Omni-Dimensional State Space Model-driven SAM for Pixel-level Anomaly Detection

Chao Huang[1,2] , **Qianyi Li**[3] , **Jie Wen**[4] , **Bob Zhang**[1] *
[1]University of Macau
[2]Sun Yat-sen University, Shenzhen Campus
[3]Ningbo University
[4]Harbin Institute of Technology, Shenzhen
huangchao_08@126.com, 2849736080@qq.com, jiewen_pr@126.com, bobzhang@um.edu.mo

Abstract

Pixel-level anomaly detection is indispensable in industrial defect detection and medical diagnosis. Recently, Segment Anything Model (SAM) has achieved promising results in many vision tasks. However, direct application of the SAM to pixel-level anomaly detection tasks results in unsatisfactory performance, meanwhile SAM needs the manual prompt. Although some automatically prompt-based SAM has been proposed, these automated prompting approaches merely utilize partial image features as prompts and fail to incorporate crucial features such as multi-scale image features to generate more suitable prompts. In this paper, we propose a novel Omni Dimensional State Space Model-driven SAM (ODS-SAM) for pixel-level anomaly detection. Specifically, the proposed method adopts the SAM architecture, ensuring easy implementation and avoiding the need for fine-tuning. A State-Space Model-based residual Omni Dimensional module is designed to automatically generate suitable prompts. This module can effectively leverage multi-scale and global information, facilitating an iterative search for optimal prompts in the prompt space. The identified optimal prompts are then fed into SAM as high-dimensional tensors. Experimental results demonstrate that the proposed ODS-SAM outperforms state-of-the-art models on both industrial and medical image datasets.

1 Introduction

Pixel-level anomaly detection [Huang *et al.*, 2022d; Huang *et al.*, 2024b] is a technique that precisely identifies anomaly pixels in images, which has emerged as an indispensable tool for industrial defect detection and medical diagnosis. However, its reliance on the expertise and proficiency of individual analysts poses a significant challenge. Consequently, it is necessary that develop an automated method to replace the previously labor-intensive and technically demanding manual detection processes [Huang *et al.*, 2022b].

Deep learning models have emerged as promising tools for pixel-level anomaly detection due to their ability to learn complex image features. However, the widespread prevalence of Pixel-level anomaly detection algorithms tailored to specific imaging modalities or objectives hinders progress. While the specialization of existing methods can achieve high accuracy in controlled settings, it comes at the cost of limited adaptability and generalizability across diverse anomaly scenarios [Mamonov *et al.*, 2014]. Vision Foundation Models (VFMs) have captivated the attention of researchers due to their remarkable generalization capabilities across diverse downstream tasks. Among these, the Segment Anything Model (SAM) [Kirillov *et al.*, 2023] has emerged as a pioneering contribution to image segmentation due to the widespread acclaim for its ability to generate precise object masks in both fully automated and interactive scenarios. However, pixel-level anomaly detection, a critical subfield of image segmentation, poses unique challenges that may impede the performance of SAM in this field. This stems from the inherent disparities between natural images and medical images. Some works [Li *et al.*, 2025; Liu *et al.*, 2024] have demonstrated that the direct application of the SAM to pixel-level anomaly detection tasks results in unsatisfactory performance. Therefore, it's necessary to perform specific improvements and optimizations to SAM so that improve the performance and generalization ability of the model in pixel level anomaly detection tasks.

Moreover, the work of [Cai *et al.*, 2024] has shown the significance of multi-scale features for pixel-level anomaly detection. Nevertheless, we found that the original SAM lacks a multi-scale feature extraction component. This is due to SAM being an interactive segmentation model that only requires segmenting objects in a specified region. However, when dealing with an image containing numerous anomaly targets, performing interactive segmentation for each target individually becomes inefficient. Consequently, some methods [Xie *et al.*, 2024; Shaharabany *et al.*, 2023] have been proposed to automatically generate SAM prompts without manual intervention. Nonetheless, these automated prompting approaches merely utilize partial image features as prompts and fail to incorporate crucial features such as multi-scale image features to generate more suitable prompts. This limitation results in the generated prompts that cannot fully capture anomalous targets in the image, thereby affecting the accuracy and

*Corresponding author

robustness of anomaly detection. Therefore, the challenge faced in this article is how to design a prompt encoder that can effectively integrate multi-scale features as prompts to improve its ability to detect anomalies in complex scenes.

To address this challenge, we propose a novel end-to-end method for pixel-level anomaly detection in industrial and medical images. Specifically, we propose a novel prompt encoder to generate prompts based on the input image fed into SAM. This contrasts with the rudimentary prompts (point, bounding box, and mask) adopted in the vanilla SAM architecture. We propose a State-Space Model-based residual Omni-Dimensional module (SSMODC) for the prompt encoder. This module can effectively leverage multi-scale and global information, facilitating an iterative search for optimal prompts in the prompt space. The identified optimal prompts are then fed into SAM as high-dimensional tensors. Notably, our method does not alter the SAM, which ensures ease of implementation and avoids finding an optimal training schedule for fine-tuning SAM. Extensive experiments demonstrate that with our proposed prompt encoder, the detection performance of the SAM is significantly improved. It surpasses state-of-the-art models.

The contributions of this work can be summarized as:

1) We propose a novel Omni Dimensional State Space Model-driven SAM (ODS-SAM) for pixel-level anomaly detection. Specifically, ODS-SAM utilizes an innovative prompt encoder to adaptively generate prompts based on input images, effectively integrating multi-scale features. This adaptive prompt generation mechanism significantly improves the quality of prompts and the accuracy of detection.

2) We design a State-Space Model-based residual Omni-Dimensional module (SSMODC), which effectively utilizes multi-scale and global information to iteratively search for the optimal prompt in the prompt space. The introduction of the SSMODC module enables the model to comprehensively understand and respond to the features of the input image, improving the accuracy and robustness of anomaly detection.

3) Experimental results on medical and industrial image datasets demonstrate that our proposed ODS-SAM outperforms several state-of-the-art methods on the task of pixel-level anomaly detection.

2 Related Work

2.1 Pixel-level Anomaly Detection

Pixel-level anomaly detection is an anomaly detection[Zhang *et al.*, 2022; Huang *et al.*, 2021a; Huang *et al.*, 2021b; Wang *et al.*, 2025; Huang *et al.*, 2022c; Huang *et al.*, 2022a; Huang *et al.*, 2024c; Huang *et al.*, 2025] task aimed at accurately identifying abnormal regions in an image, which has attracted attention from both academia and industry due to its broad application prospects. Initial efforts [Bae and Yoon, 2015; Reiss *et al.*, 2021] employed elemental features like texture, color, shape, and appearance, but their performance is often hindered by the inherent limitations of these low-level features. The advent of deep learning [Chen *et al.*, 2020] revolutionized the pixel-level anomaly detection task, and deep neural network-based methods achieve remarkable progress. Notably, encoder-decoder architecture networks, exemplified by U-Net [Ronneberger *et al.*, 2015] and the enhanced version U-Net++ [Zhou *et al.*, 2018], have demonstrated significant performance gains by leveraging multi-level feature integration to generate high-resolution detection results. Additionally, researchers have further enhanced the effectiveness of anomaly detection models by incorporating supplementary boundary information [Fang *et al.*, 2019]. These methods represent the conventional deep learning approaches. In contrast, the proposed method leverages the powerful generalization and spatial awareness capabilities of the visual foundation model SAM, thus achieving accurate pixel-level anomaly localization and detection without the need for additional information.

2.2 Segment Anything Model

SAM (Segment Anything Model, SAM) [Kirillov *et al.*, 2023] is a groundbreaking prompt-based image segmentation model. It is trained on the extensive SA-1B dataset. This dataset has an astounding tens of millions of image-annotation pairs. It grants the model exceptional zero-shot generalization capabilities. SAM adopts a Transformer-based architecture, demonstrating remarkable effectiveness in natural language processing and image recognition tasks. Specifically, SAM employs a Vision Transformer (ViT)-based image encoder [Dosovitskiy, 2020] to extract image embeddings, a prompt encoder to integrate user interaction through diverse prompt modes, and a lightweight mask decoder to predict segmentation masks by fusing image embeddings and prompt embeddings. However, existing SAM requires relatively high-quality prompts (i.e., points, boxes, and masks) to achieve satisfactory performance in segmentation tasks. In previous methods, the required prompts are generated from ground-truth labels during testing [Wei *et al.*, 2023]. Nonetheless, creating accurate and reliable prompts still necessitates specific domain expertise, which may not be readily accessible. Moreover, low-quality prompts arising from noisy annotations can significantly deteriorate segmentation accuracy. Consequently, exploring automatic prompt mechanisms aims to establish a robust adaptive framework to alleviate the performance of SAM variability and facilitate more reliable and accurate results across various segmentation tasks. This paper proposes a prompt encoder that generates prompts tailored to the given prompt for SAM input images. Specifically, it extracts multi-scale features, crucial for pixel-level anomaly detection tasks, as prompts for SAM. Additionally, the prompt encoder can identify the optimal prompt mode by optimizing parameters during training.

3 Methodxiangology

3.1 Preliminaries: State Space Model

State Space Model (SSM) originates from continuous systems that transform a sequence $x(t) \rightarrow y(t)$ through a hidden state function $h(t) \in \mathbb{R}^N$. It can be expressed as the following equations:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{1}$$

$$y(t) = \mathbf{C}h(t), \tag{2}$$

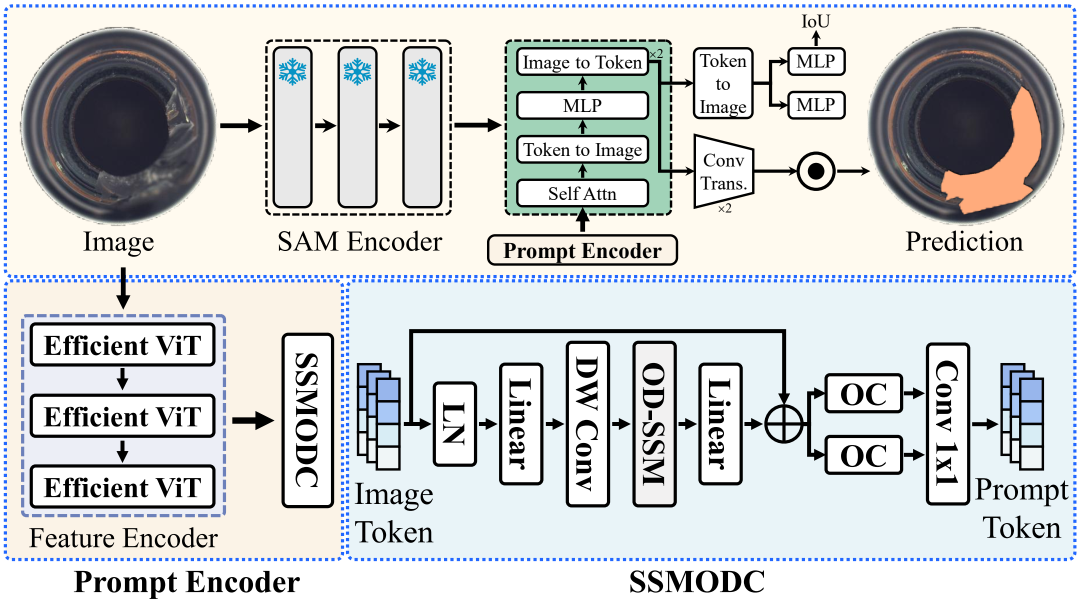

Figure 1: Overview of the proposed method. First, given the input image, the SAM encoder is adopted to extract the vision feature. Additionally, we adopt a feature encoder containing n Efficient Vit block to extract the feature of the abnormal object. Subsequently, the proposed SSMODC is adopted to generate suitable prompts, which are generated by leveraging both global and multi-scale information. Finally, both prompt and visual features are fed into the SAM decoder to generate the final anomaly mask. The "OC" is the omni-dimensional convolution and the "Dw Conv" is the depthwise convolution.

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is a state matrix, and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{N}$ are the projection matrix.

S4 and Mamba [Gu and Dao, 2023] are discrete versions of the aforementioned continuous system, which adopt a time scale parameter $\mathbf{\Delta A}$ and $\mathbf{\Delta B}$ that converts continuous parameters $\mathbf{A}, \mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}, \overline{\mathbf{B}}$. The zero-order hold is adopted as the discretization rule, defined as follows:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \tag{3}$$

$$\overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{J}) \cdot \Delta\mathbf{B}, \tag{4}$$

where $\mathbf{J}$ is a unit matrix.

After discretizing $\mathbf{A}$ and $\mathbf{B}$, Eq. 1 can be written as follows:

$$h'(t) = \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t), \tag{5}$$

$$y(t) = \mathbf{C}h(t). \tag{6}$$

Furthermore, a global convolution is adopted to compute the output, defined as:

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{AB}}, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \tag{7}$$

$$y = x * \overline{\mathbf{K}}, \tag{8}$$

where L denotes the length of the input sequence and $\mathbf{K} \in \mathbb{R}^{L}$ is a structured convolution kernel.

3.2 Overview

The proposed method is shown in Fig 1. Specifically, it consists of three parts, the SAM encoder Enc_{SAM}, the proposed prompt encoder Enc_{p}, and the SAM decoder Dec_{SAM}. Enc_{SAM} is adopted to extract visual features. We freeze for both components during the training phase to maintain the powerful feature extraction capabilities of Enc_{SAM} and the remarkable spatial reasoning prowess of Dec_{SAM}. Subsequently, the proposed State-Space Model-based residual Omni-Dimensional module is adopted to comprehensively explore the global information of anomalous targets from multiple directions. Moreover, it captures the edge details of anomalies through multi-scale representation. Thus, the prompt encoder can deeply understand anomalous features at different scales while iteratively searching for the optimal prompt in the prompt space, achieving more accurate anomaly detection results. Finally, both prompt and visual features are fed into the SAM decoder to generate the final anomaly mask.

3.3 State-Space Model-based residual Omni-Dimensional module

Previous research has demonstrated that leveraging multi-scale representations to capture contextual information can effectively improve edge segmentation accuracy. Furthermore, global information offers more holistic scene understanding, which plays a pivotal role in detecting the overall structure of anomalous objects. However, current pixel-level anomaly detection methods fail to integrate global and rich contextual information containing multi-scale information effectively. Thus it is hard to accurately segment anomalous objects in complex scenarios, particularly while both local details and global context demand comprehensive consideration. To address these limitations, we propose the State-Space Model-based residual Omni-Dimensional module, which encodes these two types of information as prompts and feeds them into the SAM model for pixel-level anomaly detection.

Specifically, we adopt the state-space model (SSM) to extract global information. SSM can process sequence data with linear complexity, which significantly improves the efficiency of sequence processing. However, SSM is hard to capture global information due to the fixed scanning direction. Furthermore, SSMs exhibit a propensity to forget previous inputs, causing the model to prioritize recent inputs during processing at the expense of early information in the sequence. This tendency impedes the model's ability to adequately focus on the middle region of images in image segmentation tasks, potentially leading to the oversight of anomalous information. To address these limitations, we introduce the Omni-directional State-Space Model (OD-SSM). OD-SSM improves the extraction of global information by scanning the input sequence in multiple directions, which can simultaneously pay attention to the middle region and corners of the image. Consequently, it can effectively avoid the omission of anomalous information. Additionally, we adopted a depthwise separable convolution (DW Conv) layer before OD-SSM to retain local details. It can extract near-neighbor features before modeling global information. Depthwise separable convolution can effectively capture salient features in local regions. Notably, this refinement does not impede the proposed method extract global information, which ensures the integration of local details with global information. As shown in Fig 2, we adopt Omni-Dimensional Dynamic Convolution (OD Conv) [Li *et al.*, 2022] with different dilated rates to extract multi-scale information. OD Conv incorporates a multi-dimensional dynamic attention mechanism to optimize convolution kernel shapes and enhance feature sensitivity. This mechanism empowers convolution kernels to adaptively focus on crucial information within input data, which bolsters the representative capacity of the model. Moreover, the model effectively captures multi-scale features by adopting ODConv with varying dilation rates, which can learn rich representations from local to global contexts. This complements the global information extracted by SSM and provides the decoder with more comprehensive prompts.

3.4 Loss Function

We adopt a composite loss function in training. This function combined the Binary Cross-Entropy (BCE) loss and Dice loss in a weighted manner. This combined function can balance the prediction accuracy and boundary alignment, which improves training efficiency and segmentation accuracy for pixel-level classification tasks.

The BCE and Dice function can be defined as follows:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right], \quad (9)$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}\hat{y}_i y_i + \epsilon}{\sum_{i=1}^{N}\hat{y}_i + \sum_{i=1}^{N} y_i + \epsilon} \quad (10)$$

where N is the total number of samples, y_i is the ground truth label for the i-th pixel, and $\hat{y}_i$ is the predicted probability for the i-th pixel. Additionally, ϵ is a small constant added for numerical stability.

The final combined loss function $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{Dice} + \lambda_2 \mathcal{L}_{BCE}. \quad (11)$$

Based on experiments, we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$.

4 Experiments

4.1 Dataset

MVTec AD [Bergmann *et al.*, 2020] is adopted to evaluate the proposed method for pixel-level anomaly detection in industrial scenes. It is a large-scale dataset for industrial anomaly detection, which contains 1258 abnormal images in 15 categories. The MoNuSeg dataset [Kumar *et al.*, 2017] comprises 30 microscopy images from 7 organs in the training set, with annotations for 21,623 individual nuclei. To align with prior work, we resized the all images to 512 × 512. The Gland segmentation (GlaS) challenge [Sirinukunwattana *et al.*, 2017] encompasses 85 images for training and 80 images for testing. All images in experiments are resized to 224 × 224. Moreover, four Polyp datasets (Kvasir-SEG [Jha *et al.*, 2020], ClinicDB [Bernal *et al.*, 2015], ColonDB [Tajbakhsh *et al.*, 2015], and ETIS [Silva *et al.*, 2014]) are adopted to evaluate the proposed method. The experiment setup is followed [Fan *et al.*, 2020].

4.2 Implementation Details

During the training, we adopted the Adam with an initial learning rate of 0.001 and a weight decay regularization parameter of 1e-5. The batch size is 10 and trained on an NVIDIA A100 GPU. The maximum number of epochs is 200. SAM pre-trained weights adopted in all experiments is based on ViT-H.

4.3 Evaluation Metrics

To align with prior research, we evaluate the performance of all methods on the MVTec dataset using the following two metrics: mean E-measure (mE) [Fan *et al.*, 2021] and mean absolute error (MAE). Additionally, we assess the performance of all methods on the GlaS, MoNusEG, and four polyp datasets using Dice and IoU [Li *et al.*, 2023] metrics.

4.4 Comparison Methods

On the medical image datasets, comparison methods include U-Net [Ronneberger *et al.*, 2015], U-Net++ [Zhou *et al.*, 2018], SFA [Fang *et al.*, 2019], MSEG [Huang *et al.*, 2021c], DCRNet [Yin *et al.*, 2022], ACSNet [Liu *et al.*, 2021], PraNet [Fan *et al.*, 2020], EU Net [Patel *et al.*, 2021], SANet [Wei *et al.*, 2021], COMMA [Shin *et al.*, 2022], SAM-EG [Trinh *et al.*, 2024], LViT [Li *et al.*, 2023], and methods from literature [Huang *et al.*, 2024a], Axial Attention [Wang *et al.*, 2020], MedT [Valanarasu *et al.*, 2021], UCTransNet [Wang *et al.*, 2022], 3P-SEG [Shaharabany and Wolf, 2022], MedAdaptor-SAM [Wu *et al.*, 2023]. The comparison methods on industrial images include SFA [Fang *et al.*, 2019], ACSNet [Zhang *et al.*, 2020], PraNet [Fan *et al.*, 2020], and methods from literature [Huang *et al.*, 2022b], AutoSAM [Shaharabany *et al.*, 2023], I-MedSAM [Wei *et al.*, 2023].

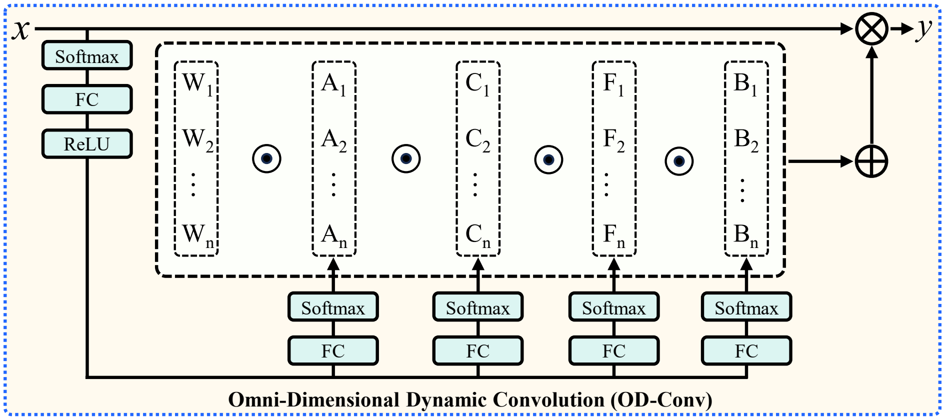

Figure 2: The overview of the Omni-Dimensional Dynamic Convolution (OD-Conv).

Method		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
U-Net	MAE ↓	0.010	0.029	0.006	0.016	0.042	0.011	0.018	0.012	0.005	0.004	0.020	0.037	0.009	0.028	0.011	0.014
	mE ↑	0.679	0.629	0.562	0.915	0.719	0.727	0.866	0.903	0.811	0.431	0.801	0.620	0.882	0.832	0.562	0.724
U-Net++	MAE ↓	0.009	0.022	0.005	0.013	0.043	0.010	0.015	0.010	0.004	0.004	0.014	0.040	0.009	0.020	0.010	0.012
	mE ↑	0.701	0.742	0.641	0.933	0.566	0.818	0.874	0.911	0.857	0.508	0.877	0.560	0.910	0.928	0.690	0.775
SFA	MAE ↓	0.031	0.083	0.025	0.062	0.133	0.031	0.031	0.077	0.015	0.007	0.095	0.045	0.009	0.037	0.023	0.041
	mE ↑	0.735	0.726	0.552	0.768	0.596	0.690	0.832	0.583	0.737	0.750	0.747	0.738	**0.941**	0.883	0.718	0.739
ACSNet	MAE ↓	0.022	0.037	0.006	0.016	0.758	0.011	0.013	0.009	**0.003**	0.003	0.022	0.033	0.008	0.029	0.009	0.036
	mE ↑	0.772	0.825	0.788	0.961	0.189	0.848	**0.938**	0.960	**0.940**	0.833	0.809	0.786	0.923	0.898	0.775	0.840
PraNet	MAE ↓	0.010	0.024	0.008	0.024	0.034	0.011	0.018	0.015	0.005	0.006	0.021	0.042	0.011	0.037	0.011	0.015
	mE ↑	0.826	0.828	0.808	0.907	0.616	0.885	0.899	0.902	0.886	0.729	0.885	0.692	0.930	0.850	0.818	0.844
Huang et al.	MAE ↓	0.006	**0.018**	**0.004**	0.012	0.032	0.008	0.013	0.008	0.004	0.003	0.013	0.027	0.009	0.017	0.010	0.011
	mE ↑	0.850	**0.881**	0.765	0.924	0.838	0.855	0.900	0.939	0.878	0.780	0.930	0.729	0.920	0.940	0.806	0.860
AutoSAM	MAE ↓	0.004	0.022	0.017	0.014	0.063	**0.007**	0.016	0.006	0.004	0.005	0.012	0.007	0.008	0.018	0.007	0.014
	mE ↑	0.900	0.791	0.429	0.954	0.390	0.900	0.860	0.906	0.915	0.789	0.939	0.806	0.918	**0.948**	0.859	0.820
I-MedSAM	MAE ↓	0.006	0.024	0.010	0.012	0.105	0.007	0.017	0.007	0.004	0.002	0.011	0.007	0.007	0.019	0.006	0.016
	mE ↑	0.878	0.831	0.841	0.954	0.590	**0.912**	0.880	**0.930**	0.904	0.903	0.947	0.839	0.930	0.947	0.842	0.875
Our	MAE ↓	**0.004**	0.020	0.007	**0.011**	**0.023**	0.010	**0.012**	**0.006**	0.004	**0.002**	**0.010**	**0.003**	**0.007**	**0.017**	**0.005**	**0.009**
	mE ↑	**0.947**	0.833	**0.850**	**0.968**	**0.913**	0.893	0.892	0.908	0.914	**0.905**	**0.950**	**0.876**	0.914	0.936	**0.865**	**0.904**

Table 1: Quantitative results on MVTec AD. The digits indicate categories: 'pill', 'cable', 'capsule', 'tile', 'transistor', 'carpet', 'wood', 'hazelnut', 'leather', 'screw', 'metal nut', 'toothbrush', 'zipper', 'bottle', and 'grid', respectively. ↑ (or ↓) indicates that the higher (or the lower) the better.

4.5 Comparison with State-of-the-Art Methods

Result on MVTec Dataset

As shown in Table 1, the proposed method outperforms all existing SOTA methods on the MVTec AD dataset. Specifically, the proposed method achieves the best mean MAE and mean squared error (mE) results. The method achieves an MAE of 0.009, which outperforms the SOTA method [Huang *et al.*, 2022b] by 18.2%. Moreover, our method increases mE by 5.1% compared to the SOTA method [Huang *et al.*, 2022b]. These results demonstrate that the proposed method can accurately detect pixel-level anomalies on industrial datasets.

Result on Medical Image Dataset

Table 2 and Table 3 demonstrate that the proposed method outperforms other methods across all medical image datasets. Specifically, compared with existing SOTA methode SAM-EG, our method achieved 0.6% and 1.4% improvement in Dice and IoU scores on the Kvasir33 dataset, respectively. Surprisingly, our method achieves significant improvement, with Dice and IoU scores increasing by 3.68% and 5.19% on the MoNuSeg dataset, respectively. These results underscore the superiority of our method for pixel-level anomaly detection in medical images.

Method	Kvasir33		Clinic		Colon		ETIS	
	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)
U-Net	81.8	74.6	82.3	75.5	51.2	44.4	39.8	33.5
U-Net++	82.1	74.3	79.4	72.9	48.3	41.0	40.1	34.4
SFA	72.3	61.1	70	60.7	46.9	34.7	29.7	21.7
MSEG	89.7	83.9	90.9	86.4	73.5	66.6	70.0	63.0
DCRNet	88.6	82.5	89.6	84.4	70.4	63.1	55.6	49.6
ACSNet	89.8	83.8	88.2	82.6	71.6	64.9	57.8	50.9
PraNet	89.8	84.0	89.9	84.9	71.2	64.0	62.8	56.7
EU-Net	90.8	85.4	90.2	84.6	75.6	68.1	68.7	60.9
SANet	90.4	84.7	91.6	85.9	75.3	67.0	75.0	65.4
COMMA	90.4	86.0	91.6	87.1	75.4	68.9	71.1	64.8
SAM-EG	91.5	86.2	**93.1**	**87.9**	77.4	68.9	75.7	68.1
Ours	**92.1**	**87.6**	**93.0**	**87.5**	**79.3**	**71.3**	**78.6**	**71.5**

Table 2: Quantitative comparison of the proposed and other SOTA methods on Kvasir33, Clinic, Colon, and ETIS datasets.

Method	MoNuSeg		GlaS	
	Dice(%)	IoU(%)	Dice(%)	IoU(%)
FCN	28.84	28.71	-	-
U-Net	79.43	65.99	75.12	75.12
U-Net++	79.49	66.04	79.03	79.03
Axial Attention	76.83	62.49	-	-
MedT	79.55	66.17	88.85	78.93
FCN-Hardnet	79.52	66.06	89.37	82.09
UCTransNet	79.87	66.68	89.84	82.24
3P-SEG	80.13	67.09	91.19	84.34
MedAdaptor-SAM	80.34	67.33	92.02	85.88
Our	**84.02**	**72.52**	**92.74**	**87.01**

Table 3: Quantitative comparison of the proposed method and other SOTA methods on MoNuSeg and GlaS datasets

SAM	OD-SSM	OD Conv	MoNuSeg		GlaS	
			Dice(%)	IoU(%)	Dice(%)	IoU(%)
✓			82.43	70.17	92.10	86.02
✓	✓		82.99	71.01	92.58	86.62
✓		✓	83.36	71.59	92.60	86.81
✓	✓	✓	**84.02**	**72.52**	**92.74**	**87.01**

Table 4: Effectiveness of each proposed component on MoNuSeg and GlaS

4.6 Ablation Studies

Effectiveness of Each Component

Table 4 reports the ablation experimental results about the effectiveness of each component. These results indicate that all of these components are effective for the proposed method. Our ablation study aimed to assess the contributions of OD-SSM and multi-scale OD Conv to medical image segmentation. When employed individually, OD-SSM led to modest performance gains on MoNuSeg and GlaS, validating its effectiveness. OD-SSM enhances global context extraction by scanning the input sequence multi-directionally. Simultaneously, it ensures comprehensive image coverage, which enhances the perception ability of anomalies.

Furthermore, the exclusive utilization of multi-scale OD Conv results in substantial performance improvements. It indicates that multi-scale representations can improve edge segmentation accuracy due to the multi-scale feature contains richer contextual information.

The combination of OD-SSM and OD Conv yielded the best performance, demonstrating a synergistic effect. Extracting global context can provide a more comprehensive scene understanding, which improves the ability of the model to parse the overall structure of anomalies. Meanwhile, OD Conv effectively integrated global and multi-scale contextual information, enhancing segmentation accuracy in complex scenarios.

Impact of different OD Conv quantities

Table 5 presents an ablation study for varying numbers of OD Conv. Specifically, the model exhibited optimal performance when equipped with two ODDC convolutional layers. This superior performance can be attributed to the following reasons: firstly, the integration of two ODDC layers could provide a balanced capacity for feature extraction. Secondly, the dual-layer configuration may enhance the model to capture multi-scale and multi-dimensional features, which is essential for the nuanced detection of anomalies within the pixel-level domain. Furthermore, the results suggest that the addition of more than two ODDC layers does not significantly improve the performance of the model, which is due to the diminishing returns on the complexity added to the network architecture.

OD Conv	MoNuSeg		GlaS	
	Dice(%)	IoU(%)	Dice(%)	IoU(%)
1	83.63	71.97	92.37	86.38
2	**84.02**	**72.52**	**92.74**	**87.01**
3	82.52	70.32	92.63	86.93

Table 5: Impact of different OD Conv quantities on MoNuSeg and GlaS

Number	MoNuSeg		GlaS	
	Dice(%)	IoU(%)	Dice(%)	IoU(%)
1	**84.02**	**72.52**	**92.74**	**87.01**
2	83.21	71.35	92.61	86.77
3	83.14	71.22	92.53	86.66

Table 6: Impact of different OD-SSM quantities on MoNuSeg and GlaS

Methods	Time complexity (GFLOPs)	Parameters (10^6)
AutoSAM	80.314	88.569
I-MedSAM	648.060	92.520
Our ODSSAM	**53.902**	**53.687**

Table 7: Time and space complexity analysis.

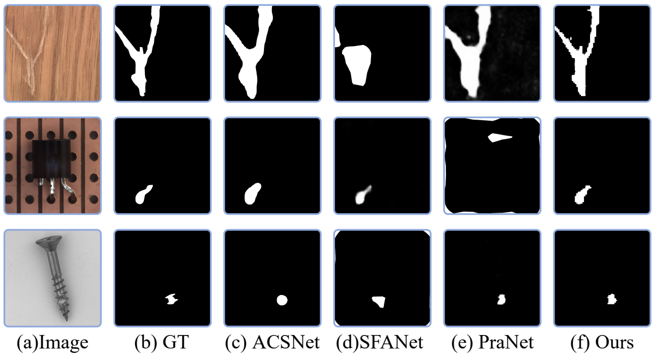

Figure 3: Visualization results of different models on MVTec dataset.

Impact of Different OD-SSM Quantities

In the ablation study, we further explored the impact of the number of Omni-directional State-Space Models (OD-SSMs) on the performance of pixel-level anomaly detection. The results are shown in Table 6. We constructed model variants with 1, 2, and 3 layers of OD-SSMs on the MoNuSeg and GlaS datasets, respectively. It demonstrates that when the model incorporates a single OD-SSM layer, it achieves Dice coefficients of 84.02% on MoNuSeg and 92.74% on GlaS, corresponding to IoU scores of 72.52% and 87.01%, respectively. These results suggest that a solitary OD-SSM layer is sufficient to capture anomalous features within images and produce reasonably accurate anomaly segmentation. Increasing the number of OD-SSM layers to two resulted in a slight decrease in the Dice coefficient on MoNuSeg to 83.21% and a corresponding decrease in IoU to 71.35%. On GlaS, the Dice coefficient was 92.61% and the IoU was 86.77%. These findings suggest that while an additional OD-SSM layer can provide supplementary global information, it offers diminishing returns in terms of performance improvement, and may even introduce a marginal performance degradation.

Time and Space Complexity Analysis

A comparative analysis of time complexity was conducted between the proposed method and two existing automatic prompt SAMs, AutoSAM and I-MedSAM. The results are shown in Table 7. Due to the strategy of freezing SAM image encoder and decoder adopted by each method, we focus on comparing the time complexity of learnable modules. Computational complexity (measured in GFLOPs) was chosen as the metric, as it directly reflects the number of floating-point operations required by the algorithm. In addition, the space complexity of learnable modules is provided for comparison in the form of parameter quantities. It can be seen that the proposed method is significantly lower than AutoSAM in terms of the number of parameters and computational complexity of the learnable module, reducing them by 39.4% and 32.9%. This result indicates that our method can achieve better anomaly detection results at lower complexity.

Visualization Results

Figures 4 and 5 show the visualization results of our method on the MVTec AD dataset and medical image dataset, respectively. The visualization results show that the proposed method can achieve excellent pixel-level anomaly detection performance in diverse and challenging industrial and medical image scenarios. This achievement is attributed to the carefully constructed modules in this work. Specifically, the SSMODC module effectively searches and determines the best prompt for SAM to use. This method maintains the parameters of each component of SAM unchanged during the training process, aiming to preserve SAM's excellent spatial perception ability, thus enabling high-precision anomaly detection on different datasets.

5 Conclusion

This work presents a novel pixel-level anomaly detection approach based on the Segment Anything Model (SAM), enhanced by the integration of the State-Space Model-based residual Omni-Dimensional module (SSMODC). Unlike conventional pixel-level anomaly detection techniques, our approach eliminates the need for manual prompt generation and the fine-tuning of SAM, thereby streamlining the implementation process while maintaining the generalization capabilities of the model. It can accurately capture anomalous regions within images by combining multi-scale features and global information, especially in complex scenarios. Extensive experimental results confirms the superior performance of our method on multiple datasets, particularly on the MVTec AD, MoNuSeg, and GlaS datasets, where it outperforms existing state-of-the-art methods across various metrics, including mean Absolute Error (MAE) and mean E-measure (mE). Ablation studies further substantiate the effectiveness of the proposed SSMODC module and multi-scale Omni-Dimensional Convolution.

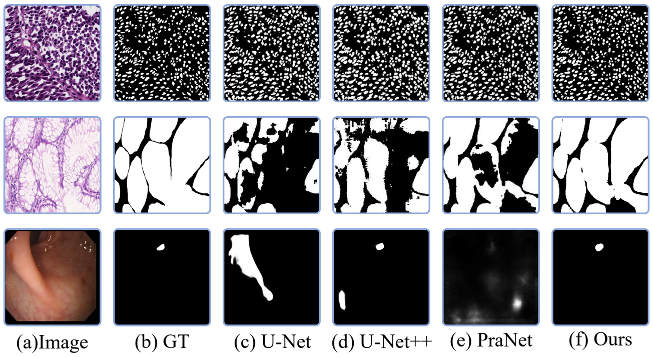

Figure 4: Visualization results of different models on Medical Image dataset.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62301621, in part by Shenzhen Science and Technology Program (No. 20231121172359002), in part by Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011398), and in part by Science and Technology Development Fund, Macao S.A.R (FDCT) 0028/2023/RIA1.

References

[Bae and Yoon, 2015] Seung-Hwan Bae and Kuk-Jin Yoon. Polyp detection via imbalanced learning and discriminative feature learning. *IEEE TMI*, 34(11):2379–2393, 2015.

[Bergmann *et al.*, 2020] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, pages 4183–4192, 2020.

[Bernal *et al.*, 2015] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[Cai *et al.*, 2024] Weichao Cai, Weiliang Huang, Lin Tian, Chao Huang, and Jingwen Yan. Multi-scale global attention for abnormal geological hazard segmentation. *IEEE Sensors Journal*, 2024.

[Chen *et al.*, 2020] Zhi Chen, Jingjing Li, Yadan Luo, Zi Huang, and Yang Yang. Canzsl: Cycle-consistent adversarial networks for zero-shot learning from natural language. In *WACV*, pages 874–883, 2020.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fan *et al.*, 2020] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273. Springer, 2020.

[Fan *et al.*, 2021] 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.

[Fang *et al.*, 2019] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *MICCAI 2019, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 302–310. Springer, 2019.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Huang *et al.*, 2021a] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE TII*, 18(8):5171–5179, 2021.

[Huang *et al.*, 2021b] Chao Huang, Zehua Yang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, and Yaowei Wang. Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection. *IEEE TCYB*, 52(12):13834–13847, 2021.

[Huang *et al.*, 2021c] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *arXiv preprint arXiv:2101.07172*, 2021.

[Huang *et al.*, 2022a] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE TCYB*, 54(5):3197–3210, 2022.

[Huang *et al.*, 2022b] Chao Huang, Chengliang Liu, Zheng Zhang, Zhihao Wu, Jie Wen, Qiuping Jiang, and Yong Xu. Pixel-level anomaly detection via uncertainty-aware prototypical transformer. In *ACM MM*, pages 521–530, 2022.

[Huang *et al.*, 2022c] Chao Huang, Yabo Liu, Zheng Zhang, Chengliang Liu, Jie Wen, Yong Xu, and Yaowei Wang. Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection. In *ACM MM*, pages 307–315, 2022.

[Huang *et al.*, 2022d] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE TNNLS*, 34(11):9389–9403, 2022.

[Huang *et al.*, 2024a] Chao Huang, Weichao Cai, Qiuping Jiang, and Zhihua Wang. Multimodal representation distribution learning for medical image segmentation. In *IJCAI*, pages 4156–4164, 2024.

[Huang *et al.*, 2024b] Chao Huang, Yushu Shi, Bob Zhang, and Ke Lyu. Uncertainty-aware prototypical learning for anomaly detection in medical images. *Neural Networks*, 175:106284, 2024.

[Huang *et al.*, 2024c] Chao Huang, Jie Wen, Chengliang Liu, and Yabo Liu. Long short-term dynamic prototype alignment learning for video anomaly detection. In *IJCAI*, pages 866–874, 2024.

[Huang *et al.*, 2025] Chao Huang, Weiliang Huang, Qiuping Jiang, Wei Wang, Jie Wen, and Bob Zhang. Multimodal evidential learning for open-world weakly-supervised video anomaly detection. *IEEE TMM*, 2025.

[Jha *et al.*, 2020] 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 *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020.

[Kirillov *et al.*, 2023] 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.

[Kumar *et al.*, 2017] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE TMI*, 36(7):1550–1560, 2017.

[Li *et al.*, 2022] Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. In *ICLR*, pages 1–20, 2022.

[Li *et al.*, 2023] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE TMI*, 2023.

[Li *et al.*, 2025] Jingtao Li, Ting Chen, Xinyu Wang, Yanfei Zhong, and Xuan Xiao. Adapting the segment anything model for multi-modal retinal anomaly detection and localization. *Information Fusion*, 113:102631, 2025.

[Liu *et al.*, 2021] An-An Liu, Hongshuo Tian, Ning Xu, Weizhi Nie, Yongdong Zhang, and Mohan Kankanhalli. Toward region-aware attention learning for scene graph generation. *IEEE TNNLS*, 33(12):7655–7666, 2021.

[Liu *et al.*, 2024] Jiaqi Liu, Kai Wu, Qiang Nie, Ying Chen, Bin-Bin Gao, Yong Liu, Jinbao Wang, Chengjie Wang, and Feng Zheng. Unsupervised continual anomaly detection with contrastively-learned prompt. In *AAAI*, volume 38, pages 3639–3647, 2024.

[Mamonov *et al.*, 2014] Alexander V Mamonov, Isabel N Figueiredo, Pedro N Figueiredo, and Yen-Hsi Richard Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE TMI*, 33(7):1488–1502, 2014.

[Patel *et al.*, 2021] Krushi Patel, Andrés M Bur, and Guanghui Wang. Enhanced u-net: A feature enhancement network for polyp segmentation. In *2021 18th conference on robots and vision (CRV)*, pages 181–188. IEEE, 2021.

[Reiss *et al.*, 2021] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *CVPR*, pages 2806–2814, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Shaharabany and Wolf, 2022] Tal Shaharabany and Lior Wolf. End-to-end segmentation of medical images via patch-wise polygons prediction. In *MICCAI*, pages 308–318. Springer, 2022.

[Shaharabany *et al.*, 2023] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023.

[Shin *et al.*, 2022] Wooseok Shin, Min Seok Lee, and Sung Won Han. Comma: Propagating complementary multi-level aggregation network for polyp segmentation. *Applied Sciences*, 12(4):2114, 2022.

[Silva *et al.*, 2014] 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*, 9:283–293, 2014.

[Sirinukunwattana *et al.*, 2017] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[Tajbakhsh *et al.*, 2015] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, 35(2):630–644, 2015.

[Trinh *et al.*, 2024] Quoc-Huy Trinh, Hai-Dang Nguyen, Bao-Tram Nguyen Ngoc, Debesh Jha, Ulas Bagci, and Minh-Triet Tran. Sam-eg: Segment anything model with egde guidance framework for efficient polyp segmentation. *arXiv preprint arXiv:2406.14819*, 2024.

[Valanarasu *et al.*, 2021] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *MICCAI 2021, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24*, pages 36–46. Springer, 2021.

[Wang *et al.*, 2020] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, pages 108–126. Springer, 2020.

[Wang *et al.*, 2022] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *AAAI*, volume 36, pages 2441–2449, 2022.

[Wang *et al.*, 2025] Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. In *AAAI*, volume 39, pages 21017–21025, 2025.

[Wei *et al.*, 2021] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *MICCAI 2021, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 699–708. Springer, 2021.

[Wei *et al.*, 2023] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with segment anything. *arXiv preprint arXiv:2311.17081*, 2023.

[Wu *et al.*, 2023] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[Xie *et al.*, 2024] Bin Xie, Hao Tang, Bin Duan, Dawen Cai, and Yan Yan. Masksam: Towards auto-prompt sam with mask classification for medical image segmentation. *arXiv preprint arXiv:2403.14103*, 2024.

[Yin *et al.*, 2022] Zijin Yin, Kongming Liang, Zhanyu Ma, and Jun Guo. Duplex contextual relation network for polyp segmentation. In *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*, pages 1–5. IEEE, 2022.

[Zhang *et al.*, 2020] Ruifei Zhang, Guanbin Li, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu. Adaptive context selection for polyp segmentation. In *MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 253–262. Springer, 2020.

[Zhang *et al.*, 2022] Dasheng Zhang, Chao Huang, Chengliang Liu, and Yong Xu. Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. *IEEE Signal Processing Letters*, 29:1197–1201, 2022.

[Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.