, a text-to-image encoder-decoder model for the simultaneous extraction of visual and textual information for vehicle track retrieval. The model consists of an encoder network that enforces the two modalities into a common latent space and a decoder network that performs an inverse mapping to the text descriptions. The method exploits visual semantic attributes of a target vehicle along with a cycle-consistency loss. The proposed method employs both intra-modal and inter-modal relationships to improve retrieval performance. Our system yields competitive performance achieving the 7th position in the Natural Language-Based Vehicle Retrieval public track of the 2021 NVIDIA AI City Challenge. We demonstrate that the proposed TIED model obtains six times higher Mean Reciprocal Rank (MRR) than the baseline, achieving an MRR of 15.48. The code and models will be made publicly available.

1. Introduction

Vehicle track retrieval from traffic cameras [9] is an essential component of upstream systems aiming for urban planning and traffic-flow control. Large-scale retrieval of vehicle tracks is difficult to obtain with conventional image or video retrieval methods, due to the immense variety of motion patterns and vehicle semantics that need to be considered. Descriptions for these tracks in Natural Language (NL) is an appealing alternative method to enable the retrieval system to directly interact with human-given descriptions [33, 2]. The objective of NL-based vehicle track retrieval [9] is to match a given NL description to the corresponding vehicle track. The NL description is given as one or more text queries, and the vehicle tracks are a sequence of frames from a single camera, where the location of the vehicle is known. This task combines visual and textual modalities, thus solutions should simultaneously account for intra- and inter-modality challenges. Vehicle tracks include a wide variety of vehicle types, colors, and motion types. NL queries often have variations and ambiguities, since different people can describe the same vehicle seman-

Figure 1: Qualitative vehicle retrieval results using the baseline method (Rows 1, 3) and our method (Rows 2, 4) of a frame from the retrieved tracks. The queries are "A blue sedan runs down the street." and "A red cargo truck pulls a yellow cement mixer."

tics and actions differently. An additional complexity to the problem is introduced by requiring to identify vehicle maneuvers over a time interval. Contrary to NL-based image or object retrieval [14, 11], an NL-based track retrieval system should address the time dimension of the task, as indicated by the related NL-based visual object tracking task defined in literature [23, 8].

We propose a system that can jointly leverage both language and visual modalities by extracting feature embeddings and then aligning them with cycle-consistent intra- and inter-modality metric loss functions. Because additional cues increase retrieval performance, we additionally use the visual semantics of the vehicles, *i.e*. color, and type information. The contributions of this work are summarized as follows.

- A Text-to-Image Encoder-Decoder (TIED) network that maps both visual and textual inputs to a latent space and jointly maps it back to the text queries.
- A cross-modal training objective that models both intra/inter-modal relations between the image and language queries as well the cycle-consistent objective.
- Additionally, we also present a semi-automated method to extract attributes from language descriptions.

2. Related Work

A. Vehicle Re-Identification. Vehicle Re-Identification (Re-ID) is an important topic in the context of smart cities and traffic management, which relates to vehicle retrieval from natural language descriptions. In this task, the system should match images of the same vehicle instance across different camera views and locations. The key dataset is the CityFlow dataset [9, 32], which is extended by the CityFlow-NL dataset [9].

A summary of the best performing methods is found in [26]. Noticeable trends from these methods are the extraction of additional attributes [6, 39, 3] (color, type, orientation), the deployment of state-of-the-art classification networks such as ResNet-IBN [27], and the combination of classification and metric losses [38, 31, 10].

The work of Zhu *et al.* [39] exploits additional attributes at the vehicle and geographic level. They propose a multi-task network that learns to identify vehicles, orientation, and cameras. Meanwhile, the authors of [12] deploy a strong baseline architecture [25] in conjunction with a two-step training process that leverages the availability of synthetic data. The best performing system [38] uses both style transform and content manipulation, to reduce the synthetic-to-real domain gap and enhance the training data. These improvements, together with camera and orientation-aware models, yield excellent performance for vehicle Re-ID.

However, while vehicle Re-ID and text-to-image retrieval share some challenges, the latter requires specialized solutions and architectures. These are necessary to accurately model cross-domain relationships and reduce the domain gap. We discuss some of the common approaches below.

B. Text-to-Image retrieval. In cross-modal retrieval, the objective is to identify the correspondences between a set of query and database elements, belonging to two different modalities. Since each modality is different, the embeddings produced by specific feature extractors (text, images, video) will not be inherently aligned. Therefore, the most frequent approach in the literature is to construct a common feature space [7, 16, 34]. We summarize some of the recent solutions to this problem below.

The authors of [4] propose to bridge the domain gap by imposing a cycle-consistent transformation from the text and image domains. To this end, they train a two-branch model for text-to-image embedding translation and vice versa. A GRU processes the textual inputs, whereas a CNN processes the visual inputs. The embeddings are compared and the triplet hinge loss [7, 18, 15] is minimized. An additional reconstruction loss is employed to reduce the error between the original sentence and the one generated by the visual feature embedding. While the previous method uses global image information, most of the solutions present in the literature deploy a more localized approach. In these works, the images and accompanying textual descriptions may contain more than one single object. Therefore, bounded image regions or patches are fed through the CNN to learn saliency as well as word-to-image semantic relationships [10, 21, 35, 37]. These systems deploy region proposal systems like Faster R-CNN [30] with bottom-up attention [1] to find salient and semantically relevant regions. In [21], these regions are then processed by a Graph Convolutional Network and a GRU to produce semantically enhanced features. The authors of [35] eschew the relational graph by employing a novel message-passing module between the textual and visual branches. The features are then fused and the branches are trained jointly. Similarly, explicit inter-modal and intra-modal attention is learned in [37]. Meanwhile, the system presented in [10] does not employ a region proposal network, but instead directly partitions the image into patches. This is done to create a visual analog of the tokens used by the language model. Hence, each patch of the image resembles each word of a sentence. The textual tokens and the patches are then fed into BERT [5] and trained using the triplet hinge loss. Recent work has extended the text-to-image retrieval into a text-to-video retrieval problem, presenting solutions

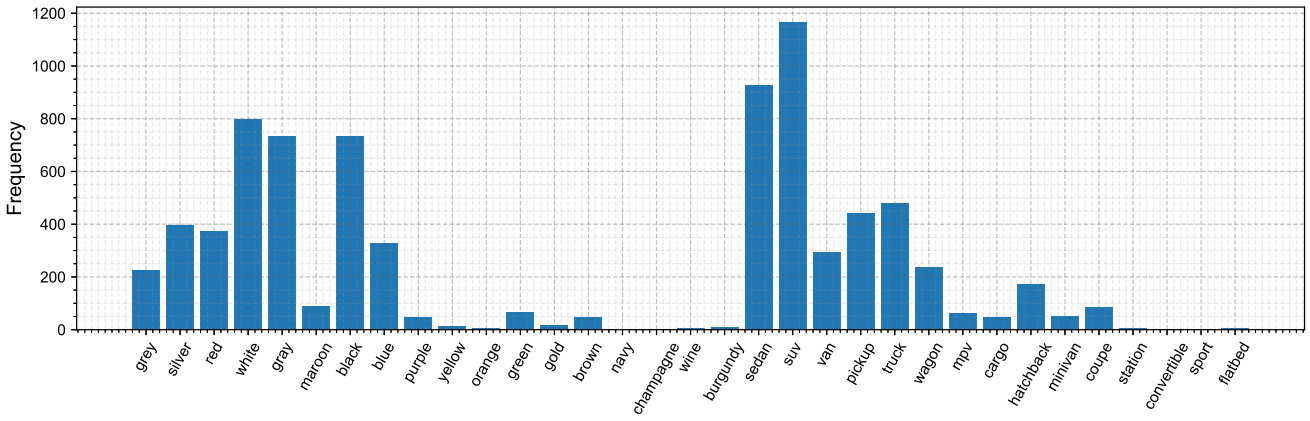

Figure 2: Word frequency for the additional attributes on the training set. We consider 18 color and 15 vehicle type attributes for training our model.

that exploit temporal and language correlations [19, 22] by producing video-level representations/predictions. However, we limit our efforts and consider only the individual image frames as in [9].

C. Vehicle retrieval from natural language. While the works summarized above perform text-to-image retrieval, there is a fundamental difference between the scenes analyzed in previous work and those present in the CityFlow-NL dataset [9]. The available temporal and geographic relationships are unique for the multi-camera CityFlow-NL dataset. Other datasets for language and vision tasks such as Flicker30K [36], Flicker30K Entities [28] and Visual Genome [17] provide textual information on the scene. This can include actions and qualities of the imaged objects and their surroundings. However, the CityFlow-NL dataset also contains other temporal information that does not occur in the above-mentioned datasets. These temporal relationships make it possible to describe the maneuvers that the depicted vehicles take, e.g. 'turns right'. While this information can differentiate the specific vehicles, it also makes individual frames less informative, since part of the textual information describes actions that occur over many images.

In [9], an architecture is presented to bridge the modality gap existent between textual and visual features. Each modality is represented separately by a ResNet50 [12] branch for visual features and by BERT for language features. This network is trained jointly and minimizes the distance between the modality-specific embeddings.

The previously described model achieves moderate success in producing discriminative cross-modal features. However, it does not explicitly learn identifying information such as color or vehicle type. Instead, this is learned intrinsically by minimizing the distance between the textual and visual embeddings. To yield better features, we propose to leverage additional attributes present in the text corpus as well as deploying a multi-task network that attempts to reduce the modality gap.

3. Natural Language attribute pre-processing

The proposed approach primarily consists of two parts. In the first part, additional labels are generated using a semi-automated method. The second part comprises an encoder-decoder architecture with cycle consistent inter/intra-modal losses as the objective regularization function.

One of the key components of our work is the extraction of identifying features from textual data and their exploitation during training. We extract simple labels, specifically color and vehicle type, in a semi-automated manner from the textual descriptions of the tracks. This is motivated by the sentence structure presented by the annotations and the frequency of specific words in the dataset.

As a preliminary step, each sentence in the training dataset is split into words. They are converted into lowercase, while stopwords are removed. Lemmatization is done to avoid counting conjugations and plural forms as different words. Afterward, we calculate the frequency of the remaining words across the training corpus. Figure 2 depicts the frequency of the extracted attributes in the corpus after preliminary processing. From these pre-processed words, we extract additional attributes by comparing the words in each textual description against a collection of words for either color or vehicle type. This results in 18 colors and 15 car types. Needless to say, a consensus is not always present in the annotations. Due to variations in illumination, certain colors can be confused with one another. For example, one annotator may consider a car to be gray while another may assess it as silver. To facilitate attribute extraction, we produce a multi-label attribute. Each image track is accompanied by two binary vectors l_c, and l_t for color and

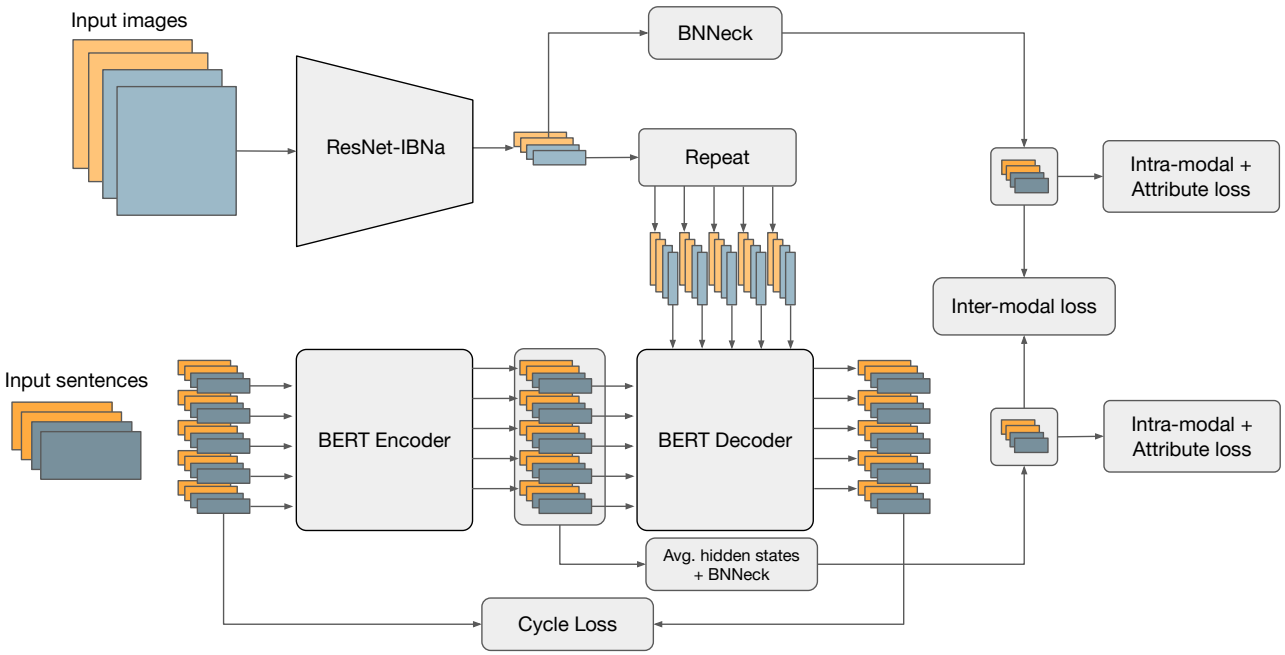

Figure 3: Overview of the TIED architecture. The input image is supplied to the image encoder to produce image-level embeddings. Similarly, the input sentence (at the right) is fed to the BERT encoder to obtain language-level embeddings. Both of the generated embeddings are fed to the BERT decoder to map back to the input language inputs. During test time, the embeddings after BNNeck are extracted across the entire database, and similarity is computed to obtain the top tracks. Each color represents different tracks.

type, respectively. If a word relating to either of these attributes is present in the textual descriptions, its corresponding position is set to unity. Therefore, in the aforementioned case of conflicting descriptions, the car is both silver and gray. Additionally, when extracting these attributes, sentences are split into parts using common connective words ('follow', 'behind', 'before', 'after'), and only the first part is used. This is done to avoid introducing features from vehicles not associated with the track being analyzed, but present in part of the frames.

4. Text-to-Image Encoder-Decoder

The text-to-image encoder-decoder model consists of two parts, an image encoder to extract visual embeddings and a language encoder to extract textual embeddings. A decoder model that jointly models the visual and textual embeddings that reconstructs the input text.

4.1. Model

As both input modalities are different, this requires both to be mapped to a common latent space. Let $\mathcal{I} = \{c_1, c_2, c_3, ..., c_n\}$, be the set of image crops from a video clip, and $\mathcal{Q} = \{q_1, q_2, q_3, ..., q_n\}$ its corresponding language queries. Our goal is to learn a common latent space between two different modalities X and Y. Given an image model F and a language model G, the objective is to learn a mapping $F: X \rightarrow L$ and $G: Y \rightarrow L$, where X denotes the image modality, Y is the language modality and L is the latent space.

Image Model. To generate the image-level embeddings, ResNet50-IBN-a is pretrained on ImageNet and utilized as the image-level model. Each crop c_i is fed to the image model, and the output from the last block is pooled to produce a 2,048-sized embedding. This is followed by a fully connected layer (FC layer) to produce a 768-dimensional embedding output. Finally, a BNNeck [25] is applied to obtain a 512-dimensional embedding, and a classification head (FC layer) is utilized to produce the label predictions.

Language Model. For the language model, a pretrained BERT with encoder-decoder architecture is utilized. This consists of a contextual embedding model that encodes the textual descriptions associated with each query. Its purpose is to produce highly descriptive representations from the

text inputs. The encoder-decoder architecture is typically used for sequence-to-sequence modeling tasks [24, 20], such as language translation, image captioning, etc. For the BERT model, we use word piece tokenization as in [5]. For generating the text embeddings, the text inputs are fed through the BERT encoder. The resulting outputs are averaged and are supplied to a BNNeck to minimize both multi-label classification (attribute) loss and embedding distances. Additionally, the proposed architecture also uses a BERT decoder to reduce the cross-modal distance between the textual and image embeddings. The decoder attempts to predict the sentence tokens based on the original input, as well as the embeddings generated through the image encoder. Figure 3 depicts the overview of the TIED architecture.

Joint Decoder. Because the mapping from the input modality space to the latent space is ill-posed, an inverse mapping from the latent space to the language modality is added. The inverse mapping can be formulated as

$$G_d(F(c_i), G_e(q_i)) \approx q_i, \tag{1}$$

where crop $c_i \in \mathcal{I}$ and $q_i \in \mathcal{Q}$. The function F is the image encoder, G_e is BERT encoder and G_d is BERT decoder. The BERT decoder jointly optimizes the image and text embeddings by reconstructing the input tokens. In essence, the BERT decoder acts as a regularization, performing cycle-consistent learning conjointly. Note that our network can be seen as the input textual embedding being conditioned by an image embedding and a textual sequence-to-sequence model. The conditioned mapping is specified by the expression $F \circ G_d$: $X \rightarrow Y$ and the sequence-to-sequence mapping by $G_e \circ G_d$: $Y \rightarrow Y$. The operator '$\circ$' denotes the function composition. Both F and G_e map to a latent representation that is used during retrieval.

4.2. Training objective

For training the network, several different objectives are utilized, mapping intra/inter-modal relationships. Multi-label classification losses are applied to preserve the visual and textual cues. For attribute classification, we apply cross-entropy loss across both vehicle color and type. The attribute loss is given as

$$\mathcal{L}_{\text{attr}} = \sum_{i}^{N} t_i \cdot \log(y_i), \tag{2}$$

where y_i is the predicted label probability of the attribute i and t_i is the target label. The attribute loss is applied across both the language as well as the image modalities. To model both inter-modal and intra-modal relations, we employ the triplet loss function, which is given as

$$\mathcal{L}_{\text{triplet}}(a, p, n) = \sum_{a,p,n} \max[D_{a,p} - D_{a,n} + m, 0], \tag{3}$$

where $D_{a,p}$ is the distance between anchor (a) and positive sample (p), and $D_{a,n}$ is the distance between anchor and negative sample (n). The triplet margin is denoted as m.

Depending on the inter/intra-modal loss, the anchor, positive, and the negative sample can be an image crop or a text query. Both the inter-modal $\mathcal{L}_{\text{inter}}$ and intra-modal loss $\mathcal{L}_{\text{intra}}$ are defined as

$$\mathcal{L}_{\text{inter}} = \lambda_1 \mathcal{L}_{\text{triplet}}(a_q, p_c, n_c) + \lambda_2 \mathcal{L}_{\text{triplet}}(a_c, p_q, n_q), \tag{4}$$

$$\mathcal{L}_{\text{intra}} = \lambda_3 \mathcal{L}_{\text{triplet}}(a_q, p_q, n_q) + \lambda_4 \mathcal{L}_{\text{triplet}}(a_c, p_c, n_c), \tag{5}$$

where a_q, p_q, n_q, and a_c, p_c, n_c are the anchor, positive and negative samples for the query and image crops, respectively. The $\mathcal{L}_{\text{inter}}$ minimizes the distance across the image and text modalities, whereas the $\mathcal{L}_{\text{intra}}$ minimizes the distances between each modality itself. The inter-modal loss minimizes the distance from the text descriptions to the images and vice-versa, generating a rich representation in latent space. The final objective function models the relation within a modality as well as against others. Therefore, the final training objective is given as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} + \lambda_5 \mathcal{L}_{\text{attr}} + \lambda_6 \mathcal{L}_{\text{cycle}}, \tag{6}$$

where $\mathcal{L}_{\text{cycle}}$ predicts the input tokens that are supplied to the BERT encoder, utilizing both image and language embeddings. The coefficients $\lambda_1, ..., \lambda_6$ are the weights assigned to each of the losses. Section 5.3 compares each of the losses independently, including $\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{inter}}$ and $\mathcal{L}_{\text{attr}}$. We show that each loss function plays a vital role in improving the retrieval performance.

5. Experiments

5.1. Dataset and evaluation metrics

In this work, we use the CityFlow-NL [9] dataset. This is a multi-camera multi-track vehicle dataset composed of 2498 training and 530 test tracks. Each of those is composed of a variable number of frames, averaging roughly 75 frames per track. Three natural language descriptions of the vehicle and the maneuver ('goes straight', 'turns right') accompany the tracks. We produce a validation set from the original by selecting 500 tracks. We report on the Recall@N (R@N) and Mean Reciprocal Rank (MRR), defined as

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r_i}, \tag{7}$$

where N is the number of queries and r_i is the rank of the first relevant element in the database.

5.2. Implementation details

Our TIED architecture is composed of a BERT [5] encoder-decoder for processing natural language and a

Attribute	Intra-modal	Inter-modal	MRR	R@5	R@10
Encoder Model					
✓			1.2	0.0	1.4
	✓		1.5	0.0	2.8
		✓	24.2	39.4	55.8
	✓	✓	32.4	51.0	68.0
✓		✓	32.1	51.4	68.8
✓	✓	✓	33.4	49.2	68.0
TIED Model					
✓	✓	✓	31.9	47.4	65.4

Table 1: Performance of each of the losses and their combinations on the encoder-based and TIED models. The listed results are reported on the validation set.

ResNet50-IBN-a [27] backbone for the visual inputs. The branches are trained jointly using Stochastic Gradient Descent with a learning rate of 2×10^{-4} and decays by an order of magnitude after a fixed number of epochs. The model is trained for 250 epochs. For the intra-/inter-modal losses, the triplet margin is set to 1.2, and batch hard sampling [13] is employed for mining the triplets. For each mini-batch, 4 frames from 8 different tracks are sampled. The coefficient of the attribute classification, intra/inter-modal and the cycle losses are set to $\lambda_1 = 2, \lambda_2 = 2, \lambda_3 = 1, \lambda_4 = 1, \lambda_5 = 2, \lambda_6 = 1$. These values are selected empirically. The input images are resized to 224 $\times$ 224 pixels and are augmented with horizontal flips and color jitter.

At test time, 512-dimension descriptors are extracted and are L_2 normalized. The query-to-track distances are computed per frame and averaged as in [9]. The final embeddings are computed after ensembling four identical models, trained with a learning-rate decay threshold of 50, 50, 60, and 70 epochs, respectively. The experiments are conducted on a GTX 1080Ti GPU using PyTorch.

Attribute	Intra-modal	Inter-modal	MRR	R@5	R@10
Encoder Model					
✓	✓	✓	14.5	22.6	36.6
TIED Model					
✓		✓	15.5	22.8	40.0
✓	✓	✓	15.0	21.9	37.7

Table 2: Performance of the proposed encoder-based and TIED models on the test set. The addition of intra-modal losses is detrimental on the overall test set.

5.3. Ablation studies

We conduct ablation experiments to study the impact of the several losses applied for training. For the ablation experiments, the TIED model without the decoder is considered. Table 1 summarizes the performance with MRR, R@5, and R@10. By applying attribute or intra-modal losses only, it is evident that the model has poor performance. This is because each query is assigned non-unique attributes. Essentially, this means that each attribute is associated with several queries. Similarly, the intra-modal loss only minimizes the distance within the same modality, which fails to learn the similarity between the image and language inputs. Only the inter-modal loss offers good performance independently since it exploits the relation between both image and language modality.

When combining either intra-modal losses with inter-modal or attribute losses, the performance improves by 8% MRR. The attribute loss further separates each language and image modality in feature space, thereby improving performance when combined with the inter-modal loss. Similarly, the addition of intra-modal losses regularizes the features by separating features within the same modality. By combining all the three losses, the performance improves on our validation set to reach an MRR of 33.4.

The results with cycle loss are provided when the TIED model is applied. On our validation set of 500 queries, it lowers performance. However, our submission on the test set shows that the addition of the decoder improves the performance. Although the intra-modal loss is beneficial on the validation set, the results on the private test set show that it was not helpful. The results on the private test set are shown in Table 2.

5.4. Results on the 2021 AI City Challenge

The proposed method is employed to generate retrieval rankings on the test set. The results are submitted to the NL-based Vehicle Retrieval track of the 2021 AI City Challenge. The top positions and our results are summarized in Table 3. We have obtained an MRR of 15.48 without using any additional data. This is a significant improvement over the baseline MRR of 2.69. Compared to the top teams, our performance has only a difference of less than 0.65 MRR to the teams from position two until six. A visual comparison of the retrieval results between the baseline model and our best model is depicted in Figure 4.

6. Discussion

During the competition, we have conducted experiments with several models that are potentially effective at the text-to-image retrieval task. Our submissions suggest that there

[1]Accessed April 11, 2021: https://eval.aicitychallenge.org/aicity2021/submission/leaderboard.

(a) Queries: “A green go to the straight.”, “A green van runs on the street.”, “A green wagon crossing the intersection.”

(b) Queries: “White mini cooper.”, “Mini cooper keep straight on the road.”, “A white hatchback goes straight at the street followed by another white vehicle.”

(c) Queries: “A white SUV drives towards the intersection.”, “A white SUV runs down the street.”, “A white SUV runs down the street.”

(d) Queries: “A red SUV drives up a hill in the left lane.”, “A red SUV runs across an intersection.”, “A dark-red SUV is going straight.”

Figure 4: Retrieval results for the baseline [9] model (top rows) and our best-performing model (bottom rows).

are methods that performed well on our validation set, but are not effective on the 50% test set. At the same time, some methods perform poorly on the validation set but not on the private test set. We discuss this aspect briefly here so that it can benefit future versions of the challenge. In our experiments, CLIP [29] and label smoothing perform poorly on the validation set as well as the 50% test set. However, both methods offer good performance on the overall test set. We have also experimented with multi-head models as well as video-based models. Both models offer high R@10 performance with a lower MRR on the full test set. Generally in our case, we attribute the differences in performance due to overfitting on the validation set. The additional models from our work will also be made publicly available.

7. Conclusion

In this paper, we propose TIED, a text-to-image encoder-decoder model that leverages both language and visual inputs to improve text-to-vehicle retrieval. The proposed

Rank	Team name	MRR
1	Alibaba-UTS	18.69
2	TimeLab	16.13
3	SBUK	15.94
7	VCA (ours)	15.48
	Baseline [9]	2.69

Table 3: Final ranking[1] of the top teams and our results for the NL-based Vehicle Retrieval track of the 2021 AI City Challenge. The MRR is reported on the private test set.

method maps both inputs to latent spaces and utilizes the joint information of visual and textual embeddings to reconstruct the text queries. The TIED model is trained with a combination of intra/inter-modal losses as well as the attribute and cycle-consistent losses to improve performance. The inter-modal loss enforces the embeddings from the different modalities to a common multi-modal feature space, as validated from the experiments. We have also performed ablation experiments to study the impact of each component of the final objective function. Finally, the proposed system yields comparable performance to the top runner-up positions in the 2021 NVIDIA AI City Challenge, achieving the 7th position in the Natural Language-Based Vehicle Retrieval public track.

References

[1] Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 2

[2] Daniel Paul Barrett, Andrei Barbu, N Siddharth, and Jeffrey Mark Siskind. Saying what you're looking for: Linguistics meets video search. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2069–2081, 2015. 1

[3] Tsai-Shien Chen, Man-Yu Lee, C. Liu, and S. Chien. Viewpoint-aware channel-wise attentive network for vehicle re-identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2448–2455, 2020. 2

[4] M. Cornia, L. Baraldi, H. Tavakoli, and R. Cucchiara. Towards cycle-consistent models for text and image retrieval. In *ECCV Workshops*, 2018. 2

[5] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2, 5

[6] V. Eckstein, Arne Schumann, and A. Specker. Large scale vehicle re-identification by knowledge transfer from simulated data and temporal attention. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2626–2631, 2020. 2

[7] Fartash Faghri, David J. Fleet, J. Kiros, and S. Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 2

[8] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Real-time visual object tracking with natural language description. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 700–709, 2020. 2

[9] Qi Feng, Vitaly Ablavsky, and S. Sclaroff. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. *ArXiv*, abs/2101.04741, 2021. 1, 2, 3, 5, 6, 7, 8

[10] Dehong Gao, Linbo Jin, B. Chen, Minghui Qiu, Yi Wei, Y. Hu, and H. Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. 2

[11] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Schmidt Feris. Dialog-based interactive image retrieval. 2018. 2

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *ArXiv*, abs/1603.05027, 2016. 2, 3

[13] A. Hermans, Lucas Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737, 2017. 6

[14] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 2

[15] A. Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[16] Ryan Kiros, R. Salakhutdinov, and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *ArXiv*, abs/1411.2539, 2014. 2

[17] R. Krishna, Yuke Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, Stephanie Chen, Yannis Kalantidis, L. Li, D. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 3

[18] Kuang-Huei Lee, X. Chen, G. Hua, H. Hu, and Xiaodong He. Stacked cross attention for image-text matching. *ArXiv*, abs/1803.08024, 2018. 2

[19] Jie Lei, Linjie Li, L. Zhou, Zhe Gan, Tamara L. Berg, M. Bansal, and Jing jing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *ArXiv*, abs/2102.06183, 2021. 3

[20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 5

[21] Kunpeng Li, Yulun Zhang, K. Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4653–4661, 2019. 2

[22] Xirong Li, F. Zhou, Chaoxi Xu, Jiaqi Ji, and G. Yang. Sea: Sentence encoder assembly for video retrieval by textual queries. *ArXiv*, abs/2011.12091, 2020. 3

[23] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6495–6503, 2017. 2

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

[25] Hao Luo, Youzhi Gu, Xingyu Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, 2019. 2, 4

[26] M. Naphade, Shuo Wang, D. Anastasiu, Z. Tang, Ming-Ching Chang, Xiaodong Yang, L. Zheng, Anuj Sharma, R. Chellappa, and Pranamesh Chakraborty. The 4th ai city challenge. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2665–2674, 2020. 2

[27] Xingang Pan, Ping Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 2, 6

[28] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015. 3

[29] A. Radford, J. W. Kim, Chris Hallacy, A. Ramesh, G. Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, J. Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ArXiv*, abs/2103.00020, 2021. 7

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 2

[31] Clint Sebastian, Raffaele Imbriaco, E. Bondarev, and P. H. With. Dual embedding expansion for vehicle re-identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2475–2484, 2020. 2

[32] Zheng Tang, M. Naphade, Ming-Yu Liu, X. Yang, Stan Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J. Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8798, 2019. 2

[33] Stefanie Tellex and Deb Roy. Towards surveillance video search by natural language query. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 1–8, 2009. 1

[34] L. Wang, Y. Li, J. Huang, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:394–407, 2019. 2

[35] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, J. Yan, X. Wang, and J. Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5763–5772, 2019. 2

[36] Peter Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3

[37] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and S. Li. Context-aware attention network for image-text retrieval. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3533–3542, 2020. 2

[38] Zhedong Zheng, Minyue Jiang, Zhigang Wang, J. Wang, Zechen Bai, Xuanmeng Zhang, Xin Yu, Xiao Tan, Y. Yang, Shilei Wen, and Errui Ding. Going beyond real data: A robust visual representation for vehicle re-identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2550–2558, 2020. 2

[39] Xiangyu Zhu, Zhenbo Luo, Pei Fu, and Xiang Ji. Voc-reld: Vehicle re-identification based on vehicle-orientation-camera. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2566–2573, 2020. 2