# Promptable Anomaly Segmentation with SAM Through Self-Perception Tuning

**Hui-Yue Yang[1], Hui Chen[2*], Ao Wang[1], Kai Chen[1], Zijia Lin[1], Yongliang Tang[3], Pengcheng Gao[3], Yuming Quan[3], Jungong Han[4], Guiguang Ding[1]**

[1]School of Software, Tsinghua University
[2]BNRist, Tsinghua University
[3]LUSTER LightTech Co., Ltd.
[4]Department of Automation, Tsinghua University
yanghuiyue18@outlook.com, {jichenhui2012,chenkai2010.9,quanym,jungonghan77}@gmail.com,
wa22@mails.tsinghua.edu.cn, linzijia07@tsinghua.org.cn, yongliangtang@lusterinc.com,
gaopengcheng15@mails.ucas.ac.cn, dinggg@tsinghua.edu.cn

## Abstract

Segment Anything Model (SAM) has made great progress in anomaly segmentation tasks due to its impressive generalization ability. However, existing methods that directly apply SAM through prompting often overlook the domain shift issue, where SAM performs well on natural images but struggles in industrial scenarios. Parameter-Efficient Fine-Tuning (PEFT) offers a promising solution, but it may yield suboptimal performance by not adequately addressing the perception challenges during adaptation to anomaly images. In this paper, we propose a novel **S**elf-**P**erception **T**uning (**SPT**) method, aiming to enhance SAM's perception capability for anomaly segmentation. The SPT method incorporates a self-drafting tuning strategy, which generates an initial coarse draft of the anomaly mask, followed by a refinement process. Additionally, a visual-relation-aware adapter is introduced to improve the perception of discriminative relational information for mask generation. Extensive experimental results on several benchmark datasets demonstrate that our SPT method can significantly outperform baseline methods, validating its effectiveness.

**Code** — https://github.com/THU-MIG/SAM-SPT
**Extended version** — https://arxiv.org/pdf/2411.17217

## 1 Introduction

Anomaly segmentation aims to automatically locate and segment abnormal regions within images of industrial products, which is crucial for improving production efficiency and product quality (Zhang et al. 2023; Chen et al. 2023c; Yang et al. 2024). In real-world scenarios, the sheer volume of industrial products and the wide spectrum of anomaly types present significant challenges for anomaly segmentation tasks. Consequently, constructing a highly generalized anomaly segmentation model has become a focal point of interest in the field (Cao et al. 2023; Jeong et al. 2023).

Recently, thanks to the emergence of pivotal vision foundation models like the Segment Anything Model (SAM) (Kirillov et al. 2023), significant advancements have been made in various computer vision fields. SAM, pre-trained on an extensive dataset SA-1B (Kirillov et al. 2023), has showcased remarkable zero-shot visual perception capabilities through prompting (Zou et al. 2024; Sun et al. 2024; Ke et al. 2024; Wang et al. 2024, 2023). Consequently, extensive research is underway to leverage SAM's robust generalization abilities for advancing anomaly segmentation tasks.

(a) Natural images (b) Industrial anomaly images

Figure 1: Illustration of the domain shift issue. SAM performs well on natural images but poorly on out-of-domain industrial anomaly images.

For example, SAA (Cao et al. 2023) represents a pioneering effort that applies SAM to anomaly segmentation tasks. Initially, it utilizes prompt-guided object detection methods like GroundingDINO (Liu et al. 2023) to generate prompt-conditioned box-level regions that highlight desired anomaly areas. These boxes are then fed into SAM as prompts to generate final predictions for anomaly segmentation. SAA+ (Cao et al. 2023) further introduces hybrid prompts that incorporate domain-specific expertise with contextual image information to mitigate ambiguities inherent in language prompts. In contrast, CLIPSAM (Li et al. 2024) replaces GroundingDINO with CLIP (Radford et al. 2021), leveraging its enhanced capability to provide precise localization information as prompts for SAM.

Despite the remarkable progress achieved, existing methods often overlook a critical issue of domain shift (Farahani et al. 2021; Xiong et al. 2023) when adapting SAM to downstream tasks (Zhang et al. 2024; Chen et al. 2023b; Huang et al. 2023). Specifically, SAM is primarily trained on nat-

*Corresponding author.

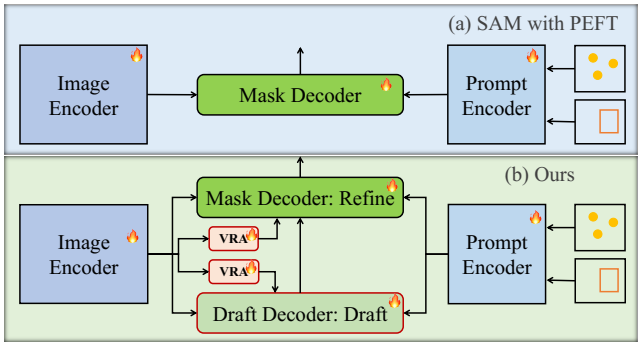


Figure 2: A comparison between SAM with PEFT methods and our promptable anomaly segmentation model with self-perception tuning. VRA denotes the VRA-Adapter.

ural image datasets, which enables excellent performance in general visual understanding tasks. However, it typically suffers from significant degradation when applied to out-of-domain industrial scenarios (see Fig. 1). Therefore, existing methods compensates for SAM's limited perception of domain-specific distributions by offering more contextual prompts. However, the fundamental issue of domain shift is still insufficiently considered in the context of anomaly segmentation.

Fine-tuning is a straightforward approach to enhance SAM's adaptation in anomaly segmentation tasks. Current research often employs parameter-efficient fine-tuning (PEFT) methods like LoRA (Hu et al. 2021) and Adapter (Houlsby et al. 2019) to tune foundation models. These methods introduce a minimal number of trainable parameters while keeping the bulk of SAM's parameters frozen. Despite notable improvements observed in our experiments, simply applying traditional PEFT methods to anomaly segmentation may obtain suboptimal results without taking the task challenge into consideration. Particularly, diverse imaging conditions, varied anomaly types across different industrial products, and the complex appearances of anomalies can greatly hinder the perception ability during adaptation. These challenges hinder the robustness of segmentation across different prompts as observed in our experiments (Fig. 5). Therefore, a question is raised: *How to enhance the perception ability of SAM to broaden its generalization capability for anomaly segmentation?*

In this paper, we propose a promptable anomaly segmentation model with SAM through a novel **S**elf-**P**erception **T**uning (**SPT**) method. Compared with conventional PEFT methods, the proposed SPT enhances perception capabilities by leveraging both **external** knowledge and **internal** priors, both of which are inherited from the model itself. Specifically, as shown in Fig. 2, we firstly design a Self-Draft Tuning (SDT) strategy, in which SAM initially generate a coarse draft of the anomaly mask, followed by a mask refinement process. This draft can serve as an external yet coarse perception knowledge of the anomaly mask, boosting the mask decoder with additional priors beyond the provided prompts and learned visual patterns. Furthermore, we introduce a Visual-Relation-Aware Adapter (VRA-Adapter) to enhance the internal perception knowledge of discriminative relation information during decoding. By capturing the relationship among different regions, our adapter can enable decoders to achieve a more nuanced and detailed understanding of anomaly patterns, resulting in more consistent and accurate anomaly segmentation. We show that our method can accommodate different PEFT methods and offer a flexible tuning framework. Experimental results on several industrial datasets show that compared to baseline methods, our SPT can enhance SAM's robust generalization for anomaly segmentation under different prompts, well demonstrating its effectiveness and superiority.

To summarize, our contributions are as follows:

- We introduce a promptable anomaly segmentation model with SAM through a novel Self-Perception Tuning (SPT) method. SPT can improve the perception capability of anomalies, enhancing the robust generalization of SAM for anomaly segmentation under various prompts.
- We design a self-draft tuning strategy to boost the mask decoder with external yet coarse perception of anomaly mask. A visual-relation-aware adapter is further introduced to enhance the internal perception of discriminative relation information for anomaly mask decoding.
- Extensive experiments across various industrial datasets for anomaly segmentation tasks show that our method can achieve state-of-the-art performance, well demonstrating the effectiveness of our method.

## 2 Related Work

**Foundation models for anomaly segmentation.** Large pre-trained models like CLIP and SAM have significantly advanced anomaly segmentation. WinCLIP (Jeong et al. 2023) first leverages CLIP's vision-language representation for zero-shot and few-shot anomaly detection. CLIP-AD (Chen et al. 2023c) refines CLIP for better zero-shot detection of unseen anomalies, while AnomalyCLIP (Zhou et al. 2023) utilizes CLIP's pre-training for effective anomaly detection without extensive task-specific training. SAA (Cao et al. 2023) uses GroundingDINO and SAM for segmentation and designs rules to filter out anomalies that meet specific criteria. CLIPSAM (Li et al. 2024) integrates CLIP's semantic understanding with SAM's segmentation capabilities, enhancing anomaly detection in complex scenes. These methods collectively demonstrate the effectiveness of integrating large pre-trained models into anomaly detection frameworks. However, they generally directly apply the foundation models into anomaly segmentation tasks, neglecting the critical issue of domain shift between pre-trained datasets and downstream tasks. In contrast, our work aims to construct a robust SAM with great generalization for anomaly segmentation tasks through fine-tuning strategy, which is complementary to the existing methods.

**Parameter-efficient fine-tuning (PEFT).** These methods address the high computational costs of adapting large pre-trained models into downstream tasks (Hu et al. 2021; Xiong et al. 2025; Houlsby et al. 2019; Hao et al. 2023). Adapter

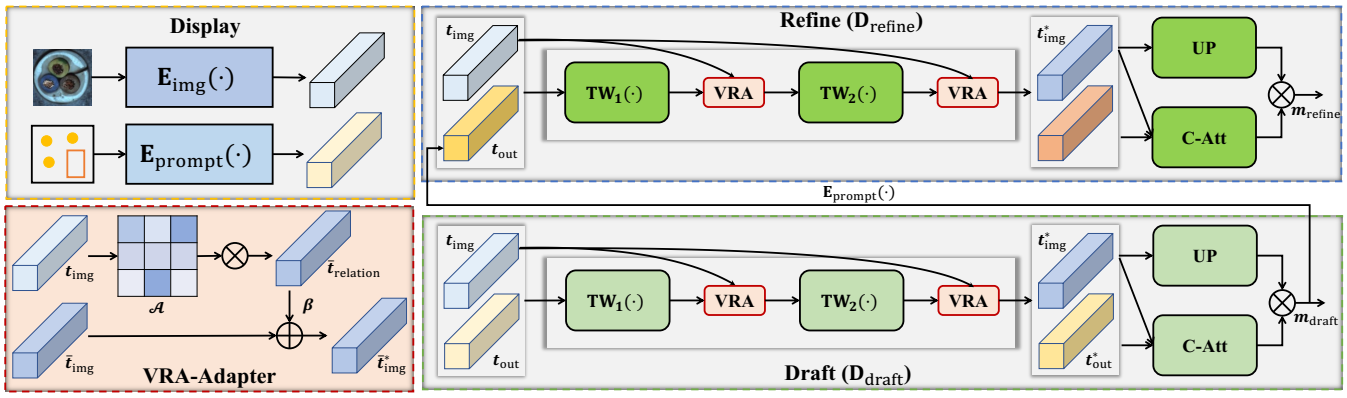


Figure 3: Overview of the proposed Self-Perception Tuning (SPT) framework, which applies a self-draft tuning (SDT) strategy and visual-relation-aware adapters (VRA-Adapter) to enhance the perception ability of SAM. SDT consists of three phases, *i.e.,* display, draft, and refine. VRA is an abbreviation for VRA-Adapter.

(Houlsby et al. 2019) introduces a small amount of trainable layers to each pre-trained layer, enabling efficient task-specific adaptation. LoRA (Hu et al. 2021) further reduces trainable parameters by injecting low-rank matrices into the model's layers. Adaptformer(Chen et al. 2022) introduces lightweight modules that add minimal parameters to the pre-trained ViTs. SamAdapter (Chen et al. 2023a) applies PEFT to SAM models, improving performance in underperforming scenes and challenging tasks. Recent advances in PEFT methods include DoRA (Liu et al. 2024) which introduces a weight-decomposed low-rank adaptation approach for efficient fine-tuning. NOLA (Koohpayegani et al. 2024) compresses LoRA using a linear combination of random basis for more efficient training. These methods demonstrate the ongoing evolution and versatility of PEFT techniques in the field of model adaptation. Despite notable improvement observed in our experiments, simply applying these PEFT methods to anomaly segmentation may take no consideration of the perception difficulties during adaptation, leading to suboptimal results. Therefore, we propose a self-perception tuning method to enhance the perception ability of SAM. As illustrated in our experiments, our SPT can accommodate different traditional PEFT methods, offering a flexible and effective fine-tuning framework.

# 3 Methodology

Here, we introduce a novel self-perception tuning (SPT) method to adapt SAM into the anomaly segmentation tasks with enhanced perception ability, as illustrated in Fig. 3.

## 3.1 Preliminary

**SAM** The Segment Anything Model (SAM) is an interactive image segmentation model based on user-provided prompts such as points, boxes, or text. It consists of an image encoder, a prompt encoder, and a mask decoder. The image encoder first extracts the image feature representation by a Vision Transformer (ViT) (Dosovitskiy et al. 2020). Then, SAM uses the prompt encoder to represent the user-provided prompts with prompt embeddings. Finally, the mask decoder utilizes the image features and prompt embeddings to generate the segmentation mask for the target object.

SAM aims to achieve high flexibility and generality, making it adaptable to various segmentation tasks without the need for task-specific fine-tuning. However, due to the domain shift between pre-trained datasets and practical industrial scenarios, directly applying SAM for anomaly segmentation unavoidably encounters performance degradation.

## 3.2 Self-Draft Tuning

Directly applying conventional PEFT methods (*e.g.,* LoRA) to SAM for anomaly segmentation tasks may lead to suboptimal performance, as these methods do not account for the task challenge to the perception ability of SAM. To mitigate this issue, the proposed SPT method firstly introduces a self-draft tuning (SDT) strategy to enhance the perception outcome of SAM by utilizing the external knowledge of the model itself. In SDT, we augment SAM with a draft decoder alongside its the original mask decoder. When given an image, SAM firstly generates a coarse draft of the anomaly mask using the draft decoder. Subsequently, the mask decoder refines this draft, resulting in more accurate segmentation. For clarity, we organize the SDT pipeline into three stages: display, draft, and refine.

**Display.** In the display phase, SAM uses the image encoder to extract the image features $\boldsymbol{e}_{\text{img}}$ for the input image $\boldsymbol{I}$ and uses the prompt encoder to generate various types of prompt embeddings for user-provided prompts $\boldsymbol{P}$, including sparse embedding $\boldsymbol{e}_{\text{sparse}}$, dense embedding $\boldsymbol{e}_{\text{dense}}$, and position embedding $\boldsymbol{e}_{\text{pos}}$:

$$\boldsymbol{e}_{\text{img}} = \mathbf{E}_{\text{img}}(\boldsymbol{I}), \{\boldsymbol{e}_{\text{sparse}}, \boldsymbol{e}_{\text{dense}}, \boldsymbol{e}_{\text{pos}}\} = \mathbf{E}_{\text{prompt}}(\boldsymbol{P}) \quad (1)$$

where $\mathbf{E}_{\text{img}}$ denotes the image encoder and $\mathbf{E}_{\text{prompt}}$ denotes the prompt encoder. We augment both encoders with learnable PEFT modules to capture discriminative feature representations for anomaly segmentation.

**Draft.** The draft decoder shares the same structure as the original mask decoder and is initialized with pre-trained weights of the latter, maintaining considerable segmentation ability for draft. It utilizes separated Two-Way Transformer

blocks to manage the input information including the image feature $\boldsymbol{e}_{\text{img}}$, the prompt embeddings $\{\boldsymbol{e}_{\text{sparse}}, \boldsymbol{e}_{\text{dense}}, \boldsymbol{e}_{\text{pos}}\}$, the IoU token $\boldsymbol{e}_{\text{iou}}$ and mask tokens $\boldsymbol{e}_{\text{mask}}$:

$$\begin{aligned} \boldsymbol{t}_{\text{img}} &= \boldsymbol{e}_{\text{img}} + \boldsymbol{e}_{\text{dense}} \\ \boldsymbol{t}_{\text{out}} &= [\boldsymbol{e}_{\text{sparse}}, \boldsymbol{e}_{\text{iou}}, \boldsymbol{e}_{\text{mask}}] \\ \boldsymbol{t}^*_{\text{img}}, \boldsymbol{t}^*_{\text{out}} &= \mathbf{TW}_2(\mathbf{TW}_1(\boldsymbol{t}_{\text{img}}, \boldsymbol{e}_{\text{pos}}, \boldsymbol{t}_{\text{out}})) \end{aligned} \tag{2}$$

where $\mathbf{TW}(\cdot)$ with subscripts is the process of the Two-Way Transformer Block. $[\cdot]$ means token concatenation. The final mask can be generated by a dot product between the updated mask token and the upsampled image features:

$$\boldsymbol{e}^*_{\text{mask}} = \text{C-Att}(\boldsymbol{t}^*_{\text{out}}, \boldsymbol{t}^*_{\text{img}})[-1], \boldsymbol{m}_{\text{draft}} = \boldsymbol{e}^*_{\text{mask}} \times \text{UP}(\boldsymbol{t}^*_{\text{img}}) \tag{3}$$

where C-Att$(\cdot,\cdot)$ is a token-to-image cross attention. The original SAM has 3 mask tokens. Here, we only use the first token for mask generation by default following previous works (Ke et al. 2024). Combining Eq. 2 and Eq. 3, the coarse draft generation can be formalized as follows:

$$\boldsymbol{m}_{\text{draft}} = \mathbf{D}_{\text{draft}}(\boldsymbol{e}_{\text{img}}, \boldsymbol{e}_{\text{sparse}}, \boldsymbol{e}_{\text{dense}}, \boldsymbol{e}_{\text{pos}}) \tag{4}$$

where $\mathbf{D}_{\text{draft}}$ denotes the draft decoder. We use PEFT methods to adapt the draft decoder to the anomaly segmentation.

**Refine.** In the refine phase, the original mask decoder of SAM aims to perceive all available information to generate a refined mask for prediction. Therefore, we derive the representation for the coarse draft of anomaly mask, *i.e.,* $\boldsymbol{m}_{\text{draft}}$, through the prompt encoder $\mathbf{E}_{\text{prompt}}$:

$$\boldsymbol{e}_{\text{draft}} = \mathbf{E}_{\text{prompt}}(\boldsymbol{m}_{\text{draft}}) \tag{5}$$

Noting that $\boldsymbol{e}_{\text{draft}}$ captures the coarse perception of the anomaly mask for SAM. Consequently, we can use it as a external knowledge to refine the mask generation for the anomaly segmentation. The final mask for anomaly segmentation is derived by:

$$\boldsymbol{m}_{\text{refine}} = \mathbf{D}_{\text{refine}}(\boldsymbol{e}_{\text{img}}, \boldsymbol{e}_{\text{sparse}}, \boldsymbol{e}_{\text{draft}}, \boldsymbol{e}_{\text{pos}}) \tag{6}$$

where $\mathbf{D}_{\text{refine}}$ follows the same procedure as $\mathbf{D}_{\text{draft}}$ in Eq. 4 but using the original mask decoder of SAM. $\boldsymbol{e}_{\text{draft}}$ is injected as the dense embedding. The PEFT method is applied to adjust the mask decoder to refine the anomaly mask.

By reviewing the SDT framework, we observe that SAM's feature representations for images and prompts are tailored to industrial scenarios during the display phase. Subsequently, the decoders are fine-tuned to align with the anomaly generation task through a draft-then-refine procedure, effectively enhancing SAM's perception capability for anomaly segmentation.

### 3.3 Visual-Relation-Aware Adapter

The Visual-Relation-Aware Adapter (VRA-Adapter) aims to enhance the mask decoders by integrating the visual relationships among different regions into the decoding process. It starts by evaluating visual relationships within the image and then uses a relation-aware adapter to propagate this relation knowledge to the decoder.

**Visual relation evaluation** is based on the insight that inter-region relationships within an image provide valuable information for anomaly detection (You et al. 2022; Yao et al. 2023). Therefore, we firstly measure such relationships using the cosine similarity metric:

$$\mathbf{S} = \text{cosine}(\boldsymbol{e}_{\text{img}}, \boldsymbol{e}_{\text{img}}) \tag{7}$$

where $\boldsymbol{e}_{\text{img}}$ is the feature embeddings corresponding to each image region derived in Eq. 1. The diagonal elements of $\mathbf{S}$ is set to $-\infty$ to avoid trivial self-similarities. Subsequently, we apply a softmax function to obtain the relation matrix:

$$\mathcal{A} = \text{softmax}(\mathbf{S}) \tag{8}$$

The relation matrix is further refined using a thresholding mechanism controlled by a parameter $\alpha$ divided by the feature dimensionality $d$, ensuring that only the most significant relations are retained:

$$\mathcal{A}^* = \begin{cases} \mathcal{A}_{ij} & \text{if } \mathcal{A}_{ij} \geq \frac{\alpha}{d} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

**Relation-aware adapter** aims to integrate this relation information to enhance the quality of anomaly segmentation. Considering that neither decoder, *i.e.,* $\mathbf{D}_{\text{draft}}$ and $\mathbf{D}_{\text{refine}}$, applies self-attention to the image embeddings $\boldsymbol{e}_{\text{img}}$, attention to image regions may gradually shift or even diminish over the course of decoding, leading to inconsistent anomaly segmentation. Therefore, we use a relation-aware adapter to complement the perception of relationship among image regions for the decoding process.

We denote the updated image features output by the Two-Way Transformers ($\mathbf{TW}(\cdot)$ in Eq. 2) as $\bar{\boldsymbol{t}}^1_{\text{img}}$ and $\bar{\boldsymbol{t}}^2_{\text{img}}$. For ease of explanation, we use $\bar{\boldsymbol{t}}_{\text{img}}$ to refer to these features. Then, leveraging the relation matrix $\mathcal{A}$, we can aggregate features with high relationships for each image feature via the attention mechanism, resulting in the relation-aware image features for decoding:

$$\bar{\boldsymbol{t}}_{\text{relation}} = \mathcal{A}\bar{\boldsymbol{t}}_{\text{img}} \tag{10}$$

Subsequently, we combine the relation-aware image feature with the origin image feature $\bar{\boldsymbol{e}}_{\text{img}}$ using a scale vector $\boldsymbol{\beta}$:

$$\bar{\boldsymbol{t}}^*_{\text{img}} = \bar{\boldsymbol{t}}_{\text{img}} + \boldsymbol{\beta}\bar{\boldsymbol{t}}_{\text{relation}} \tag{11}$$

The scale vector $\boldsymbol{\beta}$ is dynamically adjusted during training to regulate the integration of visual relation information.

By leveraging the internal perception knowledge of visual relationships, *i.e.,* $\mathcal{A}$, the image feature $\bar{\boldsymbol{t}}^*_{\text{img}}$ becomes more consistent and discriminative for regions with similar visual patterns. This enhancement significantly improves the discrimination between anomalous and normal regions, enabling a more accurate and detailed understanding of anomalies, and thus leading to better performance.

### 3.4 Promptable Anomaly Segmentation Model

We fine-tune SAM using the proposed self-perception tuning method, resulting in a promptable anomaly segmentation model. The optimization objectives combine cross-entropy loss and Dice loss over both the draft and final refined masks:

$$\begin{aligned} \mathcal{L} = &\mathcal{L}_{\text{CE}}(\mathbf{m}_{\text{draft}}, \mathbf{Y}) + \mathcal{L}_{\text{dice}}(\mathbf{m}_{\text{draft}}, \mathbf{Y}) \\ &+ \mathcal{L}_{\text{CE}}(\mathbf{m}_{\text{refine}}, \mathbf{Y}) + \mathcal{L}_{\text{dice}}(\mathbf{m}_{\text{refine}}, \mathbf{Y}) \end{aligned} \tag{12}$$

where $\mathbf{Y}$ denotes the ground truth for the input image given the prompt.

| Method | One box | | Multiple boxes | | Point=5 | | Point=10 | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mIoU | mBIoU | mIoU | mBIoU | mIoU | mBIoU | mIoU | mBIoU | mIoU | mBIoU |
| zero-shot | 56.3 | 50.7 | 63.0 | 57.9 | 44.1 | 39.8 | 44.4 | 39.7 | 52.0 | 47.0 |
| LoRA | 65.1 | 58.8 | 69.9 | 64.8 | 62.4 | 57.5 | 68.7 | 62.9 | 66.5 | 61.0 |
| $\text{SPT}_{\text{LoRA}}$ | **67.1** | **60.7** | **71.7** | **66.3** | **64.6** | **59.8** | **70.1** | **64.5** | **68.4** | **62.8** |
| DoRA | 65.3 | 58.8 | 70.1 | 64.7 | 62.8 | 57.9 | 67.7 | 62.0 | 66.5 | 60.9 |
| $\text{SPT}_{\text{DoRA}}$ | **66.9** | **60.6** | **71.4** | **66.4** | **65.3** | **60.4** | **70.5** | **64.9** | **68.5** | **63.1** |
| Adapter | 65.1 | 59.2 | 70.0 | 65.3 | 61.8 | 56.9 | 67.2 | 61.4 | 66.0 | 60.7 |
| $\text{SPT}_{\text{Adapter}}$ | **66.0** | **59.5** | **71.2** | **65.9** | **63.4** | **58.7** | **69.3** | **63.7** | **67.5** | **62.0** |

Table 1: Performance comparison under different evaluation modes (%). "Avg." is the average scores of four kinds of prompts.

# 4 Experiments

## 4.1 Experiment Setups

**Datasets.** To ensure generalization across various industrial products and anomaly types with different prompts, we collect approximately 15,000 industrial anomaly images from real-world factories as training dataset. For the evaluation, we use six standard benchmark datasets, including MVTec (Bergmann et al. 2019), VisA (Zou et al. 2022), MTD (Huang, Qiu, and Yuan 2020), KSDD2 (Božič, Tabernik, and Skočaj 2021), BTAD (Mishra et al. 2021), and MPDD (Jezek et al. 2021). The training dataset includes a variety of imaging conditions, product types, and anomaly classes that are distinct from those in the test datasets, ensuring a fair evaluation of generalization. More details are provided in the extended version.

**Implementation details.** We initialize the SAM model with the officially provided pre-trained weights, utilizing three different backbone sizes: ViT-H, ViT-L, and ViT-B. The input image is resized to $1024 \times 1024$. During training, for all models, the learning rate is set to $1 \times 10^{-3}$ and is reduced after 10 epochs. All models are trained for 16 epochs using 8 NVIDIA 3090 GPUs with a batch of 8 images. The $\alpha$ in VRA-Adapter remains robust within the range of 0 to 0.5, depending on the specific PEFT method and model size. The rank of adapter is set to 8 for all models by default.

**Evaluation metrics.** To ensure that the constructed anomaly segmentation models can maintain the promptable functionality and generalization as SAM, we design three types of evaluation mode with different prompts: (1) one box, which highlights one or multiple defects with a single box; (2) multiple boxes, where each defect is assigned its own box, and (3) points. We randomly sample either 5 or 10 points for the prompt of points. We believe that this comprehensive evaluation effectively simulates user behavior in real-world scenarios, making our method highly applicable in practice. Two widely used metrics for promptable segmentation tasks, *i.e.,* mean Intersection over Union (mIoU) and mean Boundary Intersection over Union (mBIoU), are leveraged to evaluate the segmentation result, following previous works (Kirillov et al. 2023; Ke et al. 2024).

## 4.2 Comparison with State-of-the-art Methods

We choose zero-shot SAM and representative PEFT methods as baseline methods, including LoRA (Hu et al. 2021), DoRA (Liu et al. 2024), and Adapter (Houlsby et al. 2019). Since our method can accommodate the PEFT methods, we introduce three variants of our method according to its adopted PEFT method, having $\text{SPT}_{\text{LoRA}}$, $\text{SPT}_{\text{DoRA}}$, and $\text{SPT}_{\text{Adapter}}$. We provide comparison results based on ViT-B as backbone. More baseline methods and results of ViT-L/H are left in the extended version.

As illustrated in Table 1, our method significantly outperform various baseline approaches. Notably, compared to the zero-shot performance of the vanilla SAM model, our three SPT variants significantly enhance anomaly segmentation across six benchmark datasets, achieving over 15% improvement in mIoU and mBIoU on average. This substantial improvement is due to our SPT method's effectiveness in addressing the domain shift issue for SAM. Additionally, compared to conventional PEFT baselines, our SPT method consistently delivers large performance gains. Specifically, $\text{SPT}_{\text{DoRA}}$ surpasses DoRA by an average of 2.0% in mIoU and 2.2% in mBIoU. In the Point=10 setting, these improvements increase to 2.8% in mIoU and 2.9% in mBIoU when comparing $\text{SPT}_{\text{DoRA}}$ to DoRA. These gains can be attributed to the proposed self-draft tuning framework and the visual-relation-aware adapter.

## 4.3 Model Analysis

For model analysis, we report the overall performance at the box level, point level, and the average. $\text{SPT}_{\text{LoRA}}$ based on ViT-B is leveraged here by default.

| SDT | VRA | box-level | | point-level | | Avg. | |
|---|---|---|---|---|---|---|---|
| | | mIoU | mBIoU | mIoU | mBIoU | mIoU | mBIoU |
| zero-shot | | 59.7 | 54.3 | 44.3 | 39.8 | 52.0 | 47.0 |
| PEFT | | 67.5 | 61.8 | 65.6 | 60.2 | 66.5 | 61.0 |
| ✓ | | 68.6 | 62.9 | 66.7 | 61.6 | 67.6 | 62.2 |
| | ✓ | 68.1 | 62.3 | 65.3 | 60.1 | 66.7 | 61.2 |
| ✓ | ✓ | **69.4** | **63.5** | **67.4** | **62.2** | **68.4** | **62.8** |

Table 2: Ablation study of each component in the proposed SPT (%). VRA is an abbreviation for VRA-Adapter.

**Ablation study.** To verify the effectiveness of each component in our SPT, we introduce zero-shot SAM and SAM tuned by PEFT method as competitors. As shown in Table 2, we can see that each component can substantially contribute

| SDT | VRA | All (M) | Trainable (M) | ratio (%) | memory (M) | throughput (fps) |
|---|---|---|---|---|---|---|
| zero-shot | | 89.393 | - | - | - | 8.67 |
| PEFT | | 89.719 | 0.326 | 0.364 | 7904 | 8.69 |
| ✓ | | 93.635 | 0.371 | 0.396 | 7924 | 8.27 |
| | ✓ | 89.725 | 0.327 | 0.364 | 7907 | 8.46 |
| ✓ | ✓ | 93.636 | 0.372 | 0.397 | 7926 | 7.98 |

Table 3: Cost and efficiency analysis of each component.

| Method | box-level | | point-level | | Avg. | |
|---|---|---|---|---|---|---|
| | mIoU | mBIoU | mIoU | mBIoU | mIoU | mBIoU |
| zero-shot SAM | 59.7 | 54.3 | 44.3 | 39.8 | 52.0 | 47.0 |
| zero-shot SDT | 59.9 | 55.1 | 46.0 | 41.7 | 53.0 | 48.4 |
| SPT w $\mathbf{D}_{\text{mask}}$ 2× | 67.8 | 62.0 | 66.2 | 61.2 | 67.0 | 61.6 |
| SPT w $\mathbf{D}_{\text{draft}} \times 1$ | 69.4 | 63.5 | 67.4 | 62.2 | **68.4** | 62.8 |
| SPT w $\mathbf{D}_{\text{draft}} \times 2$ | 68.8 | 63.2 | 67.5 | 62.5 | 68.1 | **62.9** |

Table 4: The impact of the self-draft tuning (%).

to the performance gains. Benefited from the nature of draft-then-refine process, the proposed self-draft tuning (SDT) can outperform conventional PEFT method with over 1.0% improvement for both metrics, indicating the advantage of SDT. Combined with VRA-Adapter, the gains can reach to 1.9% mIoU and 1.8% mBIoU, demonstrating the positive effect of VRA-Adapter. We also present the cost and efficiency analysis for each component. From Table 3, we can observe that the proposed SDT and VRA-Adapter only introduce small extra learnable parameters, leading to marginal memory assumption increase during training and tiny inference throughput decrease. These evidences can well demonstrate the effectiveness of each component in SPT.

**Analysis of the self-draft tuning.** First, we further explore the benefits of the draft-then-refine process by comparing zero-shot SAM with zero-shot SDT, which applies SDT to SAM without PEFT tuning. As shown in Table 4, zero-shot SDT substantially achieves improvements across all metrics, resulting in gains of 1.0% in mIoU and 1.4% in mBIoU on average. The results from this zero-shot comparison effectively highlight the superiority of SDT, as confirmed by the ablation study. Additionally, we analyze the impact of using more $\mathbf{D}_{\text{draft}}$ with our SPT, as well as the effect of employing a single $\mathbf{D}_{\text{mask}}$ that is applied twice during both training and inference with shared LoRA parameters (denoted by "SPT w $\mathbf{D}_{\text{mask}}$ 2×"). We can see that the performance of "SPT w $\mathbf{D}_{\text{draft}} \times 2$" is comparable to that of "SPT w $\mathbf{D}_{\text{draft}} \times 1$", indicating that a single $\mathbf{D}_{\text{draft}}$ is sufficient to achieve remarkable performance. However, the performance of "SPT w $\mathbf{D}_{\text{mask}}$ 2×" is inferior to that of "SPT w $\mathbf{D}_{\text{draft}} \times 1$", highlighting the necessity of the draft decoder in our proposed strategy.

**Accurate draft enhances the performance.** To gain further insights into the draft-then-refine process, we investigate how the perception degradation of each part of the model affects the overall segmentation performance. As shown in Table 5, we leave out PEFT modules, *i.e.,* LoRA here, in the image encoder $\mathbf{E}$, draft decoder $\mathbf{D}_{\text{draft}}$, and mask decoder $\mathbf{D}_{\text{refine}}$ separately. We can observe that (1) compared with "only $\mathbf{D}_{\text{refine}}$", "only $\mathbf{D}_{\text{draft}}$" shows better performance. Compared with "no PEFT" (*i.e.,* None), "only $\mathbf{D}_{\text{draft}}$" exhibits significant performance gains. These evidences demonstrate that the performance can be enhanced with more accurate draft result. (2) compared with "no decoder", "no encoder" suffers from more performance drops, indicating the importance of adapting representation of SAM to the industrial images. It also implicitly reveal that draft on inaccurate information can still encounter segmentation degradation. (3) enhancing the perception ability for all parts can achieve the optimal adaptation performance.

| PEFT | box-level | | point-level | | Avg. | |
|---|---|---|---|---|---|---|
| | mIoU | mBIoU | mIoU | mBIoU | mIoU | mBIoU |
| None | 59.9 | 55.1 | 46.0 | 41.7 | 53.0 | 48.4 |
| only $\mathbf{D}_{\text{refine}}$ | 63.0 | 58.3 | 58.6 | 54.5 | 60.8 | 56.4 |
| only $\mathbf{D}_{\text{draft}}$ | 61.6 | 57.0 | 62.7 | 57.1 | 62.1 | 57.0 |
| no decoder | 66.5 | 60.5 | 61.6 | 55.9 | 64.0 | 58.2 |
| no encoder | 63.5 | 59.3 | 61.4 | 57.6 | 62.4 | 58.4 |
| All | **68.6** | **62.9** | **66.7** | **61.6** | **67.6** | **62.2** |

Table 5: The impact of accommodating PEFT (%).

| VRA-Adapter | box-level | | point-level | | Avg. | |
|---|---|---|---|---|---|---|
| | mIoU | mBIoU | mIoU | mBIoU | mIoU | mBIoU |
| None | 68.6 | 62.9 | 66.7 | 61.6 | 67.6 | 62.2 |
| $\mathbf{TW}_1$ | 68.8 | 63.1 | 67.5 | **62.4** | 68.2 | **62.8** |
| $\mathbf{TW}_2$ | 69.0 | 63.2 | 67.0 | 61.9 | 68.0 | 62.5 |
| $\mathbf{TW}_1 + \mathbf{TW}_2$ | **69.4** | **63.5** | **67.4** | 62.2 | **68.4** | **62.8** |

Table 6: The impact of VRA-Adapter (%).

| Method | AUROC | AP | max-F1 |
|---|---|---|---|
| zero-shot | 73.6 | 31.1 | 41.2 |
| Ours | **74.8** | **33.1** | **42.5** |

Table 7: Analysis of automatic prompts (%).

**Analysis of the VRA-Adapter.** We investigate the impact of VRA-Adapter with respect to its replacement. We consider three cases that placing it in $\mathbf{TW}_1$, $\mathbf{TW}_2$, or both. As illustrated in Table 6, by comparing with no VRA-Adapter, we observe that the performance can be enhanced as long as the VRA-Adapter is applied. Besides, we can see that (1) placing it in an earlier layer is more beneficial than in a later layer, and (2) using it in both layers yields the best performance. These observations effectively demonstrate the positive impact of incorporating visual relations into the decoder.

**SPT also works using automatic prompts.** Here we evaluate the effectiveness of our method when using automatic prompts generated by GroundingDINO as SAA+ (Cao et al. 2023). Specifically, we directly replace the SAM model in SAA+ with our constructed model using the official code[1]. For convenience and direct comparison, we employ the same

[1]https://github.com/caoyunkang/Segment-Any-Anomaly

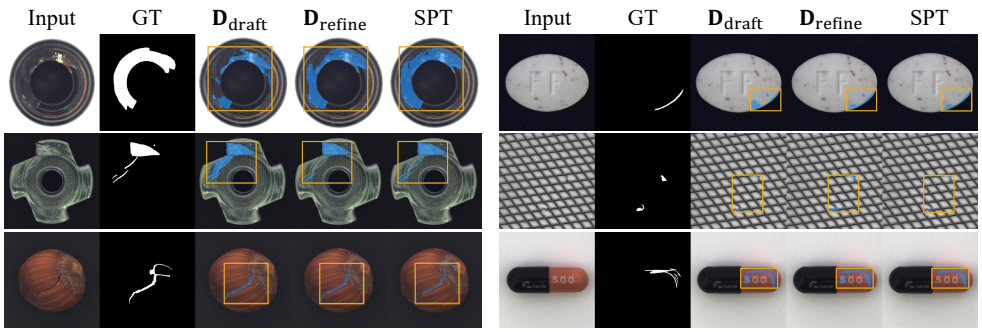


Figure 4: Examples for comparison among components in SPT. We use $\mathbf{D}_{\text{draft}}$ and $\mathbf{D}_{\text{refine}}$ for analysing SDT and SPT for analysing VRA-Adapter. GT denotes the ground truth.

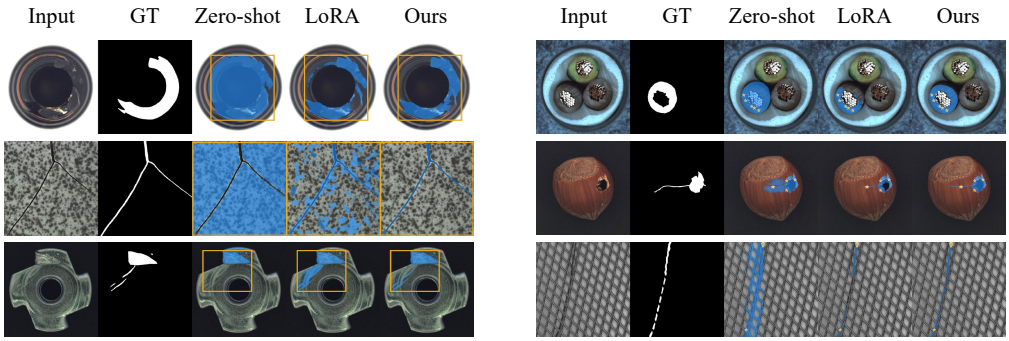


Figure 5: Qualitative analysis using different prompts. We provide examples with box-level prompts (left) and point-level prompts (right). GT means ground truth.

human-crafted prompts as those used in SAA+, even though these prompts are known to be sensitive to different models and datasets. As illustrated in Table 7, our method can consistently outperform the zero-shot SAM across different metrics with a maximal performance improvement of 2.0% (AP). These results well demonstrate the effectiveness of our method using the automatic prompts.

### 4.4 Qualitative Analysis

**Visualization analysis for components in SPT.** For intuitively understanding the advantage of each component in SPT, we visualize the anomaly mask generated by $\mathbf{D}_{\text{draft}}$, $\mathbf{D}_{\text{refine}}$ and SPT for analysing SDT and VRA-Adapter. As shown in Fig. 4, the masks produced by $\mathbf{D}_{\text{draft}}$ provide a coarse outline of the anomalous regions, but they may include some extraneous areas or miss certain parts. After going through the refine step, the anomaly mask generated by $\mathbf{D}_{\text{refine}}$ becomes more accurate and precise. Enhanced by VRA-Adapter, our SPT can capture more accurate mask with high correlation, compared with $\mathbf{D}_{\text{refine}}$. These evidences well support the effectiveness of each component.

**Visualization of anomaly segmentation using different prompts.** To demonstrate the superiority of our method in the generalization using different prompts, we visualize and compare segmentation results of different methods. As shown in Fig. 5, the original SAM model lacks the ability to recognize defects, often resulting in localizing continuous regions that are not anomalies. Although LoRA improves the performance to some extent, it performs worse than our method. For example, in the point-level prompt mode, LoRA tends to simply segment areas exactly covered by points. In contrast, our method can discover anomalies that the points do not fully covers, demonstrating its superior robustness in such scenarios. Such generalization can be largely attributed to the designed visual-relation-aware adapter which can enhance the relationship among anomalous regions, leading to consistent and accurate anomaly segmentation.

## 5 Conclusion

In this paper, we propose a novel Self-Perception Tuning (SPT) method to adapt SAM for practical anomaly segmentation scenarios. Unlike the original SAM and conventional parameter-efficient fine-tuning (PEFT) methods, our approach aims to enhance SAM's perception capabilities during adaptation to address domain shift issues. Specifically, SPT incorporates a self-drafting tuning strategy, which first generates a coarse draft of the anomaly mask and then undergoes a refinement process. Additionally, we introduce a visual-relation-aware adapter to improve the perception of discriminative relation information for mask generation. Extensive experimental results on several benchmark datasets show that our SPT method achieves state-of-the-art performance, validating its effectiveness.

## Acknowledgments

This work was supported by National Science and Technology Major (No. 2022ZD0119401), National Natural Science Foundation of China (No. 61925107, 62271281, 62021002).

## References

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Božič, J.; Tabernik, D.; and Skočaj, D. 2021. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129: 103459.

Cao, Y.; Xu, X.; Sun, C.; Cheng, Y.; Du, Z.; Gao, L.; and Shen, W. 2023. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Chen, T.; Zhu, L.; Deng, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; and Mao, P. 2023a. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3367–3375.

Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Wang, Y.; Li, Z.; Sun, L.; Mao, P.; and Zang, Y. 2023b. SAM Fails to Segment Anything?–SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. *arXiv preprint arXiv:2304.09148*.

Chen, X.; Zhang, J.; Tian, G.; He, H.; Zhang, W.; Wang, Y.; Wang, C.; Wu, Y.; and Liu, Y. 2023c. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. *arXiv preprint arXiv:2311.00453*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Farahani, A.; Voghoei, S.; Rasheed, K.; and Arabnia, H. R. 2021. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877–894.

Hao, T.; Chen, H.; Guo, Y.; and Ding, G. 2023. Consolidator: Mergable Adapter with Group Connections for Visual Adaptation. In *The Eleventh International Conference on Learning Representations*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Cao, Y.; Li, T.; Juefei-Xu, F.; Lin, D.; Tsang, I. W.; Liu, Y.; and Guo, Q. 2023. On the robustness of segment anything. *arXiv preprint arXiv:2305.16220*.

Huang, Y.; Qiu, C.; and Yuan, K. 2020. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1): 85–96.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19606–19616.

Jezek, S.; Jonak, M.; Burget, R.; Dvorak, P.; and Skotak, M. 2021. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, 66–71. IEEE.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2024. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Koohpayegani, S. A.; Navaneet, K.; Nooralinejad, P.; Kolouri, S.; and Pirsiavash, H. 2024. NOLA: Compressing LoRA using Linear Combination of Random Basis. arXiv:2310.02556.

Li, S.; Cao, J.; Ye, P.; Ding, Y.; Tu, C.; and Chen, T. 2024. ClipSAM: CLIP and SAM Collaboration for Zero-Shot Anomaly Segmentation. *arXiv preprint arXiv:2401.12665*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, S.-Y.; Wang, C.-Y.; Yin, H.; Molchanov, P.; Wang, Y.-C. F.; Cheng, K.-T.; and Chen, M.-H. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 01–06. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sun, Y.; Chen, J.; Zhang, S.; Zhang, X.; Chen, Q.; Zhang, G.; Ding, E.; Wang, J.; and Li, Z. 2024. VRP-SAM: SAM with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23565–23574.

Wang, A.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2023. Repvit-sam: Towards real-time segmenting anything. *arXiv preprint arXiv:2312.05760*.

Wang, A.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2024. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15909–15920.

Xiong, Y.; Chen, H.; Hao, T.; Lin, Z.; Han, J.; Zhang, Y.; Wang, G.; Bao, Y.; and Ding, G. 2025. Pyra: Parallel yielding re-activation for training-inference efficient task adaptation. In *European Conference on Computer Vision*, 455–473. Springer.

Xiong, Y.; Chen, H.; Lin, Z.; Zhao, S.; and Ding, G. 2023. Confidence-based Visual Dispersal for Few-shot Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11621–11631.

Yang, H.-Y.; Chen, H.; Liu, L.; Lin, Z.; Chen, K.; Wang, L.; Han, J.; and Ding, G. 2024. Context Enhancement with Reconstruction as Sequence for Unified Unsupervised Anomaly Detection. In *ECAI 2024*, 2098–2105. IOS Press.

Yao, X.; Li, R.; Qian, Z.; Luo, Y.; and Zhang, C. 2023. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6803–6813.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584.

Zhang, H.; Su, Y.; Xu, X.; and Jia, K. 2024. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23385–23395.

Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; and Chen, T. 2023. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3914–3923.

Zhou, Q.; Pang, G.; Tian, Y.; He, S.; and Chen, J. 2023. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, 392–408. Springer.