

# 2nd Place Winning Solution for the CVPR2023 Visual Anomaly and Novelty Detection Challenge: Multimodal Prompting for Data-centric Anomaly Detection

Yunkang Cao<sup>1\*</sup> Xiaohao Xu<sup>1\*</sup> Chen Sun<sup>1</sup> Yuqi Cheng<sup>1</sup>

Liang Gao<sup>1</sup> Weiming Shen<sup>1§</sup>

<sup>1</sup> State Key Laboratory of Digital Manufacturing Equipment and Technology,

Huazhong University of Science and Technology, China

{cyk\_hust, sun\_chen, chengyuqi, gaoliang}@hust.edu.cn  
xxh11102019@outlook.com, wshen@ieee.org

## Abstract

This technical report introduces the winning solution of the team Segment Any Anomaly for the CVPR2023 Visual Anomaly and Novelty Detection (VAND) challenge. Going beyond uni-modal prompt, e.g., language prompt, we present a novel framework, i.e., Segment Any Anomaly + (SAA+), for zero-shot anomaly segmentation with multi-modal prompts for the regularization of cascaded modern foundation models. Inspired by the great zero-shot generalization ability of foundation models like Segment Anything, we first explore their assembly (SAA) to leverage diverse multi-modal prior knowledge for anomaly localization. Subsequently, we further introduce multimodal prompts (SAA+) derived from domain expert knowledge and target image context to enable the non-parameter adaptation of foundation models to anomaly segmentation. The proposed SAA+ model achieves state-of-the-art performance on several anomaly segmentation benchmarks, including VisA and MVTec-AD, in the zero-shot setting. We will release the code of our winning solution for the CVPR2023 VAND challenge at <https://github.com/caoyunkang/Segment-Any-Anomaly><sup>1</sup>

## 1. Introduction

Anomaly segmentation [4–7] have gained great popularity in industrial quality control [8], medical diagnoses [9], etc. We focus on the setting of zero-shot anomaly segmentation (ZSAS) on images, which aims at utilizing neither normal nor abnormal samples for segmenting any anomalies in countless objects.

Recently, foundation models, e.g., SAM [3] and CLIP [10], exhibit great zero-shot visual perception abil-

\*Equal Contribution.

§Corresponding Author.

<sup>1</sup>The extended-version paper with more details is available at [1].

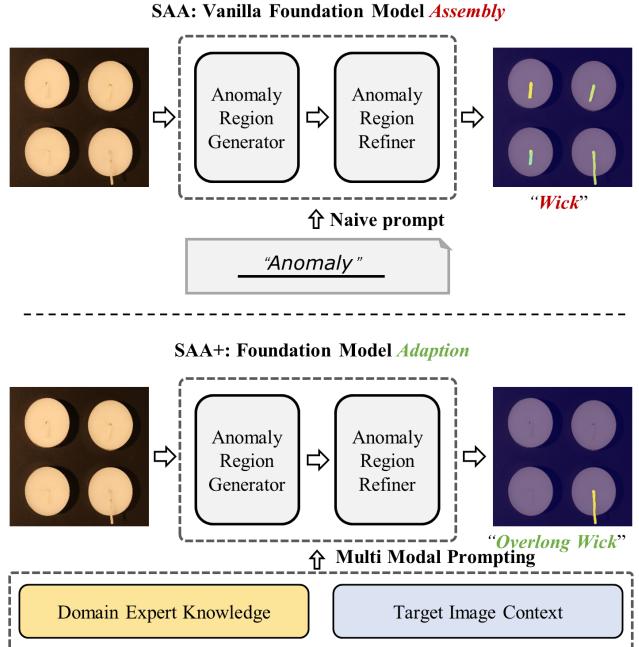


Figure 1. Towards segmenting any anomaly without training, we first construct a vanilla baseline (Segment Any Anomaly, SAA) by prompting into a cascade of anomaly region generator (e.g., a prompt-guided object detection foundation model [2]) and anomaly region refiner (e.g., a segmentation foundation model [3]) modules via a naive class-agnostic language prompt (e.g., “Anomaly”). However, SAA shows the severe false-alarm problem, which falsely detects all the “wick” rather than the ground-truth anomaly region (the “overlong wick”). Thus, we further strengthen the regularization of foundation models via multimodal prompts in the revamped model (Segment Any Anomaly +, SAA+), which successfully helps identify the anomaly region.

ties by retrieving prior knowledge stored in these models via prompting [11–15]. In this work, we first construct a vanilla baseline, i.e., Segment Any Anomaly (SAA), by cas-

cading prompt-guided object detection [2] and segmentation foundation models [3], which serve as Anomaly Region Generator and Anomaly Region Refiner, respectively. Following the practice to unlock foundation model knowledge [16, 17], naive language prompts, *e.g.*, “defect” or “anomaly”, are utilized to segment desired anomalies for a target image. In specific, the language prompt is used to prompt the Anomaly Region Generator to generate prompt-conditioned box-level regions for desired anomaly regions. Then these regions are refined in the Anomaly Region Refiner to produce final predictions, *i.e.*, masks, for anomaly segmentation.

However, as is shown in Figure 1, vanilla foundation model assembly (SAA) tends to cause significant false alarms, *e.g.*, SAA wrongly refers to all wicks as anomalies whereas only the overlong wick is a real anomaly, which we attribute to the *ambiguity* brought by naive language prompts. Firstly, conventional language prompts may become ineffective when facing the domain shift between pre-training data distribution of foundation models and downstream datasets for anomaly segmentation. Secondly, the degree of “anomaly” for a target depends on the object context, which is hard for coarse-grained language prompts, *e.g.*, “an anomaly region”, to express exactly.

Thus, to reduce the language ambiguity, we incorporate domain expert knowledge and target image context in our revamped framework, *i.e.*, Segment Any Anomaly + (SAA+). Specifically, expert knowledge provides detailed descriptions of anomalies that are relevant to the target in open-world scenarios. We utilize more specific descriptions as in-context prompts, effectively aligning the image content in both pre-trained and target datasets. Besides, we utilize target image context to reliably identify and adaptively calibrate anomaly segmentation predictions [18, 19]. By leveraging the rich contextual information present in the target image, we can accurately associate the object context with the final anomaly predictions.

## 2. Starting from Vanilla Foundation Model Assembly with Language Prompt

### 2.1. Problem Definition: Zero-shot Anomaly Segmentation (ZSAS)

The goal of ZSAS is to perform anomaly segmentation on new objects without requiring any corresponding object training data. ZSAS seeks to create an anomaly map  $\mathbf{A} \in [0, 1]^{h \times w \times 1}$  based on an empty training set  $\emptyset$ , in order to identify the anomaly degree for individual pixels in an image  $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$  that includes novel objects. The ZSAS task has the potential to significantly reduce data requirements and lower real-world inspection deployment costs.

### 2.2. Baseline: Segment Any Anomaly (SAA)

For ZSAS, we start by constructing a vanilla foundation model assembly, *i.e.*, Segment Any Anomaly (SAA), as shown in Fig. 1, which consists of an Anomaly Region Generator and an Anomaly Region Refiner.

#### 2.2.1 Anomaly Region Generator

There we base the architecture of the region detector on a text-guided open-set object detection architecture for visual grounding. Specifically, given the bounding-box-level region set  $\mathcal{R}^B$ , and their corresponding confidence score set  $\mathcal{S}$ , the module of anomaly region generator (Generator) can be formulated as,

$$\mathcal{R}^B, \mathcal{S} := \text{Generator}(\mathbf{I}, \mathcal{T}) \quad (1)$$

#### 2.2.2 Anomaly Region Refiner

To generate pixel-wise anomaly segmentation results, we propose Anomaly Region Refiner to refine the bounding-box-level anomaly region candidates into an anomaly segmentation mask set through SAM [3]. SAM accepts the bounding box candidates  $\mathcal{R}^B$  as prompts and obtain pixel-wise segmentation masks  $\mathcal{R}$ . The module of the Anomaly Region Refiner (Refiner) can be formulated as follows,

$$\mathcal{R} := \text{Refiner}(\mathbf{I}, \mathcal{R}^B) \quad (2)$$

Till then, we obtain the set of regions in the form of high-quality segmentation masks  $\mathcal{R}$  with corresponding confidence scores  $\mathcal{S}$ . We summarize the framework (SAA) as follows,

$$\mathcal{R}, \mathcal{S} := \text{SAA}(\mathbf{I}, \mathcal{T}_n) \quad (3)$$

where  $\mathcal{T}_n$  is a naive class-agnostic language prompt, *e.g.*, “anomaly”, utilized in SAA.

### 2.3. Observation: Vanilla Language Prompt Fails to Unleash the Power of Foundation Models

We present some preliminary experiments to evaluate the efficacy of vanilla foundation model assembly for ZSAS. Despite the simplicity and intuitiveness of the solution, we observe a *language ambiguity* issue. Specifically, certain language prompts, such as “anomaly”, may fail to detect the desired anomaly regions. For instance, as depicted in Fig. 1, all “wick” is erroneously identified as an anomaly by the SAA with the “anomaly” prompt. We propose introducing multimodal prompts generated by domain expert knowledge and the target image context to reduce language ambiguity, thereby achieving better ZSAS performance.

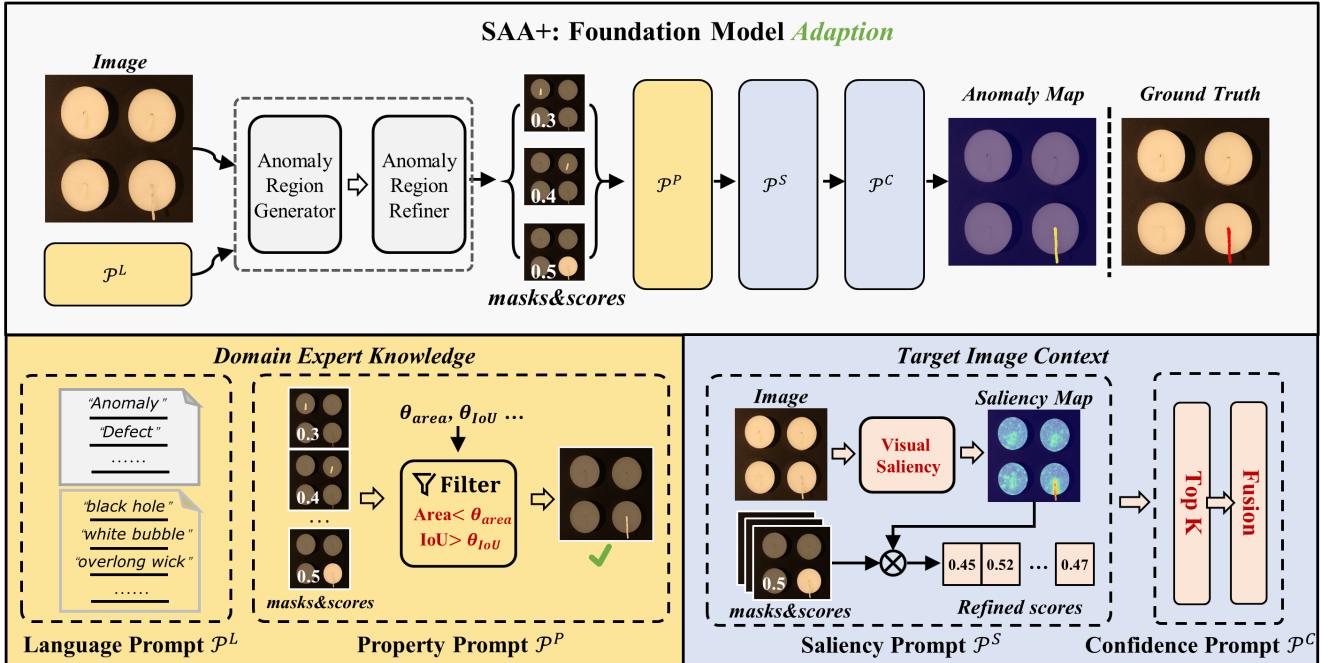


Figure 2. **Overview of the proposed Segment Any Anomaly + (SAA+) framework.** We adapt foundation models to zero-shot anomaly segmentation via multimodal prompt regularization. In specific, apart from naive class-agnostic language prompts, the regularization comes from both domain expert knowledge, including more detailed class-specific language and object property prompts, and target image context, including visual saliency and confidence ranking-related prompts.

### 3. Adapting Foundation Models to Anomaly Segmentation with Multi-modal Prompts

To address language ambiguity in SAA and improve its ability on ZSAS, we propose an upgraded version called SAA+, incorporating multimodal prompts, as shown in Fig. 2. In addition to leveraging the knowledge gained from pre-trained foundation models, SAA+ utilizes both domain expert knowledge and target image context to generate more accurate anomaly region masks. We provide further details on these multimodal prompts below.

#### 3.1. Prompts Generated from Domain Expert Knowledge

To address language ambiguity, we leverage domain expert knowledge that contains useful prior information about the target anomaly regions. Specifically, although experts may not provide a comprehensive list of potential open-world anomalies for a new product, they can identify some candidates based on their past experiences with similar products. Domain expert knowledge enables us to refine the naive “anomaly” prompt into more specific prompts that describe the anomaly state in greater detail. In addition to language prompts, we introduce property prompts to complement the lack of awareness on specific properties like “count” and “area” [20] in existing foundation models [20].

##### 3.1.1 Anomaly Language Expression as Prompt

To describe potential open-world anomalies, we propose designing more precise language prompts. These prompts are categorized into two types: class-agnostic and class-specific prompts.

**Class-agnostic prompts** ( $\mathcal{T}_a$ ) are general prompts that describe anomalies that are not specific to any particular category, e.g., “anomaly” and “defect”.

**Class-specific prompts** ( $\mathcal{T}_s$ ) are designed based on expert knowledge of abnormal patterns with similar products to supplement more specific anomaly details. We use prompts already employed in the pre-trained visual-linguistic dataset, e.g., “black hole” and “white bubble”, to query the desired regions. This approach reformulates the task of finding an anomaly region into locating objects with a specific anomaly state expression, which is more straightforward to utilize foundation models than identifying “anomaly” within an object context.

By prompting SAA with anomaly language prompts  $\mathcal{P}^L = \{\mathcal{T}_a, \mathcal{T}_s\}$  derived from domain expert knowledge, we generate finer anomaly region candidates  $\mathcal{R}$  and corresponding confidence scores  $\mathcal{S}$ .

### 3.1.2 Anomaly Object Property as Prompt

Current foundation models [2] have limitations when it comes to referring to objects with specific property descriptions, such as size or location [20], which are important for describing anomalies, such as “The small black hole on the left.” To incorporate this critical expert knowledge, we propose using anomaly property prompts formulated as rules rather than language. Specifically, we consider the location and area of anomalies.

**Anomaly Location.** Anomalies typically locate within the inspected objects. To guarantee this, we calculate the intersection over union (IoU) between the potential anomaly regions and the inspected object. By applying an expert-derived IoU threshold, denoted as  $\theta_{IoU}$ , we filter out anomaly candidates with IoU values below this threshold, retaining regions that are more likely to be abnormal.

**Anomaly Area.** The size of an anomaly, as reflected by its area, is also a property that can provide useful information. In general, anomalies should be smaller than the size of the inspected object. Experts can provide a suitable threshold value  $\theta_{area}$  for the specific type of anomaly being considered. Candidates with areas unmatched with  $\theta_{area} \cdot \text{ObjectArea}$  can then be filtered out.

By combining the two property prompts  $\mathcal{P}^P = \{\theta_{area}, \theta_{IoU}\}$ , we can filter the set of candidate regions  $\mathcal{R}$  to obtain a subset of selected candidates  $\mathcal{R}^P$  with corresponding confidence scores  $\mathcal{S}^P$  using the filter function (Filter),

$$\mathcal{R}^P, \mathcal{S}^P := \text{Filter}(\mathcal{R}, \mathcal{P}^P) \quad (4)$$

## 3.2. Prompts Derived from Target Image Context

Besides incorporating domain expert knowledge, we can leverage the information provided by the input image itself to improve the accuracy of anomaly region detection. In this regard, we propose two prompts induced by image context.

### 3.2.1 Anomaly Saliency as Prompt

Predictions generated by foundation models like [2] using the prompt “defect” can be unreliable due to the domain gap between pre-trained language-vision datasets [21] and targeted anomaly segmentation datasets [8, 22]. To calibrate the confidence scores of individual predictions, we propose Anomaly Saliency Prompt mimicking human intuition. In specific, humans can recognize anomaly regions by their discrepancy with their surrounding regions [23], *i.e.*, visual saliency could indicate the anomaly degree. Hence, we calculate a saliency map ( $\mathbf{s}$ ) for the input image by computing the average distances between the corresponding pixel feature

( $\mathbf{f}$ ) and its  $N$  nearest neighbors,

$$\mathbf{s}_{ij} := \frac{1}{N} \sum_{\mathbf{f} \in N_p(\mathbf{f}_{ij})} (1 - \langle \mathbf{f}_{ij}, \mathbf{f} \rangle) \quad (5)$$

where  $(i, j)$  denotes to the pixel location,  $N_p(\mathbf{f}_{ij})$  denotes to the  $N$  nearest neighbors of the corresponding pixel, and  $\langle \cdot, \cdot \rangle$  refers to the cosine similarity. We use pre-trained CNNs from large-scale image datasets [24] to extract image features, ensuring the descriptiveness of features. The saliency map indicates how different a region is from other regions. The saliency prompts  $\mathcal{P}^S$  are defined as the exponential average saliency value within the corresponding region masks,

$$\mathcal{P}^S := \left\{ \exp\left(\frac{\sum_{ij} \mathbf{r}_{ij} \mathbf{s}_{ij}}{\sum_{ij} \mathbf{r}_{ij}}\right) \mid \mathbf{r} \in \mathcal{R}^P \right\} \quad (6)$$

The saliency prompts provide reliable indications of the confidence of anomaly regions. These prompts are employed to recalibrate the confidence scores generated by the foundation models, yielding new rescaled scores  $\mathcal{S}^S$  based on the anomaly saliency prompts  $\mathcal{P}^S$ . These rescaled scores provide a combined measure that takes into account both the confidence derived from the foundation models and the saliency of the region candidate. The process is formulated as follows,

$$\mathcal{S}^S := \{p \cdot s \mid p \in \mathcal{P}^S, s \in \mathcal{S}^P\} \quad (7)$$

### 3.2.2 Anomaly Confidence as Prompt

Typically, the number of anomaly regions in an inspected object is limited. Therefore, we propose anomaly confidence prompts  $\mathcal{P}^C$  to identify the  $K$  candidates with the highest confidence scores based on the image content and use their average values for final anomaly region detection. This is achieved by selecting the top  $K$  candidate regions based on their corresponding confidence scores, as shown in the following,

$$\mathcal{R}^C, \mathcal{S}^C := \text{Top}_K(\mathcal{R}^P, \mathcal{S}^P) \quad (8)$$

Denote a single region and its corresponding score as  $\mathbf{r}^C$  and  $s^C$ , we then use these  $K$  candidate regions to estimate the final anomaly map,

$$\mathbf{A}_{ij} := \frac{\sum_{\mathbf{r}^C \in \mathcal{R}^C} \mathbf{r}_{ij}^C \cdot s^C}{\sum_{\mathbf{r}^C \in \mathcal{R}^C} \mathbf{r}_{ij}^C} \quad (9)$$

With the proposed multimodal prompts ( $\mathcal{P}^L, \mathcal{P}^P, \mathcal{P}^S$ , and  $\mathcal{P}^C$ ), SAA is regularized and updated into our final framework, *i.e.*, Segment Any Anomaly + (SAA+), which makes more reliable anomaly predictions.

Table 1. Quantitative comparisons between SAA+ and other concurrent methods on zero-shot anomaly segmentation. Best scores are highlighted in **bold**.

Dataset	WinClip [17]	ClipSeg [16]	UTAD [23]	SAA	SAA+
MVTec	31.65	25.42	23.48	23.44	<b>39.40</b>
VisA	14.82	14.32	6.95	12.76	<b>27.07</b>

## 4. Experiments

In this section, we first assess the performance of SAA/SAA+ on several anomaly segmentation benchmarks. Then, we extensively study the effectiveness of individual multimodal prompts.

### 4.1. Experimental Setup

**Datasets.** We leverage two datasets with pixel-level annotations: VisA [22] and MVTec-AD [8], both of which comprise a variety of object subsets, *e.g.*, circuit boards.

**Evaluation Metrics.** ZSAS performance is evaluated in terms of max-F1-pixel ( $\mathcal{F}_p$ ) [17], which measures the F1-score for pixel-wise segmentation at the optimal threshold.

**Implementation Details.** We adopt the official implementations of GroundingDINO [2] and SAM [3] to construct the vanilla baseline (SAA). Details about the prompts derived from domain expert knowledge are explained in the supplementary material. For the saliency prompts induced from image content, we utilize the WideResNet50 [25] network, pre-trained on ImageNet [24], and set  $N = 400$  in line with prior studies [23]. For anomaly confidence prompts, we set the hyperparameter  $K$  as 5 by default. Input images are fixed at a resolution of  $400 \times 400$ .

### 4.2. Main Results

**Methods for Comparison.** We compare our final model, *i.e.*, Segment Any Anomaly + (SAA+) with several concurrent state-of-the-art methods, including WinClip [17], UTAD [23], ClipSeg [16], and our vanilla baseline (SAA). For WinClip, we report its official results on VisA and MVTec-AD. For the other three methods, we use official implementations and adapt them to the ZSAS task.

**Quantitative Results:** As is shown in Table 1, SAA+ method outperforms other methods in  $\mathcal{F}_p$  by a significant margin. Although WinClip [17], ClipSeg [16], and SAA also use foundation models, SAA+ better unleash the capacity of foundation models and adapts them to tackle ZSAS. The remarkable performance of SAA+ meets the expectation to segment any anomaly without training.

**Qualitative Results:** Fig. 3 presents qualitative comparisons between SAA+ and previous competitive methods, where SAA+ achieves better performance. Moreover, the visualization shows SAA+ is capable of detecting all kinds of anomalies.

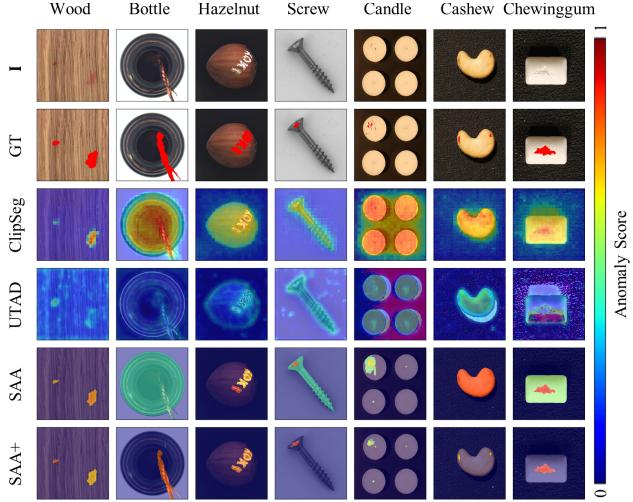


Figure 3. Qualitative comparisons on zero-shot anomaly segmentation for ClipSeg [16], UTAD [23], SAA, and SAA+ on VisA [22], and MVTec-AD [8].

Table 2. Ablation Study. Best scores are highlighted in **bold**.

Model Variants	VisA	MVTec-AD
w/o $\mathcal{P}^L$	23.29	36.49
w/o $\mathcal{P}^P$	19.28	24.43
w/o $\mathcal{P}^S$	19.39	38.79
w/o $\mathcal{P}^C$	26.70	38.68
Full Model (SAA+)	<b>27.07</b>	<b>39.40</b>

### 4.3. Ablation study

In Table 2, we perform component-wise analysis to ablate specific prompt designs in our framework, which verifies the effectiveness of all the multimodal prompts, including language prompt ( $\mathcal{P}^L$ ), property prompt ( $\mathcal{P}^P$ ), saliency prompt ( $\mathcal{P}^S$ ), and confidence prompt ( $\mathcal{P}^C$ ).

## 5. Conclusion

In this work, we explore how to *segment any anomaly* without any further training by unleashing the full power of modern foundation models. We owe the struggle of adapting foundation model assembly to anomaly segmentation to the prompt design, which is the key to controlling the function of off-the-shelf foundation models. Thus, we propose a novel framework, *i.e.*, Segment Any Anomaly +, to leverage multimodal prompts derived from both expert knowledge and target image context to regularize foundation models free of training. Finally, we successfully adapt multiple foundation models to tackle zero-shot anomaly segmentation, achieving new SoTA results on several benchmarks.

## References

- [1] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly

- without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.
- [2] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [3] 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. *arXiv preprint arXiv:2304.02643*, 2023.
- [4] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022.
- [5] Yunkang Cao, Qian Wan, Weiming Shen, and Liang Gao. Informative knowledge distillation for image anomaly segmentation. *Knowledge-Based Systems*, 248:108846, 2022.
- [6] Qian Wan, Yunkang Cao, Liang Gao, Weiming Shen, and Xinyu Li. Position encoding enhanced feature mapping for image anomaly detection. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 876–881. IEEE, 2022.
- [7] Yunkang Cao, Xiaohao Xu, Zhaoge Liu, and Weiming Shen. Collaborative discrepancy optimization for reliable image anomaly localization. *IEEE Transactions on Industrial Informatics*, pages 1–10, 2023.
- [8] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.
- [9] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69:101952, 2021.
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [11] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 105–124. Springer, 2022.
- [12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022.
- [13] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.
- [14] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. *arXiv preprint arXiv:2211.11720*, 2022.
- [15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int J Comput Vis*, 130(9):2337–2348, 2022.
- [16] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022.
- [17] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. *arXiv preprint arXiv:2303.14814*, 2023.
- [18] Xiaohao Xu, Jinglu Wang, Xiang Ming, and Yan Lu. Towards robust video object segmentation with adaptive object calibration. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1–10, 2022.
- [19] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2946–2954, 2022.
- [20] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023.
- [21] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*. Jülich Supercomputing Center, 2021.
- [22] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *Proceedings of the European Conference on Computer Vision*, 2022.
- [23] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023.
- [24] Geoffrey E Hinton, Alex Krizhevsky, and Ilya Sutskever. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25(1106–1114):1, 2012.
- [25] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.