MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Deep Learning for Video Salient Object Detection

A thesis presented in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy
in
Computer Science

School of Mathematical and Computational Sciences,
Massey University, Albany, Auckland,
New Zealand

Tao Jiang

February 2025

Life was like a box of chocolates.
You never know what you're gonna get.

–Forrest Gump

Abstract

Video Salient Object Detection (VSOD) is a fundamental task in video analysis, focusing on identifying and segmenting the most visually prominent objects in dynamic scenes. In this thesis, we propose three novel deep learning-based approaches to enhance VSOD performance in complex environments.

Firstly, we introduce an Inheritance Enhancement Network (IENet), a Transformer-based framework designed to improve the integration of long-term spatial-temporal dependencies. We propose a unidirectional cross-frame enhancement mechanism, ensuring consistent and orderly information propagation across frames while minimizing interference. Extensive experiments demonstrate that IENet significantly improves detection accuracy in complex video scenes.

Secondly, we present a Knowledge-sharing Hierarchical Memory Fusion Network (KHMF-Net) to address the challenges of scribble-supervised VSOD, where annotations are sparse and ambiguous. Our approach utilizes a memory bank-based hierarchical encoder-decoder architecture to reduce error accumulation and mitigate background distractions. Additionally, we introduce a dual-attention knowledge-sharing strategy, enabling the model to refine predictions by leveraging learned information across frames. This approach effectively enhances feature consistency and improves the separation of foreground and background objects.

Thirdly, we propose the Multimodal Energy Prompting Network (MEPNet), which leverages optical flow and depth as implicit prompts to fine-tune a Segment Anything Model (SAM) for improved VSOD performance. We first introduce a Spectrogram Energy Generator (SEG), which extracts energy-driven prompts that enhance the spatial-temporal representation of salient objects. Furthermore, the Modality-Energy Adapter (MEA) effectively integrates these prompts into SAM, improving the model's ability to capture motion and structural cues. Extensive evaluations show that MEPNet effectively incorporates multimodal information, resulting in more robust and precise VSOD outcomes.

In summary, we propose three innovative approaches to improve VSOD in complex scenes. Each method undergoes extensive evaluation on benchmark datasets, achieving superior performance compared to existing state-of-the-art models. Our contributions provide new insights into advancing video-based salient object detection, paving the

way for more robust and efficient VSOD frameworks.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my main supervisor, Professor Ruili Wang, for his invaluable academic guidance and unwavering support throughout my doctoral journey. His profound expertise and insightful advice have been instrumental in shaping my research, helping me navigate complex challenges and develop innovative solutions. His scholarly vision and meticulous attention to detail have continuously inspired me to strive for academic excellence.

I am also deeply grateful to my co-supervisor, Dr. Feng Hou, whose encouragement and constructive feedback have played a crucial role in my research progress. His support has not only strengthened my confidence but has also provided me with the motivation to push beyond my limits. His insightful perspectives and practical guidance have helped me overcome numerous challenges, fostering both my academic and personal growth.

My sincere appreciation extends to my colleagues and friends, whose stimulating discussions and collaborative spirit have enriched my research experience. Their intellectual curiosity and willingness to exchange ideas have sparked new insights and encouraged me to explore novel directions. The vibrant research environment they have helped create has been invaluable to my academic journey.

Lastly, I am profoundly grateful to my parents, Yunlai Jiang and Jin'e Zhang, whose unwavering love and unconditional support have been my greatest source of strength. Their endless encouragement, patience, and belief in my abilities have been a constant source of motivation, guiding me through both the challenges and triumphs of this journey. This achievement would not have been possible without their steadfast support.

Publications

The following research papers have been published during my PhD study:

1. **Tao Jiang**, Yi Wang, Feng Hou and Li-li Liu, Enhancing video salient object detection via SAM-based multimodal energy prompting, In *Pattern Analysis and Applications* ,2025, URL: https://doi.org/10.1007/s10044-025-01531-9.

2. **Tao Jiang**, Feng Hou, and Ruili Wang, IENet: inheritance enhancement network for video salient object detection, In *Multimedia Tools and Applications*,2024, URL: https://doi.org/10.1007/s11042-024-18408-4.

3. **Tao Jiang**, Feng Hou, Yi Wang, Guangzhu Chen and Ruili Wang, Knowledge-Sharing Hierarchical Memory Fusion Network for Scribble-Supervised Video Salient Object Detection, In *Pattern Recognition Letters*, 2025, URL: https://doi.org/10.1016/j.patrec.2025.06.003.

4. **Tao Jiang**, Feng Hou, and Yi Wang, Multimodal Energy Prompting for Video Salient Object Detection, In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pp. 1-8, 2024, URL: https://doi.org/10.1145/3696409.3700196.

5. **Tao Jiang**, Ming Zong, Yujun Ma, Feng Hou, and Ruili Wang, MobileACNet: ACNet-based lightweight model for image classification. In *International Conference on Image and Vision Computing New Zealand*, pp. 361-372, 2022, URL:https://doi.org/10.1007/978-3-031-25825-126.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an overview of the thesis. Section 1.1 introduces the background of video salient object detection and reviews related work. Section 1.2 explores the limitations of existing approaches, which motivate our research. The research objectives are then presented in Section 1.3, followed by the organization and structure of the thesis in Section 1.4.

1.1 Overview of video salient object detection

Video Salient Object Detection (VSOD) aims to identify the most visually prominent objects in a video sequence by generating a saliency map for each frame, where pixel intensity represents the likelihood of belonging to a salient object [1, 2]. Inspired by the Human Visual System (HVS) [3, 4], VSOD simulates human attention mechanisms, enabling the efficient extraction of critical visual information while filtering out background distractions. Compared to static image-based SOD, VSOD must handle both the spatial features within each frame and the temporal dynamics across consecutive frames, making it a significantly more complex task. Given its ability to highlight meaningful objects in videos, VSOD has been widely applied in video retrieval [5], video segmentation [6], video summarization [7], autonomous driving [8] and video surveillance [9].

Video Salient Object Detection (VSOD) can be categorized based on supervision levels and input modalities. Fully supervised VSOD relies on dense, pixel-wise annotations to train deep learning models, achieving high accuracy but requiring extensive labeling

efforts. Weakly supervised VSOD leverages sparse annotations, such as scribbles or points-level labels, to reduce annotation costs while maintaining reasonable performance. Self-supervised VSOD employs contrastive learning or pretext tasks to learn robust spatial-temporal representations without explicit human annotations, while unsupervised VSOD eliminates reliance on labeled data by using intrinsic cues such as motion contrast and object uniqueness. From the perspective of input modalities, VSOD can be broadly divided into single-modality methods and multi-modality methods. Single-modality VSOD relies solely on RGB frames to extract spatial and temporal saliency cues. In contrast, multi-modality VSOD incorporates complementary information from additional sources, such as depth(RGB-D), and optical flow, to enhance robustness in complex environments. RGB-D VSOD utilizes depth information to improve structural understanding, and optical flow-based VSOD exploits motion dynamics to better differentiate salient objects from background motion.

Compared to static image-based SOD, Video Salient Object Detection (VSOD) faces unique challenges due to the dynamic nature of video scenes, which require capturing both motion and contextual features. Additionally, the demand for pixel-level annotations in video datasets makes the labeling process labor-intensive and time-consuming [10, 11]. Early VSOD methods relied on hand-crafted saliency features and spatiotemporal fusion techniques [12–14]. However, the complexity of video sequences and diverse movement patterns limited these methods' performance and their ability to capture long-term temporal information. With the rise of deep learning, 2D convolutional networks were adopted to extract spatial and temporal features from videos [15–17]. To enhance temporal modeling, Long Short-Term Memory (LSTM) networks were introduced to capture long-range dependencies [18]. Recent advancements have proposed multi-frame multi-scale encoder-decoder architectures that refine VSOD predictions using adversarial learning, edge extraction, and dense residual blocks [19–21]. In contrast to single-stream models, bi-stream-based approaches incorporate optical flow to capture explicit temporal information, enabling more effective spatiotemporal fusion [22–24]. Despite these advances, achieving effective spatial and temporal interactions in the decoding stage remains a key challenge.

While these methods have shown success in VSOD, they heavily rely on large, densely annotated training datasets, which are expensive and time-consuming to create. Traditional unsupervised VSOD methods [25–27] primarily rely on handcrafted features, limiting their effectiveness in real-world applications. Although semi-supervised ap-

proaches [28, 29] utilize a small amount of manually labeled data along with pseudo labels generated from existing saliency models, obtaining high-quality labeled data is still crucial for generating reliable pseudo labels. In contrast to methods that depend entirely or partially on fully annotated datasets, weakly supervised methods [30, 31] enable end-to-end training without requiring dense annotations, significantly reducing labeling costs.

Overall, the evolution of VSOD methodologies reflects a continuous pursuit of robust spatiotemporal feature representations and effective spatial-temporal interaction modeling, with hybrid architectures playing a crucial role in advancing the field. Our exploration primarily focuses on three key VSOD paradigms: fully supervised, weakly supervised, and multimodal VSOD, each addressing different challenges in balancing annotation efficiency, feature fusion, and detection robustness.

1.2 Motivations of this research

Despite significant progress in video salient object detection, several challenges persist in the field.

- **Insufficient synergy between spatial and temporal information**.

  Recent VSOD approaches employ memory-based [21, 32] and Transformer-based [33, 34] architectures for spatial-temporal modeling. However, memory-based methods often treat historical frames as unordered features, ignoring structured spatial-temporal relationships. Meanwhile, Transformers, despite capturing long-range dependencies, lack explicit inductive biases for spatial-temporal interaction, making it difficult to fully exploit motion cues. These limitations hinder the effective integration of spatial and temporal information, reducing VSOD robustness in dynamic scenes.

- **Limitations of Scribble Annotations in Weakly-Supervised VSOD**. Scribble-based VSOD provides a cost-effective alternative to pixel-wise annotations, yet it lacks complete object structures and boundary details due to the sparsity of scribble labels. Existing methods [35, 36] mitigate this limitation by leveraging adjacent-frame interactions to facilitate short-term feature sharing. However, these approaches struggle to capture long-term contextual

dependencies, making them ineffective in handling complex motion patterns. Furthermore, error accumulation from inaccurate pseudo-labels can progressively distort segmentation quality, leading to cascading segmentation failures over time.

- **Limitations of Multimodal Feature Fusion in VSOD**. Recent VSOD methods have explored optical flow and depth as complementary modalities to enhance feature extraction. However, existing approaches [37, 38] often treat these modalities as mere extensions of the RGB stream, failing to fully exploit their distinct semantic and structural contributions. Suboptimal fusion strategies lead to noisy feature integration, limiting the ability to capture meaningful motion cues and depth-aware spatial relationships. To overcome these limitations, a more robust multimodal feature integration mechanism is needed to effectively harness the complementary nature of these modalities and improve segmentation accuracy.

1.3 Research Objectives

The objectives of this thesis are delineated as follows:

- **Objective 1** is to improve the integration of spatial and temporal dependencies in VSOD by fully exploiting long-term context and frame-aware temporal modeling through unidirectional cross-frame enhancement. The primary goal is to ensure consistent and orderly information propagation across frames while minimizing interference.

- **Objective 2** is to enhance the reliability of scribble-supervised VSOD by addressing the limitations of sparse and incomplete annotations. The primary goal is to mitigate error accumulation and improve segmentation consistency by leveraging long-term contextual dependencies and optimizing memory-based feature propagation.

- **Objective 3** is to enhance multimodal feature integration in VSOD by leveraging optical flow and depth as implicit prompts to fine-tune the pre-trained SAM. The primary goal is to fully exploit their complementary dynamic and structural information while mitigating noise interference through a spectrogram energy-driven prompting strategy.

1.4 Contributions

Throughout this thesis, we focus on three key sub-tasks of Video Salient Object Detection (VSOD): enhancing spatial-temporal integration in fully-supervised VSOD (Chapter 1), improving segmentation consistency in scribble-supervised VSOD (Chapter 2), and leveraging multimodal information to improve VSOD performance (Chapter 3). The contributions in each of these chapters are summarized as follows:

(i) Inheritance Enhancement Network for Video Salient Object Detection

- We propose to exploit frame-aware temporal relations of salient features for V-SOD by making full use of historical multi-frame attention maps. To achieve this, we introduce a novel Heritable Multi-Frame Attention (HMA) module. The HMA module effectively leverages long-term contextual dependencies by employing stacked transformer layers, homologous attention mechanisms, and a unidirectional cross-frame correlation enhancement approach.
- We propose an auxiliary attention loss leveraging inherited attention maps to guide the network towards focusing more on target regions, thereby improving the model's accuracy.
- We propose a Transformer-based Inheritance Enhancement Network (IENet) for V-SOD. IENet takes full advantage of the cumulative advantages of multiple HMAs on different levels of feature maps to significantly enhance video detection performance using the proposed attention loss.

(ii) Knowledge-sharing Hierarchical Memory Fusion Network for Scribble-supervised Video Salient Object Detection

- We propose a Knowledge-sharing Hierarchical Memory Fusion Network (KHMF-Net) for weakly-supervised video salient object detection (V-SOD) by addressing error accumulation and background distractions.
- We design an Hierarchical Memory Bank (HMB) to archive historical segmentation at multiple confidence levels, effectively capturing spatial and temporal contexts for efficient salient object expansion.
- We integrate an Adaptive Memory Fusion (AMF) module that first consolidates multi-level memory features, followed by the Interactive Equalized Matching

(IEM) module, which enhances query performance and distinguishes between background and target objects.

- We devise a dual-attention knowledge-sharing strategy based on a dual-attention mechanism to optimize IEM by effectively utilizing sparse annotations.

(iii) Multimodal Energy Prompting for Video Salient Object Detection

- We propose a Multimodal Energy Prompting Network (MEPNet), which leverages implicit prompts derived from optical flow and depth within a pre-trained SAM framework. By utilizing SAM's segmentation capability, our method effectively integrates dynamic and structural cues, improving the extraction and generalization of task-specific VSOD features.

- We present the Spectrogram Energy Generator (SEG), which draws inspiration from the concept of energy in speech processing. SEG applies the Short-Time Fourier Transform (STFT) to compute spectrogram energy, thereby effectively minimizing noise interference and outperforming the direct use of optical flow and depth as prompts in VSOD tasks.

- We propose a Circular High-frequency Filter (CHF) that refines high-frequency details in RGB images. Unlike previous handcrafted filters, CHF adaptively adjusts its radius using learnable parameters, enabling automatic and precise filtering.

- Our method achieves state-of-the-art performance on five VSOD benchmark datasets, delivering the most precise overall detection results, particularly excelling in high integrity, accurate localization, and precise boundary delineation.

1.5 Organization of this thesis

Literature reviews of the deep learning-based Video Salient Object Detection methods are presented in each chapter corresponding to the proposed three novel deep learning approaches.

The rest of this thesis is organized as follows:

Chapter 2 presents the proposed inheritance enhancement network based on Transformer, which is based on our work published in the *Multimedia Tools and Applications*

2024, titled "IENet: inheritance enhancement network for video salient object detection" [39].

Chapter 3 presents the proposed memory bank-based encoder-decoder hierarchical architecture, to mitigate background distractions and reduce error accumulation, based on our work submitted in the *Pattern Recognition Letters 2025*, titled "Knowledge-sharing Hierarchical Memory Fusion Network for Scribble-supervised Video Salient Object Detection".

Chapter 4 presents the proposed data augmentation method, which is based on our work submitted in the *Multimedia Tools and Applications 2025*, titled "Multimodal Energy Prompting for Video Salient Object Detection".

Chapter 5 concludes this thesis and discusses future work and directions.

Note that references related to each chapter are listed at the end of each chapter, and the Statements of Contributions are inserted at the beginning of each relevant chapter.

References

[1] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, 2009.

[2] Zhigang Tu, Zuwei Guo, Wei Xie, Mengjia Yan, Remco C Veltkamp, Baoxin Li, and Junsong Yuan. Fusing disparate object signatures for salient object detection in video. *Pattern Recognition*, 72:285–299, 2017.

[3] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3239–3259, 2021.

[4] Ashish Kumar Gupta, Ayan Seal, Mukesh Prasad, and Pritee Khanna. Salient object detection techniques in computer vision—a survey. *Entropy*, 22(10):1174, 2020.

[5] Jiande Sun, Xiaocui Liu, Wenbo Wan, Jing Li, Dong Zhao, and Huaxiang Zhang. Video hashing based on appearance and attention features fusion via dbn. *Neurocomputing*, 213:84–94, 2016.

[6] Huazhu Fu, Dong Xu, and Stephen Lin. Object-based multiple foreground

segmentation in rgbd video. *IEEE Transactions on Image Processing*, 26(3):1418–1427, 2017.

[7] Hugo Jacob, Flávio LC Pádua, Anisio Lacerda, and Adriano CM Pereira. A video summarization approach based on the emulation of bottom-up mechanisms of visual attention. *Journal of Intelligent Information Systems*, 49:193–211, 2017.

[8] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.

[9] Chun-Rong Huang, Yun-Jung Chang, Zhi-Xiang Yang, and Yen-Yu Lin. Video saliency map detection by dominant camera motion removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(8):1336–1349, 2014.

[10] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16835, 2021.

[11] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 670–678, 2022.

[12] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE Transactions on Image Processing*, 23(9):3910–3921, 2014.

[13] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[14] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. SCOM: Spatiotemporal constrained optimization for salient object detection. *IEEE Transactions on Image Processing*, 27(7):3345–3357, 2018.

[15] Kao Zhang and Zhenzhong Chen. Video saliency prediction based on spatial-temporal two-stream network. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3544–3557, 2018.

[16] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE Transactions on Image Processing*, 27(10):5002–5015, 2018.

[17] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via

fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

[18] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper ConvLSTM for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–731, 2018.

[19] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in Neural Information Processing Systems*, 30, 2017.

[20] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10869–10876, 2020.

[21] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[22] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3243–3252, 2018.

[23] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7274–7283, 2019.

[24] Nian Liu, Kepan Nan, Wangbo Zhao, Xiwen Yao, and Junwei Han. Learning complementary spatial–temporal transformer for video salient object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[25] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3402, 2015.

[26] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[27] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2017.

[28] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7284–7293, 2019.

[29] Yongri Piao, Chenyang Lu, Miao Zhang, and Huchuan Lu. Semi-supervised video salient object detection based on uncertainty-guided pseudo labels. *Advances in Neural Information Processing Systems*, 35:5614–5627, 2022.

[30] Yi Tang, Wenbin Zou, Zhi Jin, Yuhuan Chen, Yang Hua, and Xia Li. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):1973–1984, 2018.

[31] Shuyong Gao, Haozhe Xing, Wei Zhang, Yan Wang, Qianyu Guo, and Wenqiang Zhang. Weakly supervised video salient object detection via point supervision. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3656–3665, 2022.

[32] Jiahao Hong, Wei Zhang, Zhiwei Feng, and Wenqiang Zhang. Dual Cross-Attention for Video Object Segmentation via Uncertainty Refinement. *IEEE Transactions on Multimedia*, 2022.

[33] Kan Huang, Chunwei Tian, Jingyong Su, and Jerry Chun-Wei Lin. Transformer-based cross reference network for video salient object detection. *Pattern Recognition Letters*, 160:122–127, 2022.

[34] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5912–5921, 2021.

[35] Jian Wang, Siyue Yu, Bingfeng Zhang, Xinqiao Zhao, Ángel F García-Fernández, Eng Gee Lim, and Jimin Xiao. Cross-frame feature-saliency mutual reinforcing for weakly supervised video salient object detection. *Pattern Recognition*, page 110302, 2024.

[36] Binwei Xu, Qiuping Jiang, Xing Zhao, Chenyang Lu, Haoran Liang, and Ronghua Liang. Multidimensional Exploration of Segment Anything Model for Weakly Supervised Video Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[37] Jingjing Li, Wei Ji, Size Wang, Wenbo Li, et al. DVSOD: RGB-D video salient object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Junhao Lin, Lei Zhu, Jiaxing Shen, Huazhu Fu, Qing Zhang, and Liansheng Wang. ViDSOD-100: A New Dataset and a Baseline Model for RGB-D Video Salient Object Detection. *International Journal of Computer Vision*, pages 1–19, 2024.

[39] Tao Jiang, Yi Wang, Feng Hou, and Ruili Wang. IENet: inheritance enhancement network for video salient object detection. *Multimedia Tools and Applications*, pages 1–20, 2024.

Chapter 2

IENet: Inheritance Enhancement Network for Video Salient Object Detection

Effective utilization of spatiotemporal information is essential for improving the accuracy and robustness of Video Salient Object Detection (VSOD). However, current methods have not fully utilized historical frame information, ultimately resulting in insufficient integration of complementary semantic information. To address this issue, we propose a novel Inheritance Enhancement Network (IENet) based on Transformer. The core of IENet is a Heritable Multi-Frame Attention (HMA) module, which fully exploits long-term context and frame-aware temporal modeling in feature extraction through unidirectional cross-frame enhancement. In contrast to existing methods, our heritable strategy is based on the unidirectional inheritance model using attention maps which ensure the information propagation for each frame is consistent and orderly, avoiding additional interference. Furthermore, we propose an auxiliary attention loss by using inherited attention maps to direct the network to focus more on target regions.The experimental results of our IENet reveal its effectiveness in handling challenging scenes on five popular benchmark datasets. For instance, in the cases of VOS and DAVSOD, our method achieves 0.042% and 0.070% for MAE compared to other competitive models. Particularly, IENet excels in inheriting finer details from historical frames even in complex environments. Note that the content pre- sented in this chapter has been published in the Multimedia Tools and Applications.

2.1 Introduction

Video salient object detection (VSOD) is a crucial task in video processing that mimics the human visual attention system by identifying the most captivating objects in dynamic videos. This task has broad applications in a wide range of computer vision tasks, such as object segmentation [1, 2], medical analysis [3], and visual tracking [4]. In contrast to static image SOD, VSOD has more challenges because it has to analyze vast amounts of video data and model the dependency between multiple frames.

During the early stages of VSOD research, most methods focused on extracting spatial and temporal features using Fully Convolutional Networks (FCNs) [5] or LSTM-based models [6, 7]. These approaches, while effective, processed temporal data frame by frame, impeding their ability to adequately grasp extended dependencies spanning spatial positions across multiple frames. In response to this challenge, researchers turned their attention to Space-Time Memory-based techniques [8–11], which utilize all historical frames as the reference, and improved segmentation performance by capturing the long-range dependencies information. Nonetheless, these methods discard spatial and temporal contexts when constructing memory. Instead, they treat it as a collection of features from disordered pixels across all reference frames. Yan et al. [12] on the other hand, utilized frame grouping for identifying space and time contexts through a non-locally enhanced strategy. However, this method overlooked the diverse strengths of temporal relationships among frames. Lu et al. [13] employed a bidirectional dynamic fusion strategy to interact with spatial and temporal information between two consecutive frames. However, this method only concentrates on adjacent frames and omits information from more distant historical frames, leading to insufficient comprehension of long-term memory information.

Recently, VSOD has been promoted by Transformer [16]. Transformer was originally introduced to address the challenge of modeling relationships between word sequences in machine translation, especially long-range dependencies. The computer vision extension model headed by ViT (Vision Transformer) [17] demonstrates the potential of the Transformer in solving vision-related problems. ViT is capable of establishing long-term context dependencies among all input sequence elements and reducing restrictive inductive biases. By employing the multi-head attention mechanism, ViT enables adaptive attention and contextual understanding, facilitating the capture of significant image features and relationships from various positions. While attention-

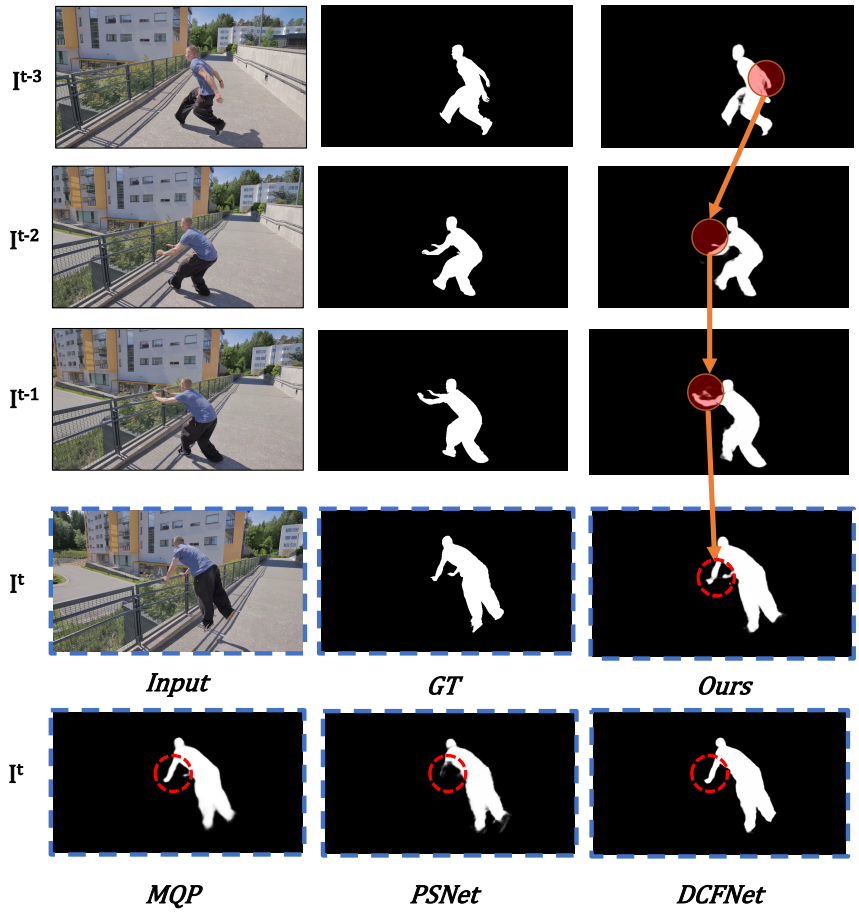

Figure 2.1: Comparison of VSOD models. The man's right arm is clear in the frame (I^{t-3}) but becomes unrecognizable due to torso occlusion in subsequent frames. Consequently, methods such as MQP [14], PSNet [15], and DCFNet [13] fail to accurately detect the right arm frame in I^t. In contrast, our model can learn the right arm features from historical frames. This is helpful for understanding and tracking the arm in subsequent frames.

based aggregation [18, 19] and Transformer-based reference [16, 20] have been employed in several methods, yielding promising improvements. However, these approaches have not adequately accounted for robust temporal dependencies between frames in lengthy videos.

To effectively leverage the intrinsic temporal and spatial relationships among frames, we propose a novel Transformer-based Inheritance Enhancement Network (IENet) for VSOD. Specifically, we propose a Heritable Multi-Frame Attention (HMA) module. Our HMA uses stacked transformer layers with homologous attention [21] and employs a unidirectional cross-frame correlation enhancement approach to effectively exploit

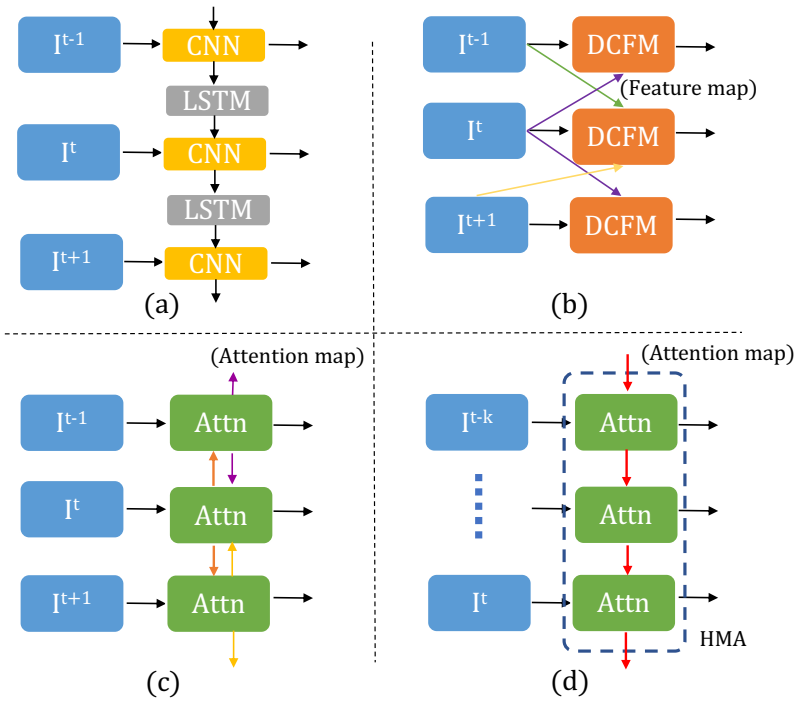

Figure 2.2: Fusion Strategies: (a) Unidirectional fusion using ConvLSTM; (b) Bidirectional fusion based on Dynamic context-sensitive filtering module (DCFM)[13]; (c) Bidirectional fusion based on attention maps; (d) Fusion using the proposed HMA, integrating unidirectional inheritance of attention maps from historical frames.

long-term contextual dependencies during feature representation extraction. IENet holistically explores temporal relationships among frames object inference via integration of HMA modules and achieves salient object inferencing through an end-to-end refinement network with residual connections.

In contrast to the LSTM structure shown in Figure 2.2 (a) and the dynamic context-sensitive filtering module(DCFM) which utilizes bidirectional dynamic fusion strategy depicted in Figure 2.2 (b) used in DCFNet [13], our HMA employs unidirectional propagation based on attention maps, ensuring a continuous and effective impact of historical frames on subsequent frames. Due to the temporal nature of video data, there is a common misconception that bidirectional shared attention maps offer natural advantages by transferring information between adjacent (previous and subsequent) frames. However, this intuition is contradicted by the fact that the future frame may impose negative effects on the current frame. For instance, as shown in Figure **??**, in frame I^t, the movement of the person's body obscures a substantial portion of his right arm. This makes it challenging to detect it with just one frame accurately. The fusion

method uses bidirectional shared attention maps (in Figure 2.2 (c)), propagating the attention map of the future frame, in which there is no information about the missing arm, to the current frame, thereby suppressing the recognition accuracy of the current frame. Therefore, the uncertainty of the future frame may adversely impact on the current frame. On the contrary, our proposed model, IENet, using the HMA fusion strategy, can learn sustainable features from historical frames and continuously provide more attention to the right arm features in subsequent frames through unidirectional inheritance, thereby making corrections timely. This characteristic allows the model to capture long-term contextual dependencies more comprehensively.

We further propose a novel auxiliary attention loss function to expedite model convergence and enhance robustness by using inherited attention maps to direct the network to focus more on target regions. This loss can enhance inherited attention outcomes, ensuring the model concentrates on crucial object regions.

Overall, our Transformer-based IENet exhibits superior performance to existing state-of-the-art solutions, as demonstrated through extensive evaluations of five widely-used benchmarks.

Concretely, the contributions of this work are three-fold:

- We propose to exploit frame-aware temporal relations of salient features for VSOD by making full use of historical multi-frame attention maps. To achieve this, we introduce a novel Heritable Multi-Frame Attention (HMA) module. The HMA module effectively leverages long-term contextual dependencies by employing stacked transformer layers, homologous attention mechanisms, and a unidirectional cross-frame correlation enhancement approach.

- We propose an auxiliary attention loss leveraging inherited attention maps to guide the network towards focusing more on target regions, thereby improving the model's accuracy.

- We propose a Transformer-based Inheritance Enhancement Network (IENet) for VSOD. IENet takes full advantage of the cumulative advantages of multiple HMAs on different levels of feature maps to significantly enhance video detection performance using the proposed attention loss.

- Extensive experiments on five popular benchmark datasets (i.e., DAVIS2016,

YouTube-VOS, ViSal, SegTrack-V2, and DAVSOD) demonstrate the superiority of IENet which outperforms other state-of-the-art methods for the VSOD task. Particularly, IENet excels at capturing fine details from historical frames even in complex environments.

The rest of this chapter is structured as shown below: Section 2.2 briefly reviews related VSOD approaches in this chapter. Section 2.3 explains the details of IENet. Section 2.4 demonstrates and discusses the proposed method through quantitative and qualitative experiments. Section 2.5 outlines the contributions of this chapter.

2.2 Related work

2.2.1 Video Salient Object Detection

Significant progress has been made in salient object detection thanks to recent advances in deep convolutional networks [22]. Due to the complexity of temporal feature extraction, video salient object detection (VSOD) faces more challenges than salient object detection in static images [23–25]. Early studies primarily focus on developing hand-crafted saliency feature extractors and spatiotemporal fusion methods [26–28]. In contrast, Fully Convolutional Networks (FCNs) have been recognized and adopted for implementing an end-to-end trainable architecture [5, 29].

To capture long-range temporal dependencies, Song et al. [7] utilized Long Short-Term Memory (LSTM) in VSOD. Li et al. [6] introduced a flow-guided recurrent encoder that utilizes an optical flow network to work with a feature extractor with ConvLSTM for feature refinement, improving temporal coherence. Fan et al. [30] explored capturing a saliency-shift-aware module and a ConvLSTM-based method that learns the attention shift of humans. To facilitate effective video salient object recognition, Gu et al. [18] utilized a pyramid-constrained self-attention module and non-local technique. Ren et al. [31] constructed a semi-curriculum learning strategy and fused spatial and temporal information to eliminate learning ambiguities. Zhang et al. [13] proposed a dynamic VSOD network that adapts context-aware convolution kernels for a weakly-supervised method based on scribble annotations. However, the convolutional layers in ConvLSTM modules limit their ability to capture and propagate long-term context dependencies due to the inherent locality of the convolutional filters. In addition, the separate memory mechanisms [32] between the layers of the ConvLSTM modules can

result in a lack of effective capture and propagation of intrinsic temporal and spatial information between frames.

To address this, Oh et al. [8] proposed an approach that utilizes space-time memory to build a memory bank that shares feature memory from all historical frames with the current frame. While this method incorporates non-local mechanisms to include memory frames, simply concatenating historical frames may not be sufficient to capture the complex interactions between different time nodes. Hong et al. [9] utilized dual cross-attention to create a shared feature memory of all historical frames that are equally important to the current frame. However, the influence of closer historical frames on the current frame should be more significant than others, which aligns more with the temporal nature of videos. Lin et al. [33] introduced a Query-Memory Aggregation module to refine the initial segmentation result by aggregating information from multiple queries and memories, focusing on different frame histories. Xie et al. [34] developed a hierarchical memory module that stores information at multiple spatial and temporal scales. This enables the model to capture short-term and long-term dependencies between frames. However, space-time memory has limitations in accurately segmenting objects in complex scenes or videos with long-term variations. For example, when an object is occluded or disappears from the scene, it is challenging to segment accurately using only past frames.

2.2.2 Vision Transformer

Transformer [35] has been first proposed as a solution to the sequence-to-sequence problem, e.g., machine translation. More recently, researchers have sought to apply this architecture to vision tasks. The first work in this area is ViT [17], which applies the Transformer to image classification and demonstrates the potential of this approach for vision tasks.

Following this success, several variants of the vision transformer have been proposed, e.g., SwinTransformer [36], PvT [37], T2T-ViT [38]. In addition, the Transformer has been adapted for use in dense pixel prediction tasks, such as video salient object detection. Mei et al. [39] extended the DETR approach from a 2D attention transformer to a 3D attention mechanism that exploits both spatial and temporal relationships, albeit with extensive computations. Su et al. [40] provided a unified architecture for group-based segmentation transformers to recognize video saliency objects. However,

the approach only utilizes an Intra-MLP module with a single frame for inference and lacks inter-frame linking. Huang et al. [20] utilized a pure vision Transformer to extract multi-resolution token feature representations. Duke et al. [16] adopted a multi-frame group history to generate attentional maps but neglected the temporal relationships between history frames. Huang et al. [41] employed Self-attention mechanisms to provide channel-wise spatiotemporal representations, allowing learning of feature relations over global contexts. The above methods incorporate Transformer structures that primarily focus on capturing temporal relationships between frames. However, they do not fully exploit the effect of history multi-frames on the current one in long-term context temporal modeling.

Existing methods such as ConvLSTM-based [30, 32], dynamic filtering-based [13], and attention-based approaches [16, 20, 41], have been employed to handle complex scenarios. However, they struggle to adequately capture intuitive long-term temporal modeling. Specifically, they lack consideration for the expected gradual reduction in the impact of historical frames on the current frame over time. To address this limitation, we propose to consider intrinsic connections between frames to infer and build frame-aware temporal relationships. In particular, a Heritable Multi-Frame Attention (HMA) module is designed to take into account the effect of history multi-frames on the current one to enhance temporal relationships and improve the exploitation of video source characteristics. Notably, unidirectional attention map propagation in HMA ensures that noises from future frames are not introduced to the current frame due to bidirectional interaction. Furthermore, the HMA adopts homologous attention [21], enabling feature representation extraction using long-term contextual temporal modeling enhanced across frames and improving the model's robustness. Then, we construct a network (IENet) utilizing multiple HMAs at different levels of feature maps. IENet effectively leverages HMA cumulative benefits, resulting in a discernible enhancement in video detection performance.

2.3 Method

2.3.1 Overview

The whole framework is presented in Figure 2.3, which employs an encoder-decoder architecture. First, a spatial feature extractor extracts spatial saliency features from

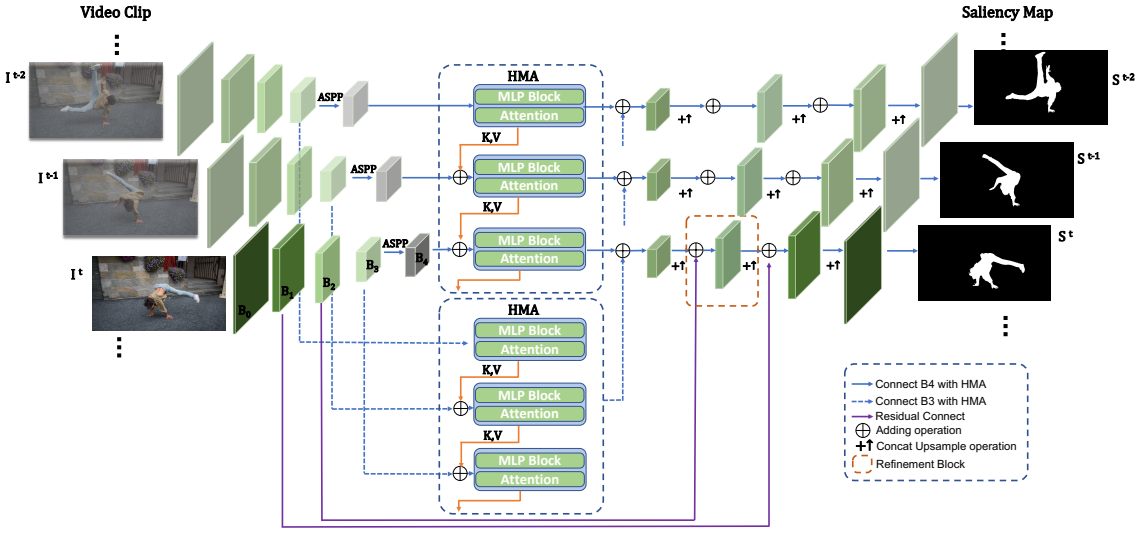

Figure 2.3: Architecture overview of our method. The group-based input frames are initially processed through the Feature Extractor, yielding multiscale feature maps B via the pyramid-dilated convolution. Then the Heritable Multi-Frame Attention (HMA) module (refer to Figure 2.4 for more details) captures long-term context relationships in the model. Finally, the Decoder Block employs the residual connected refinement module to produce the final saliency maps. Symbol '+' denotes the addition operation. Details of HMA are illustrated in Figure 2.4.

the original input images. Then, the features are encoded with low-level characteristics connected to high-level features. After that, the features are flattened and input into the Heritable Multi-Frame Attention (HMA) module, which enhances the long-term temporal relationships from multiple frames in the history to the current frame. Finally, a series of decoder layers are employed in the residual refinement blocks to generate the final saliency map for detecting salient objects.

2.3.2 Spatial Feature Extractor

The spatial feature extractor takes a video clip consisting of $n+1$ consecutive frames $\{I^{t-n}, I^{t-n+1}, ..., I^t\} \in \mathbb{R}^{3\times H_0\times W_0}$ as input. In this study, ResNet-50 [42] is employed as the backbone network for feature extraction. To produce four feature maps with varying spatial resolutions and channel numbers, feature maps B_0, B_1, B_2, and B_3 are generated. Drawing inspiration from [14], the convolutional layers in the last layer are replaced with pyramid-dilated convolutions [43] utilizing a $rate = 2$. So that the receptive field remains the same. Following this, an atrous spatial pyramid pooling (ASPP) [44] layer to generate B_4 is attached to enable extraction of both image-level

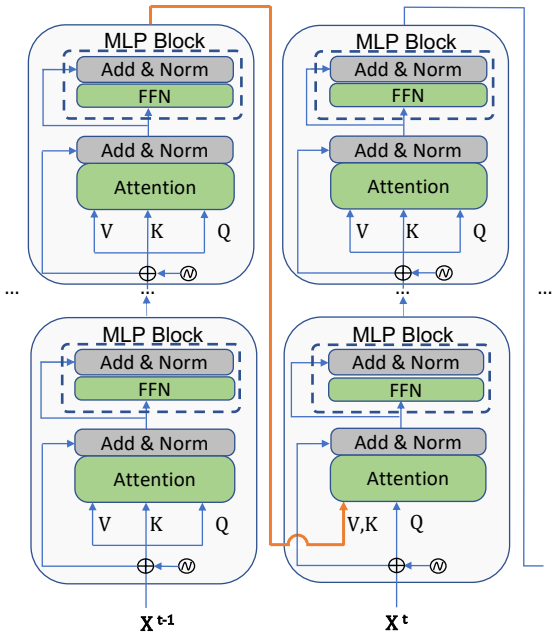

Figure 2.4: Structure of the Heritable Multi-Frame Attention (HMA) module.

global context and multiscale contextual information. Outputs of the four blocks are denoted by B_1^t, B_2^t, B_3^t, and output of ASPP is denoted by B_4^t, respectively:

$$B^{(t-n)\to t} = \Phi\left(I^{(t-n)\to t}\right) \in \mathbb{R}^{D\times H\times W}, \tag{2.1}$$

where $[H_0, W_0]$ and $[H, W]$ refer to the spatial dimensions of the input and output feature maps, respectively, and D is the channel number after the ASPP layer.

2.3.3 Heritable Multi-Frame Attention Module (HMA)

The spatial feature extractor is specifically designed to learn spatial saliency; however, it does not fully utilize the temporal context between historical frames in the video source. In previous works, multiple frames have been simply combined to feed sequential co-evolution [12, 45]. However, the extracted temporal information neglects the historical effect of different frames on the current frame in a video sequence. Considering the motion behavior of inertia and continuity in praxeology, the different historical frames impart varying degrees of space-time information to the current frame. To further enhance the modeling of temporal information from historical to current frames, we propose a Heritable Multi-Frame Attention (HMA) module. The HMA

module aims to incorporate more long-term contextual information and is depicted in Figure 2.4.

The HMA module consists of L-stacked transformer layers with homologous attention [21], which effectively captures multi-frame features. Specifically, given queries q, keys k and values v, the multi-head attention mechanism is formulated as follows:

$$MultiHead(q, k, v) = Concat(h_1, ..., h_n)W^0 \tag{2.2}$$

In contrast to the a Vanilla Transformer [38], the HMA of B_i^t inherits keys k and values v from the output of the previous frame encoder B_i^{t-1}. This normalization approach enables the utilization of previous contextual information to clarify certain parts of the input, thereby guaranteeing the assignment of each pixel to a query vector:

$$\begin{aligned} h_i &= Homo\text{-}attn\left(W_i^Q B^t, W_i^K B^{t-1}, W_i^V B^{t-1}\right) \\ &= Softmax\left(\mathcal{T}_k\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_k}}\right)\right)\tilde{V}, \end{aligned} \tag{2.3}$$

where $W_i^Q \in \mathbb{R}^{D_m \times D_k}$, $W_i^K \in \mathbb{R}^{D_m \times D_k}$, $W_i^V \in \mathbb{R}^{D_m \times D_v}$, $W^0 \in \mathbb{R}^{D_k \times D_m}$ are parameter matrices, and d_k is the dimension of keys, which is identical to standard self-attention settings. Here, we let $h = 8, D_m = 512$ and $D_k = D_v = D_m/h = 64$ be default settings. Homologous attention selects top-k similar tokens from the keys for each query to compute attention maps, which could explore more sensitive temporal information than vanilla attention. $\mathcal{T}_k(.)$ denotes the row-wise top-k selection operator:

$$(\mathcal{T}_k(A))_{ij} = \begin{cases} A_{ij} & A_{ij} \in \textit{top-k (row } j\textit{)} \\ -\infty & \textit{otherwise} \end{cases}. \tag{2.4}$$

Our approach differs from prior work, which utilizes ConvGRU modules [46] to bidirectionally [13, 47] search for spatial and temporal information between frames. Although cross-frame bidirectional designs may have advantages for sequential co-evolution, they neglect the influence of other different historical frames on the current frame in a video sequence. Cross-frame bidirectional designs transfer spatial-temporal information from the current frame to the previous one, which may introduce redundancies and noises. This can harm the network's overall performance. Moreover, using convolutional layers

in ConvGRU may introduce additional inductive bias, thereby limiting the ability of the network to generalize to more diverse data.

2.3.4 Integration

As shown in Figure 2.3, IENet utilizes a residual connection mechanism, enabling the integration of low-level spatial information into pixel-wise saliency inference during the decoding process. We employ a cascade of three refinement blocks [12] that are interconnected with a layer in the spatial feature extractor through a connection layer. Its purpose is to alleviate the negative effects of spatial detail loss during downsampling. It is critical to maintain consistency between the resolutions of the two feature maps so that an upsampling operation is performed when necessary through bilinear interpolation. The output saliency map is denoted as $\{S^{t-n}, S^{t-n+1}, ..., S^{t}\}$.

2.3.5 Loss Function

The loss for IENet consists of a main loss and an auxiliary loss, formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda \mathcal{L}_{attn}, \tag{2.5}$$

where λ represents a non-negative parameter utilized to manage the relative significance of the attention loss.

We calculate the main loss $\mathcal{L}_{\text{main}}$ using a mixed loss function that integrates the individual losses between each input frame. The formula of $\mathcal{L}_{\text{main}}$ can be outlined as follows:

$$\begin{aligned} \mathcal{L}_{main} &= L(SM_{pred}\ ,\ SM_{gt}) \\ &= L(M(I;\theta)\ ,\ SM_{gt}) \end{aligned} \tag{2.6}$$

where SM_{gt} represents ground-truth saliency maps, and $SM_{pred} = M(I;\theta)$ represents the predicted saliency maps. These predicted saliency maps are obtained by inputting the images I into our model M with the parameter configuration θ. The symbol $L(\cdot,\cdot)$ refers to the formulas employed for computing the hybrid loss, which can be defined as follows:

$$L(\cdot,\cdot) \;=\; L_{bce}(\cdot,\cdot) + \; L_{dice}(\cdot,\cdot)\;, \tag{2.7}$$

where $L_{bce}(\cdot,\cdot)$ and $L_{dice}(\cdot,\cdot)$ denote the BCE [48] and Dice loss [49], respectively. The detailed computation formula for these two losses are as follows:

$$L_{bce}(X,Y) \;=\; -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(x_i) \;+\; (1-y_i)log(1-x_i)]\;, \tag{2.8}$$

$$L_{dice}(X,Y) \;=\; 1 - \frac{2\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i}\;, \tag{2.9}$$

where X denotes one of the predicted outcomes SM_{pred}, while Y corresponds to SM_{gt}, and N signifies the total pixel count within X or Y. The utilization of BCE loss contributes to the convergence of all pixels, independent of their labels. Incorporating the Dice loss quantifies the level of overlap between X and Y. By incorporating this loss function, our model can yield enhanced detection results.

To enhance the model's resilience in complex environments, we propose an attention loss, denoted as $\mathcal{L}_{attn}$, which works in conjunction with the primary loss $\mathcal{L}_{main}$. This loss aims to effectively filter noise from the attention maps inherited by the HMA modules, ensuring the model focus on salient object regions. It achieves this by inheriting the attention maps through bilinear interpolation and computing a fused loss in combination with the ground truth (GT). The attention loss is formulated as follows:

$$\mathcal{L}_{attn} \;=\; L(SM_{attn}\;,\; SM_{gt}), \tag{2.10}$$

where SM_{attn} represents the attention maps after bilinear interpolation, resulting in a binary image. We calculate BCE and Dice loss for both SM_{attn} and SM_{gt} using Formula 2.7. Our attention loss as an auxiliary loss is coupled with the main loss $\mathcal{L}_{main}$ for training our model.

2.4 Experiments

In this section, we evaluate the performance of our model on five widely used public V-OSD datasets. We also conduct extensive ablation studies to highlight the significance

Table 2.1: Quantitative comparisons with state-of-the-art models on five widely used VSOD datasets. Symbols ↑ and ↓ donate larger and smaller is better, respectively. Symbol '-' means that results are not provided. The best results for each dataset are shown in bold.

Method	DAVIS [50]			VOS [51]			ViSal [27]			SegV2 [52]			DAVSOD [30]		
	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$
SCOM [28]$_{TIP'2018}$	0.048	0.832	0.783	0.162	0.712	0.690	0.122	0.762	0.831	0.030	0.815	0.764	0.220	0.599	0.464
MBNM [53]$_{ECCV'2018}$	0.031	0.887	0.861	0.099	0.742	0.670	0.020	0.898	0.883	0.026	0.809	0.716	0.159	0.637	0.520
MGA [54]$_{ICCV'2019}$	0.023	0.910	0.892	-	-	-	0.017	0.936	0.933	0.030	0.865	0.821	0.084	0.738	0.640
SSAV [30]$_{CVPR'2019}$	0.028	0.893	0.861	0.073	0.819	0.742	0.020	0.943	0.939	0.023	0.851	0.801	0.092	0.724	0.603
RCRNet [12]$_{ICCV'2019}$	0.027	0.886	0.848	0.051	0.873	0.833	0.026	0.922	0.906	0.035	0.842	0.781	0.087	0.741	0.653
DSNet [55]$_{ECCV'2020}$	0.018	**0.914**	**0.981**	-	-	-	0.013	0.949	0.950	0.028	0.875	0.832	-	-	-
PCSA [18]$_{AAAI'2020}$	0.022	0.902	0.880	0.065	0.827	0.747	0.017	0.946	0.940	0.025	0.865	0.810	0.086	0.741	0.655
FSNet [56]$_{ICCV'2021}$	0.020	0.920	0.907	-	-	-	-	-	-	0.023	0.870	0.772	0.072	0.773	0.825
ResueVOS [57]$_{CVPR'2021}$	0.019	0.883	0.865	-	-	-	0.020	0.928	0.933	0.025	0.844	0.832	-	-	-
TransVOS [39]$_{preprint'2021}$	0.018	0.885	0.869	-	-	-	0.021	0.917	0.928	0.024	0.816	0.800	-	-	-
DCFNet [13]$_{ICCV'2021}$	0.018	0.914	0.900	0.060	0.846	0.791	0.012	0.949	0.953	0.017	0.852	0.853	0.074	0.741	0.660
MQP [58]$_{IEEE'2021}$	0.018	0.916	0.904	0.069	0.828	0.768	0.016	0.942	0.939	0.018	0.882	0.841	0.075	0.770	0.703
PSNet [15]$_{ICCV'2022}$	**0.016**	0.919	0.907	-	-	-	0.012	**0.954**	**0.955**	**0.016**	0.889	0.852	0.074	0.765	0.678
UFO [40]$_{ICCV'2023}$	0.036	0.864	0.828	-	-	-	**0.011**	0.953	0.940	0.022	**0.892**	0.863	-	-	-
IENet(Ours)	0.017	0.912	0.896	**0.042**	**0.882**	**0.845**	0.012	0.947	0.944	**0.016**	0.882	**0.865**	**0.070**	**0.775**	**0.862**

of various methods incorporated into our model. The performance of our best model, utilizing optimal hyperparameters, is then compared to that of current state-of-the-art models and evaluated using the same assessment methodologies.

2.4.1 Implementation Details

2.4.1.1 Model Training

Firstly, the spatial feature extractor weights in our model are initialized using ResNet-50 [42] pre-trained on ImageNet [59] as the backbone. To train the model, two video-specific datasets, namely YouTube-VOS [51] and DAVIS2016 [50], are merged and used as training sets. We employ HMA with B_3 and B_4 to inherit more long-term context temporal information in IENet. As for the model hyperparameters, Adam [60] is used as the optimizer with β_1=0.9 and β_2=0.999, and the initial learning rate is set to 1e-5. The momentum is 0.925, and the weight decay is 0.0005. Training was accelerated by dual NVIDIA A100 GPUs, ensuring efficient convergence for all experiments.

2.4.1.2 Model Testing

We follow the standard benchmarks [30, 50] to test our model on the test set of DAVIS2016 [50], DAVSOD [30], the test set of YouTube-VOS [51], the Whole of ViSal [27], and the whole of SegTrack-V2 [52].

DAVIS2016 [50] is the most popular VSOD dataset and contains 50 densely annotated sequences (30 for training and 20 for validation), exhibiting varied moving objects.

YouTube-VOS [51] is a large-scale dataset consisting of 200 indoor and outdoor videos for VSOD. It includes 7,650 pixel-wise labeled key frames and 116,103 frames in total.

ViSal [27] is the first dataset earmarked for VSOD. It includes 17 video clips ranging from 30 to 100 frames in length and 193 carefully annotated frames.

SegTrackv2 [52] is among the earliest VSOD datasets and contains 14 sequences and 976 annotated frames. Each sequence involves 1-6 moving objects.

DAVSOD [30] is purpose-built for the VSOD task. It is the most challenging VSOD dataset with consistent visual attention, high-quality annotations, and various features.

2.4.2 Evaluation

We adopt three widely-used criteria to evaluate the performances of our model and competing methods, *i.e.*, Mean Absolute Error (MAE [61]), S-measure (S_m) [62], and F-measure (F_β) [63].

Mean Absolute Error(MAE) [61] represents the absolute difference between ground truth G and prediction map S in all pixel locations, and is defined as:

$$MAE \;=\; \frac{1}{N}\sum_{p=1}^{N} \mid S_p \;-\; G_p \mid , \tag{2.11}$$

where p represents a pixel, and N is the number of pixels on the map.

S-measure (S_m) [62] reflects the structure similarity between predicted salient objects and the ground truth, is defined as:

$$S = \gamma S_0 + (1-\gamma) S_r \ , \tag{2.12}$$

where γ provides a balance between the object-aware similarity S_o and region-aware similarity S_r. We utilized the default value (γ=0.5) proposed in [62].

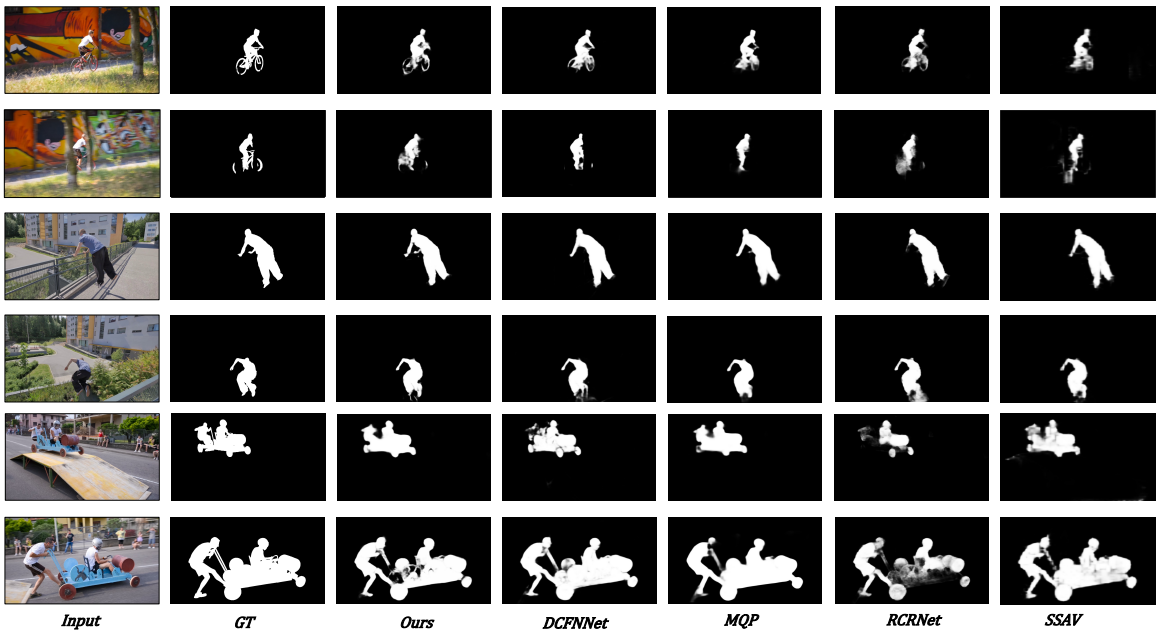

Figure 2.5: Qualitative comparison of our method with state-of-the-art video salient object detection methods on DAVIS. It can be observed that our approach produces saliency maps with superior accuracy across a variety of video scenes.

F-measure(F_β) [63] takes into account both *Precision* and *Recall* values to evaluate the performance of saliency detection models. It is defined as:

$$F_\beta = \frac{(1+\beta^2)\ \ Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{2.13}$$

where β^2 as is the case with the majority of previous image-based models, as recommended in [64].

2.4.3 Comparisons with State-of-the-art Models

In this section, we compare our proposed approach with state-of-the-art models, examining both quantitative and qualitative aspects.

Quantitative Evaluation. To demonstrate the effectiveness of our method, we compare it with 14 video salient object detection methods including: SCOM [28], MBNM [53], MGA [54], SSAV [30], RCRNet [12], DSNet [55], PCSA [18], FSNet [56], ResueVOS [57], TransVOS [39], DCFNet [13], MQP [58], UFO [40], and PSNet [15]. The quantitative evaluation of *MAE*, S_m, and F_β is shown in Table 2.1. For a fair

Table 2.2: Ablation study on the HMA with various configurations.

Configuration	DAVSOD			VOS		
	$MAE \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$
BL	0.0830	0.7428	0.7232	0.0560	0.7801	0.8405
BL-Mul	0.0729	0.7502	0.7167	0.0580	0.8672	0.8433
BL-Add	0.0726	0.7648	0.7489	0.0600	0.8736	**0.8527**
BL-Mul-HomoAttn	0.0711	0.7726	0.7591	0.0572	0.8701	0.8389
BL-Add-HomoAttn	**0.0702**	**0.7755**	**0.8616**	**0.0418**	**0.8819**	0.8449

comparison, we employ a widely-used evaluation toolbox [65]. It can be observed that our method achieves superior performance against almost all models.

Qualitative Evaluation. For qualitative evaluation, a comparison between our method and state-of-the-art (SOTA) techniques on the DAVIS2016 dataset has been presented in Figure 2.5. Two frames of three complex scenes have been selected for each test set. Apparently, our saliency maps have clearer boundaries and effectively remove a significant amount of background noise. This is particularly evident in situations where objects exhibit dynamic changes and subtle differences, as demonstrated by the accurate saliency predictions and preservation of fine details in the *soapbox* sequence (Rows 1,2). Additionally, our approach demonstrates robustness in handling complex scenarios, such as occluded dynamic objects in *bmx-tree* (Rows 3,4), by effectively utilizing spatiotemporal cues for accurate detection of salient objects. When presented with objects that exhibit a larger range of motion, our method is able to capture more dynamic details. A prominent example of this is the human arm in the *parkour* sequence (Rows 5,6).

Our IENet also shows strong robustness. As shown in Table 2.1, our model outperforms most existing modes consistently across five standard datasets. This demonstrates that our model is robust to the variations of datasets. Additionally, in Figure 2.5, we present a comparison between our model and four state-of-the-art models using three complex scenes from the DAVIS dataset. This comparison leads to the conclusion that our model exhibits remarkable robustness in scenarios involving high-speed object movement, significant motion, and approaching trajectories.

2.4.4 Ablation Analyses

To demonstrate the superiority of our IENet, a series of ablation experiments are conducted on both DAVSOD and VOS datasets. **Configuration of HMA** Table 2.2

Table 2.3: Ablation study on various IENet structures.

Method	DAVSOD			VOS		
	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$
IENet-B3	0.0782	0.7613	0.7174	0.0590	0.8728	0.8308
IENet-B4	0.0801	0.7595	0.7391	0.0541	0.8755	0.8215
IENet-B3-B4	**0.0702**	**0.7755**	**0.8616**	**0.0418**	**0.8819**	**0.8449**

Table 2.4: Comparison for using unidirectional and bidirectional strategies for IENet.

Method	DAVSOD			VOS		
	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$
IENet-bidirectional	0.0721	0.7194	0.6825	0.0611	0.7943	**0.8484**
IENet-unidirectional	**0.0702**	**0.7755**	**0.8616**	**0.0418**	**0.8819**	0.8449

illustrates the results of vairous configurations of HMA. Firstly, a vanilla transformer architecture is constructed as the baseline model (BL). We then set different HMA architectures which are fused by element-wise multiplication (BL-Mul), and addition (BL-Add) in HMA, respectively. Among these, BL-Mul and BL-Add employ element-wise multiplication and addition, respectively, to manipulate the attention maps of the two frames. The table also includes variations that employ the homologous attention mechanism in combination with BL-Mul and BL-Add. It is evident from the results that the architecture equipped with BL-Add-HomoAttn HMA achieved the best performance among all the alternatives.

2.4.4.1 Structures of IENet

Furthermore, multiple structures of our IENet are designed with different blocks in HMA in Table 2.3. The IENet-B3 and IENet-B4 models are constructed by combining HMA after B_3 and B_4, respectively. The IENet-B3-B4 denotes the model implementing HMA after both B_3 and B_4. It can be observed that the HMA leverages multiscale contextual in-context temporal information and construct salient maps for the entire video, ultimately leading to enhanced performance.

In Table 2.4, we present the results of a bidirectional model to validate whether unidirectional inheritance in long videos is more intuitive from a cognitive standpoint. We set the bidirectional model by swapping the attention maps between the two frames. By comparing the performance of the bidirectional model with the unidirectional inheritance approach, we verify the effectiveness of capturing long-term context and

Table 2.5: Ablation study on the attention loss.

Method	DAVSOD			VOS		
	$MAE \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$
$\lambda = 0$	0.0721	0.7194	0.6825	0.0539	0.8446	0.7939
$\lambda = 0.3$	**0.0702**	**0.7755**	**0.8616**	**0.0418**	**0.8819**	**0.8449**
$\lambda = 0.5$	0.0748	0.7137	0.5858	0.0568	0.8442	0.7898

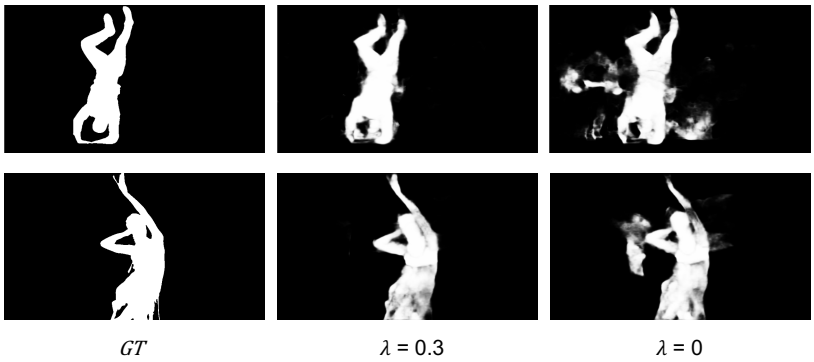

Figure 2.6: We manage the relative significance of the attention loss by a non-negative parameter λ. The auxiliary attention loss not only aids in monitoring the model's convergence but also contributes to constraining attention regions, thereby reducing the influence of noise.

frame-aware temporal modeling in feature extraction through unidirectional cross-frame enhancement.

2.4.4.2 Loss Setting

To assess the significance of the auxiliary attention loss, we create three weight values of λ for the auxiliary attention loss. As shown in Table 2.5, the effectiveness of the auxiliary attention loss is evident in our IENet. By observing Figure 2.6, we notice that the auxiliary attention loss not only aids in monitoring the model's convergence but also contributes to constraining attention regions, thereby reducing the influence of noise.

2.4.5 Limitations and Discussion

Our IENet demonstrates significant effectiveness in scenarios where the object exhibits significant motion, i.e., the object moves quickly in the video. However, the advantages of IENet are less apparent when the target moves slowly or exhibits minimal variation.

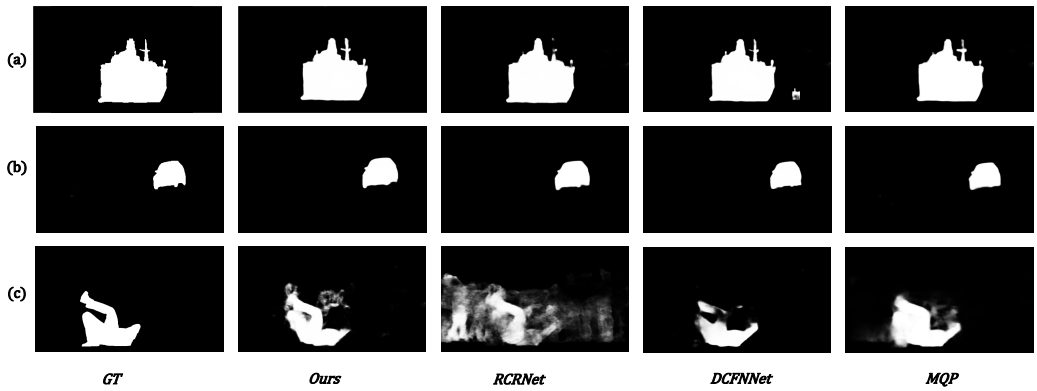

Figure 2.7: (a) and (b) are from **ViSal**, representing *boats* and *cars* which move slowly or exhibits minimal variation, respectively. (c) and (d) are inference results in the *breakdance* sequence of **DAVIS**. Our model's results demonstrate clearer contours and more detailed features compared to other methods.

As illustrated in Figure 2.7 (a) and (b), depicting two scenes from the ViSal dataset, when the *boat* (a) and *car* (b) exhibit slow movement, our results are not significantly better than other models.

Our model is prone to be affected by noise interference when the background features closely resemble the target. For instance, as shown in Figure 2.7 (c), in the *breakdance* sequence of DAVIS, the audience in the background has similar static appearance features to the target, and most models experience confusion in handling these resembling features. Although our attention loss is effectively utilized as an auxiliary loss to mitigate noise generation, this phenomenon still exists in our IENet. Currently, this represents a common challenge in VSOD models. To alleviate the aforementioned issue, we recommend expanding and refining existing datasets, and enhancing the model's generalization capabilities in subsequent steps.

Moreover, inspired by recent studies in the field of image-based SOD [66–68] that address the issue of incomplete object representation within a single image through the utilization of part-object relational visual saliency, we plan to extend part-object awareness from individual frames to long sequences by exploring collaborative 3D routing across multiple frames. This will be achieved by implementing collaborative 3D routing, enabling a more accurate segmentation of the target object from background elements that exhibit similarities with the target. We anticipate that this approach will yield complete salient objects with fine details and reduced noise.

2.5 Conclusion

In this paper, we propose an inheritance enhancement network (IENet) for VSOD. The core of IENet is the Heritable Multi-Frame Attention (HMA) module designed to utilize long-term context and temporal relationships in feature extraction through unidirectional cross-frame enhancement. HMA ensures consistent information propagation without interference and enables collaborative feature learning with frame-aware temporal relationships. Moreover, we propose an attention loss function that refines inherited attention, boosting the model's recognition of the target regions and reducing unnecessary attention. Experimental results on five widely-used benchmarks demonstrate that IENet can achieve superior performance than state-of-the-art in video salient object detection. While HMA has revealed that unidirectional inheritance has an impact on the context of long-term dependency, there still exists a phenomenon of noise generation when background features closely resemble target features. Future research aims to address this challenge by enhancing the model's generalization capabilities through the expansion and refinement of datasets. Furthermore, we will delve deeper into the exploration of HMA module and the auxiliary attention loss mechanism, with the objective of applying them to a broader spectrum of visual tasks.

References

[1] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2016.

[2] Zhe Chen, Ruili Wang, Zhen Zhang, Huibin Wang, and Lizhong Xu. Background–foreground interaction for moving object detection in dynamic scenes. *Information Sciences*, 483:65–81, 2019.

[3] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

[4] Hefeng Wu, Guanbin Li, and Xiaonan Luo. Weighted attentional blocks for probabilistic object tracking. *The Visual Computer*, 30(2):229–243, 2014.

[5] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via

fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

[6] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3243–3252, 2018.

[7] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper ConvLSTM for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–731, 2018.

[8] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[9] Jiahao Hong, Wei Zhang, Zhiwei Feng, and Wenqiang Zhang. Dual Cross-Attention for Video Object Segmentation via Uncertainty Refinement. *IEEE Transactions on Multimedia*, 2022.

[10] Kan Huang and Zhijing Xu. Lightweight video salient object detection via channel-shuffle enhanced multi-modal fusion network. *Multimedia Tools and Applications*, pages 1–15, 2023.

[11] Hong-Bo Bi, Di Lu, Hui-Hui Zhu, Li-Na Yang, and Hua-Ping Guan. STA-Net: spatial-temporal attention network for video salient object detection. *Applied Intelligence*, 51:3450–3459, 2021.

[12] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7284–7293, 2019.

[13] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1553–1563, 2021.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[15] Runmin Cong, Weiyu Song, Jianjun Lei, Guanghui Yue, Yao Zhao, and Sam

Kwong. Parallel symmetric network for video salient object detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.

[16] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5912–5921, 2021.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[18] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10869–10876, 2020.

[19] Zhiheng Zhou, Yongfan Guo, Junchu Huang, Ming Dai, Ming Deng, and Qingjun Yu. Superpixel attention guided network for accurate and real-time salient object detection. *Multimedia Tools and Applications*, 81(27):38921–38944, 2022.

[20] Kan Huang, Chunwei Tian, Jingyong Su, and Jerry Chun-Wei Lin. Transformer-based cross reference network for video salient object detection. *Pattern Recognition Letters*, 160:122–127, 2022.

[21] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22742–22751, 2023.

[22] Pourya Shamsolmoali, Xiaofang Li, and Ruili Wang. Single image resolution enhancement by efficient dilated densely connected residual network. *Signal Processing: Image Communication*, 79:13–23, 2019.

[23] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10031–10040, 2023.

[24] Navjot Singh, Rinki Arya, and RK Agrawal. A novel position prior using fusion of rule of thirds and image center for salient object detection. *Multimedia Tools and Applications*, 76:10521–10538, 2017.

[25] Runzhao Yao, Shaoyi Du, Teng Wan, and Wenting Cui. A robust registration

algorithm based on salient object detection. *Multimedia Tools and Applications*, 81(24):34387–34400, 2022.

[26] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE Transactions on Image Processing*, 23(9):3910–3921, 2014.

[27] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[28] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. SCOM: Spatiotemporal constrained optimization for salient object detection. *IEEE Transactions on Image Processing*, 27(7):3345–3357, 2018.

[29] Qingping Zheng, Ying Li, Ling Zheng, and Qiang Shen. Progressively real-time video salient object detection via cascaded fully convolutional networks with motion attention. *Neurocomputing*, 467:465–475, 2022.

[30] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting More Attention to Video Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8546–8556, 2019.

[31] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. In *European Conference on Computer Vision*, pages 212–228. Springer, 2020.

[32] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in Neural Information Processing Systems*, 30, 2017.

[33] Fanchao Lin, Hongtao Xie, Yan Li, and Yongdong Zhang. Query-memory re-aggregation for weakly-supervised video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2038–2046, 2021.

[34] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen

Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[38] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token ViT: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[39] Jianbiao Mei, Mengmeng Wang, Yeneng Lin, Yi Yuan, and Yong Liu. Video object segmentation with transformers. *arXiv:2106.00588*, 2021.

[40] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. A Unified Transformer Framework for Group-based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection. *arXiv:2203.04708*, 2022.

[41] Kan Huang, Ge Li, and Shan Liu. Learning channel-wise spatio-temporal representations for video salient object detection. *Neurocomputing*, 403:325–336, 2020.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[43] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3402, 2015.

[44] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017.

[45] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188, 2021.

[46] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv:1511.06432*,

2015.

[47] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16835, 2021.

[48] Shruti Jadon. A survey of loss functions for semantic segmentation. In *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2020.

[49] Lucas Fidon, Wenqi Li, Luis C. Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sébastien Ourselin, and Tom Vercauteren. Generalised Wasserstein Dice Score for Imbalanced Multi-class Segmentation Using Holistic Convolutional Networks. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 64–76, Cham, 2018. Springer International Publishing.

[50] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[51] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2017.

[52] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video Segmentation by Tracking Many Figure-Ground Segments. In *IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[53] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 207–223, 2018.

[54] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7274–7283, 2019.

[55] Jing Liu, Jiaxiang Wang, Weikang Wang, and Yuting Su. Dynamic spatiotemporal network for video salient object detection. *Digital Signal Processing*, 130:103700, 2022.

[56] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4922–4933, 2021.

[57] Hyojin Park, Jayeon Yoo, Seohyeong Jeong, Ganesh Venkatesh, and Nojun Kwak. Learning dynamic network using a reuse gate function in semi-supervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8405–8414, 2021.

[58] Chenglizhao Chen, Jia Song, Chong Peng, Guodong Wang, and Yuming Fang. A novel video salient object detection method via semisupervised motion quality perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2732–2745, 2021.

[59] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[60] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[61] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740. IEEE, 2012.

[62] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017.

[63] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE, 2009.

[64] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[65] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-Induced Co-Saliency Detection. In *European Conference on Computer Vision (ECCV)*, 2020.

[66] Yi Liu, Dingwen Zhang, Qiang Zhang, and Jungong Han. Part-Object relational visual saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3688–3704, 2021.

[67] Qiang Zhang, Mingxing Duanmu, Yongjiang Luo, Yi Liu, and Jungong Han. Engaging part-whole hierarchies and contrast cues for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3644–3658, 2021.

[68] Yi Liu, Dingwen Zhang, Nian Liu, Shoukun Xu, and Jungong Han. Disentangled capsule routing for fast part-object relational saliency. *IEEE Transactions on Image Processing*, 31:6719–6732, 2022.

Chapter 3

Knowledge-sharing Hierarchical Memory Fusion Network for Scribble-supervised Video Salient Object Detection

Scribble annotations, being more accessible compared to pixel-wise labels, are increasingly used for tasks such as video salient object detection (VSOD). However, due to their limited foreground coverage and imprecise boundary information, using scribble annotations suffers from background distractions and error accumulation, which can cascade and negatively affect the segmentation in subsequent frames during weakly-supervised video salient object detection (VSOD). In response to the issues outlined above, we propose a novel Knowledge-sharing Hierarchical Memory Fusion Network (KHMF-Net) for scribble-supervised VSOD. To reduce error accumulation, we design a Hierarchical Memory Bank (HMB) for KHMF-Net to archive historical segmentation at multiple confidence levels and capture both spatial and temporal contexts, facilitating an efficient salient object expansion process. In addition, KHMF-Net incorporates an Adaptive Memory Fusion (AMF) module, which first integrates multi-level memory features, followed by the Interactive Equalized Matching (IEM) module to ensure precise query performance and enhance the ability to distinguish between background and target objects, thereby mitigating background distractions. Furthermore, we devise a dual-attention knowledge-sharing strategy to optimize IEM, enhancing the accuracy of similarity computation during the matching process. Experimental results demonstrate that KHMF-Net benefits from a robust hierarchical memory architecture and its ability to clearly distinguish between background and target, achieving state-of-the-art performance on three public scribble-labeled datasets and even outperforming some fully supervised methods. Note that the content presented in this chapter has been published in the Pattern Recognition Letters.

3.1 Introduction

Video Salient Object Detection (VSOD) identifies visually distinct regions in video frames, simulating human attention to highlight important information in complex visuals. This human-like focus is crucial for applications such as video object segmentation [1], video summarization [2], object tracking [3], and autonomous driving [4].

In recent years, fully supervised deep learning-based VSOD models have achieved remarkable performance [5–7]. However, these models require pixel-level annotations, which are laborious and time-consuming to create in video datasets. To balance labeling efficiency and performance, research focuses on VSOD methods with sparse annotations, including scribbles [8, 9] or points [10, 11]. Point supervision annotates specific points on salient objects and backgrounds but may inherently miss important details about location and shape [12]. Scribble supervision uses two strokes to annotate areas of the salient object and background. Although less detailed than pixel-wise annotations, scribbles cover larger areas and provide sketchier postures than points [13, 14]. This balances labeling efficiency and annotation density, making scribble supervision favorable for weakly supervised VSOD. We focus on *scribble-based VSOD* in this work.

Compared to pixel-wise annotations, the primary challenge in scribble-based VSOD approaches is the lack of complete object structures and boundary details in scribble annotations. This limitation arises from the fact that a single stroke cannot capture the full complexity of an object's shape, thus covering only a small portion of the target and making it susceptible to background distractions. Existing models [12, 15] for scribble-based VSOD primarily leverage adjacent-frame interactions that focus on short-term contextual dependencies. These interactions help mitigate the limitations of scribble labels by facilitating feature sharing across nearby frames. However, because these models rely on a limited number of frames, they struggle to capture long-term contextual dependencies, making it challenging to handle scenarios that require more extensive temporal information. In response to these limitations, the Segment Anything Model (SAM) [16] has been integrated into propagation-based models [17, 18] to produce pseudo-labels for detecting salient objects and effectively encoding temporal information from historical frames. However, the sparsity of the scribble labels often results in incomplete target information. For instance, a single scribble might only cover the main part of an object, making it difficult to depict complex limb movements.

When an incorrect mask is stored in memory, it becomes uncorrectable and can progressively distort segmentation in subsequent frames, with small frame-by-frame inaccuracies accumulating into significant errors over time [19, 20].

We propose a novel Knowledge-sharing Hierarchical Network (KHMF-Net), a memory bank-based encoder-decoder hierarchical architecture, to mitigate background distractions and reduce error accumulation. To reduce error accumulation, our KHMF-Net integrates a Hierarchical Memory Bank (HMB) that archives historical masks at different confidence levels, enabling the capture of long-term contextual dependencies. Specifically, our HMB manages the initial scribble in the first frame, a high-confidence region, and entire predicted salient maps. HMB effectively reduces error accumulation during expansion by using historical data as a reference and propagating the predicted masks. Furthermore, we propose an Adaptive Memory Fusion (AMF) module that adaptively blends varying confidence levels to provide a reliable reference throughout the expansion process, alleviating parameter pressure during matching. Concurrently, we introduce an Interactive Equalized Matching (IEM) module to mitigate background distractions. The IEM utilizes Reference-Wise (R-W) Softmax to ensure equal contribution from all pixels in the reference frame to the prediction of the query frame, preventing excessive reliance on background information from the reference frame. Additionally, we utilize a knowledge sharing mechanism based on a dual-attention mechanism for more efficient knowledge transfer in the matching process for IEM. This mechanism extracts high-performance attention features within the Teacher Attention module and efficiently transfers them to the Student Attention module, facilitating the extraction of intricate information from sparse annotations and enhancing overall segmentation accuracy. To evaluate the performance of our method, we conduct experiments using the DAVIS [21], DAVSOD [22] and Youtube-VOS [23] datasets.

The results showcase the promising performance of our approach. The contributions of this work can be summarized as follows:

(i) We propose a Knowledge-sharing Hierarchical Memory Fusion Network (KHMF-Net) for weakly-supervised video salient object detection (VSOD) by addressing error accumulation and background distractions.

(ii) We design an Hierarchical Memory Bank (HMB) to archive historical segmentation at multiple confidence levels, effectively capturing spatial and temporal contexts for efficient salient object expansion.

(iii) We integrate an Adaptive Memory Fusion (AMF) module that first consolidates multi-level memory features, followed by the Interactive Equalized Matching (IEM) module, which enhances query performance and distinguishes between background and target objects.

(iv) We devise a dual-attention knowledge-sharing strategy based on a dual-attention mechanism to optimize IEM by effectively utilizing sparse annotations.

The rest of this chapter is structured as shown below: Section 3.2 briefly reviews related VSOD approaches in this chapter. Section 3.3 explains the details of KHMF-Net. Section 3.4 demonstrates and discusses the proposed method through quantitative and qualitative experiments. Section 3.5 outlines the contributions of this chapter.

3.2 Related work

3.2.1 Fully supervised video salient object detection

The predominant Video Salient Object Detection (VSOD) approaches focus on fully supervised models that leverage spatial and temporal information. Wang et al. [24] utilized a Fully Convolutional Network (FCN) to model short-term spatiotemporal information, using pairs of adjacent frames as input. To effectively capture long-term spatiotemporal characteristics, previous studies [25, 26] employed recurrent networks to establish more comprehensive temporal connections. Furthermore, various methods have been devised to explore dynamic information, enhancing spatiotemporal features. By incorporating multiple motion-guided attention modules and optical flow techniques, prior research [27, 28] demonstrated improved performance detecting moving objects within optical flow imagery. Liu et al. [29] proposed a dynamic contextual correlation filtering module that effectively addresses challenges faced in scenes with minimal dynamic changes. Additionally, several methods have delved into attention mechanisms to identify salient regions more effectively. For instance, Fan et al. [22] proposed a saliency-shift-aware module to learn attention by utilizing human eye fixation data. Similarly, Gu et al. [30] introduced a Pyramid Constrained Self-Attention module to enhance temporal information capture by selectively amplifying specific pixels within a defined range. Zhao et al. [31] developed a memory-based

network to reduce computational demands and enhance the quality of saliency maps by extracting pertinent temporal information from historical frames. While considerable advancements have been made in fully supervised VSOD, the major challenge hindering further progress in VSOD lies in the significant time and cost involved in annotations.

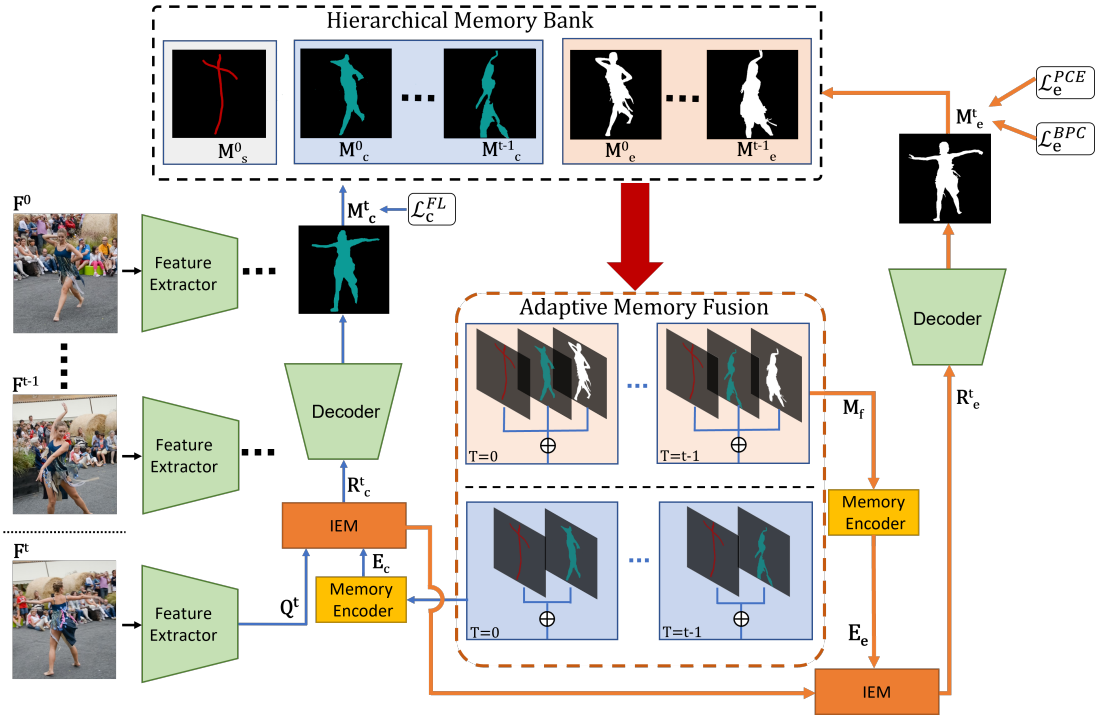

Figure 3.1: Illustration of our Knowledge-sharing Hierarchical Memory Fusion Network (KHMF-Net), featuring a memory bank-based encoder-decoder architecture. The Hierarchical Memory Block (HMB) maintains three confidence levels: the initial scribble map (M_s), the High-confidence region (M_c, blue line), and the full predicted target (M_c, orange line). The Adaptive Memory Fusion (AMF) module and Memory Encoder integrate these confidence levels (E_c and E_e) and match them with the query frame through Interactive Equalized Matching (IEM).

Scribble supervised image salient object detection

To reduce annotation costs without compromising model performance, several methods have explored the use of weak labels for salient object detection. Among various weak annotation methods, scribble supervision [13] provides a good balance between annotation efficiency and covering larger areas with rough outlines. Training models with scribble labels has shown promise in balancing annotation cost and model performance. Lin et al.[32] used scribble annotations to train a graphical model for propagating information from labeled to unlabeled areas. Zhang et al.[13] proposed a

weakly supervised VSOD network incorporating an auxiliary edge detection task and a gated structure-aware loss to enhance object structure recovery. Xu et al. [33] further improved scribble-based approaches by employing dual-modal edge guidance and an active scribble-boosting strategy. However, these methods are limited by their focus on image-based approaches, which struggle to effectively utilize contextual information in video sequences.

Scribble supervised video salient object detection Using scribble labels has recently emerged as a promising direction in VSOD. Zhao et al.[8] introduced a weakly supervised VSOD network utilizing eye-movement scribble annotations and bidirectional LSTM for improved temporal information extraction. To reduce computational costs, Huang et al.[20] proposed a scribble attention module that optimizes unlabeled pixels and rectifies inaccurate segmentation regions. Similarly, Wang et al. [34] developed an approach that maps scribbles onto object contours to generate high-quality pseudo-labels. Despite these advancements, challenges such as error accumulation, model instability, or collapse can arise during the frame-by-frame propagation of generated masks [19, 20].

Knowledge-sharing strategy Knowledge Distillation [35] was initially designed to transfer knowledge from a larger teacher model to a smaller student model. In Transformer architectures, attention distillation [36] transfers attention knowledge from teacher to student models but typically requires both models to have the same number of attention heads. Wang et al.[37] addressed differences in attention head numbers and dimensions using interpolation and aggregation techniques. More recently, Zhao et al.[38] proposed a knowledge-sharing strategy to align attention maps between teacher and student models effectively. Unlike prior approaches, our method enhances the student model's performance during inference by leveraging high-quality attention maps modulated with scribble annotation masks, rather than relying directly on ground truth foreground-background masks.

3.3 Method

3.3.1 Overview

This section introduces our proposed KHMF-Net for scribble-supervised VSOD, designed to reduce error accumulation and enhance resilience to background distractions. As shown in Figure 3.1, the Hierarchical Memory Block (HMB) captures long-term spatiotemporal context for each video sequence, while the Adaptive Memory Fusion (AMF) module integrates saliency maps at varying confidence levels, reducing memory load and mitigating errors. Additionally, the Interactive Equalized Matching (IEM) module calculates feature similarity to balance pixel contributions and ensure consistent confidence. Our Knowledge-sharing strategy, based on a dual-attention mechanism, refines IEM for improved similarity computation.

3.3.2 Hierarchical Memory Bank

Our KHMF-Net maintains a hierarchical memory of historical mask states, consisting of three levels within our Hierarchical Memory Block (HMB). The first level stores the initial target scribble map, denoted as M_s, which the annotator provides. The second level contains the confidence region denoted as $M_c \in (0,1)\mathbb{R}^{T\times H\times W}$, which is generated by integrating the initial target scribble region from previous frames into the confidence region. The last level stores the predicted probability maps $M_e \in (0,1)\mathbb{R}^{T\times H\times W}$, which are used to segment the entire predicted target in previous frames. Using guidance from the initial target scribble memory and the existing confidence region memory, our HMB effectively captures the confidence region of the query frames. Furthermore, HMB expands the recently captured confidence region to include the predicted probability maps while considering the attributes of the confidence region and all previous memory data. This capability allows the HMB to efficiently mine the extended spatiotemporal context dependencies across the entire video sequence. The detailed methods for obtaining M_c, M_e are provided in Sec. 3.3.4.

3.3.3 Adaptive Memory Fusion

To effectively capture confidence levels, we design an Adaptive Memory Fusion (AMF) module to conduct a confidence-driven fusion operation with the memory bank. The AMF operation is performed as the weighted addition of the memory maps M_s, M_c, and M_e, each with varying confidence parameters, according to the following equation:

$$M_f = M_s + \alpha \cdot M_c + (1 - \alpha) \cdot M_e, \tag{3.1}$$

where $\alpha \in [0, 1]$ is learnable non-negative parameters used to adaptively control the weight assigned to each memory map based on its confidence. We forcibly assign M_e a weight of $(1 - \alpha)$, which prevents the high-confidence region M_c from dominating the fusion process. This approach ensures that the weights for M_e are not excessively neglected. This design choice is crucial to prevent the high-confidence region M_c from overpowering the fusion process. Instead of employing a crude concatenation approach, we use this weighted addition operation to efficiently reduce matrix dimensionality during the matching process. By incorporating various regions of different confidence levels, this operation can significantly improve the model's generalization. Notably, confidence scores are not assigned to the scribble map M_s as it inherently receives the highest confidence levels directly from the annotation.

3.3.4 Interactive Equalized Matching

To incorporate historical target information into the memory frame features effectively and facilitate the capture of high-confidence regions in the query frames as well as the expansion of the entire predicted target, we perform a memory encoding operation [39] before inputting to the Interactive Equalized Matching (IEM). It enabled our model to obtain the **e**ncoded memory features, denoted as $E_e \in \mathbb{R}^{1\times H\times W\times C}$ and $E_c \in \mathbb{R}^{1\times H\times W\times C}$ from $M_f \in \mathbb{R}^{T\times H\times W\times C}$. Specifically, the E_c corresponds to the case where $\lambda_\alpha = 0$ in Equation 3.1.

High-confidence region($\mathbf{M_c}$) IEM module matches the features of the t-th query frame $Q^t \in \mathbb{R}^{1\times H\times W\times C}$ with the encoded memory features E_c to obtain the query representation of confidence region $R_c^t \in \mathbb{R}^{1\times H\times W\times C}$. The captured confidence region representation in the query frame is computed as:

$$R_c^t = \Phi_c^{\text{Match}}(E_c, Q^t), \tag{3.2}$$

where Φ_c^{Match} denotes the memory matching function of confidence region. Subsequently, a segmentation head is constructed based on R_c^t to segment the confidence region and generate the salient map M_c^t for the t-th frame, defined by:

$$M_c^t = \Phi_c^{\text{Head}}(R_c^t, f^t). \tag{3.3}$$

The segmentation head operation, referred to as Φ_c^{Head}, takes into account the intermediate-layer features f^t of the query frame and provides additional spatial details. This process is visualized by the blue line in Figure 3.1.

Entire predicted target($\mathbf{M_e}$) To expand the entire predicted target in the query frame, we employ the closest captured confidence region as a reference along with the preceding memory information. The historical entire target information E_e and the high-confidence region of the query frame R_c^t are fed into the Interactive Equalized Matching (IEM) module, resulting in $R_e^t \in \mathbb{R}^{1\times H\times W\times C}$, which models the complete target information in the query frame.

$$R_e^t = \Phi_e^{\text{Match}}(E_e, R_c^t), \quad M_e^t = \Phi_e^{\text{Head}}(R_e^t, f^t), \tag{3.4}$$

where R_e^t and M_e^t represent the t-th query representation outputted by IEM and the segmentation head, respectively. The complete process follows the orange line depicted in Figure 3.1.

Memory matching The IEM module decodes the target state within the query frame, using the information fused from the memory bank as a reference. Through pixel-level feature matching, this module generates query representations that are used for subsequent segmentation tasks. Previous VSOD methods [12, 40] employed a transformer decoder architecture to build a surjective matching [41] module for calculating similarity. However, the absence of boundarydetails in scribble annotations results in matching outcomes that are influenced by confusing pixels, such as background distractions, in the reference frame. IEM equalizes the influence of reference frame pixels, mitigating the impact of background distractions. We use two IEM modules to process E_c and E_e, respectively. For simplicity, in the following, we illustrate IEM using the example of high-confidence regions E_c. The single-pixel locations of the embedded features extracted from the historical reference frame $E_{c(p)}$ and the query frame $Q_{(q)}^t$ are denoted as p and q, respectively. The affinity matrix A_c and A_e are

Table 3.1: Quantitative comparisons with state-of-the-art models on VSOD datasets. 'Sup.' denotes supervision type: 'F' (fully supervised), 'Un' (unsupervised), 'P' (point supervised), 'S' (scribble supervised). 'OF' indicates optical flow as a multi-source. ↑ and ↓ indicate higher and lower values are better, respectively. '-' means unavailable results.

Method	publisher	Sup.	Modality	DAVIS			DAVSOD			VOS		
				MAE↓	S_m↑	F_β↑	MAE↓	S_m↑	F_β↑	MAE↓	S_m↑	F_β↑
RCRNet [42]	ICCV19	F	RGB	0.027	0.886	0.848	0.087	0.741	0.653	0.051	0.873	0.833
PCSA [43]	AAAI20	F	RGB	0.022	0.902	0.880	0.086	0.741	0.655	0.065	0.827	0.747
MQP [44]	TCSVT21	F	OF+RGB	0.018	0.916	0.904	0.075	0.770	0.703	-	-	-
MMN [45]	TIP24	F	OF+RGB	0.020	0.897	0.877	0.065	0.777	0.708	0.069	0.828	0.768
IENet [46]	MTA24	F	RGB	0.017	0.912	0.896	0.069	0.775	0.862	0.042	0.882	0.845
GF [47]	TIP15	Un	OF+RGB	0.100	0.688	0.569	0.167	0.553	0.334	0.162	0.615	0.506
SAG [48]	CVPR15	Un	OF+RGB	0.103	0.676	0.515	0.184	0.565	0.370	0.172	0.619	0.482
SSOD [13]	CVPR20	S	RGB	0.044	0.795	0.734	0.101	0.672	0.556	0.106	0.682	0.648
WSVSOD [8]	CVPR21	S	OF+RGB	0.037	0.828	0.779	0.103	0.705	0.605	0.091	0.750	0.666
WSP [49]	ACMM22	P	OF+RGB	0.040	0.808	0.754	0.094	0.718	0.622	**0.074**	0.739	**0.796**
PSW [40]	IP23	S	RGB	0.031	**0.875**	0.854	0.084	0.738	0.638	0.075	0.716	**0.795**
CFMR [15]	PR24	S	OF+RGB	0.031	0.851	0.814	0.089	0.720	0.626	0.092	**0.751**	0.677
SAM-SNet [12]	TCSVT24	S	RGB	**0.028**	0.873	**0.883**	**0.071**	**0.770**	**0.682**	-	-	-
Ours		S	RGB	**0.022**	**0.890**	**0.886**	**0.070**	**0.739**	**0.713**	**0.073**	**0.752**	0.783

obtained by calculating the similarity between each p and q in the reference and query frames as follows:

$$A_c = sim(p, q) = E_{c(p)} \cdot Q^t_{(q)}, \tag{3.5}$$

$$A_e = sim(p, q) = E_{e(p)} \cdot R^t_{c(q)}. \tag{3.6}$$

It should be noted that in the processing of the orange line, R^t_c is a query to calculate similarity with E_e. To incorporate an equalization mechanism within the affinity matrix, we apply a Reference-Wise (R-W) Softmax along the query dimension as follows:

$$A \leftarrow \text{Softmax}(A). \tag{3.7}$$

Through the equalized matching process described above, the scores of all query frame pixels are normalized, ensuring equal contribution from each reference frame pixel in the prediction. The normalization process proportionately decreases the scores of heavily referenced pixels in the reference frame, like the background, to prevent an overreliance on the information from the reference frame. This essential suppression of matching scores for frequently referenced pixels is pivotal in minimizing distractions arising from significant errors.

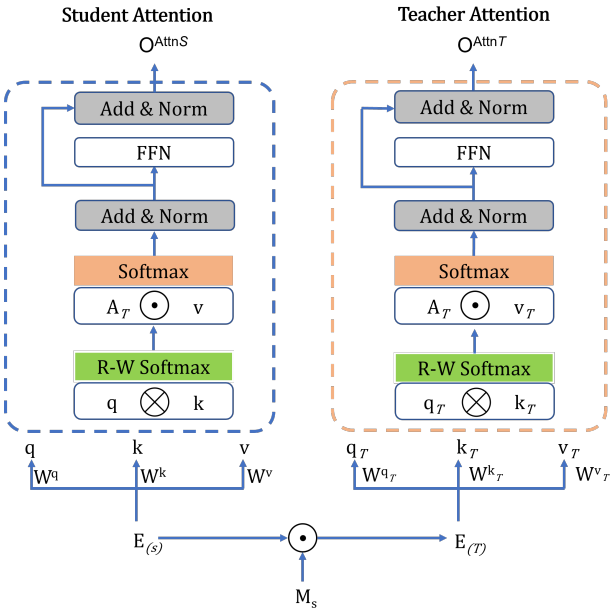

Figure 3.2: Knowledge-sharing strategy based on dual-attention.

3.3.5 Knowledge-sharing strategy

Inspired by the Knowledge Sharing **DE**tection **TR**ansformer (KS-DETR) [38] in object detection, we devise a knowledge-sharing strategy based on a dual-attention mechanism to enhance the matching process in segmentation with sparse annotations. Our approach facilitates the sharing of refined attention maps from *Teacher* to *Student* model during training, enhancing the learning process with scribble annotations (M_s) as additional attentional cues. The dual-attention mechanism is structured around two principal components: a Student Attention module and a Teacher Attention module. The structure of this dual-attention, as illustrated in Figure 3.2, operates with two sets of triplet parameters (q, k, v) containing replicated elements.

The **Student Attention**, represented as $AttnS$, obtains q, k, and v from $E_{(S)} \in \mathbb{R}^{HW \times d}$, which is flattened into tokens by $E_{c/e}$.

$$[q; k; v] = [EW^q; EW^k; EW^v], \tag{3.8}$$

where $W^q \in \mathbb{R}^{d_{model} \times d_q}$,$W^k \in \mathbb{R}^{d_{model} \times d_k}$, and $W^v \in \mathbb{R}^{d_{model} \times d_v}$ are the parameters of the scaled dot-product attention, and $d_{model} = d/head$.

In the **Teacher Attention** ($AttnT$), the pixels of the scribble map M_s are assigned a

value of 1, while the remaining pixels are set to 0. Multiplying $E_{c/e}$ and M_s element-wise and then flattening the result to obtain $E_{(T)}$ achieves the integration of M_s as additional attentional cues to enhance the learning process, as shown:

$$E_{(T)} = E_{c/e} \odot M_s, \tag{3.9}$$

where the subscript T indicates *Teacher*, as inspired by knowledge distillation [35]. The projections of the Teacher feature $E_{(T)}$ are used to obtain q_T, k_T, and v_T, as described:

$$[q_T; k_T; v_T] = [E_{(T)}W^{q_T}; E_{(T)}W^{k_T}; E_{(T)}W^{v_T}]. \tag{3.10}$$

Using the two groups of q, k, v, we generate the outputs O^{AttnS} and O^{AttnT} corresponding to $AttnS$ and $AttnT$, respectively. By sharing high-quality A_T from $AttnT$ to $AttnS$ as shown in Figure 3.2, O^{AttnS} is improved as they have to adapt themselves to fit with A_T through backpropagation. Shared decoder layers individually process these outputs to maintain separate training for each attention. The Teacher Attention $AttnT$ and the MLP are excluded during inference to avoid introducing additional parameters and computational overhead.

3.3.6 Loss Function

For scribble annotations, there are a great number of unlabeled pixels. Hence, we employ a composite loss function to train the model for segmenting the fully predicted target. The model parameters are optimized through the integration of Focal Loss [50], Partial Cross-Entropy Loss [51], and Bidirectional Prediction Consistency [19] as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_c^{FL} + \lambda_2 \mathcal{L}_e^{PCE} + \lambda_3 \mathcal{L}_e^{BPC}. \tag{3.11}$$

Here, λ_1, λ_2, and λ_3 are hyperparameter of weights that balance the loss functions. In our implementation, these weights are tuned to 1, 1, and 0.3, respectively, following the configuration used in previous works [8].

Focal Loss. The confidence region in the query frame is similar to the target scribble region in the initial frame, despite variations in the target's appearance across frames. To ensure accurate segmentation of the confidence region, we employ focal loss [50], which calculates probability maps of the confident region and the predicted target using target scribbles as supervision. This approach helps prevent misclassification

of pixels near the target scribble and addresses the imbalance between positive and negative pixels. It is defined as follows:

$$\mathcal{L}_c^{FL} = \frac{1}{N}\sum_{t=0}^{L-1}\sum_{(i,j)\in s^t}\phi^{fl}(M_c^{t,(i,j)}, Y_c^{t,(i,j)}). \tag{3.12}$$

In the above equation, ϕ^{fl} denotes the focal loss function. M_c^t and Y^t represent the probability maps of the confident region and the ground truth for the t-th frame, respectively. The notation (i, j) indicates the pixel index. The set s^t encompasses the index set of all pixels, excluding those ignored in the t-th frame. Additionally, N signifies the total count of pixels, excluding the ignored ones, while L denotes the length of the sequence training sample.

Partial Cross-entropy Loss. We reference the work by WSVSOD [8] and incorporate the use of partial cross-entropy loss (PCE) [51]. The PCE computes explicitly the cross-entropy loss within the area annotated by scribbles, excluding the other regions. Mathematically, it is defined as:

$$\mathcal{L}_e^{PCE} = -\frac{1}{N'}\sum_{t=0}^{L-1}\sum_{(i,j)\in\Omega_t} log(M_e^{t,(i,j),c}), \tag{3.13}$$

Here, $M_e^{t,(i,j),c}$ denotes the predicted probability that the pixel located at position (i, j) in the t-th frame is ground-truth object identity c. The symbol Ω_t refers to the set of indices corresponding to labeled pixels within the t-th frame. Furthermore, N signifies the total number of labeled pixels across the dataset.

Bidirectional Prediction Consistency. We employ the Bidirectional Prediction Consistency loss (BPC) [19] to ensure coherence between forward and backward predictions in segmentation. By enforcing a consistency constraint between forward and backward predictions, the BPC uncovers and leverages frame-level dependencies in a sequential training sample. Formally, the BPC loss for a sequence training sample is defined as follows:

$$\mathcal{L}_e^{BPC} = \frac{1}{N''}\sum_{t=0}^{L-1}\sum_{(i,j)}\|M_e^{t,(i,j)} - B_e^{t,(i,j)}\|^2. \tag{3.14}$$

Here, M_e^t and B_e^t represent the computed probability maps for the t-th frame in the

forward and backward predictions, respectively.

3.4 Experiments

We assess the performance of our proposed model on three public VSOD benchmark datasets: DAVIS [21], VOS [23], and DAVSOD [22]. Three evaluation criteria are employed to assess both our method and the competing methods, which consist of Mean Absolute Error (MAE), S-measure (S_m), and F-measure (F_β).

3.4.1 Implementation details

Model details: We employ a pretrained ResNet-50 model [52] on the ImageNet [53] as our backbone architecture. We capture the output from the *conv4* layer for feature extraction. Our segmentation head comprises three refinement modules, each is integrated with skip connections to the backbone's intermediate layers. This progressive configuration systematically enhances the spatial dimensions of the query representation from $\frac{H}{16} \times \frac{H}{16}$ to $\frac{H}{4} \times \frac{H}{4}$. To generate the final probability map with dimensions $H \times W$, we employ a convolutional layer, followed by an interpolation operation and a softmax layer.

Training We implement a two-stage training approach for our model, following the practice of several existing methodologies [54]. Initially, the model undergoes training on static images utilizing the S-DUTS saliency dataset [13], followed by creating sequence training samples using image augmentation. Subsequently, in the second stage, the model undergoes further training on video sequences sourced from DAVIS [21] and Youtube-VOS [55]. Further details on these datasets can be found in Chapter 2.

Inference Our model removes the need for Teacher Attention operations during inference, decreasing extra parameters and computational overhead. The inference stage relies solely on the student attention mechanism, which utilizes the standard scaled dot-product attention to generate output features.

3.4.2 Comparisons with State-of-the-art Models

Quantitative Comparison Table 3.1 presents the results of our method compared to the other SOTA methods. Our KHMF-Net outperforms all weakly supervised

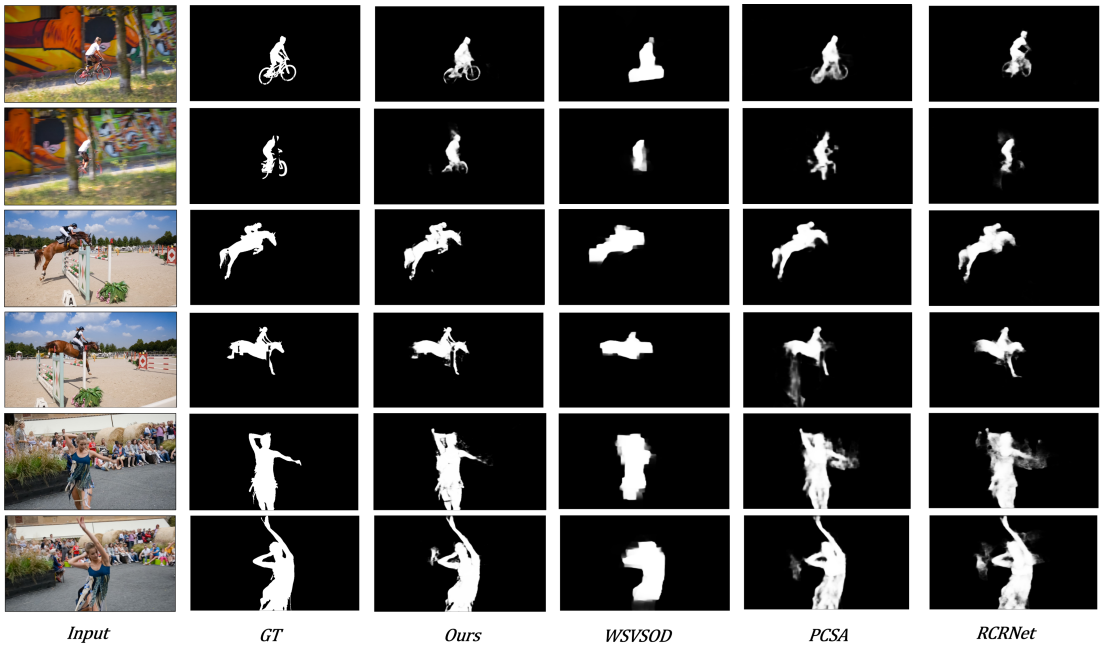

Figure 3.3: Visual comparisons with state-of-the-art VSOD methods. Each column represents a method, and each row shows the saliency map of a frame on DAVIS. The second column displays the object-level ground truth (GT). Notably, some of our results outperform even certain fully supervised methods.

and unsupervised methods, including GF [47], SAG [48], SSOD [13], WSVSOD [8], WSVP [49], PSW [40], CFMR [15] and SAM-SNet [12]. Furthermore, it demonstrates competitiveness with fully supervised methods, such as RCRNet [42], PCSA [43], MQP [44], MMN [45], and IENet [46].

Qualitative Comparison To further assess the effectiveness of our proposed method, we present visual examples in Figure 3.3, comparing it with other advanced methods. Our model excels at capturing intricate details and edges of salient objects, surpassing the performance of the previous weakly supervised method WSVSOD [8]. Notably, our model even outperforms fully supervised methods such as PSCA [30] and RCRNet [42] in certain cases. Moreover, our approach demonstrates comparable performance to fully supervised methods in challenging scenarios characterized by complex scenes (*row 1,2*), swift motion (*row 3, 4*), and variations pose (*row 5, 6*). Impressively, as shown *row 2*, our method accurately predicts the structural information of salient objects, even when partially occluded, demonstrating the effectiveness of the Hierarchical Memory Block in preserving object information from historical frames. Additionally, as seen in *rows 3 and 4*, our method excels at localizing rapid motion of salient objects,

Table 3.2: Experimental results for seven variations of our method on the DAVIS are showcased. The superior scores are emphasized in the red font.

AMF	R-W	Dual-attn	$\mathcal{L}_{FL}$	$\mathcal{L}_{PCE}$	$\mathcal{L}_{BPC}$	MAE↓	S_m↑	F_β↑
✓	✓	✓	✓	✓	✓	**0.022**	0.890	**0.886**
Concat	✓	✓	✓	✓	✓	0.029	0.863	0.851
Add	✓	✓	✓	✓	✓	0.027	0.835	0.863
✓	×	✓	✓	✓	✓	0.026	0.851	0.848
✓	✓	Single	✓	✓	✓	0.025	**0.891**	0.882
✓	✓	✓	✓	×	✓	0.029	0.846	0.863
✓	✓	✓	✓	✓	×	0.033	0.794	0.842
✓	✓	✓	✓	×	×	0.038	0.763	0.835

further emphasizing its capability in capturing robust spatiotemporal context. Finally, in scenarios with frequent movement variations, such as dancing (*rows 5 and 6*), our model effectively captures intricate details and sharp edges of salient objects while minimizing background distractions.

3.4.3 Ablation Analyses

Ablation studies were conducted to analyze KHMF-Net, with evaluations on the DAVIS [21]. Table 3.2 shows that our method achieves the best performance when all components are included, highlighting the importance of the loss functions and modules for effective training.

3.4.4 Effect of Adaptive Memory Fusion

To evaluate the influence of the Adaptive Memory Fusion (AMF) module, we provide an experiment where it was substituted with a straightforward fusion approach that lacks the adaptive parameters α and γ, and additionally, with a direct concatenation operation applied to salient maps from the Memory Bank. The variants are defined as:

$$M_f^{add} = M_s + M_c + M_e, \tag{3.15}$$

$$M_f^{concat} = \text{Concat}(M_s, M_c, M_e). \tag{3.16}$$

Compared to our fully integrated model (as depicted in the *first row*), Replacing AMF with more straightforward methods such as concatenation (*second row*) and addition (*third row*) within the Memory Fusion Block leads to a decline in performance by 0.7% and 0.5% in terms of *MAE*, respectively. This observation underscores the

Table 3.3: Experimental results of our KHMF-Net under different memory capacity on DAVIS 2016 validation set.

	c=1	c=2	c=4	c=8	c=16
$MAE \downarrow$	0.037	0.029	**0.022**	0.026	0.028
$S_m \uparrow$	0.798	0.887	**0.890**	0.883	0.875
$F_\beta \uparrow$	0.867	0.878	**0.886**	0.882	0.880

significance of the AMF in enhancing the model's overall effectiveness. The training process for the adaptive parameters α and γ is illustrated in Figure 3.4, showing that these parameters exhibit stabilization after epoch 83.

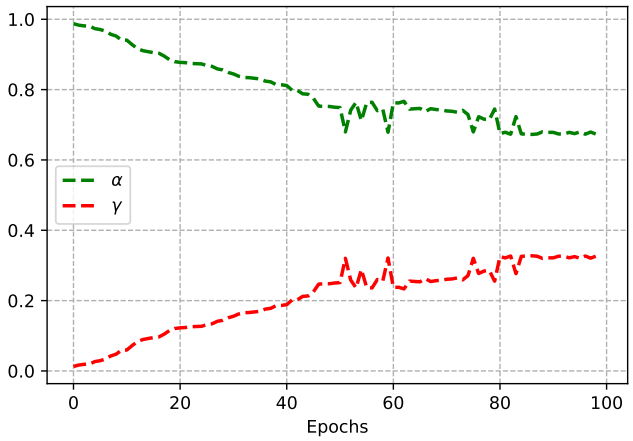

Figure 3.4: The training process for the adaptive parameters α and γ of Adaptive Memory Fusion

Interactive Equalized Matching To substantiate the effectiveness of Interactive Equalized Matching, we conducted ablation experiments on the Reference-Wise (R-W) Softmax component. Compared to the surjective matching method [41], the R-W Softmax equalizes the potential contributions to the query frame, thereby preventing excessive referencing of reference frame details like the background. Visual examples illustrating the effects of surjective matching and IEM can be found in Figure 3.5. It is evident that equalizing the matching scores aids in mitigating excessive referencing of the background, thus bolstering the robustness of the equalized matching.

Effect of the memory capacity To explore the impact of memory capacity $\boldsymbol{c}$, we evaluate our method's performance across various memory capacities and present the experimental findings in Table 3.4. Performance demonstrates enhancement as the

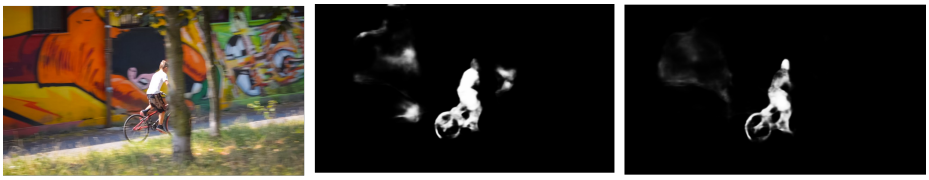

Figure 3.5: Visualization of surjective matching and IEM

memory capacity $\boldsymbol{c}$ increases until it reaches saturation around $\boldsymbol{c} = 4$.

Knowledge-sharing strategy Compared with the experiment utilizing single attention (*the fifth row*), our complete model utilizing the scribble annotation mask as an additional clue enhances the performance of salient object detection. By integrating the Knowledge-sharing strategy, our model directly shared the enhanced weights from the Teacher Attention, rather than solely mimicking these attentions.

Effect of Partial Cross-Entropy Loss In the *sixth row* of Table 3.2, we list the results of investigating the effects of excluding the smoothness loss $\mathcal{L}_{PCE}$ from the overall loss function $\mathcal{L}$. Excluding $\mathcal{L}_{PCE}$ leads to a significant performance decline of 0.7%. This outcome highlights the crucial role of dense supervision in accurately predicting the entire target mask and reaffirms the effectiveness of leveraging pairwise pixel relationships for achieving dense supervision.

Effect of Bidirectional Prediction Consistency Loss To evaluate the impact of the Bidirectional Prediction Consistency (BPC) loss, we conduct an additional experiment in which $\mathcal{L}_{BPC}$ is excluded from the overall loss function $\mathcal{L}$. The model is then trained using only the residual loss components, as illustrated in the *seventh row*. The decrease in performance highlights the importance of $\mathcal{L}_{BPC}$ in improving video salient object detection, especially in scribble-supervised learning.

3.4.5 Ablation Analyses

A series of ablation studies are conducted to analyze our proposed KHMF-Net, with evaluations performed on the DAVIS dataset [21]. Table 3.2 presents that our method achieved the highest performance when all components were included, indicating the essential nature of the loss functions and modules for successful training.

Table 3.4: Experimental results of our KHMF-Net under different memory capacity on DAVIS 2016 validation set.

	c=1	c=2	c=4	c=8	c=16
$MAE \downarrow$	0.037	0.029	**0.022**	0.026	0.028
$S_m \uparrow$	0.798	0.887	**0.890**	0.883	0.875
$F_\beta \uparrow$	0.867	0.878	**0.886**	0.882	0.880

3.4.5.1 Effect of Adaptive Memory Fusion

To evaluate the influence of the Adaptive Memory Fusion (AMF) module, we provide an experiment where it was substituted with a straightforward fusion approach that lacks the adaptive parameters α and γ, and additionally, with a direct concatenation operation applied to salient maps from the Memory Bank. The variants are defined as:

$$M_f^{add} = M_s + M_c + M_e, \tag{3.17}$$

$$M_f^{concat} = \text{Concat}(M_s, M_c, M_e). \tag{3.18}$$

3.4.5.2 Effect of Adaptive Memory Fusion

Compared to our fully integrated model (as depicted in the *first row*), Replacing AMF with more straightforward methods such as concatenation (*second row*) and addition (*third row*) within the Memory Fusion Block leads to a decline in performance by 0.7% and 0.5% in terms of *MAE*, respectively. This observation underscores the significance of the AMF in enhancing the model's overall effectiveness.

3.4.5.3 Interactive Equalized Matching

To substantiate the effectiveness of Interactive Equalized Matching, we conducted ablation experiments on the Reference-Wise (R-W) Softmax component. Compared to the surjective matching method [41], the R-W Softmax equalizes the potential contributions to the query frame, thereby preventing excessive referencing of reference frame details like the background. Visual examples illustrating the effects of surjective matching and IEM can be found in Figure 3.5. It is evident that equalizing the matching scores aids in mitigating excessive referencing of the background, thus bolstering the robustness of the equalized matching.

3.4.5.4 Effect of the memory capacity

To explore the impact of memory capacity $\boldsymbol{c}$, we evaluate our method's performance across various memory capacities and present the experimental findings in Table 3.4. Performance demonstrates enhancement as the memory capacity $\boldsymbol{c}$ increases until it reaches saturation around $\boldsymbol{c} = 4$.

3.4.5.5 Knowledge-sharing strategy

Compared with the experiment utilizing single attention (*the fifth row*), our complete model utilizing the scribble annotation mask as an additional clue enhances the performance of salient object detection. By integrating the Knowledge-sharing strategy, our model directly shared the enhanced weights from the Teacher Attention, rather than solely mimicking these attentions.

3.4.5.6 Effect of Partial Cross-Entropy Loss

In the *sixth row* of Table 3.2, we list the results of investigating the effects of excluding the smoothness loss $\mathcal{L}_{PCE}$ from the overall loss function $\mathcal{L}$. Excluding $\mathcal{L}_{PCE}$ leads to a significant performance decline of 0.7%. This outcome highlights the crucial role of dense supervision in accurately predicting the entire target mask and reaffirms the effectiveness of leveraging pairwise pixel relationships for achieving dense supervision.

3.4.5.7 Effect of Bidirectional Prediction Consistency Loss

To evaluate the impact of the Bidirectional Prediction Consistency (BPC) loss, we conduct an additional experiment in which $\mathcal{L}_{BPC}$ is excluded from the overall loss function $\mathcal{L}$. The model is then trained using only the residual loss components, as illustrated in the *seventh row*. The decrease in performance highlights the importance of $\mathcal{L}_{BPC}$ in improving video salient object detection, especially in scribble-supervised learning.

3.5 Conclusion

We have presented a novel weakly-supervised Video Salient Object Detection (VSOD) network to address error accumulation and susceptibility to background distractions.

We propose a Hierarchical Memory Bank (HMB) to store historical segmentations at three confidence levels, facilitating an effective expansion process through scribble annotation. An Adaptive Memory Fusion (AMF) module with an Interactive Equalized Matching (IEM) module is proposed to enhance query performance and ensure consistent confidence assessments across diverse conditions. We also devise a Knowledge-sharing strategy based on a dual-attention mechanism to improve matching process using sparse annotations. Experimental results demonstrate that our model effectively competes against weakly supervised VSOD methods and even outperforms some fully supervised methods, representing a significant advancement in VSOD research.

References

[1] Mingqi Gao, Feng Zheng, James JQ Yu, Caifeng Shan, Guiguang Ding, and Jungong Han. Deep learning for video object segmentation: a review. *Artificial Intelligence Review*, 56(1):457–531, 2023.

[2] Usman Muhammad, Mourad Oussalah, and Jorma Laaksonen. Saliency-based video summarization for face anti-spoofing. *Pattern Recognition Letters*, 185:190–196, 2024.

[3] Fuling Lin, Changhong Fu, Yujie He, Fuyu Guo, and Qian Tang. Learning temporary block-based bidirectional incongruity-aware correlation filters for efficient uav object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2160–2174, 2020.

[4] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

[5] Léo Maczyta, Patrick Bouthemy, and Olivier Le Meur. CNN-based temporal detection of motion saliency in videos. *Pattern Recognition Letters*, 128:298–305, 2019.

[6] Kan Huang, Chunwei Tian, Jingyong Su, and Jerry Chun-Wei Lin. Transformer-based cross reference network for video salient object detection. *Pattern Recognition Letters*, 160:122–127, 2022.

[7] Xuelu Feng, Sanping Zhou, Zixin Zhu, Le Wang, and Gang Hua. Local to global feature learning for salient object detection. *Pattern Recognition Letters*, 162:81–88, 2022.

[8] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16835, 2021.

[9] Zhengyi Liu, Xiaoshen Huang, Guanghui Zhang, Xianyong Fang, Linbo Wang, and Bin Tang. Scribble-Supervised RGB-T Salient Object Detection. *arXiv preprint arXiv:2303.09733*, 2023.

[10] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 670–678, 2022.

[11] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8843–8850, 2019.

[12] Binwei Xu, Qiuping Jiang, Xing Zhao, Chenyang Lu, Haoran Liang, and Ronghua Liang. Multidimensional Exploration of Segment Anything Model for Weakly Supervised Video Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[13] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12546–12555, 2020.

[14] Binwei Xu, Haoran Liang, Weihua Gong, Ronghua Liang, and Peng Chen. A visual representation-guided framework with global affinity for weakly supervised salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[15] Jian Wang, Siyue Yu, Bingfeng Zhang, Xinqiao Zhao, Ángel F García-Fernández, Eng Gee Lim, and Jimin Xiao. Cross-frame feature-saliency mutual reinforcing for weakly supervised video salient object detection. *Pattern Recognition*, page 110302, 2024.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[17] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-lun Chao. Segment Anything Model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*, 2023.

[19] Zikun Zhou, Kaige Mao, Wenjie Pei, Hongpeng Wang, Yaowei Wang, and Zhenyu He. Reliability-Hierarchical Memory Network for Scribble-Supervised Video Object Segmentation. *arXiv preprint arXiv:2303.14384*, 2023.

[20] Peiliang Huang, Junwei Han, Nian Liu, Jun Ren, and Dingwen Zhang. Scribble-supervised video object segmentation. *IEEE/CAA Journal of Automatica Sinica*, 9(2):339–353, 2021.

[21] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[22] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. pages 8554–8564, 2019.

[23] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2017.

[24] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

[25] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. SCOM: Spatiotemporal constrained optimization for salient object detection. *IEEE Transactions on Image Processing*, 27(7):3345–3357, 2018.

[26] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–731, 2018.

[27] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3243–3252,

2018.

[28] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7274–7283, 2019.

[29] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1553–1563, 2021.

[30] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10869–10876, 2020.

[31] Xing Zhao, Haoran Liang, Peipei Li, Guodao Sun, Dongdong Zhao, Ronghua Liang, and Xiaofei He. Motion-aware Memory Network for Fast Video Salient Object Detection. *arXiv preprint arXiv:2208.00946*, 2022.

[32] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016.

[33] Yunqiu Xu, Xin Yu, Jing Zhang, Linchao Zhu, and Dadong Wang. Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting. *IEEE Transactions on Image Processing*, 31:2148–2161, 2022.

[34] Xiongying Wang, Zaid Al-Huda, Bo Peng, and Xin Tang. Weakly Supervised Salient Object Detection by Hierarchically Enhanced Scribbles. *International Journal of Pattern Recognition and Artificial Intelligence*, 37(02):2355003, 2023.

[35] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[36] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[37] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*, 2020.

[38] Kaikai Zhao and Norimichi Ukita. KS-DETR: Knowledge Sharing in Attention

Learning for Detection Transformer. *arXiv preprint arXiv:2302.11208*, 2023.

[39] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. pages 8090–8100, 2022.

[40] Zelin Lu, Haoran Liang, Binwei Xu, and Ronghua Liang. A progressive segmentation with weight contrast label enhancement for weakly supervised video salient object detection. *IET Image Processing*, 2023.

[41] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9481–9490, 2019.

[42] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7284–7293, 2019.

[43] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10869–10876, 2020.

[44] Chenglizhao Chen, Jia Song, Chong Peng, Guodong Wang, and Yuming Fang. A novel video salient object detection method via semisupervised motion quality perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2732–2745, 2021.

[45] Xing Zhao, Haoran Liang, Peipei Li, Guodao Sun, Dongdong Zhao, Ronghua Liang, and Xiaofei He. Motion-aware memory network for fast video salient object detection. *IEEE Transactions on Image Processing*, 2024.

[46] Tao Jiang, Yi Wang, Feng Hou, and Ruili Wang. IENet: inheritance enhancement network for video salient object detection. *Multimedia Tools and Applications*, pages 1–20, 2024.

[47] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[48] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3402, 2015.

[49] Shuyong Gao, Haozhe Xing, Wei Zhang, Yan Wang, Qianyu Guo, and Wenqiang Zhang. Weakly supervised video salient object detection via point supervision. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3656–3665, 2022.

[50] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[51] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2018.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[54] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021.

[55] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

Chapter 4

Multimodal Energy Prompting for Video Salient Object Detection

Video Salient Object Detection (VSOD) aims to identify the most visually conspicuous objects in videos and extract key information from complex visual scenes. Recent studies attempt to combine optical flow and depth for complementary feature extraction. However, suboptimal fusion strategies often treat these modalities merely as extensions of the RGB stream, failing to fully leverage their unique semantic contributions. To address this limitation, we propose a novel Multimodal Energy Prompting Network (MEPNet), which utilizes implicit prompts derived from optical flow and depth within a pre-trained Segment Anything Model (SAM). This approach enhances VSOD by effectively integrating the complementary dynamic and structural information from these modalities. Particularly, we introduce a Spectrogram Energy Generator (SEG) to extract spectrogram energy from OF and depth, generating energy-driven prompts to fine-tune SAM via the Modality Energy Adapter (MEA), effectively mitigating noise interference and improving segmentation accuracy. In addition, we propose a Circular High-frequency Filter (CHF) to enhance RGB modality details using an adaptive circular mask. Extensive experiments on five VSOD benchmark datasets demonstrate that our MEPNet outperforms state-of-the-art approaches, achieving superior performance. Furthermore, our MEPNet can be generalized in the Video-Camouflaged Object Detection (VCOD) task and also achieve competitive results. Note that the content presented in this chapter has been published in the Pattern Analysis and Applications.

4.1 Introduction

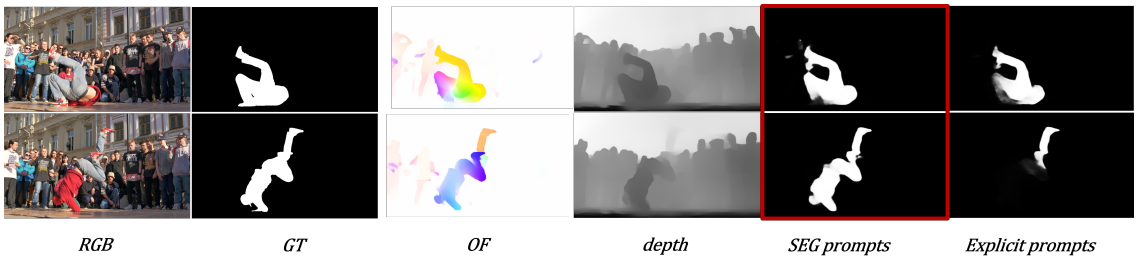

Figure 4.1: The figure shows the RGB, OF (optical flow), depth, and GT (ground truth), respectively. In complex backgrounds, both OF and depth exhibit significant noise, and treating them as equivalent or supplementary modalities overlooks their distinct semantic contributions. Furthermore, our experiments indicate that although OF and depth data can serve as either implicit or explicit prompts for SAM, the implicit prompts generated by the proposed Spectrogram Energy Generator (SEG) significantly reduce noise interference, thereby enhancing performance.

Video salient object detection (VSOD) aims to identify the most visually conspicuous objects in videos and extract key information from complex visual scenes. VSOD is essential to understand video content and has substantial implications for various real-world applications, including video compression [1], video summarization [2], autonomous driving [3], and video surveillance [4].

Despite extensive research conducted on VSOD, most existing methods [5–7] focus mainly on the synergy between temporal and spatial information. Intuitively, temporal features are typically extracted from optical flow (OF) and have been widely used to improve short-term motion details [8–10]. However, these methods, relying solely on OF and RGB data, cannot accurately discern spatial relationships and depth variances between the target object and the background in complex scenes. Depth cues provide valuable supplementary spatial position information, and have been effectively utilized in static image Salient Object Detection (SOD) [11, 12] with impressive results in feature fusion optimization. Nevertheless, spatial depth information remains insufficiently explored in the context of VSOD. Recent approaches [13, 14] have attempted to integrate OF and depth concurrently as complementary dynamic and spatial features. However, these methods often treat OF and depth merely as extensions of the RGB stream, overlooking their distinct semantic contributions. Consequently, their fusion strategies remain suboptimal, failing to fully harness the complementary

potential of these modalities to model complex video contexts.

To address the challenges associated with directly fusing OF and depth with RGB data, we propose MEPNet, an innovative Multimodal Energy Prompting Network (MEPNet), that leverages implicit prompts derived from OF and depth within the pre-trained Segment Anything Model (SAM) [15]. As a powerful foundation model for segmentation, SAM is designed to generate high-quality masks from diverse prompts, making it particularly well-suited for capturing complex visual cues in VSOD. By integrating SAM, our network improves the extraction and generalization of task-specific VSOD features, enabling more robust and adaptive in dynamic video scenes. Given that the intertwined dynamics of object movement and camera ego-motion [16] introduce noise into the OF and depth modalities (as shown in Figure 4.1) directly using them as explicit prompts for SAM may degrade segmentation accuracy. Inspired by the stable energy and rhythmic consistency observed in speech signals [17, 18], we design a Spectrogram Energy Generator (SEG) to compute Spectrogram Energy for each frame and integrate it into SAM through the Modality-Energy Adapter (MEA). Similar to how an individual's speech exhibits stable energy patterns, object motion in videos often follows an inertia-driven trajectory, resulting in consistent energy distributions across frames. SEG utilizes the Short-Time Fourier Transform (STFT) to compute these energy values, producing multimodal energy prompts that enable robust region membership assessment. The generated multimodal energy prompts mitigate noise interference and significantly outperform direct OF and depth prompting strategies.

Furthermore, we introduce a Circular High-frequency Filter (CHF) to refine fine details in the RGB domain by leveraging a SAM-driven spatio-temporal network guided by broad semantic features. Unlike previous methods [19, 20], which rely on manually defined filters, our approach employs a circular mask whose radius is adaptively adjusted through learnable parameter. This design is both straightforward and flexible, effectively overcoming the limitations of manual tuning. Extensive evaluations on five VSOD benchmark datasets demonstrate that our network consistently outperforms state-of-the-art methods, achieving superior results across multiple performance metrics.

In summary, our main contributions are as follows.

- We propose a Multimodal Energy Prompting Network (MEPNet), which lever-

ages implicit prompts derived from optical flow and depth to fine-tune a SAM via the Modality-Energy Adapter (MEA). By utilizing SAM's segmentation capability, our method effectively integrates dynamic and structural cues, improving the extraction and generalization of task-specific VSOD features.

- We design a Spectrogram Energy Generator (SEG), which draws inspiration from the concept of energy in speech processing. SEG applies the Short-Time Fourier Transform (STFT) to compute Spectrogram Energy, thereby effectively minimizing noise interference and outperforming the direct use of optical flow and depth as prompts in VSOD tasks.

- To mitigate noise interference in multimodal prompting, we design a Spectrogram Energy Generator (SEG) inspired by speech processing. SEG applies the Short-Time Fourier Transform (STFT) to compute Spectrogram Energy and reduce noise interference, outperforming the direct use of optical flow and depth as prompts in VSOD tasks.

- We propose a Circular High-frequency Filter (CHF) that refines high-frequency details in RGB images. Unlike previous hand-crafted filters, CHF adaptively adjusts its radius using learnable parameter, enabling automatic and precise filtering.

- Our method achieves state-of-the-art performance on VSOD benchmarks. Furthermore, it demonstrates strong generalization to Video Camouflage Object Detection (VCOD) tasks, surpassing existing methods on two VCOD datasets, MoCA-Mask [21] and CAD [22], thus further validating its effectiveness across various video segmentation challenges.

The following content of this chapter is: Section 4.2 briefly reviews recent WSOD methods. Section 4.3 details the proposed MEPNet. Section 4.4 gives quantitative and qualitative experiments. Section 4.5 is the conclusion of the chapter.

4.2 Related work

We provide a brief primer on Salient Object Detection in video sequences. Meanwhile, this section also reviews common prompt techniques in computer vision and Spectrogram Energy theory.

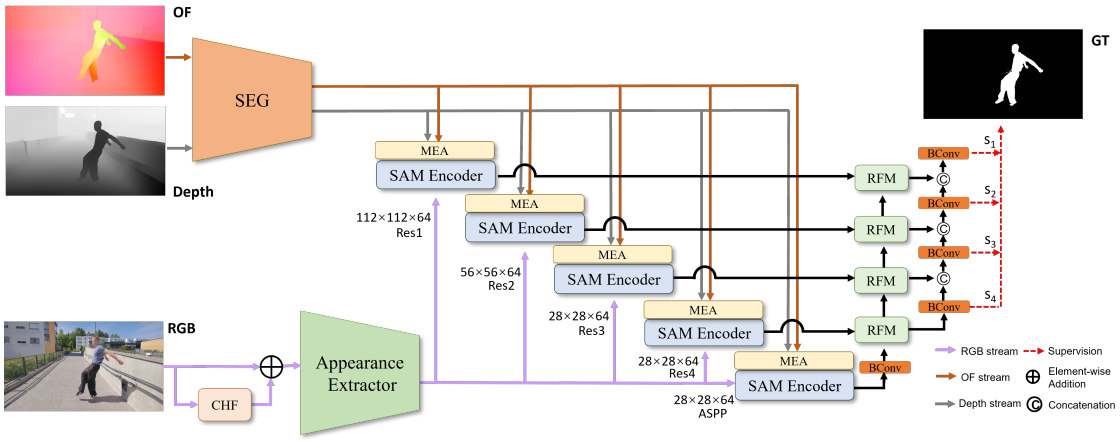

Figure 4.2: Overview of our MEPNet architecture. The Spectrogram Energy Generator (SEG) generates motion energy prompts from optical flow (OF) and depth streams, which are processed by the Modality-Energy Adapter (MEA) before being fed into the SAM encoder, alongside enhanced RGB features extracted by the Appearance Extractor. The Circular High-frequency Filter (CHF) further enhances RGB details. SAM encoder layers are fine-tuned with these prompts and features. The Refinement Fusion Module (RFM) employs *BConv* layers for multiscale feature fusion, resulting in the final salient map.

4.2.1 Video Salient Object Detection

Many deep learning methods for Video Salient Object Detection (VSOD) have achieved significant success [5–7, 23]. Several approaches extract motion details from optical flow to tackle complex real-world scenarios and enhance their VSOD performance. Some methods [8, 10, 24] uncover spatio-temporal correlations in object saliency by using optical flow as an auxiliary to guide the RGB modality. Meanwhile, some methods [25, 26] treat optical flow and RGB equally, using them together as inputs to exploit temporally concatenated deep features. Unlike in RGB-D static image SOD, where depth information is often used independently, depth information in VSOD is typically combined with optical flow as auxiliary modalities to support the RGB modality. This combination better adapts to complex backgrounds by leveraging depth and motion cues. Lin et al.[14] integrated depth and optical flow with RGB information through modality-specific branches to improve VSOD performance. Li et al.[13] introduced DCTNet+, a three-stream network using depth and optical flow as auxiliary modalities, with multimodal attention and post-fusion modules. Chen et al.[27] integrated dynamic message propagation with a multilevel strategy to fuse multimodal information for VSOD. Inspired by recent advancements, our research

aims to enhance VSOD accuracy by incorporating additional depth and optical flow information, expanding the field of multimodal VSOD.

4.2.2 Prompt in Computer Vision

Prompting originated in NLP [28] and has been successfully integrated into various computer vision applications [29]. VPT [30] uses learnable parameters as implicit prompts for transformer encoders, exceeding the full fine-tuning in many downstream tasks. Similarly, AdaptFormer [31] introduces lightweight modules to ViT, outperforming fully fine-tuned models in action recognition. Explicit prompting[19, 20], applied as task-specific knowledge, enhances model generalization in downstream tasks, particularly in scenarios with limited labeled data. ViPT[32] introduces modality-complementary prompts for multimodal tracking. In contrast, our MEPNet is designed for multimodal dense prediction in VSOD, introducing an innovative multi-prompt learning strategy that leverages energy prompts across multiple modalities. Extensive experiments validate its superior performance in multimodal VSOD.

4.2.3 Spectrogram Energy

The Short-Time Fourier Transform (STFT) is a well-established method in speech signal processing that is commonly used to capture temporal and spectral characteristics in audio signals. This technique has proven effective for tasks such as voice activity detection [33, 34], speech enhancement [18], and emotion recognition [17] by analyzing energy distribution patterns over short time intervals. Our concept of energy is different from existing measures in computer vision [35, 36], which typically rely on motion magnitude and motion direction. Our MEPNet pioneers the use of STFT to evaluate Spectrogram Energy changes across video frames, specifically to depth and optical flow data. This innovative strategy enhances motion dynamics representation and effectively leverages multimodal information to improve video salient object detection.

4.3 Method

In this section, we present MEPNet, beginning with an exploration of its multi-stream

encoder-decoder architecture and its functionality in processing various modalities. We then delve into the Spectrogram Energy Generator (SEG), which generates dynamic prompts to enhance feature representation. Following that, we discuss the SAM encoder and the Refinement Fusion Module (RFM), emphasizing their roles in achieving precise feature integration. We also analyze the contributions of the Circular High-frequency Filter (CHF) in refining RGB stream.

4.3.1 MEPNet Architecture Overview

The architecture of our MEPNet is illustrated in Figure 4.2, which shows a multi-stream encoder-decoder structure. Specifically, the Spectrogram Energy Generator (SEG) computes Spectrogram Energy from auxiliary modalities (optical flow and depth) to generate prompts. Subsequently, the SAM encoder processes the enhanced RGB features and energy prompts using a ViT-H/16 model, which features 14×14 windowed attention and four global attention blocks. To further enhance the RGB modality, we apply the Circular High-frequency Filter (CHF) to capture high-frequency components from the frequency domain. These enhanced RGB features are then refined by the Appearance Extractor. Notably, the pre-trained SAM encoder's weights are kept fixed, with minimal trainable parameters added for efficient training. The Refinement Fusion Module (RFM) predicts the final mask using a multiscale strategy to leverage higher-resolution feature maps. Additionally, *BConv* modules are employed for multiscale feature fusion, which ultimately generates the salient map.

4.3.2 Spectrogram Energy Generator(SEG)

We introduce a novel Spectrogram Energy Generator (SEG) that enhances the analysis of dynamic changes in video data. Unlike previous methods [35, 36] that rely primarily on RGB spatial data in the optical flow field to determine the motion distribution, we directly assess dynamic changes by computing the Spectrogram Energy of auxiliary modalities, specifically optical flow and depth. The magnitude spectrum, commonly used in image processing to analyze texture and edges, serves as a fundamental tool. Using the temporal features of the video, we calculate the Spectrogram Energy of optical flow $E^t_{(o)}$ and depth modality $E^t_{(d)}$ over multiple frames to effectively capture

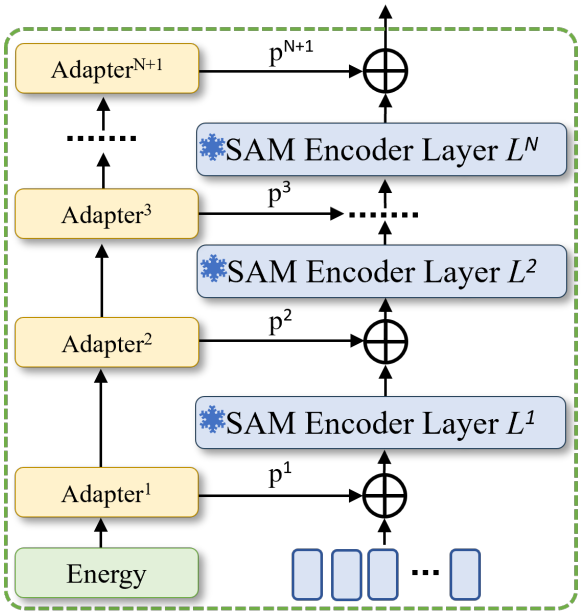

Figure 4.3: Architecture of the proposed Modality-Energy Adapter (MEA). The Spectrogram Energy tunes the RGB-modality features, and the Adapter merges these features.

dynamic changes. The computation is performed as follows:

$$E^t_{(o)} = \sqrt{\sum_f |X(f_{(o)}, t)|^2}, \tag{4.1}$$

$$E^t_{(d)} = \sqrt{\sum_f |X(f_{(d)}, t)|^2}, \tag{4.2}$$

where f and t represent the frequency and the t-th query frame, respectively, and X denotes the short-time Fourier transform (STFT). The calculated $E^t_{(o)}$ and $E^t_{(d)}$ serve as explicit prompts to fine-tune the SAM encoder.The optical flow and depth information frequently involve noise resulting from the complex interplay between object motion, camera ego-motion, and intricate backgrounds. Our method of Spectrogram Energy filtering effectively eliminates noise stemming from irregular energy levels, yielding more precise prompts for prompt learning and enhancing the model's guidance accuracy.

4.3.3 Modality-Energy Adapter (MEA)

We integrate Spectrogram Energy as an additional auxiliary modality, temporally synchronized and spatially aligned with the RGB stream. As shown in Figure 4.3, the MEA module first inputs the energy flow $E_{(o)}$ and $E_{(d)}$ into a patch embedding layer. The E is then mapped and flattened into a D-dimensional latent space, forming Energy tokens of the same dimension as the RGB tokens. To ensure efficient and effective adaptation across all layers, we use an adapter comprising two MLPs with an activation function in between. In this configuration, the i-th adapter processes the energy information E^i to generate the prompt P^i by:

$$P^i = MLP_{up}(GELU(MLP^i_{tune}(E^i_{(o)} + E^i_{(d)}))). \tag{4.3}$$

Here, the linear layers MLP^i_{tune} are employed to produce task-specific prompt for each adapter, while MLP_{up} serves as a shared up-projection layer to adjust the dimensions of transformer features uniformly across all adapters. The prompt token sequences P^i originate from the $(l+1)$-th adapter block, where i corresponds to the number of transformer layers in the SAM encoder. By incorporating stage-wise adapter blocks at each stage, we effectively leverage Spectrogram Energy information from diverse levels and modalities. The direct integration of prompts into the intermediate features of the base model facilitates the seamless application of our network to pre-trained foundation models for VSOD.

Notably, unlike prompt-tuning approaches such as SAM-Adapter [20] that involve trainable prompt-learning networks and prediction heads, all RGB-modal relevant network parameters in our model, including patch embedding and SAM Encoder Layers, remain frozen.

4.3.4 Circular High-frequency Filter (CHF)

The processing of task-specific information in videos and images, particularly within the RGB stream, often involves leveraging texture or frequency cues. In terms of frequency information, a widely used approach is to extract high-frequency components as described in [19, 37], which capture fine details and serve as complementary signals to the visual content. Specifically, this method involves suppressing low-frequency

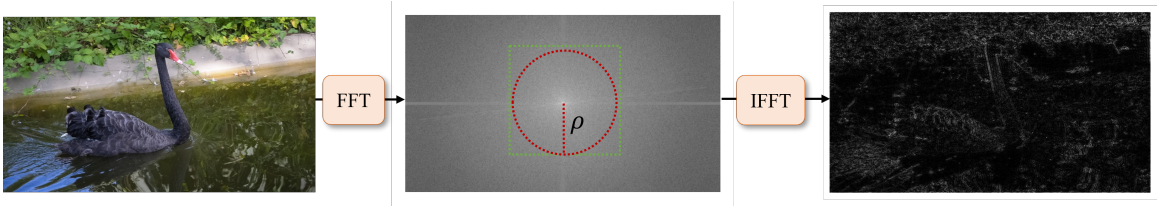

Figure 4.4: An example of using the Circular High-frequency Filter (CHF) to improve image quality. The RGB image is transformed to frequency domain representation by the Fast Fourier Transform (FFT) at first, then high frequency components are filtered out, and the InverseFast Fourier Transform (IFFT) is applied to reconstruct the image. Compared to previous methods that suppress low-frequency components within a rectangular central region (green box), our approach adopts a circular central region (orange circle), which provides better coverage of the low-frequency distribution.

components in the frequency domain using the Fast Fourier Transform (FFT), as recommended in [19]. To obtain high-frequency details, the low-frequency coefficients are first shifted to the center at $(\frac{H}{2}, \frac{W}{2})$; then create a binary mask $M_h \in \{0,1\}^{H\times W}$ and apply it to the frequency component z based on a mask ratio τ:

$$M_h^{i,j}(\tau) = \begin{cases} 0, & \frac{4|(i-\frac{H}{2})(j-\frac{W}{2})|}{HW} \leq \tau \\ 1, & \text{otherwise} \end{cases}. \tag{4.4}$$

Here, H and W denote the height and width of the frame I. The symbol τ denotes the surface ratio of the masked regions, as the *green box* region in Figure 4.4. The Inverse Fast Fourier Transform (IFFT) is used to reconstruct the image from the masked frequency components. We use z to represent the frequency component of I, defining $z = \text{FFT}(I)$ and $I = \text{IFFT}(z)$. The computation of high-frequency components I_{hfc} can be expressed as:

$$I_{hfc} = \text{IFFT}(z\text{M}_h(\tau)). \tag{4.5}$$

However, the 2D frequency spectrum shown in Figure 4.4 reveals that the low frequencies concentrate in the central area, forming an approximately circular pattern. Utilizing an equivalently sized circular region (*orange* in Figure 4.4) to represent low frequencies can provide more precise masking of the low-frequency zone, aligning closely with intuitive perception.

In contrast to conventional fixed hyperparameter approaches [19, 20], we introduce a learnable parameter ρ, to dynamically mask the extraction of high-frequency compo-

nents. The masking process is defined as follows:

$$M_h^{i,j}(\tau) = \begin{cases} 0, & \frac{\sqrt{(i-\frac{H}{2})^2+(j-\frac{W}{2})^2}}{HW} \leq \rho \\ 1, & \text{otherwise} \end{cases}. \tag{4.6}$$

Using learnable parameter ρ allows for adaptive adjustment though backpropagation in the capture of detailed information, thereby enhancing the flexibility and effectiveness of the method in varying datasets and conditions. Our CHF utilizes a circular mask based on Euclidean distance, which more precisely aligns with the radial distribution of low frequencies.

4.3.5 Refinement Fusion Module and Loss

Based on the concepts introduced in previous work [38], our approach incorporates a multiscale strategy to maximize the utilization of high-resolution feature maps in the object decoder [38, 39]. This strategy employs masked attention to accurately distinguish semantics between the foreground and background. Specifically, the Refinement Fusion Module (RFM) uses masked cross-attention to merge dual-scale features through aggregation. As depicted in Figure 4.2, *BConv* is applied to adjust the dual-scale features before implementing masked attention. By leveraging shared regions with rich information and minimal background noise from the dual-scale features, our method effectively utilizes the valuable content in these regions.

To optimize the performance of our MEPNet, we employ a combination of binary cross-entropy (BCE) loss and intersection-over-union (IoU) loss [40]. The total loss function is defined as:

$$l_{total} = \sum_{i=1}^{4} \left(\frac{1}{2^{i-1}} \right) (l_{bce}(S_i, G) + l_{iou}(S_i, G)), \tag{4.7}$$

where S_i represents the output resized to the input size from the i-th decoder layer, and G signifies the ground truth. The terms $l_{bce}(S_i, G)$ and $l_{iou}(S_i, G)$ correspond to the binary cross-entropy (BCE) loss and the intersection-over-union (IoU) loss, respectively. The BCE loss is defined as:

$$l_{bce}(S_i, G) = -\frac{1}{N} \sum_{j=1}^{N} \left[G_j \log(S_{ij}) + (1 - G_j) \log(1 - S_{ij}) \right], \tag{4.8}$$

Table 4.1: Quantitative comparisons with state-of-the-art models on five widely used VSOD datasets. Symbols ↑ and ↓ donate larger and smaller is better, respectively. Symbol '-' means that results are not provided. We use red and blue to indicate the two best scores.

Method	DAVIS [41]			VOS [42]			FBMS [43]			SegV2 [44]			DAVSOD [45]		
	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$
SSAV [45]$_{ICCV'19}$	0.028	0.893	0.861	0.091	0.786	0.704	0.040	0.879	0.865	0.023	0.851	0.798	0.084	0.755	0.659
MGAN [46]$_{ICCV'19}$	0.022	0.913	0.893	0.069	0.807	0.743	0.026	0.912	0.909	0.024	0.895	0.840	0.079	0.757	0.663
PCSA [47]$_{AAAI'20}$	0.022	0.902	0.880	0.065	0.828	0.747	0.040	0.868	0.837	0.024	0.828	0.747	0.086	0.741	0.656
FSNet [48]$_{ICCV'21}$	0.020	0.920	0.907	0.108	0.703	0.659	0.041	0.890	0.888	0.025	0.870	0.806	0.072	0.773	0.685
DCFNet [49]$_{ICCV'21}$	0.016	0.914	0.900	0.061	0.845	0.791	0.039	0.873	0.840	0.015	0.883	0.839	0.074	0.741	0.660
DFMNet [50]$_{ACMM'21}$	0.025	0.898	0.869	-	-	-	0.034	0.889	0.880	-	-	-	0.072	0.774	0.684
UGPL [51]$_{NeurIPS'22}$	0.020	0.910	0.895	0.078	0.764	0.766	0.027	0.900	0.892	0.025	0.860	0.803	0.074	0.749	0.658
MGTNet [52]$_{SPL'22}$	0.015	0.925	0.918	0.062	0.835	0.766	0.033	0.901	0.890	-	-	-	0.064	0.796	0.721
DMPNet [27]$_{TOMM'23}$	0.021	0.905	0.888	-	-	-	0.038	0.894	0.888	0.014	0.893	0.849	0.069	0.746	0.655
CoSTFormer [8]$_{TNNLS'23}$	0.014	0.921	0.903	0.081	0.812	0.793	0.036	0.889	0.885	0.016	**0.904**	0.870	0.061	0.806	0.731
UFGS [53]$_{TMM'23}$	0.013	0.921	0.907	-	-	-	0.033	0.888	0.887	0.012	0.901	0.867	-	-	-
DCTNet+ [13]$_{NeurIPS'24}$	**0.012**	**0.930**	**0.922**	**0.056**	**0.858**	**0.802**	**0.026**	**0.916**	**0.918**	**0.010**	**0.931**	**0.917**	**0.055**	**0.818**	**0.754**
***MEPNet** (Ours)*	**0.011**	**0.931**	**0.920**	**0.045**	**0.860**	**0.803**	**0.024**	**0.915**	**0.923**	**0.009**	**0.931**	**0.913**	**0.050**	**0.808**	**0.736**

where S_{ij} is the predicted probability for pixel j in layer i, G_j is the ground-truth label for pixel j, and N is the total number of pixels. The IoU loss is given by:

$$l_{iou}(S_i, G) = 1 - \frac{\sum_{j=1}^{N}(S_{ij} \cdot G_j)}{\sum_{j=1}^{N}(S_{ij} + G_j - S_{ij} \cdot G_j)}, \tag{4.9}$$

where S_{ij} and G_j are as defined above. By integrating the Refinement Fusion Module with the loss function, we ensure efficient utilization of multiscale features, minimizing discrepancies in pixel-wise and region-wise predictions compared to the ground-truth masks.

4.4 Experiments

In this section, we begin by describing the datasets used for training and evaluation, highlighting their characteristics and relevance to VSOD. Next, we provide implementation details, including training configurations and model setting details. We then present comprehensive quantitative and qualitative analyses to demonstrate the effectiveness of our approach, followed by ablation studies to validate the contribution of each proposed component.

4.4.1 Datasets

We conduct experiments and assess the performance of our proposed model on five benchmark VSOD datasets.

DAVIS2016[41], **YouTube-VOS**[42], **SegV2**[44], and **DAVSOD**[45] are employed in our experiments, as previously detailed in Chapter 2. Additionally, we also utilize **FBMS**[43], which includes 59 video sequences and 720 annotated frames, featuring multiple moving objects, some of which may remain stationary during certain periods.

In addition, we also evaluate our methods on two Video Camouflaged Object detection (VCOD) benchmark datasets.

MoCA-Mask [21] is the largest dataset with pixel-level annotations in the VCOD field, optimized based on the MoCA dataset, comprising a total of 71 sequences for training and 16 sequences for testing.

CAD[22] is a small camouflaged animal testing dataset that consists of 9 short video sequences with 181 hand-labeled masks provided for every 5th frame.

4.4.2 Evaluation Metrics

We utilize three evaluation metrics to evaluate our model's performance in both VSOD and VCOD tasks. We use **Mean Absolute Error** (MAE) [54], **Structural Measure** (S_m) [55] and **mean F-measure** (F_β) [56] to evaluate SOD models. Further details on these metrics can be found in Chapter 2.

Following the conventions of VCOD, we adopt two additional evaluation metrics to assess the performance of our model.

Mean E-measure (E_m), a metric that assesses the alignment between model predictions and true labels by integrating pixel-level sensitivity and image-level errors.

Mean Intersection over Union($mIoU$) measures the overlap and consistency between predicted areas and true areas.

The excellent performance indicated by these metrics (lower MAE, higher S_m, F_β, E_m, and $mIoU$ scores) collectively signifies the model's exceptional performance.

4.4.3 Implementation Details

For each video frame, we generate a synthetic depth map using DPT [57] within our MEPNet. Optical flow maps are computed using RAFT [58]. The trimodal inputs, consisting of RGB, depth and optical flow data, are resized to 448×448 with a batch

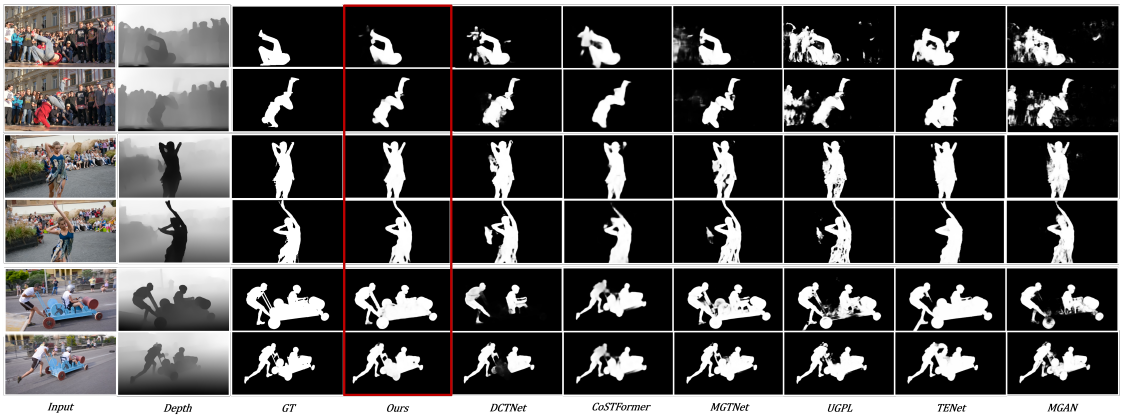

Figure 4.5: Qualitative comparison of our model and SOTA methods on conventional VSOD benchmarks.

size of 8. Each stream employs ResNet-34 [59] to capture multiscale representations. Following [10], we enhance the final layer by integrating the Atrous Spatial Pyramid Pooling (ASPP) module [60]. The multimodal energy prompts are concatenated and fed into the SAM encoder. We use the SGD algorithm for optimization, with initial learning rates set to 1e-4 for the backbones and 1e-3 for other parts. Data augmentation techniques such as random flipping and cropping are applied.

4.4.4 Comparison with State-of-the-arts

To evaluate the effectiveness of our proposed MEPNet, we perform a comparative assessment against 12 SOTA methods on VSOD benchmarks, including SSAV [45], MGAN [46], PCSA [47], FSNet [48], DCFNet [49], DFMNet [50], UGPL [51], MGT-Net [52], DMPNet [27], CoSTFormer [8], UFGS [53], and DCTNet+ [61]. The quantitative results across five VSOD datasets are presented in Table 4.1. The results demonstrate that our method consistently outperforms nearly all competing models, highlighting its effectiveness in VSOD.

For qualitative evaluation, Figure 4.5 presents a visual comparison between our MEPNet and other SOTA models, demonstrating its ability to generate sharper saliency maps with effectively suppressed backgrounds. In scenarios where noise degrades depth information (*Row 1* and *Row 2*), MEPNet effectively reduces background interference, yielding clearer and more distinguishable foreground objects. For complex backgrounds (*Row 3* and *Row 4*), MEPNet accurately separates salient objects from background elements, showcasing superior segmentation precision. Additionally, in dynamic scenes

with fast-moving objects (*Row 5* and *Row 6*), MEPNet captures motion details with high fidelity, maintaining robust saliency detection even under challenging conditions. In general, the visual results highlight the consistent superiority of MEPNet in diverse scenarios, demonstrating its robustness and precision in handling intricate visual challenges.

Table 4.2: Ablation on the architecture on DAVIS 2016 validation set. The proposed Energy prompting strategy performs more effectively.

Method	Trianable Param.	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$
SAM Encoder (w/o prompt)	6.11M	0.241	0.855	0.814
Implicit Prompt	6.24M	0.220	0.872	0.851
Explicit Prompt(hand-crafted)	6.52M	0.017	0.907	0.883
Energy Prompt (w/o optical flow)	6.31M	0.014	0.915	0.901
Energy Prompt (w/o depth)	6.31M	0.016	0.911	0.908
Energy Prompt	6.42M	0.013	0.924	**0.921**
Energy Prompt (w/ CHF)	6.42M	**0.011**	**0.931**	0.920

Table 4.3: Experimental results of Circular High-frequency Filter under different τ on DAVIS 2016 validation set.

	Rectangle Mask		Circular Mask		
	τ=10%	τ=25%	τ=10%	τ=25%	$\tau=\rho$
$MAE\downarrow$	0.0137	0.0126	0.0125	0.0118	**0.0112**
$S_m\uparrow$	0.9251	0.9280	0.9279	0.9301	**0.9313**
$F_\beta\uparrow$	0.9065	0.9189	0.9048	0.9190	**0.9204**

4.4.5 Ablation Study

We conduct thorough ablation studies by removing or replacing components from the full implementation of MEPNet on DAVIS to evaluate the contribution of each key component.

4.4.5.1 Architecture Design

To verify the effectiveness of the proposed visual prompting architecture, we modify it into different variants. As shown in Table 4.2, we use the SAM Encoder without prompting as the baseline. For the implicit prompt, we follow the approach described in [30], aligning the number of prompt tokens with the number of image embed tokens. We also compare implicit prompts with hand-crafted rules as explicit prompts. Although

Table 4.4: Quantitative comparisons with state-of-the-art models on two widely used VCOD datasets. Symbols ↑ and ↓ donate larger and smaller is better, respectively. We use red and blue to indicate the two best scores.

Method	CAD [22]					MoCA-Mask-TE [21]				
	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$E_m\uparrow$	$mIoU\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$E_m\uparrow$	$mIoU\uparrow$
RCRNet [65]$_{ICCV'19}$	0.043	0.627	0.287	0.666	0.229	0.025	0.597	0.174	0.025	0.137
PNS-Net [66]$_{MICC'21}$	0.043	0.678	0.396	0.720	0.308	0.038	0.576	0.134	0.562	0.133
MG [67]$_{ICCV'21}$	0.370	0.484	0.314	0.558	0.260	0.095	0.547	0.165	0.537	0.141
SLT-Net [21]$_{CVPR'22}$	0.028	0.704	0.524	**0.912**	0.438	0.021	0.656	0.357	**0.785**	0.310
IMEX [68]$_{TMM'24}$	0.033	0.684	0.452	0.813	0.370	0.020	0.661	0.371	**0.778**	0.319
TSP-SAM [69]$_{CVPR'24}$	0.031	0.681	0.500	0.853	0.393	**0.012**	0.673	**0.400**	0.766	**0.345**
CQF [70]$_{TCSVT'24}$	**0.022**	**0.732**	**0.556**	0.755	**0.445**	**0.011**	**0.683**	0.388	0.752	0.334
MEPNet (Ours)	**0.021**	**0.752**	**0.608**	**0.855**	**0.491**	**0.011**	**0.677**	**0.415**	0.773	**0.351**

using hand-crafted heuristic cues such as chromaticity, intensity, and texture [62–64] achieve good results, they still could not exceed the performance of using auxiliary modalities as prompts. Our Energy Prompt configurations, whether integrating optical flow or depth individually, outperform existing prompting methods. Notably, the full Energy Prompt setup, which incorporates both optical flow and depth prompts, achieves the best results when combined with CHF, highlighting its effectiveness in enhancing model accuracy.

4.4.5.2 Effectiveness of Spectrogram Energy

Using Spectrogram Energy extracted from the optical flow and depth information as prompts for the SAM Encoder significantly enhances the model's performance. To validate the efficacy of our Spectrogram Energy prompt, we introduce Energy Prompt tokens for each auxiliary modality individually. As demonstrated in Table 4.2, utilizing the Spectrogram Energy from either optical flow or depth alone enhances the model's performance. However, the fusion of motion energy from both optical flow and depth results in even more accurate energy prompts, effectively guiding the SAM Encoder. This synergy leverages depth's provision of relative positional information, which compensates for interference caused by the object's motion relative to the camera. The MEPNet model demonstrates a favorable trade-off between precision and efficiency, demonstrating notable improvements in the metrics MAE, S_m, and F_β.

4.4.5.3 Effectiveness of High-frequency Filter

We further evaluate the different mask shapes and the hyper-parameter mask ratio τ introduced in Section 4.3.4. From Table 4.3, it is evident that when masking out

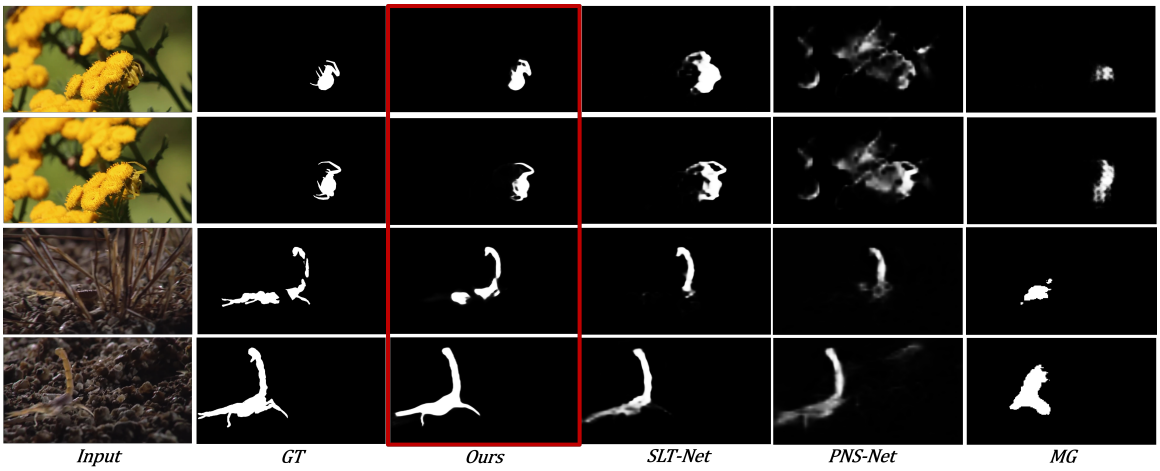

Figure 4.6: Qualitative comparison of our model and SOTA methods on VCOD benchmarks.

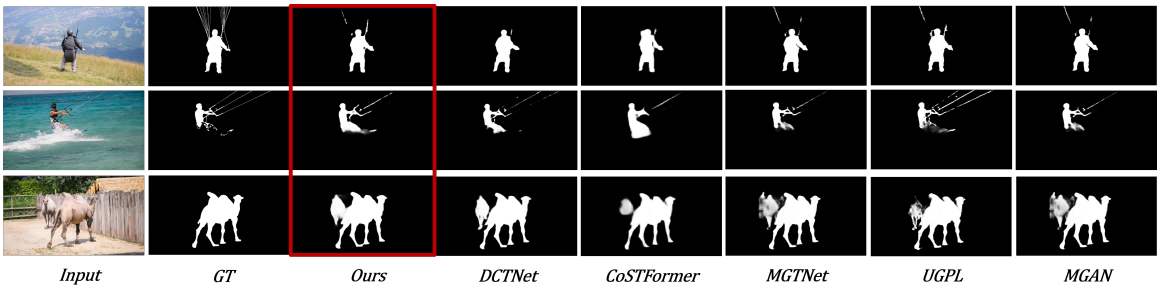

Figure 4.7: Failure cases.

10% and 25% of the central pixels in the spectrum, the circular mask consistently performs better than the rectangular mask. Specifically, the circular mask with a learnable parameter $\tau = \rho$ yields the best results in all metrics, with the lowest MAE and the highest S_m and F_β values. This indicates that the shape of the mask and the adaptability of the mask area play a crucial role in the effectiveness of the High-frequency Filter. Our experiments show that setting τ manually does not achieve optimal results. Instead, using learnable parameter θ, the model can adaptively select the mask area to achieve the best performance. This adaptability allows the model to fine-tune the high-frequency components more precisely, resulting in improved performance of saliency metrics.

4.4.5.4 Generalizability

Our MEPNet model demonstrates superior performance on VSOD. Meanwhile, we conduct extension experiments in the VCOD task to evaluate the generalization of our model, shown in Table 4.4. We compare our method with seven current state-of-the-art VCOD methods: CQF [70],TSP-SAM [69],IMEX [68],SLT-Net [21], MG [71], PNS-Net [66] and RCRNet [65]. Additionally, Figure 4.6 presents a qualitative comparison of our model against SOTA methods on VCOD benchmarks. In particular, our method achieves significant improvements not only in capturing salient objects but also in effectively detecting camouflaged targets in complex scenes. These results confirm its strong generalization ability across diverse visual conditions, demonstrating robustness in both highly distinguishable and subtle object regions.

4.4.6 Limitations

Figure 4.7 illustrates several representative cases in which MEPNet encounters challenges in VSOD. In particular, our model occasionally misidentifies non-salient foreground objects as salient due to inherent ambiguities in complex video scenes. This issue primarily arises from the limitations of existing ground truth (GT) annotations, which may not always align with human perception, leading to inconsistencies during training and evaluation.

In addition, structural limitations in saliency detection emerge when dealing with thin and elongated structures. For example, in *Rows 1* and *2*, the parachute control lines are too narrow for the model to accurately preserve. Motion-induced noise further

contributes to segmentation challenges. In *Row 2*, water splashes created by the surfer introduce high-frequency interference, causing the model to over-segment the surrounding area. Similarly, in *Row 3*, the presence of multiple salient animals leads to confusion in assigning primary and secondary salient objects. It should be noted that such limitations are not unique to MEPNet; other state-of-the-art VSOD models, including DCTNet [61], CoSTFormer [8], MGTNet [52], UGPL [51], and MGAN [46], also struggle with these cases. Future work will focus on refining prompt-based learning strategies to enhance robustness in complex environments and exploring a unified framework that extends beyond VSOD to VCOD.

4.5 Conclusion

In this paper, we proposed MEPNet, a novel multimodal energy prompting network for Video Salient Object Detection (VSOD). To address the limitations of directly using optical flow and depth as prompts due to noise interference, we introduced the Spectrogram Energy Generator (SEG), which extracts implicit prompts by modeling Spectrogram Energy, enhancing spatial-temporal representation learning. Additionally, the Modality-Energy Adapter (MEA) effectively integrates these prompts into a pre-trained SAM encoder, while the Circular High-frequency Filter (CHF) refines RGB feature extraction, capturing fine details for more robust detection. Experimental results clearly demonstrate that our network significantly outperforms existing methods, highlighting the efficacy of Spectrogram Energy prompting techniques in enhancing VSOD and VSOD accuracy.

In future work, we aim to expand our research beyond VSOD to explore Video Camouflaged Object Detection (VCOD), which presents additional challenges due to the subtle and deceptive nature of camouflaged objects. By investigating the shared characteristics between VSOD and VCOD, we aspire to develop a unified framework capable of effectively detecting both salient and camouflaged objects in videos. Such a model would enhance adaptability across different video analysis tasks and contribute to a more generalized understanding of object detection in dynamic and complex environments.

References

[1] Hadi Hadizadeh and Ivan V Bajić. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2013.

[2] Georgios Evangelopoulos, Athanasia Zlatintsi, Georgios Skoumas, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, and Yannis Avrithis. Video event detection and summarization using audio, visual and text saliency. In *2009International Conference on Acoustics, Speech, and Signal Processing*, pages 3553–3556. IEEE, 2009.

[3] Ludovic Simon, Jean-Philippe Tarel, and Roland Brémond. Alerting the drivers about road signs with poor visual saliency. In *2009 IEEE Intelligent Vehicles Symposium*, pages 48–53. IEEE, 2009.

[4] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *International conference on machine learning*, pages 597–606. PMLR, 2015.

[5] Tao Jiang, Yi Wang, Feng Hou, and Ruili Wang. IENet: inheritance enhancement network for video salient object detection. *Multimedia Tools and Applications*, pages 1–20, 2024.

[6] Runmin Cong, Weiyu Song, Jianjun Lei, Guanghui Yue, Yao Zhao, and Sam Kwong. Parallel symmetric network for video salient object detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.

[7] Chenglizhao Chen, Jia Song, Chong Peng, Guodong Wang, and Yuming Fang. A novel video salient object detection method via semisupervised motion quality perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2732–2745, 2021.

[8] Nian Liu, Kepan Nan, Wangbo Zhao, Xiwen Yao, and Junwei Han. Learning complementary spatial–temporal transformer for video salient object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[9] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16835, 2021.

[10] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7274–7283, 2019.

[11] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. RGBD salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017.

[12] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 646–662. Springer, 2020.

[13] Jingjing Li, Wei Ji, Size Wang, Wenbo Li, et al. DVSOD: RGB-D video salient object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Junhao Lin, Lei Zhu, Jiaxing Shen, Huazhu Fu, Qing Zhang, and Liansheng Wang. ViDSOD-100: A New Dataset and a Baseline Model for RGB-D Video Salient Object Detection. *International Journal of Computer Vision*, pages 1–19, 2024.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[16] Qi Dai, Vaishakh Patil, Simon Hecker, Dengxin Dai, Luc Van Gool, and Konrad Schindler. Self-supervised object motion and depth estimation from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1004–1005, 2020.

[17] Zhaojie Luo, Shoufeng Lin, Rui Liu, Jun Baba, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Decoupling Speaker-Independent Emotions for Voice Conversion via Source-Filter Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:11–24, 2023.

[18] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR, 2020.

[19] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19434–19445, 2023.

[20] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. SAM fails to segment anything?–SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, medical

image segmentation, and more. *arXiv preprint arXiv:2304.09148*, 2023.

[21] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13873, 2022.

[22] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 433–449. Springer, 2016.

[23] Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4259–4267, 2021.

[24] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3243–3252, 2018.

[25] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

[26] Mingzhu Xu, Ping Fu, Bing Liu, and Junbao Li. Multi-stream attention-aware graph convolution network for video salient object detection. *IEEE Transactions on Image Processing*, 30:4183–4197, 2021.

[27] Baian Chen, Zhilei Chen, Xiaowei Hu, Jun Xu, Haoran Xie, Jing Qin, and Mingqiang Wei. Dynamic message propagation network for RGB-D and video salient object detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(1):1–21, 2023.

[28] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[29] Guyue Hu, Bin He, and Hanwang Zhang. Compositional prompting video-language models to understand procedure in instructional videos. *Machine Intelligence Research*, 20(2):249–262, 2023.

[30] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[31] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[32] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Conference*, pages 9516–9526, 2023.

[33] Jinq Horng Teo, Shuai Cheng, and Massimo Alioto. Low-energy voice activity detection via energy-quality scaling from data conversion to machine learning. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67(4):1378–1388, 2020.

[34] Selma Özaydın. Examination of energy based voice activity detection algorithms for noisy speech signals. *Avrupa Bilim ve Teknoloji Dergisi*, pages 157–163, 2019.

[35] Mingzhu Xu, Bing Liu, Ping Fu, Junbao Li, Yu Hen Hu, and Shou Feng. Video salient object detection via robust seeds extraction and multi-graphs manifold propagation. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2191–2206, 2019.

[36] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. SCOM: Spatiotemporal constrained optimization for salient object detection. *IEEE Transactions on Image Processing*, 27(7):3345–3357, 2018.

[37] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022.

[38] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Conference*, pages 1290–1299, 2022.

[39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[40] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016.

[41] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus

Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[42] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2017.

[43] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, 2013.

[44] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[45] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting More Attention to Video Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8546–8556, 2019.

[46] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021.

[47] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10869–10876, 2020.

[48] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4922–4933, 2021.

[49] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1553–1563, 2021.

[50] Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun Zhao. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In *Proceedings of the29th ACM International Conference on Multimedia*, pages 731–740, 2021.

[51] Yongri Piao, Chenyang Lu, Miao Zhang, and Huchuan Lu. Semi-supervised video

salient object detection based on uncertainty-guided pseudo labels. *Advances in Neural Information Processing Systems*, 35:5614–5627, 2022.

[52] Dingyao Min, Chao Zhang, Yukang Lu, Keren Fu, and Qijun Zhao. Mutual-guidance transformer-embedding network for video salient object detection. *IEEE Signal Processing Letters*, 29:1674–1678, 2022.

[53] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, 2023.

[54] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740. IEEE, 2012.

[55] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017.

[56] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE, 2009.

[57] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[58] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow . In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[60] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[61] Yukang Lu, Dingyao Min, Keren Fu, and Qijun Zhao. Depth-cooperated trimodal network for video salient object detection. In *2022 IEEE International Conference*

on Image Processing (ICIP), pages 116–120. IEEE, 2022.

[62] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *2011 International Conference on Computer Vision*, pages 898–905. IEEE, 2011.

[63] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision*, 98:123–145, 2012.

[64] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 223–230. IEEE, 2010.

[65] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7284–7293, 2019.

[66] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 142–152. Springer, 2021.

[67] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188, 2021.

[68] Wenjun Hui, Zhenfeng Zhu, Guanghua Gu, Meiqin Liu, and Yao Zhao. Implicit-explicit Motion Learning for Video Camouflaged Object Detection. *IEEE Transactions on Multimedia*, 2024.

[69] Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, and Yao Zhao. Endow SAM with Keen Eyes: Temporal-spatial Prompt Learning for Video Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19058–19067, 2024.

[70] Zelin Lu, Liang Xie, Xing Zhao, Binwei Xu, Haoran Liang, and Ronghua Liang. A Weakly-supervised Cross-domain Query Framework for Video Camouflage Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[71] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie.

Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188, 2021.

Chapter 5

Summary

This chapter provides concluding remarks on the thesis, synthesizing our contributions to Video Salient Object Detection. In Section 5.1, we review the methods proposed throughout our research, highlighting their key innovations and effectiveness. Subsequently, in Section 5.2, we explore potential avenues for future research, outlining how these methods can be further refined and extended to broader contexts, thereby offering guidance for subsequent work in the field.

5.1 Research Summary

In this thesis, we have investigated Video Salient Object Detection (VSOD) from multiple perspectives, encompassing fully supervised, scribble-supervised, and multi-modal approaches. Our overarching goal is to enhance the reliability and accuracy of salient object detection in dynamic video contexts. The core methodologies and key contributions presented in this work are summarized as follows:

Chapter 2 introduces an Inheritance Enhancement Network (IENet) to address the inadequate utilization of historical frame information in VSOD. IENet incorporates a Heritable Multi-Frame Attention (HMA) module that employs unidirectional cross-frame enhancement to fully leverage long-term context and frame-aware temporal modeling. This heritable strategy ensures consistent, orderly propagation of attention maps across frames, minimizing interference. An auxiliary attention loss is further incorporated to guide the network in focusing on target regions based on inherited

attention maps. The results of the experiments on five benchmark datasets validate the effectiveness of IENet in retaining finer details from historical frames, thereby enhancing accuracy and robustness in complex scenarios.

Chapter 3 introduces a Knowledge-sharing Hierarchical Memory Fusion Network (KHMF-Net) designed to mitigate the limitations of scribble annotations in VSOD. By incorporating a Hierarchical Memory Bank (HMB) to archive historical segmentation at multiple confidence levels, KHMF-Net effectively reduces error accumulation and promotes salient object expansion. To further distinguish background from target objects, the Adaptive Memory Fusion (AMF) module is coupled with an Interactive Equalized Matching (IEM) strategy and a dual-attention knowledge-sharing mechanism, improving both accuracy and robustness.

Chapter 4 proposes a Multimodal Energy Prompt Network (MEPNet) to extend VSOD beyond purely spatio-temporal cues by capitalizing on optical flow and depth information. Guided by learned knowledge from the Segment Anything Model (SAM), MEPNet employs a Spectrogram Energy Generator (SEG) to dynamically prompt the SAM encoder, facilitating more efficient modeling of video content. In addition, a Circular High-frequency Filter (CHF) is introduced to adaptively enhance RGB modality details while avoiding constraints associated with hand-crafted design. Experimental evaluations on five benchmark datasets demonstrate that MEPNet outperforms existing approaches, affirming the efficacy of combining multimodality prompts with advanced filtering techniques. Furthermore, the model exhibits robust generalization across diverse scenarios, surpassing state-of-the-art methods in this domain.

To sum up, this thesis explores innovative strategies to enhance VSOD by addressing critical challenges under fully supervised, scribble-supervised, and multimodal contexts. Our approaches have improved accuracy across six public VSOD datasets significantly. Evaluations on multiple benchmark datasets confirm the efficacy of these methods. Our works have been published in top-tier conferences and journals (*e.g., ACM Multimedia, Pattern Recognition Letter, and ACM International Conference on Multimedia in Asia*), advancing the field of VSOD with practical and effective solutions.

5.2 Future work and directions

In this section, we discuss two directions of potential research: self-supervised learning, as detailed in Section 5.1, and Video Camouflaged Object Detection, discussed in Section 5.2.

5.2.1 Self-supervised learning

In our research, we employed both fully supervised and scribble-supervised approaches for action recognition [1, 2]. However, the reliance on extensively annotated datasets remains challenging, motivating the exploration of self-supervised learning methods [3, 4]. By leveraging techniques such as the Segment Anything Model (SAM) [5] in pretraining, we can extract meaningful representations from unlabeled data. Meanwhile, exploiting temporal and spatial structures within videos further enhances the robustness of VSOD systems [1]. Moreover, integrating contrastive learning and predictive modeling facilitates the discovery of discriminative features, paving the way for more efficient and scalable VSOD frameworks in real-world scenarios.

5.2.2 Video Camouflaged Object Detection

In addition, we will investigate Video Camouflaged Object Detection (VCOD), which builds on image-based Camouflaged Object Detection (COD) by leveraging multi-frame inputs and temporal cues. While VSOD targets visually prominent objects, VCOD focuses on deliberately hidden ones, making it more challenging by requiring both static and dynamic temporal information [6, 7]. To address these challenges, we will explore multi-scale feature alignment [8] and the Spatial-Mamba Block [9], thereby enhancing spatial dependency modeling and improving detection performance. By integrating these modules, we aim to boost accuracy and efficiency even under occlusion, complex backgrounds, and subtle motion cues, ultimately advancing the field of camouflaged object detection in videos.

References

[1] Tao Jiang, Yi Wang, Feng Hou, and Ruili Wang. IENet: inheritance enhancement network for video salient object detection. *Multimedia Tools and Applications*, pages 1–20, 2024.

[2] Tao Jiang, Feng Hou, and Yi Wang. Multimodal Energy Prompting for Video Salient Object Detection. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–8, 2024.

[3] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188, 2021.

[4] Yi-Wen Chen, Xiaojie Jin, Xiaohui Shen, and Ming-Hsuan Yang. Video salient object detection via contrastive features and attention modules. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1320–1329, 2022.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[6] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[7] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13873, 2022.

[8] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom-NeXt: A Unified Collaborative Pyramid Network for Camouflaged Object Detection . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[9] Rui Xu, Shu Yang, Yihui Wang, Bo Du, and Hao Chen. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*, 2024.