SAM2-UNeXT: An Improved High-Resolution Baseline for Adapting Foundation Models to Downstream Segmentation Tasks

Xinyu Xiong[1*] Zihuang Wu[2*] Lei Zhang[1] Lei Lu[3] Ming Li[4] Guanbin Li[1**]

[1]Sun Yat-sen University, [2]Jiangxi Normal University, [3]Hainan University, [4]Shandong Inspur Database Technology Co., Ltd, *Equal Contribution, **Corresponding Author

Abstract

Recent studies have highlighted the potential of adapting the Segment Anything Model (SAM) for various downstream tasks. However, constructing a more powerful and generalizable encoder to further enhance performance remains an open challenge. In this work, we propose SAM2-UNeXT, an advanced framework that builds upon the core principles of SAM2-UNet while extending the representational capacity of SAM2 through the integration of an auxiliary DINOv2 encoder. By incorporating a dual-resolution strategy and a dense glue layer, our approach enables more accurate segmentation with a simple architecture, relaxing the need for complex decoder designs. Extensive experiments conducted on four benchmarks, including dichotomous image segmentation, camouflaged object detection, marine animal segmentation, and remote sensing saliency detection, demonstrate the superior performance of our proposed method. The code is available at `https://github.com/WZH0120/SAM2-UNeXT`.

1 Introduction

Foundation models are playing an increasingly pivotal role in computer vision [1], natural language processing [34], intelligent medicine [21], autonomous driving [50], and other domains [23, 54, 59]. In the field of image segmentation, the Segment Anything Model (SAM) [22, 42] family has sparked significant interest. Traditional small segmentation networks typically devote substantial design effort to complex decoder modules. However, a fundamental limitation persists: once the knowledge is lost in the encoding stage, it cannot be fully recovered during decoding. In contrast, foundation models leverage their large parameter capacity and sophisticated pretraining strategies to learn high-quality representations, enabling accurate segmentation performance even with relatively simple decoder architectures.

Although foundation models demonstrate strong generalization capabilities, task-specific adaptation, such as parameter-efficient fine-tuning (PEFT) [55], remains important for many downstream applications [62, 63]. Recent methods have achieved promising results by incorporating lightweight adapters [4, 18], LoRA modules [19], or similar components into the encoder, often in conjunction with decoder refinement strategies [53, 57]. Nevertheless, relying solely on SAM still results in limited generalization in some scenarios. For example, linear probing [8] of the SAM encoder on the ImageNet [6] classification yields significantly lower accuracy compared to other large models such as CLIP [41] and DINOv2 [35]. A plausible explanation is that SAM's class-agnostic segmentation pretraining induces a representational bias that favoring fine-grained local details while capturing limited global semantic context.

Building upon the above analysis, we propose SAM2-UNeXT, a unified and extensible framework that synergistically integrates multiple foundation models, including SAM2 [42] and DINOv2 [35], to harness their complementary strengths in detail perception and semantic representation. The proposed SAM2-UNeXT offers the following key benefits:

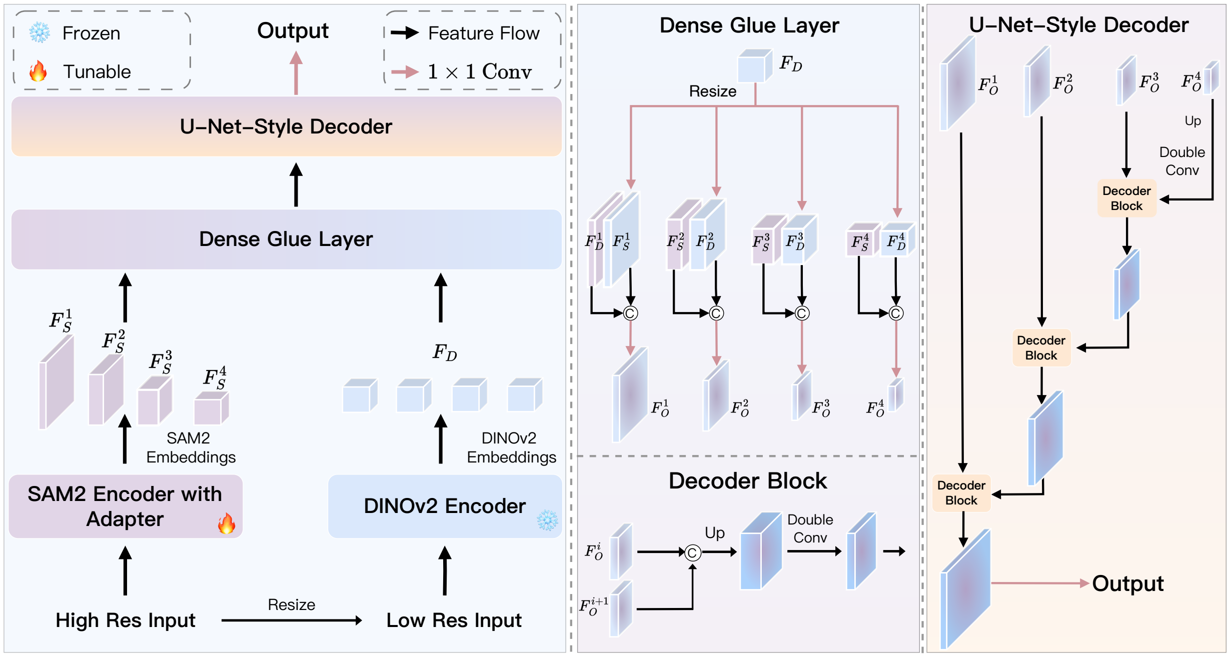

Figure 1 | Overview of the proposed SAM2-UNeXT.

- **Simplicity.** SAM2-UNeXT simplifies any additional attention design and focuses on a lightweight and efficient encoder fusion strategy.
- **Scalability.** With support for dynamic resolution adjustment and flexible auxiliary encoder configurations, SAM2-UNeXT can be readily adapted to a broad range of downstream tasks.
- **Effectiveness.** Extensive experiments on four public benchmarks demonstrate that SAM2-UNeXT consistently achieves strong segmentation performance across diverse scenarios under limited training epochs.

2 Method

As illustrated in Figure 1, the proposed architecture consists of four key components: a SAM2 encoder, a DINOv2 encoder, a dense glue layer, and a U-Net-style decoder.

2.1 SAM2 Encoder

In this stage, we closely follow the practice of SAM2-UNet [56], adapting the Hiera [44] encoder from SAM2 and freezing all of its original parameters. Parameter-efficient fine-tuning (PEFT) is performed by inserting lightweight adapters [18] before each Hiera block. The adapter adopts a simple "MLP-GeLU-MLP-GeLU" structure with a 32-channel bottleneck.

2.2 DINOv2 Encoder

Compared with the Segment Anything series [22, 42], DINOv2 [35] serves as a more general-purpose vision foundation model. Trained through self-supervised learning, it demonstrates strong transferability across a wide range of vision tasks, including classification, segmentation, and depth estimation. In line with the original implementation, we freeze all DINOv2 parameters and do not apply any parameter-efficient fine-tuning strategies to balance training efficiency and performance.

2.3 Dual-Resolution Design

A straightforward approach to combining the two large encoders is to process inputs at the same resolution; however, this is computationally inefficient. In particular, for DINOv2, which relies on standard self-attention mechanisms [7], increasing the input resolution leads to a substantial rise in computational cost. Considering that SAM focuses on fine-grained local details, whereas DINOv2 emphasizes global semantic understanding, we adopt a dual-resolution strategy: the SAM encoder operates on a higher-resolution input (H_h, W_h), while the DINOv2 encoder processes a lower-resolution input (H_l, W_l).

2.4 Dense Glue Layer

Unlike the hierarchical design of Hiera, the vanilla Vision Transformer [7] architecture adopted by DINOv2 produces non-hierarchical, scale-consistent embeddings at every layer. A common approach [3, 58] to leverage such transformer features is to enhance the final feature map of a hierarchical encoder. Instead, we employ a dense fusion strategy inspired by the observation that DINOv2 exhibits strong zero-shot capabilities: its encoded representations become highly interpretable after principal component analysis, effectively highlighting the foreground of interest without any fine-tuning. In other words, these features can be regarded as spatial attention maps enriched with global semantic information.

Based on this, we first apply four 1×1 convolutions to align the channel dimension of the DINOv2 features (1024 channels for DINOv2-L) with those of the four stages of the SAM2 encoder features (144, 288, 576, and 1152 channels for Hiera-L). Next, the DINOv2 features are resized to match the spatial dimensions of each corresponding SAM2 feature map and fused via simple channel-wise concatenation. Finally, the concatenated features are compressed to 128 channels through 1×1 convolutions to improve training efficiency.

2.5 U-Net-Style Decoder

In this stage, we mostly follow the design of SAM2-UNet by replacing the original transformer-based decoder in SAM2 with a U-Net-style decoder [43], where each decoder block consists of two consecutive "Conv–BN–ReLU" layers. The major difference is that we introduce an additional partial decoder that without feature concatenation, resulting in a total of four decoder stages. This modification increases the resolution of the final segmentation feature map to one-half (rather than one-quarter) of the high-resolution input, which is advantageous for tasks that are sensitive to boundary segmentation accuracy.

3 Experiment

3.1 Datasets and Benchmarks

We conduct experiments on four public benchmarks spanning a range of segmentation tasks:

Dichotomous Image Segmentation. We use the DIS5K [39] dataset for evaluation. The training set (DIS-TR) contains 3,000 images, while the evaluation is conducted on five subsets: DIS-VD (470), DIS-TE1 (500), DIS5K-TE2 (500), DIS-TE3 (500), and DIS-TE4 (500). Performance is measured using four metrics: S-measure (S_α) [9], weighted F-measure (F_β^w) [32], mean E-measure (E_ϕ) [11], and mean absolute error (MAE).

Camouflaged Object Detection. We evaluate on four datasets: CHAMELEON [45], CAMO [24], COD10K [10], and NC4K [31]. The unified training set consists of 4,040 images (3,040 from COD10K and 1,000 from CAMO). The remaining CHAMELEON (76), CAMO (250), COD10K (2,026), and NC4K (4,121) images are used for testing. We report results using S-measure (S_α), adaptive F-measure (F_β), mean E-measure (E_ϕ), and mean absolute error (MAE).

Marine Animal Segmentation. Two datasets are used for this task: MAS3K [26], with 1,769 training images and 1,141 test images; and RMAS [13], with 2,514 training images and 500 test images. Evaluation is based on five metrics: mIoU, S-measure (S_α), weighted F-measure (F_β^w), mean E-measure (E_ϕ), and mean absolute error (MAE).

Remote Sensing Saliency Detection. We use two datasets: EORSSD [61], with 1,400 training images and 600 test images; and ORSI-4199 [47], with 2,000 training images and 2,199 test images. Five metrics are used for evaluation: S-measure (S_α), mean F-measure (F_β^{mean}), max F-measure (F_β^{max}), mean E-measure (E_ϕ), and mean absolute error (MAE).

Table 1 | Results on dichotomous image segmentation.

Methods	DIS-VD [39]				DIS-TE1 [39]				DIS-TE2 [39]			
	S_α	F_β^w	E_ϕ	MAE	S_α	F_β^w	E_ϕ	MAE	S_α	F_β^w	E_ϕ	MAE
U^2Net [38]	0.785	0.656	0.809	0.089	0.762	0.601	0.783	0.085	0.798	0.676	0.825	0.083
HRNet [49]	0.767	0.641	0.824	0.095	0.742	0.579	0.797	0.088	0.784	0.664	0.840	0.087
IS-Net [39]	0.813	0.717	0.856	0.074	0.787	0.662	0.820	0.074	0.823	0.728	0.858	0.070
UDUN [37]	0.838	0.763	0.892	0.059	0.817	0.720	0.864	0.059	0.843	0.768	0.886	0.058
BiRefNet [63]	0.898	0.854	0.931	0.038	0.885	0.819	0.911	0.037	0.900	0.857	0.930	0.036
SAM2-UNeXT	**0.910**	**0.864**	**0.938**	**0.034**	**0.892**	**0.829**	**0.917**	**0.034**	**0.916**	**0.873**	**0.941**	**0.030**
Methods	DIS-TE3 [39]				DIS-TE4 [39]				DIS-TE(1-4) [39]			
	S_α	F_β^w	E_ϕ	MAE	S_α	F_β^w	E_ϕ	MAE	S_α	F_β^w	E_ϕ	MAE
U^2Net [38]	0.823	0.721	0.856	0.073	0.814	0.707	0.837	0.085	0.799	0.676	0.825	0.082
HRNet [49]	0.805	0.700	0.869	0.080	0.792	0.687	0.854	0.092	0.781	0.658	0.840	0.087
IS-Net [39]	0.836	0.758	0.883	0.064	0.830	0.753	0.870	0.072	0.819	0.726	0.858	0.070
UDUN [37]	0.865	0.809	0.917	0.050	0.849	0.792	0.901	0.059	0.844	0.772	0.892	0.057
BiRefNet [63]	0.919	0.893	0.955	0.028	0.900	0.864	0.939	0.039	0.901	0.858	0.934	0.035
SAM2-UNeXT	**0.926**	**0.897**	**0.956**	**0.027**	**0.909**	**0.867**	**0.944**	**0.037**	**0.911**	**0.867**	**0.940**	**0.032**

Table 2 | Results on camouflaged object detection.

Methods	CHAMELEON [45]				CAMO [24]				COD10K [10]				NC4K [31]			
	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 [10]	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 [33]	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 [36]	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 [16]	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 [56]	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
SAM2-UNeXT	**0.942**	**0.916**	**0.972**	**0.013**	**0.903**	**0.891**	**0.941**	**0.036**	**0.924**	**0.867**	**0.964**	**0.013**	**0.923**	**0.903**	**0.954**	**0.022**

Table 3 | Results on marine animal segmentation.

Methods	MAS3K [26]					RMAS [13]				
	$mIoU$	S_α	F_β^w	E_ϕ	MAE	$mIoU$	S_α	F_β^w	E_ϕ	MAE
C2FNet [46]	0.717	0.851	0.761	0.894	0.038	0.721	0.858	0.788	0.923	0.026
OCENet [29]	0.667	0.824	0.703	0.868	0.052	0.680	0.836	0.752	0.900	0.030
ZoomNet [36]	0.736	0.862	0.780	0.898	0.032	0.728	0.855	0.795	0.915	0.022
MASNet [13]	0.742	0.864	0.788	0.906	0.032	0.731	0.862	0.801	0.920	0.024
SAM2-UNet [56]	0.799	0.903	0.848	0.943	0.021	0.738	0.874	0.810	0.944	0.022
SAM2-UNeXT	**0.853**	**0.926**	**0.900**	**0.960**	**0.014**	**0.774**	**0.883**	**0.841**	**0.949**	**0.019**

3.2 Implementation Details

Our method is implemented in PyTorch and trained on a NVIDIA RTX 4090 GPU with 24 GB of memory. We use the AdamW optimizer with an initial learning rate of 0.0002 and apply cosine learning rate decay to stabilize training. The overall loss function [52] consists of a weighted cross-entropy loss ($\mathcal{L}_{\mathrm{BCE}}^{\omega}$) and a

Table 4 | Results on remote sensing saliency detection.

Methods	ORSI-4199 [47]					EORSSD [61]				
	S_α	F_β^{mean}	F_β^{max}	E_ϕ	MAE	S_α	F_β^{mean}	F_β^{max}	E_ϕ	MAE
EMFINet [64]	0.859	0.810	0.817	0.902	0.045	0.932	0.851	0.874	0.960	0.008
ACCoNet [25]	0.868	0.861	0.865	0.934	0.031	0.929	0.855	0.884	0.965	0.007
ERPNet [65]	0.865	0.839	0.854	0.917	0.037	0.925	0.827	0.874	0.937	0.008
AESINet [60]	0.870	0.863	0.868	0.936	0.031	0.935	0.850	0.879	0.965	0.006
SFANet [40]	0.876	0.866	0.871	0.939	0.029	0.935	0.868	0.883	0.973	0.006
SAM2-UNeXT	**0.887**	**0.873**	**0.886**	**0.942**	**0.026**	**0.948**	**0.892**	**0.905**	**0.978**	**0.004**

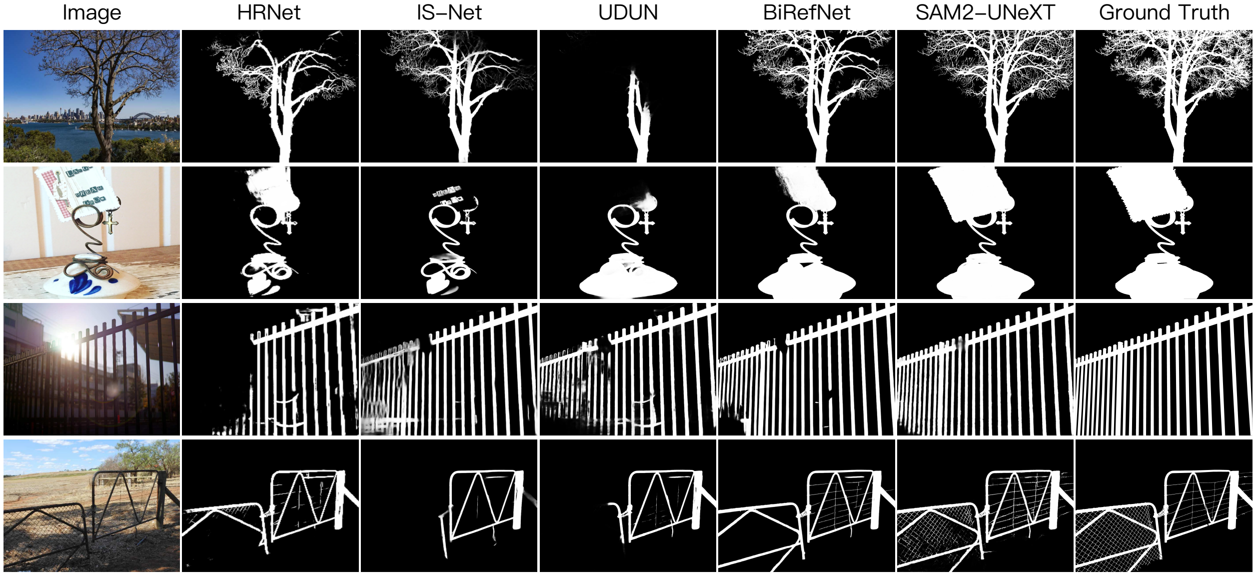

Figure 2 | Visualization results on dichotomous image segmentation.

weighted IoU loss ($\mathcal{L}_{\mathrm{IoU}}^{\omega}$). Two data augmentation strategies, including random horizontal and vertical flipping, are employed during training. Unless otherwise specified, we adopt the large version of SAM2 and DINOv2. The input resolutions are set to $(H_h, W_h) = (1024, 1024)$ for the SAM2 branch and $(H_l, W_l) = (448, 448)$ for the DINOv2 branch. All models are trained with a batch size of 1 for 20 epochs across all tasks.

3.3 Comparison with State-of-the-Art Methods

In this subsection, we first analyze the quantitative results across multiple benchmarks, followed by qualitative visual comparisons for dichotomous image segmentation.

Dichotomous Image Segmentation. The results are presented in Table 1, SAM2-UNeXT achieves steady performance gains over the second-best method, BiRefNet. Specifically, on the DIS-VD subset, our method improves the S-measure by 1.2%.

Camouflaged Object Detection. The results are presented in Table 2. Compared with SAM2-UNet, the new SAM2-UNeXT achieves consistent improvements across all metrics. For example, on the CHAMELEON dataset, SAM2-UNeXT improves the S-measure by 2.8%.

Marine Animal Segmentation. The results are presented in Table 3. SAM2-UNeXT significantly outperforms existing methods by a large margin. For instance, on the MAS3K dataset, our method improves the mIoU by 5.4%.

Remote Sensing Saliency Detection. The results are presented in Table 4. SAM2-UNeXT outperforms all

Table 5 | Results on different auxiliary encoders.

Methods	MAS3K [26]				
	$mIoU$	S_α	F_β^w	E_ϕ	MAE
w/o aux	0.832	0.918	0.882	0.949	0.017
ResNet-101 [17]	0.840	0.917	0.885	0.955	0.017
PVTv2-b5 [51]	0.833	0.917	0.878	0.952	0.018
DINOv2-S [35]	0.836	0.917	0.881	0.950	0.017
DINOv2-B [35]	0.843	0.921	0.890	0.955	0.015
DINOv2-L	**0.853**	**0.926**	**0.900**	**0.960**	**0.014**

Table 6 | Results on different resolutions.

Resolution		MAS3K [26]				
H	L	$mIoU$	S_α	F_β^w	E_ϕ	MAE
352	352	0.820	0.913	0.871	0.952	0.016
1024	224	0.842	0.920	0.888	0.957	0.015
1024	672	**0.853**	0.924	0.897	**0.960**	**0.014**
1024	**448**	**0.853**	**0.926**	**0.900**	**0.960**	**0.014**

competing methods on both datasets. Notably, on the ORSI-4199 dataset, our method achieves a 1.1% improvement in S-measure.

Qualitative Comparison. Figure 2 illustrates visual comparisons on the dichotomous image segmentation task. Our method demonstrates superior segmentation accuracy across diverse scenarios: fine-grained tree branches (row 1), complex multi-object compositions (row 2), light variations (row 3), and scenes with grid structures and shadow interference (row 4). SAM2-UNeXT effectively handles curved edges, thin structures, and subtle visual boundaries, delivering better segmentation results even under challenging conditions.

3.4 Discussion

In this section, we analyze the design choices of SAM2-UNeXT using MAS3K as a representative benchmark.

3.4.1 Impact of Auxiliary Encoder

We investigate the effect of different auxiliary encoder designs, as shown in Table 5:

Row 1. The auxiliary encoder is removed. In this setting, the model roughly becomes a high-resolution variant of SAM2-UNet [56]. Although it performs better than the low-resolution version of SAM2-UNet, its accuracy remains lower than that of configurations with an auxiliary encoder.

Row 2 & 3. We use ResNet-101 [17] and PVTv2-b5 [51] as auxiliary encoders with parameters trainable. The results show marginal improvements compared to the setting without an auxiliary encoder, suggesting limited benefits from these conventional backbones under a simple fusion strategy.

Row 4 & 5. We replace the auxiliary encoder with frozen small and base versions of DINOv2 [35]. The results indicate that larger variants generally yield better performance.

3.4.2 Impact of Dynamic Resolution

We also explore the effect of different resolution combinations, as shown in Table 6:

Row 1. Both SAM2 and DINOv2 encoders operate at a uniform low resolution of 352×352. This setting results in the lowest performance among all tested configurations, though it still surpasses the original SAM2-UNet baseline.

Row 2. The high resolution is fixed at 1024×1024 for the SAM2 branch, while the low resolution for the DINOv2 branch is reduced to 224×224. A slight performance drop is observed compared to the 448×448 setting, but it still outperforms the uniform 352×352 case.

Row 3. The high resolution remains at 1024×1024, while the low resolution is increased to 672×672. The performance difference is negligible compared to the 448×448 setting, but inference cost increases significantly, making this configuration less practical.

4 Related Work

4.1 Fusing Foundation Models

Integrating different foundation models has become a common strategy in recent years. Many Vision-Language Models (VLMs) [2, 66] are composed of a vision encoder paired with a Large Language Model (LLM), enabling flexible combinations tailored to diverse applications. For the SAM series, there have been several efforts [20, 27, 48] to enhance language understanding by incorporating CLIP [41]. Other works have focused on improving few-shot segmentation capabilities by integrating pretrained vision encoders like DINOv2 [35], exemplified by Matcher [30]. The study most related to ours is [58], which also introduces an auxiliary DINOv2 encoder to form a U-shaped architecture. However, their focus lies in the design of more sophisticated decoder design, such as content-guided attention [5] and wavelet convolution [12].

4.2 Image Segmentation

Image segmentation, viewed as a pixel-level classification task, can be broadly categorized into binary segmentation [62], semantic segmentation [15], instance segmentation [14], and panoptic segmentation [28]. This work focuses on binary segmentation, where all foreground pixels are assigned to a single class, and the remaining pixels are treated as background. Binary segmentation underpins many important application domains, including dichotomous image segmentation [37, 63], camouflaged object detection [16, 53], marine animal segmentation [13, 26], and remote sensing saliency detection [40, 60]. Most existing methods tend to design task-specific decoders for each segmentation scenario. In contrast, our proposed method introduces a unified framework capable of achieving state-of-the-art performance across multiple binary segmentation tasks with a single model architecture.

5 Conclusion

In this paper, we presented SAM2-UNeXT, a simple yet effective framework that integrates two powerful foundation models: SAM2 and DINOv2, through a decoupled resolution strategy. This design leverages the complementary feature biases of each model, resulting in enhanced segmentation performance. Extensive experiments on four benchmark datasets demonstrate the effectiveness and generalizability of our approach. Moreover, SAM2-UNeXT offers high customizability, making it well-suited for adaptation to a wide range of downstream tasks. By adjusting the dynamic resolution configuration or incorporating alternative auxiliary encoders, the framework holds promise for extending SAM2-based models to previously underexplored segmentation scenarios.

References

[1] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, page 103280, 2024.

[4] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *ICCVW*, pages 3367–3375, 2023.

[5] Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, 33:1002–1015, 2024.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[8] Miguel Espinosa, Chenhongyi Yang, Linus Ericsson, Steven McDonagh, and Elliot J Crowley. There is no samantics! exploring sam as a backbone for visual understanding tasks. *arXiv preprint arXiv:2411.15288*, 2024.

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017.

[10] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020.

[11] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6(6):5, 2021.

[12] Shahaf E Finder, Roy Amoyal, Eran Treister, and Oren Freifeld. Wavelet convolutions for large receptive fields. In *ECCV*, pages 363–380. Springer, 2024.

[13] Zhenqi Fu, Ruizhe Chen, Yue Huang, En Cheng, Xinghao Ding, and Kai-Kuang Ma. Masnet: A robust deep marine animal segmentation network. *IEEE Journal of Oceanic Engineering*, 2023.

[14] Wenchao Gu, Shuang Bai, and Lingxing Kong. A review on 2d instance segmentation based on deep neural networks. *Image and Vision Computing*, 120:104401, 2022.

[15] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020.

[16] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pages 22046–22055, 2023.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799. PMLR, 2019.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[20] Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *CVPR*, pages 3205–3215, 2024.

[21] Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

[22] 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 *ICCV*, pages 4015–4026, 2023.

[23] Peiwen Lai, Weizhi Zhong, Yipeng Qin, Xiaohang Ren, Baoyuan Wang, and Guanbin Li. Llm-driven multimodal and multi-identity listening head generation. In *CVPR*, pages 10656–10666, 2025.

[24] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019.

[25] Gongyang Li, Zhi Liu, Dan Zeng, Weisi Lin, and Haibin Ling. Adjacent context coordination network for salient object detection in optical remote sensing images. *IEEE Transactions on Cybernetics*, 53(1):526–538, 2022.

[26] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2303–2314, 2021.

[27] Wenxue Li, Xinyu Xiong, Peng Xia, Lie Ju, and Zongyuan Ge. Tp-drseg: Improving diabetic retinopathy lesion segmentation with explicit text-prompts assisted sam. In *MICCAI*, pages 743–753. Springer, 2024.

[28] Xinye Li and Ding Chen. A survey on deep learning-based panoptic segmentation. *Digital Signal Processing*, 120: 103283, 2022.

[29] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *WACV*, pages 1445–1454, 2022.

[30] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *ICLR*, 2024.

[31] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pages 11591–11601, 2021.

[32] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014.

[33] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, pages 8772–8781, 2021.

[34] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[36] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2160–2170, 2022.

[37] Jialun Pei, Zhangjun Zhou, Yueming Jin, He Tang, and Pheng-Ann Heng. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In *ACM MM*, pages 2139–2147, 2023.

[38] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

[39] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, pages 38–56. Springer, 2022.

[40] Yueqian Quan, Honghui Xu, Renfang Wang, Qiu Guan, and Jianwei Zheng. Orsi salient object detection via progressive semantic flow and uncertainty-aware refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021.

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[44] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, pages 29441–29454. PMLR, 2023.

[45] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished Manuscript*, 2(6):7, 2018.

[46] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *IJCAI*, pages 1025–1031, 2021.

[47] Zhengzheng Tu, Chao Wang, Chenglong Li, Minghao Fan, Haifeng Zhao, and Bin Luo. Orsi salient object detection via multiscale joint region and boundary model. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.

[48] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *CVPRW*, pages 3635–3647, 2024.

[49] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020.

[50] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *ICRA*, pages 6687–6694. IEEE, 2024.

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[52] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020.

[53] Zihuang Wu, Xinyu Xiong, Guangwei Gao, Hongwei Li, and Hua Chen. Hfs-sam2: Segment anything model 2 with high-frequency feature supplementation for camouflaged object detection. *IEEE Signal Processing Letters*, 2025.

[54] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.

[55] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024.

[56] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024.

[57] Wenhao Xu, Shuchen Zheng, Changwei Wang, Zherui Zhang, Chuan Ren, Rongtao Xu, and Shibiao Xu. Samamba: Adaptive state space modeling with hierarchical vision for infrared small target detection. *Information Fusion*, page 103338, 2025.

[58] Yimin Xu, Fan Yang, and Bin Xu. Dsu-net: An improved u-net model based on dinov2 and sam2 with multi-scale cross-model feature enhancement. *arXiv preprint arXiv:2503.21187*, 2025.

[59] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[60] Xiangyu Zeng, Mingzhu Xu, Yijun Hu, Haoyu Tang, Yupeng Hu, and Liqiang Nie. Adaptive edge-aware semantic interaction network for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[61] Qijian Zhang, Runmin Cong, Chongyi Li, Ming-Ming Cheng, Yuming Fang, Xiaochun Cao, Yao Zhao, and Sam Kwong. Dense attention fluid network for salient object detection in optical remote sensing images. *IEEE Transactions on Image Processing*, 30:1305–1317, 2020.

[62] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *International Journal of Computer Vision*, 132(10):4157–4234, 2024.

[63] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3, 2024.

[64] Xiaofei Zhou, Kunye Shen, Zhi Liu, Chen Gong, Jiyong Zhang, and Chenggang Yan. Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[65] Xiaofei Zhou, Kunye Shen, Li Weng, Runmin Cong, Bolun Zheng, Jiyong Zhang, and Chenggang Yan. Edge-guided recurrent positioning network for salient object detection in optical remote sensing images. *IEEE Transactions on Cybernetics*, 53(1):539–552, 2022.

[66] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.