Myriad: Large Multimodal Model by Applying Vision Experts for Industrial Anomaly Detection

Yuanze Li[1,2,*] Haolin Wang[1,2,*] Shihao Yuan[1] Ming Liu[1]
Debin Zhao[1] Yiwen Guo[3] Chen Xu[2] Guangming Shi[2] Wangmeng Zuo[1,2]
sqlyz@hit.edu.cn why_cs@outlook.com csshihao@outlook.com csmliu@outlook.com
dbzhao@hit.edu.cn guoyiwen89@gmail.com xc.xc@qq.com gmshi@xidian.edu.cn cswmzuo@gmail.com
[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2] Pazhou Lab Huangpu, Guangzhou, China
[3] Independent Researcher

Abstract

Existing industrial anomaly detection (IAD) methods predict anomaly scores for both anomaly detection and localization. However, they struggle to perform a multi-turn dialog and detailed descriptions for anomaly regions, e.g.*, color, shape, and categories of industrial anomalies. Recently, large multimodal (*i.e.*, vision and language) models (LMMs) have shown eminent perception abilities on multiple vision tasks such as image captioning, visual understanding, visual reasoning,* etc.*, making it a competitive potential choice for more comprehensible anomaly detection. However, knowledge about anomaly detection is absent in existing general LMMs, while training a specific LMM for anomaly detection requires a tremendous amount of annotated data and massive computation resources. In this paper, we propose a novel large **m**ultimodal model by appl**y**ing vision expe**r**ts for **i**ndustrial **a**nomaly **d**etection (dubbed **Myriad**), which leads to definite anomaly detection and high-quality anomaly description. Specifically, we adopt MiniGPT-4 as the base LMM and design an Expert Perception module to embed the prior knowledge from vision experts as tokens that are intelligible to Large Language Models (LLMs). To compensate for the errors and confusion of vision experts, we introduce a domain adapter to bridge the visual representation gaps between generic and industrial images. Furthermore, we propose a Vision Expert Instructor, which enables the Q-Former to generate IAD domain vision-language tokens according to the vision expert prior. Extensive experiments on MVTec-AD and VisA benchmarks demonstrate that our proposed method not only performs favorably against state-of-the-art methods under the 1-class and few-shot settings, but also provides definite anomaly prediction along with detailed descriptions in the IAD domain. IAD instruction dataset, source code, and pre-trained models will be publicly available at* https://github.com/tzjtatata/Myriad.

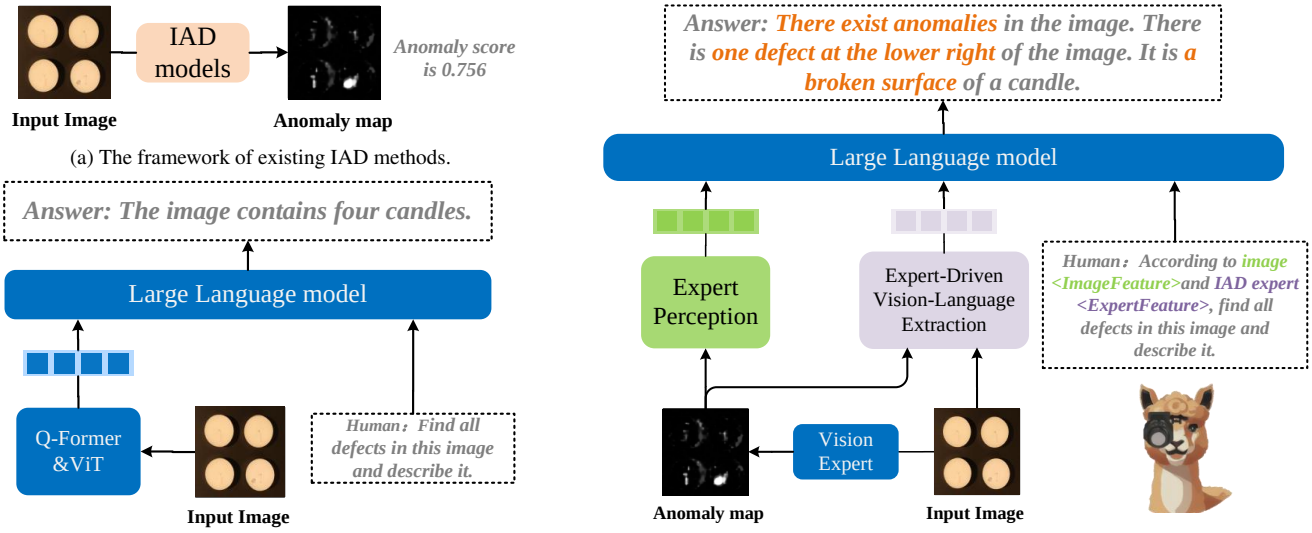
1. Introduction

Industrial anomaly detection (IAD) aims to classify and localize defects in industrial manufacturing, and plays a critical role in ensuring the efficient and reliable operation of complex industrial systems. By identifying abnormal patterns or behaviors in industrial processes, anomaly detection techniques enable timely intervention, maintenance, and optimization, thereby enhancing overall productivity and minimizing downtime.

In the recent few years, industrial anomaly detection have made significant progress, including feature-embedding based methods [6, 17, 21] and reconstruction-based methods [30, 32]. However these methods [3, 6, 17, 21, 30, 32–34] have focused on providing an anomaly score and anomaly segmentation map for each sample, leading to the over-dependence of manually selected thresholds for presenting anomalies and their potential locations. Moreover, these methods fall short in providing detailed descriptions of specific information about anomalies, such as their locations, categories, colors, and severity. Consequently, factories encounter challenges in effectively utilizing the results of these models to identify and summarize the characteristics of unseen anomalies, thereby impeding their ability to enhance manufacturing processes. Also those methods model different anomaly scenes independently, leading to weak capability of practical deployment and excessive resource consumption.

Most recently, there has been rapid advancements in large multimodel models (LMMs) built upon large language models (LLMs). Owing to their exceptional language com-

*Work done during an internship at Pazhou Lab Huangpu.



(a) The framework of existing IAD methods.

(b) The framework of MiniGPT-4.

(c) The framework of our proposed Myriad.

Figure 1. Existing IAD methods only predict anomaly maps and anomaly scores without definite results and comprehension description while LMMs like MiniGPT-4 cannot generate IAD-domain description. By incorporating pre-trained IAD models as vision experts, our proposed Myriad can perceive IAD domain knowledge via Expert Perception Sec. 3.2 and Expert-Driven Vision-Language Extraction Sec. 3.3 according to vision experts priors. Our proposed Myriad not only produce definite detection but also detailed description on anomaly content.

prehension capabilities, LLMs, *e.g.*, GPT-4, LLaMA [26], and Vicuna [4], have demonstrated the ability to perform tasks such as summarization, paraphrasing, and instruction following in zero-shot scenarios. Moreover, large multimodal models (LMMs) like MiniGPT-4 [38], LLaVA [16], and Otter [14], which build upon the foundation of LLM, exhibit a comprehensive understanding across both linguistic and visual modalities. Therefore, constructing large multimodal models for industrial anomaly detection is an effective approach to resolving the aforementioned issues in the IAD domain. However, when it comes to industrial anomaly images, the limitations arise due to the inadequacy of prior domain knowledge and corresponding multimodal datasets. Thus LMMs often face challenges in accurately detecting the presence of anomalies, such as broken on the surface of candle as shown in Fig. 1.

To constructing a large multimodal model for industrial anomaly detection, collecting a large scale dataset with industrial domain knowledge is tedious and time-consuming. Existing IAD methods estimate anomaly maps, which contain enough anomaly information, such as potential anomaly location and extent. Therefore, pre-trained IAD models can act as vision expert, and then provide prior knowledge for large multimodal model. In this paper, we propose a novelty LMM for industrial anomaly detection, **Myriad**, that leverages the extraordinary visual comprehension abilities of LMMs and the rich prior knowledge provided by pre-trained IAD models. The first challenge is make LLM comprehend anomaly segmentation maps provided by the vision expert. To solve this issue, we introduce a trainable encoder, referred to as Vision Expert Tokenizer (VE-Tokenizer), to embed the vision expert's segmentation output to tokens that LLM can understand. Vision experts sometimes exist error, which can lead to wrong response of LLM. Thus Myriad is built on MiniGPT-4 [38] as a base multimodal model to perform vision-language comprehension, which is constructed upon Vicuna with a ViT backbone from EVA-CLIP [9] and a query transformer (Q-Former) from pretrained BLIP-2 [15]. We design a domain adapter upon the vision encoder to enhance the industrial visual representation. We also propose a Vision Expert Instructor to encoder the predicted anomaly maps as query tokens, which further interact with domain-specific visual representation via cross attention in Q-Former. By considering both expert prior and its corresponding IAD domain vision-language representation, Our proposed Myriad finally produce detailed description for industrial anomaly by perceiving prior knowledge from vision expert.

Extensive experiments are conducted on both MVTec-AD and VisA datasets, which show our proposed Myriad performs favorably against than not only vision experts but also other state-of-the-art methods. Simultaneously, our proposed Myriad generates definite judgement and detailed description on industrial anomaly domain for given image.

Our contributions are summarized as follows:

- We propose a novel large vision-language model by ap-

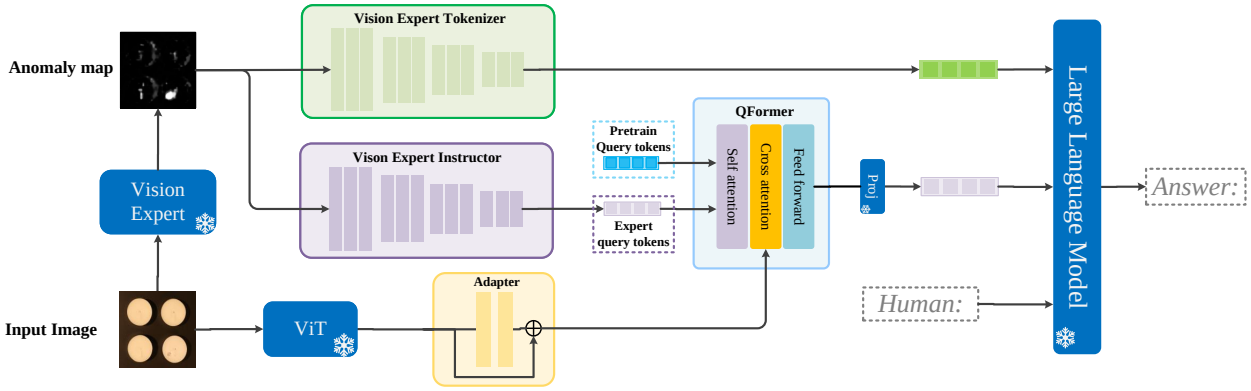


Figure 2. The architecture of our proposed Myriad. Given an input industrial image, Vision Expert estimates anomaly map which contains prior knowledge. The Vision Expert Tokenizer embeds the anomaly map into vision expert tokens to make LLM perceive prior knowledge. To compensate for the errors and confusions of vision experts, we propose Vision Expert Instructor to provide expert query tokens which enable Q-Former generate IAD domain vision-language tokens according to vision expert prior. Note that adapter is used to extract visual feature for industrial anomaly detection.

plying vision experts for industrial anomaly detection termed by Myriad. By incorporating arbitrary existing IAD models as vision experts to provide prior knowledge, Myriad produces comprehensive descriptions of IAD task.

- We design Vision Expert Tokenizer to make LMMs can receive prior knowledge. An Expert-Driven Vision-Language Extraction module is further proposed to extract domain-specific vision-language representation to further achieve accurate anomaly detection.
- Extensive experiments show that our proposed Myriad can receive prior knowledge from vision experts and further outperforms state-of-the-art methods on both MVTec-AD [1] and VisA [39] benchmarks.

2. Related Work

2.1. Industrial Anomaly Detection.

Given a industrial RGB or gray-scale images, industrial anomaly detection (IAD) aims to determinate potential anomaly and its location. Feature embedding based methods [5, 6, 17, 21, 34] has make great progress recently. [5, 6, 21] mainly extract the features of normal samples with pre-trained models [12, 23, 24] and construct a memory bank, which used to estimate anomaly extent during inference phase. Patchcore [21] introduces greedy coreset subsampling to achieve faster inference and less redundancy. Under the guidance of teacher networks (pre-trained models), the student networks [7, 25, 34] learn the normal sample representation and then the anomaly results are predicted by comparing the characteristics representation between the teacher and student networks. Pre-trained models may exhibit domain biases when applied to industrial images. SimpleNet [17] addresses this issue by mapping the pre-trained feature space to a domain-specific feature space using a fully-connected layer. Additionally, a binary discriminator is trained to detect outliers.

Reconstruction-based methods [18, 30, 32, 33] aims to reconstruct normal images from samples in the training phase. In the inference phase, the anomaly maps are predicted by comparing with original and reconstructed images pixel by pixel. Some methods [29, 31, 35] also use an unified model to achieve multiple class anomaly detection instead of separate models for each class [17, 21, 30, 33]. Language-guided methods [3, 8, 13] rely on pre-trained multimodal models, such as CLIP [20], Region CLIP [36], and Imagebind [10], which establish a foundation for visual-language integration. WinCLIP [13] and AprilGAN [3] construct two sets of prompts for abnormal and normal samples, respectively. By comparing the pixel-wise well-aligned vision-textual features with the textual features of both normal and abnormal prompts, it enables to provide the anomaly scores on all positions. AnoVL [8] design test-time adaptation (TTA) to refine features and further improve anomalies location. These models exhibit strong zero-shot or few-shot capabilities owing to the remarkable generalization ability of multi-model models.

However existing IAD methods only generate potential anomalies without determinant detection and detailed descriptions. To solve these issues, we incorporate large vision-language model for industrial anomaly detection.

2.2. Large Multimodal Models.

Motivated by the impressive cognitive abilities exhibited by Large multimodal Models [2, 15, 16, 19, 27, 28, 37, 38], researchers have embarked on investigating ways for transferring these capacities to the realm of visual perception. In this context, BLIP-2 [15] proposes the utilization of an image encoder to encode visual features, which are fed into LLMs alongside text prompts. Building upon this founda-

tion, LLaVA [16] and Mini-GPT4 [38] initially prioritize the establishment of alignment between image and text features, followed by in-context instruction tuning. Expanding on these models, Shikra [2] and Kosmos-2 [19] further enhance the grounding capabilities by referencing objects or regions of interest using definite coordinates or specialized tokens, as opposed to providing detailed textual descriptions. Nevertheless, the application of these Large Vision-Language Models (LMMs) in industrial anomaly detection encounters challenges stemming from the absence of domain-specific knowledge. Concurrent to our work, AnomalyGPT [11] designs a LLM-based image decoder to generate anomaly map and employs prompt embedding to finetune the LMM. But it still fail to utilize vision comprehension capacity of large multimodal models. In contrast, this work aims to address this limitation by pursuing the following objectives: 1) proficiently incorporating domain knowledge as embeddings alongside the textual and visual embeddings, and 2) extracting well-aligned vision-language features with the aforementioned domain knowledge.

3. Method

In this section, we first introduce the brief architecture of Myriad in Sec. 3.1, which consist of Expert Perception and Expert-Driven Vision-Language Extraction. Then we introduce design of Expert Perception in Sec. 3.2 and Expert-Driven Vision-Language Extraction in Sec. 3.3. Finally, we present more details about the training data preparation and training hyper-parameters in Sec. 3.4

3.1. Model Architecture

Previous IAD methods usually predict anomaly map and score for industrial images. These methods only detect potential anomaly defects instead of definite judgment and cannot provide important description for IAD task, such as anomaly location, extent and shape. Owing to the ability of comprehensive understanding, we introduce large multimodal models for industrial anomaly detection. However, existing LMMs are still failed to detect anomalies without large-scale IAD dataset. Previous IAD methods predict anomaly map where each pixel indicates the probability of anomaly presence ranged from 0 to 1. We argue that the anomaly map contains enough IAD prior knowledge which is essential for constructing large multimodal model for industrial anomaly detection. Thus the pre-trained IAD models can act as vision expert and extent LMM with the ability of producing industrial domain description and achieving anomaly detection.

Specifically, our proposed **Myriad** utilizes the MiniGPT-4 [38] as the base large multimodal model which is built upon Vicuna with pre-trained ViT backbone from EVA-CLIP [9] and a query transformer (Q-Former) from pre-trained BLIP-2 [15]. Given an industrial image $\mathbf{I}$, MiniGPT-4 produce general description without prior knowledge of industrial anomaly detection. To make Myriad further achieve IAD tasks, we design Expert Perception Sec. 3.2 and Expert-Driven Vision-Language Extraction Sec. 3.3 to receive the prior knowledge from vision expert. Expert Perception embeds the anomaly maps as expert tokens $\mathbf{E_t}$, which can be perceived by LLM. Expert-Driven Vision-Language Extraction enhances vision-language representation $\mathbf{E_{vl}}$ on IAD domain according to the expert prior. In this way, given a text instruction and a input industrial image $\mathbf{I}$, Myriad generates the text response including anomaly detection results and further detailed descriptions.

3.2. Expert Perception

The pre-trained IAD models predict anomaly maps, which contain sufficient prior knowledge in industrial anomaly detection. To make LLM receive above prior, we design Vision Expert Tokenizer (VE-Tokenizer). It aims to embed the anomaly maps into textual tokens which are able to be understood by LLM. As shown in Fig. 2, VE-Tokenizer contains several blocks which consist of a convolution with 3×3 kernel, a ReLU as the activation function and a max pooling, in order to map the input anomaly maps $M \in \mathbb{R}^{H\times W}$ into vision expert tokens $T \in \mathbb{R}^{D_{VE}\times D_{LLM}}$ where D_{LLM} is the dimension of LLM and D_{VE} is the amount of vision expert embeddings. In our experiments, D_{VE} is set to 9 by default which indicates a anomaly maps with 3×3 resolution. Note that the VE-Tokenizer only take anomaly map as inputs. It makes the vision expert tokens be decoupled from training domain. Therefore Myriad achieves generalization performance in unseen scenarios.

3.3. Expert-Driven Visual-Language Extraction

As MiniGPT-4 lacks of vision-language alignment for the industrial image-text pairs, the text responses of LMM excessively rely on the prediction of vision experts. Thus merging additional representation about anomaly detection into large language model is also an essential manner to achieve anomaly detection task significantly.

Here we propose Expert-Driven Vision-Language Extraction to feed vision-language representation on industrial anomaly detection into LLM. Firstly, after obtain visual features via frozen ViT from EVA-CLIP, we introduce a trainable adapter to enhance the visual representation on industrial anomaly detection. It contain a residual block with 2 convolution layers. Then taking anomaly map as input, the proposed Vision Expert Instructor generates expert query tokens tailored to original Q-former query tokens. These expert queries interacts with visual embeddings through cross-attention layers in Q-Former, to generate enhanced visual-language representation according to prior knowledge from vision experts. With expert-driven visual-language extraction, Myriad not only performs ac-

curate anomaly detection with high-quality anomaly maps with vision experts but also achieves comparable performance when the vision experts fails.

3.4. IAD Instruction Dataset

Due to the lack of detailed annotations and few amount of IAD datasets, we construct several instruction templates for the industrial anomaly detection task. For instance, we can use "*According to image <Img> <ImageFeature> <Img> and domain expert <Img> <ExpertFeature> </Img>, find out if there are defects in this image.* " where <ImageFeature> is the visual tokens produced from Expert-Driven Vision-Language Extraction and <ExpertFeature> is expert tokens extracted from Expert Perception.

To produce visual-language pairs in IAD datasets, we follow the similar anomaly simulated methods in Natural Synthetic Anomalies [22] to generate synthetic anomalies from normal samples, and the same augment hyperparameters in AnomalyGPT [11]. NSA is built upon the cut-paste method where several random regions in the source image are cut and pasted to a random position in the target image as anomalies. Note that we do not import extra datasets as the source image for NSA generation.

4. Experiments

Datasets Our experiments are performed on the MVTec-AD [1] and VisA dataset [39]. The MVTec-AD dataset [1] consists of 3629 samples in the train set and 1725 samples in the test set, respectively. MVTec-AD contains 15 sub-datasets across different types of industrial products, including 5 textual sub-datasets and 10 object sub-datasets, making it the classical dataset in industrial anomaly detection. The VisA dataset [39] contains $9,621$ normal and $1,200$ anomalous samples, and cover 12 objects. Different from MVTec-AD, there are always multiple objects, more complex background and fewer normal samples in VisA. Therefore, VisA becomes a popular benchmark with larger challenge. For both MVTec-AD and VisA, we follow standard 1-class data split setting [39], where only normal samples are visible during training and all abnormal samples are used for testing.

Evaluation Metrics Myriad can produce an anomaly map and a definite anomaly detection without anomaly scores. Therefore, we first follow the previous works to report I-AUROC and P-AUROC to measure the performance of image-level anomaly detection and pixel-level anomaly localization with the anomaly map, respectively. For full comparison between Myriad and previous IAD works, we also report the mean accuracy across all sub-datasets. Specifically, the threshold used to compute accuracy for previous IAD works is determined with the max scores of normal samples on the k-fold evaluations.

Implementation Details We utilize MiniGPT-4 as a base large multimodal model which has a Vicuna-7B/13B as inferential LLM, a pre-trained Q-former from BLIP2, the linear layer to align visual-language representation from MiniGPT-4 and a pre-trained EVA-CLIP with a VIT-B/14 backbone.

For vision experts, we utilize different vision experts under three settings: zero-shot, few-shot and typical 1-class setting. We use AprilGAN [3] as a zero-shot anomaly detector. Under the zero-shot setting, the expert and Myriad is trained on VisA when testing on MVTec-AD; When applying to VisA, we use all test and training data for training. We mainly compare with PatchCore, AnomalyGPT and PaDiM under the few-shot setting. These methods all have a memory bank but different matching strategies or augmentations with few references. For classical 1-class setting, Myriad can utilize SimpleNet, UniAD, PatchCore, PaDiM and AnomalyGPT as vision experts. We use a default wide-resnet50 as the simplenet backbone. With Simplenet, the images first resize as 329 and then center crop to 288×288 to achieve the original performance of Simplenet. With AnomalyGPT, we follow the default few-shot mode in their works, applying ImageBind as the default backbone. Note that we do not apply the rotation augmentation to the references, in order to improve inference speed when both train and test.

During training, our strategies are similar as MiniGPT-4. The whole training process lasts for 32000 steps with a batch size of 8. AdamW with 0.05 weight decay are employed as the optimizer. A linear warm-up with 400 steps starts with a small learning rate $1e^{-6}$ and reach the maximum learning rate $1e-4$. The cosine learning rate decay strategy is applied and the minimum learning rate is 0. All experiments are conducted on NVIDIA A100 GPU.

4.1. Quantitative Results

In this section, we report the quantitative comparison between Myriad and previous IAD works, including feature-embedding-based, reconstruction-based and language-guided methods. Thanks to different vision experts, Myriad can perform as a zero-shot, few-shot or 1-class anomaly detector.

1-class Industrial Anomaly Detection We first report 1-class results for MVTec-AD [1] on Table. Tab. 1. With the 1-class language-guided vision expert, we achieves superior accuracy than the state-of-the-art methods, *e.g.* outperform PatchCore by about 5.5 % (94.4 % vs. 88.9 %) and SimpleNet by about 1.4 % (94.4 % vs. 93.0 %). Specially, we use the same language-guided model as AnomalyGPT to generate anomaly maps. This expert achieves only 90.3 % accuracy on MVTec-AD datasets, while our method achieves 94.4 % accuracy and also outperform AnomalyGPT by about 1.1 % (94.4 % vs. 93.3 %).

Method	Image-AUC	Pixel-AUC	Accuracy
PaDiM	95.3	97.4	76.5
PatchCore	99.0	**98.1**	89.2
SimpleNet	**99.6**	98.1	93.0
UniAD	97.6	97.0	89.3
AnomalyGPT	97.4	93.1	93.3
Myriad (ours)	97.4	93.1	**94.4**

Table 1. 1-class anomaly detection results on MVTec-AD dataset. The best-performing method is in **bold**.

Zero-shot Industrial Anomaly Detection With the zero-shot vision experts, such as AprilGAN [3], Myriad can also perform zero-shot predictions. As shown in Table. Tab. 2, AprilGAN performs 87.6 % and 94.2 % zero-shot pixel-level AUROC on MVTec-AD and VisA respectively. With the language-guided anomaly maps, Myriad reaches 87.7 % and 81.8 % mean accuracy on MVTec-AD and VisA. Meanwhile, AprilGAN gets 69.3 % and 61.8 % accuracy with a threshold selected by k-fold evaluation.

Few-shot Industrial Anomaly Detection In few-shot setting, we use a memory-bank expert which use few samples as memory which achieves 73.3 % accuracy on VisA dataset and 82.8 % on MVTec-AD dataset with only one-shot. Experiment results are reported in Table. Tab. 2. Myriad with this few-shot vision expert improves the accuracy by about 7.2 % (80.5 % vs. 73.3 %) on VisA and about 4.6 % (87.4 % vs. 82.8 %) on MVTec-AD. We outperform the state-of-the-art methods including PaDiM, PatchCore, and AnomalyGPT. Specially, Myriad surpass AnomalyGPT with 1.3 % (87.4 % vs. 86.1 %) and 3.3 % (80.5 % vs. 77.4 %) on MVTec-AD and VisA respectively. Further, we compare Myriad on 2-shot and 4-shot setting and still perform the comparable or outperform accuracy to the state-of-the-art methods.

4.2. Qualitative Examples

The qualitative results on an example of objects sub-datasets in MVTec-AD are shown in Fig. 3. Myriad provides a definite judgement (**exist**), the category and number of the defect (**a missing piece**), and the fine-grained visual perception (**shadow cast**). Note that in MVTec-AD, the shadow cast is not defined as an anomaly. Previous IAD methods may ignore this unseen anomaly, while Myriad can customize its behavior via adding prompts before the instructions, such as "*You can ignore the shadow cast when detect anomalies.*"

We further examine Myriad on the texture sub-datasets in MVTec-AD, as shown in Fig. 4. A part of a leather is demonstrated and a huge cut locates on the right of the image. Myriad not only predicts the existence of the defect,

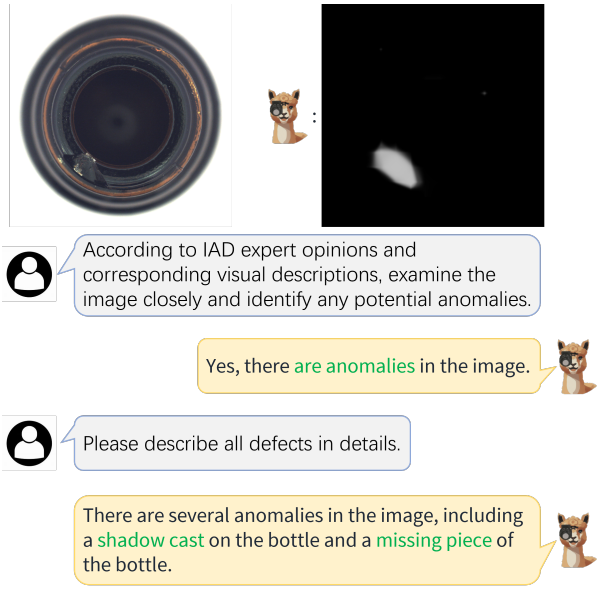


Figure 3. Qualitative results on an example of object sub-datasets in MVTec-AD. Myriad performs the ability of anomaly description.

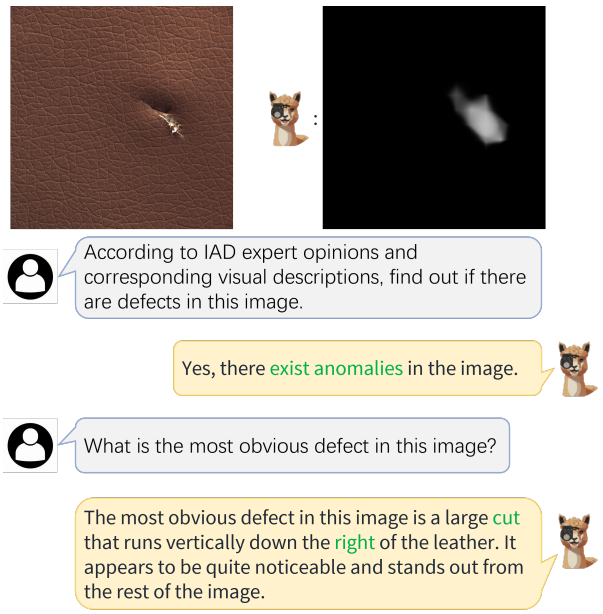


Figure 4. Qualitative results on an example of texture sub-datasets in MVTec-AD. Myriad provides the definite anomaly detection and sufficient descriptions.

but also provide the category (**cut**), location (**the right of the image**, the kind of products (**leather**) and the severity (**large**). Advanced customization can be done.

We also show a normal hazel nut, in order to illustrate the adequate information from Myriad. Although the rough surface of the nut can be a large source of interference for

Setup	Method	MVTec-AD			VisA		
		Image-AUC	Pixel-AUC	Accuracy	Image-AUC	Pixel-AUC	Accuracy
0-shot	AprilGAN	86.2 (± 0.5)	87.6 (± 0.1)	69.3 (± 0.4)	78.0 (± 0.4)	94.2 (± 0.3)	61.8 (± 0.2)
	Myriad (ours)	86.2 (± 0.5)	87.6 (± 0.4)	**87.7** (± **0.2**)	78.0 (± 0.4)	94.2 (± 0.3)	**81.8** (± **0.7**)
1-shot	SPADE	81.0 (± 2.0)	91.2 (± 0.4)	-	79.5 (± 4.0)	95.6 (± 0.4)	-
	PaDiM	75.5 (± 1.0)	90.0 (± 0.4)	57.4 (± 1.6)	59.6 (± 1.8)	91.3 (± 0.3)	45.0 (± 0.6)
	PatchCore	84.2 (± 1.2)	92.4 (± 0.5)	65.0 (± 1.8)	76.8 (± 1.6)	93.5 (± 0.6)	60.0 (± 1.0)
	WinCLIP	93.1 (± 2.0)	95.2 (± 0.5)	-	83.8 (± 4.0)	**96.4** (± **0.4**)	-
	AnomalyGPT	**94.1** (± **1.1**)	**95.3** (± **0.1**)	86.1 (± 1.1)	**87.4** (± **0.8**)	96.2 (± 0.1)	77.4 (± 1.0)
	Myriad (ours)	**94.1** (± **1.1**)	**95.3** (± **0.1**)	**87.4** (± **0.9**)	**87.4** (± **0.8**)	96.2 (± 0.1)	**80.5** (± **1.2**)
2-shot	SPADE	82.9 (± 2.6)	92.0 (± 0.3)	-	80.7 (± 5.0)	96.2 (± 0.4)	-
	PaDiM	78.2 (± 0.6)	92.1 (± 0.4)	56.9 (± 0.8)	65.5 (± 1.5)	93.2 (± 0.1)	46.4 (± 0.7)
	PatchCore	87.1 (± 0.8)	94.1 (± 0.2)	68.4 (± 2.3)	80.4 (± 0.7)	95.0 (± 0.2)	61.8 (± 1.2)
	WinCLIP	94.4 (± 1.3)	**96.0** (± **0.3**)	-	84.6 (± 2.4)	**96.8** (± **0.3**)	-
	AnomalyGPT	**95.5** (± **0.8**)	95.6 (± 0.2)	84.8 (± 0.8)	**88.6** (± **0.7**)	96.4 (± 0.1)	77.5 (± 0.3)
	Myriad (ours)	**95.5** (± **0.8**)	95.6 (± 0.2)	**85.4** (± **0.7**)	**88.6** (± **0.7**)	96.4 (± 0.1)	**78.9** (± **0.5**)
4-shot	SPADE	84.8 (± 2.5)	92.7 (± 0.3)	-	81.7 (± 3.4)	96.6 (± 0.3)	-
	PaDiM	80.9 (± 0.9)	94.0 (± 0.2)	57.9 (± 1.2)	69.6 (± 1.5)	94.4 (± 0.1)	48.0 (± 1.4)
	PatchCore	89.5 (± 1.3)	94.9 (± 0.2)	72.5 (± 1.8)	82.2 (± 0.8)	96.0 (± 0.1)	63.1 (± 0.4)
	WinCLIP	95.2 (± 1.3)	96.2 (± 0.3)	-	87.3 (± 1.8)	**97.2** (± **0.2**)	-
	AnomalyGPT	**96.3** (± **0.3**)	**96.2** (± **0.1**)	85.0 (± 0.3)	**90.6** (± **0.7**)	96.7 (± 0.1)	77.7 (± 0.4)
	Myriad (ours)	**96.3** (± **0.3**)	**96.2** (± **0.1**)	**85.3** (± **0.3**)	**90.6** (± **0.7**)	96.7 (± 0.1)	**78.3** (± **0.3**)

Table 2. Few-shot IAD results on MVTec-AD and VisA datasets. Results are listed as the average of 5 runs and the best-performing method is in **bold**. The results for SPADE, PaDiM, PatchCore and WinCLIP are reported from [13]. The models are trained and evaluated on across datasets from MVTec-AD and VisA.

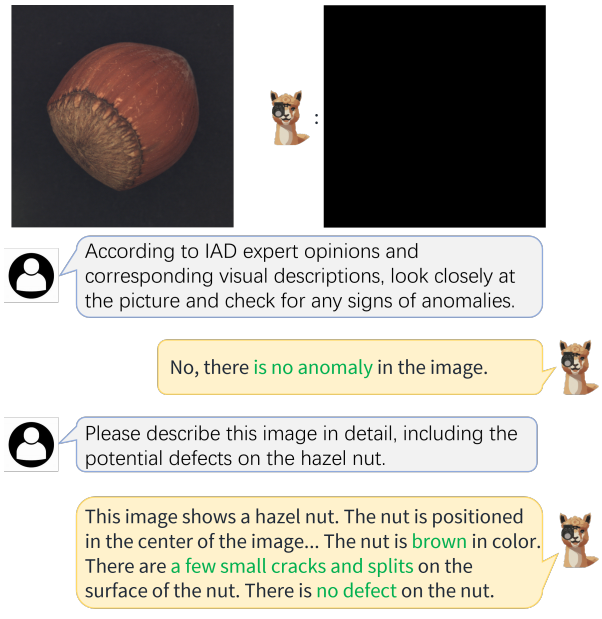


Figure 5. Qualitative results on a normal sample in MVTec-AD.

anomaly detection, our model still give the right answer and further offer the description of the image.

Examples on VisA are illustrated in Fig. 6 and Fig. 7. In Fig. 6, the vision expert provides a coarse anomaly map for the broken on the chewing gum. Myriad can understand the suggestion of vision experts and point out that there is only one hole in the chewing gum. When the vision expert fails to provide correct anomaly map, Myriad can still figure out the anomaly. As shown in Fig. 7, there is a small broken on the upper left macaroni which is ignored by the vision expert. Still Myriad outputs accurate responses which not only the category of defect (**crack**), but also the position (**upper left**). More qualitative examples will be reported in our appendix.

4.3. Ablation Studies

In this section, we conduct ablation studies to explore each components in our proposed Myriad. As shown in Table. Tab. 3, VE-Tokenizer with MiniGPT-4 have a acceptable performance as 93.2 % on MVTec-AD 1-class anomaly detection and 74.0 % on VisA one-shot setting. While adding adapter, VE-Tokenizer still bootstrap its 1-

Adapter	VE-Tokenizer	VE-Instructor	MVTec-AD (1-class)	VisA (1-shot)
			Accuracy	Accuracy
	✓		93.2	74.0
✓			89.5	65.0
✓	✓		94.0	75.0
✓		✓	92.0	72.0
✓	✓	✓	**94.4**	**80.5**

Table 3. Results of ablation studies on architecture.

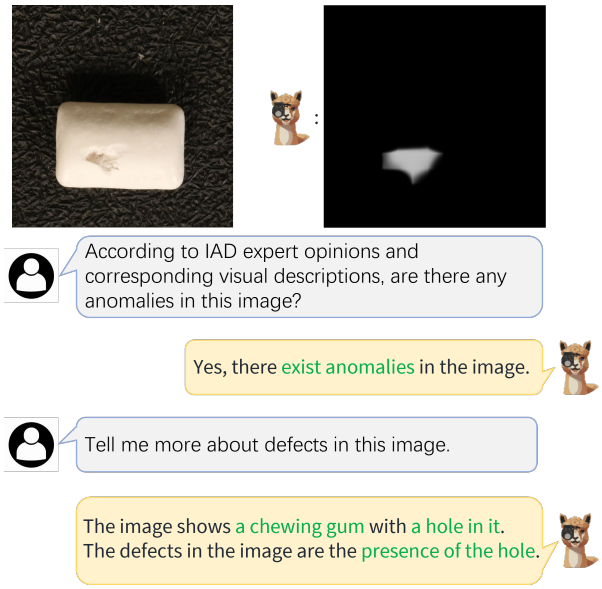


Figure 6. Qualitative results on VisA, with good segmentation maps from vision experts.

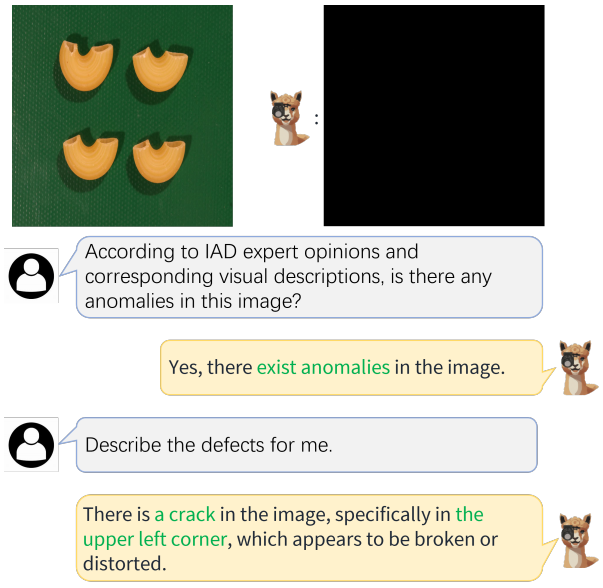


Figure 7. Qualitative results on VisA, with bad prediction from vision experts.

class performance from 89.5 % to 94.0 % and one-shot accuracy near 10.0 % (75.0 % vs. 65.0 %). Thus VE-Tokenizer can transform the potential anomaly areas into prior knowledge which is intelligible to LLM. By providing IAD visual feature via domain adapter, our proposed Myriad can further improve detection accuracy. In the meantime, the performances in both 1-class and one shot accuracy drop without VE-Tokenizer and VE-Instructor separately. It demonstrate that IAD domain vision-language representation and prior knowledge from vision expert contribute to industrial anomaly detection task complementarily. Above variants prove that receiving IAD prior knowledge from vision expert is crucial method for constructing IAD large multimodal model.

5. Conclusion

Existing industrial anomaly detection methods only predict potential anomaly results and fail to provide definite judgement and detailed description for anomaly information. In this paper, we propose a novelty large multimodal model, dubbed Myriad, to achieve industrial anomaly detection task. By introducing estimated anomaly maps from pre-trained IAD methods, our Myriad perceive IAD prior knowledge without constructing large-scale IAD datasets. We first design Vision Expert Tokenizer to encode anomaly map into expert tokens and feed them into large language models. As vision experts sometimes make mistakes, we further propose Expert-Driven Vision-language Extraction to provide domain-specific vision-language tokens for LMMs. Specifically, We design Vision Expert Instructor to generate expert query tokens, which used for extracting vision-language embeddings via interacting with visual features through cross attention in Q-Former. Experiments show that our proposed Myriad not only achieves superior performance than both vision experts and state-of-the-art methods, but also provide detailed description for industrial anomaly detection.

References

[1] 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.

[2] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[3] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[5] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[6] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021.

[7] 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 (CVPR)*, pages 9737–9746, 2022.

[8] Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2023.

[9] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[11] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. *arXiv preprint arXiv:2308.15366*, 2023.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[14] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[17] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023.

[18] Fanbin Lu, Xufeng Yao, Chi-Wing Fu, and Jiaya Jia. Removing anomalies as noises for industrial defect localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16166–16175, 2023.

[19] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306, 2023.

[20] 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.

[21] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

[22] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer, 2022.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[25] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24511–24520, 2023.

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[27] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023.

[28] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023.

[29] Xincheng Yao, Ruoqi Li, Zefeng Qian, Yan Luo, and Chongyang Zhang. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6803–6813, 2023.

[30] Haonan Yin, Guanlong Jiao, Qianhui Wu, Borje F Karlsson, Biqing Huang, and Chin Yew Lin. Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection. *arXiv preprint arXiv:2307.08059*, 2023.

[31] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

[32] Hui Zhang, Zheng Wang, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Denoising diffusion for anomaly detection. *arXiv preprint arXiv:2303.08730*, 2023.

[33] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023.

[34] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023.

[35] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3924–3933, 2023.

[36] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[37] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*, 2023.

[38] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[39] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022.