Generative-based Fusion Mechanism for Multi-Modal Tracking Supplementary Material

Zhangyong Tang[1], Tianyang Xu[1], Xiaojun Wu[1]*, Xue-Feng Zhu[1], Josef Kittler[2]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu, PR. China
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
zhangyong_tang_jnu@163.com; {tianyang.xu;wu_xiaojun}@jiangnan.edu.cn; j.kittler@surrey.ac.uk

Introduction

To deliver a precise introduction of the proposed generative-based fusion mechanism for multi-modal tracking (GMMT), some of the details are exhibited in this supplementary material.

- **A.** The introduction of the discriminator, D, in the conditional generative adversarial network (CGAN).
- **B.** Benchmarks and metrics.
- **C.** More details when implementing GMMT.
- **D.** Intuitive comparison between the original fusion block and the proposed GMMT.
- **E.** More visualisation of the tracking results.
- **F.** The quantitative results produced during the analysis of n, the number blocks in the U-shape network.
- **G.** The quantitative results on GTOT.
- **H.** The quantitative results produced by multiple baseline trackers and benchmarks, as well as the comparison between UNet (Song, Meng, and Ermon 2020) and UViT (Bao et al. 2023) architectures.
- **I.** The quantitative results produced during the analysis of s, the number of steps the reverse diffusion process takes.
- **J.** Analysis of λ, which is a balance between the original tracking loss and the generative loss.

A. Discriminator in CGAN

Since the generator G kept the same for all the variants, it remains in the next section and only the discriminator D is introduced in this part, which is shown in Fig. 1. D receives the output of G and four convolutional blocks are embedded for dimension reduction. Later, a sigmoid activation is employed to transfer the output to the interval [0,1]. The output serves the possibility to be a real sample. Each convolutional block contains a convolutional layer, a batch normalisation layer, and a ReLu activation. Notably, there is no activation in the last convolution block.

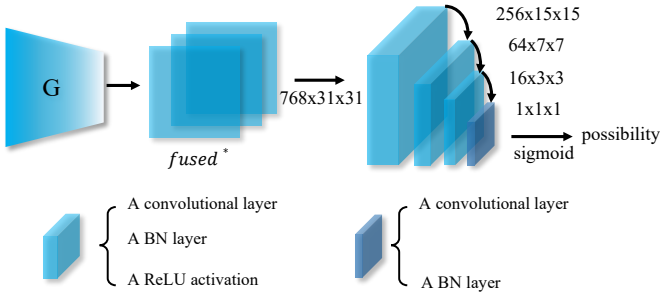


Figure 1: Architecture of the discriminator embedded in our CGAN-based GMMT.

B. Benchmark and Metrics

The effectiveness of GMMT is verified on GTOT (Li et al. 2016), RGBT234 (Li et al. 2019), LasHeR (Li et al. 2022), and RGBD1K (Zhu et al. 2023b).

GTOT is a pioneering RGB-T dataset, including 7.8K image pairs. The evaluation metrics are precision rate (PR) and success rate (SR). PR measures the percentage of frames with the distance between centres of the predicted and ground truth bounding box below a threshold, 5 in this benchmark. SR represents the ratio of frames being tracked with the overlap between the predicted and ground truth bounding box above zero.

RGBT234 is an extended version of GTOT by involving some videos in some special scenarios, like the summer days. It includes 234 multimodal video pairs and employs the same metrics as GTOT do.

LasHeR is a larger and widely-used benchmark in the RGB-T tracking field, and its testing split consists of 245 video pairs. PR, SR and the normalised precision rate (NPR) are used for benchmarking. NPR (Muller et al. 2018) is a modified version of PR since PR can be easily affected the image resolution and the size of the ground truth bounding box.

RGBD1K is the largest benchmark in the RGB-D tracking field. It contains around 50 videos in the test set, with most of them have a length of 3000 frames. Since the objects in this benchmark may disappear and reappear, the metric named Recall is employed to calculate the rate of the object being successfully tracked. The definition of being success-

*Corresponding author

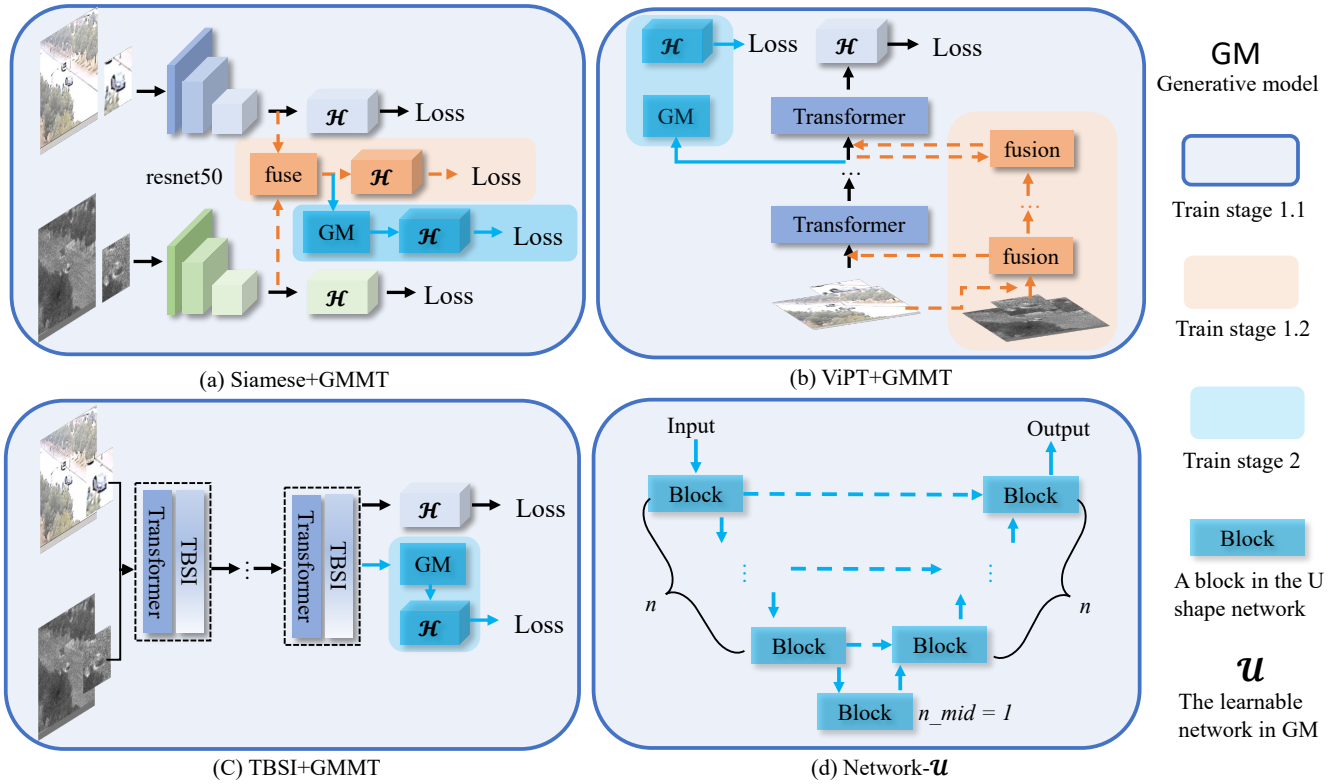


Figure 2: The implementation of GMMT on multiple baseline trackers (a)Siamese, (b)ViPT, (c)TBSI, and (d) the architecture of the learnable network $\mathcal{U}$ in GMMT. We follow the official implementation of ViPT and TBSI while the Siamese baseline is a self-designed one.

fully tracked means the overlap between the predicted and ground truth is above zero. Precision is involved for measuring how accuracy the predicted bounding boxes are. Later, F-score, a comprehensive metric, is further introduced by taking both PR and RE into account.

C. Implementing Details

The experimental details are introduced in this part. Basically, the effectiveness of GMMT is demonstrated on three baseline methods, *i.e.*, a self-designed Siamese tracker, the ViPT (Zhu et al. 2023a) and TBSI (Hui et al. 2023), which are sequentially introduced in the following paragraphs.

The first selected baseline is a Siamese tracker constructed based on the advanced RGB tracker SiamBAN (Chen et al. 2020). SiamBAN consists of a ResNet50 (He et al. 2016) backbone, a Neck block, and three tracking heads. In our design, only the output of the third residual layer of ResNet50 is maintained for efficiency, and the number of tracking heads is also reduced to 1. As a multi-modal tracker, the above baseline is duplicated once, which is further applied to the TIR data. The overall architecture is shown in Fig. 2(a). As Fig. 2(a) shows, the SiamBAN baseline is trained at first. Later, the fusion block (a convolutional layer) as well as its corresponding tracking head are optimised with other parts frozen. Our generative model (GM) is embedded in our GMMT and a tracking head is trained at last.

The second baseline tracker is ViPT, as shown in Fig. 2b. Basically, an offline-trained model is provided by the authors. However, it performs 51.9 on the SR while 52.5 in the published manuscript, which leads to the retraining of the fusion blocks (also termed prompt layers). After that, our GMMT and the extra tracking head are appended and optimised.

With the least modification, the application on the third baseline, TBSI, is illustrated in Fig. 2(c). The training procedure of GMMT is activated after the parameters of the official-provided model are frozen.

Notably, our GMMT is applied to the features from search image patches, and almost all the crucial configurations during the training procedure are displayed in Table. 2.

After the training of GMMT, it is then evaluated on the multi-modal benchmarks. To present the difference between the typical fusion method and GMMT, the pseudo-code is programmed in Table. 1. Effortlessly, the GMMT can be distinguished from the typical fusion methods. Firstly, the original discriminative-based fusion block is abandoned and, instead, a generative-based fusion block is activated. Secondly, our GMMT can be executed iteratively, which property is unseen in the existing fusion methods.

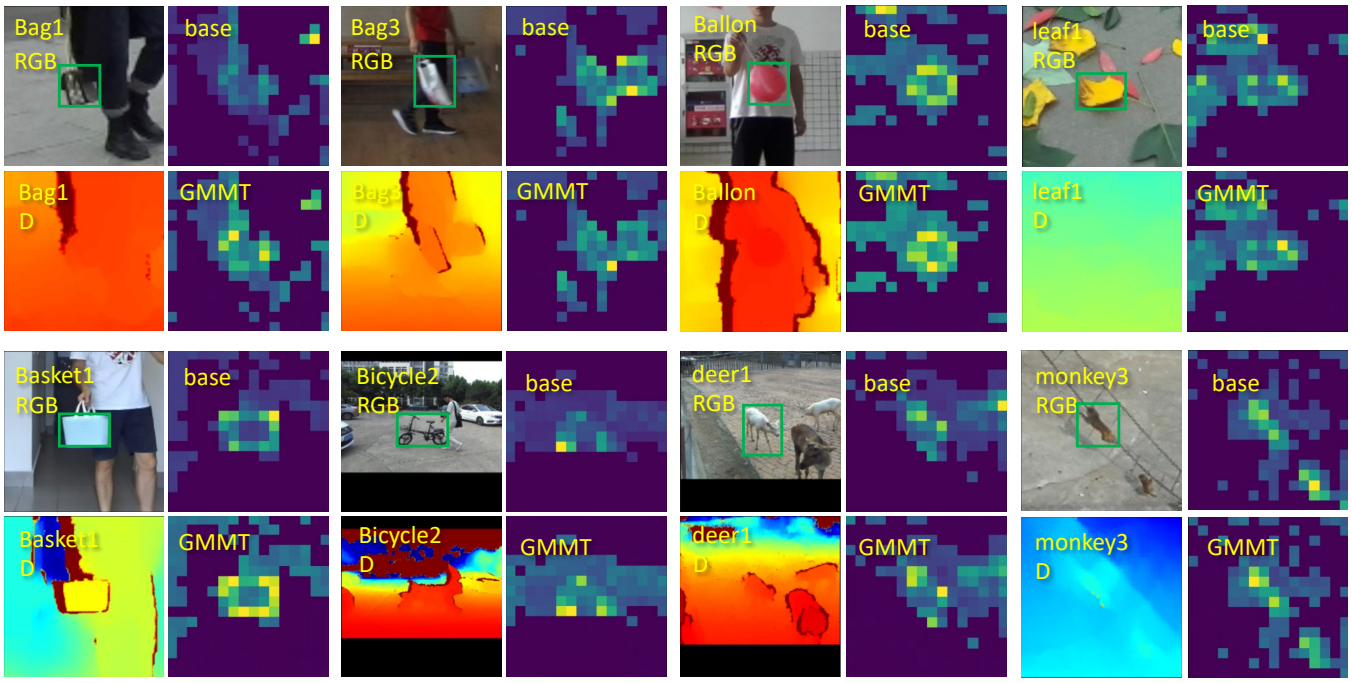
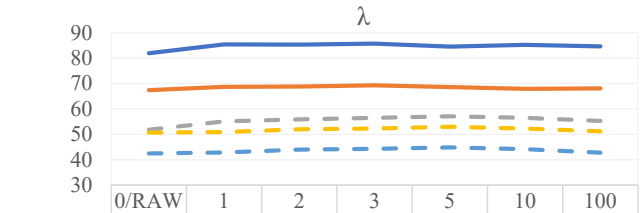


Figure 3: Visualisation of the feature embeddings on the challenging RGBD1K benchmark.

	Pseudo code when testing
Input:	Current features from each modality $f_{rgb,t}$, $f_{tir,t}$, Noise z,Tracking Head $\mathcal{H}$, Generator G Feature extractor and fusion block $\mathcal{F}$
Track:	IF Type == 'Typical': $fused = \mathcal{F}(f_{rgb}, f_{tir})$ ELIF Type == 'GMMT': while iteration $fused^* = \mathcal{G}(z, f_{rgb}, f_{tir})$ $z = fused^*$ $fused = fused^*$
Output:	Prediction of current frame $\mathcal{P}$=$\mathcal{H}(fused)$

Table 1: Pseudo code of GMMT.



	0/RAW	1	2	3	5	10	100
PR	81.9	85.4	85.3	85.7	84.5	85.2	84.6
SR	67.4	68.7	68.8	69.3	68.6	67.9	68.1
PR	51.8	55.1	55.9	56.5	57.1	56.5	55.3
NPR	50.7	50.9	52.0	52.3	53.0	52.3	51.2
SR	42.5	42.9	44.0	44.3	44.9	44.2	42.8

Figure 4: Analysis of λ.

D. Intuitive Comparison for GMMT

To exhibit the superiority of the proposed GMMT, the comparison between the features with and without GMMT is intuitively provided in Fig. 3, with all examples sampled from the RGBD1K benchmark. In the instances sampled from $Basket1$, $Bicycle2$, the response in the target regions are significantly enhanced. In other samples, the highest responses appear in the incorrect places, while they are corrected with the equipment of our GMMT. This indicates that the features generated by GMMT have better discrimination, as well as a higher possibility to predict compact bounding boxes.

E. Visualisation of Tracking Results

More intuitive tracking results are provided in Fig. 5. Under more challenging scenarios, the TBSI+GMMT and ViPT+GMMT still perform significantly better than the baseline methods, TBSI and ViPT, respectively, which further demonstrates the superiority of the proposed GMMT.

F. Number of Blocks

As shown in Fig. 2(d), the depth of the embedding U-shape network is determined by a hyper-parameter n, which represents the number of blocks. The exploration of network depth is a long-lasting topic in the deep learning era. Therefore, the value of n is investigated in this part, and Table. 3 gives the quantitative results on RGBD1K, with both the UNet (Song, Meng, and Ermon 2020) and UViT (Bao et al.

Configurations	Stage 1.1	Stage 1.2	Stage 2
Batchsize	8		
	-	32	
	-	-	32
Epoch	20		100
	-	100	
	-	-	100
Learnable blocks	Backbone,Neck, RPN	Fusion block	GMMT
	-	Promt layers	GMMT
	-	-	GMMT
Optimiser	SGD	SGD	SGD
	-	ADAMW	SGD
	-	-	SGD
Base LR	0.005		
	-	0.0004	0.005
	-	-	0.005
Learning rate	1∼5:0.001-0.005, 6∼20:0.005-0.00005	1∼5:0.001-0.005, 6∼20:0.005-0.00005	1∼20:0.001-0.005, 21∼100:0.005-0.00005
	-	-	1∼20:0.001-0.005, 21∼100:0.005-0.00005
	-	-	1∼20:0.001-0.005, 21∼100:0.005-0.00005
Weight decay	0.0001		
	-	0.0001	
	-	-	0.0001
Momentum	0.9		
	-	No	0.9
	-	-	0.9
T	-	-	1000
Samper	-	-	DDIM

Table 2: Experimental configurations on three baseline methods during the multi-stage training scheme. For each hyper-parameter, there are three rows, which are held for the Siamese baseline, ViPT, and TBSI, sequentially.

UViT n	1	2	3	4	5	6	7	8	9
PR↑	51.8	52.5	52.8	53.8	53.3	53.5	53.2	55.9	OoM
RE↑	54.6	55.4	55.7	57.0	56.3	56.5	56.4	59.0	OoM
F-score↑	53.2	53.9	54.2	55.4	54.8	54.9	54.7	57.4	OoM
UNet n	3		6		9		12		15
PR↑	53.5		54.0		54.7		53.5		OoM
RE↑	56.4		57.1		57.9		56.6		OoM
F-score↑	55.0		55.5		56.2		55.0		OoM

Table 3: Analysis of n in the U-shape network. OoM is the abbreviation of out of memory.

2023) are involved. When using UViT, the best performance is achieved when n equals 8, achieving 57.4 on F-score. When UNet is selected, the best performance is 1.2% worse than UViT, reaching 56.2 on F-score. Compared to the state-of-the-art tracker SPT, the performance of these two variants is better, especially the UViT-based GMMT, owning an increment of 1.3%.

Besides, the performance of our baseline method ViPT* is 49.2, 52.0, and 50.6 on PR, RE and F-score, respectively. That is to say, no matter which inner network is chosen and how many blocks are stacked, our GMMT can consistently boost the tracking performance.

As to the architecture of each block in UNet and UViT, it is beyond the scope of this manuscript, and please refer to (Song, Meng, and Ermon 2020) and (Bao et al. 2023), respectively.

G. Results on GTOT

In this part, we provide the attribute analysis and overall results on GTOT (Li et al. 2016). In general, based on TBSI (Hui et al. 2023), the proposed GMMT can boost its performance to a new state-of-the-art, reaching 78.5 and 93.6 on SR and PR, respectively.

H. Comparison of UNet and UViT

Since the proposed GMMT makes less limitation to the network architecture, two popular architectures are involved for comparison, *i.e.*, UNet (Song, Meng, and Ermon 2020) and UViT (Bao et al. 2023). The quantitative results are showcased in Table. 6 by conducting the experiments on three

s	1	2	3	4	5	6	7	8	9	15	20	30	40
PR↑	85.7	84.5	85.0	85.2	84.9	85.1	83.9	84.9	83.8	84.4	85.3	84.7	85.6
SR↑	69.3	68.4	68.5	68.5	68.5	68.4	67.9	68.4	68.0	68.1	68.7	68.4	68.9

Table 4: Quantitative analysis of diffusion steps.

	CAT	CMPP	SiamCDA	JMMAC	ADRNet	MaNet++	HMFT	MacNet	APFNet	TBSI	TBSI+GMMT
TC	71.0/90.0	72.9/**93.8**	68.5/82.6	70.5/88.6	73.9/90.6	70.7/89.9	73.2/89.2	69.4/89.5	71.6/90.4	75.0/90.7	**77.6**/*92.6*
OCC	69.2/89.9	71.6/**94.7**	69.4/82.2	68.7/84.0	69.9/87.9	70.1/89.0	72.2/88.1	68.5/88.2	71.3/90.3	75.9/91.8	**76.9**/*92.7*
LSV	68.0/85.0	70.0/91.2	74.8/91.5	74.6/90.3	70.8/85.5	69.3/86.6	75.4/89.1	67.1/84.9	71.2/87.8	*77.4*/*93.8*	**79.1/94.5**
FM	65.4/83.9	68.6/91.7	72.0/86.6	75.4/88.6	67.3/82.4	70.3/86.7	74.3/84.8	65.8/82.6	68.4/**96.5**	*76.0*/91.6	**77.4**/*91.8*
LI	72.3/89.2	74.3/92.4	76.4/92.4	76.5/*95.3*	76.2/91.9	73.1/91.7	76.9/94.3	72.9/90.0	74.8/91.4	*77.1*/94.1	**80.1/96.5**
SO	69.9/84.7	72.5/**98.1**	69.1/87.4	*73.8*/*95.2*	72.5/94.4	69.9/93.9	71.6/92.5	69.2/95.1	71.3/94.3	71.9/90.3	**75.3**/92.8
DEF	75.5/92.5	**78.8**/94.6	72.7/87.9	76.2/**96.4**	77.9/94.3	74.4/93.8	74.8/94.0	76.2/93.2	78.0/94.6	75.0/91.5	*78.4*/*94.7*
ALL	71.7/88.9	73.8/*92.6*	73.2/87.7	73.2/90.2	73.9/90.5	72.3/90.1	74.9/91.3	71.2/88.6	73.7/90.5	*75.9*/91.5	**78.5/93.6**

Table 5: Attribute-based Success/Precision score on GTOT dataset.

Benchmark	Method	PR↑	NPR↑	SR↑	Δ
GTOT	Siamese	84.0	-	67.0	
GTOT	Siamese+GMMT(U)	85.7	-	69.3	+2.3%
LasHeR	Siamese	50.9	47.4	39.8	
LasHeR	Siamese+GMMT(U)	57.1	53.0	44.9	+5.1%
LasHeR	ViPT*	65.0	61.6	52.4	
LasHeR	ViPT*+GMMT(U)	65.9	62.4	52.7	+0.3%
LasHeR	ViPT*+GMMT(V)	66.4	63.0	53.0	+0.6%
LasHeR	TBSI	69.2	65.7	55.6	
LasHeR	TBSI+GMMT(U)	70.7	67.0	56.6	+1.0%
LasHeR	TBSI+GMMT(V)	70.5	66.6	56.3	+0.7%
LasHeR	BD^2Track	56.0	-	43.2	

Table 6: Effectiveness analysis on multiple baseline trackers and benchmarks. U and V are the abbreviation of UNet and UViT, respectively.

tracking baselines and two benchmarks. From this table, we can see that both the UNet and UViT can be embedded into our GMMT with tracking performance better than the variant without our GMMT. Besides, compared to the only one published RGB-T method that employs the DM, $\mathrm{BD}^2\mathrm{Track}$ (Fan et al. 2023), a significant performance gap can be witnessed, which demonstrates the correctness our design.

I. Number of Diffusion Steps

Table. 4 displays the quantitative results during the analysis of reverse diffusion steps s. From this table, no positive correlation between the performance and s is observed. Meanwhile, a larger s will cost more time and computational resources, which harms the efficiency significantly. Therefore, s is set to 1.

J. Analysis of λ

λ is a factor banding the tracking and generation tasks. Fig. 4 shows the analysis on GTOT (solid) and LasHeR (dashed) benchmarks, with λ chosen from (0,1,2,3,5,10,100). 0 represents the generative loss is inactivated and the network $\mathcal{U}$ is optimized by the $Loss_{track}$, which is inherited from the baseline methods. When λ is a nonzero value, stable improvements can be found on all the metrics. However, the performance suffers a heavy degradation when λ=100. At this time, the $Loss_{track}$ is too small and its influence on $\mathcal{U}$ is negligible. Therefore, the above phenomenon indicates that the strong supervision of the tracking task is crucial and should not be ignored.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China (Grant NO.62106089, 62020106012, 62332008, 62336004, U1836218), the 111 Project of Ministry of Education of China (Grant No.B12018), the Fundamental Research Funds for the Central Universities (JUSRP123030) and the UK EPSRC (EP/N007743/1, MURI/EPSRC/DSTL, EP/R018456/1).

References

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023. All are Worth Words: A ViT Backbone for Diffusion Models. In *CVPR*.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese Box Adaptive Network for Visual Tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6667–6676.

Fan, S.; He, C.; Wei, C.; Zheng, Y.; and Chen, X. 2023. Bayesian Dumbbell Diffusion Model for RGBT Object Tracking With Enriched Priors. *IEEE Signal Processing Letters*, 30: 873–877.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hui, T.; Xun, Z.; Peng, F.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Bridging Search Region Interaction With Template for RGB-T Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13630–13639.

Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; and Lin, L. 2016. Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking. *IEEE Transactions on Image Processing*, 25(12): 5743–5756.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96: 106977.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2022. LasHeR: A Large-Scale High-Diversity Benchmark for RGBT Tracking. *IEEE Transactions on Image Processing*, 31: 392–404.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, 300–317.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023a. Visual Prompt Multi-Modal Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9516–9526.

Zhu, X.-F.; Xu, T.; Tang, Z.; Wu, Z.; Liu, H.; Yang, X.; Wu, X.-J.; and Kittler, J. 2023b. RGBD1K: A Large-Scale Dataset and Benchmark for RGB-D Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3870–3878.

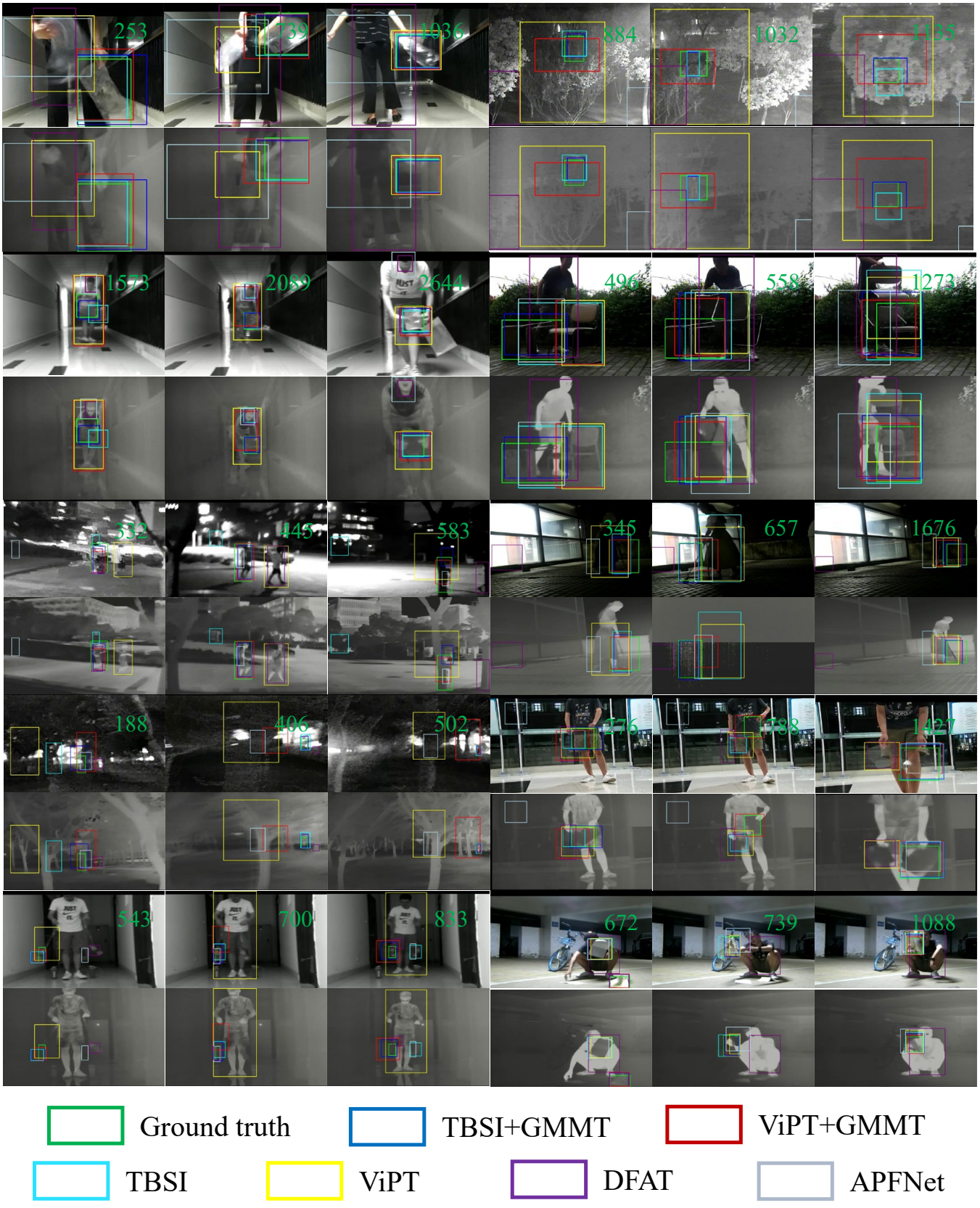


Figure 5: Qualitative results on LasHeR. The exhibited image pairs are sampled from video $leftthyalinepaperfrontpants$, $hyalinepaperfrontface$, $boytakingbasketballfollowing$, $darktreesboy$, $leftbottle2hang$, $boyride2path$, $leftchair$, $broom$, $large$, $leftexcersicebookyellow$, which are introduced in a top-down and left-right way.