SAMWave: Adapting Segment Anything Model to Difficult Tasks

Saurabh Yadav*[1]
saurabhy@iiitd.ac.in
Avi Gupta*[1]
avig@iiitd.ac.in
Koteswar Rao Jerripothula[2,1]
kotesrj@iitk.ac.in

[1] IIIT Delhi
New Delhi, India
[2] IIT Kanpur
Kanpur, India

Abstract

The emergence of large vision models has propelled significant advances in various domains. The Segment Anything Model (SAM), a leading model for image segmentation, exemplifies these advances, outperforming traditional methods. However, such foundation models often suffer from performance degradation when applied to complex tasks for which they are not trained. Existing methods typically employ adapter-based fine-tuning strategies to adapt SAM for difficult tasks and leverage high-frequency features extracted from the Fourier domain. However, our analysis reveals that these approaches offer limited benefits due to constraints in their feature extraction techniques. To overcome this, we propose ***SAMWave***, a novel and interpretable approach that utilizes the wavelet transform to extract richer, multi-scale high-frequency features from input data. Extending this, we introduce complex-valued adapters capable of capturing complex-valued information via complex wavelet transforms. By adaptively integrating these wavelet coefficients, SAMWave enables SAM's encoder to capture more relevant information for dense prediction. Empirical evaluations on four difficult vision tasks demonstrate that SAMWave significantly outperforms existing adaptation methods. This superior performance is consistent across both the SAM and SAM2 backbones and holds for both real and complex-valued adapter variants, highlighting the efficiency, flexibility, and interpretability of our proposed method for adapting segment anything models. Code is available at https://github.com/saurabhya/SAMWave.

1 Introduction

Large vision models [13, 40, 42] have become the dominant solution for numerous computer vision tasks, significantly advancing fields like semantic segmentation [23, 46] and object detection. However, their substantial parameter counts, while enabling high performance on their trained tasks, lead to sub-optimal generalization to other related problems, a challenge particularly evident in low-level vision tasks. For instance, while SAM (Segment Anything Model) [23] excels at semantic segmentation, it struggles with tasks like camouflaged object detection, polyp detection, defocus blur detection and other such difficult tasks.

*equal contribution

The difficulty of developing a single, universally high-performing model for all such difficult tasks has led to growing interest in adapting large models [3, 28, 29, 34, 44, 62]. A common approach is to use finetuning strategies; however, it faces several challenges, including catastrophic forgetting [1, 4, 6, 17, 51, 54] and the reliance on large-scale datasets for adequate adaptation.

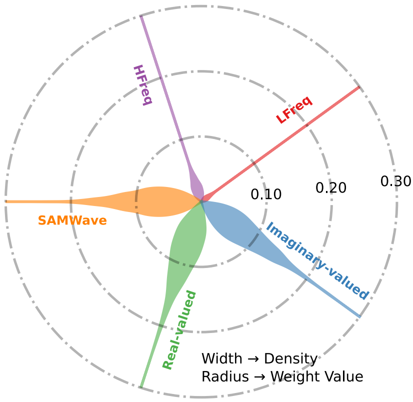

Figure 1: Learned weights for the previous high frequency extraction method, compared with SAMWave's real and complex adapter weights. Unlike the previous method, SAMWave's weights are not concentrated around 0.

In this work, we investigate the optimal adaptation of large models to related tasks, focusing particularly on SAM, which comprises a powerful encoder and a prompt-conditioned decoder. We hypothesize that the encoder can produce a sufficiently rich latent representation, such that the decoder can generate an effective segmentation mask for a new task–given appropriate supervision. Previous works [7, 32] have shown the efficacy of providing high-frequency features to the encoder for adapting a frozen decoder. These methods are typically performed by transforming input images into the Fourier domain and applying a predefined mask to emphasize high-frequency components. Different images possess varying frequency distributions, and a fixed mask leads to suboptimal feature selection. Our empirical study, presented in Tab. 1, supports this observation–demonstrating that simply inverting the mask yields comparable performance across tasks. Fig. 1 further illustrates this issue. This motivates our departure from rigid, predefined Fourier-domain masks toward a more adaptive feature selection mechanism. To this end, we propose a novel high-frequency feature extraction method using wavelet transform, which offers superior spatial and frequency localization [24, 43, 55]. Specifically, we extract horizontal (I_{lh}), vertical (I_{hl}), and diagonal (I_{hh}) features from an image I, combining them into a high-frequency image $I_{HF} = I_{lh} + I_{hl} + I_{hh}$, which provides finer details and texture to the encoder. Wavelet transform is advantageous due to its ability to capture frequency and spatial information, allowing for a more precise representation of image details than Fourier-based methods. We further extend this approach with complex wavelet transforms and complex-valued adapters. Complex-valued representations offer a more comprehensive way to capture image information, potentially leading to improved feature extraction and adaptation, as demonstrated in various applications[59, 65]. We are the first to show the adaptation of large vision models using a complex-valued approach. Our contributions are summarized as follows: ❷ *We demonstrate the limitations of using the Fourier domain with predefined masks for high-frequency feature extraction.* ❷ *We introduce a Wavelet High-Frequency (WHF) module that leverages the wavelet transform for effective finetuning.* ❸ *We propose a complex-valued extension of our method using complex wavelet transform and complex-valued adapters.* ❹ *Our experiments show significant performance improvements over existing tuning methods for SAM and SAM2 across four difficult tasks.*

2 Related Works

Visual Adapters: The paradigm of adapting large pre-trained models for downstream tasks using adapter-based finetuning, initially popular in NLP [18], has effectively extended to computer vision [9, 30]. Adapter-based fine-tuning is a highly effective form of parameter-efficient fine-tuning. Instead of updating all the weights of a massive pre-trained model, this approach freezes the original model and inserts small, learnable neural network modules called adapters between its layers. During training for a new downstream task, only the weights of these newly added adapters are updated. This method offers a compelling balance between performance and computational efficiency. Recent work has shown its applicability to vision foundation models like SAM [23], notably through visual prompting methods [7, 32]. While prior adaptation methods often prioritize global context, our work focuses on incorporating crucial local, high-frequency information via wavelets to enhance SAM's performance on low-level vision tasks.

Low-level Vision Tasks: Our work addresses several low-level vision tasks where fine-grained details are critical for accurate localization. **Camouflaged Object Detection (COD)** involves identifying objects that are seamlessly blended into their background, a significant challenge due to the low contrast and texture similarity between the object and its surroundings [15, 16, 27]. Success in this task requires models to meticulously capture subtle boundary cues and texture discontinuities [41, 66]. **Defocus Blur Detection (DBD)** aims to precisely segment out-of-focus regions in an image. This relies on detecting subtle variations in blurriness across the scene, where the transition from sharp to blurry areas, often characterized by local edge information, is a key indicator for segmentation [22, 50, 58]. **Shadow Detection (SD)** requires distinguishing shadows from actual objects or dark surfaces by analyzing illumination patterns and local context [19, 73]. The task is challenging as shadows can alter an object's appearance, and accurately identifying their soft or sharp boundaries is crucial for robust scene understanding. **PolyP Detection (PD)** focuses on the precise delineation of polyp regions from endoscopic images, a critical step for assisting clinicians in the early detection of colorectal abnormalities [48, 61]. The task is demanding due to the highly variable shape, size, and subtle appearance of polyps against the surrounding tissue.

While some methods [32] use generic high-frequency components for low-level tasks, we propose a more principled approach using specific wavelet decompositions to provide targeted local feature guidance to SAM for these diverse low-level vision challenges.

3 Methodology

In this section, we describe our proposed network: SAMWave. The idea is to leverage the high-frequency features obtained from the wavelet transform and use them to guide the Segment-Anything model (SAM) [23] for more precise segmentation in various tasks.

Table 1: Result of low frequency (LFreq) & high frequency (HFreq) on camouflaged object detection on three separate datasets. We also show the difference in the performance of both approaches.

	CHAMELEON				CAMO				COD10K			
	$\mathcal{S}_m \uparrow$	$\mathcal{E}_\phi \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{E}_\phi \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{E}_\phi \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$
HFreq	**0.896**	**0.919**	0.824	**0.033**	**0.847**	**0.873**	**0.765**	**0.070**	**0.883**	**0.918**	0.801	**0.025**
LFreq	0.888	0.903	**0.830**	0.034	0.824	0.843	**0.765**	0.078	0.872	0.901	**0.803**	0.027
	+ 0.008	+ 0.016	- 0.006	+ 0.001	+ 0.023	+ 0.030	0.000	+ 0.008	+ 0.009	+ 0.017	- 0.002	+ 0.002

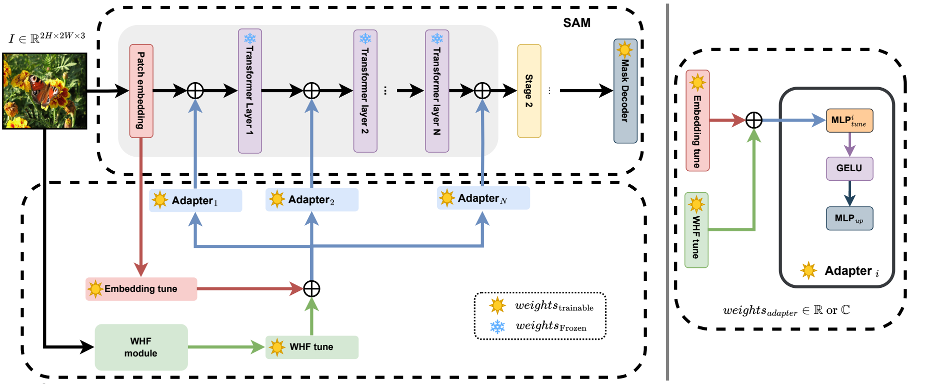

Figure 2: Overview of SAMWave for difficult vision tasks. **Left** shows that the image is passed through SAM and our WHF (wavelet high frequency) module, where the high-frequency features are extracted. The extracted details are then combined with patch embeddings obtained from the transformer encoder. Note that the encoder is frozen; we only updated the decoder for our training. **Right** shows the overview of a single adapter used; for complex adapters, we replace the weights of MLP layers with complex values.

3.1 Motivation for High-Frequency Feature Utilization

Precise segmentation, particularly the accurate delineation of object boundaries, is a challenging task that fundamentally relies on capturing fine-grained image details. These details are predominantly encoded within the high-frequency components of the image signal. Therefore, effectively leveraging high-frequency features is crucial for achieving high-quality segmentation masks.

Previous methods aiming to incorporate high-freqency information, such as those in [7, 32], have often relied on extracting these features in the Fourier domain using a predefined, fixed mask. However, a significant limitation of this *"one mask fits all"* approach is its inability to adapt to the diverse characteristics and varying distributions of high-frequency information across different images and tasks. Consequently, a static Fourier mask may not effectively isolate the relevant high-frequency cues necessary for robust fine-grained segmentation in all scenarios.

We conducted an empirical study to investigate the efficacy of this fixed-mask approach for capturing essential high-frequency information. We compared the performance using features extracted via the standard high-frequency mask (HFreq) against those extracted by inverting this mask to obtain a low-frequency mask (LFreq). Our experiments focused on the challenging task of camouflaged object detection, and the results are summarized in Tab. 1. While HFreq shows superior performance on *9* out of *12* metrics compared to LFreq, a closer examination reveals that the average performance gain offered by HFreq is remarkably small, merely **0.009**. Despite high frequencies theoretically being more aligned with boundary information, this marginal difference suggests that a fixed Fourier mask is not an optimal strategy for extracting discriminative high-frequency features that significantly enhance segmentation performance across varied data. This observation motivates the need for a more adaptive and practical approach to harness high-frequency information to guide models like SAM towards more precise segmentation.

3.2 SAM as the Backbone

Our SAMWave framework employs the Segment-Anything Model (SAM) [23] as its core backbone, leveraging its robust pre-trained capabilities for various segmentation tasks. SAM comprises three main components: an image encoder, a prompt encoder, and a mask decoder.

The image encoder, a pre-trained ViT-H/16, processes the input image $I \in \mathbb{R}^{H\times W\times 3}$ to produce dense embeddings $E_{\text{img}} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 256}$. These embeddings are spatially downscaled by $16\times$ and have 256 channels after a 1×1 convolution. The prompt encoder processes input prompts (such as dense features from our method), typically using convolutions, and these prompt embeddings are element-wise added to E_{img}. The mask decoder, a lightweight transformer-based network inspired by [5, 11], takes the combined embeddings and predicts the final segmentation mask via upsampling and a linear classifier.

In SAMWave, we freeze the weights of the pre-trained image encoder and train only the parameters of the mask decoder to adapt SAM's segmentation generation to the task-specific guidance.

3.3 Wavelet High-Frequency Module (WHF)

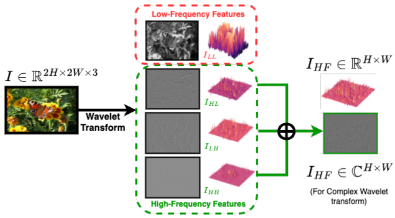

Figure 3: An overview of the WHF module, we first use the wavelet transform to obtain high- and low-frequency features. Then, we combine the high-frequency features to create a feature map. For complex-valued features, the methodology remains the same but utilizes a complex wavelet transform for the initial decomposition step.

The Wavelet High-Frequency (WHF) module extracts rich high-frequency spatial features using the Discrete Wavelet Transform (DWT). Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, we first resize it to $I' \in \mathbb{R}^{2H\times 2W\times 3}$. This upsampling compensates for the DWT's inherent downsampling, ensuring output subbands match the original $H\times W$ resolution.

Applying a 2D DWT to I' decomposes it into four subbands at $H\times W\times 3$ resolution: I'_{ll} (approximation), I'_{lh} (horizontal details), I'_{hl} (vertical details), and I'_{hh} (diagonal details). $I' \xrightarrow{\text{DWT}} \{I'_{ll}, I'_{lh}, I'_{hl}, I'_{hh}\}$ Focusing on fine details, we exclude the low-frequency I'_{ll} subband and combine the high-frequency detail subbands $(I'_{lh}, I'_{hl}, I'_{hh})$ to form the composite high-frequency map I_{HF}. $I_{HF} = I'_{lh} + I'_{hl} + I'_{hh} \in \mathbb{R}^{H\times W\times 3}$ captures edge and texture information (Fig. 3).

A key advantage of this wavelet-based approach is its inherent flexibility. While we describe the process using standard real-valued DWT, the WHF module can seamlessly incorporate complex wavelet transforms, such as the Dual-Tree Complex Wavelet Transform (DT-CWT)[33]. The core procedural steps—resizing, wavelet decomposition into subbands, selection and combination of high-frequency subbands, and linear projection—remain identical. The only change lies in the specific wavelet basis functions used in the decomposition, allowing for adaptation to properties captured by complex wavelets[47] (*e.g.*, phase information, shift invariance) without altering the module's overall structure.

3.4 SAMWave Architecture

SAMWave enhances SAM's segmentation capabilities by integrating learned high-frequency wavelet features *via* lightweight adapters. As illustrated in Fig. 2, the architecture consists of three key modules–Embedding Tune, WHF Tune, and Adapters–which are integrated into the SAM backbone.
Embedding Tune: This module generates a learned spatial feature map $\psi_{PE} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times\gamma}$ from SAM's image patch embeddings $\mathbf{F}_{\text{emb}}$ using a learnable MLP layer $\phi_{PE}(.;\theta_{PE})$:

$$\psi_{PE} = \phi_{PE}(\mathbf{F}_{\text{emb}};\theta_{PE})$$

WHF Tune: Complementary to ψ_{PE}, this module processes the high-frequency feature map I_{HF} from the WHF module (§3.3). Depending on the wavelet, I_{HF} can be real or complex. Since the features I_{HF} are 2D spatial features, they are divided into patches and projected using a convolution layer to obtain ψ_{whf}, matching ψ_{PE}'s dimensions.
Real/Complex Adapters: Adapters are placed within SAM's image encoder layers. For layer i, the adapter input is the channel concatenation $\Psi = [\psi_{PE}, \psi_{whf}] \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 2\gamma}$ (or complex). The output $\mathbf{P}^{(i)} = \phi_{up}^{(i)}(\sigma(\phi_{tune}^{(i)}(\Psi)))$, with $\mathbf{P}^{(i)} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_{img}^{i}}$.

$\psi_{tune}^{(i)}$ and $\psi_{up}^{(i)}$ are MLPs (standard or complex), and the activation function (σ) is GeLU (or split GeLU for complex). $\mathbf{P}^{(i)}$ is then added with the layer's feature map $\mathbf{F}_{img}^{i}$.

$$\overline{\mathbf{F}}_{img}^{(i)} = \mathbf{F}_{img}^{(i)} + \mathbf{P}^{(i)}$$

In case of complex adapters, the real and imaginary parts of $\mathbf{P}^{(i)}$ are both added separately to $\mathbf{F}_{img}$. The $\mathbf{P}^{(i)}$ feature is obtained for each layer separately and then added to the respective layer's feature map. It provides robustness to learn features required for adaptation.
SAMWave benefits from: 1) **Feature Complementarity:** combining SAM's semantic features with explicitly learned high-frequency details; 2) **Wavelet Flexibility:** easily integrates various wavelets, including complex ones with $\mathbb{C}$-Adapters to leverage properties like shift invariance and phase information for improved robustness on challenging tasks.

3.5 Objective Function

Following [7, 32], we optimize the tasks in SAMWave by minimizing the loss between the predicted binary mask ($\hat{m}$) and ground truth (m). Specifically, we use binary cross-entropy loss ($\mathcal{L}_{bce}$)[12] for defocus blur detection, balanced binary cross-entropy loss ($\mathcal{L}_{bbce}$) for shadow detection, and summation of binary cross-entropy and intersection-over-union loss ($\mathcal{L}_{bbce} + \mathcal{L}_{iou}$) for camouflaged object detection. The loss functions are formulated below as follows:

$$\mathcal{L}_{bce}(\hat{m},m) = -\big(m * log(\hat{m}) + (1-m) * log(1-\hat{m})\big) \tag{1}$$

$$\mathcal{L}_{bbce}(\hat{m},m) = -\zeta * \big(m * log(\hat{m}) + (1-m) * log(1-\hat{m})\big) \tag{2}$$

$$\zeta = \frac{\sum m + \varepsilon}{\sum m + \sum(1-m) + \varepsilon} \tag{3}$$

$$\mathcal{L}_{iou}(\hat{m},m) = 1 - \frac{\sum(\hat{m} * m)}{\sum(\hat{m}+m) - \sum(\hat{m} * m)} \tag{4}$$

Table 2: Comparison of Camouflaged Object Detection performance on three datasets against task-specific and finetuning methods. Our SAMwave approach demonstrates improved performance across all datasets compared to existing methods.(■: best, ■: second best)

Method	CHAMELEON $\mathcal{S}_m \uparrow$	CHAMELEON $\mathcal{E}_\phi \uparrow$	CHAMELEON $\mathcal{F}_\beta^w \uparrow$	CHAMELEON $\mathcal{M} \downarrow$	CAMO $\mathcal{S}_m \uparrow$	CAMO $\mathcal{E}_\phi \uparrow$	CAMO $\mathcal{F}_\beta^w \uparrow$	CAMO $\mathcal{M} \downarrow$	COD10K $\mathcal{S}_m \uparrow$	COD10K $\mathcal{E}_\phi \uparrow$	COD10K $\mathcal{F}_\beta^w \uparrow$	COD10K $\mathcal{M} \downarrow$
SINet[15]	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
RankNet[36]	0.846	0.913	0.767	0.045	0.712	0.791	0.583	0.104	0.767	0.861	0.611	0.045
JCOD[26]	0.870	0.924	-	0.039	0.792	0.830	-	0.082	0.800	0.872	-	0.041
PFNet[38]	0.882	0.942	0.810	0.033	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040
FBNet[31]	0.888	0.939	0.828	0.032	0.783	0.839	0.702	0.081	0.809	0.889	0.684	0.035
FSPNet[20]	-	-	-	-	0.856	0.899	0.799	0.050	0.851	0.895	0.735	0.026
SAM[23]	0.727	0.734	0. 639	0.081	0.684	0.687	0.606	0.132	0.783	0.798	0.701	0.050
SAM2[45]	0.359	0.375	0.115	0.357	0.350	0.411	0.079	0.311	0.429	0.505	0.115	0.218
EVP[32]	0.871	0.917	0.795	0.036	0.846	0.895	0.777	0.059	0.843	0.907	0.742	0.029
SAM-Adapter[7]	0.896	0.919	0.824	0.033	0.847	0.873	0.765	0.070	0.883	0.918	0.801	0.025
SAM2-Adapter[8]	0.915	0.955	**0.889**	0.018	0.855	0.909	0.810	0.051	0.899	0.950	0.850	0.018
SAM as Backbone												
Real-Valued Adapters												
Ours (daubechies (db))	0.922	0.947	0.856	0.023	0.865	0.903	0.790	0.057	0.900	0.936	0.824	0.021
Ours (coiflet (cf))	0.923	0.952	**0.890**	0.022	0.856	0.888	0.807	0.060	**0.903**	0.939	0.857	0.019
Ours (haar (ha))	0.923	0.954	0.866	0.023	0.868	0.905	0.798	0.056	0.899	0.934	0.819	0.022
Ours (symlet (sym))	0.915	0.947	0.845	0.027	0.846	0.880	0.777	0.067	0.897	0.938	0.822	0.021
Complex-Valued Adapters												
Ours (Symmetric-b (sy-b))	**0.931**	0.961	0.883	0.020	0.856	0.896	0.797	0.058	0.902	0.940	0.843	0.019
Ours (Symmetric-a (sy-a))	**0.928**	0.957	0.881	0.019	0.858	0.893	0.800	0.057	0.902	0.939	0.841	0.019
SAM2 as Backbone												
Real-Valued Adapters												
Ours (daubechies (db))	0.912	0.952	0.880	0.018	**0.870**	**0.927**	0.828	**0.045**	**0.905**	**0.956**	0.861	**0.016**
Ours (coiflet (cf))	0.917	**0.963**	0.883	**0.017**	0.866	0.923	0.824	0.048	**0.905**	**0.957**	**0.862**	**0.016**
Ours (haar (ha))	0.918	0.962	**0.889**	**0.017**	0.869	0.926	**0.829**	**0.047**	0.900	0.952	**0.865**	**0.017**
Ours (symlet (sym))	0.917	0.961	0.887	**0.017**	**0.871**	**0.929**	**0.831**	**0.045**	**0.905**	0.954	0.859	**0.016**
Complex-Valued Adapters												
Ours (Symmetric-b (sy-b))	0.917	**0.966**	**0.890**	**0.016**	0.864	0.921	0.824	0.049	0.902	0.955	0.858	**0.017**
Ours (Symmetric-a (sy-a))	0.917	0.960	0.881	0.019	0.864	0.921	0.826	0.048	**0.903**	0.955	0.856	**0.016**

4 Experimental Details and Results

We evaluate our approach on four difficult vision tasks: camouflaged object detection, shadow detection, defocus blur detection, and polyp detection, and compare the performance with current state-of-the-art methods. We also perform ablation studies to analyze the effectiveness of our approach.

4.1 Datasets and Evaluation Metrics

Camouflaged Object Detection. Following [7, 32], we evaluate on COD10K [15], CAMO [25], and CHAMELEON [53]. Metrics: S-measure ($\mathcal{S}_m$) [14], mean E-measure ($\mathcal{E}_\phi$), weighted F-measure ($\mathcal{F}_\beta^w$) [37], and Mean Absolute Error ($\mathcal{M}$).

Shadow Detection. Following [7, 32], we evaluate on SBU [63] and ISTD [64]. Metric: Balanced Error Rate (BER).

Defocus Blur Detection. Following [7, 32], we evaluate on CUHK [49] and DUT [67]. Metrics: F-measure ($\mathcal{F}_\beta$) and Mean Absolute Error ($\mathcal{M}$).

PolyP Detection. Following [61], we evaluate on Kvasir-SEG [21], CVC-ClinicDB [2], ColonDB [56], ETIS [52]. Metrics: $mDice$ and $mIoU$.

Additional details about the datasets, including their train-test split, are provided in the supplementary.

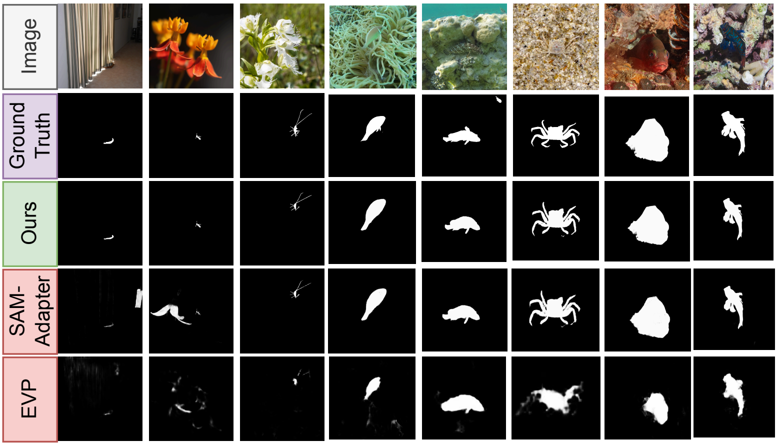

Figure 4: Visual comparison of our proposed SAMWave with existing methods for the COD task.

4.2 Implementation Details

All experiments are implemented in PyTorch on an NVIDIA A100 GPU. We follow a similar training setting as [7, 32] with an image size of 1024×1024. We use pre-trained SAM and SAM2 networks as the backbone. We use the AdamW optimizer for weight updates and use a learning rate of 0.0002.

4.3 Experimental Results

Camouflaged Object Detection: We compare our approach with existing task-specific methods [31, 35, 39] and efficient tuning methods [7, 32]. Tab. 2 shows that SAMWave significantly outperforms these methods across three datasets, consistently with SAM and SAM2 backbones. To demonstrate the efficacy of our adaptive approach, we show that performance improvement over previous methods is consistent across different wavelets. Despite vanilla SAM2 performing poorly compared to SAM and other methods in Tab. 2, its performance is significantly boosted when combined with our approach. We provide the qualitative results in Fig. 4.

Table 3: Comparison of state-of-the-art approaches on shadow detection.(■: best, ■: second best)

Dataset	DSD [illegible]	FDRNet [illegible]	MTMT [illegible]	EVP [32]	SAM-Adapter[7]	SAM2-Adapter[8]	Ours: SAM as Backbone, Real-valued Adapters				SAM as Backbone, Complex-valued Adapters		SAM2 as Backbone, Real-valued Adapters				SAM2 as Backbone, Complex-valued Adapters	
							db	cf	ha	sym	sy-b	sy-a	db	cf	ha	sym	sy-b	sy-a
ISTD ($BER\downarrow$)	2.17	1.55	1.72	1.35	1.43	1.43	1.24	1.24	**1.01**	1.04	1.09	1.15	1.91	1.24	**0.92**	1.22	2.60	1.32
SBU ($BER\downarrow$)	3.45	3.04	3.15	4.31	-	-	**2.93**	3.01	3.00	3.07	3.06	3.06	3.04	5.28	**2.99**	3.77	4.43	4.32

Shadow Detection: We compare the performance of our approach against existing methods across two benchmark datasets. In both cases, our method yields substantial performance improvements. Although the performance improvements are relatively minor on the ISTD dataset when using Daubechies and Coiflet wavelets compared to other wavelet variants, the improvement remains significant, demonstrating the robustness and generalizability of our adaptive approach.

Table 4: Comparison with state-of-the-art approaches on blur detection.(■: best, ■: second best)

Dataset	Metric	DeFusion-Net[illegible]	BTBNet [illegible]	CENet [58]	DAD [70]	EFENet [59]	EVP [32]	Ours SAM Real-valued db	SAM Real-valued cf	SAM Real-valued ha	SAM Real-valued sym	SAM Complex-valued sy-b	SAM Complex-valued sy-a	SAM2 Real-valued db	SAM2 Real-valued cf	SAM2 Real-valued ha	SAM2 Real-valued sym	SAM2 Complex-valued sy-b	SAM2 Complex-valued sy-a
DUT	$\mathcal{F}_\beta \uparrow$	0.823	0.827	0.817	0.794	0.854	0.890	0.888	0.886	0.880	0.885	0.892	0.893	0.896	0.893	0.894	0.893	**0.907**	**0.902**
DUT	$\mathcal{M} \downarrow$	0.118	0.138	0.135	0.153	0.094	0.068	0.068	**0.067**	0.108	0.106	**0.067**	0.066	**0.051**	**0.051**	**0.051**	**0.051**	**0.052**	0.056
CHUK	$\mathcal{F}_\beta \uparrow$	0.818	0.889	0.906	0.884	0.914	**0.928**	0.922	0.916	0.911	0.911	0.917	0.912	0.923	0.924	**0.925**	0.924	**0.928**	**0.925**
CHUK	$\mathcal{M} \downarrow$	0.117	0.082	0.059	0.079	0.053	0.045	0.040	0.046	0.070	0.069	0.045	0.046	**0.036**	0.039	**0.037**	0.038	**0.036**	0.039

Defocus Blur Detection: For the Blur Detection task, we primarily compare our method with EVP[32], as SAM-Adapter[7] does not report results for this task. As shown in Tab. 4, our approach achieves the best results using complex-valued adapters with the SAM2 backbone. While we outperform EVP in all cases, our method demonstrates significant improvements over other existing approaches as well.

Table 5: Comparison with state-of-the-art methods on PolyP detection. (■: best, ■: second best)

Dataset	Metric	M²UNet [60]	SAM Adapter [7]	SAM2 Adapter [8]	SAM-EG [61]	Ours SAM Real-valued db	SAM Real-valued cf	SAM Real-valued ha	SAM Real-valued sym	SAM Complex-valued sy-b	SAM Complex-valued sy-a	SAM2 Real-valued db	SAM2 Real-valued cf	SAM2 Real-valued ha	SAM2 Real-valued sym	SAM2 Complex-valued sy-b	SAM2 Complex-valued sy-a
ClinicDB	*mDice* ↑	**0.901**	-	-	**0.931**	0.855	0.652	0.866	0.854	0.832	0.840	0.887	0.881	0.885	0.874	0.862	0.846
ClinicDB	*mIoU* ↑	**0.853**	-	-	**0.879**	0.789	0.449	0.794	0.786	0.761	0.772	0.826	0.821	0.827	0.816	0.800	0.784
ColonDB	*mDice* ↑	0.767	-	-	0.774	0.743	0.635	0.724	0.737	0.726	0.748	0.764	0.769	0.769	0.778	**0.781**	**0.782**
ColonDB	*mIoU* ↑	0.684	-	-	0.689	0.668	0.386	0.647	0.661	0.652	0.671	0.696	0.703	0.697	0.704	**0.710**	**0.707**
Kvasir	*mDice* ↑	0.907	0.850	0.873	**0.915**	0.889	0.812	0.901	0.900	0.901	0.902	**0.917**	0.913	0.911	0.908	0.899	0.902
Kvasir	*mIoU* ↑	0.855	0.776	0.806	**0.862**	0.818	0.618	0.827	0.828	0.834	0.838	**0.867**	0.861	0.859	0.860	0.844	0.847
ETIS	*mDice* ↑	0.670	-	-	0.757	0.688	0.445	0.685	0.693	0.680	0.732	0.780	0.749	0.784	**0.802**	**0.797**	0.750
ETIS	*mIoU* ↑	0.595	-	-	0.681	0.610	0.223	0.603	0.622	0.611	0.654	0.709	0.683	0.720	**0.737**	**0.730**	0.676

PolyP Detection: Tab. 5 presents quantitative results for the PolyP detection task. Compared to existing approaches[7, 8, 60, 61], our method achieves significant improvements across nearly all datasets. We observe further gains when using the SAM2 backbone, especially with complex-valued adapters. While our performance is comparable to that of [61], it is important to emphasize that–unlike their task-specific model–our adaptive approach is designed to generalize across multiple low-level vision tasks.
$\mathbb{R}$-valued *vs* $\mathbb{C}$-valued Adapters: Tab. 2, 3, 4, and 5 show that, when using the SAM backbone, complex-valued adapters perform comparably to—or in some cases better than—their real-valued adapters. Furthermore, when switching to the SAM2 backbone, complex-valued adapters consistently yield improved results, highlighting the enhanced compatibility and representational capacity of SAM2 for this task.

4.4 Ablation Study

We conduct three ablation studies to highlight our contribution. First, we remove the WHF module and adapters from the SAM backbone while keeping everything else the same, and then, we provide each of the wavelet transform subbands. Lastly, we replace the wavelets used in our approach with other wavelets and observe the result.

Effect of adapters: We remove high-frequency features and patch embeddings from the backbone and train the SAM model. Tab. 6 shows the result for the same. Although this

Table 6: The results of our ablation study highlight the importance of adapters for efficient finetuning; it also highlights the importance of each low-frequency subband (*LL*) and the rest of the high-frequency subbands (*HL*, *LH*, *HH*) in the image in the wavelet domain. The ablation experiment is done for camouflage object detection, and SAM is taken as the backbone.

Method				CHAMELEON				CAMO				COD10K			
				$\mathcal{S}_m \uparrow$	$\mathcal{E}_\phi \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{E}_\phi \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{E}_\phi \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$
SAM Finetuned				0.796	0.802	0.676	0.062	0.750	0.756	0.639	0.105	0.789	0.817	0.596	0.049
LL	*HL*	*LH*	*HH*	Evaluating each component											
✔	✗	✗	✗	0.907	0.937	0.842	0.029	**0.856**	0.884	0.785	0.062	0.891	0.929	0.817	0.024
✗	✔	✗	✗	0.895	0.927	0.834	0.030	0.851	0.882	0.781	0.063	0.887	0.924	0.816	0.024
✗	✗	✔	✗	0.896	0.915	0.838	0.030	0.854	0.887	0.790	0.063	0.891	0.928	0.838	0.023
✗	✗	✗	✔	0.908	0.935	0.848	0.028	0.853	0.883	0.783	0.064	0.893	0.928	0.818	0.023
✗	✔	✔	✔	**0.923**	**0.952**	**0.890**	**0.022**	**0.856**	**0.888**	**0.807**	**0.060**	**0.903**	**0.939**	**0.857**	**0.019**

helped improve performance over SAM, it lags significantly.

Effect of each subband: In §3.3, we claim that when combining all three high-frequency subbands (*HL*, *LH*, *HH*), there is a significant improvement in the performance. Hence, we perform an ablation study by considering each subband individually and observing its significance to assert our claim. As expected, we see performance degradation for all three datasets when provided with low-frequency features (*LL*). Similarly, when we provide each high-frequency subband, we see a similar pattern in performance loss. However, we only see the best performance when all three are combined.

Effect of each wavelet: To ensure that the results obtained are not due to any specific property of a single wavelet, we use four different wavelets and show their results. The comparative result are shown in Tab. 2, 3, 4, 5. The observation is similar across all four tasks; we see more performance gain for some wavelets, but overall, using any wavelet provides better information than previous methods.

5 Conclusion

This work presents a new efficient fine-tuning method for large vision models, particularly SAM and SAM2, for solving four difficult vision tasks. Additionally, we show that a complex-valued adapter can be used to fine-tune large vision models efficiently. We first show that the existing method of providing high-frequency information to the network is not optimal and needs improvement. We propose a WHF module, which converts the input image to a wavelet domain and combines the high-frequency subbands. When provided to the networks, this new feature map significantly improves over existing methods.

Acknowledgement

This work was supported in part by the Infosys Centre for Artificial Intelligence, IIIT-Delhi, and in part by Junior Research Fellowship (JRF) scheme of the University Grants Commission (UGC).

References

[1] Sara Babakniya, Zalan Fabian, Chaoyang He, Mahdi Soltanolkotabi, and Salman Avestimehr. A data-free approach to mitigate catastrophic forgetting in federated class

incremental learning for vision tasks. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d160ea01902c33e30660851dfbac5980-Abstract-Conference.html.

[2] Jorge Bernal, Francisco Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez de Miguel, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Medical Imaging Graph.*, 43:99–111, 2015. doi: 10.1016/J.COMPMEDIMAG.2015.02.007. URL https://doi.org/10.1016/j.compmedimag.2015.02.007.

[3] David Brüggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring relational context for multi-task dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15849–15858. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01557. URL https://doi.org/10.1109/ICCV48922.2021.01557.

[4] Xinzi Cao, Xiawu Zheng, Guanhong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16880–16889, June 2024.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13. URL https://doi.org/10.1007/978-3-030-58452-8_13.

[6] Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ho-Shing Ip, and Sam Kwong. Saving 100x storage: Prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/708e0d691a22212e1e373dc8779cbe53-Abstract-Conference.html.

[7] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, 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, and more, 2023.

[8] Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting

segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more, 2024. URL https://arxiv.org/abs/2408.04579.

[9] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=plKu2GByCNW.

[10] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5610–5619. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00565. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_A_Multi-Task_Mean_Teacher_for_Semi-Supervised_Shadow_Detection_CVPR_2020_paper.html.

[11] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021.

[12] Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134(1):19–67, 2005. doi: 10.1007/S10479-005-5724-Z. URL https://doi.org/10.1007/s10479-005-5724-z.

[13] 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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[14] 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 (ICCV)*, Oct 2017.

[15] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] Avi Gupta, Koteswar Rao Jerripothula, and Tammam Tillo. CIRCOD: co-saliency inspired referring camouflaged object discovery. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pages 8313–8323. IEEE, 2025. doi: 10.1109/WACV61041.2025.00806. URL https://doi.org/10.1109/WACV61041.2025.00806.

[17] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

URL http://papers.nips.cc/paper_files/paper/2023/hash/376276a95781fa17c177b1ccdd0a03ac-Abstract-Conference.html.

[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 Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

[19] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7454–7462. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00778. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Direction-Aware_Spatial_Context_CVPR_2018_paper.html.

[20] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5557–5566. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00538. URL https://doi.org/10.1109/CVPR52729.2023.00538.

[21] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve, editors, *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II*, volume 11962 of *Lecture Notes in Computer Science*, pages 451–462. Springer, 2020. doi: 10.1007/978-3-030-37734-2_37. URL https://doi.org/10.1007/978-3-030-37734-2_37.

[22] Ali Karaali and Cláudio R. Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Trans. Image Process.*, 27(3):1126–1137, 2018. doi: 10.1109/TIP.2017.2771563. URL https://doi.org/10.1109/TIP.2017.2771563.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://doi.org/10.1109/ICCV51070.2023.00371.

[24] Cansu Korkmaz, A. Murat Tekalp, and Zafer Dogan. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. *CoRR*, abs/2402.19215, 2024. doi: 10.48550/ARXIV.2402.19215. URL https://doi.org/10.48550/arXiv.2402.19215.

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

[26] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10071–10081. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00994. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Uncertainty-Aware_Joint_Salient_Object_and_Camouflaged_Object_Detection_CVPR_2021_paper.html.

[27] Aixuan Li, Jing Zhang, Yunqiu Lyu, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[28] Jin Li, Yaoming Wang, Xiaopeng Zhang, Bowen Shi, Dongsheng Jiang, Chenglin Li, Wenrui Dai, Hongkai Xiong, and Qi Tian. Ailurus: A scalable vit framework for dense prediction. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/62c9aa4d48329a85d1e36d5b6d0a6a32-Abstract-Conference.html.

[29] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Learning multiple dense prediction tasks from partially annotated data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18857–18867. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01831. URL https://doi.org/10.1109/CVPR52688.2022.01831.

[30] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 280–296. Springer, 2022. doi: 10.1007/978-3-031-20077-9_17. URL https://doi.org/10.1007/978-3-031-20077-9_17.

[31] Jiaying Lin, Xin Tan, Ke Xu, Lizhuang Ma, and Rynson W. H. Lau. Frequency-aware camouflaged object detection. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(2): 61:1–61:16, 2023. doi: 10.1145/3545609. URL https://doi.org/10.1145/3545609.

[32] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19434–19445. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01862. URL https://doi.org/10.1109/CVPR52729.2023.01862.

[33] Ange Lou, Benjamin Planche, Zhongpai Gao, Yamin Li, Tianyu Luan, Hao Ding, Terrence Chen, Jack Noble, and Ziyan Wu. Darenerf: Direction-aware representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5031–5042, June 2024.

[34] Yuxiang Lu, Shalayiding Sirejiding, Yue Ding, Chunlin Wang, and Hongtao Lu. Prompt guided transformer for multi-task dense prediction. *IEEE Trans. Multim.*, 26: 6375–6385, 2024. doi: 10.1109/TMM.2024.3349865. URL https://doi.org/10.1109/TMM.2024.3349865.

[35] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11591–11601. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01142. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lv_Simultaneously_Localize_Segment_and_Rank_the_Camouflaged_Objects_CVPR_2021_paper.html.

[36] Yunqiu Lyu, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[37] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[38] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8772–8781. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00866. URL https://openaccess.thecvf.com/content/CVPR2021/html/Mei_Camouflaged_Object_Segmentation_With_Distraction_Mining_CVPR_2021_paper.html.

[39] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8772–8781, June 2021.

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2160–2170, June 2022.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[43] Michaël Ramamonjisoa, Michael Firman, Jamie Watson, Vincent Lepetit, and Daniyar Turmukhambetov. Single image depth prediction with wavelet decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11089–11098. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01094. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ramamonjisoa_Single_Image_Depth_Prediction_With_Wavelet_Decomposition_CVPR_2021_paper.html.

[44] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12159–12168. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01196. URL https://doi.org/10.1109/ICCV48922.2021.01196.

[45] 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 Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

[46] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

[47] Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G. Baraniuk. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18507–18516, June 2023.

[48] Vanshali Sharma, Abhishek Kumar, Debesh Jha, Manas Kamal Bhuyan, Pradip K. Das, and Ulas Bagci. Controlpolypnet: Towards controlled colon polyp synthesis

for improved polyp segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 2325–2334. IEEE, 2024. doi: 10.1109/CVPRW63382.2024.00238. URL https://doi.org/10.1109/CVPRW63382.2024.00238.

[49] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 2965–2972. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.379. URL https://doi.org/10.1109/CVPR.2014.379.

[50] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 657–665. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298665. URL https://doi.org/10.1109/CVPR.2015.7298665.

[51] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2990–2999, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/0efbe98067c6c73dba1250d2beaa81f9-Abstract.html.

[52] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283—293, March 2014. ISSN 1861-6410. doi: 10.1007/s11548-013-0926-3. URL https://hal.archives-ouvertes.fr/hal-00843459/file/IJCARS_R2_vf.pdf.

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

[54] Xiang Song, Kuang Shu, Songlin Dong, Jie Cheng, Xing Wei, and Yihong Gong. Overcoming catastrophic forgetting for multi-label class-incremental learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 2378–2387. IEEE, 2024. doi: 10.1109/WACV57701.2024.00238. URL https://doi.org/10.1109/WACV57701.2024.00238.

[55] Gilbert Strang and Truong Q. Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, 1997. ISBN 978-0-9614088-7-9.

[56] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Medical Imaging*, 35(2):630–644, 2016. doi: 10.1109/TMI.2015.2487997. URL https://doi.org/10.1109/TMI.2015.2487997.

[57] Chang Tang, Xinzhong Zhu, Xinwang Liu, Lizhe Wang, and Albert Y. Zomaya. Defusionnet: Defocus blur detection via recurrently fusing and refining multi-scale

deep features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2700–2709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00281. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Tang_DeFusionNET_Defocus_Blur_Detection_via_Recurrently_Fusing_and_Refining_Multi-Scale_CVPR_2019_paper.html.

[58] Chang Tang, Xinzhong Zhu, Xinwang Liu, Lizhe Wang, and Albert Y. Zomaya. Defusionnet: Defocus blur detection via recurrently fusing and refining multi-scale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2700–2709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00281. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Tang_DeFusionNET_Defocus_Blur_Detection_via_Recurrently_Fusing_and_Refining_Multi-Scale_CVPR_2019_paper.html.

[59] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep complex networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1T2hmZAb.

[60] Quoc-Huy Trinh, Nhat-Tan Bui, Trong-Hieu Nguyen Mau, Minh-Van Nguyen, Hai-Minh Phan, Minh-Triet Tran, and Hai-Dang Nguyen. M2unet: Metaformer multi-scale upsampling network for polyp segmentation. In *31st European Signal Processing Conference, EUSIPCO 2023, Helsinki, Finland, September 4-8, 2023*, pages 1115–1119. IEEE, 2023. doi: 10.23919/EUSIPCO58844.2023.10290110. URL https://doi.org/10.23919/EUSIPCO58844.2023.10290110.

[61] Quoc-Huy Trinh, Hai-Dang Nguyen, Nguyen Ngoc Bao Tram, Debesh Jha, Ulas Bagci, and Minh-Triet Tran. Sam-eg: Segment anything model with egde guidance framework for efficient polyp segmentation. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. URL https://papers.bmvc2024.org/0472.pdf.

[62] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7):3614–3633, 2022. doi: 10.1109/TPAMI.2021.3054719. URL https://doi.org/10.1109/TPAMI.2021.3054719.

[63] T. F. Y. Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *European Conference on Computer Vision*, 2016. URL https://api.semanticscholar.org/CorpusID:17623309.

[64] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[65] Saurabh Yadav and Koteswar Rao Jerripothula. Fccns: Fully complex-valued convolutional networks using complex-valued color model and loss function. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 10655–10664. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00981. URL https://doi.org/10.1109/ICCV51070.2023.00981.

[66] Bowen Yin, Xuying Zhang, Qibin Hou, Bo-Yuan Sun, Deng-Ping Fan, and Luc Van Gool. Camoformer: Masked separable attention for camouflaged object detection. *arXiv preprint arXiv:2212.06570*, 2022.

[67] Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3080–3088. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00325. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhao_Defocus_Blur_Detection_CVPR_2018_paper.html.

[68] Wenda Zhao, Bowen Zheng, Qiuhua Lin, and Huchuan Lu. Enhancing diversity of defocus blur detectors via cross-ensemble network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8905–8913. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00911. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zhao_Enhancing_Diversity_of_Defocus_Blur_Detectors_via_Cross-Ensemble_Network_CVPR_2019_paper.html.

[69] Wenda Zhao, Xueqing Hou, You He, and Huchuan Lu. Defocus blur detection via boosting diversity of deep ensemble networks. *IEEE Trans. Image Process.*, 30: 5426–5438, 2021. doi: 10.1109/TIP.2021.3084101. URL https://doi.org/10.1109/TIP.2021.3084101.

[70] Wenda Zhao, Cai Shang, and Huchuan Lu. Self-generated defocus blur detection via dual adversarial discriminators. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6933–6942. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00686. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Self-Generated_Defocus_Blur_Detection_via_Dual_Adversarial_Discriminators_CVPR_2021_paper.html.

[71] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. Distraction-aware shadow detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5167–5176. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00531. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zheng_Distraction-Aware_Shadow_Detection_CVPR_2019_paper.html.

[72] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson W. H. Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada,*

October 10-17, 2021, pages 4682–4691. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00466. URL https://doi.org/10.1109/ICCV48922.2021.00466.

[73] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson W. H. Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 4682–4691. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00466. URL https://doi.org/10.1109/ICCV48922.2021.00466.