RGB-D Tracking via Hierarchical Modality Aggregation and Distribution Network

Boyue Xu
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
xuby@smail.nju.edu.cn

Yi Xu
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
yxu1025@smail.nju.edu.cn

Ruichao Hou
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
rc_hou@smail.nju.edu.cn

Jia Bei*
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
beijia@nju.edu.cn

Tongwei Ren
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
rentw@nju.edu.cn

Gangshan Wu
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
gswu@nju.edu.cn

ABSTRACT

The integration of dual-modal features has been pivotal in advancing RGB-Depth (RGB-D) tracking. However, current trackers are less efficient and focus solely on single-level features, resulting in weaker robustness in fusion and slower speeds that fail to meet the demands of real-world applications. In this paper, we introduce a novel network, denoted as HMAD (Hierarchical Modality Aggregation and Distribution), which addresses these challenges. HMAD leverages the distinct feature representation strengths of RGB and depth modalities, giving prominence to a hierarchical approach for feature distribution and fusion, thereby enhancing the robustness of RGB-D tracking. Experimental results on various RGB-D datasets demonstrate that HMAD achieves state-of-the-art performance. Moreover, real-world experiments further validate HMAD's capacity to effectively handle a spectrum of tracking challenges in real-time scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**; *Artificial intelligence*; *Computer vision*.

KEYWORDS

RGB-D , single object tracking , multi-modal fusion , attention mechanism

1 INTRODUCTION

RGB-D tracking is a type of multimodal object tracking [11, 15, 16, 29], which combines RGB and depth data. RGB offers visual details like color and texture but is limited to 2D dimensions. Depth information, providing the distance from the camera to the object, enables accurate 3D object localization. This combination is widely used in tasks like saliency detection, object detection, and tracking [9, 10, 32, 40, 41].

Among these tasks, RGB-D tracking presents the most valuable and challenging applications. It plays a crucial role in human-computer interaction, robotic environmental perception, and other
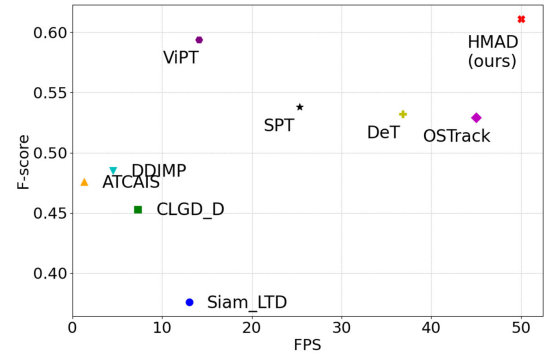
*Corresponding author.

Figure 1: Comparison results with representative trackers on DepthTrack dataset.

domains [1, 17, 25, 35]. The fundamental requirement of this task is to efficiently utilize the complementary dual-modal features in real-time to address complex tracking scenarios. Especially in complex indoor environments with intricate backgrounds, similar targets, and dim lighting, the optimal use of these dual-modal complementary features becomes the central challenge for RGB-D tracking.

Current RGB-D trackers primarily focus on modal fusion methods [36, 37, 40]. Fusion methods have evolved from simple additive or weighted fusion of both modalities [36] to using Transformer for fusion [40] or applying prompt learning to guide the fusion process [39]. While these methods have gradually improved fusion quality, they have introduced two challenges: (1) Can the increasingly complex fusion methods meet the real-time requirements of RGB-D tracking tasks? This concern is particularly relevant in contexts where RGB-D tracking is commonly used, such as robotics and human-computer interaction device, where computational power is limited. (2) Are all features suitable for RGB-D tracking tasks? Is it possible to select appropriate features for the tracking task?

We propose the HMAD tracker specifically designed to address the aforementioned issues. To address the first issue, the HMAD

tracker adopts the classical RGB tracking framework DIMP [3] as a baseline. This framework has strong feature discrimination capability and uses less computational cost compared to the complex Transformer structure. Additionally, HDMA itself is a simple and effective hierarchical modality fusion method, avoiding the substantial speed loss that complex fusion methods can entail. For the second problem, we propose a hierarchical modality aggregation and distribution network to extract rich color and texture features provided by RGB at the shallow feature level, and semantic features such as contour distance provided by depth information at the deep feature level. By effectively integrating these features, we extract as many valid features as possible while using minimal resources, thus enhancing the robustness of RGB-D tracking. Experimental results show that the HMAD tracker performs well on major RGB-D tracking datasets. Importantly, it achieves a tracking speed of 15 FPS on edge device, meeting real-time tracking requirements.

The main contributions of this paper is summarized as follows:

- We propose a real-time RGB-D tracker HMAD based on the DIMP framework, which guarantees high tracking performance while achieving real-time tracking speed.
- We propose a hierarchical modality aggregation and distribution network capable of fully extracting and integrating effective features from both RGB and depth information.
- We validate the effectiveness of the HMAD on existing RGB-D tracking datasets. Moreover, we conduct real-world experiments on real-world edge device to demonstrate the effectiveness and real-time capabilities of the tracker in real-world scenarios.

2 RELATED WORK

2.1 Single Object Tracking

Two primary strategies dominate the field of single-object tracking: Siamese-based and tracking-by-detection methods. Siamese-based methods, represented by SiamFC [2], utilize template correlation within a search area to identify tracking targets. Subsequent enhancements to this approach include SiamRPN [4] and SiamRPN++ [22] by Li *et al.*, which integrate RPN [31] networks and feature pyramids to increase model performance. Further advances have seen the incorporation of Transformer [33] structures, as demonstrated in Chen *et al.*'s TransT [6] tracking network, providing improved global association capabilities for tracking. Subsequently, Cui *et al.* delved into the Transformer architecture, going a step further to replace the traditional convolutional neural network backbone with a Transformer structure [33], which led to the development of the MixFormer [7]. This innovation greatly enhanced the effectiveness of feature extraction and achieved state-of-the-art results in the domain of generic single-object tracking. In a parallel approach, Lin *et al.* adopted a similar line of thinking, designing a new Transformer-based backbone network for tracking, as described in their work SwinTrack [23].

Contrastingly, tracking-by-detection methods, exemplified by MDNet [28] from Nam *et al.*, employ a discriminator model to locate tracking target within input image. Extensions of this approach include the ATOM [8] by Danelljan *et al.*, which integrates both target estimation and online update modules to increase robustness. Bhat *et al.*'s DiMP [3] method introduced a discriminative learning loss, enhancing target identification. Moreover, the Keep-track [26] method by Mayer *et al.* incorporates a target association module that is specialized for scenarios with a large number of similar interfering targets.

2.2 RGB-D Tracking

In the domain of RGB-D tracking, there are two main strategies: methods based on correlation filters, and methods relying on transfer RGB tracking [41].

The former approach includes the work of researchers such as Camplani *et al.*, who proposed a real-time tracker [5] based on the Kernelized Correlation Filters (KCF) [14] algorithm, which integrates RGB and depth information to manage complex situations such as occlusions. This work demonstrates promising results. Kart *et al.* seek to transform short-term RGB trackers into RGB and depth trackers by proposing a general framework [19] that integrates an occlusion detection module based on depth segmentation into a correlation filter framework [18]. In addition, Harika *et al.* introduced multimodal fusion at different layers to integrate the features of RGB images and depth maps [12].
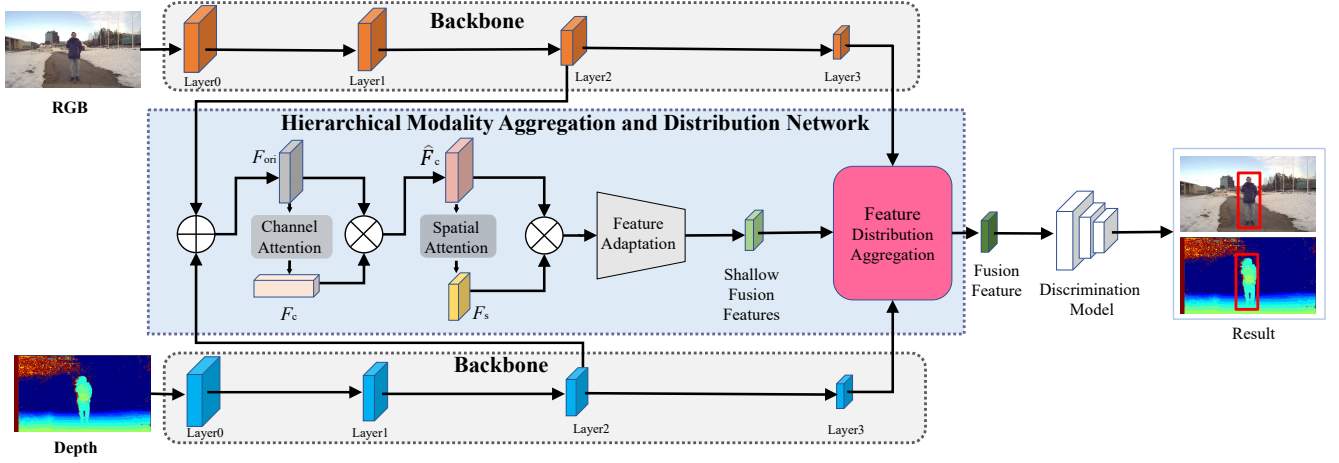
On the other hand, among the transfer RGB tracker based methods, Yan *et al.* performed fusion of RGB and depth features on end-to-end RGB tracker ATOM [8] and DiMP [3]. They used a straightforward addition method to merge the deep features of RGB and depth images [36]. Xue *et al.* went a step further and used the more advanced Transformer tracking framework for feature fusion [40]. They leverage the powerful feature association capability of Transformer to achieve better results, although the complexity of the Transformer structure results in slower tracking. Additionally, some researchers applied prompt learning to RGB-D tracking [37, 39]. By utilizing minimal prompt information, they were able to transfer single-modality tracking methods to RGB-D tracking, substantially reducing the complexity and challenges of training. This approach represents a novel utilization of existing methodologies, expanding the applicability to more advanced and diverse tracking scenarios.

3 METHODOLOGY

3.1 Network Architecture

We propose an RGB-D single object tracking method namde HMAD, based on the DIMP [3]. This method incorporates a hierarchical modality aggregation and distribution network to integrate features from both modalities. These fused features are then used to train a discrimination model, which subsequently predicts the classification score and location of the tracking target. The flow of the method is illustrated in Figure 2.

The input of the HMAD tracker is the corresponding RGB and depth images, which are merged through a hierarchical modality aggregation and distribution network. This process consists of two parts: the first part involves attention-based [34] shallow feature extraction, while the second part involves feature distribution and fusion. The former is responsible for extracting the effective components in shallow features, while the latter is responsible for distributing and fusing shallow features with deep features. The fused features are then used for final target discrimination and tracking.

Figure 2: The framework of HMAD, consists of backbone , hierarchical modality aggregation and distribution network and a target discrimination model.

Our method uses random samples from video sequences to train a discriminative model derived from the DIMP [3]. The fused features are fed into the discrimination model initialization module, which initializes the target region. This module adopts a method of precise pooling to generate $W \times H \times N$ feature filters. It then iteratively optimizes the initialized filters with background information from the target region and generates the final discriminative model, which is used in the testing phase to predict the classification score and location of the tracked target.

In order to maintain the robust generalization ability of the discrimination model during tracking, the HMAD tracker employs the online update strategy of DIMP [3] to update the discrimination model. During prediction, for the first frame with annotations, a data augmentation strategy is used to construct an initial set of K samples, given the accuracy of the annotations. The trained discrimination model is then applied to these K samples for further gradient descent, initially satisfying the feature distribution of the sequence. As tracking progresses, if the confidence in the tracked target is high, the network opts to discard the oldest samples, maintaining a maximum of L samples. Throughout the tracking process, the discrimination model is optimized using the saved samples every P frames. This ensures that the discrimination model always satisfies the feature distribution of the sequence and does not lose its discriminative power over the target features due to environmental disturbances or target scale variations.

3.2 Hierarchical Modality Aggregation and Distribution Network

The hierarchical modality aggregation and distribution network can be divided into two parts: attention-based shallow feature extraction and feature distribution fusion. Their respective roles are to extract the effective parts of shallow features and effectively fuse deep and shallow features.

Attention-based shallow feature extraction first extracts shallow features from the second layer of the ResNet50 [13] backbone network and uses an attention mechanism to fuse the shallow features, which uses CBAM [34] attention module. The module contains a channel attention component and a spatial attention component. The experiments of the CBAM show that places channel attention before spatial attention can achieve better performance. Therefore, we use the same design. To facilitate reader comprehension, we briefly introduce channel attention and spatial attention. For further details, please refer to CBAM [34]. The channel attention first apply max-pooling and mean-pooling to the features. Subsequently, the channel attention distributes weights across the channels, employing a multi-layer perceptron for this process. The output feature vectors from these two parts are then summed element-wise and activated using a sigmoid function. The formulation governing the channel attention mechanism is given as follows:

$$F_c = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F_{ori})) + \mathrm{MLP}(\mathrm{MaxPool}(F_{ori}))), \quad (1)$$

where F_c represents the feature processed by the channel attention mechanism, $\sigma(\cdot)$ denotes the sigmoid activation function, and $\mathrm{MLP}(\cdot)$ stands for a multi-layer perceptron; $\mathrm{AvgPool}(\cdot)$ represents average pooling and $\mathrm{MaxPool}(\cdot)$ correspond to max pooling operations.

The role of spatial attention is to enhance the meaningful local regions within features. The feature output from channel attention is first concatenated after max pooling and average pooling operations in sequence, and then output after convolution and activation operations. The spatial attention formulation is as follows:

$$F_s = \sigma\left(f^{7\times7}([\mathrm{AvgPool}(\hat{F}_c); \mathrm{MaxPool}(\hat{F}_c)])\right), \quad (2)$$

where F_s represents the feature processed by the spatial attention mechanism, $\sigma(\cdot)$ denotes the Sigmoid activation function, and $f^{7\times7}$ is a convolution operation with a 7×7 kernel. $\mathrm{AvgPool}(\cdot)$ and $\mathrm{MaxPool}(\cdot)$ are the average pooling and max pooling operations respectively.
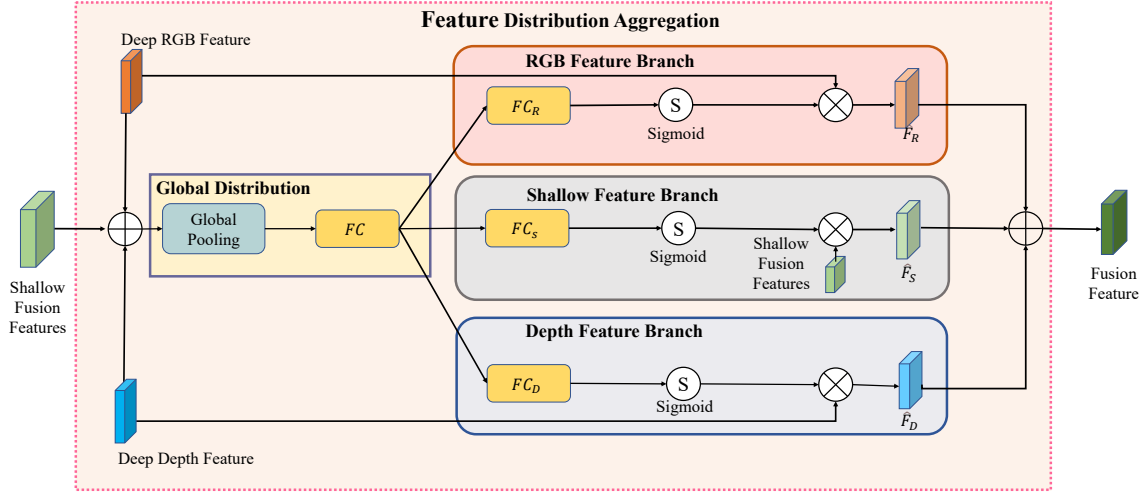
Figure 3: The details of the feature distribution.

After the attention mechanism, the parts of shallow features that are beneficial for the tracking task are greatly enhanced. These features not only complement the lack of texture information and detailed features in deep features, but also guide deep features to pay more attention to the parts that require more attention. This module uses convolutional pooling operations to scale it to the same size as the deep features, which are then fed into the feature distribution module together with the deep features.

The inputs to this module include deep RGB features, deep depth features, and shallow fusion features. After adding the three features, they are passed into the global average pooling operation. This operation obtains a global feature vector through average pooling and projection operations to comprehensively consider the three aspects of features. Then this feature vector is divided into three parts, which are the RGB feature branch, the depth feature branch, and the shallow feature branch. Each branch goes through a projection operation and activation operation, and outputs a weight for each channel of the branch. This weight describes in detail the importance of each channel of the three features. Multiplying this weight by the three-channel input feature elementally, the adjusted deep and shallow features can be obtained. These features not only comprehensively consider the necessary feature information globally but also contain contour and texture features, which greatly assist the subsequent tracking tasks. The formula for global distribution is as follows:

$$F_g = FC\left(GP\left(F_R \oplus F_D \oplus F_S\right)\right), \quad (3)$$

where $FC(\cdot)$ represents fully connected layer, $GP(\cdot)$ represents global pooling; F_R, F_D and F_S represent deep RGB features, deep depth features, and shallow fusion features, respectively; $\oplus$ denotes element-wise addition. After obtaining the distribution features of the three branches, the final fused features can be obtained again using element-wise addition.

$$\hat{F}_i = F_i \otimes \sigma\left(FC_i\left(F_{global}\right)\right), i \in \{R, D, S\}, \quad (4)$$

where $\hat{F}_i$ represents the processed corresponding feature; $FC_i(\cdot)$ represents fully connected layer in different branches; $\otimes$ denotes the element-wise multiplication operation.

4 EXPERIMENTS

4.1 Datasets and Metrics

We conduct comparison experiments with existing high-performance trackers on two popular RGB-D tracking datasets, DepthTrack [36] and RGBD1K [40]. Given that the datasets predominantly consists of extended temporal sequences, evaluation metrics for long-term tracking have been selected to assess performance. These include precision (Pr), recall (Re), and F-score [21, 24].

Precision is calculated by the Gaussian Mixture Distribution between all frame output boxes and the given correct output boxes. The sum of all computed Gaussian Mixture Distributions is divided by the total frame count to determine tracking precision. The precision is calculated as follows:

$$\Pr\left(\tau_\theta\right) = \frac{1}{N_p} \sum_t \Omega\left(A_t\left(\theta_t\right), G_t\right), t \in \{t : A_t\left(\theta_t\right) \neq \emptyset\}, \quad (5)$$

where $\Pr\left(\tau_\theta\right)$ represents precision, $A_t\left(\theta_t\right)$ represents the tracker's output, G_t represents the ground truth, and $\Omega(\cdot)$ represents the intersection of the two. The sum is taken over all non-empty predicted results.
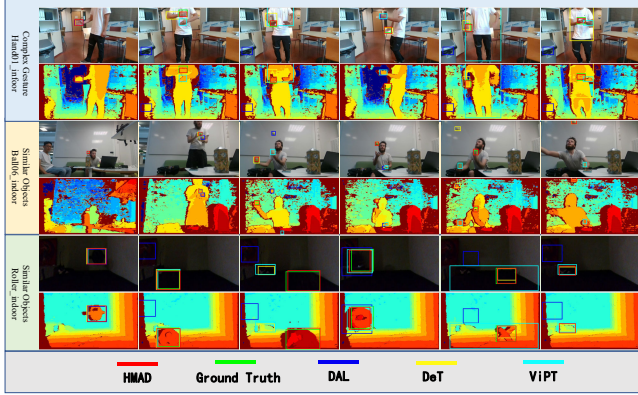
Recall is calculated by the Gaussian Mixture Distribution between all frame output boxes and the given correct output boxes. The sum of all computed Gaussian Mixture Distributions is divided by the total frame count where targets are present to determine tracking recall:

$$\Re\left(\tau_\theta\right) = \frac{1}{N_g} \sum_t \Omega\left(A_t\left(\theta_t\right), G_t\right), t \in \{t : G_t \neq \emptyset\}, \quad (6)$$

where $\Re\left(\tau_\theta\right)$ represents recall, $A_t\left(\theta_t\right)$ represents the tracker's output. The sum is taken over all non-empty ground truth results.

Table 1: Comparision between HDMA and the state-of-the-arts trackers. The best results are highlighted in red and the second are highlighted in green. The performance is evaluated in terms of precision (Pr), recall (Re), F-score and frame per second (FPS).

Method	DepthTrack			RGBD1K			FPS
	Pr	Re	F-score	Pr	Re	F-score	
Siam_LTD [27]	0.342	0.418	0.376	0.543	0.318	0.398	13.0
DAL [30]	0.512	0.369	0.429	0.562	0.407	0.472	-
CLGS_D [27]	0.369	0.584	0.453	-	-	-	7.3
ATCAIS [27]	0.455	0.500	0.476	0.511	0.451	0.479	1.3
DDIMP [27]	0.469	0.503	0.485	0.557	0.534	0.545	4.7
Drefine [20]	-	-	-	0.532	0.462	0.494	-
OSTrack [38]	0.522	0.536	0.529	-	-	-	-
DeT [36]	0.506	0.560	0.532	0.438	0.419	0.428	36.8
SPT [40]	0.549	0.527	0.538	0.545	0.578	0.561	25.3
ProTrack [37]	0.583	0.573	0.578	-	-	-	-
ViPT [39]	0.596	0.592	0.594	-	-	-	14.1
HMAD	0.626	0.597	0.611	0.573	0.552	0.562	50.0

Figure 4: Qualitative comparison between HDMA and other trackers on three challenging sequences in DepthTrack dateset.

F-score is divided by the summary of Pr and Re and then multiplied by two to obtain the tracking F-score:

$$F\text{-}score\left(\tau_\theta\right) = 2\frac{\Pr\left(\tau_\theta\right)\text{Re}\left(\tau_\theta\right)}{\left(\Pr\left(\tau_\theta\right)+\text{Re}\left(\tau_\theta\right)\right)}, \qquad (7)$$

where $\text{Re}\left(\tau_\theta\right)$ represents the corresponding recall; $\Pr\left(\tau_\theta\right)$ represents the corresponding precision.

4.2 Implementation Details

All experimental methods are trained on a server with an 5.2GHz CPU and a RTX-3090 GPU with 24GB memory. The proposed tracker was then deployed on the NVIDIA Jetson AGX Orin platform, which boasts a 1.6GHz CPU and 2,000 stream processors GPU. In terms of parameter settings, we utilize an Adam optimizer with a learning rate of 2e-4 for optimization. The network loss function is divided into two parts, namely the classification loss and the IoU loss. Both of these are online update loss functions that change as

the model is updated online. During prediction, for the first annotated frame, the initial set of samples is set to 15, the threshold for adding samples is set to 0.6, and the maximum number of samples is set to 50.

4.3 Compararison with the State-of-the-Arts

To verify the effectiveness of HMAD, we compare HMAD on the test set of DepthTrack [36] and RGBD1K [40] with 11 state-of-arts methods, including DDIMP [27], ATCAIS [27], Siam_LTD [27], OSTrack [38], CLGS_D [27], Drefine [20], DAL [30], DeT [36], Pro-Track [37], SPT [40] and ViPT [39]. All the tracking speed are derived from the published papers.

The results of the proposed tracker on DepthTrack [36] are presented in Table 1. The proposed method outperforms all existing techniques, including the prompt learning based method ViPT [39]. Specifically, we observe improvements of 3.0%, 0.5% and 1.7% in Pr, Re, and F-score, respectively. Moreover, our method achieves the fastest tracking speed among the existing methods.

The comparative results on the RGBD1K [40] dataset are shown in Table 1. The proposed tracker surpasses all other methods. Compared to the second-placed SPT method, our method leads by 2.8% and 0.1% in PR and F-score, respectively. The fact that the proposed tracker achieves the best performance on the larger, more complex RGBD1K dataset sufficiently demonstrates its superiority. To intuitively demonstrate the effectiveness of HMAD, we select three challenging sequences from DepthTrack dataset [36] for visualization analysis, as shown in Figure 4. In the Hand01_indoor sequence, the main challenges are complex hand gesture changes and similar targets. HMAD accurately and robustly tracks the target, while other methods suffer from tracking losses and errors. In the Ball06_indoor sequence, the primary challenges include numerous similar targets and rapid movement, HMAD stably tracked the correct target. In the Rolle_indoor sequence, the main challenge is the dimly lit environment, and HMAD effectively overcame this problem, achieving stable tracking.
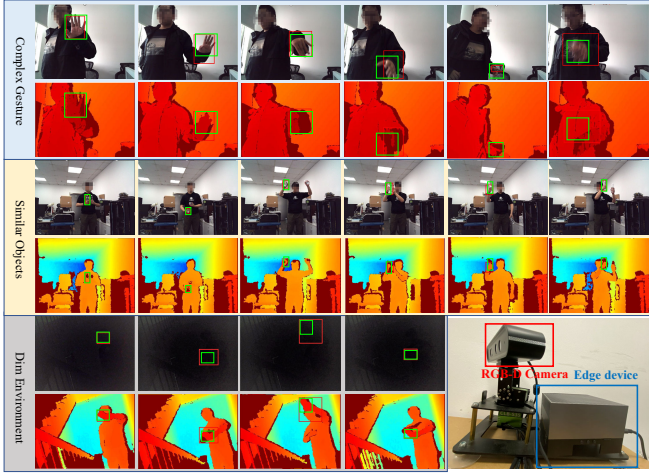
Figure 5: Real-world test results of the proposed tracker, the tracking results are marked in red and the ground truth are marked in green.

Table 2: Ablation study on different components. The performance is evaluated on the DepthTrack test set in terms of precision (Pr), recall (Re) and F-score.

Variations	Pr	Re	F-score
baseline	0.548	0.525	0.536
w/o distribution	0.571	0.545	0.557
w/o attention	0.596	0.562	0.579
HMAD	0.626	0.597	0.611

4.4 Ablation Study

To evaluate the specific impact of each module within our proposed method, we conduct an ablation experiment using the test set of the DepthTrack [36]. Table 2 presents the results of our ablation experiment. The baseline model used in this study directly add RGB and depth data, resulting in Pr, Re, and F-score values of 0.548, 0.525 and 0.536 respectively. We use the attention module [34] independently on top of the baseline, resulting in improvements of 2.3%, 2.0% and 2.1% for Precision, Recall, and F-score, respectively. Similarly, the incorporation of the feature distribution aggregation module separately led to performance gains of 4.8%, 3.7% and 4.1% for Precision, Recall, and F-score, respectively. The complete HDMA achieve a score of 0.626, 0.597 and 0.611 respectively. The ablation study allows us to investigate the impact of each module on the overall performance of the proposed method.

4.5 Real-World Tests

In this section, we deploy the proposed tracker on an actual edge device to validate its performance in the real world. We conduct experiments using the NVIDIA Jetson AGX Orin platform and the Orbbec Astra PRO RGB-D dual-modality camera, The device used is shown in the bottom right corner of Figure 5. Notably, this camera can effectively detect distances up to eight meters with an error

of less than three millimeter, meeting the requirements of real-world application scenarios. Meanwhile, the edge device used in this system has a computing power of approximately one-third of the mainstream GPU RTX-3090, which places higher requirements on real-time performance for the method.

RGB-D tracking is primarily used in the real world for human pose estimation and behavior recognition. To evaluate the performance of the proposed tracker in these application scenarios, we simulated real-world scenarios and added corresponding challenges. As shown in rows one and two of Figure 5, in the commonly used gesture tracking task in human-machine interaction, the proposed method can stably track complexly varying hand targets. The images in row three and four demonstrate tracking in scenes with similar targets; when there are two similar targets in the scene, the proposed method can still stably track the correct target. The last two rows show the tracking effect in a dim environment; even when the RGB modality is essentially ineffective, the proposed method still successfully tracks the target using depth information.

Throughout all experiments, the tracker consistently maintained a tracking speed of 15 FPS on the edge device. These experiments comprehensively illustrate that the proposed tracker not only attained excellent metrics on offline datasets but also demonstrated real-time tracking capabilities and exceptional performance in real-world environments and on edge device.

5 CONCLUSIONS

In this paper, we proposed a novel HMAD tracker for real-time and robust RGB-D tracking tasks. We proposed a hierarchical modality aggregation and distribution network that efficiently fuses multi-level features from RGB and depth modalities. Experimental results demonstrated that HMAD achieves state-of-the-art performance on multiple datasets. Moreover, real-world test verified that HMAD can effectively handle various challenges presented in real-world and achieves real-time effectiveness.

REFERENCES

[1] Md. Shahinur Alam, Ki-Chul Kwon, and Nam Kim. 2021. Implementation of a Character Recognition System Based on Finger-Joint Tracking Using a Depth Camera. *IEEE Transactions on Human-Machine Systems* 51, 3 (2021), 229–241.

[2] L. Bertinetto, J. Valmadre, Joo F. Henriques, A. Vedaldi, and Phs Torr. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In *European Conference on Computer Vision Workshops*.

[3] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte. 2019. Learning Discriminative Model Prediction for Tracking. In *IEEE International Conference on Computer Vision*.

[4] L. Bo, J. Yan, W. Wei, Z. Zheng, and X. Hu. 2018. High Performance Visual Tracking with Siamese Region Proposal Network. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[5] M. Camplani, S. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt. 2015. Real-time RGB-D Tracking with Depth Scaling Kernelised Correlation Filters and Occlusion Handling. In *British Machine Vision Conference*.

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. 2022. Mixformer: End-to-end Tracking with Iterative Mixed Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[8] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. 2020. ATOM: Accurate Tracking by Overlap Maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[9] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. 2021. Siamese Network for RGB-D Salient Object Detection and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 5541–5559.

[10] Jingfan Guo, Tongwei Ren, and Jia Bei. 2016. Salient Object Detection for RGB-D Image Via Saliency Evolution. In *IEEE International Conference on Multimedia and Expo*.

[11] Stefan Haag, Bharanidhar Duraisamy, Wolfgang Koch, and Jürgen Dickmann. 2018. Radar and Lidar Target Signatures of Various Object Types and Evaluation of Extended Object Tracking Methods for Autonomous Driving Applications. In *International Conference on Information Fusion*.

[12] Harika, Narumanchi, Dishant, Goyal, Nitesh, Emmadi, Praveen, and Gauravaram. 2018. Hierarchical Multi-modal Fusion FCN with Attention Model for RGB-D Tracking. In *IEEE International Conference on Cloud Engineering*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[14] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2014. High-speed Tracking with Kernelized Correlation Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2014), 583–596.

[15] Ruichao Hou, Tongwei Ren, and Gangshan Wu. 2022. MIRNet: A Robust RGBT Tracking Jointly with Multi-Modal Interaction and Refinement. In *IEEE International Conference on Multimedia and Expo*.

[16] Ruichao Hou, Boyue Xu, Tongwei Ren, and Gangshan Wu. 2023. MTNet: Learning Modality-aware Representation with Transformer for RGBT Tracking. In *IEEE International Conference on Multimedia and Expo*.

[17] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang. 2015. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[18] Uğur Kart, Joni-Kristian Kämäräinen, Jiří Matas, Lixin Fan, and Francesco Cricri. 2018. Depth Masked Discriminative Correlation Filter. In *International Conference on Pattern Recognition*.

[19] Uur Kart, Joni Kristian Kmrinen, and Jií Matas. 2019. How to Make an RGBD Tracker?. In *European Conference on Computer Vision Workshops*.

[20] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, et al. 2021. The Ninth Visual Object Tracking Vot2021 Challenge Results. In *IEEE International Conference on Computer Vision*.

[21] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. 2019. The Seventh Visual Object Tracking VOT2019 Challenge Results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

[22] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan. 2020. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[23] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. 2022. Swintrack: A Simple and Strong Baseline for Transformer Tracking. In *Advances in Neural Information Processing Systems*.

[24] Alan Lukeźič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. 2020. Performance Evaluation Methodology for Long-term Single-object Tracking. *IEEE Transactions on Cybernetics* 51, 12 (2020), 6305–6318.

[25] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. 2022. MultiScan: Scalable RGBD Scanning for 3D Environments with Articulated Objects. In *Advances in Neural Information Processing Systems*.

[26] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. 2021. Learning Target Candidate Association to Keep Track of What Not to Track. In *IEEE International Conference on Computer Vision*.

[27] A. Memarmoghadam. 2021. The Eighth Visual Object Tracking VOT2020 Challenge Results. In *European Conference on Computer Vision Workshops*.

[28] Hyeonseob Nam and Bohyung Han. 2016. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[29] J Peršić, L Petrović, I Marković, and I Petrović. 2021. Online Multi-sensor Calibration Based on Moving Object Tracking. *Advanced Robotics* 35, 3-4 (2021), 130–140.

[30] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, and Jiří Matas. 2021. DAL: A Deep Depth-aware Long-term Tracker. In *International Conference on Pattern Recognition*.

[31] S. Ren, K. He, R. Girshick, and J. Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (2017), 1–9.

[32] Tongwei Ren and Ao Zhang. 2019. *RGB-D Salient Object Detection: A Review*. Springer International Publishing, Cham, 203–220.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*.

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional Block Attention Module. In *European Conference on Computer Vision*.

[35] Peng Wu, Runze Guo, Xiaozhong Tong, Shaojing Su, Zhen Zuo, Bei Sun, and Junyu Wei. 2022. Link-RGBD: Cross-guided Feature Fusion Network for RGBD Semantic Segmentation. *IEEE Sensors Journal* 22, 24 (2022), 24161–24175.

[36] S. Yan, J. Yang, J. Kpyl, F. Zheng, A. Leonardis, and J. K. Kmrinen. 2021. DepthTrack : Unveiling the Power of RGBD Tracking. In *IEEE International Conference on Computer Vision*.

[37] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. 2022. Prompting for Multi-modal Tracking. In *the ACM International Conference on Multimedia*.

[38] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2022. Joint Feature Learning and Relation Modeling for Tracking: A One-stream Framework. In *European Conference on Computer Vision*.

[39] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. 2023. Visual Prompt Multi-Modal Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[40] Xue-Feng Zhu, Tianyang Xu, Zhangyong Tang, Zucheng Wu, Haodong Liu, Xiao Yang, Xiao-Jun Wu, and Josef Kittler. 2023. RGBD1K: A Large-scale Dataset and Benchmark for RGB-D Object Tracking. In *AAAI Conference on Artificial Intelligence*.

[41] Xue-Feng Zhu, Tianyang Xu, and Xiao-Jun Wu. 2022. Visual Object Tracking on Multi-modal RGB-D Videos: A Review. *arXiv preprint arXiv:2201.09207* (2022), 1–5.