Type-to-Track: Retrieve Any Object via Prompt-based Tracking

Pha Nguyen[1], Kha Gia Quach[2], Kris Kitani[3], Khoa Luu[1]

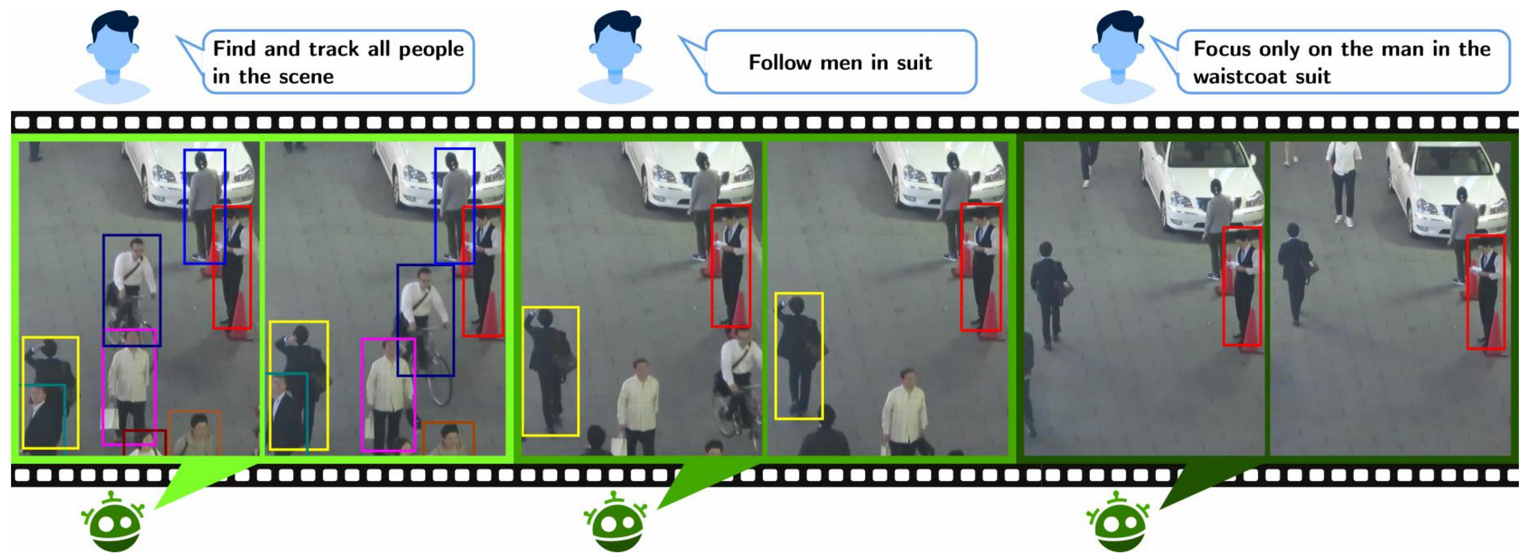
[1]CVIU Lab, University of Arkansas [2]pdActive Inc. [3]Robotics Institute, Carnegie Mellon University

https://uark-cviu.github.io

NEURAL INFORMATION PROCESSING SYSTEMS

NEW ORLEANS 2023

Type-to-Track Paradigm



An example of the responsive *Type-to-Track*. The user provides a video sequence and a prompting request. During tracking, the system is able to track the target subjects and iteratively responds to the request.

- New ***Grounded Multiple Object Tracking*** dataset named GroOT that is more advanced than existing tracking datasets.
- GroOT contains videos with ***various types of multiple objects*** and ***detailed textual descriptions of 256K words***.
- **Five new benchmarking protocols** and ***three new metrics*** for prompt-based visual tracking.
- New framework ***MENDER*** as ***the first efficient approach***.

Visual Object Tracking Benchmarks

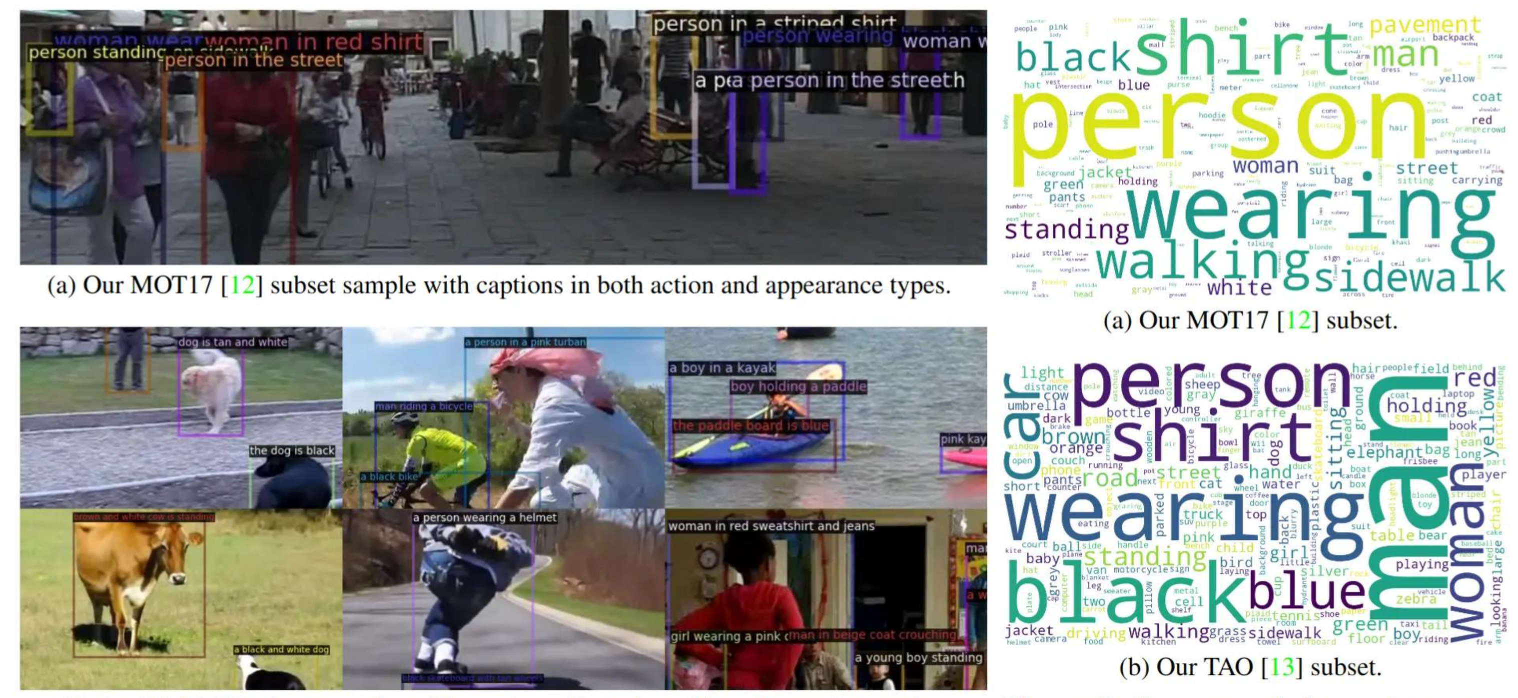
Datasets	Task	NLP	#Videos	#Frames	#Tracks	#AnnBoxes	#Words	#Settings
OTB100 [8]	SOT	✗	100	59K	100	59K	-	-
VOT-2017 [9]	SOT	✗	60	21K	60	21K	-	-
GOT-10k [10]	SOT	✗	10K	1.5M	10K	1.5M	-	-
TrackingNet [11]	SOT	✗	**30K**	**14.43M**	**30K**	**14.43M**	-	-
MOT17 [12]	**MOT**	✗	14	11.2K	1.3K	0.3M	-	-
TAO [13]	**MOT**	✗	1.5K	**2.2M**	8.1K	0.17M	-	-
MOT20 [14]	**MOT**	✗	8	13.41K	3.83K	2.1M	-	-
BDD100K [15]	**MOT**	✗	**2K**	318K	**130.6K**	**3.3M**	-	-
LaSOT [6]	SOT	✓	1.4K	**3.52M**	1.4K	**3.52M**	9.8K	1
TNL2K [7]	SOT	✓	2K	1.24M	2K	1.24M	10.8K	1
Ref-DAVIS [16]	VOS	✓	150	94K	400+	-	10.3K	**2**
Refer-YTVOS [17]	VOS	✓	**4K**	1.24M	**7.4K**	131K	**158K**	**2**
Ref-KITTI [18]	**MOT**	✓	18	6.65K	-	-	3.7K	1
GroOT (Ours)	**MOT**	✓	**1,515**	**2.25M**	**13.3K**	**2.57M**	**256K**	**5**

Comparison of current datasets. # denotes the number of the corresponding item. **Bold** numbers are the best number in each sub-block, while **highlighted** numbers are the best across all sub-blocks.

Most existing datasets and benchmarks for object tracking are ***limited in their coverage and diversity*** of language and visual concepts. Additionally, the prompts in the existing Grounded SOT benchmarks ***do not contain variations in covering many objects in a single prompt***, which limits the application of existing trackers in practical scenarios.

To address this, we present ***a new dataset and benchmarking metrics*** to support the emerging trend of the Grounded MOT, where the goal is to ***align language descriptions with fine-grained regions or objects in videos***.

Dataset Overview



(a) Our MOT17 [12] subset sample with captions in both action and appearance types.

(a) Our MOT17 [12] subset.

(b) Our TAO [13] subset samples with captions. **Best viewed in color and zoom in.**

(b) Our TAO [13] subset.

Figure 3: Some words in our language

Example sequences and annotations in our dataset..

Datasets		#Videos	#Frames	#Tracks	#AnnBoxes	#Words	Parts
MOT17**	Train	7	5,316	546*	112,297*	3,792	(1)
	Test	7	5,919	785*	188,076*	5,757	(2)
	Total	14	11,235	1,331*	300,373*	9,549	
TAO**	Train	500	764,526	2,645	54,639	19,222	(3)
	Val	993	1,460,666	5,485	113,112	39,149	(4)
	Test	914	2,221,846	7,972	164,650	-	
	Total	2,407	4,447,038	16,089	332,401	58,371	
MOT20**	Train	4	8,931	2,332*	1,336,920*	-	(5)
	Test	4	4,479	1,501*	765,465*	-	(6)
	Total	8	13,410	3,833*	2,102,385*	-	
GroOT**	**nm**	1,515	2,249,837	13,294	2,570,509	21,424	*all*
	syn	1,515	2,249,837	13,294	2,570,509	53,540	*all*
	def	1,515	2,249,837	13,294	2,570,509	99,218	*all*
	cap	1,507	2,236,427	9,461	468,124	67,920	*w/o MOT20*
	retr	993	1,460,666	1,952	-	13,935	*uses (4)*

all uses (1, 2, 3, 4, 5, 6) and *w/o MOT20* uses (1, 2, 3, 4).

* Statistics from the official site, including objects other than human.

Statistics of *GroOT*'s settings.

Our proposed Type-to-Track paradigm is distinct in its focus on ***responsively*** and ***conversationally*** tracking ***any objects*** in videos, maintaining the temporal motions of multiple objects of interest.

Class-agnostic Evaluation Metrics

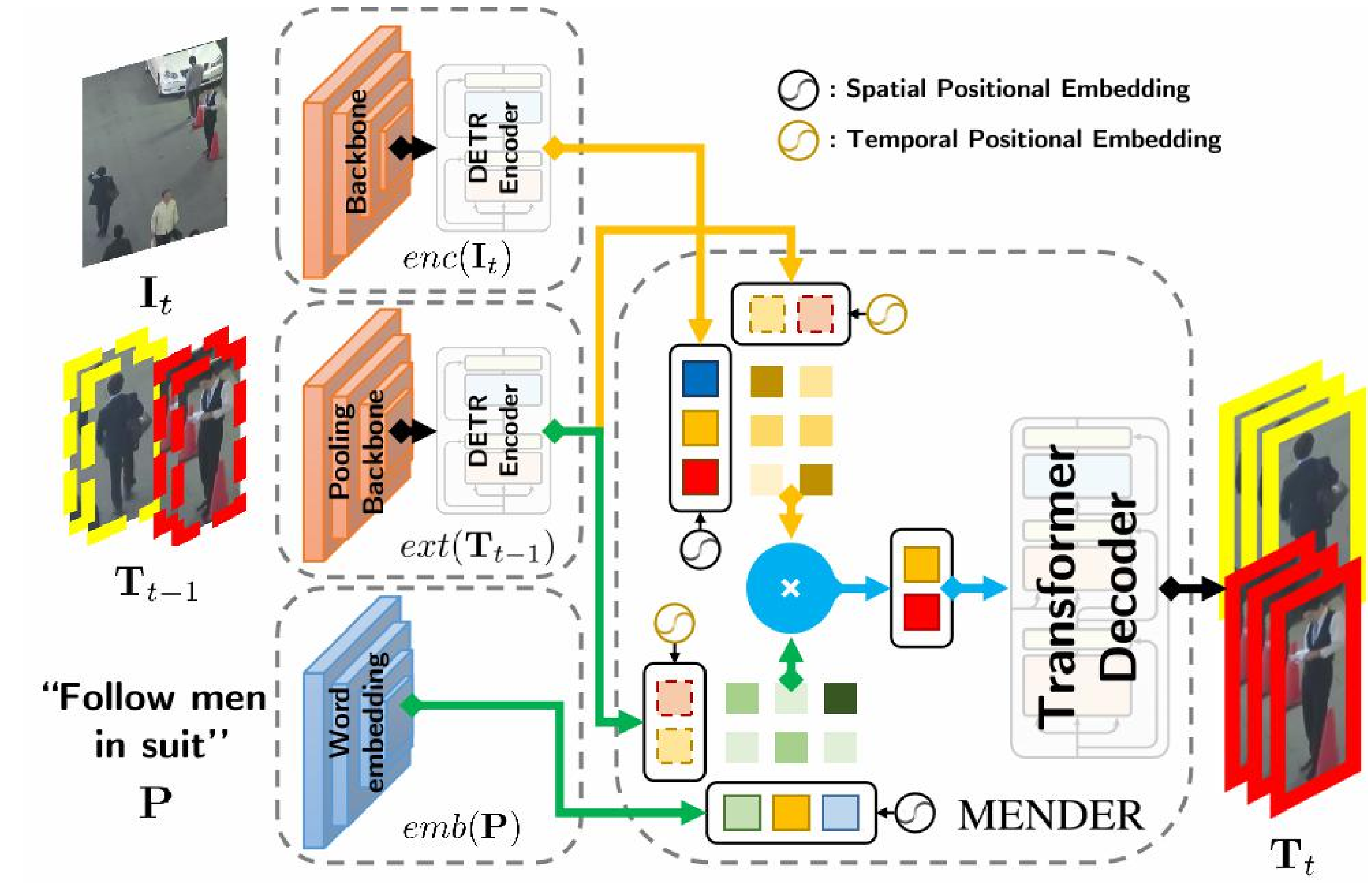
$$\text{MOTA} = \frac{1}{|CLS^n|} \sum_{cls}^{CLS^n} \left(1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDS}_t)}{\sum_t \text{GT}_t}\right)_{cls}, \text{CA-MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDS}_t)_{CLS^1}}{\sum_t (\text{GT}_{CLS^1})_t} \tag{1}$$

$$\text{IDF1} = \frac{1}{|CLS^n|} \sum_{cls}^{CLS^n} \left(\frac{2 \times \text{IDTP}}{2 \times \text{IDTP} + \text{IDFP} + \text{IDFN}}\right)_{cls}, \quad \text{CA-IDF1} = \frac{(2 \times \text{IDTP})_{CLS^1}}{(2 \times \text{IDTP} + \text{IDFP} + \text{IDFN})_{CLS^1}} \tag{2}$$

$$\text{HOTA} = \frac{1}{|CLS^n|} \sum_{cls}^{CLS^n} \left(\sqrt{\text{DetA} \cdot \text{AssA}}\right)_{cls}, \quad \text{CA-HOTA} = \sqrt{(\text{DetA}_{CLS^1}) \cdot (\text{AssA}_{CLS^1})} \tag{3}$$

where CLS^n is the category, set size n is reduced to 1 by combining all elements: $CLS^n \rightarrow CLS^1$.

MENDER for MOT by Prompts



The structure of our proposed *MENDER*. It employs a visual backbone to extract visual features and a word embedding to extract textual features. We model the tracklet-prompt correlation instead of the region-prompt to avoid unnecessary computation caused by no-object tokens.

Quantitative Results

P	sim	CA-MOTA	CA-IDF1	MT	IDs	mAP	FPS
		GroOT - MOT17 Subset					
nm	✗/✓	67.00	71.20	544	1352	0.876	10.3
syn	✗/✓	65.10	71.10	554	1348	0.874	10.3
def	✗	67.00	72.10	556	1343	0.876	5.8
	✓	**67.30**	**72.40**	**568**	**1322**	**0.877**	**10.3**
cap	✗	58.20	53.20	**289**	1751	0.674	3.4
	✓	**59.50**	**54.80**	201	**1734**	**0.688**	**7.8**
		GroOT - TAO Subset					
nm	✓	27.30	37.20	3523	4284	0.212	11.2
syn	✓	25.70	36.10	3212	5048	0.198	11.2
def	✗	15.20	27.30	2452	6253	0.154	6.2
	✓	**16.80**	**27.70**	**2547**	**6118**	**0.158**	**10.5**
cap	✗	20.30	31.80	2943	5242	0.188	4.3
	✓	**20.70**	**32.00**	**3103**	**5192**	**0.184**	**8.7**
retr	✗	32.40	38.40	630	3238	0.423	7.6
	✓	**32.90**	**39.30**	**645**	**3194**	**0.430**	**11.5**
		GroOT - MOT20 Subset					
nm	✗/✓	72.40	67.50	823	2498	0.826	7.6
syn	✗/✓	70.90	65.30	809	2509	0.823	7.6
def	✗	72.90	67.70	**823**	2489	0.826	4.3
	✓	**72.10**	**67.10**	812	**2503**	**0.825**	**7.6**

Approach	CA-MOTA	CA-IDF1	MT	IDs	mAP	FPS
	GroOT - MOT17 Subset					
MDETR + TFm	62.60	64.70	519	1382	0.793	2.2
MENDER	**65.10**	**71.10**	**554**	**1348**	**0.874**	**10.3**
MDETR + TFm	62.60	64.70	519	1382	0.793	2.2
MENDER	**67.30**	**72.40**	**568**	**1322**	**0.877**	**10.3**
MDETR + TFm	44.80	45.20	193	1945	0.619	2.1
MENDER	**59.50**	**54.80**	**201**	**1734**	**0.688**	**7.8**
	GroOT - TAO Subset					
MDETR + TFm	21.30	33.20	2945	5834	0.184	3.1
MENDER	**25.70**	**36.10**	**3212**	**5048**	**0.198**	**11.2**
MDETR + TFm	14.60	21.40	1944	6493	0.137	3.1
MENDER	**16.80**	**27.70**	**2547**	**6118**	**0.158**	**10.5**
MDETR + TFm	15.30	23.60	2132	6354	0.156	3.0
MENDER	**20.70**	**32.00**	**3103**	**5192**	**0.182**	**8.7**
MDETR + TFm	25.70	26.40	513	3993	0.387	3.1
MENDER	**32.90**	**39.30**	**645**	**3194**	**0.430**	**11.5**
	GroOT - MOT20 Subset					
MDETR + TFm	61.20	60.40	784	2824	0.732	1.9
MENDER	**70.90**	**65.30**	**809**	**2509**	**0.823**	**7.6**
MDETR + TFm	68.00	66.30	763	2975	0.783	1.9
MENDER	**72.10**	**67.10**	**812**	**2503**	**0.825**	**7.6**

References

Dongming Wu et al. Referring multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023

Jonathon Luiten et al. Hota: A higher order metric for evaluating multi-object tracking. International journal of Computer Vision. 2021

Aishwarya Kamath et al. Mdetr-modulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021

Tim Meinhardt et al. Trackformer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022

Project page: