



NEURAL INFORMATION
PROCESSING SYSTEMS

2025



Collaborating Vision, Depth, and Thermal Signals for Multi-Modal Tracking: Dataset and Algorithm

**Xue-Feng Zhu¹, Tianyang Xu¹, Yifan Pan¹, Jinjie Gu¹,
Xi Li², Jiwen Lu³, Xiao-Jun Wu¹, Josef Kittler⁴**

¹ Jiangnan University

² Zhejiang University

³ Tsinghua University

⁴ University of Surrey



江南大学
JIANGNAN UNIVERSITY



浙江大学
Zhejiang University



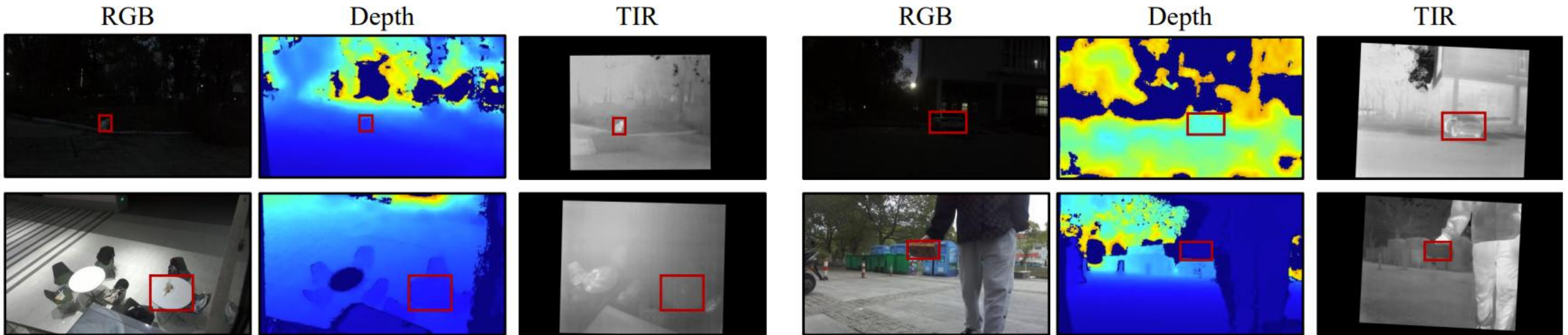
清华大学
Tsinghua University



UNIVERSITY OF
SURREY

● Motivation

- Single-modal or Dual-modal trackers (RGB-D / RGB-T) fail in complex real-world scenes.
- Need for tri-modal fusion (RGB + Depth + Thermal) to boost robustness.



● RGBDT500 Dataset

- **500 videos** with synchronized **RGB, Depth, and Thermal** frames.
- **400 train / 100 test** sequences, **66 object classes**, total **203K frames (1080p)**.
- **Precise spatial–temporal alignment** and **bounding box annotations**.
- Publicly available: <https://xuefeng-zhu5.github.io/RGBDT500/>

Dataset	Modalities			Sequences	Frames	Classes	Training Set	Publication
	RGB	Depth	TIR					
UAV123 [1]	✓	✗	✗	123	113K	-	✗	ECCV 2016
GOT-10k [14]	✓	✗	✗	10K	1.5M	563	✓	IEEE TPAMI 2019
LaSOT [9]	✓	✗	✗	1400	3.5M	70	✓	CVPR 2019
GTOT [19]	✓	✗	✓	50	15.8K	9	✗	IEEE TIP 2016
RGBT210 [22]	✓	✗	✓	210	104.7K	22	✗	ACM MM 2017
RGBT234 [20]	✓	✗	✓	234	116.7K	22	✗	PR 2019
LasHeR [21]	✓	✗	✓	1224	734.8K	32	✓	IEEE TIP 2021
VTUAV [51]	✓	✗	✓	500	1.7M	13	✓	CVPR 2022
PTB [36]	✓	✓	✗	100	21K	26	✗	CVPR 2013
CDTB [30]	✓	✓	✗	80	101.9K	21	✗	ICCV 2019
DepthTrack [48]	✓	✓	✗	200	294.5K	90	✓	ICCV 2021
ARKitTrack [52]	✓	✓	✗	455	229.7K	144	✓	CVPR 2023
RGBDT500	✓	✓	✓	500	203.7K	66	✓	NeurIPS 2025

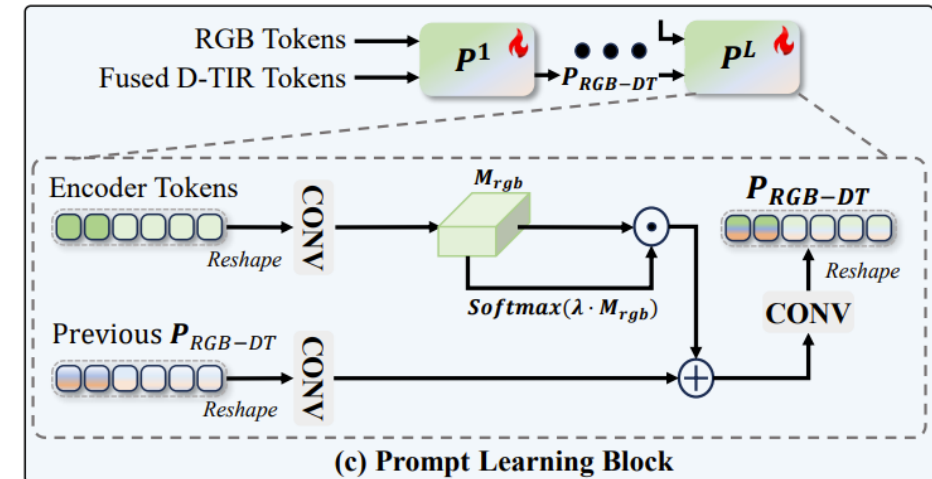
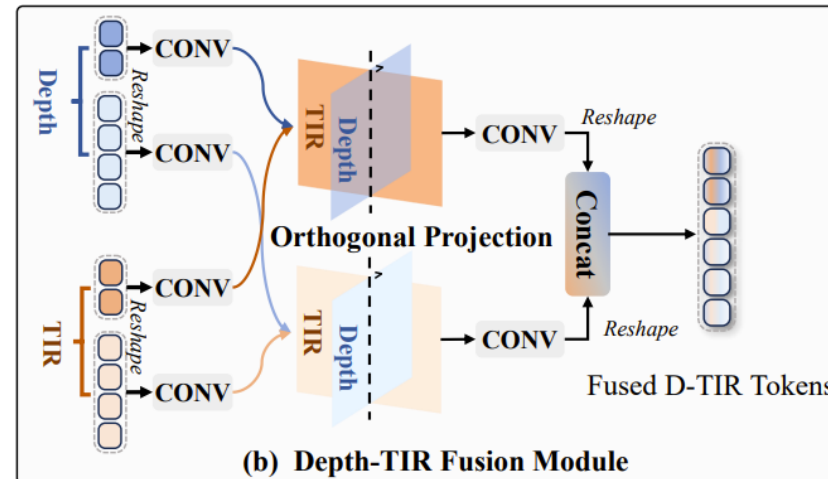
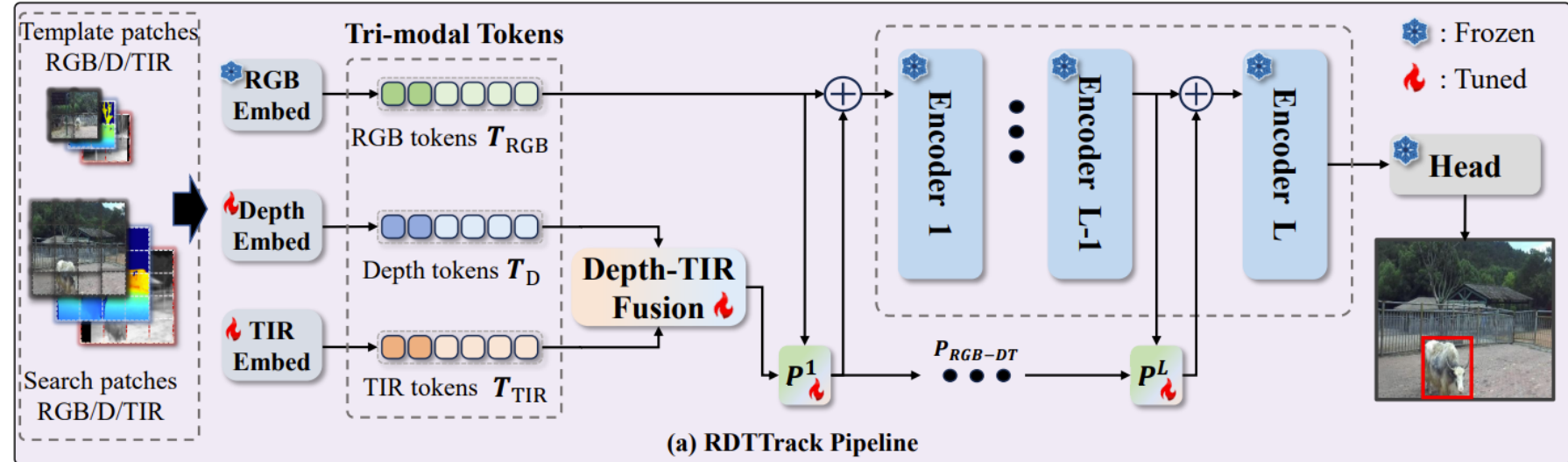


● Baseline: RDTTrack

Key Idea: Leverage pre-trained RGB tracker + Prompt Learning for tri-modal input.

Components:

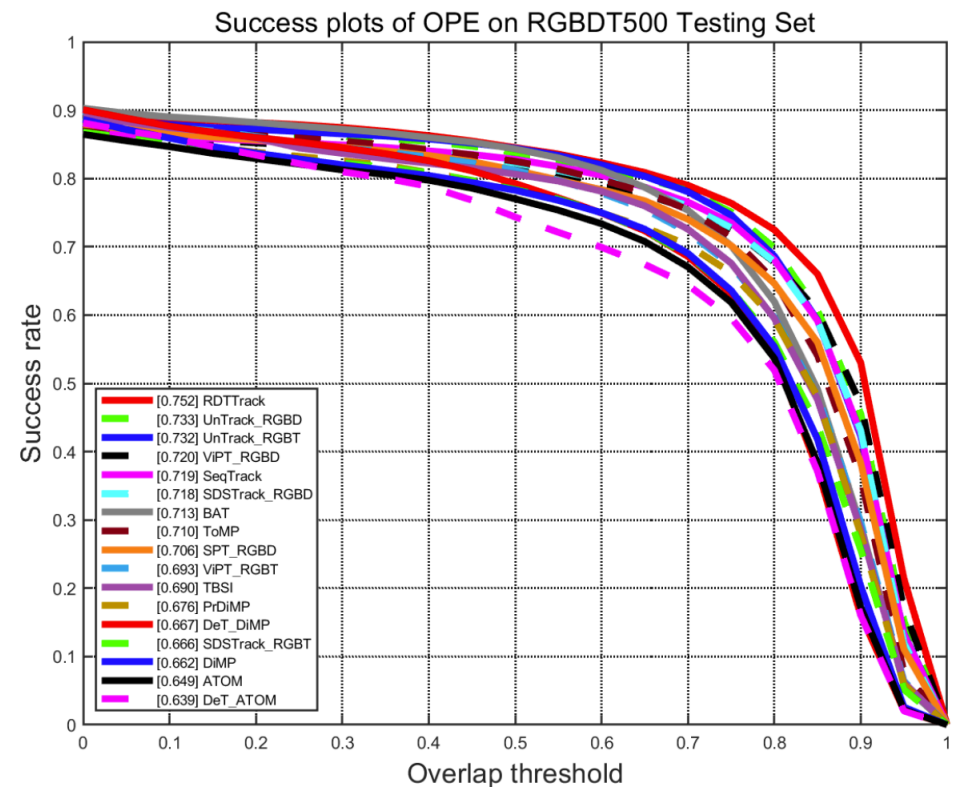
- **Depth-TIR Fusion Module:** orthogonal projection $\rightarrow$ remove redundancy.
- **Prompt Learning Block:** inject fused depth-thermal info into RGB stream.
- Frozen RGB model, only train fusion & prompt modules.



● Experimental Results

Benchmarking experiments:

Tracker	Input Modality	AUC	DP	Publication Year
ATOM [6]	RGB	0.649	0.695	2019
DiMP [2]	RGB	0.662	0.710	2019
PrDiMP [7]	RGB	0.676	0.698	2020
ToMP [31]	RGB	0.710	0.747	2022
OTrack [49]	RGB	0.694	0.736	2022
SeqTrack [4]	RGB	0.719	0.766	2023
MixFormer [5]	RGB	0.732	0.781	2024
DeT_ATOM [48]	RGB+D	0.639	0.665	2021
DeT_DiMP [48]	RGB+D	0.667	0.700	2021
SPT [57]	RGB+D	0.706	0.757	2023
ViPT_RGBD [53]	RGB+D	0.720	0.759	2023
SDSTrack_RGBD [13]	RGB+D	0.718	0.763	2024
UnTrack_RGBD [42]	RGB+D	0.733	0.776	2024
TBSI [15]	RGB+T	0.690	0.749	2023
ViPT_RGBT [53]	RGB+T	0.693	0.752	2023
SDSTrack_RGBT [13]	RGB+T	0.666	0.708	2024
BAT [3]	RGB+T	0.713	0.782	2024
UnTrack_RGBT [42]	RGB+T	0.732	0.790	2024
DCEvo+OTrack [29]	RGB+T	0.706	0.741	2025
RDTTrack(Ours)	RGB+D+T	0.752	0.792	2025



Tracking Efficiency:

Tracker	RDTTrack	BAT	TBSI	SDSTrack	Un-Track
FPS	76.6	26.8	32.5	20.9	41.2

● Ablation & Insights

Ablation Studies:

- Without Depth–TIR Fusion $\rightarrow$ $\downarrow$ AUC by 1.9%.
- Without Orthogonal Projection $\rightarrow$ $\downarrow$ AUC by 2.0%.
- Tri-modal > any dual-modal variant.

Module ablation		
Tracker	AUC	DP
w/o D-TIR OP	0.733	0.773
w/o α & β	0.739	0.779
RDTTrack	0.752	0.792

Modal ablation			
Tracker	Modality of Training Set	AUC	DP
STARK [47]	RGB	0.692	0.732
SPT_RGBD [57]	RGB+Depth	0.706	0.757
SPT_RGBT [57]	RGB+Thermal	0.730	0.784
OSTrack [49]	RGB	0.694	0.736
ViPT_RGBD [53]	RGB+Depth	0.719	0.762
ViPT_RGBT [53]	RGB+Thermal	0.729	0.768
RDTTrack	RGB+Depth	0.737	0.776
RDTTrack	RGB+Thermal	0.734	0.774
RDTTrack	RGB+Depth+Thermal	0.752	0.792

Insights:

- Orthogonal fusion avoids redundancy between depth and thermal cues.
- Prompt learning efficiently adapts pre-trained models to new modalities.
- Tri-modal integration yields best robustness and accuracy.

● Conclusion

Contributions: <https://xuefeng-zhu5.github.io/RGBDT500/>

- **First Tri-modal tracking dataset:** RGBDT500.
- **Novel Tri-modal tracker:** RDTTrack.
- **Achieves state-of-the-art accuracy and high speed.**
- **Promotes general multi-modal fusion** for robust real-world tracking.

Future Work:

- **Flexible and Input-Agnostic Fusion:** Build an adaptive model for variable sensor inputs
- **Dataset Expansion:** Add more modalities to RGBDT500, like event or LiDAR

THANKS!