RGB-D Video Object Segmentation via Enhanced Multi-store Feature Memory

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

Ruichao Hou*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
rc_hou@smail.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 RGB-Depth (RGB-D) Video Object Segmentation (VOS) aims to integrate the fine-grained texture information of RGB with the spatial geometric clues of depth modality, boosting the performance of segmentation. However, off-the-shelf RGB-D segmentation methods fail to fully explore cross-modal information and suffer from object drift during long-term prediction. In this paper, we propose a novel RGB-D VOS method via multi-store feature memory for robust segmentation. Specifically, we design the hierarchical modality selection and fusion, which adaptively combines features from both modalities. Additionally, we develop a segmentation refinement module that effectively utilizes the Segmentation Anything Model (SAM) to refine the segmentation mask, ensuring more reliable results as memory to guide subsequent segmentation tasks. By leveraging spatio-temporal embedding and modality embedding, mixed prompts and fused images are fed into SAM to unleash its potential in RGB-D VOS. Experimental results show that the proposed method achieves state-of-the-art performance on the latest RGB-D VOS benchmark.

CCS CONCEPTS

• **Computing methodologies** → **Video segmentation**; *Artificial intelligence*; *Computer vision.*

KEYWORDS

RGB-Depth, Video Object Segmentation, Memory Mechanism, Segment Anything Model

*Corresponding author.

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06
https://doi.org/10.1145/3652583.3658036

ACM Reference Format:
Boyue Xu, Ruichao Hou, Tongwei Ren, and Gangshan Wu. 2024. RGB-D Video Object Segmentation via Enhanced Multi-store Feature Memory. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652583.3658036

1 INTRODUCTION

Video object segmentation (VOS) [43] aims to continuously segment specific object masks given in the first frame throughout the video sequence, which has numerous applications in autonomous driving [19, 23, 33], 3D reconstruction [20, 26, 44], surveillance [36, 45].

RGB VOS methods face various challenges in scenarios such as extreme illumination, complex backgrounds, and occlusion. To address these issues, RGB-Depth (RGB-D) VOS [46] introduces depth modality which offers additional spatial geometric clues for more robust segmentation. However, as shown in Figure 1(a), RGB-D VOS methods which use template to guide fusion and segmentation may encounter challenges in complex scenarios and initial template may not guide the subsequent segmentation in long-term videos, decreasing the robustness of segmentation.

Memory mechanism is used to address these issues in RGB VOS [4, 5, 14, 21, 41, 42], which mine the spatio-temporal information and appearance features from previous segmentation results to guide subsequent segmentation. Compared to traditional approaches, memory mechanism effectively addresses the problem of poor feature representation caused by continuous propagation relying solely on adjacent frames. Among all memory-based methods, XMem [4] achieves satisfactory results by constructing a multi-store memory network inspired by Atkinson-Shiffrin memory model [9]. It consists of sensory memory, working memory and long-term memory that significantly improves the robustness of segmentation through effective insertion of memory content. XMem effectively inserts memory content, significantly improving the robustness of segmentation.

Inspired by XMem, we propose an enhanced multi-store memory network for RGB-D VOS. However, there are two challenges that need to be addressed: (1) How to adaptively fuse complementary information from RGB-D modalities? (2) How to estimate the

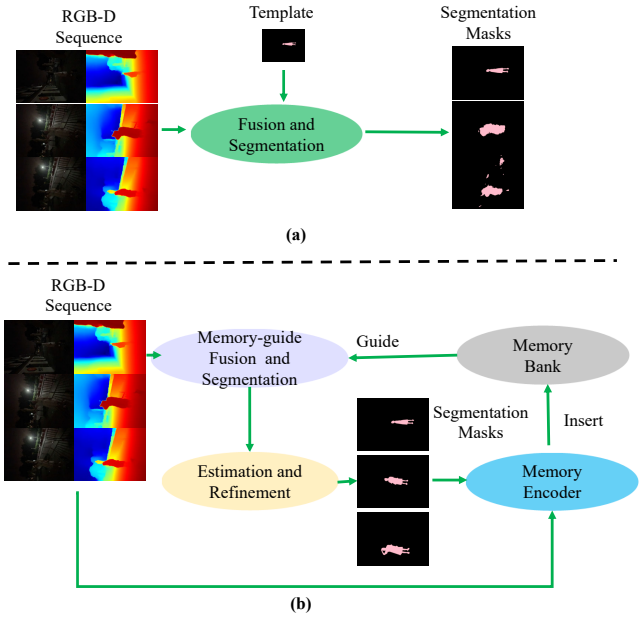

Figure 1: Comparison of the framework between different RGB-D VOS methods. (a) RGB-D VOS methods without memory, which use template to guide fusion and segmentation. (b) The proposed method which use memory to guide fusion and segmentation.

quality and reliability of inserted segmentation results as guidance for accurate subsequent segmentation?

As shown in Figure 1(b), the hierarchical modality selection and fusion (HMSF) is proposed to fuse RGB-D features under the guide of memory. HMSF is capable of extracting the complementary features from both modalities and selectively fusing multi-modal features, making it suitable for RGB-D feature fusion during segmentation and content encoding in a multi-store memory network. To enhance the reliability of segmentation results, we attempt to leverage the powerful segmentation capabilities of the Segment Anything Model (SAM) [18] for refinement. To the best of our knowledge, we are the first to introduce SAM into the RGB-D VOS. Specifically, we take the fused image and the mixed prompt as the input of SAM. This paper focuses on generating pixel-wise RGB-D fused images and effective prompts to reduce the false positive pixels or regions, thereby guiding more accurate segmentation. To this end, we propose spatio-temporal embedding and modality embedding to further improve performance. The spatio-temporal embedding integrates historical information from the entire video segmentation process to provide prompts for SAM, while modality embedding effectively combines RGB-D images through a pixel-wise fusion strategy. The fused weights are calculated by the quality of modality from HMSF and the significance of the region to facilitate the segmentation performance and refine the memory content. Experimental results demonstrate that the proposed method achieves the best results on the latest RGB-D VOS dataset.

The main contributions of this paper are summarized as follows:

- We propose a novel RGB-D VOS method based on multi-store feature memory, enabling robust segmentation of RGB-D sequences in diverse and complex scenarios.
- We present the HMSF to effectively fuse hierarchical and cross-modality features. Also, it can facilitate the encoding of memory.
- We develop a segmentation refinement module that incorporates spatio-temporal embedding and modality embedding with SAM, significantly enhancing the reliability of segmentation results.

2 RELATED WORK

2.1 Semi-supervised VOS

Video object segmentation encompasses unsupervised VOS [8, 32], semi-supervised VOS [4, 21, 38, 42], and interactive VOS [3, 24]. Semi-supervised VOS annotates the object masks in the first frame to guide the segmentation of the rest frames. The semi-supervised VOS methods can be categorized into motion-aware VOS methods and detection-aware VOS methods.

Motion-aware VOS methods utilize optical flow or CNN-based learning for mask refinement. Optical flow-based methods employ motion cues to estimate pixel changes over time [6, 17], while CNN-based learning methods refine the object mask frame-to-frame by leveraging temporal information [28] or combining RNNs [37] with current mask predictions. Although these methods show promising performance, they still lack robustness when dealing with challenging attributes, *i.e.*, occlusion or fast motion in long-term videos.

Detection-aware VOS methods primarily involve learning an appearance model for pixel-wise object detection and segmentation. For instance, Caelles *et al.* [2] employ fully convolutional neural networks on static images to segment objects by fine-tuning the first frame of the video sequence. Some other methods [15, 29] employ pixel-matching techniques to segment objects based on feature matching with a template. Yang *et al.* [41, 42] embed multiple objects into the same embedding space, then uniformly and simultaneously propagate all object embedding. Xmem [4], inspired by human memory mechanisms, offers an innovative propagation method, avoiding the out-of-memory crash. These methods perform well in handling occlusion or fast motion scenarios. However, distinguishing cross-temporal similar objects remains challenging.

In RGB-D VOS, Zhao *et al.* [46] pioneer the emerging area by introducing the first benchmark known as ARKitTrack, which opens up a new frontier. However, there is room for decreasing the complexity of the fusion model and improving its performance in long-term segmentation. We propose a multi-store feature memory for robust RGB-D segmentation.

2.2 RGB-D Video Object Tracking

As RGB-D VOS is a relatively new domain with limited prior work, it is beneficial to provide an introduction to the related field of RGB-D tracking [27, 36, 40, 49, 50], which shares a similar processing scheme.

RGB-D tracking is a type of multi-modality object tracking[10, 12, 13, 34, 35]. Song *et al.* [30] establish the first RGB-D tracking

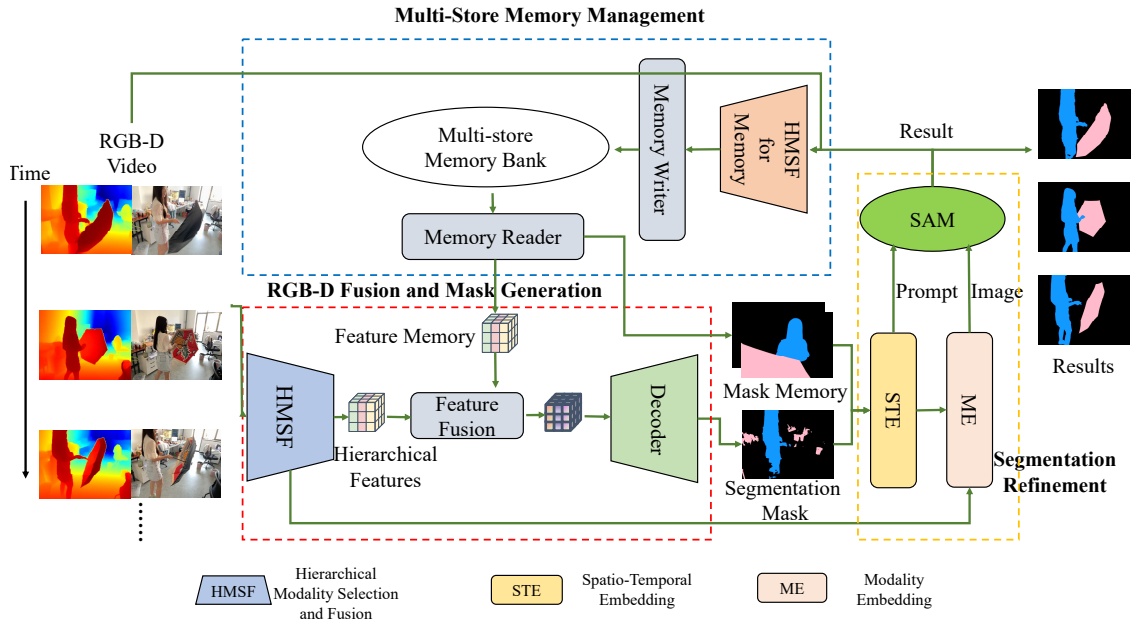

Figure 2: The framework of the proposed method, consists of RGB-D fusion and mask generation, segmentation refinement and multi-store memory management.

dataset, pioneering the task of RGB-D tracking. Zhong *et al.* [48] introduce a method for extracting key points from dense depth maps for use in RGB-D tracking. Bibi *et al.*[1] employ appearance models and 3D spatial motion models to estimate object positions. Kart *et al.* [16] integrate discriminative correlation filters (DCF) to propose a general framework for RGB-D tracking. Qian *et al.* [27] address the occlusion challenge by embedding depth into deep features and training discriminators. Zhao *et al.* [47] employ depth to generate object masks and accurately cut out prediction results. Yan *et al.* [39] explore pseudo-color images derived from depth data and investigate the integration with RGB images. With the advancement of transformer, Zhu *et al.* [50] incorporate transformers to fuse multi-modal features. Moreover, some researchers [40, 49] apply prompt learning to adapt existing RGB trackers for RGB-D tracking.

3 METHODOLOGY

3.1 Network Architecture

As illustrated in Figure 2, the proposed method comprises three modules: RGB-D fusion and mask generation module, segmentation refinement module, and multi-store memory management module. Specifically, the RGB-D fusion and mask generation module aims to fuse RGB-D dual-modality features and integrate them with multi-store feature memory to produce segmentation results. The segmentation refinement module flexibly utilizes the SAM [18] to refine segmentation results and ensure more accurate results as memory to guide subsequent segmentation. The multi-store memory management module encodes and stores both RGB-D images and segmentation results as feature memory.

The proposed method utilizes dual-modality image sequences as input, which are initially processed by the HMSF for feature extraction and fusion. The output from HMFS is then combined with the feature memory in the multi-store memory bank, followed by decoding to obtain accurate segmentation results. To further improve the segmentation performance and guide the subsequent segmentation, we incorporate SAM [18] to refine segmentation masks. In particular, we estimate the reliability of segmentation masks and generate mixed prompts for SAM to maintain the consistency of the given prompts via spatio-temporal embedding. Next, we fuse the RGB-D images using a pixel-wise strategy by using modality embedding. Subsequently, the refined masks are encoded via the HMSF for memory and inserted into the multi-store memory bank that captures historical information of segmentations, thus preventing error accumulation in long-term videos.

3.2 RGB-D Fusion and Mask Generation

To extract and integrate hierarchical complementary features from different modalities, we propose hierarchical modality selection and fusion. As shown in Figure 3(a), the HMSF firstly extracts features of different modalities using ResNet50 [11]. Then it adaptively fuses dual-modality hierarchical features across layer 1 to layer 3 of ResNet50 via modality selection and fusion, which combines the shallow features and deep features to enhance the multi-modality feature representation. The fused hierarchical features F_4, F_8, and F_{16} are utilized for segmentation.

The details of modality selection and fusion are shown in Figure 4. Initially, the RGB and depth features from the current layer undergo global average pooling to extract global features.

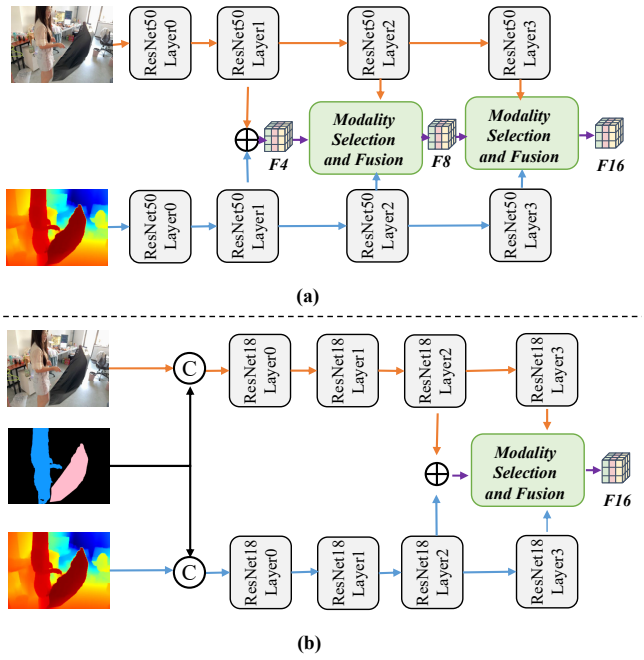

Figure 3: The detail of hierarchical modality selection and fusion.(a) Hierarchical modality selection and fusion used in RGB-D fusion and mask generation module. (b) Hierarchical modality selection and fusion for memory used in multi-store memory management module.

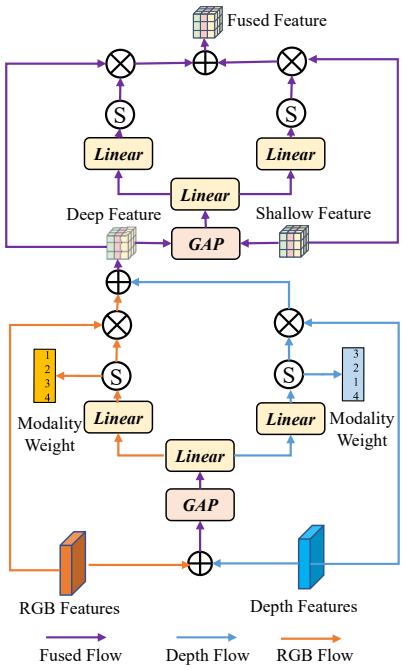

Figure 4: The details of modality selection and fusion.

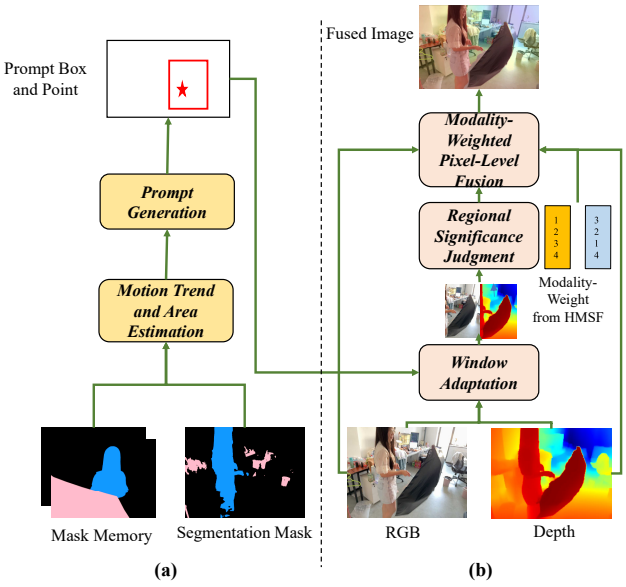

Figure 5: The details of segmentation refinement. (a) The details of spatio-temporal embedding, which generates mixed prompts. (b) The details of modality embedding, which generates fused images.

These global features are then distributed through the RGB and depth flows to generate channel-level modality weights, which are crucial for selectively fusing dual-modality features and guiding pixel-wise fusion within the segmentation refinement module. Next, the modality weights are multiplied by the corresponding feature channels to obtain discriminative features. Deep features are derived by summing the above discriminative features, which is formulated as follows:

$$F_{global} = FC\left(GAP\left(F_{RGB} \oplus F_D\right)\right), \tag{1}$$

$$\hat{W}_i = \sigma\left(FC_i\left(F_{global}\right)\right), \quad i \in \{RGB, D\}, \tag{2}$$

$$\hat{F}_i = F_i \otimes W_i, \quad i \in \{RGB, D\}, \tag{3}$$

$$F_{deep} = \hat{F}_{RGB} \oplus \hat{F}_D, \tag{4}$$

where FC$(\cdot)$ is the linear layer; GAP$(\cdot)$ represents global average pooling; F_{RGB}, F_D, W_{RGB}, W_D represent RGB features, depth features, RGB weights, depth weight respectively; $\oplus$ denotes element-wise addition and $\otimes$ denotes the element-wise multiplication operation.

Subsequently, both shallow features and deep features are fused separately using the same way. Modality selection and fusion aim at fusing hierarchical features from multi-scale layers to expand the receptive field and thus enhance the representational power of the fused features.

Once obtaining the hierarchical features, we retrieve the best-matched features from the memory bank and merge them with the hierarchical features for decoding to generate segmentation masks. Specifically, the features stored in the memory are concatenated with F_{16}. Subsequently, these hierarchical features are decoded

into a segmentation mask using a series of upsampling operations. The decoder and memory retrieval operations follow XMem [4].

3.3 Segmentation Refinement

To improve the reliability of segmentation results and guide subsequent segmentation, we propose a segmentation refinement module that flexibly leverages SAM. This module incorporates spatio-temporal embedding to estimate the quality of segmentation, generating mixed prompts for guiding SAM. Moreover, it employs modality embedding to evaluate the reliability of different modalities, thereby generating pixel-wise fused RGB-D images that are enriched with cross-modal information for SAM.

Spatio-Temporal Embedding. The spatio-temporal embedding leverages historical segmentation results to estimate the reliability of current segmentation and generate mixed prompts, which are composed of box prompt and point prompt. While the box-only prompt can designate the object area, it often struggles to accurately specify the object in complex backgrounds. On the other hand, the point-only prompt cannot precisely predict the scale of objects. To bridge this gap between these two types of prompts, the mixed prompt capitalizes on their respective strengths, unleashing SAM's potential for accurate segmentation.

As shown in Figure 5(a), spatio-temporal embedding considers historical spatial trends of the objects, which involves motion trends and area estimation. Deviation from the expected motion trend in the current segmentation result indicates a potential error in object positioning. In such cases, the cluster center from the previous mask memory is used to generate the point prompt. Conversely, if the current position is deemed highly reliable, the cluster center of the current mask serves as the point prompt. Specifically, we define a threshold for significant object shift denoted as M. When the deviation between the current position and the historical trend exceeds M, it implies a potential error prompting the generation of point prompts using stored memory content. Conversely, if the deviation falls within M, the point prompt is generated using the cluster center of the current position. Additionally, area estimation is performed, if there is a considerable disparity in mask area between the segmentation mask and mask memory, it may indicate unreliable segmentation and we generate the box prompt using the outer enclosing box of mask memory. Otherwise, the outer enclosing box of the current segmentation mask is used as box prompt. The mixed prompt consisting of both box and point information is then fed into SAM to provide guidance for modality embedding and generate reliable dual-modality images.

Modality Embedding. The modality embedding explores the collaborative and heterogeneous nature of different modalities, integrating RGB-D complementary information and providing SAM with pixel-wise fused RGB-D images. In complex scenarios, the fused weights learned from the training set may yield suboptimal fused results. Therefore, we enhance the reliability of the fused weights by combining regional significance assessment with the modality weights of HMSF.

As illustrated in Figure 5(b), we first crop the dual-modality images to an appropriate size based on the box prompt from the spatio-temporal embedding to extract the region of interest and minimize excessive background interference. Subsequently, we estimate the significance of the depth image within the cropped region. If the significance rate is low, incorporating depth modality might introduce segmentation errors. Conversely, if the significance rate is high, we can employ the modality weights obtained from the HMSF for pixel-wise fusion. The regional significance assessment involves calculating the entropy of color distribution in a pseudo-color transformed version of the depth image, which can be calculated as follows:

$$E = -\sum_{i=0}^{255}\sum_{j=0}^{255}\sum_{k=0}^{255} H[i,j,k] \cdot \log_2(H[i,j,k]), \tag{5}$$

where E represents the value of entropy, with a higher value indicating a higher significance rate; $H[i,j,k]$ denotes the proportion of pixels in the color space with red, green, and blue channel values of i, j, and k, respectively. After estimating the significance of the region, we determine whether to fuse depth information. Subsequently, using the modality weights obtained from the HMSF, we perform a weighted element-wise addition of the dual-modality images. This aims to meet the input requirement of the SAM while also maintaining effective fusion of the modalities [31]. Ultimately, this process provides the SAM with a fused image for segmentation refinement.

3.4 Multi-Store Memory Management

Memory mechanism in VOS facilitates the extraction of appearance features and establishment of spatio-temporal connections from previous segmentation results. To enhance the robustness of RGB-D VOS, we adopt the multi-store memory model introduced by XMem [4], modifying it to suit the requirements of dual-modalities.

The multi-store memory model, inspired by Atkinson-Shiffrin memory model [9], divides memory into sensory memory, working memory, and long-term memory. The network encodes images and segment results at distinct frequencies and stores them in each feature memory. During retrieval, a similarity matrix is computed based on the current frame's features to match relevant content from the multi-store memory.

To enhance the reliability of segmentation masks as memory by encoding complementary features of RGB-D images, we propose the HMSF for memory as the memory encoder. The proposed method utilizes modality selection and fusion to encode the refined masks from the segmentation refinement module along with RGB-D images as the memory content. As shown in Figure 3(b), we employ the ResNet18 [11] as the backbone for encoder. The refined segmentation results are concatenated with RGB and depth images respectively, and the modality selection and fusion are used to extract features from layer 2 and layer 3. These fused features are stored in memory according to the following formula:

$$M = Fu((RGB_2 \oplus D_2), RGB_3, D_3), \tag{6}$$

where $Fu(\cdot)$ represents modality selection and fusion; M represents the memory features; RGB_2, D_2, RGB_3, and D_3 denote the second and third layer features of RGB and depth, respectively. This way allows the multi-store memory to adapt to the multi-modal memory content of RGB-D, effectively preventing errors in memory content that could mislead subsequent segmentation.

3.5 Loss Function

Following XMem [4], we employ bootstrapped cross-entropy loss and dice loss as loss functions [25]. The bootstrapped cross-entropy loss can be calculated as:

$$\mathcal{L}_{bce} = \frac{1}{|S_l|} \sum_{m_i^l, g_i \in S_l} \left\{ F\left(m_i^l\right) < \eta \right\} \mathbf{C}\left(g_i, F\left(m_i\right)\right), \tag{7}$$

where $F(\cdot)$ is the output probability for a labeled example m_i ; g_i is ground truth, and $C(\cdot)$ represents the cross-entropy loss. The dice loss can be calculated as:

$$\mathcal{L}_d = 1 - \frac{2|m \cap g|}{|m| + |g|}, \tag{8}$$

where m and g represent predict mask and ground truth mask, respectively. The total loss is calculated as the sum of the bootstrapped cross-entropy loss and the dice loss:

$$\mathcal{L}_{total} = \mathcal{L}_{bce} + \mathcal{L}_d. \tag{9}$$

4 EXPERIMENTS

4.1 Datasets and and Metrics

The ARKitTrack [46] is the most recent RGB-D VOS dataset, which is utilized for comparative experiments against state-of-the-art methods. This dataset consists of more than 200 pairs of RGB-D sequences of 1920 × 1440 resolution which collected in real-world scenarios. Each video sequence contains synchronized and aligned RGB frames and depth maps. The sequences encompass various challenging scenarios such as similar objects, occlusions, and extreme illumination.

We use $\mathcal{J}_\mathcal{M}$, $\mathcal{F}_\mathcal{M}$, and $\mathcal{J}\&\mathcal{F}$ measure [43] as evaluation metrics used in these experiments. Specifically, $\mathcal{J}_\mathcal{M}$ denotes region similarity and is calculated as the intersection over union (IoU) between the predicted object segmentation mask and the ground truth. It can be calculated using the following formula:

$$\mathcal{J}_\mathcal{M} = \frac{M \cap G}{M \cup G}, \tag{10}$$

where M is the predicted mask, while G denotes the ground truth. $\mathcal{F}_\mathcal{M}$ stands for contour accuracy which is calculated based on the precision and recall of the contour. It can be calculated as follows:

$$\mathcal{F}_\mathcal{M} = \frac{2P_c R_c}{P_c + R_c}, \tag{11}$$

where P_c represents precision and R_c denotes recall. The metric $\mathcal{J}\&\mathcal{F}$ is the average of $\mathcal{J}_\mathcal{M}$ and $\mathcal{F}_\mathcal{M}$:

$$\mathcal{J}\&\mathcal{F} = \frac{\mathcal{J}_\mathcal{M} + \mathcal{F}_\mathcal{M}}{2}. \tag{12}$$

4.2 Implementation Details

The proposed model is trained on a server equipped with a 5.2GHz CPU and four 3090 GPUs with a total of 96GB of memory. During training, we employ the AdamW optimizer [22] with a learning rate set to $1e^{-5}$ and a batch size of 8. The model undergoes a total of 120K iterations on the training set. All other training parameters are consistent with the baseline [4]. We utilize the ViT-H version weights of SAM [18]. Furthermore, the threshold for regional significance E is set to 6, while the threshold for significant object shift M is set to 500 pixels. The parameter amount of the proposed method without SAM is 64.9M, when the input images are paired 1920×1440 RGB-D dual-modal images, the inference efficiency is about 1.5 FPS.

Table 1: Comparison results of the proposed method against the competing methods on ARKitTrack test set. The upper section lists RGB methods, and the lower section includes RGB-D methods. The best results are highlighted in bold.

Methods	Year	$\mathcal{J}_\mathcal{M} \uparrow$	$\mathcal{F}_\mathcal{M} \uparrow$	$\mathcal{J}\&\mathcal{F} \uparrow$
STCN [5]	2021	0.489	0.560	0.525
AOT [41]	2021	0.555	0.627	0.582
RPCM [38]	2022	0.492	0.527	0.509
QDMN [21]	2022	0.276	0.337	0.306
XMem [4]	2022	0.541	0.565	0.553
STCN_RGBD [5]	2021	0.498	0.574	0.537
XMem_RGBD [4]	2022	0.617	0.680	0.649
SAMTrack_RGBD [7]	2023	0.445	0.463	0.454
ARKitTrack [46]	2023	0.625	0.698	0.662
Ours	-	**0.673**	**0.723**	**0.698**

4.3 Comparison with the State-of-the-Art

To validate the effectiveness of the proposed method, we conduct comparative experiments with seven state-of-the-art methods, including STCN [5], AOT [41], RPCM [38], QDMN [21], XMem [4], SAMTrack [7], and ARKitTrack [46]. In addition, we modify Xmem and SAMTrack by adding a depth branch, which is then fused with the RGB branch to form Xmem_RGBD and SAMTrack_RGBD. To ensure fairness in the experimental evaluation, all methods are fine-tuned on the training dataset provided by ARKitTrack [46].

The comparison results are presented in Table 1, illustrating that our method achieves the best performance in all three metrics. In comparison to the state-of-the-art method ARKitTrack [46], our method exhibits improvements of 4.8%, 2.5%, and 3.6% in $\mathcal{J}_\mathcal{M}$, $\mathcal{F}_\mathcal{M}$, and $\mathcal{J}\&\mathcal{F}$, respectively. These advancements can be primarily attributed to the enhanced multi-store memory, which effectively extracts appearance features and establishes historical connections of segmentation results, thereby enhancing segmentation reliability significantly. When compared to baseline, XMem [4], the performance of the proposed method has even more substantial improvements with increases of 13.7%, 6.8%, and 14.5% in the three metrics, respectively. The proposed segmentation refinement module is primarily responsible for this improvement, as it effectively corrects segmentation errors and prevents their inclusion in the multi-store memory, thereby avoiding subsequent misleading segmentation.

Compared to SAMTrack, which also utilizes the SAM for segmentation guidance, our method shows notable improvements due to the incorporation of spatio-temporal embedding and modality embedding. The spatio-temporal embedding generates accurately mixed prompts, while the modality embedding fuses informative RGB-D images with a pixel-wise strategy. In contrast, SAMTrack only performs full-image segmentation on a single modality, which limits its efficient utilization of SAM. Moreover, the HMSF extracts and fuses the complementary features from each

Table 2: Ablation study on different components. HMSF represents hierarchical modality selection and fusion, STE and ME denote spatio-temporal embedding and modality embedding respectively, The best results are highlighted in bold.

HMSF	STE	ME	$\mathcal{J}_M \uparrow$	$\mathcal{F}_M \uparrow$	$\mathcal{J}\&\mathcal{F} \uparrow$
			0.617	0.680	0.649
✓			0.637	0.691	0.664
✓	✓		0.651	0.702	0.677
✓	✓	✓	**0.673**	**0.723**	**0.698**

Table 3: Comparison of different thresholds, E represents the threshold for regional significance and M is the threshold for significant object shift, all results in the table are $\mathcal{J}\&\mathcal{F}$. The best result is highlighted in bold.

	E=4	E=6	E=8
M=300	0.673	0.687	0.654
M=500	0.681	**0.698**	0.666
M=700	0.666	0.767	0.659

modality, thus improving the representation ability of features, and further enhancing the performance of RGB-D segmentation. The comparison experiments fully demonstrate the effectivenes of the proposed method.

4.4 Ablation Study

Components Analysis. To further validate the effectiveness of each component, we conduct ablation experiments with four versions of the proposed method.

As shown in Table 2, in the first row, we add depth and RGB directly in the RGB-D fusion and mask generation module, as well as the multi-store memory management module. The corresponding three metrics are 0.617, 0.680, and 0.649, respectively. For the second row, instead of direct addition in RGB-D fusion mask generation and multi-store memory management module, we introduce HMSF for improved performance. The three metrics show respective increases of 2%, 1.1%, and 1.5%. These results indicate that our HMSF effectively integrates dual-modality information and encodes the memory content, increasing the robustness of segmentation in complex scenarios. In the third row, we introduce a spatio-temporal embedding to generate mixed prompts, only RGB modality is fed into SAM. The metrics show improvements of 1.5%, 1.1%, and 1.3%, respectively. This demonstrates that spatio-temporal embedding can enhance SAM's segmentation quality by providing reliable mixed prompts enriched with complementary dual-modalities information. The final row involves the modality embedding, which utilizes RGB-D dual-modalities for segmentation refinement. The final scores reach 0.673, 0.723, and 0.698, respectively. These results demonstrate that the modality embedding employs a pixel-wise fusion strategy based on the fused weights learned from HMSF and region significance. Thus, this process effectively provides informative segmentation images for SAM, enriched with complementary dual-modalities information. The component analysis indicates the effectiveness of each component.

Parameter Analysis. In this section, we conduct an analysis of the parameters set for the proposed method, focusing on the thresholds set for spatio-temporal embedding and modality embedding in segmentation refinement. These thresholds include the threshold for significant object shift M and the threshold for regional significance E.

We conduct a total of nine sets of experiments, varying the value of M at 300 pixels, 500 pixels, and 700 pixels, and E at 4, 6, and 8, respectively. As indicated in Table 3, the best results are achieved when M is set to 500 pixels and E to 6. The parameter analysis suggests that the significant object shift M is primarily used to estimate whether an object is segmented incorrectly based on its spatial change, hence determining whether to generate prompts from memory masks or current masks. If this value is set too low, normal object movements may be misclassified as significant shifts, leading to erroneous generation of prompts from historical information and missing the actual position of objects which can cause segmentation errors. Conversely, if the threshold is set too high, segmentation errors may be ignored thereby missing opportunities for correction.

The regional significance threshold E estimates the amount of information contained in the depth modality through entropy measurement, it determines whether fusion with RGB is necessary. If this value is set too low, it may lead to the integration of potentially misleading depth information with low distinctiveness into the RGB image. On the other hand, setting it too high may disregard informative depth images containing discriminative features. The parameter analysis fully proves the rationality of parameter settings.

Qualitative Analysis. To demonstrate the advantages of the proposed method, we conduct a qualitative analysis, as depicted in Figure 6. Given the lack of test results provided by most existing methods, we primarily employed results visualization to validate the efficacy of the segmentation refinement module.

In the first sequence, the objects are specified as two books with confusing appearances and similar depths in the depth modality. More importantly, their depth closely resembled that of the background which posed a significant challenge for segmentation. Without segmentation refinement, our method initially segments correctly but gradually confuses these objects due to unreliable segmentation results being stored in memory leading to error accumulation and eventual complete confusion between them. However, with segmentation refinement utilizing spatio-temporal embedding and mixed prompt information generation through SAM upon identifying any errors during the segmentation process followed by pixel-wise fusion using modality embedding generates accurate segmentations which are then stored in memory for guiding subsequent tasks.

In the second sequence, the objects are specified as a person and her bag in the frame. The segmentation process is complicated by challenges such as the person's movements, which can occlude the backpack. Additionally, the presence of multiple similar-looking persons in the frame can lead to confusion with the segmentation

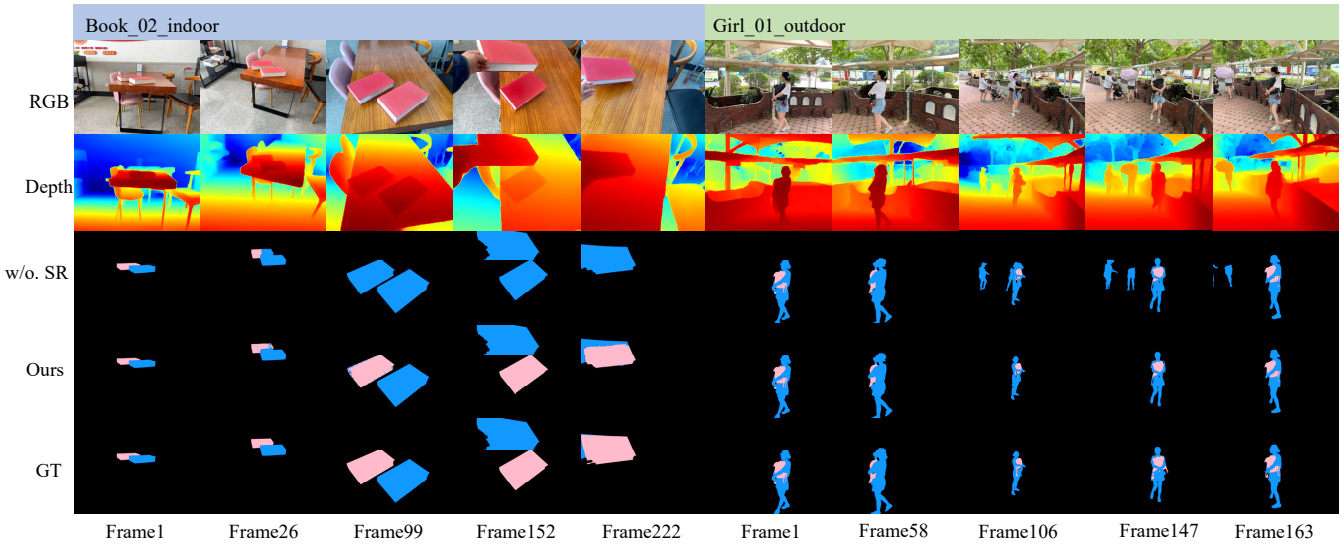

Figure 6: Qualitative comparison of the proposed method in handling different challenging scenarios, the first row shows the RGB images, the second row presents the depth images, the third row depicts the segmentation results without segmentation refinement, the fourth row illustrates the results of the proposed method, and the final row shows the ground truth.

object. The method without segmentation refinement might confuse one person for another. Our method leverages the segmentation refinement module to reliably segmentation and identify each object.

The qualitative analysis visually demonstrates the effectiveness of our method in complex scenarios.

5 CONCLUSION

In this paper, we proposed a novel RGB-D VOS method based on an enhanced multi-store memory. Specifically, we proposed a segmentation refinement module to improve the reliability of segmentation results by incorporating spatio-temporal embedding and modality embedding which provide mixed prompts and pixel-wise fused images for SAM. Additionally, we presented the HMSF for the selective fusion of hierarchical features across both modalities, thereby enhancing the modality interaction and memory encoding. The proposed method achieved state-of-the-art performance on the latest RGB-D VOS dataset ARKitTrack.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Adel Bibi, Tianzhu Zhang, and Bernard Ghanem. 2016. 3D Part-Based Sparse Tracker with Automatic Synchronization and Registration. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[2] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. 2017. One-Shot Video Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[3] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. 2018. Blazingly Fast Video Object Segmentation with Pixel-wise Metric Learning. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[4] Ho Kei Cheng and Alexander G Schwing. 2022. Xmem: Long-term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In *European Conference on Computer Vision.*

[5] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. 2021. Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation. In *Neural Information Processing Systems.*

[6] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. 2017. Segflow: Joint Learning for Video Object Segmentation and Optical Flow. In *IEEE International Conference on Computer Vision.*

[7] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and Track Anything. *arXiv preprint arXiv:2305.06558* (2023).

[8] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. 2017. Fusionseg: Learning to Combine Motion and Appearance for Fully Automatic Segmentation of Generic Objects in Videos. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[9] Hermann Ebbinghaus. 2013. Memory: A contribution to experimental psychology. *Annals of Neurosciences* 20, 4 (2013), 155.

[10] M. Feng and J. Su. 2022. Learning Reliable Modal Weight with Transformer for Robust RGBT Tracking. *Knowledge Based Systems* 249 (2022), 108945.

[11] 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.*

[12] 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.*

[13] 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.*

[14] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. 2021. Learning Position and Target Consistency for Memory-Based Video Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[15] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. 2018. Videomatch: Matching Based Video Object Segmentation. In *European Conference on Computer Vision.*

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

[17] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. 2017. Lucid Data Dreaming for Object Tracking. In *The DAVIS Challenge on Video Object Segmentation.*

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment Anything. In *IEEE International Conference on Computer Vision*.

[19] Jiaqi Li, Yiran Wang, Zihao Huang, Jinghong Zheng, Ke Xian, Zhiguo Cao, and Jianming Zhang. 2023. Diffusion-Augmented Depth Prediction with Sparse Annotations. In *ACM International Conference on Multimedia*.

[20] Caixia Liu, Dehui Kong, Shaofan Wang, Jinghua Li, and Baocai Yin. 2020. DLGAN: Depth-Preserving Latent Generative Adversarial Network for 3D Reconstruction. *IEEE Transactions on Multimedia* 23 (2020), 2843–2856.

[21] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. 2022. Learning Quality-Aware Dynamic Memory for Video Object Segmentation. In *European Conference on Computer Vision*.

[22] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101* (2017).

[23] Zeyu Ma, Yang Yang, Guoqing Wang, Xing Xu, Heng Tao Shen, and Mingxing Zhang. 2022. Rethinking Open-World Object Detection in Autonomous Driving Scenarios. In *ACM International Conference on Multimedia*.

[24] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. 2018. Deep extreme cut: From extreme points to object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[25] Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. Semi-supervised Semantic Segmentation with Cross-Consistency Training. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[26] Zihao Pan, Junyi Hou, and Lei Yu. 2023. Optimization RGB-D 3-D Reconstruction Algorithm Based on Dynamic SLAM. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–13.

[27] 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*.

[28] Gilad Sharir, Eddie Smolyansky, and Itamar Friedman. 2017. Video Object Segmentation Using Tracked Object Proposals. *arXiv preprint arXiv:1707.06545* (2017).

[29] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. 2017. Pixel-Level Matching for Video Object Segmentation Using Convolutional Neural Networks. In *IEEE International Conference on Computer Vision*.

[30] S. Song and J. Xiao. 2014. Tracking Revisited Using RGBD Camera: Unified Benchmark and Baselines. In *IEEE International Conference on Computer Vision*.

[31] Zhangyong Tang, Tianyang Xu, Hui Li, Xiao-Jun Wu, XueFeng Zhu, and Josef Kittler. 2023. Exploring Fusion Strategies for Accurate RGBT Visual Object Tracking. *Information Fusion* 99 (2023), 101881.

[32] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. 2017. Learning video object segmentation with visual memory. In *IEEE International Conference on Computer Vision*.

[33] Xiaoyang Xiao, Yuqian Zhao, Fan Zhang, Biao Luo, Lingli Yu, Baifan Chen, and Chunhua Yang. 2023. BASeg: Boundary Aware Semantic Segmentation for Autonomous Driving. *Neural Networks* 157 (2023), 460–470.

[34] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. 2022. Attribute-based Progressive Fusion Network for Rgbt Tracking. In *AAAI Conference on Artificial Intelligence*.

[35] Boyue Xu, Ruichao Hou, Jia Bei, Tongwei Ren, and Gangshan Wu. 2024. Jointly Modeling Association and Motion Cues for Robust Infrared UAV Tracking. *The Visual Computer* (2024), 1–12.

[36] Boyue Xu, Yi Xu, Ruichao Hou, Jia Bei, Tongwei Ren, and Gangshan Wu. 2023. RGB-D Tracking via Hierarchical Modality Aggregation and Distribution Network. In *ACM International Conference on Multimedia in Asia*.

[37] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. 2018. Youtube-VOS: Sequence-to-Sequence Video Object Segmentation. In *European Conference on Computer Vision*.

[38] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. 2022. Reliable Propagation-Correction Modulation for Video Object Segmentation. In *AAAI Conference on Artificial Intelligence*.

[39] 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*.

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

[41] Zongxin Yang, Yunchao Wei, and Yi Yang. 2021. Associating Objects with Transformers for Video Object Segmentation. In *Neural Information Processing Systems*.

[42] Zongxin Yang and Yi Yang. 2022. Decoupling Features in Hierarchical Propagation for Video Object Segmentation. In *Neural Information Processing Systems*.

[43] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. 2020. Video Object Segmentation and Tracking: A Survey. *ACM Transactions on Intelligent Systems and Technology* 11, 4 (2020), 1–47.

[44] Chenyangguang Zhang, Zhiqiang Lou, Yan Di, Federico Tombari, and Xiangyang Ji. 2023. Sst: Real-Time End-to-End Monocular 3D Reconstruction via Sparse Spatial-Temporal Guidance. In *IEEE International Conference on Multimedia and Expo*.

[45] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. 2020. Causal Intervention for Weakly-Supervised Semantic Segmentation. In *Advances in Neural Information Processing Systems*.

[46] Haojie Zhao, Junsong Chen, Lijun Wang, and Huchuan Lu. 2023. ARKitTrack: A New Diverse Dataset for Tracking Using Mobile RGB-D Data. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[47] Pengyao Zhao, Quanli Liu, Wei Wang, and Qiang Guo. 2021. TSDM: Tracking by SiamRPN++ with a Depth-refiner and a Mask-generator. In *International Conference on Pattern Recognition*.

[48] Bineng Zhong, Yingju Shen, Yan Chen, Weibo Xie, Zhen Cui, Hongbo Zhang, Duansheng Chen, Tian Wang, Xin Liu, Shujuan Peng, et al. 2015. Online Learning 3D Context for Robust Visual Tracking. *Neurocomputing* (2015).

[49] 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*.

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