# Shallow Features Matter: Hierarchical Memory with Heterogeneous Interaction for Unsupervised Video Object Segmentation

Xiangyu Zheng*
xyzheng23@m.fudan.edu.cn
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China

Songcheng He*
24210240166@m.fudan.edu.cn
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China

Wanyun Li
wyli22@m.fudan.edu.cn
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China

Xiaoqiang Li
xqli@shu.edu.cn
The School of Computer Engineering and Science, Shanghai University
Shanghai, China

Wei Zhang†
weizh@fudan.edu.cn
Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University
Shanghai, China

## Abstract

Unsupervised Video Object Segmentation (UVOS) aims to predict pixel-level masks for the most salient objects in videos without any prior annotations. While memory mechanisms have been proven critical in various video segmentation paradigms, their application in UVOS yield only marginal performance gains despite sophisticated design. Our analysis reveals a simple but fundamental flaw in existing methods: **over-reliance on memorizing high-level semantic features**. UVOS inherently suffers from the deficiency of lacking fine-grained information due to the absence of pixel-level prior knowledge. Consequently, memory design relying solely on high-level features, which predominantly capture abstract semantic cues, is insufficient to generate precise predictions. To resolve this fundamental issue, we propose a novel hierarchical memory architecture to incorporate both shallow- and high-level features for memory, which leverages the complementary benefits of pixel and semantic information. Furthermore, to balance the simultaneous utilization of the pixel and semantic memory features, we propose a heterogeneous interaction mechanism to perform pixel-semantic mutual interactions, which explicitly considers their inherent feature discrepancies. Through the design of Pixel-guided Local Alignment Module (PLAM) and Semantic-guided Global Integration Module (SGIM), we achieve delicate integration of the fine-grained details in shallow-level memory and the semantic representations in high-level memory. Our **H**ierarchical **M**emory with **H**eterogeneous **I**nteraction Network (**HMHI-Net**) consistently achieves state-of-the-art performance across all UVOS and video saliency detection benchmarks. Moreover, HMHI-Net consistently exhibits high performance across different backbones, further demonstrating its superiority and robustness. Project page: https://github.com/ZhengxyFlow/HMHI-Net .

## Keywords

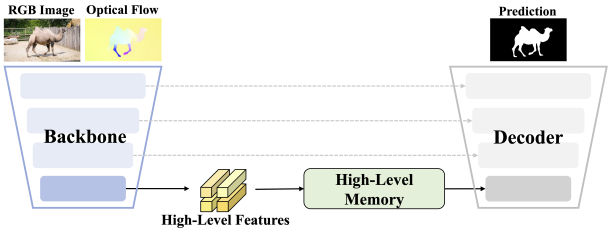
Unsupervised video object segmentation, memory mechanism, optical flow, pixel-level feature, semantic feature

## 1 Introduction

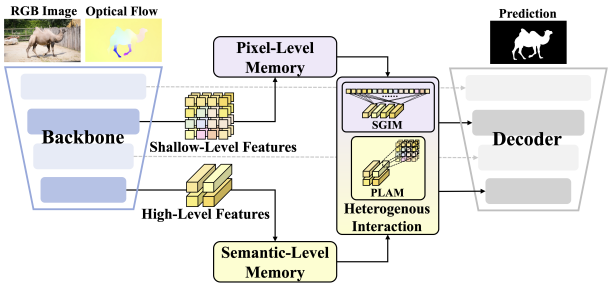
Unsupervised video object segmentation (UVOS) aims to segment the most salient object in a video sequence without any prior annotations, which makes itself a highly challenging task in visual domain. Given its ability to autonomously identify and track objects, UVOS plays a crucial role in a myriad of real-world applications.

UVOS approaches have long been confronted with a fundamental challenge: predicting pixel-wise precise segmentation with no prior knowledge. To address this issue, mainstream UVOS methods[8, 9, 20, 23, 27, 37, 42, 47, 52, 67, 72, 76] commonly incorporate optical flow as an auxiliary input to guide segmentation, and focus on designing sophisticated fusion modules to enhance performance. However, optical flow only contains short-term motion cues from two consecutive frames, which neglects the crucial long-term correspondences in video sequences. Memory mechanisms [6, 7, 24, 25, 31, 39, 48, 50, 58] have emerged as a powerful design in various video segmentation tasks due to their ability to effectively capture temporal dependencies across the video sequence. Some recent approaches[8, 13, 26, 27, 41] have explored the integration of long-term memory mechanisms into UVOS. Nevertheless, these memory-based methods have yielded only marginal performance gains despite their intricate architectures. We observe a simple yet pivotal defect in these methods: **A predominant reliance on**

*Equal contribution.
†Corresponding authors.




(a) Single-Level Memory Architecture




(b) Hierarchical Memory Architecture

**Figure 1: Illustration of the conventional single-level memory architecture and our hierarchical memory architecture. (a) Single-level memory architecture which solely relies on high-level features. (b) Hierarchical memory architecture which incorporates both shallow- and high-level features for memory.**

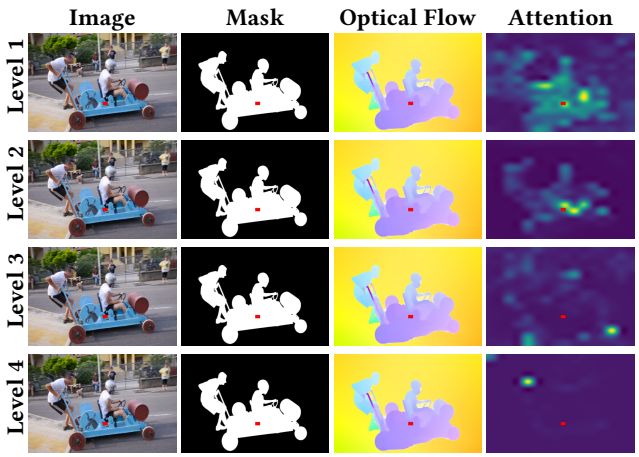



**Figure 2: Visualization of attention maps at different encoder levels. From level 1 to level 4, the model gradually focuses more on few tokens with representative semantic cues, while emphasizes less on the object details. The little read square on each map denotes the token selected for attention analysis.**

**high-level memory features, accompanied by a disregard for the fundamental limitations intrinsic to UVOS.**

Unlike semi-supervised video object segmentation (SVOS), where a pixel-wise mask of the first frame is provided as guidance, UVOS inherently lacks fine-grained object details and thereby struggles to generate pixel-wise predictions. Moreover, the compressing of raw images into compact high-level features at the multi-scale encoder further aggravates the loss of fine-grained details. And information retrieved from the high-level memory bank is gradually diluted during the bottom-up decoding phase. We investigate the information focus of different layers during the encoding process, and visualize their attention maps in Fig. 2. It can be observed that shallow encoding levels (level 1 and 2) focus more on the general pixels of foreground objects, while high encoding levels (level 3 and 4) only emphasize on few key points which best conveys object semantics. As a result, high-level memory alone can hardly compensate for the intrinsic absence of pixel-level guidance in UVOS, leading to segmentation maps with imprecision and insufficient details. This shortcoming remarkably limits the performance of previous UVOS models, particularly in complex scenarios.

To tackle the aforementioned challenge, we put forwardthe Hierarchical Memory with Heterogeneous Interaction Network (HMHI-Net) for UVOS. Firstly, we propose a simple yet effective hierarchical memory architecture which innovatively integrates both high-level and shallow-level features for memory. High-level features, which primarily encode semantic information, contribute to maintaining object consistency across frames. In contrast, shallow-level features preserve rich pixel-wise details, thereby enhancing the segmentation of fine-grained structures. We construct two separate memory banks for both features, and thereby optimize feature encoding with both semantic and pixel-level from memorized frames. Additionally, predicted masks are added to the memory banks during the memory update. By leveraging both semantic and pixel-level memory, the proposed architecture realizes frame-wise feature refinement through hierarchical guidance and produces segmentation maps with remarkable advancements.

Furthermore, we introduce a heterogeneous interaction mechanism to balance the contributions of the two memory banks and facilitate their feature mutual refinement. Due to the feature discrepancy between the shallow- and high-level memory, the improper use of their features can lead to feature misalignment and performance degradation. We address this issue by conducting shallow-high mutual interactions to facilitate their bidirectional feature refinement, which at the same time explicitly accounts for their inherent feature distinctions. Specifically, shallow-level features emphasize more on local fine-grained details, while high-level features capture global semantic representations. Accordingly, we design two specialized modules: the Pixel-guided Local Alignment Module (PLAM) and the Semantic-guided Global Integration Module (SGIM), tailored for shallow-to-high and high-to-shallow refinement, respectively. PLAM performs shallow-to-high information integration according to the relative positions of tokens, thereby preserving spatial coherence of pixel-level details and minimizing interference from unrelated details. Conversely, SGIM realizes high-to-shallow advancement by enabling broader perception filed on semantic cues in high-level tokens, which better leverages the comprehensive semantic guidance. With PLAM and SGIM, the heterogeneous interaction mechanism achieves a well-balanced shallow-high mutual interactions, and effectively optimizes both

features with their complementary nature, remarkably elevating the overall model performance.

Our contributions can be summarized as follows:

• We propose a novel hierarchical memory architecture that simultaneously incorporates shallow- and high-level features for memory, facilitating UVOS with both pixel-level details and semantic richness stored in memory banks.

• We introduce the pixel-guided local alignment module (PLAM) and the semantic-guided global integration module (SGIM), which perform heterogeneous mutual refinement between high-level and low-level features according to their feature distinctions.

• Our HMHI-Net achieves state-of-the-art performance on all UVOS and video salient object detection (VSOD) benchmarks, with 89.8% $\mathcal{J}\&\mathcal{F}$ on DAVIS-16[44], 86.9% $\mathcal{J}$ on FBMS[38] and 76.2% $\mathcal{J}$ on YouTube-Objetcs[45]. Moreover, HMHI-Net consistently delivers high performance across different backbones, underscoring its superior generalization capability and robustness.

## 2 Related Work

### 2.1 Semi-supervised Video Object Segmentation

Semi-supervised Video Object Segmentation (SVOS) aims to segment the target objects throughout the subsequent video sequence by leveraging the given object masks in the first frame. STM[39] and STCN[7] pioneered a space-time memory that computes similarity between current and past frames to propagate masks, maintaining spatio-temporal consistency. To mitigate the increasing computational cost in long video sequences, methods such as AOT[69], XMem[6], Cutie[5] and SAM2[46] introduced a layered memory design, proposed long- and short-term memory or object tokens to compresses semantically rich features from distant frames. Other approaches, such as OneVOS[31], discarded explicit memory selection and instead input all previous frames, allowing the model to dynamically select and store informative key frame features.

Although matching-based techniques demonstrate strong performance in SVOS, their direct application to UVOS proves challenging. In the UVOS setting, where the first-frame mask is unavailable, representations encoded from high-level features inherently lack sufficient spatial details.

### 2.2 Unsupervised Video Object Segmentation

Unlike SVOS, Unsupervised Video Object Segmentation (UVOS) requires segmenting the most salient object in a video without prior supervision or manual annotations. Early UVOS methods [35, 51, 53, 60, 62, 68] primarily relied on exploiting appearance consistency across frames. A recent major trend in UVOS research involves leveraging optical flow to capture object motion and promotes segmentation. Representative works[23, 43, 47, 67, 72–74, 76] have developed various fusion modules to integrate optical flow with image appearance. Typically, these methods adopt either two-stream[9] or single-stream[20, 42] backbones to extract flow and image features, which are then combined through complex fusion mechanisms. Despite their contributions, flow-based methods remain constrained by short-term motion cues, making them prone to errors under occlusion or rapid motion.

Recent works[8, 13, 26, 27] attempt to address these limitations by incorporating long-term memory in UVOS. These methods typically fuse high-level visual and motion features from reference frames with the current frame to enhance segmentation. However, they only utilize the high-level features of the reference frames, failing to capture more fine-grained pixel-level information. To address this, we store both shallow- and high-level features in memory, supplying both precise details and strong semantic priors for accurate, consistent segmentation.

## 3 Methodology

### 3.1 Overall Pipeline

We employ a hierarchical backbone as the encoder following the conventional segmentation paradigm, which generates multi-scale features for decoding. At time $t$, the encoder takes the current frame image $I_t \in \mathbb{R}^{H\times W\times 3}$ and its corresponding optical flow map $O_t \in \mathbb{R}^{H\times W\times 3}$ as inputs, and extracts multi-scale features through four hierarchical layers. $I_t^i \in \mathbb{R}^{H_iW_i\times C_i}$ and $O_t^i \in \mathbb{R}^{H_iW_i\times C_i}$ ($i \in \{1,2,3,4\}$) represents the encoded feature of image and optical flow at $i$-th layer, where $H_iW_i = \frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}$. We simply add $I_i^t$ and $O_i^t$ from the same level to form the merged feature $F_t^i \in \mathbb{R}^{H_iW_i\times C_i}$ for further processing. The general encoding phase is briefly presented as follows:

$$\begin{aligned} I_t^i,\, O_t^i &= \text{Encoder}(I_t, O_t) \\ F_t^i &= I_t^i + O_t^i, \quad i \in \{1,2,3,4\} \end{aligned} \tag{1}$$

During the encoding phase, the encoded feature $F_t^2$ and $F_t^4$ are further optimized by extracting helpful knowledge from their relevant memory banks. To be more specific, the memory bank stores numerous history frames as the reference features $R_{t-1}^i \in \mathbb{R}^{TH_iW_i\times C_i}$ ($i \in \{2,4\}$), where $T$ is the number of memorized frames and $t-1$ is the time step denoting the reference for the current frame at time $t$. As presented in (2), we adopt a unified refinement architecture for both levels to attend to either spatial details or semantic object cues from their individual reference features.

$$\begin{aligned} F_t^{2'} &= \text{Mem\_Refine}(F_t^2, R_{t-1}^2) \\ F_t^{4'} &= \text{Mem\_Refine}(F_t^4, R_{t-1}^4) \end{aligned} \tag{2}$$

We further propose the heterogeneous interaction mechanism to promote mutual refinement of $F_t^{2'}$ and $F_t^{4'}$. Specifically, we propose the pixel-guided local alignment module (PLAM) for shallow-to-high refinement. PLAM adopts structure-preserving attention mechanism to retrieve fine-grained knowledge, which preserves the spatial layout and fine-grained structural information from the shallow-level feature $F_t^{2'}$. Additionally, the semantic-guided global integration module (SGIM) applies a global attention strategy to extract semantic cues from the high-level feature map $F_t^{4'}$ and aligns them with the shallow-level representation $F_t^{2'}$. The mutual refinement process is formulated as:

$$\begin{aligned} F_t^{2''} &= \text{SGIM}(F_t^{4'}, F_t^{2'}) \\ F_t^{4''} &= \text{PLAM}(F_t^{2'}, F_t^{4'}) \end{aligned} \tag{3}$$

We then use the updated multi-scale features $[F_t^1, F_t^{2''}, F_t^3, F_t^{4''}]$ as inputs to the hierarchical decoder. The decoder progressively

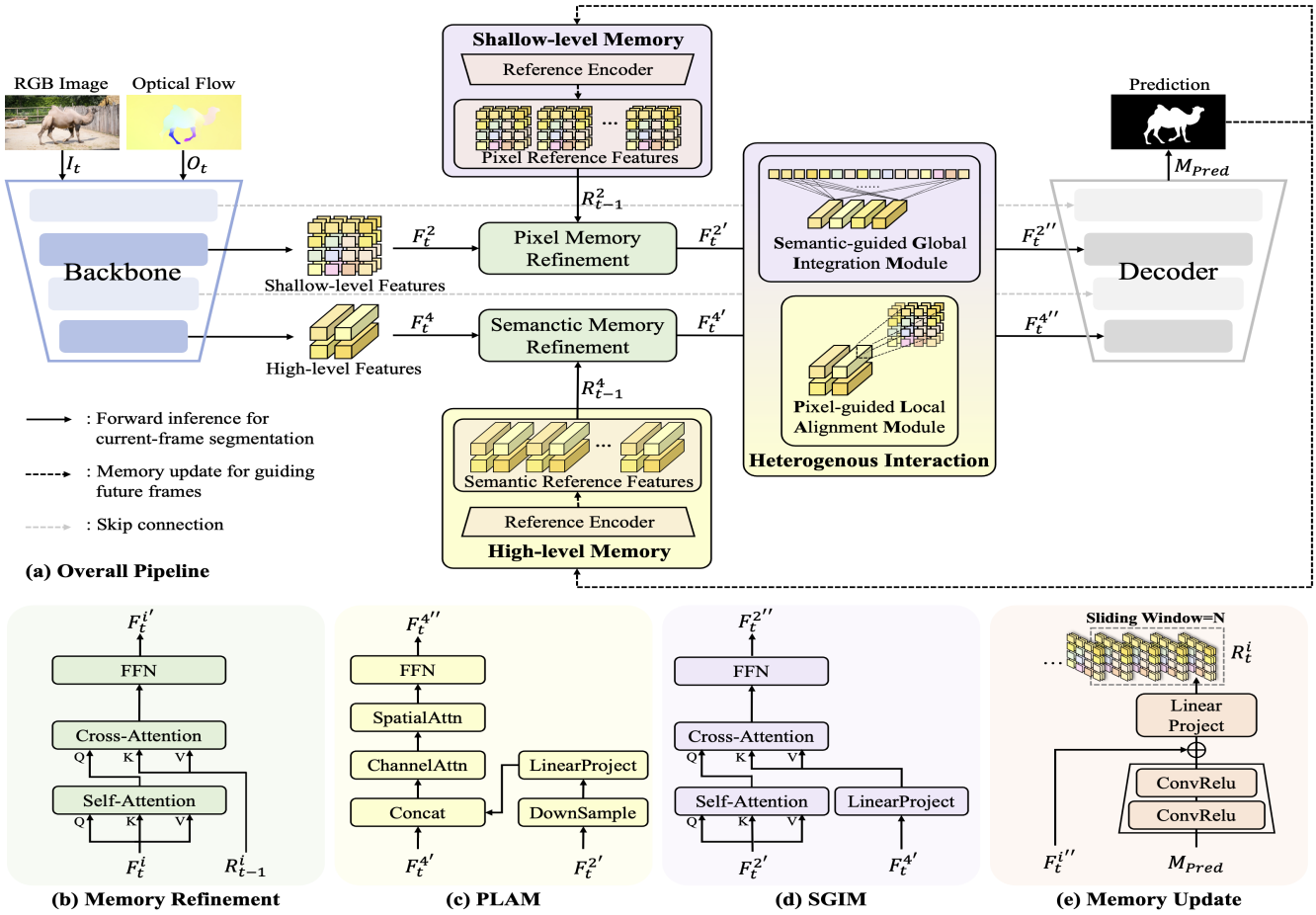



**Figure 3: (a) Overall pipeline of HMHI-Net. (b) Memory readout mechanism for frames refinement. (c) Pixel-guided local alignment module. (d) Semantic-guided global integration module. (e) Memory update mechanism with the reference encoder.**

upsamples the features from the top layer, fusing them with lower-level features in a bottom-up manner and ultimately generates the segmentation map $M_{\text{Pred}} \in \mathbb{R}^{H\times W\times 1}$. And $M_{\text{Pred}}$ is added to the memory banks together with the encoded features for memory update as follows:

$$\begin{aligned} M_{\text{Pred}} &= \text{Decoder}(F_t^1, F_t^{2''}, F_t^3, F_t^{4''}) \\ R_t^2 &= \text{Mem\_Update}(R_{t-1}^2, F_t^{2''}, M_{\text{Pred}}) \\ R_t^4 &= \text{Mem\_Update}(R_{t-1}^4, F_t^{4''}, M_{\text{Pred}}) \end{aligned} \tag{4}$$

## 3.2 Hierarchical Memory Structure

We construct our hierarchical memory architecture based on the four-layer hierarchical encoder. Normally, the first two layers are considered as the shallow level, which concentrate more on local information. And the last two layers are deemed as the high level, which mostly encode abstract representations. As illustrated in Fig.2, the second-layer feature $F_t^2$ possesses abundant pixel-level information and is encoded more sufficiently compared to $F_t^1$, which contains misleading focus on background pixels. Conversely, the fourth-layer features $F_t^4$ conveys the most compact semantic cues. Therefore, we choose the second and fourth layers to establish two separate memory banks, which correspond to the shallow- and high-level memory, respectively.

Given a current frame image $I_t$ and its optical flow $O_t$, we optimize the encoded features $F_t^i$ ($i \in \{2, 4\}$) using their corresponding reference features $R_t^i \in \mathbb{R}_{t-1}^{THW\times C_i}$ ($i \in \{2, 4\}$) from memory. In order to avoid misalignment between shallow- and high-level features, we apply the same memory mechanism to these separate memory banks. First of all, we first employ attention mechanism to feature $F_t^i$ ($i \in \{2, 4\}$) itself, to enhance $F_t^i$ by attending to its own internal representations. As expressed in (5), we projected $F_t^i$ into query $Q_{\text{sa}}^{i,t}$, key $K_{\text{sa}}^{i,t}$, and value $V_{\text{sa}}^{i,t}$ embeddings via three separate linear layers, and perform scaled dot-product attention accordingly.

$$\begin{aligned} F_t^{i'} &= \text{Attention}(Q_{\text{sa}}^{i,t}, K_{\text{sa}}^{i,t}, V_{\text{sa}}^{i,t}) \\ &= \text{Softmax}\left(\frac{Q_{\text{sa}}^{i,t\top} K_{\text{sa}}^{i,t}}{\sqrt{d}}\right) V_{\text{sa}}^{i,t} \end{aligned} \tag{5}$$

We further use the stored reference characteristics $R^i_{t-1}$ at time $t$ to optimize current feature $F^{i'}_t$. Memory banks contains the encode attributes of former predictions, which are relatively reliable supervision. We derive $Q^{i,t}_{\text{mem}}$ from the present feature $F^{i'}_t$, and $K^{i,t}_{\text{mem}}$, $V^{i,t}_{\text{mem}}$ from $R^i_t$, and calculate the per-token relationship between $F^{i'}_t$ and $R^i_t$ as shown in (6).

$$S_{\text{corr}} = (Q^{i,t\top}_{\text{mem}} K^{i,t}_{\text{mem}}) \,/\, \sqrt{d} \tag{6}$$

The attention map $S_{\text{corr}} \in \mathbb{R}^{H_2W_2 \times TH_2W_2}$ reveals the relation between the current frame and the reference frames, where each row indicates the similarity between a pixel in the current frame and all pixels in $T$ reference frames. We normalize $S_{\text{corr}}$ through a softmax function and retrieve information accordingly from $V^{i,t}_{\text{mem}}$ to refine $F^{i'}_t$. The read out attributes are added back to $F^{i'}_t$. Additionally, a feed-forward neural network (FFN) is applied to $F^{i'}_t$ to realign the feature vector back to the same representational space as those produced by the backbone. This memory readout process is formulated in (7).

$$\begin{aligned} F^{i'}_t &+= \text{Softmax}\left(S_{\text{corr}}\right) V^{i,t}_{\text{mem}} \\ F^{i'}_t &= \text{FFN}(F^{i'}_t) \end{aligned} \tag{7}$$

## 3.3 Heterogeneous Interaction Mechanism

To avoid feature misalignment between the shallow- and high-level memory banks due to their feature discrepancy, we propose the heterogeneous interaction mechanism to facilitate the pixel-semantic memory mutual interactions. The heterogeneous interaction mechanism takes advantage of the complementary nature of $F^{2'}_t$ and $F^{4'}_t$, and designs different modules for the shallow-to-high and high-to-shallow feature communications according to their character differences. Specifically, shallow features focus more on local patches, and relative positional relation is important when retrieving information from it. High-level features have more inclusive representations, which requires a broader perception field to fully exploit the semantic cues. The heterogeneous interaction mechanism consists of two key components: the pixel-guided local alignment module (PLAM) and the semantic-guided global integration module (SGIM).

**Pixel-guided Local Alignment Module**. PLAM is introduced to enrich the high-level feature $F^{4'}_t \in \mathbb{R}^{H_4W_4 \times C_4}$ with detailed structural information extracted from the shallow-level feature $F^{2'}_t \in \mathbb{R}^{H_2W_2 \times C_2}$. By incorporating fine-grained local cues into the high-level representation, the decoder is guided with structural priors from the very beginning, reducing confusion caused by background regions with similar semantics. PLAM first projects $F^{2'}_t$ to align the high-level feature $F^{4'}_t$ in dimension space via a series of downsample operations, producing the aligned features $F^{2\_\text{tmp}}_t \in \mathbb{R}^{H_4W_4 \times C_4}$ as in (8).

$$\begin{aligned} F^{2\_\text{tmp}}_t &= \text{ConvReLu}\left(F^{2'}_t\right) \\ F^{2\_\text{tmp}}_t &= \text{Linear}(F^{2\_\text{tmp}}_t) \end{aligned} \tag{8}$$

To preserve the spatial consistency from the shallow-level features, we directly concatenate the aligned features $F^{2\_\text{tmp}}_t$ with $F^{4'}_t$ at the $C$ dimension for unified representation $F^{4''}_t \in \mathbb{R}^{H_4W_4 \times 2C_4}$. Next, we employ the channel attention to re-emphasize channel-wise information which enhances the feature's semantic expressiveness at the channel level. This process is denoted as:

$$\begin{aligned} F^{4''}_t &= \text{Concat}(F^{4'}_t, F^{2\_\text{tmp}}_t) \\ F^{4''}_t &= \text{Channel\_Attn}(F^{4''}_t) \end{aligned} \tag{9}$$

Following this, spatial attention is employed to identify spatial locations that are crucial for object representation, which promotes the semantic focus on the target regions, yielding the shallow-enhanced high-level feature $F^{4''}_t$. Finally, an FFN is utilized to project the refined feature back to the original feature space as in (10), ensuring compatibility during the decoding phase.

$$\begin{aligned} F^{4''}_t &= \text{Spatial\_Attn}(F^{4''}_t) \\ F^{4''}_t &= \text{FFN}(F^{4''}_t) \end{aligned} \tag{10}$$

**Semantic-guided Global Integration Module**. To better inject high-level semantics into shallow-level features $F^{2'}_t$ and prevent dilution of semantic cues during decoding, we design the SGIM. Similarly, SGIM begins by aligning abstract high-level features $F^{4'}_t \in \mathbb{R}^{H_4W_4 \times C_4}$ to the shallow-level pixel feature space via a linear projector, producing aligned features $F^{4\_\text{tmp}}_t \in \mathbb{R}^{H_4W_4 \times C_4}$. Next, SGIM extracts $Q^{2,t}_{\text{sa}}, K^{2,t}_{\text{sa}}, V^{2,t}_{\text{sa}}$ from $F^{2'}_t$ and applies the attention mechanism to refine $F^{2'}_t$ by modeling stronger inner pixel-level relations, as expressed in (11).

$$\begin{aligned} F^{4\_\text{tmp}}_t &= \text{Linear}(F^{4'}_t) \\ F^{2'}_t &= \text{Attention}(Q^{2,t}_{\text{sa}}, K^{2,t}_{\text{sa}}, V^{2,t}_{\text{sa}}) \\ &= \text{Softmax}\left(\frac{Q^{2,t\top}_{\text{sa}} K^{2,t}_{\text{sa}}}{\sqrt{d}}\right) V^{2,t}_{\text{sa}} \end{aligned} \tag{11}$$

Although the aligned high-level features $F^{4\_\text{tmp}}_t$ share the same dimensionality as $F^{2'}_t$, they differ in spatial resolution and semantic abstraction. To integrate global semantic context into each pixel representation, we apply a global attention mechanism to $F^{4\_\text{tmp}}_t$ and $F^{2'}_t$, where $Q^{2,t}_{\text{ca}}$ is derived from $F^{2'}_t$ and $K^{4,t}_{\text{ca}}, V^{4,t}_{\text{ca}}$ are projected from $F^{4\_\text{tmp}}_t$:

$$\begin{aligned} F''_2 &+= \text{Attention}(Q^{2,t}_{\text{ca}}, K^{4,t}_{\text{ca}}, V^{4,t}_{\text{ca}}) \\ &= \text{Softmax}\left(\frac{Q^{2,t}_{\text{ca}} K^{4,t}_{\text{ca}}}{\sqrt{d}}\right) V^{4,t}_{\text{ca}} \end{aligned} \tag{12}$$

Finally, an FFN is applied to map the refined shallow-level features $F''_2$ back to their original representation space, completing the fusion process from object-level semantics to pixel-level details.

## 3.4 Memory Update

After generating the predicted mask $M_{\text{Pred}}$ of the current frame, we update two memory banks the final refined features $F''_2$ and $F''_4$, together with the predicted mask. Two simple memory encoders are employed to integrate $M_{\text{Pred}}$ into the shallow- and high-level refined features.

During memory bank updates, we adopt a sliding window strategy with maximum memory limit $N$. The memory bank $R_t^i \in \mathbb{R}^{TH_iW_i \times C_i}$, where $T \in \{1, 2, \ldots, N\}$, stores the most recent $T$ reference features. We update $R_t^i$ every $k$ frames following the first-in-first-out manner. Since no reference frame is available for the first frame in a video sequence, we only utilize the simple baseline without any memory refinement or heterogeneous interaction.

# 4 Experiments

## 4.1 Implementation Details

**Training and Inference**. Following [20, 33, 42, 75], we utilize mit_b1 as our backbone for fair comparison. We adopt the simple and efficient motion-appearance integration paradigm in [75] to avoid redundant discussion on the fusion mechanism, which is not an emphasis in this paper. We directly employ the common multi-scale decoder as [9, 42] for fair comparison. At the training stage, a sequence of five frames is selected for training in each iteration, where $k = 1$ and $T = 5$. The first frame skips the memory refinement and heterogeneous interaction modules, and only stores the shallow- and high-level features into their respective memory banks. We follow [46, 75] to adopt a combination of binary cross entropy loss, focal loss and dice loss for training. The final training loss is computed as the average of the segmentation losses from all five frames.

In the training stage, we adopt the AdamW optimizer with a learning rate of 6e-5, and train the model for 150 epochs on the YouTube-VOS datasets[66]. During fine-tuning, we set the learning rate to 1e-4 with a CosineAnnealingLR scheduler and train the model until convergence. All training and inference are conducted on four NVIDIA RTX 4090 GPUs.

For inference, $k$ is set to 1 and $T$ is set to 5 across all benchmarks for convenience as in prior works. Images are resized to $512 \times 512$ during both training and inference.

**Evaluation Metrics**. We assess model performance using a comprehensive set of metrics. For UVOS, we adopt region similarity $\mathcal{J}$, which evaluates segmentation accuracy via intersection-over-union (IoU). Boundary accuracy $\mathcal{F}$ measures the quality of mask contours through F1 score computation. Their average, $\mathcal{J}\&\mathcal{F}$, serves as the overall performance indicator.

For VSOD, we report mean absolute error (MAE) to quantify pixel-level prediction accuracy, maximum F-measure ($F_m$) to capture the best precision-recall tradeoff, enhanced-alignment measure ($E_m$) to reflect both pixel-wise and global consistency, and structure-measure ($S_m$) to evaluate region-aware and object-aware structural similarity.

## 4.2 Quantitative Results

**UVOS Performance**. We first pretrain the model on YouTube-VOS[66] and finetune the pre-trained model on DAVIS-16[44] or FBMS[38] datasets for evaluations. We directly adopts the model in the pre-training stage to test on YouTube-Object[45]. We compare our proposed model with previous UVOS approaches across these benchmarks. As shown in Table 1, our model achieves state-of-the-art performance on three widely used UVOS benchmarks: DAVIS-16[44], FBMS[38], and YouTube-Objects[45], and outperforms the most recent state-of-the-art methods by 1.6%, 3.5%, and 1.5% respectively. By analyzing Table 1, we observe that among all previous memory-free approaches, algorithms like FSNet[23] and TransportNet[72] which rely solely on high-level motion-appearance fusion typically underperform compared to those methods incorporating multi-level feature alignment, such as HFAN[42], TMO[9], and SimulFlow[20]. This implicitly highlighted the inadequacy of high-level features along for generating satisfactory results. Furthermore, memory-based methods, such as [8, 13, 27], achieve some advancements, which highlights the contribution of long-term memory in providing richer information. Our HMHI-Net incorporates both shallow- and high-level features into the long-tern memory, which enables robust and precise boundary segmentation under challenging and rapidly changing scenarios.

**VSOD Performance**. Following previous works, we further finetune HMHI-Net on a mixed dataset of DAVIS-16 and DAVSOD[12], and evaluate our model on four viedo salient object detection benchmarks: DAVIS-16, FBMS, ViSal[61], and DAVSOD. As shown in Table 2, our model achieves the best performance on all datasets and evaluation metrics with great margins, except for MAE on FBMS, where it ranks the third. Models using long-term memory, such as [13], achieves better results on metrics like $MAE$ and $F_m$, indicating the strength of long-term memory in preserving global semantic consistency. Our HMHI-Net achieves top results across multiple datasets and metrics, demonstrating its outstanding capability in saliency detection.

## 4.3 Qualitative Results

To intuitively demonstrate the capability of our model on both the UVOS and VSOD tasks, we provide qualitative visualizations of the segmentation results in several challenging scenarios in Fig. 4 and Fig. 5. Our model delivers consistently accurate and complete detections across all challenging UVOS and VSOD cases, highlighting the model's robustness and generalization capability in handling diverse and challenging conditions.

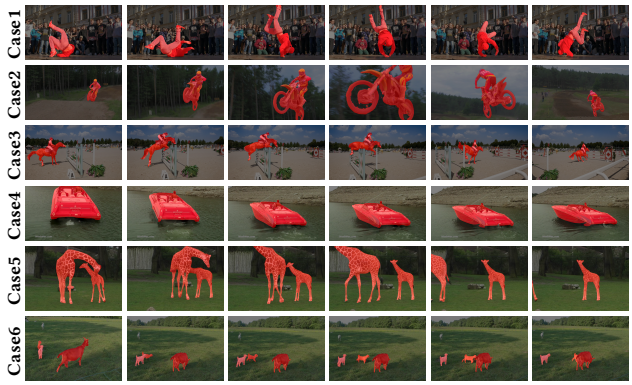



**Figure 4: Qualitative visualization of segmentation results on multiple challenging scenarios, including rapid movement, fine-grained segmentation, motion blur, multiple objects, viewpoint variation, etc. Case 1-3: *breakdance, motocross-jump, horsejump-high* from DAVIS-16; Case 4: *boat0001_00161* from YouTube-Obects; Case 5-6: *giraffes01, goats01* from FBMS.**

**Table 1: Evaluation results on three UVOS benchmarks: DAVIS-16, FBMS, and YouTube-Objects. Methods employing optical flow are marked with 'OF', while 'PP' indicates the use of post-processing. The top-performing and runner-up methods are emphasized using bold and underline formatting, respectively.**

| Method | Publication | Backbone | OF | PP | DAVIS-16 | | | FBMS | YTO |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{J}$ |
| PDB[53] | ECCV'18 | ResNet-50 | | ✓ | 75.9 | 77.2 | 74.5 | 74.0 | - |
| COSNet[35] | CVPR'19 | DeepLabv3 | | ✓ | 80.0 | 80.5 | 79.4 | 75.6 | 70.5 |
| AGNN[60] | ICCV'19 | DeepLabV3 | | ✓ | 79.9 | 80.7 | 79.1 | - | 70.8 |
| MATNet[76] | AAAI'20 | ResNet-101 | ✓ | ✓ | 81.6 | 82.4 | 80.7 | 76.1 | 69.0 |
| DFNet[74] | ECCV'20 | DeepLabv3 | | ✓ | 82.6 | 83.4 | 81.8 | - | - |
| RTNet[47] | CVPR'21 | ResNet-101 | ✓ | ✓ | 85.2 | 85.6 | 84.7 | - | 71.0 |
| TransportNet[72] | ICCV'21 | ResNet-101 | ✓ | | 84.8 | 84.5 | 85.0 | 78.7 | - |
| AMC-Net[67] | ICCV'21 | ResNet-101 | ✓ | ✓ | 84.6 | 84.5 | 84.6 | 76.5 | 71.1 |
| IMP[28] | AAAI'22 | ResNet-50 | | | 85.6 | 84.5 | 86.7 | 77.5 | - |
| HFAN[42] | ECCV'22 | Mit-b1 | ✓ | | 86.7 | 86.2 | 87.1 | - | 73.4 |
| HCPN[43] | TIP'23 | ResNet-101 | ✓ | ✓ | 85.6 | 85.8 | 85.4 | 78.3 | 73.3 |
| PMN[27] | WACV'23 | VGG-16 | ✓ | | 85.9 | 85.4 | 86.4 | 77.7 | 71.8 |
| TMO[9] | WACV'23 | ResNet-101 | ✓ | | 86.1 | 85.6 | 86.6 | 79.9 | 71.5 |
| OAST[55] | ICCV'23 | MobileViT | ✓ | | 87.0 | 86.6 | 87.4 | 83.0 | - |
| TGFormer[13] | ACMMM'23 | MobileViT | | | 86.3 | 85.8 | 86.7 | 84.0 | - |
| SimulFlow[20] | ACMMM'23 | Mit-b1 | ✓ | | 87.4 | 86.9 | 88.0 | 80.4 | 72.9 |
| HGPU[41] | TIP'24 | ResNet-101 | ✓ | | 86.1 | 86.0 | 86.2 | - | 73.9 |
| DPA[8] | CVPR'24 | VGG-16 | ✓ | | 87.6 | 86.8 | 88.4 | 83.4 | 73.7 |
| GSA[26] | CVPR'24 | ResNet-101 | ✓ | | 87.7 | 87.0 | 88.4 | 83.1 | - |
| DTTT[33] | CVPR'24 | Mit-b1 | ✓ | | 87.2 | 85.8 | 88.5 | 78.8 | - |
| GFA[52] | AAAI'24 | - | ✓ | | 88.2 | 87.4 | 88.9 | 82.4 | 74.7 |
| GFA[52] | AAAI'24 | ResNet-101 | ✓ | | 86.3 | 85.9 | 86.7 | 80.1 | 73.6 |
| **Ours** | - | Mit-b1 | ✓ | | **89.8** | **88.6** | **91.0** | **86.9** | **76.2** |

**Table 2: Quantitative comparison on the VSOD benchmarks: DAVIS-16, DAVSOD, ViSal, and FBMS. In the table, results marked with * are reproduced using the official released code. ↑ denotes that higher values indicate better performance, while ↓ implies the opposite. Numbers indicated in bold and underline represent the best and second-best scores, respectively.**

| Method | DAVSOD | | | | DAVIS-16 | | | | ViSal | | | | FBMS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $MAE\downarrow$ | $F_m\uparrow$ | $E_m\uparrow$ | $S_m\uparrow$ | $MAE\downarrow$ | $F_m\uparrow$ | $E_m\uparrow$ | $S_m\uparrow$ | $MAE\downarrow$ | $F_m\uparrow$ | $E_m\uparrow$ | $S_m\uparrow$ | $MAE\downarrow$ | $F_m\uparrow$ | $E_m\uparrow$ | $S_m\uparrow$ |
| MATNet*[76] | 0.098 | 0.628 | 0.789 | 0.707 | 0.048 | 0.752 | 0.890 | 0.776 | 0.041 | 0.891 | 0.967 | 0.863 | 0.091 | 0.751 | 0.852 | 0.760 |
| RTNet*[47] | 0.068 | 0.647 | 0.782 | 0.743 | 0.012 | 0.928 | 0.978 | 0.933 | 0.019 | 0.938 | 0.975 | 0.936 | 0.057 | 0.845 | 0.892 | 0.855 |
| FSNet[23] | 0.072 | 0.685 | 0.825 | 0.773 | 0.020 | 0.907 | 0.970 | 0.920 | - | - | - | - | 0.041 | 0.888 | 0.935 | 0.890 |
| TransportNet[72] | - | - | - | - | 0.013 | 0.928 | - | - | **0.012** | 0.953 | - | - | 0.045 | 0.885 | - | - |
| HFAN*[42] | 0.078 | 0.656 | 0.795 | 0.763 | 0.014 | 0.930 | 0.984 | 0.939 | 0.029 | 0.860 | 0.928 | 0.891 | 0.065 | 0.794 | 0.877 | 0.818 |
| TGFormer[13] | 0.065 | 0.728 | - | 0.798 | 0.011 | 0.922 | - | 0.932 | 0.011 | 0.955 | - | 0.952 | 0.026 | 0.919 | - | 0.916 |
| HCPN*[43] | 0.072 | 0.684 | 0.818 | 0.774 | 0.017 | 0.923 | 0.980 | 0.932 | 0.016 | 0.942 | 0.986 | 0.945 | 0.060 | 0.850 | 0.903 | 0.851 |
| TMO*[9] | 0.062 | 0.731 | 0.849 | 0.805 | 0.013 | 0.925 | 0.982 | 0.936 | 0.013 | 0.951 | 0.989 | 0.951 | 0.036 | 0.887 | 0.933 | 0.893 |
| OAST[55] | 0.070 | 0.712 | - | 0.786 | 0.011 | 0.926 | - | 0.935 | - | - | - | - | **0.025** | 0.919 | - | 0.917 |
| SimulFlow[20] | 0.069 | 0.722 | - | 0.771 | **0.009** | 0.936 | - | 0.937 | **0.012** | 0.943 | - | 0.946 | - | - | - | - |
| **Ours** | **0.054** | **0.801** | **0.896** | **0.847** | **0.009** | **0.947** | **0.990** | **0.951** | **0.012** | **0.962** | **0.991** | **0.960** | 0.030 | **0.946** | **0.977** | **0.930** |

## 4.4 Ablation Study

**Impact of Memory Layer Selection**. We examined the impact of memory mechanism at different encoder layers on UVOS performance. As shown in Tab. 3, incorporating memory at various levels consistently improves segmentation results. The memory applied at the second layer yields the best performance with an increase of 0.8% $\mathcal{J}\&\mathcal{F}$ on DAVIS-16, followed by that at the third layer with a 0.6% $\mathcal{J}\&\mathcal{F}$ increase. In contrast, memory at the final layer brings only marginal gains, supporting our theoretical claim that UVOS inherently lacks pixel-level supervision. Consequently, relying solely on high-level semantic memory bring about limited benefits for fine-grained object segmentation. We also tested their resource consumption, including the inference speed **Spd**, GPU memory usage **GPU** and the number of parameters **Pars**.

**Table 3: Ablation study on the impact of memory layer selection**

| Variant | DAVIS16 $\mathcal{J}\&\mathcal{F}$ | FBMS $\mathcal{J}$ | YTBOBJ $\mathcal{J}$ | **Spd** *FPS* | **GPU** *MB* | **Pars** *M* |
|---|---|---|---|---|---|---|
| baseline | 88.4 | 84.7 | 75.1 | 34.4 | 140.0 | 36.7 |
| layer = 1 | 88.9 **(+0.5)** | 86.0 **(+1.3)** | 75.7 **(+0.6)** | 9.2 | 141.5 | 37.1 |
| layer = 2 | 89.2 **(+0.8)** | 85.9 **(+1.2)** | 75.9 **(+0.8)** | 31.5 | 143.9 | 37.7 |
| layer = 3 | 89.0 **(+0.6)** | 85.2 **(+0.5)** | 75.6 **(+0.5)** | 30.9 | 156.9 | 41.1 |
| layer = 4 | 88.6 **(+0.2)** | 84.9 **(+0.2)** | 76.1 **(+1.0)** | 31.8 | 178.0 | 46.7 |

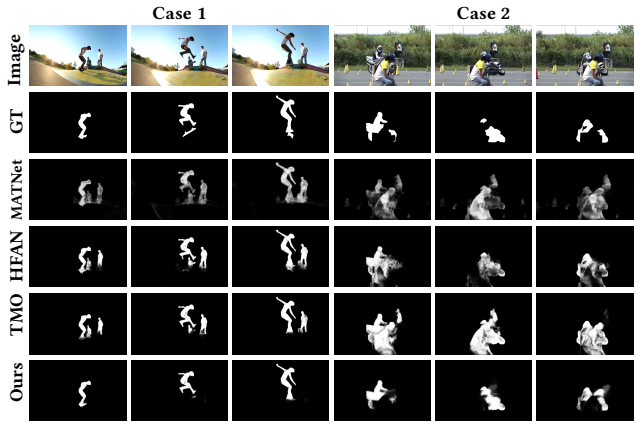



**Figure 5: Visual demonstration and comparison of the UVOS models' performance on the challenging VSOD scenarios. Case 1-2: *select_019, select_0194* from DAVSOD.**

**Effectiveness of the Proposed Modules**. As shown in Tab. 4, even without cross-memory feature interactions, the hierarchical memory architecture alone (denoted as Hier-Mem) already outperforms models with single-level memory. Introducing SGIM for high-to-shallow (abbreviated as H2S) and PLAM for shallow-to-high (abbreviated as S2H) interactions realizes the total performance improvements of 1.3% and 1.0% on DAVIS-16, respectively. And HMHI-Net achieves great improvements on all three UVOS bencmarks over the baseline model. To verify the heterogeneity of the interaction mechanisms, we swap SGIM and PLAM during shallow-high mutual refinement, which is denoted as 'Hetero' in Tab. 4. And the degradation of performance in both cases confirms the necessity of our heterogeneous interaction mechanism. Additionally, we evaluated the efficiency of HMHI-Net and different modules on a single 4090 GPU, proving that HMHI-Net reaches the speed for reality application.

**Table 4: Ablation study on the effectiveness of the proposed modules**

| Variant | | DAVIS16 $\mathcal{J}\&\mathcal{F}$ | FBMS $\mathcal{J}$ | YTBOBJ $\mathcal{J}$ | Speed *FPS* | Params *M* |
|---|---|---|---|---|---|---|
| Modules | baseline | 88.4 | 84.7 | 75.1 | 34.4 | 36.7 |
| | *w/* Multi-Mem | 89.3 **(+0.9)** | 86.0 **(+1.3)** | 76.1 **(+1.0)** | 27.8 | 47.7 |
| | + S2H *w/* PLAM | 89.4 **(+1.0)** | 86.3 **(+1.6)** | 75.6 **(+0.5)** | 27.3 | 60.0 |
| | + H2S *w/* SGIM | 89.7 **(+1.3)** | 86.5 **(+1.8)** | 75.3 **(+0.2)** | 26.9 | 48.4 |
| | HMHI-Net | 89.8 **(+1.4)** | 86.9 **(+2.1)** | 76.2 **(+1.1)** | 26.2 | 60.8 |
| Hetero | S2H *w/* SGIM | 89.1 **(-0.3)** | 85.5 **(-0.8)** | 73.9 **(-1.7)** | - | - |
| | H2S *w/* PLAM | 89.3 **(-0.4)** | 84.8 **(-1.7)** | 75.7 (+0.4) | - | - |

**Evaluation of Model Robustness Across Backbones**. Finally, to assess the robustness of our proposed design, we integrate the hierarchical memory structure and heterogeneous interaction mechanism into various backbone architectures, including swin_tiny from Swin-Transformer [34] and mit_b1, mit_b2, mit_b3 from SegFormer [65]. Mark * denotes the use of the original backbone, yet others indicates modification following [75]. Due to the large number of parameters in mit_b2 and mit_b3, baselines and HMHI-Net with these two backbones may not be fully trained. However, results presented in Tab. 5 still show consistent and significant performance improvements across all backbones, further validating the effectiveness and versatility of our approach.

**Table 5: Evaluation of model robustness across backbones**

| Variant | | DAVIS16 $\mathcal{J}\&\mathcal{F}$ | FBMS $\mathcal{J}$ | YTBOBJ $\mathcal{J}$ |
|---|---|---|---|---|
| mit_b1* | baseline | 87.8 | 82.5 | 75.3 |
| | HMHI-Net | 89.1 **(+1.3)** | 84.0 **(+3.0)** | 75.2 |
| mit_b2 | baseline | 88.6 | 86.0 | 76 |
| | HMHI-Net | 89.6 **(+1.0)** | 86.5 **(+0.5)** | 75.7 |
| mit_b3 | baseline | 88.0 | 86.4 | 76.4 |
| | HMHI-Net | 89.6 **(+1.6)** | 87.2 **(+0.8)** | 77.3 **(+0.9)** |
| swin_tiny | baseline | 88.4 | 84.7 | 73.8 |
| | HMHI-Net | 89.4 **(+1.0)** | 85.5 **(+0.8)** | 76.1 **(+2.3)** |

**Table 6: Ablation study on the influence of model inputs**

| Variant | | DAVIS16 $\mathcal{J}\&\mathcal{F}$ | FBMS $\mathcal{J}$ | YTBOBJ $\mathcal{J}$ |
|---|---|---|---|---|
| baseline | only_flow | 78.8 | 63.8 | 60.0 |
| | only_image | 83.4 | 80.4 | 73.8 |
| HMHI-Net | only_flow | 82.3 **(+3.5)** | 66.8 **(+3.0)** | 62.1 **(+2.1)** |
| | only_image | 84.4 **(+1.0)** | 82.8 **(+2.4)** | 75.8 **(+2.0)** |
| | flow & image | 89.8 | 86.9 | 76.2 |

**Influence of Model Inputs**. We also studied the contribution of optical flow and RGB images as inputs. As shown in Tab. 6, taking both features as inputs notably promotes model performance. Furthermore, even with single input, our model consistently surpasses the baseline under identical conditions with great margins. This validates HMHI-Net's capacity to effectively leverage video temporal cues to enhance segmentation.

## 5 Conclusion

We propose a simple and efficient hierarchical memory architecture with heterogeneous interaction mechanism for UVOS, which leverages both high-level features and shallow-level features for memory and performs different interactions during the shallow-high mutual refinement. Our model achieves state-of-the art performance on all UVOS and VSOD benchmarks. However, the hierarchical memory mechanism might lead to computation and storage overload, which can impact the model efficiency and is worth further investigation.

## Acknowledgments

This work was supported in part by CAAI-Lenovo Blue Sky Research Fund (No. CAAI-LXJJ 2024-05), and Scientific and Technological innovation action plan of Shanghai Science and Technology Committee (No.22511101502).

## References


[1] Olivier Barnich and Marc Van Droogenbroeck. 2011. ViBe: A Universal Background Subtraction Algorithm for Video Sequences. *IEEE Transactions on Image Processing* 20 (2011), 1709–1724. https://api.semanticscholar.org/CorpusID:783186

[2] Thomas Brox and Jitendra Malik. 2010. Object Segmentation by Long Term Analysis of Point Trajectories. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:16608752

[3] Sebastian Brutzer, Benjamin Höferlin, and Gunther Heidemann. 2011. Evaluation of background subtraction techniques for video surveillance. *CVPR 2011* (2011), 1937–1944. https://api.semanticscholar.org/CorpusID:206591471

[4] Lin Chen, Jianbing Shen, Wenguan Wang, and Bingbing Ni. 2015. Video Object Segmentation Via Dense Trajectories. *IEEE Transactions on Multimedia* 17 (2015), 2225–2234. https://api.semanticscholar.org/CorpusID:10157303

[5] Ho Kei Cheng, Seoung Wug Oh, Brian L. Price, Joon-Young Lee, and Alexander G. Schwing. 2023. Putting the Object Back into Video Object Segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 3151–3161. https://api.semanticscholar.org/CorpusID:264305820

[6] 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*. https://api.semanticscholar.org/CorpusID:250526250

[7] 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*. https://api.semanticscholar.org/CorpusID:235376958

[8] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Dogyoon Lee, and Sangyoun Lee. 2022. Dual Prototype Attention for Unsupervised Video Object Segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 19238–19247. https://api.semanticscholar.org/CorpusID:257532674

[9] Suhwan Cho, Minhyeok Lee, Seung-Hyun Lee, Chaewon Park, Donghyeon Kim, and Sangyoun Lee. 2022. Treating Motion as Option to Reduce Motion Dependency in Unsupervised Video Object Segmentation. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), 5129–5138. https://api.semanticscholar.org/CorpusID:252111197

[10] Jifeng Dai, Kaiming He, and Jian Sun. 2015. Instance-Aware Semantic Segmentation via Multi-task Network Cascades. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 3150–3158. https://api.semanticscholar.org/CorpusID:8510667

[11] A. Elgammal, Ramani Duraiswami, David Harwood, and Larry S. Davis. 2002. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proc. IEEE* 90 (2002), 1151–1163. https://api.semanticscholar.org/CorpusID:751988

[12] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. 2019. Shifting More Attention to Video Salient Object Detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 8546–8556. https://api.semanticscholar.org/CorpusID:198905006

[13] Jiaqing Fan, Tiankang Su, Kaihua Zhang, Bo Liu, and Qingshan Liu. 2023. Temporally Efficient Gabor Transformer for Unsupervised Video Object Segmentation. *Proceedings of the 31st ACM International Conference on Multimedia* (2023). https://api.semanticscholar.org/CorpusID:264492527

[14] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. 2012. Video segmentation by tracing discontinuities in a trajectory embedding. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 1846–1853. https://api.semanticscholar.org/CorpusID:2980297

[15] Ken Fukuchi, Kouji Miyazato, Akisato Kimura, Shigeru Takagi, and Junji Yamato. 2009. Saliency-based video segmentation with graph cuts and sequentially updated priors. *2009 IEEE International Conference on Multimedia and Expo* (2009), 638–641. https://api.semanticscholar.org/CorpusID:15569415

[16] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. 2017. Video Captioning With Attention-Based LSTM and Semantic Consistency. *IEEE Transactions on Multimedia* 19 (2017), 2045–2055. https://api.semanticscholar.org/CorpusID:25497516

[17] Daniela Giordano, Francesca Murabito, Simone Palazzo, and Concetto Spampinato. 2015. Superpixel-based video object segmentation using perceptual organization and location prior. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 4814–4822. https://api.semanticscholar.org/CorpusID:1481961

[18] Vitor Campanholo Guizilini and Fabio Tozeto Ramos. 2013. Online self-supervised segmentation of dynamic objects. *2013 IEEE International Conference on Robotics and Automation* (2013), 4720–4727. https://api.semanticscholar.org/CorpusID:17349052

[19] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770–778. https://api.semanticscholar.org/CorpusID:206594692

[20] Lingyi Hong, Wei Zhang, Shuyong Gao, Hong Lu, and Wenqiang Zhang. 2023. SimulFlow: Simultaneously Extracting Feature and Identifying Target for Unsupervised Video Object Segmentation. *Proceedings of the 31st ACM International Conference on Multimedia* (2023). https://api.semanticscholar.org/CorpusID:264492041

[21] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2021. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 9914–9925. https://api.semanticscholar.org/CorpusID:244729413

[22] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv* abs/2106.09685 (2021). https://api.semanticscholar.org/CorpusID:235458009

[23] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. 2021. Full-duplex strategy for video object segmentation. *Computational Visual Media* 9 (2021), 155–175. https://api.semanticscholar.org/CorpusID:236950747

[24] Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2019. Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:208176127

[25] Dejiang Kong and Fei Wu. 2018. HST-LSTM: A Hierarchical Spatial-Temporal Long-Short Term Memory Network for Location Prediction. In *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:51606411

[26] Minhyeok Lee, Suhwan Cho, Dogyoon Lee, Chaewon Park, Jungho Lee, and Sangyoun Lee. 2023. Guided Slot Attention for Unsupervised Video Object Segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 3807–3816. https://api.semanticscholar.org/CorpusID:257532769

[27] Minhyeok Lee, Suhwan Cho, Seung-Hyun Lee, Chaewon Park, and Sangyoun Lee. 2022. Unsupervised Video Object Segmentation via Prototype Memory Network. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), 5913–5923. https://api.semanticscholar.org/CorpusID:252118837

[28] Youngjo Lee, Hongje Seong, and Euntai Kim. 2021. Iteratively Selecting an Easy Reference Frame Makes Unsupervised Video Object Segmentation Easier. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:245424944

[29] Alex Levinshtein, Cristian Sminchisescu, and Sven J. Dickinson. 2012. Optimal Image and Video Closure by Superpixel Grouping. *International Journal of Computer Vision* 100 (2012), 99–119. https://api.semanticscholar.org/CorpusID:468651

[30] Wanyun Li, Jack Fan, Pinxue Guo, Lingyi Hong, and Wei Zhang. 2024. HFVOS: History-Future Integrated Dynamic Memory for Video Object Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 34 (2024), 10208–10222. https://api.semanticscholar.org/CorpusID:270039383

[31] Wanyun Li, Pinxue Guo, Xinyu Zhou, Lingyi Hong, Yangji He, Xiangyu Zheng, Wei Zhang, and Wenqiang Zhang. 2024. OneVOS: Unifying Video Object Segmentation with All-in-One Transformer Framework. *ArXiv* abs/2403.08682 (2024). https://api.semanticscholar.org/CorpusID:268379457

[32] Daizong Liu, Dongdong Yu, Changhu Wang, and Pan Zhou. 2020. F2Net: Learning to Focus on the Foreground for Unsupervised Video Object Segmentation. *ArXiv* abs/2012.02534 (2020). https://api.semanticscholar.org/CorpusID:227305215

[33] Weihuang Liu, Xi Shen, Haolun Li, Xiu-Li Bi, Bo Liu, Chi-Man Pun, and Xiaodong Cun. 2024. Depth-Aware Test-Time Training for Zero-Shot Video Object Segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 19218–19227. https://api.semanticscholar.org/CorpusID:268264338

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 9992–10002. https://api.semanticscholar.org/CorpusID:232352874

[35] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Murat Porikli. 2019. See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 3618–3627. https://api.semanticscholar.org/CorpusID:198352766

[36] T. Uma Mageswari. 2016. Density Based Multifeature Background Subtraction with Relevance Vector Machine. *Artificial Intelligent Systems and Machine Learning* 8 (2016), 272–274. https://api.semanticscholar.org/CorpusID:64841243

[37] Sabarinath Mahadevan, Ali Athar, Aljosa Osep, Sebastian Hennen, Laura Leal-Taixé, and B. Leibe. 2020. Making a Case for 3D Convolutions for Object Segmentation in Videos. *ArXiv* abs/2008.11516 (2020). https://api.semanticscholar.org/CorpusID:221319536

[38] Peter Ochs, Jitendra Malik, and Thomas Brox. [n. d.]. Ieee Transactions on Pattern Analysis and Machine Intelligence Segmentation of Moving Objects by Long Term Video Analysis. https://api.semanticscholar.org/CorpusID:12351806

[39] Seoung Wug Oh, Joon-Young Lee, N. Xu, and Seon Joo Kim. 2019. Video Object Segmentation Using Space-Time Memory Networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 9225–9234. https://api.semanticscholar.org/CorpusID:90262243

[40] Anestis Papazoglou and Vittorio Ferrari. 2013. Fast Object Segmentation in Unconstrained Video. *2013 IEEE International Conference on Computer Vision* (2013), 1777–1784. https://api.semanticscholar.org/CorpusID:3194346

[41] Gensheng Pei, Fumin Shen, Yazhou Yao, Tao Chen, Xian-Sheng Hua, and Heng Tao Shen. 2023. Hierarchical Graph Pattern Understanding for Zero-Shot VOS. *ArXiv* abs/2312.09525 (2023). https://api.semanticscholar.org/CorpusID:266335664

[42] Gensheng Pei, Fumin Shen, Yazhou Yao, Guosen Xie, Zhenmin Tang, and Jinhui Tang. 2022. Hierarchical Feature Alignment Network for Unsupervised Video Object Segmentation. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:250627320

[43] Gensheng Pei, Yazhou Yao, Fumin Shen, Daniel Huang, Xing-Rui Huang, and Hengtao Shen. 2023. Hierarchical Co-Attention Propagation Network for Zero-Shot Video Object Segmentation. *IEEE Transactions on Image Processing* 32 (2023), 2348–2359. https://api.semanticscholar.org/CorpusID:258049188

[44] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 724–732. https://api.semanticscholar.org/CorpusID:1949934

[45] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. 2012. Learning object class detectors from weakly annotated video. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 3282–3289. https://api.semanticscholar.org/CorpusID:7952817

[46] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya K. Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Doll'ar, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *ArXiv* abs/2408.00714 (2024). https://api.semanticscholar.org/CorpusID:271601113

[47] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. 2021. Reciprocal Transformations for Unsupervised Video Object Segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 15430–15439. https://api.semanticscholar.org/CorpusID:235703427

[48] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. 2021. Hierarchical Memory Matching Network for Video Object Segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 12869–12878. https://api.semanticscholar.org/CorpusID:237605255

[49] Jianbo Shi and Jitendra Malik. 1997. Normalized cuts and image segmentation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1997), 731–737. https://api.semanticscholar.org/CorpusID:14848918

[50] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, W. Liu, and Jian Yang. 2018. Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2018), 1110–1118. https://api.semanticscholar.org/CorpusID:53164806

[51] Mennatullah Siam, Chen Jiang, Steven Weikai Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, and Martin Jägersand. 2018. Video Object Segmentation using Teacher-Student Adaptation in a Human Robot Interaction (HRI) Setting. *2019 International Conference on Robotics and Automation (ICRA)* (2018), 50–56. https://api.semanticscholar.org/CorpusID:67749821

[52] Huihui Song, Tiankang Su, Yuhui Zheng, Kaihua Zhang, Bo Liu, and Dong Liu. 2024. Generalizable Fourier Augmentation for Unsupervised Video Object Segmentation. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:268692695

[53] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. 2018. Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:52954448

[54] Nasim Souly, Concetto Spampinato, and Mubarak Shah. 2017. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 5689–5697. https://api.semanticscholar.org/CorpusID:11996618

[55] Tiankang Su, Huihui Song, Dong Liu, Bo Liu, and Qingshan Liu. 2023. Unsupervised video object segmentation with online adversarial self-tuning. (2023), 688–698.

[56] Patrik Sundberg, Thomas Brox, Michael Maire, Pablo Arbeláez, and Jitendra Malik. 2011. Occlusion boundary detection and figure/ground assignment from optical flow. *CVPR 2011* (2011), 2233–2240. https://api.semanticscholar.org/CorpusID:9894725

[57] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and G. Hamarneh. 2019. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review* 54 (2019), 137 – 178. https://api.semanticscholar.org/CorpusID:204743865

[58] Andrew Tao, Karan Sapra, and Bryan Catanzaro. 2020. Hierarchical Multi-Scale Attention for Semantic Segmentation. *ArXiv* abs/2005.10821 (2020). https://api.semanticscholar.org/CorpusID:218763375

[59] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:13756489

[60] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. 2019. Zero-Shot Video Object Segmentation via Attentive Graph Neural Networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 9235–9244. https://api.semanticscholar.org/CorpusID:207968609

[61] Wenguan Wang, Jianbing Shen, and Ling Shao. 2015. Consistent Video Saliency Using Local Gradient Flow Optimization and Global Refinement. *IEEE Transactions on Image Processing* 24 (2015), 4185–4196. https://api.semanticscholar.org/CorpusID:4303753

[62] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven C. H. Hoi, and Haibin Ling. 2019. Learning Unsupervised Video Object Segmentation Through Visual Attention. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 3059–3069. https://api.semanticscholar.org/CorpusID:92995115

[63] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. 2022. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 4238–4247. https://api.semanticscholar.org/CorpusID:247315180

[64] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. 2018. CBAM: Convolutional Block Attention Module. *ArXiv* abs/1807.06521 (2018). https://api.semanticscholar.org/CorpusID:49867180

[65] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:235254713

[66] N. Xu, L. Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott D. Cohen, and Thomas S. Huang. 2018. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:52154988

[67] Shu-Hsiang Yang, Lu Zhang, Jinqing Qi, Huchuan Lu, Shuo Wang, and Xiaoxing Zhang. 2021. Learning Motion-Appearance Co-Attention for Zero-Shot Video Object Segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 1544–1553. https://api.semanticscholar.org/CorpusID:244306379

[68] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip H. S. Torr. 2019. Anchor Diffusion for Unsupervised Video Object Segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 931–940. https://api.semanticscholar.org/CorpusID:201710064

[69] Zongxin Yang, Yunchao Wei, and Yi Yang. 2021. Associating Objects with Transformers for Video Object Segmentation. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:235352901

[70] Tao Yu, Rui Zhang, Kai-Chou Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Z Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. *ArXiv* abs/1809.08887 (2018). https://api.semanticscholar.org/CorpusID:52815560

[71] Yichen Yuan, Yifan Wang, Lijun Wang, Xiaoqi Zhao, Huchuan Lu, Yu Wang, Wei Su, and Lei Zhang. 2023. Isomer: Isomerous Transformer for Zero-shot Video Object Segmentation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 966–976. https://api.semanticscholar.org/CorpusID:260887437

[72] Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu. 2021. Deep Transport Network for Unsupervised Video Object Segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 8761–8770. https://api.semanticscholar.org/CorpusID:245022221

[73] Lu Zhang, Jianming Zhang, Zhe L. Lin, Radomír Měch, Huchuan Lu, and You He. 2020. Unsupervised Video Object Segmentation with Joint Hotspot Tracking. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:226842087

[74] Mingmin Zhen, Shiwei Li, Lei Zhou, Jiaxiang Shang, Haoan Feng, Tian Fang, and Long Quan. 2020. Learning Discriminative Feature with CRF for Unsupervised Video Object Segmentation. *ArXiv* abs/2008.01270 (2020). https://api.semanticscholar.org/CorpusID:220961430

[75] Xiangyu Zheng, Wanyun Li, Songcheng He, Xiaoqiang Li, and We Zhang. 2025. Intrinsic Saliency Guided Trunk-Collateral Network for Unsupervised Video Object Segmentation. arXiv:2504.05904 [cs.CV] https://arxiv.org/abs/2504.05904

[76] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. 2020. MATNet: Motion-Attentive Transition Network for Zero-Shot Video Object Segmentation. *IEEE Transactions on Image Processing* 29 (2020), 8326–8338. https://api.semanticscholar.org/CorpusID:212633918

[77] Yunzhi Zhuge, Hongyu Gu, Lu Zhang, Jinqing Qi, and Huchuan Lu. 2024. Learning Motion and Temporal Cues for Unsupervised Video Object Segmentation. *IEEE transactions on neural networks and learning systems* PP (2024). https://api.semanticscholar.org/CorpusID:271062835