Visual Intelligence

RESEARCH Open Access

SAM2-UNet: segment anything 2 makes strong encoder for natural and medical image segmentation

Xinyu Xiong[1], Zihuang Wu[2], Shuangyi Tan[3], Wenxue Li[4], Feilong Tang[5], Ying Chen[6], Siying Li[7], Jie Ma[1] and Guanbin Li[1*]

Abstract

Image segmentation plays an important role in vision understanding. Recently, the emerging vision foundation models continuously achieved superior performance on various tasks. Following such success, in this paper, we prove that the Segment Anything Model 2 (SAM2) can be a strong encoder for U-shaped segmentation models. We propose a simple but effective framework, termed SAM2-UNet, for versatile image segmentation. Specifically, SAM2-UNet adopts the Hiera backbone of SAM2 as the encoder, while the decoder uses the classic U-shaped design. Additionally, adapters are inserted into the encoder to enable parameter-efficient fine-tuning. Preliminary experiments on various downstream tasks, such as camouflaged object detection, salient object detection, marine animal segmentation, mirror detection, and polyp segmentation, demonstrate that our SAM2-UNet can outperform existing specialized state-of-the-art methods with minimal additional complexity.

Keywords: Image segmentation, Segment anything model, U-Net, Vision foundation model

1 Introduction

Image and video segmentation [1–4] is a crucial task in the field of computer vision, serving as the foundation for various visual understanding applications. By dividing an image into meaningful regions based on specific semantic criteria, image segmentation enables a wide array of downstream tasks in both natural and medical domains, such as camouflaged object detection [5, 6], salient object detection [7, 8], marine animal segmentation [9, 10], mirror detection [11, 12], and polyp segmentation [13, 14]. Many specialized architectures have been proposed to achieve superior performance on these different tasks, while it remains an open challenge to design a unified architecture to address the diverse segmentation tasks.

The emergence of vision foundation models (VFMs) [15–18] has introduced significant potential in the field of image segmentation. Among these VFMs, a notable example is the Segment Anything Model (SAM1) [15] and its successor, Segment Anything Model 2 (SAM2) [16]. SAM2 builds upon the foundation laid by SAM1, utilizing a larger dataset for training and incorporating improvements in architectural design. However, despite these advancements, SAM2 still produces class-agnostic segmentation results when no manual prompt is provided. This limitation highlights the ongoing challenge of effectively transferring SAM2 to downstream tasks, where task-specific or class-specific segmentation is often required. Exploring strategies to enhance SAM2's adaptability and performance in these scenarios remains an important area of research.

To adapt SAM to downstream tasks, several approaches have been proposed, including the use of adapters [19, 20] for parameter-efficient fine-tuning and the integration of

*Correspondence: liguanbin@mail.sysu.edu.cn
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
Full list of author information is available at the end of the article

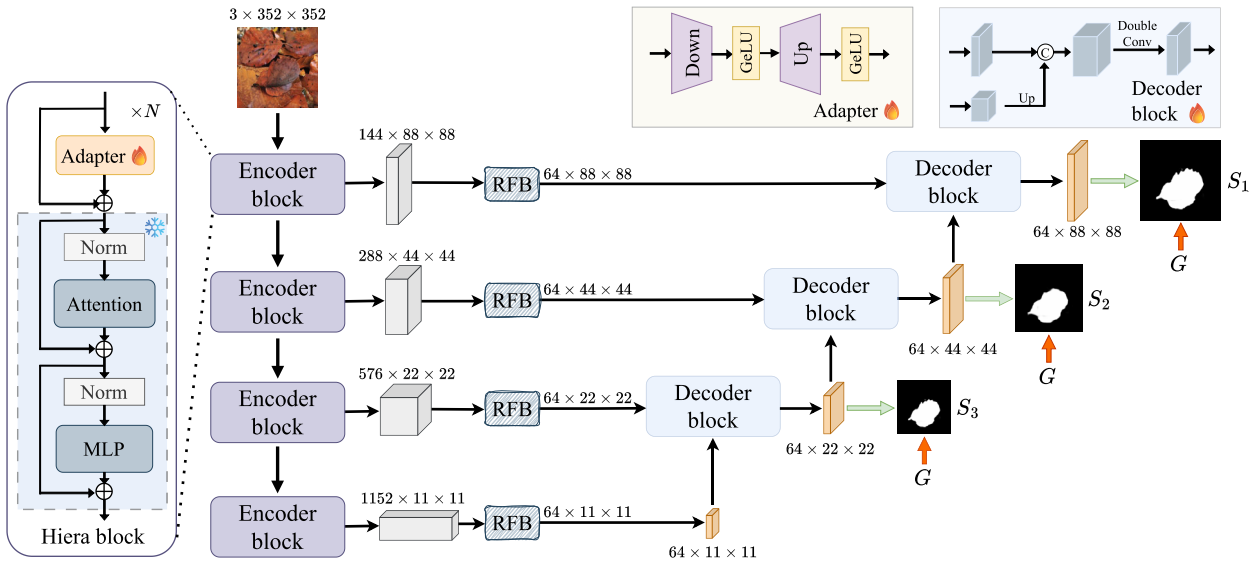

Figure 1 Overview of the proposed SAM2-UNet. Note that there are some variants of the Hiera block, and we only demonstrate a simplified structure for ease of understanding. MLP: multilayer perceptron; RFB: receptive field block

additional conditional inputs such as text prompts [21–23] or in-context samples [24, 25]. Inspired by the strong segmentation capabilities of U-Net [26] and its variants [27–29], some researchers have explored the possibility of transforming SAM into a U-shaped architecture [30, 31]. However, these efforts have often been limited by the plain structure of the vanilla ViT encoder [32], which lacks the hierarchy needed for more sophisticated segmentation tasks. Fortunately, the introduction of SAM2, which features a hierarchical backbone, opens new avenues for designing a U-shaped network with improved effectiveness.

In this paper, we propose SAM2-UNet, the benefit of which is summarized as follows:

1) Simplicity. SAM2-UNet adopts a classic U-shaped encoder-decoder architecture, known for its ease of use and high extensibility.

2) Efficiency. Adapters are integrated into the encoder to enable parameter-efficient fine-tuning, allowing the model to be trained even on memory-limited devices.

3) Effectiveness. Extensive experiments on eighteen public datasets demonstrate that SAM2-UNet delivers powerful performance across five challenging benchmarks.

2 Related works

2.1 Foundation segmentation model

In the past two years, the deep learning community has witnessed the emergence of large-scale pixel-level foundation models, with the Segment Anything Model (SAM) [6, 15, 16] family standing out as one of the most prominent examples. While these foundation models demonstrate remarkable zero-shot performance across a wide range of scenarios, deploying them for specific applications still requires significant effort, such as efficient adaptation [19, 20], model slimming [33], and multi-modal knowledge injection [34]. This paper focuses on parameter-efficient fine-tuning [35], which freezes the original model parameters and introduces learnable components to efficiently adapt the foundation models to downstream tasks.

2.2 Single-class image segmentation

Unlike semantic segmentation, which typically considers identifying multiple classes in the image, single-class image segmentation is dedicated to finding a single foreground class of interest [8, 9, 11, 14, 36, 37] in the image. For example, FEDER [8] leveraged learnable wavelets to decompose features into distinct frequency bands, offering a new perspective in feature representation for the camouflaged object detection task. MENet [36] incorporated biologically inspired mechanisms from the human visual system, including boundary sensitivity, content integrity, iterative refinement, and frequency decomposition, to enhance the detection of salient objects. MASNet [9] leveraged a novel data augmentation paradigm that systematically manipulates degradation patterns and camouflage characteristics of marine organisms. HetNet [11] proposed a hierarchical architecture that initially identifies potential mirror regions through low-level feature analysis and subsequently refines predictions by integrating high-level semantic features. CFA-Net [14] simultaneously exploited boundary information and captured hierarchical semantic details to achieve accurate polyp segmentation. Unlike the aforementioned existing practices,

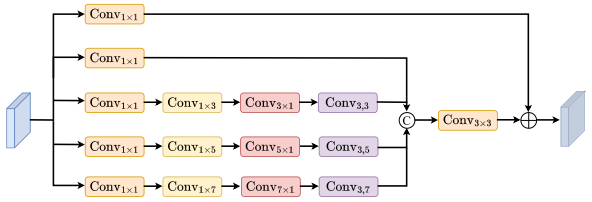

Figure 2 Overview of the receptive field block. For simplicity, normalization and activation operations are omitted in the figure

this paper shifts the focus of network design to the encoder, leveraging the powerful representation capabilities of foundational segmentation models to achieve strong performance even with a simple U-shaped decoder.

3 Method

The overall architecture of SAM2-UNet is illustrated in Fig. 1, comprising four main components: encoder, decoder, receptive field block (RFB), and adapter. Note that we discard components that are not essential for constructing a basic U-Net [26], such as memory attention, prompt encoder, memory encoder, and memory bank.

3.1 Encoder

SAM2-UNet applies the Hiera [38] backbone pretrained by SAM2. Compared with the plain ViT [32] encoder used in SAM1 [15], Hiera uses a hierarchical structure that allows multiscale feature capturing, which is more suitable for designing a U-shaped network. Specifically, given an input image $I \in \mathbb{R}^{3\times H\times W}$, where H denotes height and W denotes width, Hiera will output four hierarchical features $X_i \in \mathbb{R}^{C_i\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}} (i \in \{1,2,3,4\})$. For Hiera-L, $C_i \in \{144, 288, 576, 1152\}$.

3.2 Receptive field block

The structure of RFB is illustrated in Fig. 2, which is composed of multiple branches incorporating kernels and dilated convolution layers of varying sizes. After extracting the encoder features, we pass them through four receptive field blocks [13, 39] to reduce the channel number to 64 as well as enhance these lightweight features.

3.3 Adapter

Since the Hiera parameters may be enormous (212 million for Hiera-L), full fine-tuning may not always be feasible in terms of memory. Therefore, we freeze the parameters of Hiera and insert adapters before each multi-scale block of Hiera to achieve parameter-efficient fine-tuning. Similar to the adapter design in Refs. [40, 41], each adapter in our framework consists of a linear layer for downsampling, a GELU activation function, followed by another linear layer for upsampling, and a final GELU activation.

Table 1 Detailed information of datasets for different tasks. For camouflaged object detection, the training set is obtained by combining CAMO [42] and COD10K [43]. For polyp segmentation, the training set is obtained by combining Kvasir-SEG [44] and CVC-ClinicDB [45]

Tasks	Dataset	Training set	Test set
Camouflaged object detection	CAMO [42]	1000	250
	COD10K [43]	3040	2026
	CHAMELEON [46]	-	76
	NC4K [47]	-	4121
Salient object detection	DUTS [48]	10553	5019
	DUT-OMRON [49]	-	5168
	HKU-IS [50]	-	4447
	PASCAL-S [51]	-	850
	ECSSD [52]	-	1000
Marine animal segmentation	MAS3K [10]	1769	1141
	RMAS [9]	2514	500
Mirror detection	MSD [53]	3063	955
	PMD [54]	5096	571
Polyp segmentation	Kvasir-SEG [44]	900	100
	CVC-ClinicDB [45]	550	62
	CVC-ColonDB [55]	-	380
	CVC-300 [56]	-	60
	ETIS [57]	-	196

3.4 Decoder

The original mask decoder in SAM2 uses a two-way transformer approach to facilitate feature interaction between the prompt embedding and encoder features. In contrast, inspired by the highly customizable U-shaped structure that has proven effective in many tasks [27–29], our decoder also adheres to the classic U-Net design. It consists of three decoder blocks, each containing two "Conv-BN-ReLU" combinations, where "Conv" denotes a 3×3 convolution layer and "BN" represents batch normalization. The output feature from each decoder block passes through a 1×1 Conv segmentation head to produce a segmentation result S_i $(i \in 1,2,3)$, which is then upsampled and supervised by the ground truth mask G. Among these output results, S_1 is the final output, while S_2 and S_3 are used only during the training phase for auxiliary supervision.

3.5 Loss function

Following the approaches in Refs. [7, 13], we use the weighted IoU loss and binary cross-entropy (BCE) loss as our training objectives: $\mathcal{L} = \mathcal{L}^{w}_{IoU} + \mathcal{L}^{w}_{BCE}$. Additionally, we apply deep supervision to all segmentation outputs S_i. The total loss for SAM2-UNet is formulated as $\mathcal{L}_{total} = \sum_{i=1}^{3} \mathcal{L}(G, S_i)$.

4 Experiments

4.1 Datasets and benchmarks

Our experiments are conducted on five different benchmarks with 18 datasets in total, as shown in Table 1. All these datasets use publicly available train-test splits to ensure fair comparison.

Table 2 Comparison of results on camouflaged object detection. S_α: *S*-measure; F_β: adaptive *F*-measure; E_ϕ: mean *E*-measure; MAE: mean absolute error

Methods	CHAMELEON [46]				CAMO [42]				COD10K [43]				NC4K [47]			
	S_α	F_β	E_ϕ	MAE	S_α	F_β	E_ϕ	MAE	S_α	F_β	E_ϕ	MAE	S_α	F_β	E_ϕ	MAE
SINet [43]	0.872	0.823	0.936	0.034	0.745	0.712	0.804	0.092	0.776	0.667	0.864	0.043	0.808	0.768	0.871	0.058
PFNet [61]	0.882	0.820	0.931	0.033	0.782	0.751	0.841	0.085	0.800	0.676	0.877	0.040	0.829	0.779	0.887	0.053
ZoomNet [62]	0.902	0.858	0.943	0.024	0.820	0.792	0.877	0.066	0.838	0.740	0.888	0.029	0.853	0.814	0.896	0.043
FEDER [8]	0.903	0.856	0.947	0.026	0.836	0.807	0.897	0.066	0.844	0.748	0.911	0.029	0.862	0.824	0.913	0.042
SAM2-UNet	**0.914**	**0.863**	**0.961**	**0.022**	**0.884**	**0.861**	**0.932**	**0.042**	**0.880**	**0.789**	**0.936**	**0.021**	**0.901**	**0.863**	**0.941**	**0.029**

Table 3 Comparison of results on salient object detection

Methods	DUTS-TE [48]			DUT-OMRON [49]			HKU-IS [50]			PASCAL-S [51]			ECSSD [52]		
	S_α	E_ϕ	MAE	S_α	E_ϕ	MAE	S_α	E_ϕ	MAE	S_α	E_ϕ	MAE	S_α	E_ϕ	MAE
U2Net [63]	0.874	0.884	0.044	0.847	0.872	0.054	0.916	0.948	0.031	0.844	0.850	0.074	0.928	0.925	0.033
ICON [64]	0.889	0.914	0.037	0.845	0.879	0.057	0.920	0.959	0.029	0.861	0.893	0.064	0.929	0.954	0.032
EDN [65]	0.892	0.925	0.035	0.850	0.877	0.049	0.924	0.955	0.026	0.865	0.902	0.062	0.927	0.951	0.032
MENet [36]	0.905	0.937	0.028	0.850	0.891	0.045	0.927	0.966	0.023	0.872	0.913	0.054	0.928	0.954	0.031
SAM2-UNet	**0.934**	**0.959**	**0.020**	**0.884**	**0.912**	**0.039**	**0.941**	**0.971**	**0.019**	**0.894**	**0.931**	**0.043**	**0.950**	**0.970**	**0.020**

1) Camouflaged object detection aims to detect objects well hidden in the environment. We adopt four datasets for benchmarking, including CAMO [42], COD10K [43], CHAMELEON [46], and NC4K [47]. Four metrics are used for comparison, including *S*-measure (S_α) [58], adaptive *F*-measure (F_β) [59], mean *E*-measure (E_ϕ) [60], and mean absolute error (MAE).

2) Salient object detection aims to mimic human cognition mechanisms to identify salient objects. We adopt five datasets for benchmarking, including DUTS [48], DUT-OMRON [49], HKU-IS [50], PASCAL-S [51], and ECSSD [52]. Three metrics are used for comparison, including S_α [58], E_ϕ [60], and MAE.

3) Marine animal segmentation focuses on exploring underwater environments to find marine animals. We adopt two datasets for benchmarking, including MAS3K [10] and RMAS [9]. Five metrics are used for comparison, including mean IoU (mIoU), S_α [58], weighted *F*-measure (F_β^w) [59], E_ϕ [60], and MAE.

4) Mirror detection can identify the mirror regions in the given input image. We adopt two datasets for benchmarking, including MSD [53] and PMD [54]. Three metrics are used for comparison, including IoU, *F*-measure (F_m) [59], and MAE.

5) Polyp segmentation helps in the diagnosis of colorectal cancer. We adopt five datasets for benchmarking, including Kvasir-SEG [44], CVC-ClincDB [45], CVC-ColonDB [55], CVC-300 [56], and ETIS [57]. Two metrics are used for comparison, including mean Dice (mDice) and mIoU.

4.2 Implementation details

Our method is implemented using PyTorch and trained on a single NVIDIA RTX 4090 GPU with 24 GB of memory. We use the AdamW optimizer with an initial learning rate of 0.001, applying cosine decay to stabilize training. Two data augmentation strategies are employed: random vertical and horizontal flips. Unless otherwise specified, we use the Hiera-L version of SAM2. All input images are resized to 352 × 352, with a batch size of 12. The bottleneck channel dimension of the adapter is set to 32. The training epoch is set to 50 for camouflaged object detection and salient object detection, and to 20 for marine animal segmentation, mirror detection, and polyp segmentation. For polyp segmentation, we also adopt a multi-scale training strategy {1, 1.25} similar to Ref. [13].

4.3 Comparison with state-of-the-art methods

In this subsection, we first analyze the quantitative results across different benchmarks, followed by visual comparisons in camouflaged object detection and polyp segmentation.

4.3.1 Results on camouflaged object detection

The results are shown in Table 2. SAM2-UNet outperforms all other methods across all four benchmark datasets, achieving the highest scores in every metric. Specifically, in terms of *S*-measure, SAM2-UNet surpasses FEDER by 1.1% on the CHAMELEON dataset and by 4.8% on the CAMO dataset. On the more challenging COD10K and NC4K datasets, which have larger image counts and higher segmentation difficulty, SAM2-UNet still exceeds the performance of FEDER by 3.6% and 3.9% in *S*-measure, respectively.

4.3.2 Results on salient object detection

The results are shown in Table 3. SAM2-UNet consistently achieves the top results across all metrics. For *S*-

Table 4 Comparison of results on marine animal segmentation. mIoU: mean IoU; F_β^w: weighted F-measure

Methods	MAS3K [10]					RMAS [9]				
	mIoU	S_α	F_β^w	E_ϕ	MAE	mIoU	S_α	F_β^w	E_ϕ	MAE
C2FNet [5]	0.717	0.851	0.761	0.894	0.038	0.721	0.858	0.788	0.923	0.026
OCENet [66]	0.667	0.824	0.703	0.868	0.052	0.680	0.836	0.752	0.900	0.030
ZoomNet [62]	0.736	0.862	0.780	0.898	0.032	0.728	0.855	0.795	0.915	**0.022**
MASNet [9]	0.742	0.864	0.788	0.906	0.032	0.731	0.862	0.801	0.920	0.024
SAM2-UNet	**0.799**	**0.903**	**0.848**	**0.943**	**0.021**	**0.738**	**0.874**	**0.810**	**0.944**	**0.022**

Table 5 Comparison of results on mirror detection. F_m: F-measure

Methods	MSD [53]			PMD [54]		
	IoU	F_m	MAE	IoU	F_m	MAE
MirrorNet [53]	0.790	0.857	0.065	0.585	0.741	0.043
PMD [54]	0.815	0.892	0.047	0.660	0.794	0.032
SANet [12]	0.798	0.877	0.054	0.668	0.795	0.032
HetNet [11]	0.828	0.906	0.043	0.690	0.814	0.029
SAM2-UNet	**0.918**	**0.957**	**0.022**	**0.728**	**0.826**	**0.027**

measure, SAM2-UNet outperforms MENet by 2.9%, 3.4%, 1.4%, 2.2%, and 2.2% on the DUTS-TE, DUT-OMRON, HKU-IS, PASCAL-S, and ECSSD datasets, respectively.

4.3.3 Results on marine animal segmentation

The results are shown in Table 4. Once again, SAM2-UNet achieves the best performance across all metrics on the two benchmark datasets. Specifically, for mIoU, SAM2-UNet outperforms the second-best MASNet by 5.7% on the MAS3K dataset and by 0.7% on the RMAS dataset.

4.3.4 Results on mirror detection

The results are shown in Table 5. SAM2-UNet outshines all other comparison methods in every metric. For instance, SAM2-UNet significantly outperforms HetNet in terms of IoU on the MSD dataset, with a substantial improvement of 9%. Moreover, on the PMD dataset, SAM2-UNet surpasses HetNet by 3.8% in IoU.

4.3.5 Results on polyp segmentation

The results are shown in Table 6. SAM2-UNet demonstrates state-of-the-art performance on three out of five datasets. For example, on the Kvasir dataset, SAM2-UNet achieves a mDice score of 92.8%, surpassing CFA-Net by 1.3%. Additionally, SAM2-UNet delivers the best performance on ColonDB and ETIS, exceeding CFA-Net by 6.5% and 6.4%, respectively, in mDice. Although our performance is weaker on the ClinicDB and CVC-300 datasets, SAM2-UNet still outperforms CFA-Net by an average of 2.4% in mDice across all five datasets.

4.3.6 Visual comparison

The results are shown in Figs. 3 and 4. In camouflaged object detection, our method demonstrates superior accuracy across various scenes, such as detecting a hidden face (row 1), chameleon (row 2), caterpillar (row 3), and seahorse (row 4). For polyp segmentation, our method effectively reduces false-positive rates (row 1) and false-negative rates (row 2).

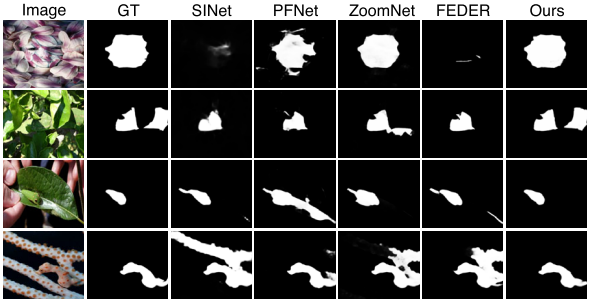

Figure 3 Visualization results on camouflaged object detection

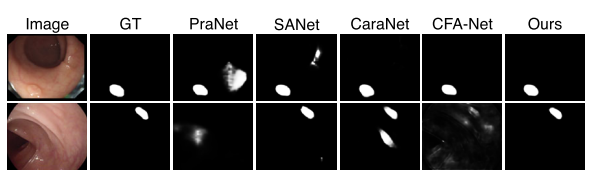

Figure 4 Visualization results on polyp segmentation

4.4 Discussion

In this section, we discuss some design choices of SAM2-UNet using MAS3K as an example.

4.4.1 Model scaling

To assess the impact of the Hiera backbone size, we tested three smaller variants: Tiny, Small, and Base+, with the results presented in Table 7. In general, larger backbones tend to yield better performance. Even with the smaller Hiera-Small backbone, SAM2-UNet still outperforms MASNet and achieves satisfactory results. As the backbone size is reduced further, SAM2-UNet produces results comparable to those of ZoomNet, even when using parameter-efficient fine-tuning. Moreover, with fewer parameters, SAM2's Hiera-Large outperforms

Table 6 Comparison of results on polyp segmentation. mDice: mean Dice

Methods	Kvasir [44]		ClinicDB [45]		ColonDB [55]		CVC-300 [56]		ETIS [57]		Average	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
PraNet [13]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567	0.801	0.739
SANet [67]	0.904	0.847	0.916	0.859	0.752	0.669	0.888	0.815	0.750	0.654	0.842	0.769
CaraNet [68]	0.913	0.859	0.921	0.876	0.775	0.700	**0.902**	**0.836**	0.740	0.660	0.850	0.786
CFA-Net [14]	0.915	0.861	**0.933**	**0.883**	0.743	0.665	0.893	0.827	0.732	0.655	0.843	0.778
SAM2-UNet	**0.928**	**0.879**	0.907	0.856	**0.808**	**0.730**	0.894	0.827	**0.796**	**0.723**	**0.867**	**0.803**

Table 7 Impact of backbone size. "Total Para." and "Lea Para." denote the total number of parameters and the number of learnable parameters in the backbone, respectively (excluding the decoder)

Methods	Tot	Lea	MAS3K [10]				
	Para.	Para.	mIoU	S_α	F_β^w	E_ϕ	MAE
Hiera-Tiny	27.1	0.3	0.735	0.868	0.788	0.911	0.032
Hiera-Small	34.3	0.4	0.760	0.882	0.814	0.924	0.028
Hiera-Base+	69.3	0.6	0.779	0.893	0.833	0.935	0.025
ViT-Large	305.2	1.6	0.737	0.872	0.799	0.935	0.026
Hiera-Large	213.8	1.7	0.799	0.903	0.848	0.943	0.021

Table 8 Impact of PEFT design

Methods	MAS3K [10]				
	mIoU	S_α	F_β^w	E_ϕ	MAE
Freeze Encoder	0.746	0.876	0.794	0.917	0.029
Mona	0.797	0.902	0.848	0.941	0.022
LMSA	0.642	0.816	0.696	0.862	0.053
Simple Adapter	0.799	0.903	0.848	0.943	0.021

the ImageNet-pretrained ViT-Large. This demonstrates the high-quality representations offered by the SAM2 pretrained Hiera backbone.

4.4.2 PEFT design

We explored other fine-tuning strategies, with the results shown in Table 8. First, we removed the adapter, leaving only the decoder trainable while completely freezing the encoder. In this case, the mIoU decreased by 5.3%, highlighting the importance of PEFT. Additionally, we tested two alternative adapter designs: Mona [69] and LMSA [70]. Mona produced results similar to those of our simple adapter, while LMSA experienced a significant performance drop. This suggests that the structure of the adapter plays a crucial role in fine-tuning effectiveness and should be carefully designed to fully harness the power of pretrained representations.

4.4.3 Input size

Table 9 presents the impact of input image resolution. The results demonstrate that input resolution significantly affects model performance. At 224 × 224, the model achieves lower scores across all metrics, indicating insufficient spatial detail for accurate segmentation. Increasing the resolution to 512 × 512 improves performance substantially, with mIoU rising to 82.4% and MAE dropping to 0.019, suggesting richer feature representation due to higher resolution. We adopt 352 × 352 as the input resolution, offering a well-balanced trade-off between accuracy and efficiency.

Table 9 Impact of input image size

Resolution	MAS3K [10]				
	mIoU	S_α	F_β^w	E_ϕ	MAE
224 × 224	0.746	0.878	0.802	0.922	0.029
512 × 512	0.824	0.912	0.869	0.949	0.019
352 × 352	0.799	0.903	0.848	0.943	0.021

5 Limitations and further improvements

Despite achieving state-of-the-art results on multiple public benchmarks with its simple architecture, SAM2-UNet still has some limitations. On one hand, it is currently designed to operate at a resolution of 352 × 352, which limits its effectiveness on high-resolution images with fine edges. On the other hand, due to the relative lack of semantic knowledge in SAM2's pretraining process, SAM2-UNet remains limited in its ability to distinguish multiple categories simultaneously. Since the release of SAM2-UNet on the preprint platform, it has garnered considerable attention within the image segmentation community, and many derivative models have been proposed. In this section, we introduce several representative works that extend the SAM2-UNet model.

5.1 Encoder improvements

Enhancements to the encoder can generally be divided into two main directions. The first focuses on parameter-efficient fine-tuning mechanisms. The original adapter in SAM2-UNet employs a simple MLP-style bottleneck, while merely increasing adapter complexity does not always yield better performance. Consequently, several studies have explored more specialized adapters designed to capture richer contextual information. For instance, SAMamba [71] introduced a feature selection adapter

(FS-Adapter), which facilitates efficient domain adaptation from natural to infrared imagery through a dual-stage feature selection process. DA2-Net [72] proposed a hierarchical low-rank adaptation (Hi-LoRA) strategy that inserts low-rank matrices into key layers of SAM2, effectively injecting domain-specific inductive biases from remote sensing data and mitigating domain shift.

The second direction involves introducing auxiliary encoders to address SAM2's limited high-level semantic representation. For example, in our official improved demo SAM2-UNeXT [73], we incorporated an additional DINOv2 [74] encoder and designed a dual-resolution learning strategy, forming a more powerful and efficient encoder that achieved further improvements across multiple downstream tasks. For example, on the MAS3K dataset, the mIoU can be further improved from 79.9% to 85.3%. Similarly, MEDU [75] integrated a parallel CNN-based encoder with feature fusion modules to combine SAM2's strong pretrained initialization with CNN's spatial generalization capacity, thereby enhancing robustness and reducing overfitting.

5.2 Decoder improvements

Decoder enhancements generally follow the design philosophy of compact, efficient segmentation networks, focusing on refined attention mechanisms to better reconstruct spatial details from encoded features. For example, AIS-FCANet [76] proposed a frequency–spatial domain context-aware fusion module (FSCM) that leverages both frequency- and spatial-domain attention mechanisms to strengthen multilevel feature fusion and enable effective cross-modal interaction. SLENet [77] designed a multiscale supervised decoder (MSSD) that enhances spatial awareness and focus through a top-down, iterative cross-scale fusion strategy.

6 Conclusion

In this paper, we propose SAM2-UNet, a simple yet effective U-shaped framework for versatile segmentation across both natural and medical domains. SAM2-UNet is designed for ease of understanding and use, featuring a SAM2 pretrained Hiera encoder coupled with a classic U-Net decoder. Extensive experiments across 18 datasets on five benchmarks demonstrate the effectiveness of SAM2-UNet. Our SAM2-UNet can serve as a new baseline for developing future SAM2 variants.

Abbreviations

COD, camouflaged object detection; MAS, marine animal segmentation; PEFT, parameter-efficient fine-tuning; RFB, receptive field block; SAM2, segment anything model 2; SOD, salient object detection; VFM, vision foundation model.

Author contributions

XX and ZW designed the concept and method of the work. All authors jointly helped experiments and analyze the results. All authors read and approved the final manuscript.

Funding information

This work is supported by the National Natural Science Foundation of China (No. 62322608), and the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515010255).

Data availability

All the datasets used in this study are publicly accessible. CAMO [42], COD10K [43], CHAMELEON [46], and NC4K [47] are from https://github.com/ChunmingHe/FEDER. DUTS [48], DUT-OMRON [49], HKU-IS [50], PASCAL-S [51], and ECSSD [52] are from https://github.com/moothes/SALOD. MAS3K [10] and RMAS [9] are from https://github.com/zhenqifu/MASNet. MSD [53] and PMD [54] are from https://github.com/Catherine-R-He/HetNet. Kvasir-SEG [44], CVC-ClincDB [45], CVC-ColonDB [55], CVC-300 [56], and ETIS [57] are from https://github.com/DengPingFan/PraNet. The code for SAM2-UNet is available at https://github.com/WZH0120/SAM2-UNet.

Declarations

Competing interests

The authors declare that they have no conflict of interest or competing interests.

Author details

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. [2]School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China. [3]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, China. [4]Thrust of Robotics and Autonomous Systems, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. [5]AIM Lab, Faculty of IT, Monash University, Melbourne, Australia. [6]School of Future Technology, South China University of Technology, Guangzhou, China. [7]School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China.

Received: 14 November 2025 Revised: 20 December 2025
Accepted: 24 December 2025 Published online: 13 January 2026

References

1. Ding, H., Liu, C., He, S., Jiang, X., & Loy, C. C. (2023). Mevis: a large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2694–2703). Piscataway: IEEE.
2. Ding, H., Liu, C., He, S., Jiang, X., Torr, P. H., & Bai, S. (2023). MOSE: a new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 20224–20234). Piscataway: IEEE.
3. Ding, H., Ying, K., Liu, C., He, S., Jiang, X., Jiang, Y.-G., Torr, P. H., & Bai, S. (2025). MOSEv2: a more challenging dataset for video object segmentation in complex scenes. arXiv preprint. arXiv:2508.05630.
4. Ding, H., Liu, C., He, S., Ying, K., Jiang, X., Loy, C. C., & Jiang, Y.-G. (2025). MeViS: a multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(12), 11400–11416.
5. Sun, Y., Chen, G., Zhou, T., Zhang, Y., & Liu, N. (2021). Context-aware cross-level fusion network for camouflaged object detection. In Z.-H. Zhou (Ed.), *Proceedings of the 30th international joint conference on artificial intelligence* (pp. 1025–1031). Palo Alto: AAAI Press.
6. Zhou, Y., Sun, G., Li, Y., Xie, G.-S., Benini, L., & Konukoglu, E. (2025). When SAM2 meets video camouflaged object segmentation: a comprehensive evaluation and adaptation. *Visual Intelligence*, *3*, 10.
7. Wei, J., Wang, S., & Huang, Q. (2020). F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 12321–12328). Palo Alto: AAAI Press.
8. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., & Li, X. (2023). Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22046–22055). Piscataway: IEEE.
9. Fu, Z., Chen, R., Huang, Y., Cheng, E., Ding, X., & Ma, K.-K. (2024). MASNet: a robust deep marine animal segmentation network. *IEEE Journal of Oceanic Engineering*, *49*, 1104–1115.

10. Li, L., Dong, B., Rigall, E., Zhou, T., Dong, J., & Chen, G. (2021). Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*, 2303–2314.
11. He, R., Lin, J., & Lau, R. W. (2023). Efficient mirror detection via multi-level heterogeneous learning. In B. Williams, Y. Chen, & J. Neville (Eds.), *Proceedings of the 37th AAAI conference on artificial intelligence* (pp. 790–798). Palo Alto: AAAI Press.
12. Guan, H., Lin, J., & Lau, R. W. (2022). Learning semantic associations for mirror detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5941–5950). Piscataway: IEEE.
13. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020). PraNet: parallel reverse attention network for polyp segmentation. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, & L. Joskowicz (Eds.), *Proceedings of the 23rd international conference on medical image computing and computer-assisted intervention* (pp. 263–273). Cham: Springer.
14. Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., & Shen, D. (2023). Cross-level feature aggregation network for polyp segmentation. *Pattern Recognition*, *140*, 109555.
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026). Piscataway: IEEE.
16. Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. (2025). SAM 2: segment anything in images and videos. In *Proceedings of the 13th international conference on learning representations* (pp. 1–44). Retrieved August 15, 2025, from https://openreview.net/forum?id=Ha6RTeWMd0.
17. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., & Huang, T. (2023). SegGPT: towards segmenting everything in context. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1130–1140). Piscataway: IEEE.
18. Li, X., Yuan, H., Li, W., Ding, H., Wu, S., Zhang, W., Li, Y., Chen, K., & Loy, C. C. (2024). OMG-Seg: is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 27948–27959). Piscataway: IEEE.
19. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., & Mao, P. (2023). Sam-adapter: adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 3367–3375). Piscataway: IEEE.
20. Zhang, K., & Liu, D. (2023). Customized segment anything model for medical image segmentation. arXiv preprint. arXiv:2304.13785.
21. Huang, D., Xiong, X., Ma, J., Li, J., Jie, Z., Ma, L., & Li, G. (2024). AlignSAM: aligning segment anything model to open context via reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3205–3215). Piscataway: IEEE.
22. Zhang, Y., Cheng, T., Hu, R., Liu, L., Liu, H., Ran, L., Chen, X., Liu, W., & Wang, X. (2024). EVF-SAM: early vision-language fusion for text-prompted segment anything model. arXiv preprint. arXiv:2406.20076.
23. Li, W., Xiong, X., Xia, P., Ju, L., & Ge, Z. (2024). TP-DRSeg: improving diabetic retinopathy lesion segmentation with explicit text-prompts assisted SAM. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, & J. A. Schnabel (Eds.), *Proceedings of the 27th international conference on medical image computing and computer-assisted intervention* (pp. 743–753). Cham: Springer.
24. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Qiao, Y., Gao, P., & Li, H. (2024). Personalize segment anything model with one shot. In *Proceedings of the 12th international conference on learning representations* (pp. 1–30). Retrieved August 15, 2024, from https://openreview.net/forum?id=6Gzkhoc6YS.
25. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., & Shen, C. (2024). Matcher: segment anything with one shot using all-purpose feature matching. In *Proceedings of the 12th international conference on learning representations* (pp. 1–22). Retrieved August 15, 2024, from https://openreview.net/forum?id=yzRXdhk2he.
26. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Proceedings of the 18th international conference on medical image computing and computer-assisted intervention* (pp. 234–241). Cham: Springer.
27. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, *39*(6), 1856–1867.
28. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al. (2024). TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, *97*, 103280.
29. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-Unet: Unet-like pure transformer for medical image segmentation. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), *Proceedings of the 17th European conference on computer vision workshops* (pp. 205–218). Cham: Springer.
30. Gao, Y., Xia, W., Hu, D., Wang, W., & Gao, X. (2024). DeSAM: decoupled segment anything model for generalizable medical image segmentation. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, & J. A. Schnabel (Eds.), *Proceedings of the 27th international conference on medical image computing and computer-assisted intervention* (pp. 509–519). Cham: Springer.
31. Xiong, X., Wang, C., Li, W., & Li, G. (2023). Mammo-SAM: adapting foundation segment anything model for automatic breast mass segmentation in whole mammograms. In X. Cao, X. Xu, I. Rekik, Z. Cui, & X. Ouyang (Eds.), *Proceedings of the 14th international workshop on machine learning in medical imaging* (pp. 176–185). Cham: Springer.
32. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th international conference on learning representations* (pp. 1–21). Retrieved August 15, 2024, from https://openreview.net/forum?id=YicbFdNTTy.
33. Sun, X., Liu, J., Shen, H., Zhu, X., & Hu, P. (2025). On efficient variants of segment anything model: a survey. *International Journal of Computer Vision*, *133*(10), 7406–7436.
34. Wang, H., Vasu, P. K. A., Faghri, F., Vemulapalli, R., Farajtabar, M., Mehta, S., Rastegari, M., Tuzel, O., & Pouransari, H. (2024). Sam-clip: merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 3635–3647). Piscataway: IEEE.
35. Xin, Y., Luo, S., Zhou, H., Du, J., Liu, X., Fan, Y., Li, Q., & Du, Y. (2024). Parameter-efficient fine-tuning for pre-trained vision models: a survey. arXiv preprint. arXiv:2402.02242.
36. Wang, Y., Wang, R., Fan, X., Wang, T., & He, X. (2023). Pixels, regions, and objects: multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10031–10040). Piscataway: IEEE.
37. Li, W., Ye, T., Xiong, X., Bai, J., Tang, F., Song, W., Xing, Z., Ju, L., Li, G., & Zhu, L. (2025). Glasswizard: harvesting diffusion priors for glass surface detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 17848–17858). Piscataway: IEEE.
38. Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al. (2023). Hiera: a hierarchical vision transformer without the bells-and-whistles. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the international conference on machine learning* (pp. 29441–29454). Retrieved August 15, 2024, from https://proceedings.mlr.press/v202/ryali23a.html.
39. Liu, S., Huang, D., & Wang, Y. (2018). Receptive field block net for accurate and fast object detection. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 385–400). Cham: Springer.
40. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the international conference on machine learning* (pp. 2790–2799). Retrieved August 15, 2024, from http://proceedings.mlr.press/v97/houlsby19a.html.
41. Qiu, Z., Hu, Y., Li, H., & Liu, J. (2023). Learnable ophthalmology SAM. arXiv preprint. arXiv:2304.13425.
42. Le, T.-N., Nguyen, T. V., Nie, Z., Tran, M.-T., & Sugimoto, A. (2019). Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, *184*, 45–56.
43. Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J., & Shao, L. (2020). Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2777–2787). Piscataway: IEEE.
44. Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., De Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-SEG: a segmented polyp dataset. In Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, & W. De Neve (Eds.), *Proceedings of the 26th international conference on multimedia modeling* (pp. 451–462). Cham: Springer.

45. Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & Vilariño, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, *43*, 99–111.
46. Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., & Kozieł, P. (2018). Animal camouflage analysis: chameleon database. *Unpublished Manuscript*, *2*, 7.
47. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., & Fan, D.-P. (2021). Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11591–11601). Piscataway: IEEE.
48. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., & Ruan, X. (2017). Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 136–145). Piscataway: IEEE.
49. Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3166–3173). Piscataway: IEEE.
50. Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5455–5463). Piscataway: IEEE.
51. Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 280–287). Piscataway: IEEE.
52. Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1155–1162). Piscataway: IEEE.
53. Yang, X., Mei, H., Xu, K., Wei, X., Yin, B., & Lau, R. W. (2019). Where is my mirror? In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8809–8818). Piscataway: IEEE.
54. Lin, J., Wang, G., & Lau, R. W. (2020). Progressive mirror detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3697–3705). Piscataway: IEEE.
55. Tajbakhsh, N., Gurudu, S. R., & Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, *35*, 630–644.
56. Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., Drozdzal, M., & Courville, A. (2017). A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, *2017*, 4037190.
57. Silva, J., Histace, A., Romain, O., Dray, X., & Granado, B. (2014). Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, *9*, 283–293.
58. Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., & Borji, A. (2017). Structure-measure: a new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision* (pp. 4548–4557). Piscataway: IEEE.
59. Margolin, R., Zelnik-Manor, L., & Tal, A. (2014). How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 248–255). Piscataway: IEEE.
60. Fan, D.-P., Ji, G.-P., Qin, X., & Cheng, M.-M. (2021). Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, *6*, 5.
61. Mei, H., Ji, G.-P., Wei, Z., Yang, X., Wei, X., & Fan, D.-P. (2021). Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8772–8781). Piscataway: IEEE.
62. Pang, Y., Zhao, X., Xiang, T.-Z., Zhang, L., & Lu, H. (2022). Zoom in and out: a mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2160–2170). Piscataway: IEEE.
63. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognition*, *106*, 107404.
64. Zhuge, M., Fan, D.-P., Liu, N., Zhang, D., Xu, D., & Shao, L. (2022). Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 3738–3752.
65. Wu, Y.-H., Liu, Y., Zhang, L., Cheng, M.-M., & Ren, B. (2022). EDN: salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, *31*, 3125–3136.
66. Liu, J., Zhang, J., & Barnes, N. (2022). Modeling aleatoric uncertainty for camouflaged object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1445–1454). Piscataway: IEEE.
67. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S. K., & Cui, S. (2021). Shallow attention network for polyp segmentation. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, & C. Essert (Eds.), *Proceedings of the 24th international conference on medical image computing and computer-assisted intervention* (pp. 699–708). Cham: Springer.
68. Lou, A., Guan, S., Ko, H., & Loew, M. H. (2022). CaraNet: context axial reverse attention network for segmentation of small medical objects. In *Proceedings of the medical imaging 2022: image processing* (pp. 81–92). Bellingham: SPIE.
69. Yin, D., Hu, L., Li, B., & Zhang, Y. (2023). Adapter is all you need for tuning visual tasks. arXiv preprint. arXiv:2311.15010.
70. Gao, S., Zhang, P., Yan, T., & Lu, H. (2024). Multi-scale and detail-enhanced segment anything model for salient object detection. In J. Cai, M. S. Kankanhalli, B. Prabhakaran, S. Boll, R. Subramanian, L. Zheng, V. K. Singh, P. César, L. Xie, & D. Xu (Eds.), *Proceedings of the 32nd ACM international conference on multimedia* (pp. 9894–9903). New York: ACM.
71. Xu, W., Zheng, S., Wang, C., Zhang, Z., Ren, C., Xu, R., & Xu, S. (2025). Samamba: adaptive state space modeling with hierarchical vision for infrared small target detection. *Information Fusion*, *124*, 103338.
72. Ning, H., He, Q., Lei, T., Cao, X., Zhang, W., Chen, Y., & Nandi, A. K. (2025). DA2-Net: integrating SAM2 with domain adaption and difference aggregation for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, *63*, 1–17.
73. Xiong, X., Wu, Z., Zhang, L., Lu, L., Li, M., & Li, G. (2025). Sam2-unext: an improved high-resolution baseline for adapting foundation models to downstream segmentation tasks. arXiv preprint. arXiv:2508.03566.
74. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: learning robust visual features without supervision. arXiv preprint. arXiv:2304.07193.
75. Wang, R., Guo, J., Zhang, J., Qi, L., & Shi, Y. (2025). Fusing dual encoders: single-source domain generalization with extremely few annotations. In J. C. Gee, D. C. Alexander, J. Hong, J. E. Iglesias, C. H. Sudre, A. Venkataraman, P. Golland, J. H. Kim, & J. Park (Eds.), *Proceedings of the 28th international conference on medical image computing and computer-assisted intervention* (pp. 289–298). Cham: Springer.
76. Xue, W., Ai, J., Zhu, Y., Chen, J., & Zhuang, S. (2025). AIS-FCANet: long-term AIS data assisted frequency-spatial contextual awareness network for salient ship detection in SAR imagery. *IEEE Transactions on Aerospace and Electronic Systems*, *61*, 15166–15171.
77. Huang, X., Sun, H., Liu, N., Zhou, H., & Yao, Y. (2025). SLENet: a guidance-enhanced network for underwater camouflaged object detection. arXiv preprint. arXiv:2509.03786.

Publisher's note