VISUAL ADAPT FOR RGBD TRACKING

Guangtong Zhang[a,b], *Qihua Liang*[*,a,b], *Zhiyi Mo*[a,b,c], *Ning Li*[a,b], *Bineng Zhong*[a,b]
[a]Key Laboratory of Education Blockchain and Intelligent Technology Ministry of Education,
Guangxi Normal University, Guilin 541004, China.
[b]Guangxi Key Lab of Multi-Source Information Mining & Security,
Guangxi Normal University, Guilin 541004, China.
[c]Guangxi Key Laboratory of Machine Vision and Intelligent Control,
Wuzhou University, Wuzhou 543002, China.

ABSTRACT

Recent RGBD trackers have employed cueing techniques by overlaying Depth modality images as cues onto RGB modality images, which are then fed into the RGB-based model for tracking. However, the direct overlaying interaction method between modalities not only introduces more noise into the feature space but also exhibits the inadaptability of the RGB-based model to mixed-modality inputs. To address these issues, we introduce *Visual Adapt for RGBD Tracking (VADT)*. Specifically, we maintain the input of the RGB-based model as the RGB modality. Additionally, we have devised a fusion module to enable modality interaction between depth and RGB features. Subsequently, a Depth Adapt module has been formulated to facilitate image interaction with the fused features. This module involves cross-attending to the obtained depth-assisted features and the RGB search frame features produced by the RGB-based model's output. Experimental results indicate that our proposed tracker achieves state-of-the-art results on various RGBD benchmark tests.

***Index Terms*—** Adapt RGBD Tracking

1. INTRODUCTION

Achieving robust and stable tracking in complex scenes remains a significant challenge in object-tracking tasks. In recent years, numerous outstanding trackers[1, 2, 3] have emerged, achieving some progress in addressing this challenge. With the proliferation of depth sensors, attempts have been made to leverage both Depth and RGB images for joint tracking. The main advantage of RGBD tracking lies in its ability to enhance the robustness and stability of tracking in complex scenarios characterized by significant changes in object appearance, occlusions, and low-light conditions.

In comparison to traditional RGB tracking, the distinct information carried by Depth images in RGBD tracking introduces modal disparities. Early RGBD trackers[4, 5, 6] independently extract RGB and Depth features, followed by depth fusion in fusion modules. For example, DAL[7] employs depth-aware convolution to enhance RGB features by extracting Depth feature information with depth awareness. With the rapid development of transfer learning in the computer vision field, Yang et al.[8] introduce ProTrack, where the Depth modality is directly overlaid on the RGB modality as input to the RGB-based model for tracking. Subsequently, Zhu et al.[9] advance this overlay approach in ViPT by embedding Depth features into a low-dimensional space via convolution before overlaying them onto the RGB features inputted to the RGB-based model for tracking. However, these prompt strategies introduce two issues: 1) The construction of Depth images differs from that of RGB images, leading to the introduction of interference information in the feature space due to the direct overlay of Depth features onto RGB images. 2) Directly overlaying Depth features onto RGB features results in a hybrid-modality image, causing adaptation issues for the RGB-based model when dealing with mixed-modality inputs.

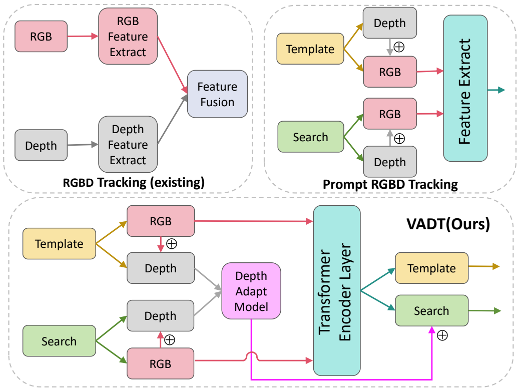

Fig. 1. We compare our proposed VADT model with existing RGBD trackers and Prompt-based RGBD trackers. Our proposed tracker employs the Depth Adapt Module, which reduces the influence of depth features on the input feature space of the model compared to previous trackers.

Qihua Liang is the corresponding author.

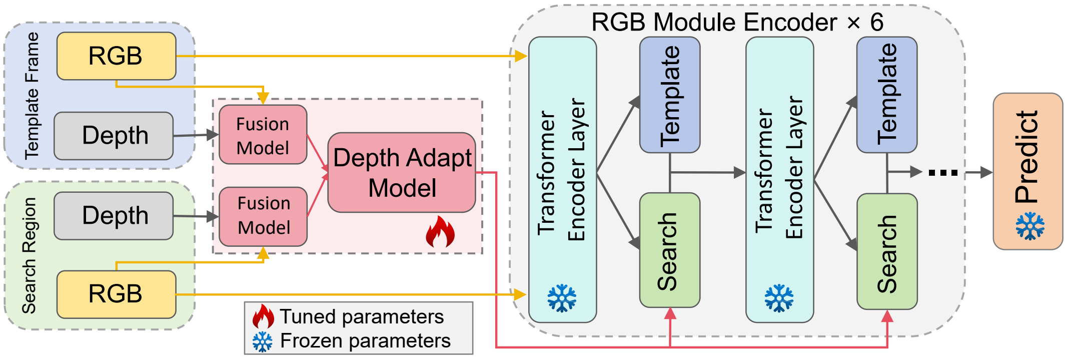

Fig. 2. Overview of the VADT pipeline. Our proposed VADT mainly consists of deep adapt modules.

To address these problems, we propose a novel RGBD tracker called Visual Adapt for RGBD Tracking (VADT). VADT differentiates itself from previous RGBD trackers, as depicted in the Fig.1. VADT refrains from directly overlaying Depth features onto RGB features. Instead, it maintains the input of the RGB-based model as RGB images, reducing the RGB-based model's inadaptability to hybrid features. Depth and RGB modalities are then fused, with hybrid-modality features undergoing Depth Adapt Model processing to generate depth-modality features. Utilizing a data pathway, depth-modality features are overlayed onto the search features outputted by the RGB-based model, thus preserving the model's template feature space. The primary contributions of this paper are as follows:

- We propose a novel Adapt RGBD tracker named Visual Adapt for RGBD Tracking (VADT). This framework maintains RGB images as input to the RGB-based model and efficiently employs Depth information to assist RGB tracking via model fine-tuning.
- We introduce a data pathway called Depth Adapt Model (DAM), capable of efficiently extracting highly relevant features from RGBD data, thereby reducing the introduction of interference information caused by direct overlay of Depth features onto RGB features.
- The proposed VADT achieves state-of-the-art performance on two popular RGBD tracking benchmarks, including DepthTrack and VOT-RGBD2022.

2. METHOD

In this work, our proposed VADT adapts a pre-trained foundation model in the RGB domain for RGBD tasks. The overall architecture of our VADT is depicted in Fig.2.

2.1. RGB-based Foundation Model

The RGB-based model we employ takes inputs in the form of two pairs of images, denoted as RGB search patches $x_{rgb} \in \mathbb{R}^{3\times H_x\times W_x}$, RGB template patches $t_{rgb} \in \mathbb{R}^{3\times H_t\times W_t}$, depth search patches $x_{depth} \in \mathbb{R}^{3\times H_x\times W_x}$, and depth template patches $t_{depth} \in \mathbb{R}^{3\times H_t\times W_t}$, respectively. Subsequently, employing a linear projection layer, we project x_{rgb}, t_{rgb}, x_{depth} and t_{depth} into a D-dimensional latent space and incorporate a one-dimensional positional P embedding to generate the ultimate RGB template feature embedding $\boldsymbol{T}^0_{RGB}$, RGB search region embedding $\boldsymbol{X}^0_{RGB}$, depth template feature embedding $\boldsymbol{T}^0_{Depth}$, depth search region embedding $\boldsymbol{X}^0_{Depth}$. We feed the RGB template patch $\boldsymbol{T}^0_{RGB}$ and RGB search patch $\boldsymbol{X}^0_{RGB}$ into the L-layer RGB-based module encoder, resulting in $\boldsymbol{T}^L_{RGB}$ and $\boldsymbol{X}^L_{RGB}$. We refer readers to OSTrack[10] for more details about our RGB-based foundation model.

2.2. Fusion Model

To enhance the richness of depth features and achieve complementarity between the RGB and depth modalities, we propose a simple yet effective fusion module as Fig.3. Firstly, we project RGB patches $\boldsymbol{T}^C_{RGB}$, $\boldsymbol{X}^C_{RGB}$ and RGBD patches $\boldsymbol{T}^C_{Depth}$, $\boldsymbol{X}^C_{Depth}$ through convolutional projections into a lower-dimensional latent embedding. Subsequently, a spatial concave operation is performed to fuse RGB features with depth features, facilitating the complementary internal representation of multimodal features through filtering and adaptation. Specifically, this involves applying λ-smoothed space across all spatial dimensions of modal features, followed by applying channel-level spatial attention masks for inter-modal interaction. Finally, we utilize convolutional operations to restore the embeddings to their original dimensions, generating corresponding patches for the mixed modality F^C_t and F^C_x. The process can be written as Eq.1:

$$\begin{aligned}
\boldsymbol{F}^c_t &= Conv\left(T^C_{Depth}\right) \oplus Softmax\left(\lambda \cdot Conv\left(T^C_{RGB}\right)\right), \\
\boldsymbol{F}^c_x &= Conv\left(X^C_{Depth}\right) \oplus Softmax\left(\lambda \cdot Conv\left(X^C_{RGB}\right)\right), \\
\boldsymbol{F}^C_t &= Conv\left(F^c_t\right), \\
\boldsymbol{F}^C_x &= Conv\left(F^c_x\right),
\end{aligned} \tag{1}$$

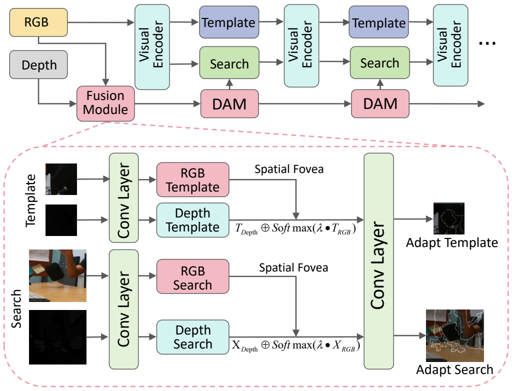

Fig. 3. The proposed Fusion Model structure. This module can achieve complementary features between RGB mode and deep mode.

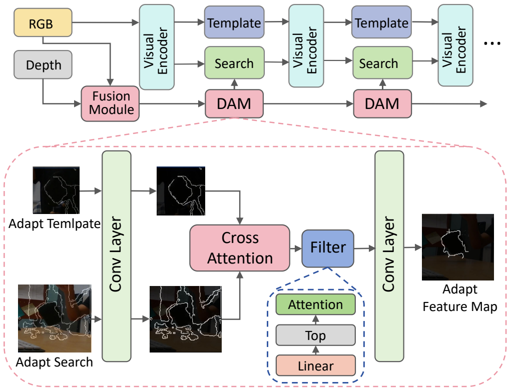

Fig. 4. The proposed Depth Adapt Model structure. This data pathway can extract key information from deep features and filter interference information.

Where c represents the number of channels projected to the low-dimensional latent embedding, with the value of 8, the value of C is 768, and λ denotes a learnable parameter.

2.3. Depth Adapt Model

Existing prompt-based RGBD trackers adopt a direct overlay of depth modal features onto RGB modal features as inputs to RGB-based models. However, this approach introduces a significant amount of noise into the sample space. Our design aims to minimize the introduction of interfering information to the RGB-based model and enhance its transferability. We propose a data pathway that operates in parallel with the Transformer Encoder layer of the RGB-based model as Fig.4. Our Depth Adapt Model, introduced alongside, takes the fused features from the fusion module as input. Initially, these mixed features are projected into a lower-dimensional latent embedding, yielding low-dimensional mixed features. These features are subjected to cross-attention to facilitate depth feature interaction, thus obtaining latent interaction embeddings. This process can be articulated as Eq.2:

$$\boldsymbol{H}_f^c = Atten\left(Conv\left(F_x^C\right), Conv\left(F_t^C\right)\right), \tag{2}$$

Subsequently, by utilizing a depth feature filter, interference between modalities is suppressed within the depth interaction, resulting in depth interaction features. This procedure can be expressed as Eq.3:

$$\begin{aligned} \boldsymbol{H}_{linear}^c &= Linear\left(H_f^c\right), \\ \boldsymbol{H}_{adapt}^c &= Top\left(H_{linear}^c \otimes Softmax\left(H_{linear}^c\right)\right), \end{aligned} \tag{3}$$

Finally, the obtained depth interaction features are mapped back to the original dimensional space, followed by a linear summation with the output RGB search frame features from the Transformer Encoder layer of the RGB-based model. This procedure can be expressed as Eq.4:

$$\boldsymbol{X}_{RGB}^{n+1} = Conv\left(H_{adapt}^c\right) \oplus X_{RGB}^n, \tag{4}$$

3. EXPERIMENTS

3.1. Experimental Settings

Our model was implemented in the Pytorch framework on a server with 1 NVIDIA V100 GPU. Our model is trained with 60 epochs, each epoch with 60,000 image pairs and each mini-batch with 64 sample pairs. To contrast the efficacy of our proposed approach, and to isolate the effect improvement resulting from model differences, we employ the OSTrack, a method for migrating to the RGB modality, as our RGB-based model in parallel with ViPT. We utilize the DepthTrack[11] training dataset as our exclusive training data source.

3.2. Main Properties and Analysis

DepthTrack. The DepthTrack[11] dataset is a comprehensive benchmark commonly used in the computer vision field, specifically for RGB-D object tracking. The DepthTrack[11] dataset uses precision (Pr) and recall (Re) to measure the accuracy and robustness of target localization. F-score, calculated by $F_{score} = \frac{2RePr}{Re+Pr}$, is the primary measure. We compare the proposed VADT with the most existing RGBD trackers, and the results are shown in Tab.1. On the DepthTrack dataset, VADT achieves an F-score of 0.610, a Recall (Re) score of 0.603, and a Precision (Pr) score of 0.606, surpassing all existing RGBD trackers.

VOT-RGBD2022. VOT-RGBD2022[19] is an important evaluation standard for contemporary RGB-D tracking. The performance of our proposed VADT on the VOT-RGBD2022 dataset is presented in Tab.1. VADT achieves an EAO score of 0.721, an Accuracy score of 0.816, and a Robust score of 0.873, demonstrating state-of-the-art performance.

Table 1. We compare our model with state-of-the-art models on the DepthTrack and VOT-RGBD2022 datasets. The best result is highlighted in red.

Type	Method	Published	DepthTrack			VOT-RGBD2022		
			F-score	Recall	Precision	EAO	Accuracy	Robust
RGB to RGBD Module	ATOM[12]	CVPR2019	-	-	-	0.505	0.698	0.688
	STARK_RGBD[13]	ICCV2021	-	-	-	0.647	0.803	0.798
	TransT[2]	CVPR2021	0.484	0.494	0.489	-	-	-
	OSTrack[10]	ECCV2022	0.529	0.522	0.536	0.676	0.803	0.833
	SBT_RGBD[14]	CVPR2022	-	-	-	0.708	0.809	0.864
RGBD Full Fine-tuning Module	SiamM_Ds[15]	ICCVW2019	0.336	0.264	0.463	-	-	-
	ATCAIS[16]	ECCVW2020	0.476	0.455	0.512	0.559	0.761	0.739
	DDiMP[16]	ECCVW2020	0.485	0.469	0.503	-	-	-
	DRefine [17]	ICCVW2021	-	-	-	0.592	0.775	0.760
	DeT[11]	ICCV2021	0.532	0.506	0.560	0.657	0.760	0.845
	SPT[18]	AAAI2023	0.538	0.549	0.527	0.651	0.798	0.851
Prompt-tuning Module	ProTrack[8]	ACMMM2022	0.578	0.573	0.583	0.651	0.801	0.802
	ViPT[9]	CVPR2023	0.594	0.596	0.592	**0.721**	0.815	0.871
	VADT	**Ours**	**0.610**	**0.603**	**0.606**	**0.721**	**0.816**	**0.873**

Table 2. For the ablation experiment of the proposed Noise Feature Filter.

Method	DepthTrack		
	F-score	Recall	Precision
No Filter	0.589	0.594	0.592
Add Filter (VADT)	**0.610**	**0.603**	**0.606**

Table 3. For the ablation experiment of the proposed Prompt Tuning Training.

Method	DepthTrack		
	F-score	Recall	Precision
Foundation (OSTrack)	0.529	0.522	0.536
ViPT	0.594	0.596	0.592
Full Fine-tuning VADT	0.554	0.558	0.555
VADT	**0.610**	**0.603**	**0.606**

3.3. Ablation Study

Noise Feature Filter. To validate the efficacy of our introduced filters, we conduct experiments without their application, and the results are shown in Tab.2. When the filters are not employed, this observation strongly underscores the substantial dissimilarities between Depth and RGB features due to their inherent characteristics, with Depth features exhibiting a notable amount of interference relative to RGB features. Our proposed filters effectively address this issue.

Prompt Tuning Training. Our proposed VADT employs the technique of Prompt Tuning for model fine-tuning. To verify the enhancement of the model's transferability in RGBD tasks, we conducted comparisons between our proposed VADT and the baseline model, the Full Fine-tuning VADT model, and the ViPT model, as presented in Tab.3. It's worth noting that the Full Fine-tuning VADT model experienced a decrease of 5.6% in F-score score, 4.5% in Recall (Re) score, and 5.1% in Precision (Pr) scores on the Depth Track test set. These comparative results strongly underscore that our VADT significantly bolsters the transferability of the RGB-based model for RGBD tasks.

4. CONCLUSION

In this work, we introduce a Prompt-based RGBD tracking model termed *Visual Adapt for RGBD Tracking (VADT)*. We enhance existing Prompt-based RGBD tracking models by designing an efficient data pathway. On the one hand, VADT employs a data pathway structure to facilitate Depth features aiding RGB features for tracking, while retaining RGB imagery as the input modality to enhance the transferability of the RGB-based model. On the other hand, VADT incorporates a novel Depth Adapt Module that efficiently extracts crucial Depth features, thereby mitigating the introduction of noise. Through an extensive set of experiments, we validate the superior performance of the proposed VADT and assess the effectiveness of each constituent component.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.U23A20383 and U21A20474), the Project of Guangxi Science and Technology (No.2022GXNSFDA035079), the Guangxi ”Bagui Scholar” Teams for Innovation and Research Project, the Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, the Guangxi Talent Highland Project of Big Data Intelligence and Application, and the Research Project of Guangxi Normal University (No.2022TD002).

6. REFERENCES

[1] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu, "Mixformer: End-to-end tracking with iterative mixed attention," *Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

[2] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu, "Transformer tracking," *Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

[3] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, Rongrong Ji, Zhenjun Tang, and Xianxian Li, "Siamban: Target-aware tracking with siamese box adaptive network," *IEEE Transactions on Pattern Analysis and Machine Intelligence,TPAMI*, 2022.

[4] Massimo Camplani, Sion Hannuna, Majid Mirmehdi, Dima Damen, Adeline Paiement, Lili Tao, and Tilo Burghardt, "Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling," in *Procedings of the British Machine Vision Conference 2015*, Jan 2015.

[5] Sion Hannuna, Massimo Camplani, Jake Hall, Majid Mirmehdi, Dima Damen, Tilo Burghardt, Adeline Paiement, and Lili Tao, "Ds-kcf: A real-time tracker for rgb-d data," *Journal of Real-Time Image Processing*, vol. 16, no. 5, pp. 1439–1458, Oct 2019.

[6] Uğur Kart, Joni-Kristian Kämäräinen, and Jiří Matas, *How to Make an RGBD Tracker*, p. 148–161, Jan 2019.

[7] Yanlin Qian, Song Yan, Alan Lukezic, Matej Kristan, Joni-Kristian Kämäräinen, and Jiri Matas, "DAL: A deep depth-aware long-term tracker," in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. 2020, pp. 7825–7832, IEEE.

[8] Jinyu Yang, Zhe Li, Feng Zheng, Aleš Leonardis, and Jingkuan Song, "Prompting for multi-modal tracking," *In Proceedings of the 30th ACM International Conference on Multimedia (ACMMM)*, Jul 2022.

[9] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu, "Visual prompt multi-modal tracking," *2023 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Mar 2023.

[10] Botao Ye, Hong Chang, Bingpeng Ma, and Shiguang Shan, "Joint feature learning and relation modeling for tracking: A one-stream framework," *European Conference on Computer Vision,ECCV*, 2022.

[11] Song Yan, Jinyu Yang, Jani Kapyla, Feng Zheng, Ales Leonardis, and Joni-Kristian Kamarainen, "Depthtrack: Unveiling the power of rgbd tracking," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[12] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg., "Atom: Accurate tracking by overlap maximization," *Conference on Computer Vision and Pattern Recognition, CVPR*, 2019.

[13] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu, "Learning spatio-temporal transformer for visual tracking," *International Conference on Computer Vision, ICCV*, 2021.

[14] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng, "Correlation-aware deep tracking," *Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

[15] Jiri Matas et al Matej Kristan, "The seventh visual object tracking vot2019 challenge results," *In ICCVW*, 2019.

[16] Ale˘s Leonardis et al. Matej Kristan, "The eighth visual object tracking vot2020 challenge results," *In ECCVW*, 2020.

[17] Ales Leonardis et al.l Matej Kristan, Jiri Matas, "The ninth visual object tracking vot2021 challenge results," *In ICCVW*, 2021.

[18] Zhangyong Tang et al. Xue-Feng Zhu, Tianyang Xu, "Rgbd1k: A large-scale dataset and benchmark for rgb-d object tracking," *Association for the Advancement of Artificial Intelligence,AAAI*, 2023.

[19] Ale˘s Leonardis et al Matej Kristan, "Transformer trackingthe tenth visual object tracking vot2022 challenge results," 2022.