NEURAL INFORMATION PROCESSING SYSTEMS

WebUOT-1M: Advancing Deep Underwater Object Tracking with A Million-Scale Benchmark

Chunhui Zhang, Li Liu, Guanjie Huang, Hao Wen, Xi Zhou, Yanfeng Wang

chunhui.zhang@sjtu.edu.cn

Presenter: Chunhui Zhang

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

香港科技大学（广州） THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

云从科技 CLOUDWALK

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Background

- Underwater object tracking (UOT) refers to the task of sequentially locating a submerged instance in an underwater video, given its initial position in the first frame.
- The underwater environment usually exhibits uneven lighting conditions, low visibility, low contrast, watercolor variations, similar distractors, camouflage, etc. posing distinct challenges for UOT compared to traditional open-air tracking tasks.

Motivations

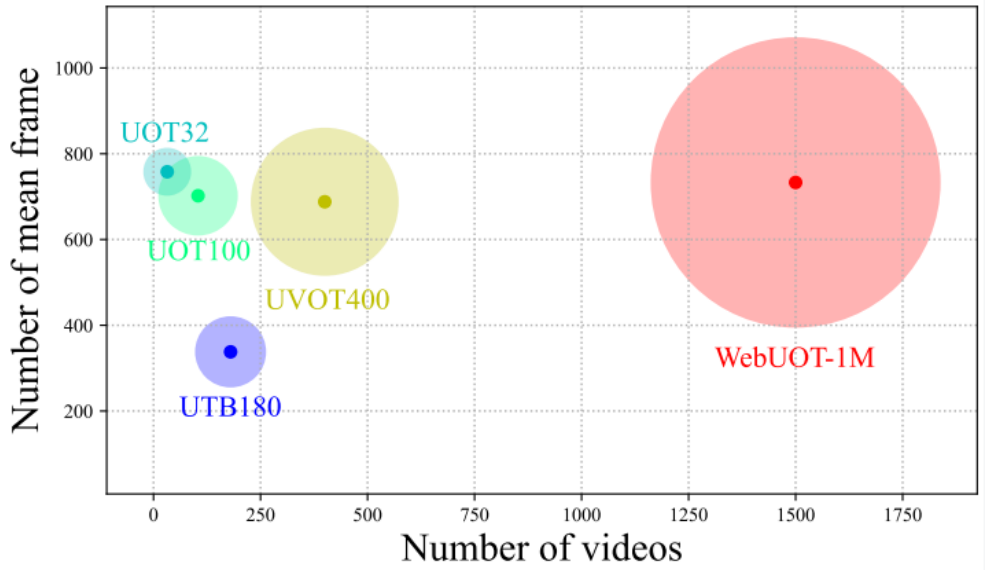
- UOT has not been thoroughly explored due to the absence of large-scale datasets, benchmarks, and challenges in gathering abundant underwater videos.

- Due to the huge appearance variation and behavioral differences among various marine animals, models trained on small-scale datasets[1-4] struggle with unseen species, leading to poor generalization performance.

[1] Landry Kezebou, et al. Underwater object tracking benchmark and dataset. HST 2019
[2] Karen Panetta, et al. Comprehensive underwater object tracking benchmark dataset and underwater image enhancement with GAN. JOE 2021
[3] Basit Alawode,et al. UTB180: A high-quality benchmark for underwater tracking. ACCV 2022
[4] Levi Cai, et al. Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. IJCV 2023

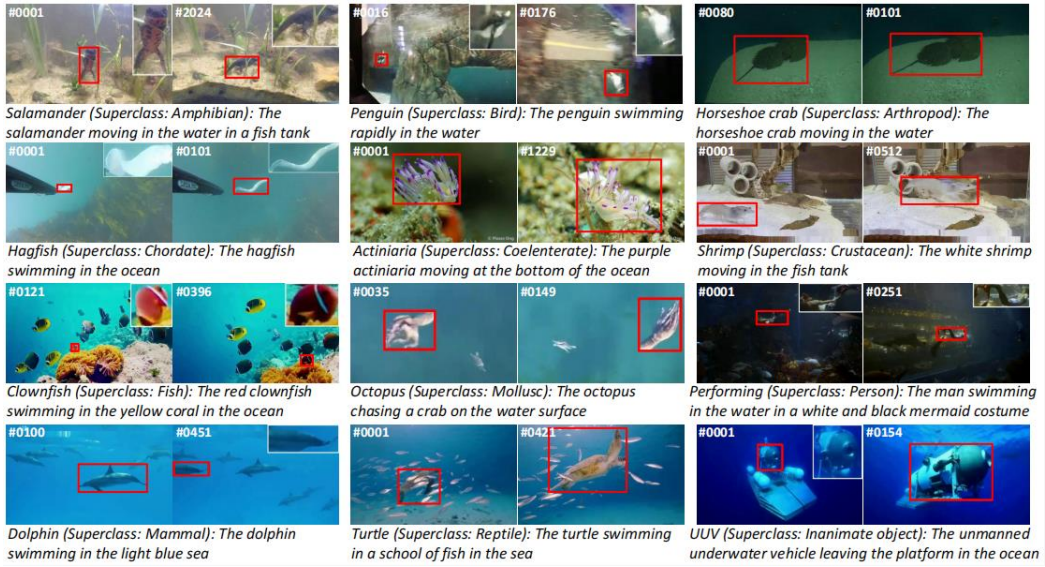
Motivations

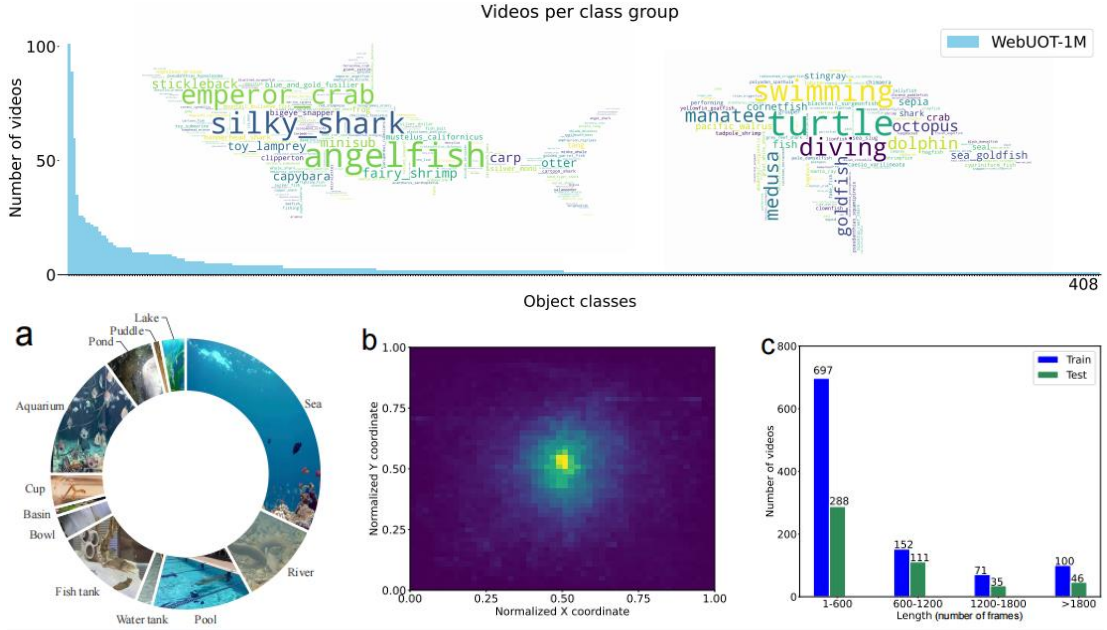
- **WebUOT-1M**: The First Million-Scale UOT Benchmark



- **OKTrack**: A framework to effectively transfer open-air domain knowledge to the UOT model through knowledge distillation

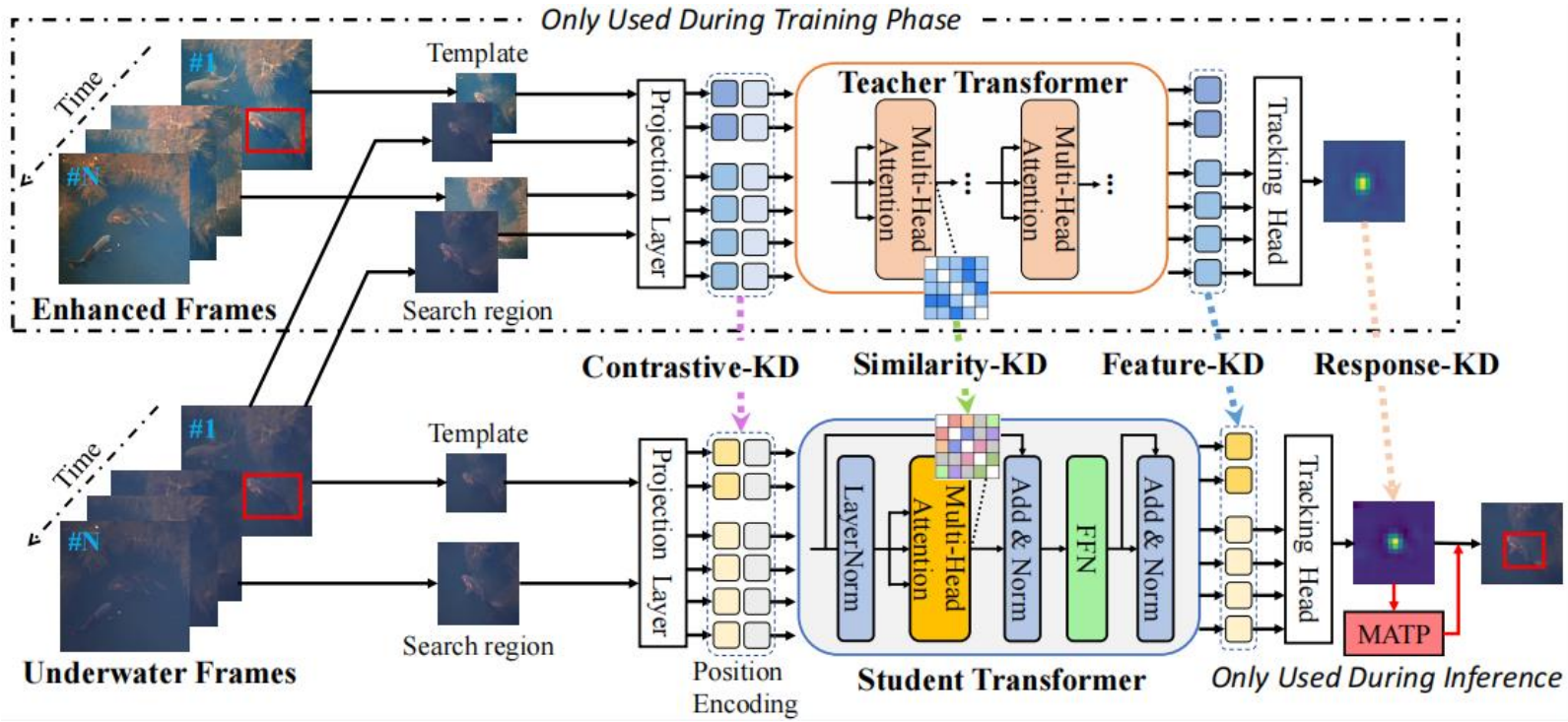
WebUOT-1M: The First Million-Scale UOT Benchmark





Dataset	Year	Videos	Classes	Attributes	Min frame	Mean frame	Max frame	Total frames	Annotated boxes	Total duration	Absent label	Language prompt	Data partition	Open source
UOT32 [36]	2019	32	-	-	283	758	1,573	24 K	24 K	16 min	✗	✗	Test	Proprietary
UOT100 [51]	2022	104	-	3	264	702	1,764	74 K	74 K	41 min	✗	✗	Test	Fully
UTB180 [2]	2022	180	-	10	40	338	1,226	58 K	58 K	32 min	✗	✗	Train/Test	Fully
VMAT [5]	2023	33	17	13	438	2,242	5,550	74 K	74 K	41 min	✗	✗	Test	Fully
UVOT400 [1]	2023	400	50	17	40	688	3,273	275 K	275 K	2.6 hours	✗	✗	Train/Test	Partially
WebUOT-1M	2024	1,500	408	23	49	733	9,985	1.1 M	1.1 M	10.5 hours	✓	✓	Train/Test	Fully

OKTrack: Omni-Knowledge Distillation Framework

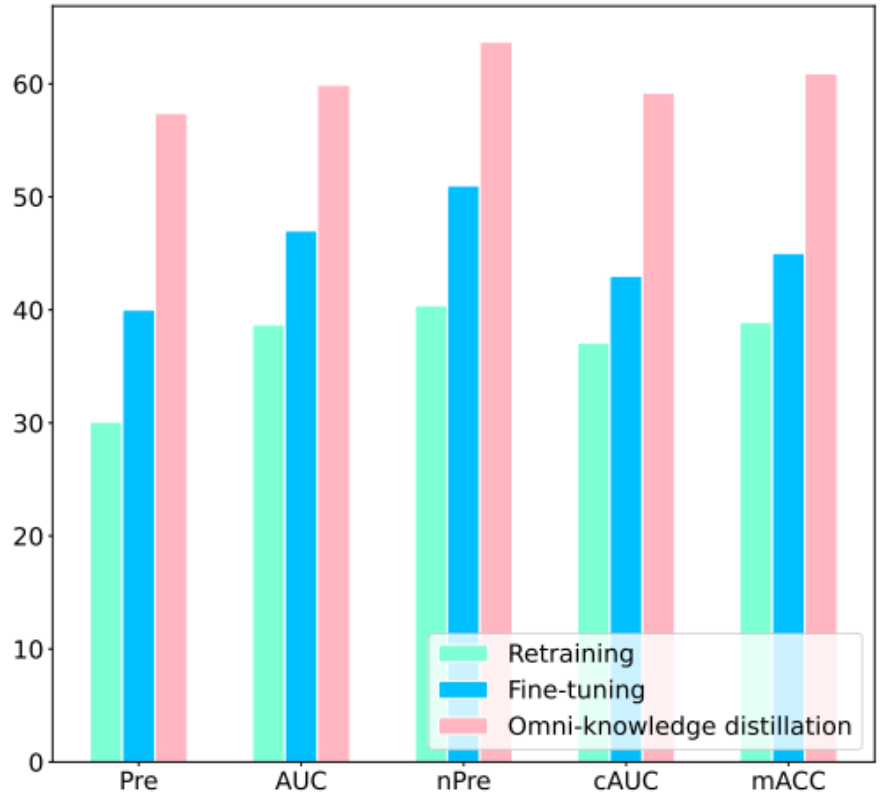


- The omni-knowledge distillation contains token contrastive representation, similarity matrix, feature embeddings, and response maps distillation losses for transferring open-air domain knowledge to underwater domain.
- A training-free motion-aware target prediction (MATP) to address model drift.

Experiments and Results

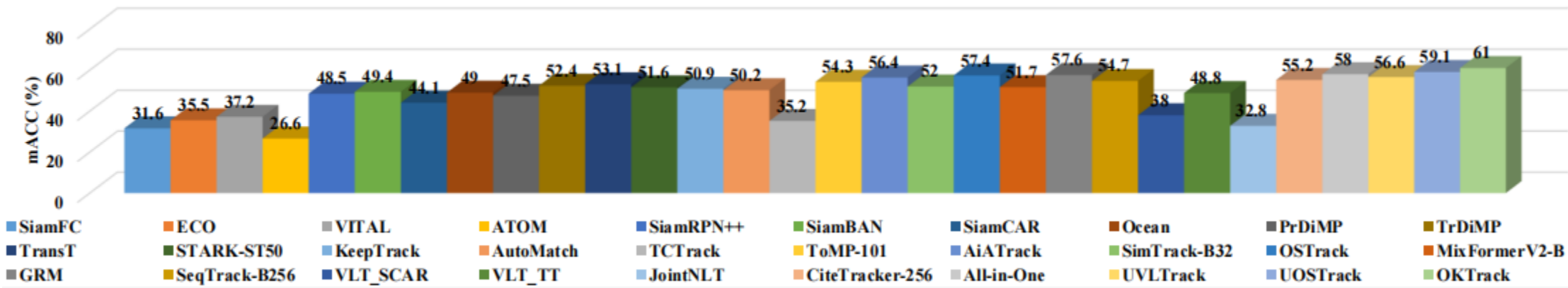
☐ Ablation Study

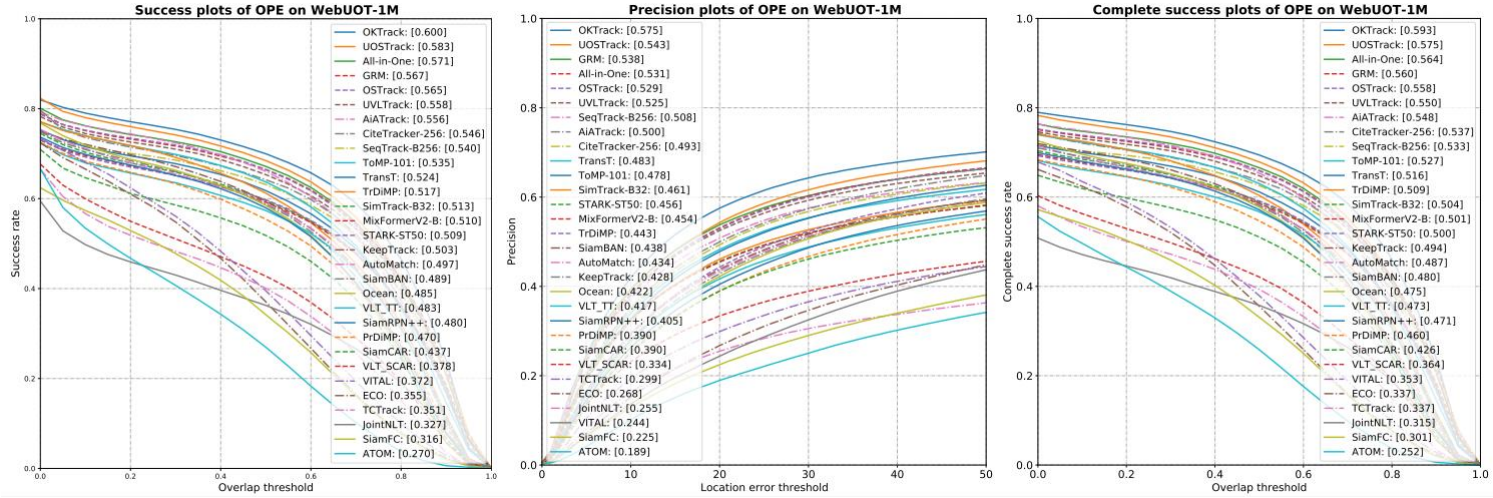
Base	CKD	SKD	FKD	RKD	MATP	UTB180	WebUOT-1M
✓						62.3/66.6	52.0/56.5
✓	✓					63.6/67.9	53.9/57.8
✓		✓				63.2/67.4	53.5/57.2
✓			✓			63.2/66.9	53.2/57.2
✓				✓		65.0/68.1	54.8/57.9
✓	✓	✓	✓	✓		65.4/68.3	55.2/58.3
✓	✓	✓	✓	✓	✓	66.0/68.5	56.1/58.9



Experiments and Results

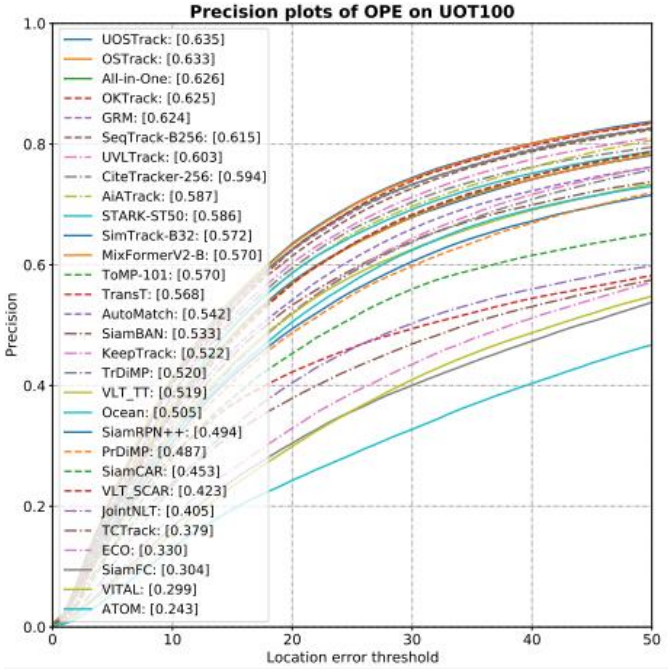
Results on WebUOT-1M

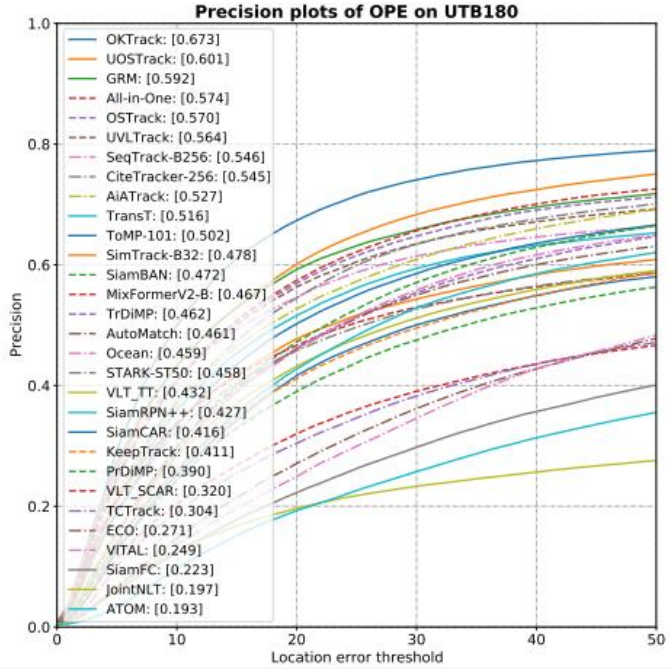


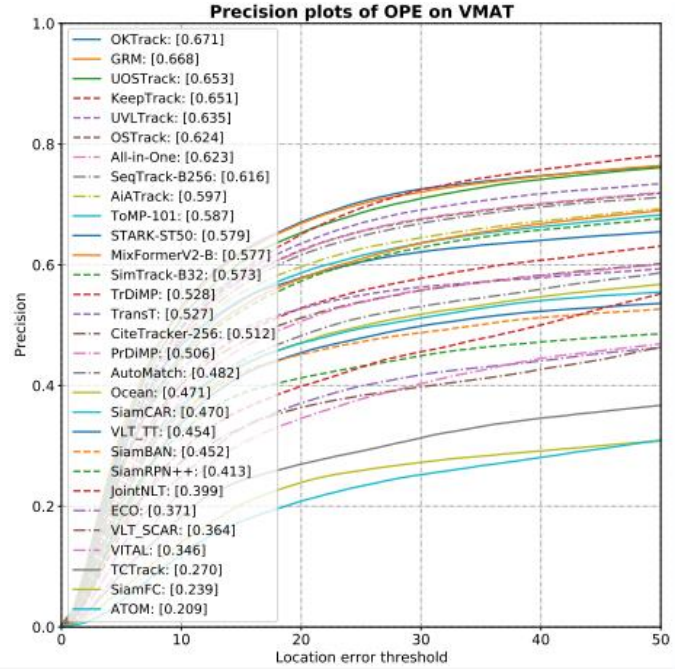


Experiments and Results

- Results on existing UOT benchmarks (UOT100, UTB180, VMAT)







Experiments and Results

- Vision-language tracking: The usage of more cues (e.g., language prompt and bounding box) can significantly boost tracking performance.

- Integrating language modality: OKTrack is a flexible and scalable baseline tracker that is not only suitable for pure visual-based UOT but can also be seamlessly extended to underwater VL tracking.

Method	Pre	nPre	AUC	cAUC	mACC
Language prompt					
JointNLT [86]	22.4	32.2	31.2	29.8	31.2
UVLTrack [44]	22.5	33.8	31.2	30.1	31.3
Language prompt + bounding box					
JointNLT [86]	25.5	34.9	32.7	31.5	32.8
VLT_{SCAR} [28]	33.4	44.0	37.8	36.4	38.0
VLT_{TT} [28]	41.7	52.1	48.3	47.3	48.8
CiteTracker-256 [40]	49.3	58.4	54.6	53.7	55.2
UVLTrack [44]	52.5	60.0	55.8	55.0	56.6
All-in-One [79]	53.1	61.5	57.1	56.4	58.0

Method	Type	#Params	FLOPs	FPS	WebUOT-1M
OKTrack	Visual-based	92.1 M	21.5 G	115	60.0/57.5/59.3/63.8/61.0
OKTrack++	VL-based	150.9 M	57.9 G	66	63.4/58.4/62.9/68.5/64.4

Conclusion

- We introduce WebUOT-1M, the first million-scale benchmark dataset featuring diverse underwater video sequences, essential for offering a dedicated platform for the development and evaluation of UOT algorithms.

- We propose a simple yet strong omni-knowledge distillation tracking approach, termed OKTrack. It is the first work to explore knowledge transfer from a teacher Transformer using underwater and enhanced frames to a student Transformer in the UOT area.

- We comprehensively benchmark the proposed approach, along with 30 trackers based on CNN, CNN-Transformer, and Transformer on both the newly proposed WebUOT-1M and existing UOT datasets.

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS

for listening

https://github.com/983632847/Awesome-Multimodal-Object-Tracking