SAM-DAQ: Segment Anything Model with Depth-guided Adaptive Queries for RGB-D Video Salient Object Detection

Jia Lin[1], Xiaofei Zhou[1*], Jiyuan Liu[1], Runmin Cong[2], Guodao Zhang[1], Zhi Liu[3*], Jiyong Zhang[1]

[1]Hangzhou Dianzi University, Hangzhou, China
[2]Shandong University, Jinan, China
[3]Shanghai University, Shanghai, China
lin_j@hdu.edu.cn, zxforchid@outlook.com, hankliu@hdu.edu.cn, rmcong@sdu.edu.cn,
guodaozhang@hdu.edu.cn, liuzhisjtu@163.com, jzhang@hdu.edu.cn

Abstract

Recently segment anything model (SAM) has attracted widespread concerns, and it is often treated as a vision foundation model for universal segmentation. Some researchers have attempted to directly apply the foundation model to the RGB-D video salient object detection (RGB-D VSOD) task, which often encounters three challenges, including the dependence on manual prompts, the high memory consumption of sequential adapters, and the computational burden of memory attention. To address the limitations, we propose a novel method, namely **S**egment **A**nything **M**odel with **D**epth-guided **A**daptive **Q**ueries (SAM-DAQ), which adapts SAM2 to pop-out salient objects from videos by seamlessly integrating depth and temporal cues within a unified framework. Firstly, we deploy a parallel adapter-based multi-modal image encoder (PAMIE), which incorporates several depth-guided parallel adapters (DPAs) in a skip-connection way. Remarkably, we fine-tune the frozen SAM encoder under prompt-free conditions, where the DPA utilizes depth cues to facilitate the fusion of multi-modal features. Secondly, we deploy a query-driven temporal memory (QTM) module, which unifies the memory bank and prompt embeddings into a learnable pipeline. Concretely, by leveraging both frame-level queries and video-level queries simultaneously, the QTM module selectively extracts temporal consistency features, iteratively updates the temporal representations of the queries. Extensive experiments are conducted on three RGB-D VSOD datasets, and the results show that the proposed SAM-DAQ consistently outperforms state-of-the-art methods in terms of all evaluation metrics.

Code — https://github.com/LinJ0866/SAM-DAQ

Introduction

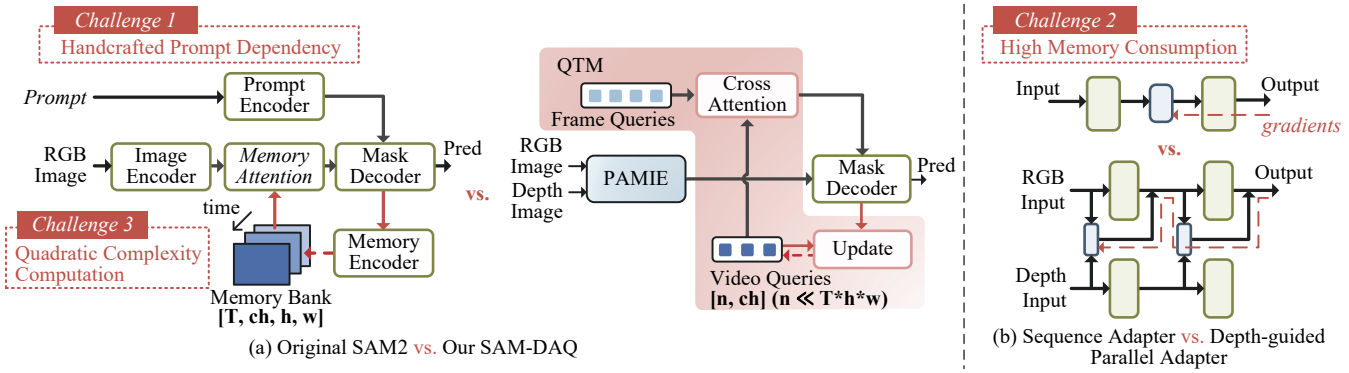
Salient object detection (SOD) (Borji et al. 2019) aims to identify and highlight the most attractive objects in visible images, which has been widely applied to many related areas, such as object tracking (Zhang et al. 2020; Zhou et al. 2021), robotics (Bao et al. 2025; Huang et al. 2025) and image retrieval (Fan, Wang, and Liang 2015). However, the conventional SOD methods relying solely on RGB data often present unsatisfactory performance when dealing with challenging scenarios, such as cluttered background, occlusions, and low light conditions.

To overcome the aforementioned limitations, some researchers have increasingly explored the introduction of multiple modalities. RGB-D SOD methods (Qu et al. 2017; Han et al. 2017; Cong et al. 2023; Bao et al. 2024) leverage depth information to provide robust spatial structure, effectively mitigating background distractions. Meanwhile, video SOD (VSOD) methods (Wang, Shen, and Shao 2017; Li et al. 2018; Liu and Liu 2023; Zhou et al. 2023) incorporate temporal cues into the SOD task, where they major in digging an effective characterization of motion cues. Embarking on this, researchers attempt to introduce both depth and temporal information to the SOD task, namely RGB-D VSOD.

Recently, the segment anything model (SAM) (Kirillov et al. 2023), which is treated as a vision foundation model for universal segmentation, has attracted more and more concerns. Embarking on SAM, SAM2 (Ravi et al. 2024) further extends its capability to video segmentation by incorporating a memory mechanism, which captures inter-frame dependencies via a memory bank. However, directly applying SAM2 to the RGB-D VSOD task introduces several challenges. Firstly, to guide the object segmentation task, SAM2 requires manual prompts (e.g., points, boxes and masks), but the RGB-D VSOD task cannot provide such information during the inference stage (Fig. 1(a)). Secondly, sequential adapters (Houlsby et al. 2019) used for prompt-free adaptation cause high GPU memory consumption during training, because the backward gradients must traverse the entire encoder (Diao et al. 2024) (Fig. 1 (b)). Finally, the memory attention mechanism in SAM2 imposes a high computational burden due to the extensive correlation computations between the current-frame features and the large memory bank.

To overcome the aforementioned limitations, we propose a Segment Anything Model with Depth-guided Adaptive Queries (SAM-DAQ), which adapts SAM2 for the RGB-D VSOD by seamlessly integrating depth and temporal cues. Specifically, we propose a parallel adapter-based multi-modal image encoder (PAMIE) that incorporates a series of

*Corresponding authors.



(a) Original SAM2 vs. Our SAM-DAQ

(b) Sequence Adapter vs. Depth-guided Parallel Adapter

Figure 1: High-level illustration of our SAM-DAQ.

depth-guided parallel adapters (DPAs). This design enables efficient fine-tuning of the frozen vision foundation model-based encoder under prompt-free conditions, while leveraging depth information as guidance to facilitate the fusion of RGB and depth features. In addition, we develop a query-driven temporal memory (QTM) module, which replaces the memory bank and prompt embedding with learnable queries (*i.e.*, object queries (Carion et al. 2020)), to capture temporal dependencies among different frames. Particularly, by leveraging both frame-level queries and video-level queries, the QTM module can selectively highlight visually attractive regions by incorporating the temporal consistency and iteratively update the temporal characterization of the queries. In this way, we can obtain effective learnable embeddings for the mask decoder. In summary, our contributions can be summarized as follows:

- We propose a Segment Anything Model with Depth-guided Adaptive Queries (SAM-DAQ) to conduct RGB-D VSOD, which effectively adapts the vision foundation model by deploying a parallel adapter-based multi-modal image encoder (PAMIE) and a query-driven temporal memory (QTM) module.
- We propose a PAMIE to enable prompt-free fine-tuning with minimal memory consumption and facilitate effective RGB-D fusion by using depth-guided parallel adapters (DPAs).
- We propose a QTM module to unify the memory bank and prompt embeddings into a learnable pipeline by using frame-level queries and video-level queries, where the former extracts visually attractive information from each frame and the latter captures temporal dependencies among different frames via cross-attention.
- Extensive experiments on three RGB-D VSOD datasets firmly demonstrate the superiority of the proposed SAM-DAQ over the state-of-the-art models.

Related Work

RGB-D Salient Object Detection

Depth information provides spatial cues resistant to contextual distractions and is widely as regarded a critical complement for RGB information. Existing RGB-D SOD methods typically follow three fusion paradigms (Mou et al. 2024), namely early fusion, middle fusion and late fusion. The early fusion strategy (Qu et al. 2017) treats the depth map as an additional channel to the RGB image, and directly concatenates the two modal images to form a 4D input. The late fusion strategy (Han et al. 2017) processes the two modalities independently, which generates the final mask by fusing their coarse predictions. However, the interaction between the modalities occurs at the output only, which limits the fusion of the two modal features. The middle fusion strategy (Cong et al. 2023; Bao et al. 2024) is widely employed to capture and exploit multi-modal correlations. For instance, Cong *et al.* (Cong et al. 2023) combine the multi-modal CNN features and then utilize the fused features to refine the transformer decoder. However, prior studies have revealed inherent limitations in the depth modality, including a lack of informative content (Hao et al. 2024) and unstable quality (Bao et al. 2024).

Video Salient Object Detection

Video SOD (VSOD) extends SOD by leveraging temporal information across frames to ensure spatial-temporal consistency. Early works (Wang, Shen, and Shao 2017; Li et al. 2018; Ji et al. 2020) utilize optical flow and attention mechanisms to capture motion cues, but their accuracy heavily depends on flow quality and motion scale (Singh, Verma, and Cheruku 2024). To better model long-term dependencies, memory-based frameworks (Oh et al. 2019; Cheng and Schwing 2022) encode the previous frame's features and their predictions, and store the generated representations in a memory bank. However, interacting with a large memory bank can introduce substantial computational consumption. To tackle the issue, query-based methods (Fang et al. 2024; Wang et al. 2023) utilize the learnable query to focus on relevant features selectively, achieving both high accuracy and efficiency.

Recent RGB-D VSOD methods aim to integrate depth and temporal cues simultaneously. For instance, Li *et al.* (Li et al. 2023) store previously fused RGB-D features into memory, extending memory networks to RGB-D settings. Mou *et al.* (Mou et al. 2024) fuse flow and multi-modal features via holistic multi-modal attentive paths (HMAPs).

Lin *et al.* (Lin et al. 2024) treat RGB, depth, and optical flow equally, and deploy intermediate supervision on their respective encoders. In addition, Suolang *et al.* (Suolang et al. 2025) propose a lightweight cross-shift module that efficiently fuses auxiliary depth and temporal cues.

Segment Anything Model

The Segment Anything Model (SAM) (Kirillov et al. 2023; Ravi et al. 2024) is a vision foundation model trained on large-scale data for universal image segmentation. While it generalizes well, its reliance on manual prompts (e.g., points, boxes), making it impractical for the video salient object detection task.

To eliminate the need for manual prompts, several works (Zhang et al. 2024; Ayzenberg, Giryes, and Greenspan 2024; Xie et al. 2025) attempt to generate pseudo-prompts from coarse masks, while others (Yang et al. 2024; Xiong et al. 2024; Xu 2025) only use SAM's encoder for feature extraction. Generally, the above methods tailor SAM for specific tasks without modifying its intrinsic architecture. Recent methods have explored parameter-efficient fine-tuning (PEFT) strategies. Wang *et al.* (Wang et al. 2024) insert adapter (Houlsby et al. 2019) between encoder blocks, while Zhong *et al.* (Zhong et al. 2024) integrate LoRA (Hu et al. 2022) into transformer layers. However, these sequential adapter structures tend to incur high memory consumption, as backward gradients must propagate through the entire encoder (Diao et al. 2024).

To extend SAM to the video segmentation task, Yue *et al.* (Yue et al. 2024) propose a flow reconstruction technique to guide SAM in object discovery. Deng *et al.* (Deng et al. 2024) present three distinct types of memory banks to mitigate the adverse effects of speckle noise and motion artifacts during memory prompting. To the best of our knowledge, no previous work has attempted to encode memory queries within the video SAM-based framework.

Method

Overview

As shown in Fig. 2, we introduce the vision foundation model and propose a novel Segment Anything Model with Depth-guided Adaptive Queries (SAM-DAQ), which organically integrates the depth and temporal cues. Our SAM-DAQ consists of three components, namely a parallel adapter-based multi-modal image encoder (PAMIE), a query-driven temporal memory (QTM) module, and the mask decoder retained from the original SAM2. Here, for simplicity, we present our method by detailing the process of a single frame. Firstly, RGB image and depth image constitute an image pair $\{\mathbf{I}_{RGB}, \mathbf{I}_D\}$, which is fed into PAMIE. Specifically, at each level, the PAMIE employs a depth-guided parallel adapter (DPA) to efficiently fine-tune the encoder under prompt-free conditions and fuse RGB and depth features. The encoder produces three-level image embeddings, namely $\mathbf{E}_I = \{\mathbf{E}_I^i\}_{i=2}^4$. Then, in the query-driven temporal memory (QTM) module, the highest-level image embeddings $\mathbf{E}_I^4$ progressively interact with static frame-level queries and iteratively update video-level queries, generating learnable embeddings $\mathbf{E}_L$. Next, by integrating both the image embeddings and the learnable embeddings, the mask decoder predicts a high-quality segmentation result $\mathbf{P}$ for the current frame. Finally, the update mechanism in the QTM module leverages the current image embeddings and the corresponding high-quality segmentation result to update the video-level queries, ensuring that temporal dependencies are effectively captured.

Parallel Adapter-based Multi-modal Image Encoder

To sufficiently leverage the powerful segmentation performance and generalization of the vision foundation model, we propose a parallel adapter-based multi-modal image encoder (PAMIE) that parallelly integrates the original image encoder with the lightweight adapters.

Given the multi-modal image pair $\{\mathbf{I}_{RGB}, \mathbf{I}_D\}$, the input is processed separately by the Hiera (Ryali et al. 2023). To address the inherent differences between RGB and depth modalities, we first use a linear depth projector (Moussely-Sergieh et al. 2018) to transform the depth input embeddings into the same feature space as the RGB modality. After that, considering the limited RGB-D VSOD data and large number of trainable parameters in the SAM2 encoder, we freeze all encoder weights and employ parameter-efficient fine-tuning (PEFT) strategies to fine-tune the model. Specifically, unlike prior efforts (Gao et al. 2024; Wang et al. 2024; Zhong et al. 2024), which insert adapters sequentially between transformer blocks or apply LoRA to the query and value projections within each attention layer, we parallelly integrate the adapters in a skip-connection way. This design not only facilitates the exploration of multi-modal information but also allows gradients to bypass the heavy transformer computations, significantly reducing memory consumption during the training stage.

Here, we design two kinds of parallel adapters. For the depth modality, the adapter connects the input and output of each Hiera block, which can be formulated as follows,

$$\begin{cases} \tilde{\mathbf{F}}_D^{i-1} = Adapter(\mathbf{F}_D^{i-1}) \\ \mathbf{F}_D^{i} = Hiera^i(\mathbf{F}_D^{i-1}) + DS(\tilde{\mathbf{F}}_D^{i-1}) \end{cases}, \tag{1}$$

where $Adapter$ denotes the adapter consisting of a down-projection linear layer followed by an activation function and an up-projection linear layer. $Hiera^i$ represents the i-th Hiera block ($i = 1, 2, 3, 4$), and DS represents the bilinear downsampling operation. Here, $\mathbf{F}_D^i$ and $\tilde{\mathbf{F}}_D^i$ are the features of the i-th Hiera block and the corresponding adapter, respectively. Note that, $\mathbf{F}_D^0$ is the output of the depth projector.

For the RGB modality, we utilize the depth-guided parallel adapter (DPA) to fuse RGB features and depth features. The whole process can be written as follows,

$$\begin{cases} \tilde{\mathbf{F}}_{RGB}^{i-1} = Adapter(Cat(\mathbf{F}_{RGB}^{i-1}, \mathbf{F}_D^{i-1})) \\ \mathbf{F}_{RGB}^{i} = Hiera^i(\mathbf{F}_{RGB}^{i-1}) + DS(\tilde{\mathbf{F}}_{RGB}^{i-1}) \\ \mathbf{F}_{RGB}^{1} = Hiera^1(\mathbf{F}_{RGB}^{0}) \end{cases}, \tag{2}$$

where Cat denotes concatenate operation and $\mathbf{F}_{RGB}^i$ is the RGB feature from the i-th Hiera block ($i = 2, 3, 4$). $\mathbf{F}_{RGB}^0$

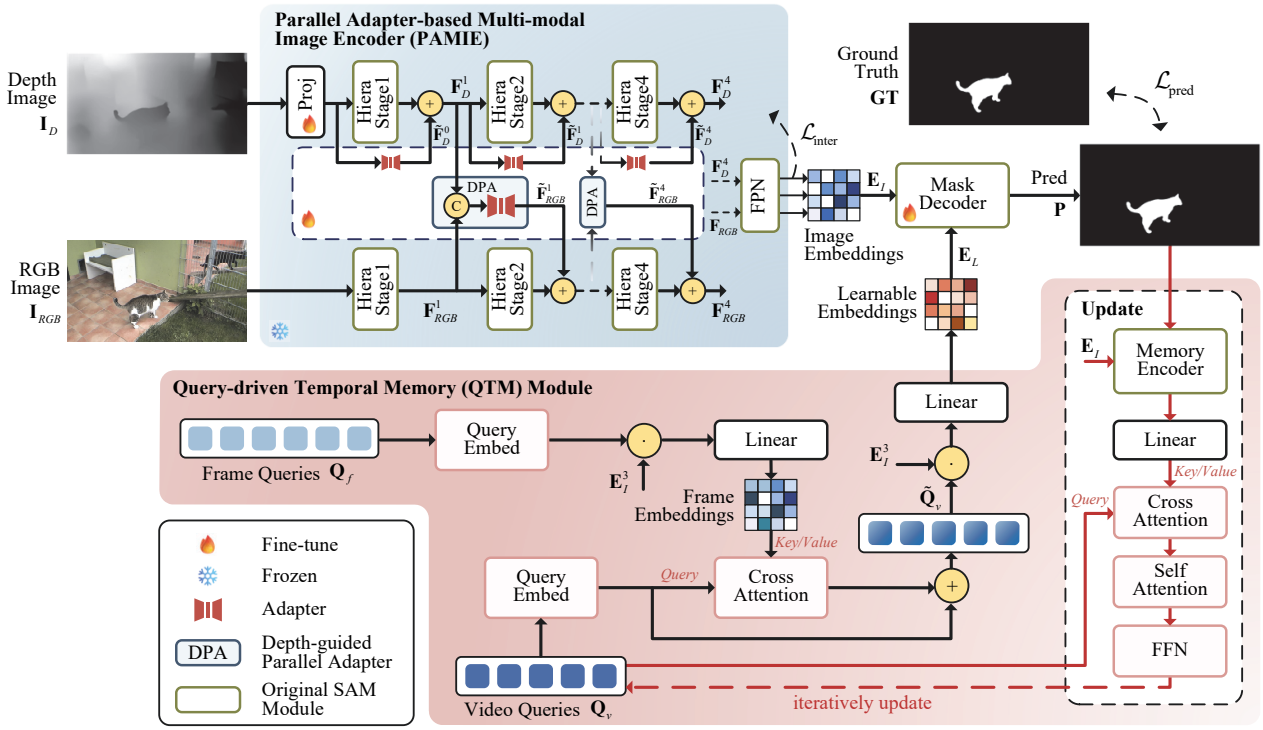


Figure 2: The overall architecture of the proposed Segment Anything Model with Depth-guided Adaptive Queries (SAM-DAQ) of a single frame.

means the input RGB image. Finally, after applying a feature pyramid network (FPN), PAMIE generates three-level image embeddings, namely $\mathbf{E}_I = \{\mathbf{E}_I^i\}_{i=2}^4$.

In addition, to further promote the fusion of the depth features and RGB features, we introduce a self-reasoning scheme that applies a lightweight convolution followed by a sigmoid activation function to each level image embedding $\mathbf{E}_I^i$, generating an intermediate prediction result, namely $\tilde{\mathbf{P}} = \{\tilde{\mathbf{P}}^2, \tilde{\mathbf{P}}^3, \tilde{\mathbf{P}}^4\}$. Notably, we only deploy the supervision to the highest-level image embeddings.

Query-Driven Temporal Memory Module

The manual prompts and the large memory bank restrict the application of the vision foundation model, especially in the VSOD task, where the former is difficult to acquire and the latter presents a high computational cost. To address the challenges, we propose a query-driven temporal memory (QTM) module, which unifies the prompt generation and the temporal modeling via a learnable query method.

We introduce two sets of learnable queries, namely frame-level queries $\mathbf{Q}_f \in \mathbb{R}^{N_f \times c}$ and video-level queries $\mathbf{Q}_v \in \mathbb{R}^{N_v \times c}$, where c is the hidden dimension, and N_f and N_v mean the number of frame-level queries and video-level queries, respectively. As shown in Fig. 2, the frame-level queries interact with the highest-level image embeddings $\mathbf{E}_I^4$, and in this way, we can acquire the saliency-related frame embeddings $\mathbf{E}_f$. Besides, to incorporate temporal context, we perform cross-attention between video-level queries and frame embeddings. Here, both $\mathbf{Q}_f$ and $\mathbf{Q}_v$ are linearly projected via a query embedding layer, resulting in $\mathbf{Q}_f'$ and $\mathbf{Q}_v'$. The following operations can be defined as follows,

$$\begin{cases} \mathbf{E}_f = Linear(\mathbf{Q}_f' \cdot \mathbf{E}_I^4) \\ \tilde{\mathbf{Q}}_v = CA(\mathbf{Q}_v', \mathbf{E}_f) + \mathbf{Q}_v' \end{cases}, \tag{3}$$

where $Linear$ is the linear projection operation and CA represents cross-attention operation. After that, we further interact the enhanced video-level queries $\tilde{\mathbf{Q}}_v$ and $\mathbf{E}_I^4$ via element-wise multiplication, generating learnable embeddings $\mathbf{E}_L \in \mathbb{R}^{N_v \times c}$, which can be used to replace SAM's original sparse prompt embeddings. Embarking on this, the mask decoder utilizes both learnable embeddings $\mathbf{E}_L$ and the image embeddings $\mathbf{E}_I$ to generate the predicted map $\mathbf{P}$ for the current frame.

Different from the static frame-level queries, video-level queries are iteratively updated via an update mechanism, which effectively captures the temporal dependencies among different frames. The update mechanism comprises three components, including a memory encoder adopted from SAM2, a linear projection, and a transformer decoder module consisting of a cross-attention, a self-attention, and a feed-forward network (FFN). Concretely, given the image embeddings $\mathbf{E}_{I,t}$ and corresponding prediction map $\mathbf{P}_t$ at frame t, we first extract memory features as follows,

$$\mathbf{F}_m = Linear(ME(\mathbf{E}_{I,t}, \mathbf{P}_t)), \tag{4}$$

where ME is the memory encoder. These features are then used for the temporal update stage, namely

$$\mathbf{Q}_{v,t+1} = \mathbf{Q}_{v,t} + FFN(SA(CA(\mathbf{Q}_{v,t}, \mathbf{F}_m))), \tag{5}$$

Methods	Referece	RDVS (Mou et al. 2024)				ViDSOD-100 (Lin et al. 2024)				DViSal (Li et al. 2023)			
		$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
HRTransNet	TCSVT'22	0.725	0.671	0.445	0.076	0.745	0.686	0.531	0.099	0.745	0.685	0.531	0.099
PICRNet	MM'23	0.743	0.728	0.535	0.074	0.873	0.830	0.738	0.038	0.715	0.670	0.568	0.147
DVSOD	NeurIPS'23	0.748	0.587	0.452	0.070	0.783	0.702	0.568	0.083	0.807	0.729	0.610	0.113
LSTA	PR'24	0.746	0.650	0.484	0.069	0.757	0.671	0.565	0.086	0.848	0.700	0.640	0.082
DPA	CVPR'24	0.675	0.666	0.445	0.096	0.848	0.817	0.715	0.051	0.796	0.724	0.635	0.102
DCTNet+	TIP'24	0.909	0.876	0.794	0.029	0.901	0.876	0.809	0.030	0.828	0.767	0.689	0.095
ATFNet	IJCV'24	0.732	0.713	0.491	0.074	0.901	0.875	0.813	0.027	0.795	0.724	0.622	0.111
MDSAM	MM'24	0.813	0.791	0.647	0.056	0.909	0.877	0.815	0.026	0.856	0.796	0.715	0.071
SAM2-UNet	Arxiv'24	0.888	0.843	0.765	0.035	0.907	0.891	0.829	0.025	0.856	0.778	0.747	0.064
MFENet	ICASSP'25	-	0.794	0.700	0.049	-	0.831	0.763	0.040	-	0.760	0.717	0.080
KAN-SAM	ICME'25	0.888	0.854	0.791	0.028	0.912	0.892	0.846	0.025	0.885	0.835	0.783	0.052
Ours	-	**0.913**	**0.879**	**0.827**	**0.026**	**0.918**	**0.894**	**0.868**	**0.020**	**0.914**	**0.840**	**0.818**	**0.046**

Table 1: Quantitative comparison results with the state-of-the-art RGB-D video salient object detection models on three representative datasets.

Methods	Trainable/Total (M)	Memory (G)	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
w/o depth projector	-	20.3	0.899	0.870	0.808	0.023
w/o parallel (sequential adapter)	17.4/236.0	91.9	0.860	0.830	0.778	0.028
w/o parallel (LoRA)	56.0/274.6	95.0	0.889	0.877	0.824	0.027
w/o multi-modal	-	17.9	0.876	0.853	0.782	0.029
Ours	**19.2/237.9**	**21.0**	**0.913**	**0.879**	**0.827**	**0.026**

Table 2: Ablation studies of our Parallel Adapter-based Multi-modal Image Encoder (PAMIE).

where FFN and SA represent the FFN and self-attention, respectively. This iterative update mechanism refines the video-level queries for the subsequent frame, which effectively builds the temporal dependencies.

Loss Function

To effectively train our model, we attempt to supervise the generation of both image embeddings and learnable embeddings. Concretely, our loss function consists of the final prediction loss $\mathcal{L}_{\text{pred}}$ and the intermediate supervision loss $\mathcal{L}_{\text{inter}}$. The loss function $\mathcal{L}_{\text{total}}$ can be written as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \alpha \cdot \mathcal{L}_{\text{inter}}, \tag{6}$$

where both $\mathcal{L}_{\text{pred}}$ and $\mathcal{L}_{\text{inter}}$ are computed by using the binary cross entropy (BCE) loss (De Boer et al. 2005), and α is the weight of the intermediate loss. The final prediction loss $\mathcal{L}_{\text{pred}}$ computes the differences between the ground-truth $\mathbf{GT}$ and the final prediction result $\mathbf{P}$. The intermediate supervision loss $\mathcal{L}_{\text{inter}}$ is computed by comparing $\mathbf{GT}$ with intermediate prediction results $\tilde{\mathbf{P}}$.

Experiments

Experiment Settings

Datasets. To comprehensively evaluate our model, we train and test our proposed method on three newly RGB-D VSOD datasets, namely RDVS (Mou et al. 2024) (4087 frames, 57 videos), ViDSOD-100 (Lin et al. 2024) (9362 frames, 100 videos) and DViSal (Li et al. 2023) (7117 frames, 237 videos). Note that, for a fair comparison, only the labeled frames in the DViSal dataset are used.

Evaluation Metrics. Following the previous works (Li et al. 2023), we adopt four widely-used metrics to evaluate the model performance, *i.e.*, E-measure (E_ξ) (Fan et al. 2018), S-measure (S_α) (Fan et al. 2017), F-measure (F_β) (Achanta et al. 2009), and mean absolute error (MAE or M) (Borji et al. 2015). *The lower the MAE, the better. For other metrics, the higher score is better.*

Implementation Details. Our SAM-DAQ is built on the large-scale configuration of SAM2 (SAM-L), with extensions for RGB-D VSOD. During the training stage, the SAM encoder parameters are frozen, and the spatial resolution of input images is resized to 1024×1024. To balance training efficiency with long-term memory modeling, we randomly sample 10 frames per video in each epoch and feed them into the network simultaneously. We adopt AdamW (Loshchilov and Hutter 2017) as the optimizer, where the learning rate and the weight decay are set to 0.0001 and 0.05, respectively. The batch size is set to 1, and the number of training iterations is 2000. Thanks to our efficient DPA, our SAM-DAQ can be trained on a single RTX-3090 (24 GB) GPU within 3 hours.

Comparison with the State-of-the-art Methods

We compare SAM-DAQ with 11 state-of-the-art RGB-D VSOD models, including HRTransNet (Tang et al. 2022), PICRNet (Cong et al. 2023), DVSOD (Li et al. 2023), LSTA (Li et al. 2024), DPA (Cho et al. 2024), DCT-

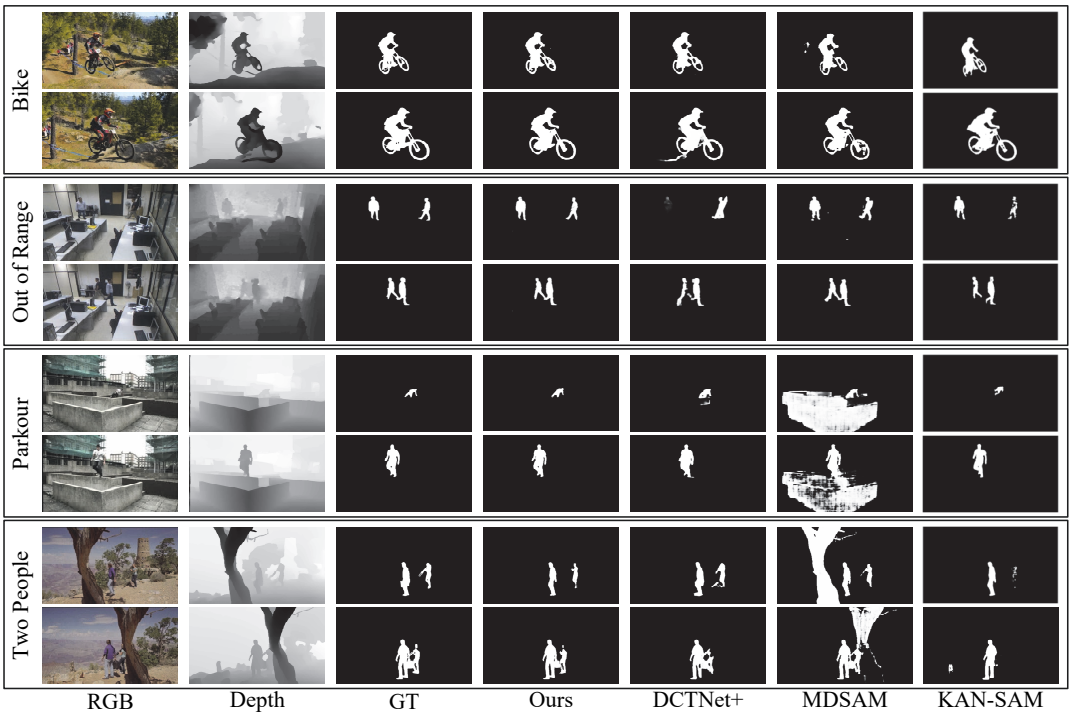
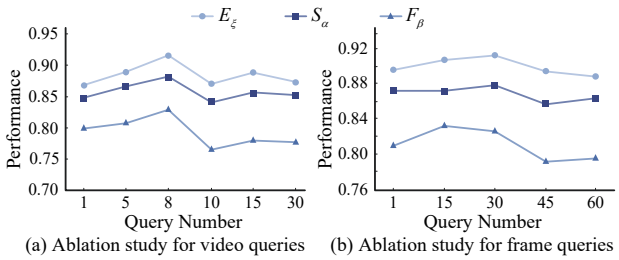


Figure 3: Qualitative comparison with the state-of-the-art RGB-D video salient object detection models on RDVS dataset.

Strategies	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
sparse only (Ours)	**0.913**	**0.879**	**0.827**	**0.026**
dense only	0.875	0.856	0.783	0.032
both	0.862	0.839	0.763	0.033

Table 3: Ablation studies of our learnable embeddings generation strategy.



(a) Ablation study for video queries (b) Ablation study for frame queries

Figure 4: Ablation studies of different query numbers.

Net+ (Mou et al. 2024), ATFNet (Lin et al. 2024), MDSAM (Gao et al. 2024), SAM2-UNet (Xiong et al. 2024), MFENet (Suolang et al. 2025), and KAN-SAM (Li et al. 2025).

The quantitative results on three RGB-D VSOD datasets are shown in Table 1. It can be observed that our SAM-DAQ achieves the best performance when compared with the cutting-edge methods. To demonstrate that the performance gain of our method is not solely due to the SAM2 backbone, we make comparisons with various SAM-based baselines. Specifically, compared with the KAN-SAM, we achieve average improvements of 1.5%, 1.0%, 2.4% and 0.003 in terms of E-measure, S-measure, F-measure and MAE, respectively. The above quantitative comparison results firmly demonstrate the effectiveness and superiority of our method.

The qualitative comparison is presented in Fig. 3, where we compare our model with DCTNet+, MDSAM, and KAN-SAM on four representative scenes ("bike", "out of range", "parkour" and "two people") from the RDVS benchmark. We observe that though MDSAM and KAN-SAM can segment objects completely, they struggle to accurately detect the salient regions without temporal modeling. DCTNet+ tends to highlight background regions, and the reason behind this can be attributed to that the quality of optical flow is highly sensitive to the foreground motion. These visualization results further validate the consistent superiority of our SAM-DAQ.

Ablation Studies

Effect of PAMIE:

To validate the effectiveness of our PAMIE, we conduct ablation studies and introduce trainable/total parameters (*i.e.*, Trainable/Total) and GPU memory usage in the training stage (*i.e.*, Memory) as additional metrics, as shown in Table 2. The results indicate that the E-measure drops from 91.3% to 89.9% when removing the depth projector (*i.e.*, w/o depth projector), which highlights the necessity of addressing the inherent differences between RGB cues

Dimensions	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
32	0.889	0.863	0.795	0.023
64 (Ours)	**0.913**	**0.879**	**0.827**	**0.026**
128	0.874	0.842	0.779	0.030
256	0.880	0.863	0.799	0.023

Table 4: Ablation studies of query hidden dimension.

Methods	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
none	0.883	0.854	0.788	0.032
SAM2	0.853	0.829	0.796	0.028
multiply	0.895	0.862	0.804	0.027
addition (Ours)	**0.913**	**0.879**	**0.827**	**0.026**

Table 5: Ablation studies of update mechanism in Query-Driven Temporal Memory (QTM) Module.

and depth cues. Replacing the parallel adapter with either a sequential adapter (*i.e.*, w/o parallel (sequential adapter)) or LoRA (*i.e.*, w/o parallel (LoRA)) increases memory usage significantly (91.9 GB and 95.0 GB, respectively), despite maintaining competitive accuracy. In stark contrast, our DPA achieves the best overall performance with only 21.0 GB of memory usage, confirming its superior trade-off between accuracy and memory efficiency. Besides, compared to the w/o multi-modal, we can see that our model still performs better than this variant. According to the above comparison results, we can confirm that our DPAs organized in a skip-connection way and explicit multi-modal fusion are crucial for effectively adapting SAM to the RGB-D VSOD task.

Effect of QTM:

To validate the effectiveness of the learnable embeddings, we conduct ablation studies, as shown in Table 3. We can see that sparse-only variant achieves superior performance when compared with the dense-only or their combination. A plausible explanation is that the queries in QTM interact via token-wise attention rather than pixel-wise convolution, making them structurally analogous to sparse embeddings in SAM pretraining. This structural consistency enables more efficient adaptation. Furthermore, in Fig. 4, we present a detailed analysis of the impact of the number of frame-level queries and video-level queries on the performance of our SAM-DAQ. Specifically, reducing the number of video-level queries to 5 results in decreases of 2.6%, 1.6%, and 0.8% in terms of E-measure, S-measure, and F-measure, respectively. Conversely, increasing the number of video-level queries to 10 introduces excessive background noise, which significantly degrades detection performance. For frame-level queries, the results indicate that 30 queries provide sufficient guidance. Overall, the best performance is achieved when the number of video-level queries and the frame-level queries are set to 8 and 30, respectively. Additionally, we conduct a comprehensive analysis of the query's hidden dimension. As shown in Table 4, when the hidden dimensions of the query are set to 64, we can obtain optimal results.

Supervised Levels $\mathbf{E}_I^2$	$\mathbf{E}_I^3$	$\mathbf{E}_I^4$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
✓			0.855	0.837	0.762	0.033
	✓		0.877	0.846	0.770	0.031
		✓	**0.913**	**0.879**	**0.827**	**0.026**
	✓	✓	0.858	0.833	0.741	0.036
✓	✓	✓	0.845	0.824	0.749	0.032

Table 6: Ablation studies of our Intermediate Supervision.

We also analyze the temporal update mechanism in the QTM module, as shown in Table 5. From the results, we can see that directly removing the update phase (*i.e.*, none), the performance drops 3% in terms of E-measure, 2.5% in terms of S-measure, 3.9% in terms of F-measure, and increases 0.006 % in terms of MAE, which proves the importance of the video-level queries update mechanism. We also compare SAM2's original memory bank (*i.e.*, SAM2) with our QTM, we can see that our QTM more effectively leverages temporal cues than traditional memory-bank baselines. Additionally, replacing our addition update strategy with a multiply update strategy, the variant multiply presents minor performance drops. This validates the effectiveness of our design.

Effect of Intermediate Supervision: To validate the effectiveness of our intermediate supervision, we conduct ablation studies, where we deploy supervision to different levels of image embeddings. As shown in Table 6, supervising only the highest-level embeddings $\mathbf{E}_I^4$ yields the best performance, while additional supervision on lower-level embeddings degrades the overall performance.

Conclusion

In this paper, we propose a novel RGB-D VSOD model, namely SAM-DAQ, which introduces the vision foundation model (*i.e.* SAM2) for RGB-D VSOD. To address the challenges of prompt dependency, high memory consumption, and computational costs, we deploy two key components, namely the PAMIE and the QTM module. The PAMIE module leverages a series of DPAs that are deployed in a skip-connection way to efficiently integrate RGB and depth features while fine-tuning the encoder under prompt-free conditions. Meanwhile, the QTM module unifies the temporal modeling and prompts into a learnable query way, eliminating the need for handcrafted prompts and mitigating the computational burden of large memory banks. Extensive experiments on three benchmark RGB-D VSOD datasets demonstrate the superiority of our SAM-DAQ framework over state-of-the-art methods. Our approach establishes a new paradigm for leveraging vision foundation models in the RGB-D video salient object detection task, offering an efficient and effective solution for the RGB-D VSOD task. In future work, we will further optimize the query-based memory for multi-object video segmentation and explore adaptive query generation strategies.

Acknowledgments

This work was supported in part by the Zhejiang Province Key R&D Project No. 2023C01046, in part by the National Natural Science Foundation of China (No. 62271180, 62471278, 62471285).

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, 1597–1604. IEEE.

Ayzenberg, L.; Giryes, R.; and Greenspan, H. 2024. Proto-SAM: One-Shot Medical Image Segmentation With Foundational Models. *arXiv preprint arXiv:2407.07042*.

Bao, L.; Zhou, X.; Lu, X.; Sun, Y.; Yin, H.; Hu, Z.; Zhang, J.; and Yan, C. 2024. Quality-aware selective fusion network for VDT salient object detection. *IEEE Transactions on Image Processing*.

Bao, L.; Zhou, X.; Zheng, B.; Cong, R.; Yin, H.; Zhang, J.; and Yan, C. 2025. IFENet: Interaction, Fusion, and Enhancement network for VDT Salient Object Detection. *IEEE Transactions on Image Processing*.

Borji, A.; Cheng, M.-M.; Hou, Q.; Jiang, H.; and Li, J. 2019. Salient object detection: A survey. *Computational visual media*, 5(2): 117–150.

Borji, A.; Cheng, M.-M.; Jiang, H.; and Li, J. 2015. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12): 5706–5722.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Cheng, H. K.; and Schwing, A. G. 2022. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, 640–658. Springer.

Cho, S.; Lee, M.; Lee, S.; Lee, D.; Choi, H.; Kim, I.-J.; and Lee, S. 2024. Dual prototype attention for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19238–19247.

Cong, R.; Liu, H.; Zhang, C.; Zhang, W.; Zheng, F.; Song, R.; and Kwong, S. 2023. Point-aware interaction and cnn-induced refinement network for RGB-D salient object detection. In *Proceedings of the 31st ACM international conference on multimedia*, 406–416.

De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of operations research*, 134: 19–67.

Deng, X.; Wu, H.; Zeng, R.; and Qin, J. 2024. MemSAM: taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9622–9631.

Diao, H.; Wan, B.; Zhang, Y.; Jia, X.; Lu, H.; and Chen, L. 2024. Unipt: Universal parallel tuning for transfer learning with efficient parameter and memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28729–28740.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*.

Fan, D. P.; Wang, J.; and Liang, X. M. 2015. Improving image retrieval using the context-aware saliency areas. *Applied Mechanics and Materials*, 734: 596–599.

Fang, H.; Zhang, T.; Zhou, X.; and Zhang, X. 2024. Learning better video query with sam for video instance segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Gao, S.; Zhang, P.; Yan, T.; and Lu, H. 2024. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9894–9903.

Han, J.; Chen, H.; Liu, N.; Yan, C.; and Li, X. 2017. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE transactions on cybernetics*, 48(11): 3171–3183.

Hao, Z.; Xiao, Z.; Luo, Y.; Guo, J.; Wang, J.; Shen, L.; and Hu, H. 2024. PrimKD: Primary Modality Guided Multimodal Fusion for RGB-D Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1943–1951.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, J.; Wu, Y.; Zhou, X.; Lin, J.; Chen, Z.; Zhang, G.; Xia, L.; and Zhang, J. 2025. Multi-Scale Adaptive Prototype Transformer Network for Few-shot Strip Steel Surface Defect Segmentation. *IEEE Transactions on Instrumentation and Measurement*.

Ji, Y.; Zhang, H.; Jie, Z.; Ma, L.; and Wu, Q. J. 2020. CASNet: A cross-attention siamese network for video salient object detection. *IEEE transactions on neural networks and learning systems*, 32(6): 2676–2690.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, G.; Xie, Y.; Wei, T.; Wang, K.; and Lin, L. 2018. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3243–3252.

Li, J.; Ji, W.; Wang, S.; Li, W.; et al. 2023. Dvsod: Rgb-d video salient object detection. *Advances in Neural Information Processing Systems*, 36: 8774–8787.

Li, P.; Zhang, Y.; Yuan, L.; Xiao, H.; Lin, B.; and Xu, X. 2024. Efficient long-short temporal attention network for unsupervised video object segmentation. *Pattern Recognition*, 146: 110078.

Li, X.; Hou, R.; Ren, T.; and Wu, G. 2025. KAN-SAM: Kolmogorov-Arnold Network Guided Segment Anything Model for RGB-T Salient Object Detection. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Lin, J.; Zhu, L.; Shen, J.; Fu, H.; Zhang, Q.; and Wang, L. 2024. ViDSOD-100: A New Dataset and a Baseline Model for RGB-D Video Salient Object Detection. *International Journal of Computer Vision*, 132(11): 5173–5191.

Liu, Z.-y.; and Liu, J.-w. 2023. Part-aware attention correctness for video salient object detection. *Engineering Applications of Artificial Intelligence*, 119: 105733.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mou, A.; Lu, Y.; He, J.; Min, D.; Fu, K.; and Zhao, Q. 2024. Salient object detection in RGB-D videos. *IEEE Transactions on Image Processing*.

Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; and Roth, S. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 225–234.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9226–9235.

Qu, L.; He, S.; Zhang, J.; Tian, J.; Tang, Y.; and Yang, Q. 2017. RGBD salient object detection via deep fusion. *IEEE transactions on image processing*, 26(5): 2274–2285.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ryali, C.; Hu, Y.-T.; Bolya, D.; Wei, C.; Fan, H.; Huang, P.-Y.; Aggarwal, V.; Chowdhury, A.; Poursaeed, O.; Hoffman, J.; et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, 29441–29454. PMLR.

Singh, H.; Verma, M.; and Cheruku, R. 2024. Dsfnet: video salient object detection using a novel lightweight deformable separable fusion network. *IEEE Transactions on Instrumentation and Measurement*.

Suolang, D.; He, J.; Tsering, W.; Fu, K.; Li, X.; and Zhao, Q. 2025. Lightweight Multi-Frequency Enhancement Network for RGB-D Video Salient Object Detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Tang, B.; Liu, Z.; Tan, Y.; and He, Q. 2022. HRTransNet: HRFormer-driven two-modality salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 728–742.

Wang, J.; Chen, D.; Wu, Z.; Luo, C.; Tang, C.; Dai, X.; Zhao, Y.; Xie, Y.; Yuan, L.; and Jiang, Y.-G. 2023. Look before you match: Instance understanding matters in video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2268–2278.

Wang, K.; Lin, D.; Li, C.; Tu, Z.; and Luo, B. 2024. Adapting Segment Anything Model to Multi-modal Salient Object Detection with Semantic Feature Fusion Guidance. *arXiv preprint arXiv:2408.15063*.

Wang, W.; Shen, J.; and Shao, L. 2017. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1): 38–49.

Xie, B.; Tang, H.; Yan, Y.; and Agam, G. 2025. RFMedSAM 2: Automatic Prompt Refinement for Enhanced Volumetric Medical Image Segmentation with SAM 2. *arXiv preprint arXiv:2502.02741*.

Xiong, X.; Wu, Z.; Tan, S.; Li, W.; Tang, F.; Chen, Y.; Li, S.; Ma, J.; and Li, G. 2024. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*.

Xu, Y. 2025. DGSUnet: An Improved Unet Model with DINO-Guided SAM2 for Multi-Scale Feature Collaboration. *arXiv preprint arXiv:2503.21187*.

Yang, S.; Bi, H.; Zhang, H.; and Sun, J. 2024. SAM-UNet: Enhancing Zero-Shot Segmentation of SAM for Universal Medical Images. *arXiv preprint arXiv:2408.09886*.

Yue, J.; Zhang, R.; Zhang, Z.; Zhao, R.; Lv, W.; and Ma, J. 2024. How SAM helps Unsupervised Video Object Segmentation? In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.

Zhang, P.; Liu, W.; Wang, D.; Lei, Y.; Wang, H.; and Lu, H. 2020. Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. *Pattern Recognition*, 100: 107130.

Zhang, X.; Liu, Y.; Lin, Y.; Liao, Q.; and Li, Y. 2024. Uv-sam: Adapting segment anything model for urban village identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22520–22528.

Zhong, Z.; Tang, Z.; He, T.; Fang, H.; and Yuan, C. 2024. Convolution meets lora: Parameter efficient finetuning for segment anything model. *arXiv preprint arXiv:2401.17868*.

Zhou, X.; Cao, W.; Gao, H.; Ming, Z.; and Zhang, J. 2023. STI-Net: Spatiotemporal integration network for video saliency detection. *Information Sciences*, 628: 134–147.

Zhou, Z.; Pei, W.; Li, X.; Wang, H.; Zheng, F.; and He, Z. 2021. Saliency-associated object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9866–9875.