Cross-modal Target Retrieval for Tracking by Natural Language

Yihao Li, Jun Yu†, Zhongpeng Cai, Yuwen Pan
University of Science and Technology of China

Abstract

Tracking by natural language specification in a video is a challenging task in computer vision. Distinct from initializing the target state only by the bounding box in the first frame, language specification has a strong potential to assist visual object trackers to capture appearance variation and eliminate semantic ambiguity of the tracked object. In this paper, we carefully design a unified local-global-search framework ***from the perspective of cross-modal retrieval****, including a local tracker, an adaptive retrieval switch module, and a target-specific retrieval module. The adaptive retrieval switch module aligns semantics from the visual signal and the lingual description of the target using three sub-modules, i.e., object-aware attention memory, part-aware cross-attention, and vision-language contrast, which achieve an automatic switch between local search and global search. When booting the global search mechanism, the target-specific retrieval module relocalizes the missing target in the image-wide range via an efficient vision-language guided proposal selector and target-text match. Numerous experimental results on three prevailing benchmarks show the effectiveness and generalization of our framework.*

1. Introduction

Tracking by natural language specification is one of the most challenging tasks in computer vision, which was first introduced to the tracking field by [26]. Classical box-query based visual object tracking aims at estimating sequential states of an arbitrary target only by means of a bounding box around the object in the first video frame [31, 32], whereas the goal of tracking by natural language is to initialize the tracker with natural language specification or assist classical visual trackers with the template to cope with tricky states of the target [41]. In spite of extensive application demand, this topic area of research has not provoked tremendous research interest. In contrast to box-query based tracking methods [1,3,20,21,42] developed in recent years, the introduction of language specification provides clear semantics of the target and alleviates certain failures in object tracking caused by abrupt appearance variation, object occlusion, and distractors, as shown in Fig. 1. Generally, visual information keeps dynamic throughout a video. The target appearance and the background distribution might dramatically change along the video stream. On the contrary, the semantic information from language is static and does not drastically vary with visual signals. The visual trackers can take advantage of this invariance to adapt to appearance variation, discriminate similar objects and avoid box drift. Moreover, language description is a natural and convenient manner for human-computer interaction, especially in multi-modal scenarios.

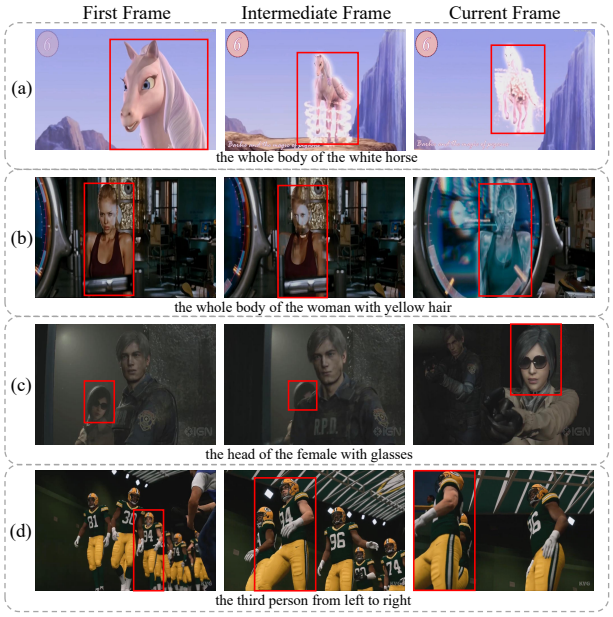

Figure 1. Examples of different challenges for tracking by natural language. The target to be tracked is annotated with a red bounding box. (a)-(b) The visual appearance of the target dramatically changes, but the semantics of the language specification never changes. (c) The target is occluded by the other foreground object. When the local tracker misses the target, we boot the global search. (d) There exists some similar distractors around the target.

However, how to exploit high-level semantics from language description to improve the performance of visual

†Corresponding author (harryjun@ustc.edu.cn)

trackers is still an open problem. Several previous proposal-based works [12, 13, 26] consider language specification as auxiliary semantic information. They fuse this information with visual signals and attempt to generate numerous object proposals related to lingual semantics, hoping to suppress the proposals containing similar objects so that they can decrease incorrect localization caused by distractors. Other grounding-based methods [41,45] refer to visual grounding algorithms that take as input natural language description to directly regress the bounding box of the target. Both prevalent paradigms adopt a global search strategy in frame images for localizing the target. Nevertheless, this scheme takes into account too much clutter background and inevitably causes box drift. In the recent method, wang et al. [41] introduced a local-global-search strategy. They design a switch module to adaptively change search strategy between local tracking and global grounding, which makes a significant improvement over all previous methods. To automatically boot the global grounding when the local tracker loses the target, they model the switch procedure as anomaly detection. However, they fuse historical target image embedding from the local tracker with language embedding in a naive manner and feed this embedding into a simple bi-directional GRUs [5]. This procedure does not explicitly consider the semantic alignment between language description and historical target images, which leads to substantial unreasonable decisions and postpones booting global grounding mechanism. As a result, the target is out of tracking completely. Furthermore, their model has a heavy reliance on the performance of the chosen grounding algorithm. Once the global grounding module feeds back incorrect clues to the local tracker, the target can no longer be captured in the following frames. Unfortunately, a powerful grounding algorithm is extremely time-consuming and data-hungry. We are not able to train a strong grounding model with both high speed and good generalization in tracking by language.

Knowledge acquired from cross-modal retrieval inspires us to solve the problems above. The semantic alignment of multi-modality has been researched for years in cross-modal retrieval where researchers focus on measuring semantic similarity between two different types of data [18]. Substantial mature algorithms on image-text retrieval/match have been designed in recent years. This provides a novel perspective for us to reconsider how to design an effective system for tracking by natural language.

In this paper, considering cross-modal retrieval has reached a relatively applicable level, we address local-global switch and global search in tracking by language as cross-modal retrieval. We design a unified system to match the language description of the target with proposal images from local search and retrieve the target from these proposals. Figure 2 shows our framework. More specifically, to align the semantics in visual proposals with language description, we first suppress the background and retain the foreground target using object-aware attention memory module. Then the following part-aware cross-attention module extracts different parts of the foreground object. These adaptive part semantics will match with different local semantics from language description by using improved transformer decoders with learnable semantic prototypes in vision-language contrast module. We also abandon visual grounding scheme for global search and present a target-specific retrieval module. This module allows us to use language as query to retrieve the target from candidates generated by the proposed vision-language guided proposal selector.

The contributions of this work can be summarized in 3 aspects:

- A novel adaptive retrieval switch module is proposed. Equipped with object-aware attention memory, part-aware cross-attention, and vision-language contrast, this module can robustly discriminate whether the target is out of the local search region.
- A target-specific retrieval module is developed to precisely capture the tracked object in a global search region. In this module, we adopt the sliding window techniques and retrieve the most possible candidates using the proposed vision-language guided proposal selector.
- Numerous experimental results on 3 prevailing benchmarks show the effectiveness and generalization of our proposed system.

2. Related Work

2.1. Tracking By Bounding Box

Given the target template in the first frame image by bounding box only, classical visual object trackers estimate the states of the target in a serious of local search regions cropped from the subsequent frames. Most existing algorithms construct a robust object appearance model for guiding deep neural networks to understand what the target looks like. One successful research branch is Siamese-network-based trackers. The pioneering work, namely siamFC [1], first exposed the advantage of siamese networks by introducing similarity learning to the tracking. Inspired by siamFC, a series of works have sprung out in recent years. SiamRPN [21] introduced Region Proposal Network(RPN) [27] to regress precise bounding boxes of the target. SiamRPN++ [20] reached higher performance by applying a spatial-aware sampling strategy. With the rise of vision transformers [8], researchers begin to explore more powerful appearance models with the attention mechanism.

STARK [42] and TransT [3] achieved state of the art by modeling long-range dependency of the target and relationship between background and foreground with transformers. Even though tracking by bounding box has made great progress, the currently developed trackers can still not be straightly applied in the real world. Because these trackers constantly miss the target because of target occlusion and appearance variation.

2.2. Tracking By natural Language

Li et al. [26] first proposed the task of tracking by natural language and designed a Lingual Specification Attention Network (LSAN) for tracking. They encoded the language query with an RNN model and extracted the visual information with a CNN model to generate two independent dynamic filters. However, the huge computational burden of the RNN module may greatly slow down the tracking speed. Unlike this, Wang et al. [40] and Feng et al. [14] embedded the natural language with a CNN to generate global proposals for tracking, among them Wang et al. [40] regarded the language cues as the extra information alongside with bounding box for tracking and Feng et al. [14] proposed to solve the problem by one-shot detection approach. Yang et al. [45] decomposed the problem into three sub-task modules: Grounding, Tracking, and Integration. With the key task "Integration", they proposed an "RT-integration" to synergistically combine the grounding and tracking and achieved effective results. In the most recent work, [41] released the TNL2K dataset and provided a baseline method based on the local-global-search strategy. Although Our framework is based on this method, the accuracy surpasses it by a large margin.

2.3. Cross-modal Retrieval

Cross-Modal Retrieval is the most basic task in cross-modal understanding. It takes one type of data as a query to retrieve another type [38], which is a very challenging task. Generally speaking, there are two solutions for common cross-modal retrieval [18]. The first idea is to fuse graphic and text features, and then learn a function that can measure cross-modal similarity through the hidden layer [19, 39]. The other method is to imply images and texts into a common feature space and obtain multi-modal representations respectively so that the similarity can be directly calculated to learn an excellent multimodal representation [9, 15, 16, 33]. Rasiwasia et al. [33] followed the latter idea and proposed a method based on canonical correlation analysis(CCA) that embeds image-text pairs as single feature vectors in a common representational space. Furthermore, Gu et al. [15] Used the Generative Adversarial Network to construct a generative adversarial task to learn a common representation across modalities, which provided a new idea for cross-modal retrieval. Unlike this, Lee et al. [19] presented a Stacked Cross Attention Network(SCAN), they used the attention interaction to get a better feature representation of the local textual-visual information and constructed the similarity function to learn under the common sorting loss. Compared with this, Wang et al. [39] designed a Scene Graph Matching(SGM) model and introduced the visual scene graph(VSG) and textual scene graph(TSG) to represent images and texts respectively, and transformed the traditional image and text retrieval problem into the matching problem of the two scene graphs. Chen et al. [2] proposed generalized pooling Operator(GPO). This operator automatically discovers the best pooling function for both image and text.

2.4. Cross-modal Transformer

Transformer [37] was first introduced by Vaswani et al and was widely used in computer vision and natural language processing. The transformer-based methods for cross-modal tasks can be roughly divided into two categories [34]: single-stream transformers [4, 23, 24, 35] and multi-stream transformers [22, 28, 36, 46, 47, 49]. As for a single-stream model, the multi-modal features are input into a single transformer block to catch the cross-modal information. VisualBERT [24] and VideoBERT [35] are the classic transformers of the single-stream models. ViLBERT [28] is a representative two-stream structure of the transformer-based model, in which they proposed a co-attention transformer layer to process both image content and natural language in separate streams. Since then, Some excellent multi-stream structures such as ActBERT [49] and DeVLBERT [47] are also followed ViLBERT by the co-attentional transformer layer.

3. Methodology

In this paper, we propose the **Ada**ptive **R**etrieval **S**witch module(**AdaRS**) and **T**arget-**s**pecific **R**etrieval module(**TSR**) from the perspective of vision-language retrieval for efficient local-global search switch and global target search in object tracking by natural language specification. Figure 2 shows the framework of our proposed system.

3.1. Local Proposal Generation

In our framework, we adopt SiamRPN++ [20] as the local tracker. The original SiamRPN++ only outputs the bounding box with the highest confidence score. However, the confidence score system in SiamRPN++ is not reliable. To further elevate the recall, we retain top-k outputs as the target proposals. All these proposals will be assessed by our designed modules.

3.2. Adaptive Retrieval Switch

Our **AdaRS** is comprised of the object-aware attention memory module(**OAM**), the part-aware cross-

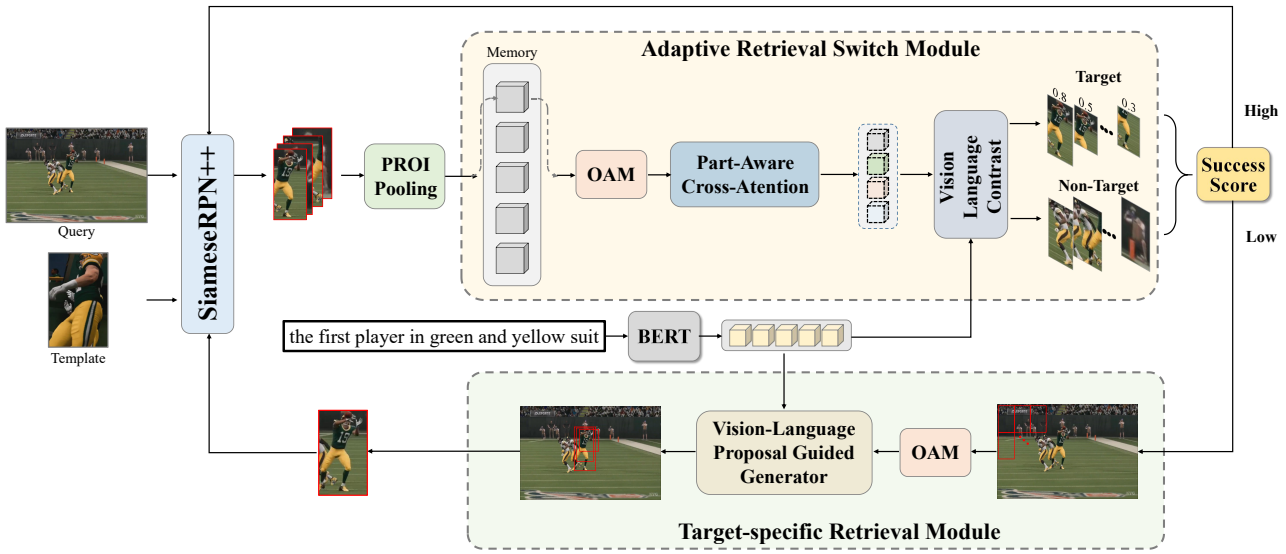

Figure 2. Overall framework of our proposed method. It consists of a adaptive retrieval switch module and target-specific retrieval module.

attention module(**PC**) and the vision-language contrast module(**VLC**). The first two modules jointly discover discriminative part semantics of the foreground object and the last module measures semantic similarity between tracking proposals and language description as vision-language retrieval among all the proposals.

For each proposal from the siamRPN++, we first obtain its representation using PROI Pooling. Then we extract the foreground information by our OAM and further obtain the part information by the PC. Here we get part features of each proposal, then we use languange description to retrieve the most possilble target.

Object-aware Attention Memory(OAM). In tracking by natural language, although objects to be tracked are arbitrary and have a tremendous discrepancy in appearance, they adhere to similar patterns which are significantly distinct from background regions. These particular patterns can be learned from data and recorded by the attention-based foreground memory. Following [30], the foreground memory stores N learnable key-value pairs $\{(\boldsymbol{k_n}, \boldsymbol{v_n})\}_{n=1}^{N}$ in order to cover various appearances of the foreground object. Each key represents a specific appearance pattern and the corresponding value denotes a foreground classifier, as shown in Fig. 3.

Given the local search region $\mathbf{S_t}$ in the frame $\mathbf{t}$, the feature maps $\mathbf{F} = [\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_i}, \ldots, \boldsymbol{f_M}] \in \mathbb{R}^{M\times H\times W\times C}$ of the proposals $\mathbf{B} = [\boldsymbol{b_1}, \boldsymbol{b_2}, \ldots, \boldsymbol{b_i}, \ldots, \boldsymbol{b_M}] \in \mathbb{R}^{M\times 4}$ from the local tracker are extracted by using PrROI Pooling proposed by [17]. For the feature map $\boldsymbol{f_i} \in \mathbb{R}^{H\times W\times C}$, we read from memory and attain a set of classifiers for each pixel. These foreground classifiers are adaptive to appearance variation and can be used to compute the foreground confidence scores with which we retain the foreground pixels and suppress the background pixels. To read from memory, we first compute query set $\mathbf{Q} = [\boldsymbol{q_{i,1}}, \boldsymbol{q_{i,2}}, \ldots \boldsymbol{q_{i,k}}, \ldots, \boldsymbol{q_{i,H\times W}}] \in \mathbb{R}^{(H\times W)\times C/16}$ for each pixel in $\boldsymbol{f_i} = [\boldsymbol{x_{i,1}}, \boldsymbol{x_{i,2}}, \ldots \boldsymbol{x_{i,k}}, \ldots, \boldsymbol{x_{i,H\times W}}] \in \mathbb{R}^{(H\times W)\times C}$ by linear projection. The similarity $s_{i,j}^n$ between each query $\boldsymbol{q_{i,j}}$ and the n-th key $\boldsymbol{k_n}$ is given as

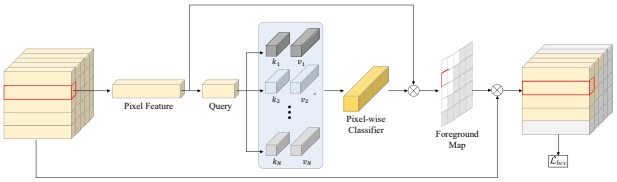

Figure 3. The architecture of OAM. This module can suppress the background features.

$$s_{i,j}^n = \frac{\beta_{i,j}^n}{\sum_{n=1}^{N}\beta_{i,j}^n},\ \beta_{i,j}^n = \frac{\boldsymbol{q_{i,j}^{\mathsf{T}}}\,\boldsymbol{k_n}}{\sqrt{C/16}} \tag{1}$$

where $n = 1, 2, \ldots, N$, $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, H\times W$, and T represents transport operator. Then we can obtain the foreground classifiers and identify the foreground pixels in $\boldsymbol{f_i}$ as

$$M_{i,j} = \boldsymbol{w_{i,j}^{\mathsf{T}}}\,\boldsymbol{x_{i,j}},\ \boldsymbol{w_{i,j}} = \sum_{n=1}^{N} s_{i,j}^n \cdot \boldsymbol{v_n} \tag{2}$$

where $\boldsymbol{x_{i,j}}$ denotes j-th pixel in $\boldsymbol{f_i}$. We can repeatedly perform the same operation on each pixel to get $\boldsymbol{M_i} \in \mathbb{R}^{H\times W}$ indicating the foreground map of $\boldsymbol{f_i}$ and compute the feature of the foreground object as

$$\boldsymbol{f_i^{obj}} = \boldsymbol{f_i} \odot \sigma(\boldsymbol{M_i}) \tag{3}$$

where $\odot$ represents element-wise multiplication and $\sigma(\cdot)$ denotes the sigmoid function. In this process, the background pixels are adaptively suppressed.

Part-aware Cross-attention(PC). A couple of proposals from the local tracker might only contain similar objects with the target. The object-aware attention memory module is not target-specific, and it will also highlight those similar foreground objects in proposals. Our goal is to distinguish the target from the similar objects with the help of language description. Although these foreground objects share a great many of appearance features, there exist a small amount of discriminative local parts that play a critical role in distinguishing the target from foreground objects. Moreover, the part-level features benefit the alignment between visual semantics and lingual semantics in the region-phrase manner. Inspired by [25], we design A learnable part prototypes $\mathbf{P} = [\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_i}, \ldots, \boldsymbol{p_A}] \in \mathbb{R}^{A\times C}$. Each of them denotes a local part pattern. To get different part features from the foreground, we adopt the modified transformer decoder.

In the previous work, Li et al. [25] retain the self-attention layer in the transformer decoder and hold the view that the self-attention layer allows the local context information propagation between prototypes during part prototype learning. However, our experiments show that the self-attention layer causes the collapse of part features. More specifically, different part prototypes generate the same part feature. Based on this observation, we abandon the self-attention layer in the transformer decoder, as shown in Fig. 4.

In cross-attention layer, each prototype aims to extract a part feature from the foreground object. By computing the similarity between each adaptive prototype and pixels in the foreground, we can decompose the specified object into different parts. Given the feature map $\boldsymbol{f_i^{obj}}$, queries arise from part prototypes $\mathbf{P}$, keys and values arise from pixels of the foreground feature map. Formally,

$$\mathbf{Q} = \mathbf{P}\mathbf{W}^Q,\ \mathbf{K}_i = \boldsymbol{f_i^{obj}}\mathbf{W}^K,\ \mathbf{V}_i = \boldsymbol{f_i^{obj}}\mathbf{W}^V \quad (4)$$

where $i = 1, 2, \ldots, M$ and $\mathbf{W}^Q \in \mathbb{R}^{C\times d}, \mathbf{W}^K \in \mathbb{R}^{C\times d}, \mathbf{W}^V \in \mathbb{R}^{C\times d}$ are linear projections.

For each foreground object $\boldsymbol{f_i^{obj}}$, we illustrate how to compute the part-aware masks by attention mechanism. We use these masks to obtain the part features. Formally,

$$\mathbf{S} = Softmax(\frac{\mathbf{Q}\mathbf{K}_i^\mathsf{T}}{\sqrt{d}}) \quad (5)$$

where $\sqrt{d}$ is a scaling factor. The corresponding part feature $\mathbf{F}_i^{part}$ is given as:

$$\mathbf{F}_i^{part} = \mathrm{FFN}(\mathbf{S}\mathbf{V}_i) \quad (6)$$

where FFN denotes the linear layers.

Vision Language Contrast(VLC). The semantic similarity between specified natural language and visual proposals from the local tracker is a useful clue. With this clue, we can precisely select the target from numerous proposals. To measure the similarity, we position this problem as cross-modal retrieval and retrieve the target with natural language from proposals. We adopt GPO [2] to obtain the visual representation and the lingual representation due to its good performance. However, GPO computes features from the global view, which does not take into account the local semantic alignment between image and text. We improve it by adding a novel module as shown in Fig. 5.

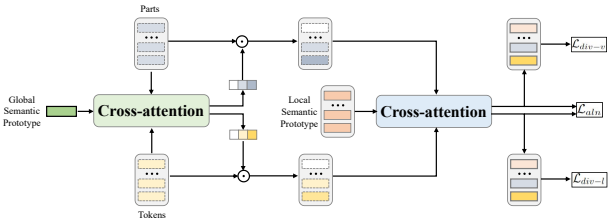

Figure 5. The proposed vision-language contrast module.

The words in language description first are embedded into feature representation using a pretrained BERT [7], which is widely used as word embedding model in natural language related task. The module takes as input part features and word features simultaneously. To capture the local semantic alignment between part features $\mathbf{F}_i^{part} \in \mathbb{R}^{A\times C}$ of i-th object proposal and word embedding $\mathbf{E} \in \mathbb{R}^{H\times C}$ of the corresponding language specification, where the H denotes the length of the sentence, we need to discover common information in two modalities first. We design a shared learnable global semantic prototype $\boldsymbol{p}_{gs}$to inquiry the common semantics between part features and word features as

$$\mathbf{Q}_{gs} = \boldsymbol{p}_{gs}\mathbf{W}_{gs}^Q \quad (7)$$

$$\mathbf{K}_i^{vgs} = \mathbf{F}_i^{part}\mathbf{W}_{gs}^K, \quad \mathbf{V}_i^{vgs} = \mathbf{F}_i^{part}\mathbf{W}_{gs}^V \quad (8)$$

$$\mathbf{K}^{lgs} = \mathbf{E}\mathbf{W}_{gs}^K, \quad \mathbf{V}^{lgs} = \mathbf{E}\mathbf{W}_{gs}^V \quad (9)$$

$$\mathbf{S}_i^{vgs} = \frac{\mathbf{Q}_{gs}(\mathbf{K}_i^{vgs})^\mathsf{T}}{\sqrt{d}}, \quad \mathbf{S}^{lgs} = \frac{\mathbf{Q}_{gs}(\mathbf{K}^{lgs})^\mathsf{T}}{\sqrt{d}} \quad (10)$$

$$\hat{\mathbf{V}}_i^{vgs} = \mathbf{V}_i^{vgs} \odot \sigma(\mathbf{S}_i^{vgs}),\ \hat{\mathbf{V}}^{lgs} = \mathbf{V}^{lgs} \odot \sigma(\mathbf{S}^{lgs}) \quad (11)$$

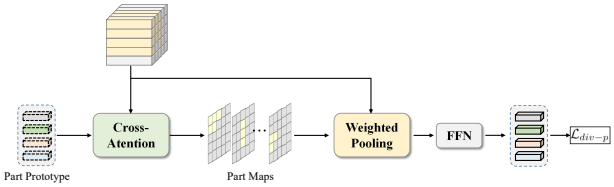

Figure 4. The architecture of PC. We use this module extract the part features of the candidates.

where $i = 1, 2, \ldots, M$, $\mathbf{W}_{gs}^{Q} \in \mathbb{R}^{C\times d}$, $\mathbf{W}_{gs}^{K} \in \mathbb{R}^{C\times d}$ and $\mathbf{W}_{gs}^{V} \in \mathbb{R}^{C\times d}$ are linear projections. $\hat{\mathbf{V}}_i^{vgs}$ and $\hat{\mathbf{V}}^{lgs}$ are shared semantics extracted by the shared global semantic prototype.

Then we feed $\hat{\mathbf{V}}_i^{vgs}$ and $\hat{\mathbf{V}}^{lgs}$ into the following standard transformer decoder with L local semantic prototypes $\mathbf{P}_{ls} = [\boldsymbol{p}_{ls,1}, \boldsymbol{p}_{ls,2}, \ldots, \boldsymbol{p}_{ls,i}, \ldots, \boldsymbol{p}_{ls,L}] \in \mathbb{R}^{L\times C}$ to further extract common local semantics in two modalities for region-phrase alignment.

$$\mathbf{Q}_{ls} = \mathbf{P}_{ls}\mathbf{W}_{gs}^{Q} \tag{12}$$

$$\mathbf{K}_i^{vls} = \hat{\mathbf{V}}_i^{vgs}\mathbf{W}_{ls}^{K}, \quad \mathbf{V}_i^{vls} = \hat{\mathbf{V}}_i^{vgs}\mathbf{W}_{ls}^{V} \tag{13}$$

$$\mathbf{K}^{lls} = \hat{\mathbf{V}}^{lgs}\mathbf{W}_{ls}^{K}, \quad \mathbf{V}^{lls} = \hat{\mathbf{V}}^{lgs}\mathbf{W}_{ls}^{V} \tag{14}$$

$$\mathbf{F}_i^{vls} = \sigma(\frac{\mathbf{Q}_{ls}(\mathbf{K}_i^{vls})^{\mathsf{T}}}{\sqrt{d}})\mathbf{V}_{vls} \tag{15}$$

$$\mathbf{F}^{lls} = \sigma(\frac{\mathbf{Q}_{ls}(\mathbf{K}^{lls})^{\mathsf{T}}}{\sqrt{d}})\mathbf{V}_{lls} \tag{16}$$

where $\mathbf{W}_{ls}^{Q} \in \mathbb{R}^{C\times d}$, $\mathbf{W}_{ls}^{K} \in \mathbb{R}^{C\times d}$ and $\mathbf{W}_{ls}^{V} \in \mathbb{R}^{C\times d}$ are linear projections. Two local semantic features from the same semantic prototype should be treated as a semantic pair. We will restrict each pair to get similar during training.

Ultimately, We concatenate the local semantic features with the global feature from GPO $\boldsymbol{f}^{gpo} \in \mathbb{R}^{O}$ along the dimension C for each modality and obtain the fine-tuned global representation $\hat{\boldsymbol{f}}_i^{vls} \in \mathbb{R}^{(L\times C)+O}$ and $\hat{\boldsymbol{f}}^{lls} \in \mathbb{R}^{(L\times C)+O}$ by fully connected layers. By computing the semantic similarity scores between each proposal and language description, AdaRS select the proposal with the highest score as the result of local search. Once the highest score is lower than the threshold, AdaRS boots the global search mechanism.

3.3. Target-specific Retrieval

Visual grounding is a common module as global search in tracking by language, whereas it is not easy to obtain a grounding model with both excellent accuracy and high speed in the scenario of tracking which requires the limited computation. Instead, we adopt the widely used sliding window techniques [29, 48] to conduct global searches. However, this manner is too time-consuming and it is almost impossible to verify each window with our AdaRS. Inspired by [43], we present a vision-language guided proposal selector to efficiently retrieve the most possible candidates from a large number of sliding windows with the information from the language description and the target image.

As shown in Fig. 6, we adopt the pretrained transformer encoder to encode target-language information. By introducing the learnable guided token, we get the semantic feature from the target template and the language specification.

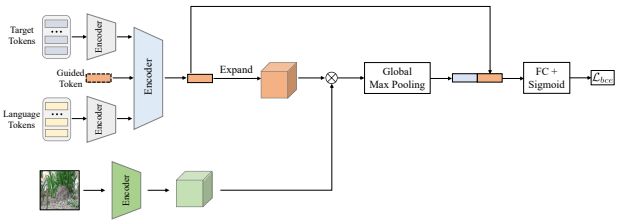

Figure 6. The proposed vision-language guided proposal selector. Here we use tiny transformer encoders to reduce the computation of this module.

Then we multiply the semantic feature and the features of window regions in the element-wise manner after the semantic feature is expanded. This operation is equivalent to attention mechanism. To improve the speed of selecting, we utilize the global max pooling to reduce computation.

When our AdaRS boots global search mechanism, we apply the selector to return top-k candidates. We average these coordinates of the selected windows as the result of global search. More importantly, this result guides the local tracker to search in the region that contains the target in the subsequent frames.

4. Training and Inference

Both of the proposed modules are trained offline. When optimizing the AdaRS, each sub-module has a corresponding loss. For object-aware attention memory, we optimize it with the binary cross-entropy loss as foreground-background classification which is different from [30], denoted as $\mathcal{L}_{bce}$. Following [25], to diversify part features and local semantic features, we adopt diversity loss as bellow.

$$\mathcal{L}_{div-p} = \frac{1}{A(A-1)}\sum_{i=1}^{A}\sum_{j=1,j\neq i}^{A}\frac{\langle \boldsymbol{f}_i^v, \boldsymbol{f}_j^v\rangle}{\|\boldsymbol{f}_i^v\|_2 \left\|\boldsymbol{f}_j^v\right\|_2}, \tag{17}$$

$$\mathcal{L}_{div-l,v} = \frac{1}{2L(L-1)}\sum_{m\in\{l,v\}}\sum_{i=1}^{L}\sum_{j=1,j\neq i}^{L}\frac{\langle \boldsymbol{f}_i^m, \boldsymbol{f}_j^m\rangle}{\|\boldsymbol{f}_i^m\|_2 \left\|\boldsymbol{f}_j^m\right\|_2} \tag{18}$$

where $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_j^v$ represent the part features from the same proposal, $\boldsymbol{f}_i^l$ and $\boldsymbol{f}_j^l$ denote the local semantic features from the same language specification. The alignment similarity is introduced to restrict the semantic alignment as Eq. (19).

$$\mathcal{S}_{corr} = \frac{\left\langle \hat{\boldsymbol{f}}^{lls}, \hat{\boldsymbol{f}}_i^{vls}\right\rangle}{\left\|\hat{\boldsymbol{f}}^{lls}\right\|_2 \left\|\hat{\boldsymbol{f}}_i^{vls}\right\|_2} + \frac{1}{L}\sum_{j=1}^{L}\frac{\left\langle \boldsymbol{f}_j^l, \boldsymbol{f}_j^v\right\rangle}{\left\|\boldsymbol{f}_j^l\right\|_2 \left\|\boldsymbol{f}_j^V\right\|_2} \tag{19}$$

where $i = 1, 2, \ldots, M$. Once we get S_{corr}, we calculate the hinge-based triplet ranking loss with online hard negative mining proposed by VSE++ [10], denoted as $\mathcal{L}_{aln}$. We jointly optimize the following loss.

$$\mathcal{L}_{total} = \mathcal{L}_{aln} + \lambda_{div}(\mathcal{L}_{div-p} + \mathcal{L}_{div-l,v}) + \lambda\mathcal{L}_{bec} \tag{20}$$

When training TSR, We sample training data from generated windows and only optimize vision-language guided proposal selector with the binary cross-entropy loss as foreground-background classification. We also add the object-aware attention memory module to the TSN. This module is trained jointly with the vision-language guided proposal selector.

In tracking by joint language and bounding box, the initial bounding box of the target is specified by a human. In tracking by natural language only , we first localize the target in the first frame using a powerful grounding model [6] and abandon this grounding model in the subsequent frames., which do not influence the speed of our tracker. The inference is listed in Algorithm 1.

Algorithm 1: Inference

Input: The language specification L_t, the initial bounding box B_t, the input image sequence $\{I_i\}_{i=1}^{F}$, the threshold T^{ars}

Output: Bounding box sequence $\{B_i\}_{i=1}^{F}$

$E_t \leftarrow \mathrm{BERT}(L_t)$;
Lingual semantics $F_t^{lls} \leftarrow \mathrm{VLC}(E_t)$;
The GPO feature $F_t^{gpo} \leftarrow GPO(E_t)$;
$B \leftarrow \{\}$;
for $i = 1, 2, 3, ..., F$ **do**
  Generate G candidates $\{C_k\}_{k=1}^{G}$ via B_t, I_i using the local tracker in the local region I_{lsr};
  $S \leftarrow \{\}$;
  for $j = 1, 2, 3, ..., G$ **do**
    The foreground feature $F_j^{obj} \leftarrow OAM(C_j)$;
    The part features $F_j^{part} \leftarrow \mathrm{PC}(F_j^{obj})$;
    Visual semantics $F_j^{vls} \leftarrow \mathrm{VLC}(F_j^{part})$;
    The GPO feature $F_j^{gpo} = GPO(F_j^{part})$;
    $S_j^{vl} \leftarrow cossim(F_t^{lls}, F_j^{vls}, F_t^{gpo}, F_j^{gpo})$;
    $S \leftarrow S \cup \{S_j^{vl}\}$;
  end
  $id \leftarrow \arg\max_i \ S_i^{vl}$;
  if $S_{id}^{vl} > T^{ars}$ **then**
    $B_i \leftarrow \{C_{id}\}$
    $B \leftarrow B \cup \{B_i\}$;
  else
    $B_i \leftarrow TSR(I_i)$;
    $B \leftarrow B \cup \{B_i\}$;
  end
end
return B

5. Experiments

5.1. Results and Comparisons

We evaluate our proposed tracker on OTB-Lang, LaSOT and TNL2K. Li et al. [26] presented two different settings of tracking by natural language only, *i.e.*, tracking by language only and tracking by joint language and box, respectively denoted as NL and NL+BBox. The former initializes the target with language only, while the latter initializes the target with bounding box and language description only provides auxiliary semantics. We validate the effectiveness of our tracker in two settings. All results are showed in Tab. 1. Experiments are also conducted using MindSpore.

Results on OTB-Lang Benchmark. Following tracking by nature language setup, Li et al. [26] in their early work annotated OTB-100 with natural language for the target, and the dataset consists of 99 videos. Only with language specification, Feng et al. [12] attain 0.78 |0.54, which is higher than other previous trackers. Our method achieves a comparable result of 0.72 |0.53. Even though our tracker does not outperform that designed by Feng et al. [12] in the NL setting, we reach state of the art in the NL+BBox setting. These experimental results on OTB-Lang demonstrate the effectiveness of our method.

Results on LaSOT Benchmark. LaSOT [11] contains 1,400 videos with auxiliary language annotation. Following the original split, we use 1,120 videos for training and 280 videos for testing. This benchmark is more challenging than OTB-Lang. The foreground objects constantly suffer occlusion and the background contains various distractors. Compared with all previous methods, our tracker achieves the best result of 0.51|0.52 in the NL setting. It also outperforms all other trackers by a large margin in the NL+BBox setting, which shows the advance of our tracker.

Results on TNL2K Benchmark. TNL2K [41] is the most recent benchmark, containing 2000 long video sequences. Each video has one sentence to describe the target and one bounding box to indiacte the localization of the target. This benchmark is more challenging than the other two benchmarks. However, our tracker still surpass all previous trackers in both tracking settings. These experiments demonstrate the effectiveness and generalization of proposed method.

5.2. Ablation Study

In this section, we conduct ablation analysis of our framework on the TNL2K dataset in the NL+BBox setting.

Effectiveness of different parts of our AdaRS. We evaluate Object-aware Attention Memory Module(OAM), Part-aware Cross-attention Module(PC) and Vision Language Contrast Module(VLC). Each module of our AdaRS acts an important role. The results are shown in Tab. 2 which shows the contribution of different modules.

Table 1. Performance on the OTB-Lang, LaSOT, TNL2K dataset. We report the results as [Prec.|Norm. Prec.|Succ. Plot].

Method	OTB-Lang				LaSOT				TNL2K			
	NL			NL+BBox			NL			NL+BBox		
Li et al. [26]	0.29	0.25		0.72	0.55		-			-		
Feng et al. [13]	0.56	0.54		0.73	0.67		-		0.56	0.50		-		
Feng et al. [12]	**0.78**	**0.54**		0.79	0.61		0.28	0.28		0.35	0.35		-		0.27	0.33	0.25
Wang et al. [40]	-			0.89	0.65		-			0.30	0.27		-			-		
GTI [45]	-			0.73	0.58		-			0.47	0.47		-			-		
Wang et al. [41]	0.24	0.19		0.88	0.68		0.49	0.51		0.55	0.51		0.06	0.11	0.11	0.42	0.50	0.42
Ours	0.72	0.53		**0.91**	**0.69**		**0.51**	**0.52**		**0.56**	**0.53**		**0.09**	**0.15**	**0.14**	**0.45**	**0.52**	**0.44**

Table 4. Recall comparisons. TSN-OAM denotes the TSN without OAM.

Method	Recall@50
baseline	0.67
TSN-OAM	0.71
TSN	**0.73**

Table 2. Performance comparison with different components.

Index	$\mathcal{OAM}$	$\mathcal{PC}$	$\mathcal{VLC}$	Prec.
1				0.410
2	✓			0.418
3	✓	✓		0.432
4	✓	✓	✓	0.451

Effectiveness of different losses of our AdaRS. We introduce four different losses to guide the AdaRS. In Tab. 3, the index-1 represents the baseline method where we only use the features from GPO to compute the hinge-based triplet ranking loss [10]. By analyzing the table, we can draw a conclusion that each loss plays an integral role when we train our AdaRS.

Table 3. Performance comparison with different losses.

Index	$\mathcal{L}_{aln}$	$\mathcal{L}_{div-p}$	$\mathcal{L}_{div-l,v}$	$\mathcal{L}_{bce}$	Prec.
1					0.418
2	✓				0.439
3	✓	✓	✓		0.447
4	✓	✓	✓	✓	0.451

Impact on recall with our TSN. To fairly compare, we only replace the TSN module in our tracker with the visual grounding model [44] which is adopted by Wang et al. [41] in their method as the global search module. The results are shown in Table 4. Our TSN significantly improves the recall of the target in the image-wide range at the cost of increasing limited computation. When the OAM is added to TSN, the recall is further elevated. Figure 7 shows results of visualization.

6. Conclusion

To solve the challenging issues in tracking by natural language specification, we propose a unified tracking framework from the perspective of cross-modal retrieval, which significantly improves the performance of the cross-modal trackers. First, we discriminate whether the local tracker loses the target or not via the adaptive retrieval switch. Once the target is out of the local search region, we then start up the target-specific retrieval to re-localize the target and update the search region for the local region. Numerous experiments on prevalent benchmarks show the strong potential of our framework in handling tracking by natural language, which can be widely used in real-world applications.

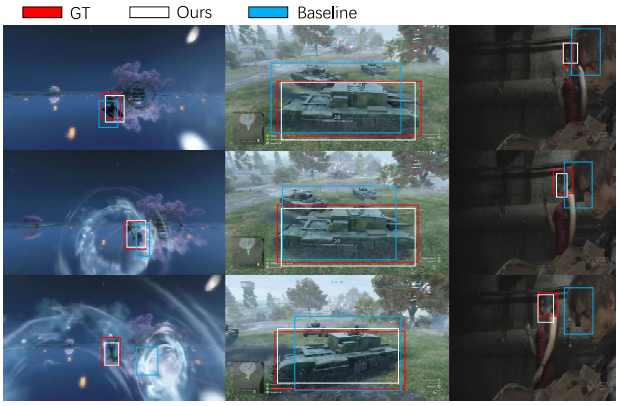

Figure 7. Visualization of the results of baseline and our method.

Acknowledgements

This work is sponsored by CAAI-Huawei MindSpore Open Fund(CAAIXSJLJJ-2021-016B), Anhui Province Key Research and Development Program (202104a05020007) and USTC Research Funds of the Double First-Class Initiative (YD2350002001). We acknowledge these funding and the corresponding supports.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 1, 2

[2] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021. 3, 5

[3] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021. 1, 3

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 3

[5] 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

[6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 7

[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. 5

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

[9] F Faghri, DJ Fleet, JR Kiros, and S Vse+ Fidler. Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 3

[10] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 6, 8

[11] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 7

[12] 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, 7, 8

[13] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Robust visual object tracking with natural language region proposal network. 2019. 2, 8

[14] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5851–5860, 2021. 3

[15] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7181–7189, 2018. 3

[16] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019. 3

[17] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018. 4

[18] Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. Comparative analysis on cross-modal information retrieval: a review. *Computer Science Review*, 39:100336, 2021. 2, 3

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 3

[20] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 1, 2, 3

[21] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 1, 2

[22] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. Semvlp: Vision-language pre-training by aligning semantics at multiple levels. *arXiv preprint arXiv:2103.07829*, 2021. 3

[23] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 3

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[25] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021. 5, 6

[26] 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. 1, 2, 3, 7, 8

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[29] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming-Hsuan Yang. Long-term correlation tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5388–5396, 2015. 6

[30] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3385–3395, October 2021. 4, 6

[31] Ajoy Mondal. Supervised machine learning approaches for moving object tracking: A survey. 1

[32] Milan Ondrašovič and Peter Tarábek. Siamese visual object tracking: A survey. *IEEE Access*, 9:110149–110172, 2021. 1

[33] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260, 2010. 3

[34] Andrew Shin, Masato Ishii, and Takuya Narihira. Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision. *International Journal of Computer Vision*, pages 1–20, 2022. 3

[35] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[36] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016. 3

[39] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1508–1517, 2020. 3

[40] Xiao Wang, Chenglong Li, Rui Yang, Tianzhu Zhang, Jin Tang, and Bin Luo. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *arXiv preprint arXiv:1811.10014*, 2018. 3, 8

[41] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021. 1, 2, 3, 7, 8

[42] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10448–10457, 2021. 1, 3

[43] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'skimming-perusal'tracking: A framework for real-time and robust long-term tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2385–2393, 2019. 6

[44] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 8

[45] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3433–3443, 2020. 2, 3, 8

[46] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. 3

[47] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020. 3

[48] Yunhua Zhang, Dong Wang, Lijun Wang, Jinqing Qi, and Huchuan Lu. Learning regression and verification networks for long-term visual tracking. *arXiv preprint arXiv:1809.04320*, 2018. 6

[49] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 3