Connecting Language and Vision for Natural Language-Based Vehicle Retrieval

Shuai Bai[1] Zhedong Zheng[2] Xiaohan Wang[3] Junyang Lin[1]
Zhu Zhang[1] Chang Zhou[1] Hongxia Yang[1] Yi Yang[2]
[1]DAMO Academy, Alibaba Group, [2]ReLER Lab, University of Technology Sydney, [3] Zhejiang University

Abstract

Vehicle search is one basic task for the efficient traffic management in terms of the AI City. Most existing practices focus on the image-based vehicle matching, including vehicle re-identification and vehicle tracking. In this paper, we apply one new modality, i.e.*, the language description, to search the vehicle of interest and explore the potential of this task in the real-world scenario. The natural language-based vehicle search poses one new challenge of fine-grained understanding of both vision and language modalities. To connect language and vision, we propose to jointly train the state-of-the-art vision models with the transformer-based language model in an end-to-end manner. Except for the network structure design and the training strategy, several optimization objectives are also revisited in this work. The qualitative and quantitative experiments verify the effectiveness of the proposed method. Our proposed method has achieved the **1st** place on the 5th AI City Challenge, yielding competitive performance* 18.69% *MRR accuracy on the private test set. We hope this work can pave the way for the future study on using language description effectively and efficiently for real-world vehicle retrieval systems. The code will be available at* https://github.com/ShuaiBai623/AIC2021-T5-CLV.

1. Introduction

Vehicle retrieval usually meets a large-scale candidate pool due to the 24/7 records, which is an important part of the intelligent transportation system for AI City. Most existing vehicle retrieval systems are based on image-to-image matching, also known as vehicle re-identification (vehicle re-id) [68]. To retrieve the target vehicle tracklets, these methods require a vehicle image query, which is not always available in the real-world scenario [25, 10]. In this report, we leverage natural language descriptions to search vehicles. Compared to image queries, natural language descriptions are more user-friendly and easier to be obtained. Besides, it enables fuzzy vehicle search and provides more flexible applications.

A common approach to performing language-based vehicle retrieval is to embed the images and descriptions to shared feature space and then rank the vehicle images based on the cross-modal similarities. Most existing methods [8] construct the visual encoder and text encoder with fixed backbones and only optimize several projection layers. In our solution, we train the cross-modal vehicle retrieval framework in an end-to-end manner. This design enables the powerful backbones to learn fine-grained vehicle attributes like vehicle types and directions. Inspired by the recent advances in cross-modal representation learning [39, 70], we adopt the symmetric InfoNCE loss [36] and instance loss [70] to jointly train the text encoder and visual encoder.

For the visual encoder, we propose a two-stream architecture to provide complimentary local details (*e.g.* color, type and size) and global information (*e.g.* motion and environment). The motivation behind the design is that the natural language sentences not only contain the information of vehicle appearance but also describe the trajectories and background. This is also the main difference between language-based vehicle retrieval and image-based vehicle retrieval. Specifically, we adopt two individual CNNs to construct the backbones of the two streams. The local stream takes the detected patches that only contain the vehicle as input, while the input for the global stream is the synthesized dynamic image with the averaged background and the trajectory of the vehicle. The outputs from the two streams are concatenated as the final visual representation. For the text encoder, we adopt the state-of-the-art transformer-based language models like BERT [7] and RoBERTa [29] as our text encoder. To enhance the robustness of the model, we propose a text augmentation approach by back-translation technique. The proposed method has achieved 18.69% MRR accuracy on the private test set of the 5th AI City Challenge on natural language-based vehicle retrieval, yielding the **1st** place on the public leader-

board.

2. Related Work

2.1. Video Retrieval via Natural Language

Natural language-based video retrieval aims to search a specific video matching the given language description from a large amount of candidate videos. Most existing works [23, 56, 37, 60, 61, 34, 63, 32, 8, 50] adopt the similarity learning [55] to learn a function (network) that can estimate the similarity between videos and language descriptions. These works encode the language by the textual feature extractor (Word2Vec [33], LSTM [16], etc.), learn video representations by the visual feature extractor (Two-Stream Network [42], C3D [3], S3D [54], etc.) and estimate the language-video similarity in a common semantic space. Further, for language representations, Xu *et al*. [56] design a compositional language model by the dependency-tree structure. Yu *et al*. [61] develop a high-level concept detector as semantic priors and apply the attention mechanism to selectively focuses on the detected word concepts. For video representations, recent methods [43, 2] utilize the well-designed Transformer architecture [47] to learn powerful video features. And some works [34, 32, 11, 9] further incorporate multi-modal features (*e.g*. motion, audio) from a video for more robust video understanding. As for the video-language interaction, Zhang *et al*. [63] exploit both low-level and high-level correspondences in the hierarchically semantic spaces, and Dong *et al*. [8] propose the multi-level encoding including global, local and temporal patterns in both videos and sentences to learn better shared representations. Besides, M6 [24], VideoBERT [44], UniViLM [31], HERO [22] and ClipBERT [21] explore the large-scale vision-language pre-training to boost comprehensive video-language understanding.

Recently, to search fine-grained video contents via natural language, researchers begin to explore moment retrieval and object retrieval in videos. Video moment retrieval [12, 15, 65] localizes a video clip corresponding to the given language, which avoids manually searching for the clip of interests in a long video. Existing approaches often pre-define a series of clip proposals by sliding windows or multi-granularity anchors, and rank these clips by visual-textual interaction and estimation, *e.g*. attention mechanism [27] and graph convolution [64]. And video object retrieval [58, 4, 67] aims to search the spatio-temporal object track (i.e. a sequence of bounding boxes) according to the language description. Early work [58] only searches the person track in multiple videos and recent approaches [71, 4] further retrieve the spatio-temporal tracks of diverse objects. Besides single-object retrieval, Huang and Shi *et al*. [18, 41] try to localize multiple objects that appear in the language description. Different from previous tasks, vehicle retrieval via natural language is one practical task for the traffic management, which retrieves the specific vehicle given single-camera tracks and corresponding language descriptions of the targets. Our method sufficiently considers the inherent attributes of the vehicles as well as global motion and environment information to search the described vehicle.

2.2. Vehicle Re-identification

Vehicle re-identification (vehicle re-id) is to find the vehicle of interest from millions of candidate images from different cameras, which can largely save the human resources as well as the time cost. Several pioneering works focus on building the large-scale dataset for subsequential learning, including VehicleID [26], VeRi-776 [28] and VehicleNet [68]. The follow-up works focus on the discriminative representation learning [28, 69] as well as mining the structure information [51, 45]. For instance, Qian *et al*. [38] and Yu *et al*. [59] propose to leverage the multi-scale information within deeply-learned models. To mine the fine-grained vehicle structure, Wang *et al*. [51] further take the keypoints into consideration and apply the structure information to the final feature aggregation part. Besides, Zheng *et al*. [68] apply the transfer learning to distill common knowledge from large-scale vehicle dataset to the specific small dataset, yielding the state-of-the-art performance. Attributes and environment conditions also have been explored in several pioneering works [30, 25]. In summary, vehicle re-identification is primarily different from the natural language-based vehicle retrieval in terms of the input modality. The two different modalities are inherently different, which is challenging in mapping heterogeneous inputs to the same semantic space. In this paper, we mainly focus on the vehicle tracklet retrieval via the natural language, but the existing vehicle re-id also gives us many inspirations in the representation learning and optimization strategies. We will provide more details in Section 3.

2.3. Data Augmentation in NLP

Data augmentation has gradually become a common practice in NLP and it brings substantial improvement due to the requirement of a large amount of training data. The augmented data should be semantic-consistent variants of the original ones. A conventional method is lexical replacement, including synonym replacement with WordNet [66, 35, 52], word embedding substitution [19, 49], and masked language modeling [13], etc. Backtranslation is an effective method to generate samples that are semantically invariant [40], and it strongly promotes the development of unsupervised machine translation [20]. Xie *et al*. [53] applied backtranslation for text classification and reached the state-of-the-art performance. Other methods include random noise injection [53, 52], syntax tree manipulation [5],

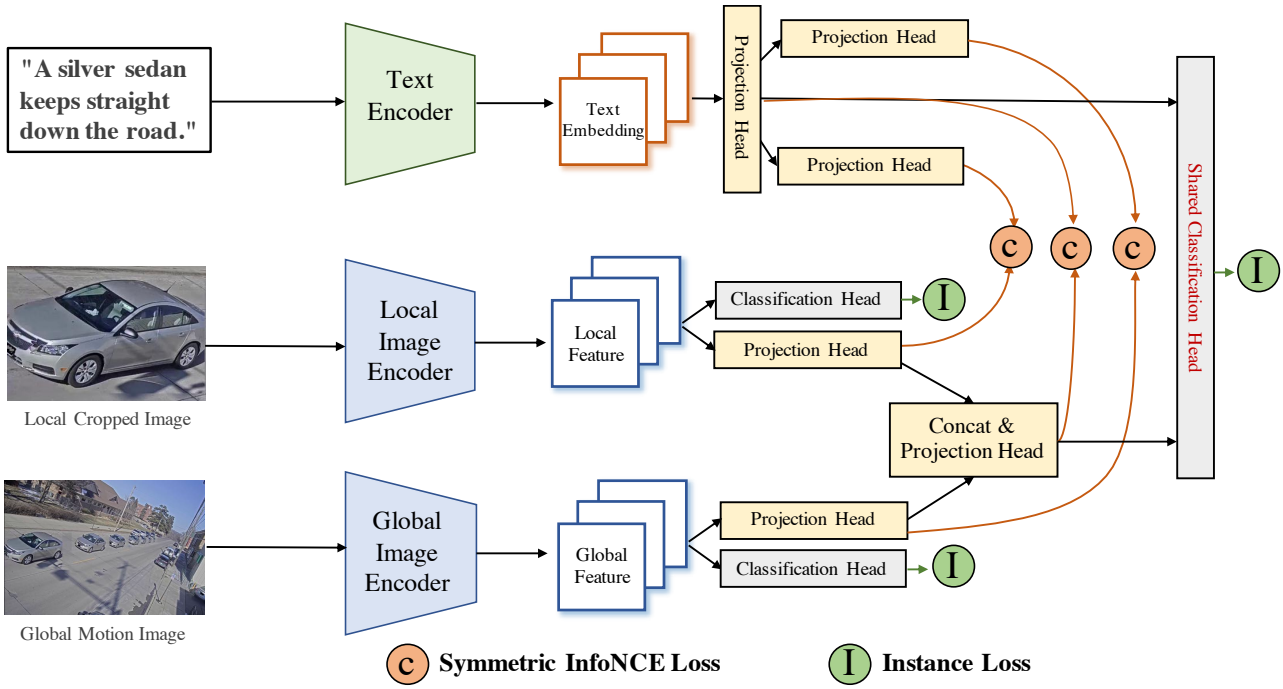

Figure 1. **The overall framework.** The vanilla network only contains one local image encoder for local cropped vehicle image and one text encoder for language description input. We further introduce the global image encoder to help learning more position information as well as the environment from motion images to ease the matching difficulty. In this paper, we also explore different optimization objectives, including Symmetric InfoNCE Loss [36] and Instance Loss [70].

mixup [14], etc.

3. Method

In this section, we provide one detailed illustration of the proposed framework. In particular, we first start with the data augmentation strategies, which include the motion modeling and description augmentation. Followed by the data augmentation, the representation learning contains the description of the network structure and optimization functions.

3.1. Data Augmentation

3.1.1 Motion and background modeling

Compared with the image-based vehicle retrieval, which applies vehicle images as queries for fine-grained appearance modeling, the natural language descriptions contain more surrounding factors and the motion information. The inherent attributes of vehicles are not enough to distinguish the specific target. For example, two white SUVs that go straight and turn left are difficult to distinguish only through the color and type of the vehicles. Therefore, the introduction of global information, such as background, is of vital importance for the accurate natural language-based vehicle retrieval. We propose a simple but effective way to introduce global information. As shown in Figure 2, we adopt background and trajectory modeling to preserve environment and motion information as a motion image. In particular, the generation of motion images consists of three steps.

Firstly, we notice that the camera position is fixed, and maintains the same angle of view in video clips. It means that the background in the same video is stable. The method continuously calculating the weighted sum of input frame can enhance the static parts and remove the moving vehicles, which is widely applied in traffic anomaly detection [1, 57]. Specifically, we calculate the mean value of each frame in the same video to generate background images. It can be formulated as:

$$B = \frac{1}{N}\sum_{i}^{N} F_i, \tag{1}$$

where F_i is the i_{th} frame, B is the background image, and N is the number of video frames. For the AICity training/test data, the environmental information is preserved in the background image, including parking lots, intersections and traffic lights.

Secondly, the motion information of the vehicle reflects

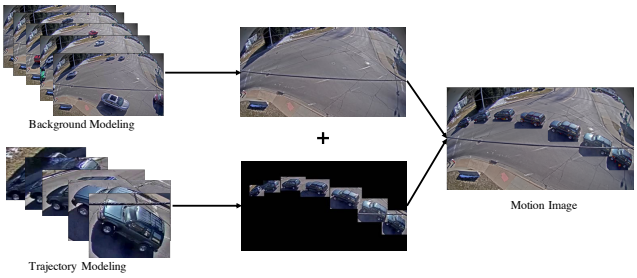

Figure 2. **Motion and background modeling.** Here we show the steps to obtain motion maps. In particular, we average the whole views into one consistent background, and then past the trajectory vehicle bounding boxes with the time gap.

that the position of the vehicle is different at different moments. We continuously cover the crop of the detection box to the trajectory image T. In particular, since the movement distance of consecutive frames is limited, we use interval frames for coverage.

$$T_{box_i} = F_{box_i}, \tag{2}$$

where box_i is the detection box in the i_{th} frame. T is the trajectory image, and it is initialized with the zero matrix as large as the video frames.

Finally, we copy the detected bounding boxes, and paste the vehicles of different timestamps to the background image as the motion image.

3.1.2 Description Augmentation

In order to provide more training data and enhance the model robustness, we apply text data augmentation. Specifically, in our practice, we use backtranslation to generate semantic invariants for the training samples. We collect all the command texts and apply translation and backtranslation with the in-house application Ali-translate. We observe the text data and find that the commands are mostly of short length and without complex syntactic structure. Translating to languages that are similar to English, such as French and German, may cause backtranslation's generating the same texts. Instead, we translate the texts to Chinese and backtranslate them to English. We demonstrate some examples of translation and backtranslation in Figure 3.

In addition, to avoid the interference caused by multiple vehicle descriptions in the text, we enhance the text by strengthening the subject. As shown in Figure 4, the target vehicle is often the subject of the text description. Therefore, we use "spacy "[1] extract the subject as a separate sentence and put it at the beginning. At the same time, the subject of the three sentences forms the fourth text description.

[1] https://spacy.io/

Text: A mid-sized black SUV drives straight down a road behind another SUV.
*Translation:*一辆中型黑色SUV在另一辆SUV后面的道路上直奔。
Back-translation: A medium-sized black SUV went straight on the road behind the other.

Figure 3. An example of the translation and back-translation as description augmentation. We first translate the original training sentence from English to Chinese, and then translate the sentence back to the English format.

"A mid-sided blue sedan **goes straight** through an **intersection** behind a **blue vehicle**."
"A black sedan **keeping straight** down the street followed by another **black vehicle**."
"A black sedan **goes down** the straight after a **blue sedan**."
" A mid-sided blue sedan. A mid-sided blue sedan **goes straight** through ……. "
" A black sedan. A black sedan **keeping straight** down the street followed ……"
" A black sedan. A black sedan **goes down** the straight after a **blue sedan**. "
" A mid-sided blue sedan. A black sedan. A black sedan "

Figure 4. An example of strengthening the subject in the text description.

3.2. Cross-Modal Representation Learning

Natural language-based vehicle retrieval aims to retrieve the specific vehicle according to the text description. These texts describe the inherent attributes of the vehicles (*e.g.*, color, type, and size), as well as external factors such as the behavior of the vehicle and the surrounding environment. At the network structure level, we construct a dual-stream image encoder for visual representation learning, the dual-stream architecture takes local detected objectives, *i.e.*, vehicle, as input and global motion images separately to pay attention to the inherent attributes and external factors of the vehicles. In addition, the pretrained text encoder is utilized to extract text embedding. We revisit several losses in terms of language-based vehicle retrieval. For instance, following the natural language supervision in CLIP [39], the symmetric InfoNCE [36] loss is adopted to learn multi-modal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings. Furthermore, we introduce the instance loss [70] to learn the instance-level features and the self-supervised barlow twins loss [62] to demand the feature to be as diverse as possible.

Network structure. Following existing re-id works, we adopt the strong networks pretrained on ImageNet [6] as the vision backbone module. We have adopted two vision structures:

- The vanilla re-id baseline[2] is used to extract the visual representation for cropped vehicle. We follow the code[3] of the 4*th* AICity vehicle re-id to pre-train the model on the data of track2 this year. In particular, the

[2] https://github.com/layumi/Person_reID_baseline_pytorch
[3] https://github.com/layumi/AICIty-reID-2020

visual backbone is SE-ResNeXt50 [17]. We then finetune the model on the track5 data with instance loss to extract the visual feature. Only the cropped vehicle images are considered in this baseline model. We extract the final 512-dim feature before the classification layer as the visual representation.

- To learn the motion information, we further adopt a dual-stream structure. The inputs consist of local cropped images and global motion images. The local cropped images are the detected vehicles cropped from a random frame. The global motion images are generated as illustrated in Section 3.1.1. The dual-stream structure contains two independent CNN encoders pretrained on ImageNet as the backbone, including SE-ResNeXt50 [17] or EfficientNet B3 [46]. In particular, for each stream, we introduce projection heads to map visual representation to the spaces of contrastive representation learning and instance fine-grained feature learning. The projection head uses a MLP with the hidden layer to obtain

$$z_i = g_i(h_i) = W_2\sigma(BN(W_2h_i)), \tag{3}$$

where BN is a Batch Normalization (BN) layer, σ is a ReLU layer, and the output dimension is 512. h_i is the visual features extracted by the backbone. As shown in Figure 1, there are three projection heads, corresponding to local detail features, global motion features and fusion features. In addition, the classification heads are applied to output the predicted possibility of different tracks. The classification head is similar with the projection head, but the output dimension is the number of tracks.

For text embeddings, we deploy pretrained BERT [7] or RoBERTa [29] as text encoder. Similar with the image encoder, the projection head is introduced to map text embeddings to the space of contrastive representation learning. But the BN is replaced with the Layer Normalization (LN) layer. Due to the limited amount of text data, the parameters of text encoder are fixed or updated with a small learning rate.

$$z_t = g_t(h_t) = W_2\sigma(LN(W_2h_t)), \tag{4}$$

where h_t is the text embeddings extracted by the pretrained model.

3.3. Optimization Objectives

3.3.1 Contrastive Loss

To maximize the cosine similarity of the image and text embeddings, we utilize symmetric InfoNCE [36] loss like CLIP [39]. Specially, we optimize the symmetric InfoNCE in three levels to achieve well-aligned with the given description: (1) local cropped image region and sentence, (2) global motion image and sentence, (3) fusion feature and sentence. We define the score function following previous work in contrastive learning:

$$\mathcal{S} = cos(z_{img}, z_{text})/\tau, \tag{5}$$

where $cos(u, v) = \frac{u^T v}{||u||||v||}$ denotes cosine similarity, and τ denotes a temperature learnable parameter initialized with 1. This maps the image and text representations into a joint embedding space.

Given a batch of M image-text pairs, it consists of $M \times M$ possible sample pairs. The symmetric InfoNCE has two parts: Text-to-Image and Image-to-Text. Text-to-Image compares one positive pair with $M-1$ Negative pairs for each text description:

$$\mathcal{L}_{t2i} = \frac{1}{M}\sum_{i=1}^{M} -\log \frac{exp(cos(z_{img,i}, z_{text,i})/\tau)}{\sum_{j=1}^{M} exp(cos(z_{img,j}, z_{text,i})/\tau)}. \tag{6}$$

Meanwhile, Image-to-Text optimizes one positive pair with $M-1$ Negative pairs for each track:

$$\mathcal{L}_{i2t} = \frac{1}{M}\sum_{i=1}^{M} -\log \frac{exp(cos(z_{img,i}, z_{text,i})/\tau)}{\sum_{j=1}^{M} exp(cos(z_{img,i}, z_{text,j})/\tau)}. \tag{7}$$

The symmetric InfoNCE is formulated as:

$$\mathcal{L}_{SNCE} = \lambda_1\mathcal{L}_{t2i} + \lambda_2\mathcal{L}_{i2t}, \tag{8}$$

where λ_1, λ_2 are the weights of Text-to-Image and Image-to-Text. Due to the evaluation of Text-to-Image manner, we set $\lambda_1 = 2, \lambda_2 = 1$.

InfoNCE loss between local cropped image and sentence. The local cropped image contains the inherent attributes of the vehicle (e.g., color, type, and size). These inherent attributes should be consistent with corresponding words in the text description, which are often the text description about the subject. In order to strengthen the relationship between the target vehicle image and the subject of the text description, we adopt the description augmentation in Section 3.1.2.

InfoNCE loss between global motion image and sentence. The global motion image reflects the motion of vehicle and the external factors of the surrounding environment. As illustrated in section 3.1.1, our motion images can effectively retain these information, and the role of these external factors can be grasped through the supervision of the global motion map. .

InfoNCE loss between fusion feature and sentence. Fusion of local and global features using nonlinear mapping takes advantage of neural networks to mine some complex

associations. During the inference, we only use the fused features as the representation of the retrieval.

At three levels, we use the symmetric InfoNCE in Eq. 8 to optimize the learning of contrastive representations, and the weight of each level is 1.

3.3.2 Instance Loss

Instance loss is one common objective in the bi-directional image-text retrieval task to capture the global discrepancy [70], and we also explore this loss in terms of the natural language-based vehicle retrieval task. We treat every track and the corresponding descriptions as one category. The optimization goal is to mapping the visual and textual input into one shared classification space. In particular, we adopted one shared classifier for both visual and textual inputs, and enforce the model to learn the mapping function.

$$\mathcal{L}_i = -log(W_{shared}\ z_i), \tag{9}$$

$$\mathcal{L}_t = -log(W_{shared}\ z_t), \tag{10}$$

where z_i and z_t are the visual and textual embedding defined in Eq. 3 and Eq. 4, and W_{shared} denotes the weight of the final linear classifier. Instance loss can be formulated as:

$$\mathcal{L}_{instance} = \mathcal{L}_i + \mathcal{L}_t. \tag{11}$$

It is worth noting that the instance loss is different from the symmetric infoNCE in whether it optimizes the cosine similarity within one mini-batch or the stored classification weights of all image-text pairs. As shown in Table 2, the instance loss is complementary to the contrastive loss, which further boosts the performance.

3.3.3 Barlow-twins Loss

Barlow-twins loss [62] is an optional loss in the proposed framework. We trained three models based on such loss for the ensemble. This loss is similar to the CLIP loss [39] but it conducts the feature multiplication in the feature channel, which results in one totally different request to the learned feature. Actually, this loss can be viewed as one regularization term. It asks the model to learn one orthogonal feature, where every channel contains a different semantic meaning from the rest channels.

4. Experiment

4.1. Dataset Analysis

Natural language (NL) description offers another useful way to specify vehicle track queries. The dataset for Natural Language-Based Vehicle Retrieval Track is built upon the CityFlow Benchmark by annotating vehicles with natural language descriptions. This dataset contains 2498 tracks

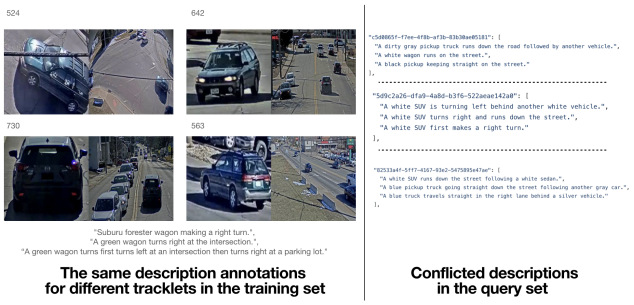

Figure 5. Noise in the training & test set. We observe two kinds of noise existed in the dataset. The identical three descriptions are annotated for different vehicle tracklets in the training set (left). Conflicting descriptions existed in the query set (right).

Rank	Team Name	MRR
1	**Alibaba-UTS (Ours)**	**0.1869**
2	TimeLab	0.1613
3	SBUK	0.1594
4	SNLP	0.1571
5	HUST	0.1564

Table 1. Competition results of AI City Natural Language-Based Vehicle Retrieval Challenge.

of vehicles with three unique natural language descriptions each. 530 unique vehicle tracks together with 530 query sets with three descriptions are curated for this challenge.

Noise in the training & test set. As shown in Figure 5, we observe that noise exists in both training and test set. The main noise is from the same three descriptions for different vehicle tracklets. There are 323 tracklets sharing the identical three sentences with another tracklets. A similar phenomenon is observed on the query set, and there are 56 queries containing the identical three textual descriptions. Such textual input compromises the training process as well as the inference accuracy. Besides, we also observe the conflicting descriptions in the query set. For instance, "turning left" and "turns right" simultaneously appear in one description group. We can not optimize such noise but deploy the mean textual feature to find the most similar samples.

Evaluation. The Vehicle Retrieval by NL Descriptions task is evaluated using standard metrics for retrieval tasks. The Mean Reciprocal Rank (MRR) is used [48]. It is formulated as :

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}, \tag{12}$$

where $rank_i$ refers to the rank position of the right track for the i_{th} text description, and Q is the set of text descriptions.

4.2. Quantitative Results

During the inference, we average all the frame features of the target in each track as track features, the embedding

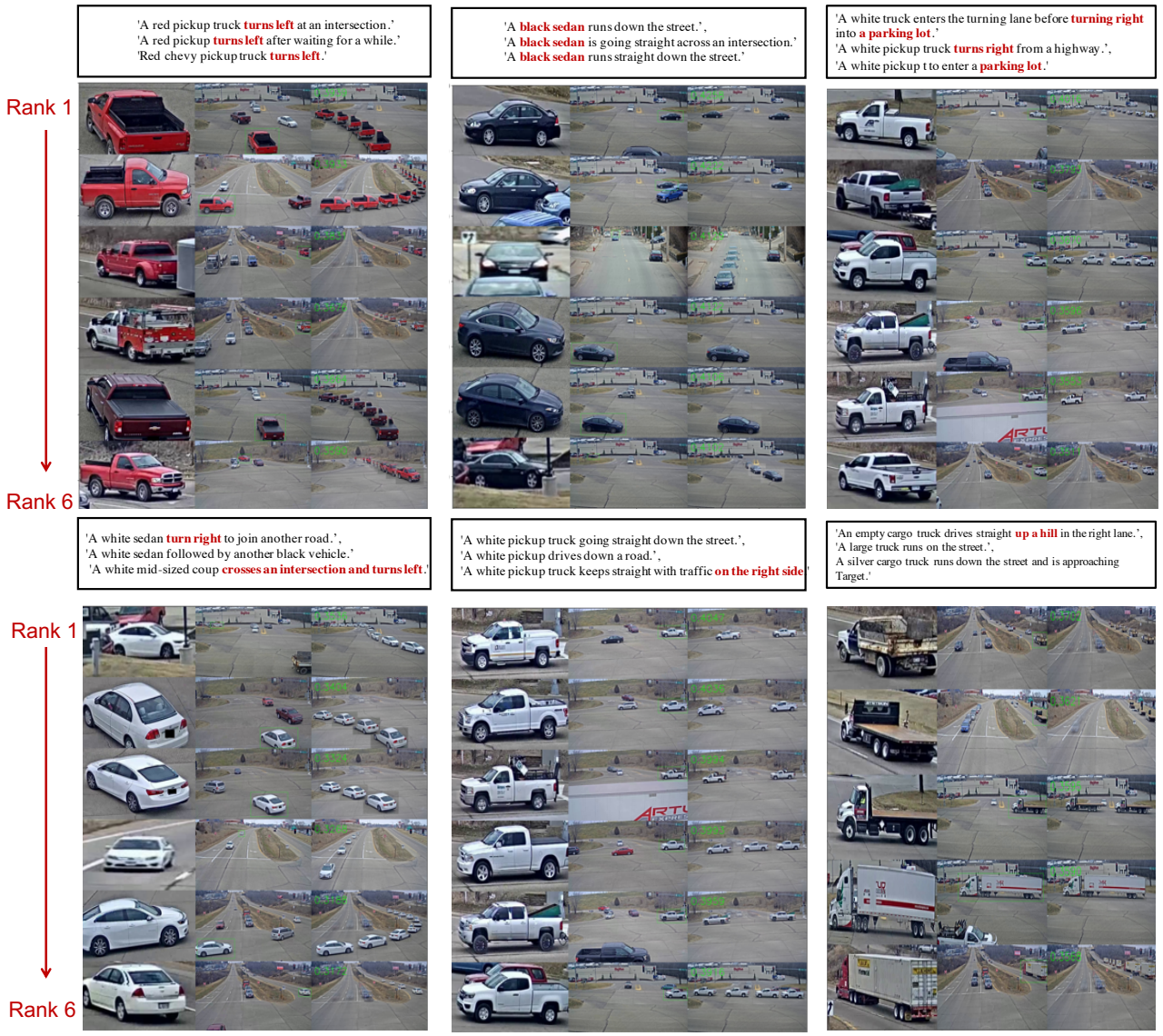

Figure 6. **Qualitative Results.** We highlight "key" words in red and show the proposed method can find the right matches with the fine-grained attention.

Method	Performance					
Baseline	✓	✓	✓	✓	✓	✓
Instance Loss		✓	✓	✓	✓	✓
Motion Image			✓	✓	✓	✓
NLP Augmentation				✓	✓	✓
Large Size& Model					✓	✓
Ensemble						✓
MRR(%)	8.25	9.65	13.21	14.56	19.27	20.77

Table 2. Ablation Study on TestA in the online evaluation system.

described by three sentences is also averaged as the query features. The cosine distance is used for ranking as the final result.

Comparsion with Other Teams. As shown in Table 1, the proposed method has achieved the state-of-the-art MRR, i.e., 0.1869, which is superior to the second-best team by a large margin. Moreover, the consistent performance on all Test datasets demonstrates the effectiveness and robustness of the proposed method.

Abation Study. As illustrated in Table 2, we perform ablation studies with different modules of our proposed method. The "Baseline" donates that CLIP [39] with ResNet50 as image encoder and BERT-BASE as text en-

coder. The "Instance Loss" optimizes the cross-entropy loss function for distinguishing each track."Motion Images" donates the dual-stream architecture with local cropped image and global motion images. The introduction of motion images gains a relative MRR improvement of 36.5%, which demonstrates the external factors and motion information play a vital role in natural language-based vehicle retrieval. Our method of motion and background modeling is a simple and effective manner to capture these information. The "NLP augmentation" consists of strengthening the subject description and backtranslation. "Large Size&Model" means that using larger pretrained models, such as RoBERTa [29] as text encoder and ResNeXt101 as image encoder. The size of image input is improved to 320×320. The obvious improvement of larger pretrained model proves the importance of Large-scale language pretraining. The well-pretrained text encoder provides a good Initialization embedding space, especially in the case of lack of text content. On the 50% Test set, we improve the CLIP baseline from 0.0825 to 0.1927 mAP MRR with single model.

4.3. Qualitative Results

Furthermore, we visualize the ranking results in Figure 6, which shows the effectiveness of the proposed method. All the top-6 samples are relevant to the query descriptions. We highlight "key" words in the description and show the proposed method can find the right matches with the fine-grained attention.

5. Conclusion

In this paper, we propose a robust natural language-based vehicle search system for smart city applications. To connect the vision and language modalities, we jointly train the state-of-the-art vision model and transformer-based language model with the symmetric InfoNCE loss and instance loss. Further, we design a two-stream architecture to incorporate both local details and global information of vehicles, and apply the text augmentation technique backtranslation to enhance the model robustness. Finally, the system achieves 18.69% MRR accuracy and reaches the first place in the natural language-based vehicle retrieval track of the 5th AICity Challenge. In the future, we will continually explore the large-scale and efficient vehicle search technique for the intelligent transportation system, such as more elaborate model architectures, more powerful optimization objectives and more abundant data augmentation methods.

References

[1] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, Ming Sun, and Junjie Yan. Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In *CVPR Workshops*, pages 117–124, 2019. 3

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017. 2

[4] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. 2019. 2

[5] Claude Coulombe. Text data augmentation made simple by leveraging NLP cloud apis. *CoRR*, abs/1812.04718, 2018. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 5

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9346–9355, 2019. 1, 2

[9] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. *arXiv preprint arXiv:2103.10699*, 2021. 2

[10] Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. 2021. 1

[11] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Eur. Conf. Comput. Vis.*, volume 5. Springer, 2020. 2

[12] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *Int. Conf. Comput. Vis.*, pages 5277–5285. IEEE, 2017. 2

[13] Siddhant Garg and Goutham Ramakrishnan. BAE: bert-based adversarial examples for text classification. In *EMNLP 2020*, pages 6174–6181, 2020. 2

[14] Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. *CoRR*, abs/1905.08941, 2019. 3

[15] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Int. Conf. Comput. Vis.*, pages 5803–5812, 2017. 2

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

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

[18] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 2

[19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. In *EMNLP 2020*, pages 4163–4174, 2020. 2

[20] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *ICLR 2018*, 2018. 2

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *arXiv preprint arXiv:2102.06183*, 2021. 2

[22] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2

[23] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2657–2664, 2014. 2

[24] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021. 2

[25] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 1, 2

[26] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, 2016. 2

[27] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, pages 15–24. ACM, 2018. 2

[28] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, 2016. 2

[29] 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. 1, 5, 8

[30] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *CVPR*, 2019. 2

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

[32] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[34] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*, pages 19–27, 2018. 2

[35] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *AAAI 2016*, pages 2786–2792, 2016. 2

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 3, 4, 5

[37] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *Eur. Conf. Comput. Vis.*, pages 651–667. Springer, 2016. 2

[38] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. *CVPR*, 2017. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 4, 5, 6, 7

[40] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *ACL 2016*, 2016. 2

[41] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10444–10452, 2019. 2

[42] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[43] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 2

[44] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Int. Conf. Comput. Vis.*, pages 7464–7473, 2019. 2

[45] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, 2018. 2

[46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 5

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

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

[49] William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *EMNLP 2015*, pages 2557–2563, 2015. 2

[50] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *CVPR*, 2021. 2

[51] 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 *ICCV*, 2017. 2

[52] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP 2019*, pages 6381–6387, 2019. 2

[53] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *NeurIPS 2020*, 2020. 2

[54] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Eur. Conf. Comput. Vis.*, pages 305–321, 2018. 2

[55] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In *Adv. Neural Inform. Process. Syst.*, volume 15, page 12. Citeseer, 2002. 2

[56] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, volume 29, 2015. 2

[57] Yan Xu, Xi Ouyang, Yu Cheng, Shining Yu, Lin Xiong, Choon-Ching Ng, Sugiri Pranata, Shengmei Shen, and Junliang Xing. Dual-mode vehicle motion pattern learning for high performance road traffic anomaly detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 145–152, 2018. 3

[58] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Int. Conf. Comput. Vis.*, pages 1453–1462, 2017. 2

[59] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106*, 2017. 2

[60] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. Video captioning and retrieval models with semantic attention. *arXiv preprint arXiv:1610.02947*, 6(7), 2016. 2

[61] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3165–3173, 2017. 2

[62] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 4, 6

[63] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Eur. Conf. Comput. Vis.*, pages 374–390, 2018. 2

[64] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1247–1257, 2019. 2

[65] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, pages 12870–12877, 2020. 2

[66] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS 2015*, pages 649–657, 2015. 2

[67] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10668–10677, 2020. 2

[68] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: learning robust visual representation for vehicle re-identification. *IEEE Transactions on Multimedia*, 2020. 1, 2

[69] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. *CVPR*, 2019. 2

[70] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020. 1, 3, 4, 6

[71] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv preprint arXiv:1805.02834*, 2018. 2