DUN: Dual-path Temporal Matching Network for Natural Language-based Vehicle Retrieval

Ziruo Sun*
Shandong University
ziruosun@foxmail.com

Xinfang Liu*
Shandong University
xinfangliu@qq.com

Xiaopeng Bi
Xidian University
bixiaopeng1996@163.com

Xiushan Nie †
Shandong Jianzhu University
niexiushan@163.com

Yilong Yin †
Shandong University
ylyin@sdu.edu.cn

Abstract

*Retrieving vehicles matching natural language descriptions from collections of videos is a novel and uniquely challenging task, requiring consideration not only of vehicle types and colors, but also of temporal relations, e.g., "A white crossover keeping straight behind a silver hatchback." To perform this task, we propose **Du**al-path Temporal Matching **N**etwork (DUN). DUN uses a pre-trained CNN and GloVe to extract visual and text features, respectively, and GRUs to mine temporal relationships in videos and sentences. Furthermore, the proposed network can attain superior performance by including techniques such as re-ranking. With its simple structure, DUN achieved second place on the AI City Challenge 2021 Track 5. The codes are available at https://github.com/okzhili/AICITY2021_Track5_DUN.*

1. Introduction

Developing realistic machine learning models of events that occur on roads, and thus the capability to retrieve specific vehicles from video data, has a wide range of applications in urban planning, traffic engineering, and law enforcement. In these application scenarios, querying using natural language is undoubtedly one of the most efficient and convenient methods possible. Therefore it is necessary to develop methods based on natural language queries enabling the tracking and retrieval of vehicles from surveillance videos taken at different locations and times.

Conventional retrieval methods are labor-intensive, and with the proliferation of large numbers of road surveillance cameras, the urgency of using computers to solve this task is becoming even more pronounced. Although this application may seem implausible on first acquaintance, the idea is gradually becoming feasible with the continuous development of computer hardware and software, especially the advanced progress of artificial intelligence in computer vision and natural language processing.

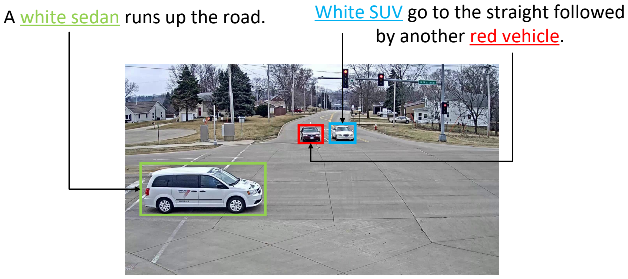
A white sedan runs up the road.
White SUV go to the straight followed by another red vehicle.

Figure 1. An example illustration of the Vehicle Retrieval (VR) by Natural Language (NL) Description task. A query sentence describes static features, such as the color and type of vehicle, as well as dynamic features such as action and state.

Formally, this task is known as Vehicle Retrieval (VR) by Natural Language (NL) Description. As shown in Fig.1, given a sentence in natural language, VR by NL aims to find clips that best match this description from a collection of road surveillance videos. This work is important for building smart cities, with applications in diverse fields such as road planning, traffic engineering, and law enforcement. In fact, the task of vehicle retrieval has been studied for some time. In contrast to the objectives of previous tasks, the task targeted in this study emphasizes the use of natural language for queries, which is considerably more challenging.

The primary problem is to develop models of video and sentence information, that is, to perform feature extraction. For visual feature extraction, commonly used pre-trained feature extractors, such as CNN networks pre-trained on

[*] Equal contribution.
[†] Corresponding authors.

ImageNet [8], are as yet not capable of expressing information on vehicles well, which can lead to a loss of detail to some extent. Moreover, modeling temporal information in videos also remains a major challenge. Although 3-dimensional convolutional networks (C3D) [33] and I3D [4] networks, which are popular in the field of action recognition, can exploit temporal information, their robustness is limited owing to the lack of sufficiently large amounts of available training data. A suitable feature extractor also needs to be carefully selected to extract textual features of query statements. Some large networks such as BERT [9] and GPT-3 [2] need to be considered with caution because of the limitations of data volume and computational cost. Once the individual obstacles in feature extraction are removed, the task of VR by NL can be transformed into a very mature cross-modal retrieval problem.

Based on the above considerations, we designed a simple network, **Du**al-path Temporal Matching **N**etwork, from scratch. Our proposed method first adopts ResNet-IBN [25], which is widely used in vehicle re-identification, to extract vehicle features from a video database. For query sentence feature extraction, DUN uses GloVe [27]. Subsequently, in order to mine temporal relationships between video frames and between words, bidirectional gated recurrent units (GRUs) [7] are applied. Finally, the features of these two modalities are mapped to a common subspace through fully connected layers. In summary, DUN has a very simple structure. Combining circle-loss and some post-processing measures, the metrics of our model nonetheless exceed the officially provided baseline by a large margin.

2. Related Works

Essentially, VR by NL is a cross-modal retrieval problem, specifically that of retrieving videos using text. Therefore, it is largely similar to general NL-based video retrieval tasks. However, VR by NL tasks can also draw on research on vehicle re-identification. In this subsection, NL-based video retrieval and vehicle re-identification will be briefly introduced.

2.1. NL-based Video Retrieval

NL-based video retrieval finds corresponding videos using a query sentence. This can take many forms. The most intuitive approach is to find the most relevant video from a collection of videos based on the query sentence, i.e., video-text retrieval; an alternative approach is to locate start and end times of a moment in a single video, i.e., natural language moment localization. Many video clips in the dataset used in this study are from the same surveillance camera and the bounding boxes of the vehicles are given; therefore, there are significant similarities with both task objectives.

Video-Text Retrieval. The study of cross-modal retrieval has an early origin [12, 18, 19], and its essence is to map the features of two different modalities to a common subspace. In this space, the similarities of different modal features reflect the proximity of their original semantics. The initial research on cross-modal retrieval focused on image-text retrieval, and video-text retrieval [10, 24, 26, 36] was studied subsequently. The models developed in the literature on video-text retrieval have relatively simple network structures, which are generally combinations of existing models and often elaborate on the loss function to improve accuracy.

Natural Language Moment Localization. Natural language moment localization has become one of the most popular research topics in machine learning in recent years [21]. Compared with the video-text retrieval task, it requires localization in a single video without traversing the video collection, which greatly reduces the number of operations and thus makes cross-modal interaction possible. Naturally, many models with very different structures, training methods, and modular structures have emerged from this task, which has greatly enriched research on machine learning methods to analyze video data.

In recent years, [1] and [13] designed the MCN and CTRL models, respectively, to accomplish moment localization. However, these models were based on a proposal window and are considered somewhat inelegant. Later, [5, 22, 38, 20] put locators after the feature interaction to reduce the number of operations. [15, 35, 3] used reinforcement learning to find the location of events by a series of steps. The boundary-aware based approach in [14, 34, 29] used a skillful method to avoid the candidate window. To reduce the cost of manual labeling, weakly supervised studies are increasingly being conducted.

2.2. Vehicle Re-identification

Vehicle re-identification (re-ID) aims to retrieve specific vehicles from different cameras with non-overlapping views. This can be regarded as a retrieval problem; given a query vehicle image, vehicle re-ID tasks aim to find all images containing a given vehicle. The trained vehicle re-ID model can be regarded as a feature extractor, which can extract vehicle features more effectively than some pre-trained models such as CNN networks pre-trained on ImageNet [8].

He *et al.* [16] proposed a multi-domain learning method to make more effective use of synthetic data. Zhu *et al.* [40] used orientation and camera similarity as penalties to obtain a final similarity to reduce the influence of background and shape. Meng *et al.* [23] proposed a parsing-based view-aware embedding network to achieve view-aware feature alignment and enhancement for vehicle re-ID tasks. Chen *et al.*[6] proposed a dedicated semantics-guided part attention network (SPAN) to robustly predict part attention masks for

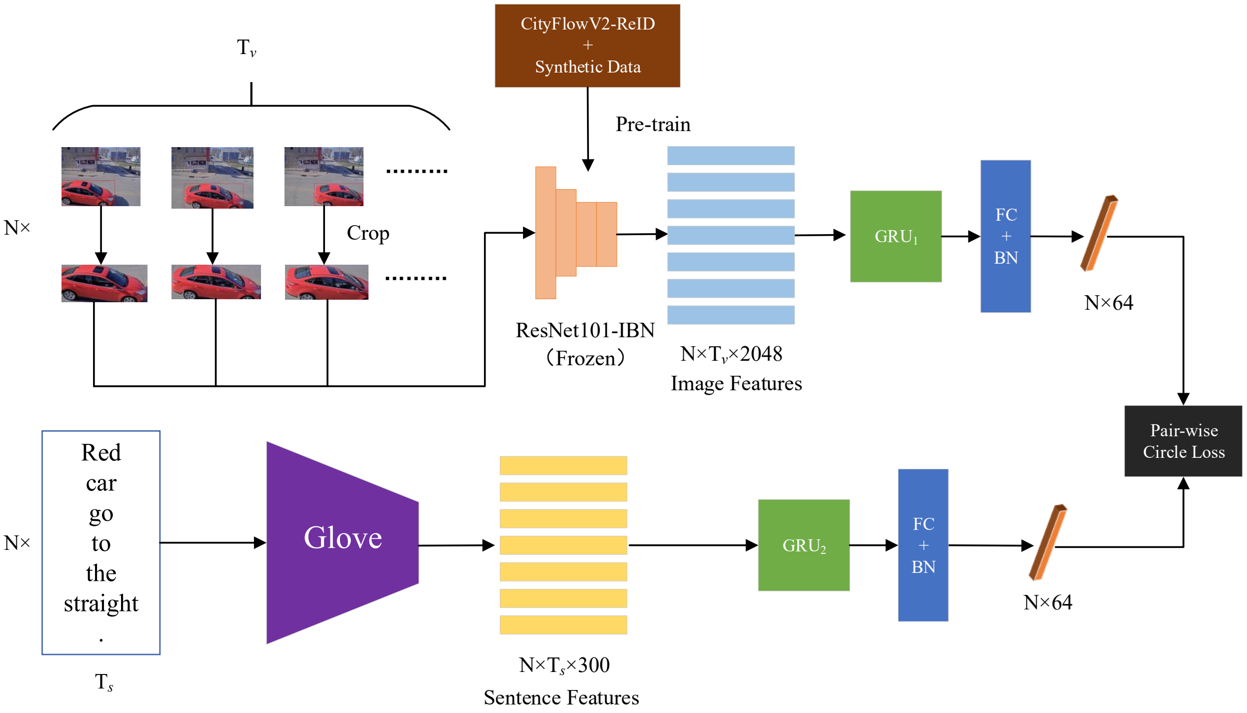

Figure 2. The workflow of Dual-path Temporal Matching Network (DUN). The video frames are fed into a pre-trained ResNet to extract features, followed by a GRU [7] network used to mine temporal relationships, and finally, a fully connected layer to obtain a feature representation of the video. The text features are represented similarly, but GloVe [27] is used to encode the word embedding in the initial stage.

different views of vehicles, given only image-level semantic labels during training.

3. Dual-path Temporal Matching Network

As shown in Fig.2, the use of Dual-path Temporal Matching Network (DUN) for VR by NL tasks can be divided into three steps, including feature extraction, model training, and post-processing. In this section, we elaborate on each in detail.

3.1. Feature Extraction

As mentioned above, to address the challenging VR by NL task, methods to best select the appropriate pre-trained features are worth considering. It is almost infeasible to train directly from scratch; not only is the computational cost very considerable, but also the performance of this approach in terms of convergence and robustness are generally insufficient. After weighing the computational cost and feature representation capability of various models, we chose a ResNet101-ibn-a network pre-trained on data of the AI City Challenge 2021 Track 2 dataset as the visual feature extractor and used GloVe [27] as the text feature extractor.

3.1.1 Visual Feature Extraction

A vehicle re-ID model was trained as a visual feature extractor to obtain a robust feature representation of the vehicle tracks. Because the data of other tracks are allowed to be used in this competition, we used the CityFlowV2-ReID dataset and synthetic data of Track 2 to train our re-ID model.

Specifically, we first combine the real dataset and a synthetic dataset based on Track 2, as well as a weakly supervised detection data augmentation method proposed by Zhu *et al.* [40] to further eliminate the background part of the real data and generate cropped data. Finally, these three parts of the data form the training set to train the vehicle re-ID model. Some common data augmentation methods, such as random erasing, random horizontal flipping, and random cropping, were utilized. For the model, we used ResNet101-ibn-a [25] pre-trained on ImageNet [8] as the backbone network of the proposed model, and generalized mean pooling (GeM) [28] as a method of feature aggregation, the performance of which was shown to be better than the commonly used global max and average pooling. In the part of the loss function, following the basic paradigm of re-ID, metric learning and classification learning are integrated, and the loss function is a combination of triplet [30, 17] and CircleSoftmax loss [31].

After training, the trained model is used to extract visual features. Specifically, the proposed method uses the provided detection box information to extract features from a cropped vehicle image instead of the original image. Finally, for each track, a set of visual features is obtained, which can be expressed as $\mathbf{V} = \{\mathbf{c}_t\}_{t=1}^{T_v} \in \mathbb{R}^{T_v \times 2048}$. Subsequently, a bidirectional GRU network is used to mine the temporal information of the feature sequences to obtain an overall representation of the video feature.

$$\mathbf{h}_v^t = BiGRU_1([\mathbf{c}_t, \mathbf{h}_v^{t-1}]). \tag{1}$$

Finally, a fully connected layer maps the feature to the common subspace.

$$f_v = \mathbf{W}_\alpha \mathbf{h}_v^{T_v} + \mathbf{b}_\alpha, \tag{2}$$

which is the final video feature used for similarity calculation.

3.1.2 Text Feature Extraction

Given a query sentence containing T_s words, GloVe [27] is used to extract word-level features, which can be denoted as $\mathbf{S} = \{\mathbf{w}_t\}_{t=1}^{T_s} \in \mathbb{R}^{T_s \times 300}$. Then, as in the case of processing video, a bidirectional GRU [7] network is applied to integrate information between words to obtain an overall feature representation of the sentence, which can be expressed as

$$\mathbf{h}_s^t = BiGRU_2([\mathbf{w}_t, \mathbf{h}_s^{t-1}]). \tag{3}$$

The output of the bidirectional GRU hidden layer at the last time step is used as the feature representation of the sentence, i.e., $\mathbf{h}_s^{T_s}$. Finally, another fully connected layer maps the features to the common subspace.

$$f_s = \mathbf{W}_\gamma \mathbf{h}_s^{T_s} + \mathbf{b}_\gamma. \tag{4}$$

where f_s represents the final sentence feature, which is used for similarity calculation.

3.2. Training

After obtaining the feature representation of the two modes of information, we perform a batch normalization operation and then use circle loss [31] for training. The original circle loss is mainly used for unimodal retrieval. For a single sample with K within-class similarity scores, i.e. $\left\{s_p^i\right\}$ $(i = 1, 2, \cdots, K)$, and L between-class similarity scores, i.e., $\left\{s_n^j\right\}$ $(j = 1, 2, \cdots, L)$, the circle loss can be expressed as

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

where Δ_n and Δ_p are between-class and within-class margins, respectively, and γ is a scale factor. s denotes the similarity between features, and cosine similarity is used in this study. α_n^j and α_p^i are non-negative weighting factors, which can be obtained by

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

In circle loss [31] , these hyperparameters are set as

$$\begin{cases} O_p = 1 + m \\ O_n = -m \\ \Delta_p = 1 - m \\ \Delta_n = m \end{cases}, \tag{7}$$

where m represents the margin. In this step, the circle loss requires only two hyperparameters, γ and m.

Although circle loss was originally designed for unimodal retrieval tasks, it can be adapted to the VR by NL task relatively simply. Specifically, because the query sentence and the video clip correspond one-to-one, within-class samples cannot be found within a single modality. Thus, we ignore the modal variability, and for any modal feature, the corresponding feature of another modality is regarded as a within-class sample, and all other features are regarded as between-class samples. This approach widens the distance between semantically distinct features. Thus, $K = 1$ and $L = 2N - 2$ in Equation 5 denote the number of positive and negative samples in a batch of size N, respectively.

3.3. Post-processing

In a unimodal retrieval task, post-processing is an effective means to improve accuracy; some post-processing methods can also be introduced to improve accuracy in multi-modal tasks. K-reciprocal encoding [39] is a common re-ranking method in unimodal tasks, which considers samples that are mutually close neighbors to be more likely to be correct samples. Similarly, in this task, when natural language and vehicle tracks are near neighbors to one another, they are considered more likely to be correct samples, and their rankings are improved accordingly.

Specifically, after training, the model is used to extract text features and track-level visual features; to calculate the initial distance matrix we calculate three distance matrices for each of the three natural language descriptions provided, and average the three distance matrices after k-reciprocal encoding re-ranking to obtain the final retrieval results.

4. Experiments

4.1. Dataset

Three datasets were used in this task, including CityFlowV2-ReID [32], and synthetic data were used to

Table 1. Ablation Study Results of DUN on CityFlow-NL test set.

Method	MRR	Recall@5	Recall@10
Baseline	0.0269	0.0264	0.0491
DUN	0.1292	0.1887	0.3283
DUN+Ensemble	0.1494	0.2377	0.3736
DUN+Ensemble+RK	**0.1613**	**0.2585**	**0.3925**

train the re-identification model, while CityFlow-NL [11] was used to train the proposed DUN.

CityFlowV2-ReID. CityFlowV2-ReID was captured by 46 cameras in a real-world traffic surveillance environment. It includes 85058 images of 880 vehicles in total; we used 440 vehicles with a total of 31238 images to train the proposed network, and the remaining 440 vehicles were used for testing.

Synthetic Data. The synthetic dataset contained 1362 vehicles with a total of 192150 images, which were generated by VehicleX [37], a publicly available 3D engine. All samples contained attribute information such as orientation, color, type, and camera.

CityFlow-NL. CityFlow-NL contains 3028 vehicle tracks collected from 40 cameras, of which 2498 vehicle tracks were used for training, and each track was annotated with three natural language descriptions, the remaining samples were used to evaluate the proposed method.

4.2. Evaluation Metrics

The vehicle retrieval by NL description task conventionally uses the mean reciprocal rank (MRR) as the main evaluation metric, which can be denoted as:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}, \tag{8}$$

where $|Q|$ is the number of query sentences and rank_i indicates the ranking of the correct answer for the i-th query sentence. Recall @ 5 and Recall @ 10 results are also reported.

4.3. Implement Details

For the vehicle re-ID model, all the images were resized to 320 × 320 pixels, and we trained the model over 12 epochs with a mini-batch of four identities and 16 images per identity. Adam was applied as the optimizer, and the initial learning rate was decayed from 3.5e-4 to 7.7e-7 using a cosine annealing scheduler. For the DUN, we set the batch size to 16 (16 tracks and 16 natural language descriptions), using cosine distance as the similarity metric, and γ and m of circle loss were set to 80 and 0.2, respectively. We trained the three models using different configurations for the final model ensemble. The first configuration uniformly samples video over 300 frames (the default DUN configuration). The second configuration samples all videos uniformly to 300. The last configuration removes the batch normalization in expectation of achieving domain adaptation on the test set. In the post-processing stage, we set $k1 = 100, k2 = 30$, and $\lambda = 0.8$ for the k-reciprocal encoding re-ranking, where $k1$ and $k2$ are number of neighbors, λ is the proportion of the original distance. Finally, the three models trained with different hyperparameters were combined in an ensemble, and the final result was obtained by averaging their distance matrices.

Table 2. Public Leaderboard of AI City 2021 Track5

Rank	Team ID	Team Name	MRR
1	132	Alibaba-UTS	0.1869
2	**17**	**TimeLab**	**0.1613**
3	36	SBUK	0.1594
4	20	SNLP	0.1571
5	147	HUST	0.1564
6	13	HCMUS	0.1560
7	53	VCA	0.1548
8	71	aiem2021	0.1364
9	87	Enablers	0.1314
10	6	Modulabs	0.1195

4.4. Performance on CityFlow-NL

As shown in Table 1, DUN achieved an MRR of 0.1292 on the CityFlow-NL test set, which significantly exceeds the given baseline, and MRR improved to 0.1613 after using the ensemble model and re-ranking (RK) strategies.

As shown in Table 2, among the performances of the top-10 team in the abovementioned competition, our team (Team ID 17) achieved an MRR score of 0.1613, taking second highest among 15 total submissions on the AI City Challenge 2021 Track 5.

5. Conclusion

In this paper, we have proposed a Dual-path Temporal Matching Network (DUN) to retrieve specific vehicle tracks from video databases using natural language. We used a pre-trained ResNet and GloVe to extract video and text features, respectively, and GRU networks to model temporal relationships. DUN has a very simple structure, but it can achieve superior performance.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of*

the IEEE international conference on computer vision, pages 5803–5812, 2017. 2

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[3] Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zheng Qin. Adversarial video moment retrieval by jointly modeling ranking and localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 898–906, 2020. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171, 2018. 2

[6] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *European Conference on Computer Vision*, pages 330–346. Springer, 2020. 2

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2, 3, 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3

[9] 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. 2

[10] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016. 2

[11] Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. 2021. 5

[12] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013. 2

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2

[14] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019. 2

[15] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8393–8400, 2019. 2

[16] Shuting He, Hao Luo, Weihua Chen, Miao Zhang, Yuqi Zhang, Fan Wang, Hao Li, and Wei Jiang. Multi-domain learning and identity mining for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 582–583, 2020. 2

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

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[19] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015. 2

[20] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020. 2

[21] Xinfang Liu, Xiushan Nie, Zhifang Tan, Jie Guo, and Yilong Yin. A survey on natural language video localization. *arXiv preprint arXiv:2104.00234*, 2021. 2

[22] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5147–5156, 2019. 2

[23] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2020. 2

[24] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer, 2016. 2

[25] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 2, 3

[26] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4594–4602, 2016. 2

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In

Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014. 2, 3, 4

[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] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2464–2473, 2020. 2

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[31] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 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. 4

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[34] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12168–12175, 2020. 2

[35] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2019. 2

[36] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 2

[37] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *The European Conference on Computer Vision (ECCV)*, page 775–791, August 2020. 5

[38] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2

[39] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 4

[40] 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. 2, 3