Cross-modal Target Retrieval for Tracking by Natural Language

Jiamin Wu, Yihao Li, Jun Yu, Zhongpen Cai, Yuwen Pan

University of Science and Technology of China

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Our Motivation

- **Tracking by Natural Language(TNL)**
- aim at tracking the foreground object in a video by joint bounding box and natural language or only by natural language.
- **Previous methods**
- Fuse visual signal with lingual information or adopt some global grounding algorithms to localize the target and upadate the target visual template in a image-wide range.
- **Problems**
- Lost the target easily and frequently.
- Not consider the local search of target.

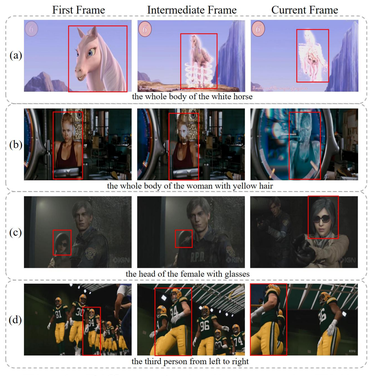
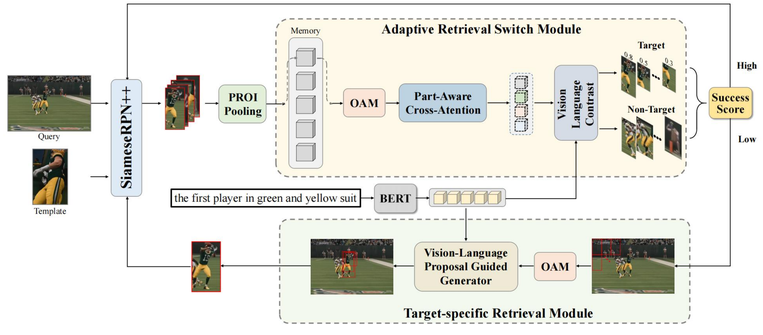


Fig 1. Different challenges in TNL

Our Architecture and Contribution

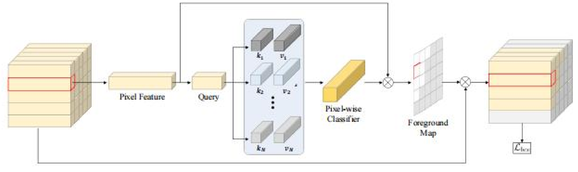


- Propose a novel **adaptive retrieval switch module** equipped with three sub-modules, which are **object-aware attention memory**, **part aware cross-attention**, and **vision-language contrast**, to incorporate local search of the target.
- A **target-specific retrieval module** is developed to precisely capture the tracked object in a global search region. we adopt the sliding window techniques and retrieve the most possible candidates using the proposed vision-language guided proposal selector.

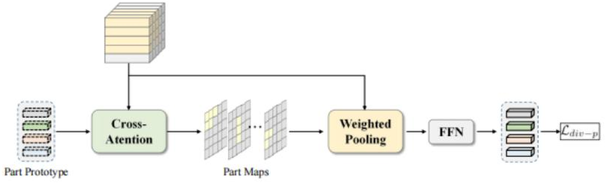
The Details of Our Method

Adaptive Retrieval Switch

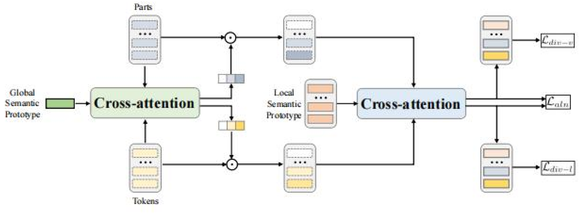
- **Object-aware attention memory(OAM)**: We adopt learnable queries to extract the foreground object from the features of the proposals. The foreground mask is learnen adaptively to indicate the target information.



- **Part-aware cross-attention**: Here we utilize a tiny transformer architecture to obtain the different parts of the potential foreground object from OAM. To achieve this, we introduce the diversity loss as denoted in the following fugure.Then we get target part features.

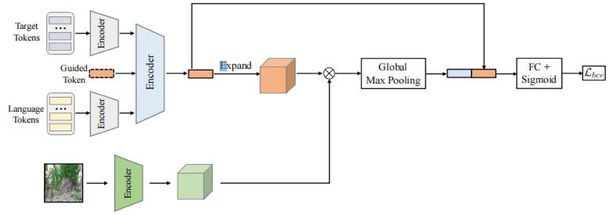


- **Vision-language Contrast:** This sub-module takes as input part features and word features from BERT simultaneously. We set multiple local semantic prototypes and a global prototype to get the relevant semantics from different modalities. Utimately, the module returns confidet scores.



Target-specific Retrieval

- We take the output from vision-language contrast module as indicator of switch to global search.We also adopt the window technique to get proposals of different location. Then we test each proposal with the proposed module as below.



Experiments

Tab 1. Results in different benchmarks.

Method	OTB-Lang				LaSOT				TNL2K			
	NL		NL+BBox		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	-		0.27	0.34 | 0.25
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**

Fig 2. Visualization of our method.