Zero-Shot learning in industrial scenarios: New Large-Scale Benchmark, Challenges and Baseline

Zekai Zhang[1], Qinghui Chen[1], Maomao Xiong[1], Shijiao Ding[1], Zhanzhi Su[2], Xinjie Yao[3], Yiming Sun[4], Cong Bai[5], Jinglin Zhang[1]*

[1]School of Control Science and Engineering, Shandong University, Jinan, China
[2] Qilu University of Technology, Jinan, China
[3] College of Intelligence and Computing, Tianjin University, Tianjin, China
[4] School of Automation, Southeast University, Nanjing, China
[5] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
{202420810, 202420785, 202314866, 202234879}@mail.sdu.edu.cn, yaoxinjie@tju.edu.cn, sunyiming@seu.edu.cn, congbai@zjut.edu.cn, jinglin.zhang@sdu.edu.cn

Abstract

Large Visual Language Models (LVLMs) have achieved remarkable success in vision tasks. However, the significant differences between industrial and natural scenes make applying LVLMs challenging. Existing LVLMs rely on user-provided prompts to segment objects. This often leads to suboptimal performance due to the inclusion of irrelevant pixels. In addition, the scarcity of data also makes the application of LVLMs in industrial scenarios remain unexplored. To fill this gap, this paper proposes an open industrial dataset and a Refined Text-Visual Prompt (RTVP) for zero-shot industrial defect detection. First, this paper constructs the Multi-Modal Industrial Open Dataset (MMIO) containing **80K+** samples. MMIO contains diverse industrial categories, **including 6 super categories and 18 subcategories**. MMIO is **the first large-scale multi-scenes pre-training dataset for industrial zero-shot learning**, and provides valuable training data for open models in future industrial scenarios. Based on MMIO, this paper provides a RTVP specifically for industrial zero-shot tasks. **RTVP has two significant advantages:** First, this paper designs an expert-guided large model domain adaptation mechanism and designs an industrial zero-shot method based on Mobile-SAM, which enhances the generalization ability of large models in industrial scenarios. Second, RTVP automatically generates visual prompts directly from images and considers text-visual prompt interactions ignored by previous LVLM, improving visual and textual content understanding. RTVP achieves **SOTA** with **42.2%** and **24.7%** AP in zero-shot and closed scenes of MMIO.

Introduction

Product defect detection tasks in industrial scenarios play an important role in ensuring the safety of users. Expert models (Wang, Bochkovskiy, and Liao 2022; Ge et al. 2021; Li et al. 2022a) in industrial scenarios usually use single-modal data from a single field and strictly follow class-aware methods, which limits the ability of model to process multi-scene data and generalize to open datasets. Recently, the development of Segment Anything series (SAMs) (Kirillov et al. 2023; Zhao et al. 2023; Zhang et al. 2023a) has shown powerful interactive and strong zero-shot capabilities in remote sensing, medicine, and other fields. The uniqueness of SAM lies in the design of human-computer interactive prompts, which allows segmentation based on user-provided point, line, and box prompts.

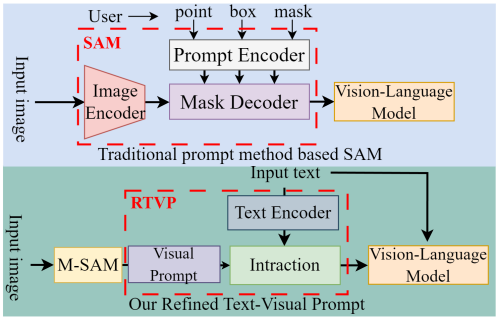

Figure 1: Comparison between traditional prompting methods and RTVP. RTVP solves the subjectivity of traditional manual prompts and introduces text to further refine the semantics.

However, since the environment, scale and distribution of images in industrial scenes are significantly different from natural scenes (**Figure 2-a**), there are many significant challenges in applying SAM's pre-training-prompt paradigm in industrial scenes. As shown in **Figure 1**, the existing SAM prompt mode needs to rely on the user's prompts (point, box, mask) (Kirillov et al. 2023) to segment objects when processing complex scenes. The user's familiarity will significantly affect specific prompts' effect and introduce irrelevant or noisy pixels.

To address the above problems, some methods (Liu et al. 2023c; Zhang et al. 2023b) combine semantic models (Xie et al. 2017; He et al. 2016) to obtain pseudo masks of objects. CPT (Yao et al. 2024) and ReCLIP (Subramanian et al. 2022) use visual prompt to establish relationships between instances. On above basis, Hu et al. (Hu, Xu, and Shi 2023) designed a sampling strategy to extract points and bounding

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

boxes from the pseudo template as prompts for SAM to segment the object image. CoCoOp (Yan et al. 2023) turns the image-generated prompt into a conditional input and dynamically combines it with the language prompt. These methods ignore false positives in pseudo masks and rely hyperparameter sensitivity. Therefore, it heavily depends on the quality of pseudo masks and has poor generalization ability. To address the above problems, this paper proposes a Refined Text-visual prompt (RTVP), which improves the zero-shot capability of VLM in industrial scenarios. Based on the zero-shot of Mobile-SAM (Zhang et al. 2023a) in natural scenes, RTVP further enhances its generalization ability in industrial scenes. As shown in **Figure 1**, considering the subjectivity and noise of user prompts, RTVP introduces expert-guided mechanism based on Mobile-SAM to generate coarse-grained segmentation features and encode the segmentation features into a low-dimensional space. Following the unique sparse nature of industrial images, this paper perform spatial-channel pixel activation on the segmentation features, extract uncertainty score of the object. Then, this paper introduces a sparse modelling sample selection strategy to extract the semantic clues from the enhanced features through the uncertainty score and obtain a refined Visual Prompt. Finally, the refined Visual Prompt is interacted with the text prompt to generate prompt embeddings of semantically specific objects. Based on the inherent capabilities of Mobile-SAM, RTVP enhances the model's visual-language understanding and generalization capabilities, especially in open-world and cross-domain scenarios.

Another challenge to applying SAM in industrial scenarios is that the existing data in industrial inspection is single, and it is impossible to find a unified multi-domain generalized industrial scene dataset. To solve the above problem, this paper creates a large-scale Multi-Modal Industrial Open dataset called MMIO-80K. MMIO consists of more than 80K+ samples converted from 18 different industrial defect datasets, including various product defects in major industrial categories such as metallurgy, automobile manufacturing, precision electronics, textiles, daily necessities, and wood processing. MMIO is tailored for the unique feature distribution in industrial zero-shot detection, effectively alleviating the lack of industrial domain expertise of LVLMs. To the best of our knowledge, MMIO-80K is the first open dataset proposed for industrial zero-shot detection, and MMIO-80K can catalyze the development of LVLMs in industrial openness.

In the detection task of MMIO closed scenes, RTVP significantly improved mAP compared to YOLOv8, YOLOv9 and other field expert models. In the zero-shot detection task in MMIO, RTVP surpasses most benchmark models. Generalization experiments on COCO and LVIS datasets also show that RTVP can generalize detection in natural scenes. In summary, the core contributions of this paper are as follows:

•**MMIO-80K:** To the best of our knowledge, this paper constructed the first object detection data set MMIO-80K for industrial open scenarios. MMIO-80K consists of more than 80K samples, effectively alleviating the lack of domain expertise in industrial open scenarios.

•**RTVP:** This paper proposes refined learnable text-visual prompts to improve the zero-shot detection capability of visual-language models in industrial scenarios. RTVP does not require users to provide specific prompts for each image, which reduces the user's usage burden and noise and effectively improves the knowledge and understanding ability of LVLMs in industrial domains.

•**Superior performance:** Extensive experiments show that RTVP has superior performance in industrial open scene detection. Therefore, RTVP represents a significant advancement in LVLMs for industrial detection, providing a general method for mutual visual language understanding in industrial scenarios.

Related Work

Application of Vision-Language Models

In recent years, pre-trained Large language models, such as GPT-4 (Achiam et al. 2023), Llava (Liu et al. 2024), etc., have shown strong zero-shot learning capabilities in natural language processing. Subsequently, pre-trained visual-language models such as CLIP (Radford et al. 2021), BLIP-2 (Li et al. 2023), etc., have been extended to computer vision. Currently, there are two methods to apply large pre-trained models. One is to use the segmentation results of large pre-trained models as prior information to assist downstream tasks, which requires additional intermediate layer fine-tuning of model. For example, Ahmadi et al. (Ahmadi et al. 2023) used the segmentation results of SAM as prior information in crack and other defect detection. Wu et al. (Wu et al. 2023) inserted the Adapter module into SAM for medical image segmentation tasks. Another method uses prompts to guide the pre-trained model transfer to the object domain. For example, Xu et al. (Xu et al. 2023) proposed an untrained evidence prompt generation method, incorporating human prior information into prompts. Zhang et al. (Jie and Zhang 2023) proposed generation network for the shadow detection to generate dense point prompts.

The above two methods easily consume computing resources and cannot guarantee the training effect of the domain layer. In addition, the visual prompts are not refined and do not consider the importance of text prompts. In contrast, in terms of generating prior information, this paper introduces an expert model to assist Mobile-SAM. In terms of prompts, this paper uses refined text-visual prompts to provide richer object semantic information.

Prompt for Zero-shot Learning

Prompt technology originated from NLP. The prompt was subsequently used to guide zero-shot learning of large pre-trained models. However, prompts often rely on artificial features, leading to user burden and noise introduction. Recently, automated prompt training methods have been widely used in zero-shot tasks. For example, AutoPrompt (Shin et al. 2020) used gradient-based methods to generate prompt templates automatically. With the development of large models, fine-grained visual prompts are widely used in zero-shot detection scenarios. For example, CPT (Yao et al. 2024) introduces coloured object boxes as markers on

Dataset	Classes	Number	Scene Category	Modal
MVTEC AD(Bergmann et al. 2020)	15	5,354	15	Image
BTAD(Mishra et al. 2021)	3	2,830	3	Image
VisA(Fernando et al. 2020)	12	10,821	12	Image
Ind(Zhu et al. 2024)	30	600,000	11	Image
MS-COCO(Lin et al. 2014)	80	118,000	1	Image
VOC2007(Everingham et al. 2015)	20	9,963	1	Image
MMIO-80K	100	21,836	18	Text,Image

Table 1: Comparison of MMIO with zero-shot industrial defect dataset and general dataset.

the image. However, the prompts contain a lot of noise. To solve this problem, FGVP (Yang et al. 2024), VRP-SAM (Sun et al. 2024) etc., proposed refined visual prompts.

However, these methods rely on the adaptability of SAM and cannot iteratively optimize the quality of prompts. In addition, these methods ignore the role of text prompts in zero-shot tasks. In contrast, this paper expresses more refined semantically specific object features through continuous optimized text-visual interaction prompts.

Multi-Modal Industrial-Open Dataset

Dataset construction

This paper creates a large-scale Multi-Modal industrial open dataset named MMIO-80K. MMIO consists of more than 80K samples converted from 18 different major industrial scenes. Among them, the glass container dataset related to daily necessities is a subset collected by this paper. **Figure 2-b** shows the glass container image acquisition equipment. This paper designs a multi-camera collaboration mechanism for cylindrical transparent objects such as glass container. Specifically, the conveyor belt conveys the container to the photoelectric gate to activating the three cameras on the left to take pictures. Then, the container will rotate 180 degrees, and the three cameras on the right will repeat the above process. A total of 625 pictures were collected for object detection and annotation. Each image has a resolution of 700 $\times$ 820 and has five categories ('oil', 'black spot', 'bubble', 'plastering thread', and 'quenched grain').

This paper obtains images and labels for the remaining 17 industrial scenarios through various enterprise open-source data. This paper delivers over 22,000 defect images initially summarized to professional technicians for quality screening, eliminates duplicate and blank images, and obtains 21,836 defect images in different fields. This paper re-adds attributes to each image based on experts' advice. This paper divides the defect into 6 super categories according to the main scenarios of intelligent manufacturing, and each super category contains multiple scene data. Since MMIO contains 80K+ defect entities, it is difficult to manually correspond the real category to the defect. Therefore, this paper uses CLIP to extract the semantic vector of each category and the cosine distance to calculate the similarity score between the semantic vector and the image-related semantics. In addition, the text-image matching pairs with the highest similarity are filtered by a threshold of 0.8. Finally, experts check the effect of text-image matching linked to 100 categories. We modified some similar objects. The number of samples of the same industrial product may be relatively

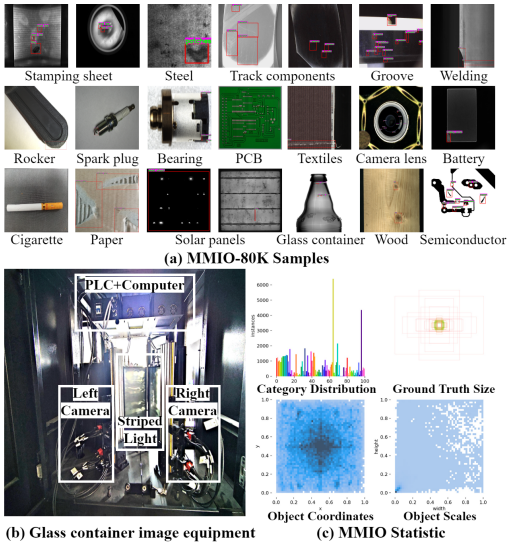

Figure 2: MMIO statistical analysis. (a) MMIO example, which includes more than 18 types of industrial scenes. (b) The proposed Glass container image equipment. (c) MMIO data statistics results.

small, and there may be similar defects: such as holes and irregular holes, but they still have similar categories. We divide them into seen and unseen categories to distinguish different categories of the same industrial product. In the end, MMIO has 21,836 images and 100 categories.

Dataset characteristics, statistics, and challenges

Features and statistics: MMIO has 6 super categories and 100 attributes. **Figure 2-c** summarizes the distribution of MMIO categories and instance sizes. Different product types and manufacturing in various industrial fields will produce different defects, among which small-sized defects account for the vast majority. In particular, MMIO stands out with rich attribute annotations, covering a wide range of standard industrial manufacturing categories, making it particularly suitable for the complex task of industrial zero-shot detection. The 100 classes in MMIO can be divided into 6 super categories: metallurgy, automobile manufacturing, precision electronics, textiles, daily necessities, and wood processing. Examples of each category are shown in **Figure 2-a**. **Table 1** compares general zero-shot industrial defect comprehensive datasets and general datasets, which indicates that MMIO has significant advantages in the number of scenes and semantic annotations. To our knowledge, MMIO is the first multi-scene defect dataset for industrial zero-shot detection. Unlike only nature scene general object detection datasets such as MS-CCOCO (Lin et al. 2014) and VOC2007 (Everingham et al. 2015), MMIO specializes in text annotation enhancement for zero-shot learning tasks, providing a more challenging and relevant benchmark for zero-shot learning.

Dataset processing: MMIO is divided into a visible class training dataset and an invisible class test set for zero-shot tasks. Among them, the visible class contains 18,811 im-

ages and annotations, and the invisible class contains 3025 images and annotations. This paper also provides a training test set for closed scenario tasks, with a test and training set ratio of 20% and 80%.

Dataset challenges: MMIO faces material changes, imaging limitations, and internal-external interference. First, the product materials vary greatly, resulting in drastic scale changes in MMIO. The defect detection model needs to perceive object areas of different scales sensitively. Secondly, environmental interference and intra-class correlation make the discriminant information of defects unclear, which requires the model to be robust. The third is the long-tail distribution problem. Industrial scenarios need to seek balanced detection effects in extremely unbalanced category distributions.

Proposed Method

Problem Definition

The goal of the zero-shot framework is to recognize objects that have never appeared in the domain under pre-training with domain-specific text $Y \subseteq \varphi^N = \{t^1, ..., t^N\}$ and image $I \subseteq \delta^N = \{I^1, ..., I^N\}$. This paper provides a training dataset D_s containing image-text pairs of C_s visible categories. Let $z_s = \{1, ..., C_s\}$ and $z_u = \{1, ..., C_u\}$ are the label sets of visible and invisible categories, respectively. $z_s \cap z_u = \phi$. $D = D_s + D_u$ is the image-label space set of visible and invisible classes. Let the text set $Y = y_s + y_u$. During training, the model extracts the semantic information of y_s contained z_s and accurately matches z_s to the relevant area of image I. A test set D_t contains D_s and D_u in the zero-shot stage. The goal of the zero-shot task is to optimize a model from D_s and detect the invisible category C_u in D_u through the user-defined invisible text prompt y_u (y_u contains the semantic information of z_u). For the visible category, this paper tests the D_s category accuracy in D_t to measure the effect of the visible class.

Refined Text-Visual Prompt

Figure 3 shows the detailed structure of the zero-shot framework proposed in this paper. Traditional zero-shot methods usually use RoIAlign (He et al. 2017) or manually set points, lines, boxes and other prompts to obtain regional prompts. However, the coarse feature area leads to excessive noise, and the prompt cannot be iteratively optimized. The RTVP includes three important innovations: expert-assisted domain adaptation, refined visual prompter, and cross-modal prompt interaction, significantly improving the understanding ability of Mobile-SAM in industrial scenarios.

Expert-assisted Domain Adaptation: Although Mobile-SAM has strong zero-shot capabilities, obtaining refined visual prompt is still challenging. The challenge is that Mobile-SAM trained in natural scenes makes it difficult to transfer to industrial scenes effectively. Previous methods usually insert learnable layers into SAM to make it conform to the feature distribution of the target domain (Wu et al. 2023; Cheng et al. 2023). However, above methods typically do not effectively supervise the adaptation layer, resulting in a poor migration effect. To solve the above problems, this paper proposes expert-assisted domain adaptation (**Figure 3-b**). Expert-assisted domain adaptation uses an expert CNN architecture trained in industrial scenes to provide expert knowledge to Mobile-SAM. The principle is that the convergence speed of the pre-trained expert CNN in industrial scenes is faster than the domain adaptation layer of Mobile-SAM, so it can quickly give Mobile-SAM adequate supervision to achieve domain transfer. The expert model is not frozen during training because the expert CNN converges quickly, and the iteratively optimized expert CNN can better provide domain expertise.

In practice, given a CNN-based domain expert model M_{Expert}, this paper first pre-trains it on visible categories in the training dataset D_s to obtain the expert model multi-scale feature $F_j, j \subseteq \{1, 2, 3, 4\}$. To enhance the multi-scale information of industrial defects, this paper performs multi-scale sampling on the output feature of Mobile-SAM to obtain the multi-scale feature $F_i, i \subseteq \{1, 2, 3, 4\}$. As shown in **Figure 3-b**, this paper masks and reconstructs the multi-scale features from Mobile-SAM to improve the robustness of domain adaptation. Unlike MAE (He et al. 2022), this paper reconstructs the domain expert information instead of the original input. This paper divides the features into N groups according to the patch size to obtain the neighbourhood multi-scale feature $F_i^N, (N \subseteq 2^k, k \subseteq N^+)$. Then, this paper uniformly masks part of the patch according to the random sampling method, and the two-dimensional index of each patch is defined as $P_i^c \subseteq [(P_1^x, P_1^y), ..., (P_N^x, P_N^y)]$. Randomly sample M of the indexes to obtain a binary mask and multiply it with F_i to get the masked multi-scale feature F_i^M. The formula for the above process is as follows:

$$\begin{aligned} Mask_{x,y}^{\mathrm{j}} &= \begin{cases} 0, (x, y \in \mathrm{S}) \\ 1, Otherwise \end{cases} \\ F_i^M &= P\left(F_i\right) \times Mask_{x,y}^{\mathrm{j}} \end{aligned} \tag{1}$$

Among them, $s = P_i^c$ is the range of the extracted M indexes. $Mask_{x,y}^{\mathrm{j}}, j \subseteq \{1, 2, 3, 4\}$ is a binary mask, and P is the image segmentation operation. After obtaining the F_i^M, this paper builds a decoder to reconstruct the masked multi-scale features to get F_i^R. F_i^R can be considered a coarse-grained visual prompt. The mask decoder consists of three layers of convolution and deconvolution. The formula is as follows:

$$F_i^R = \textstyle\sum_1^3 Relu\left(Conv\left(F_i^M\right)\right) \sum_1^3 Relu\left(Conv^T\left(F_i^M\right)\right) \tag{2}$$

To solve the problem of insufficient optimization of the existing methods, this paper introduces a multi-scale domain optimization function to optimize the domain adaptation layer. It is worth noting that this paper will stop optimizing the domain adaptation layer when the optimisation function is infinitely small. The optimization function is as follows:

$$f_{optimization} = \sum_{j=1}^{4} \sum_{i=1}^{4} \left\| F_j - F_i^R \right\| \tag{3}$$

Refined Visual Prompt: This paper uses Mobile-SAM guided by domain expert knowledge to generate coarse-grained visual prompts automatically and establishes a

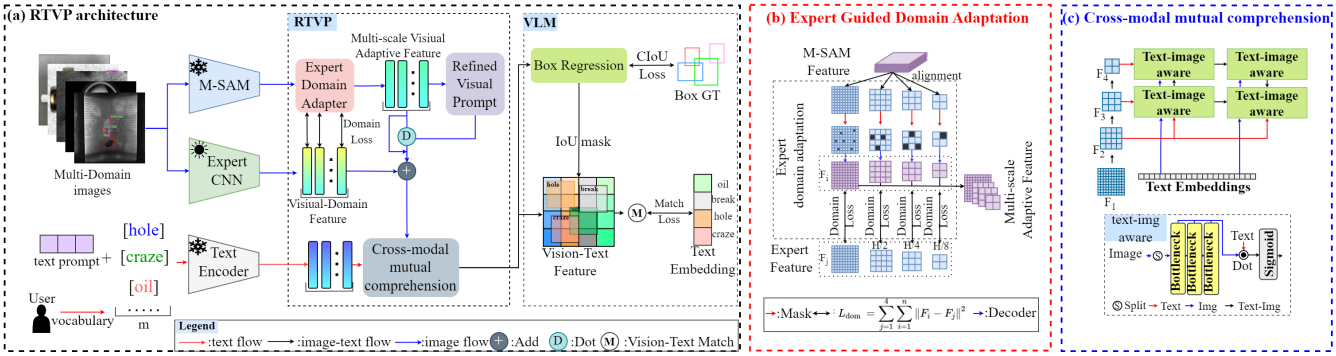

Figure 3: TRVP framework. (a) The specific architecture of the RTVP. (b) This paper constructs a CNN expert model to assist Mobile-SAM in rapid domain adaptation. Refined Visual Prompt establishes a sample selection strategy to generate refined visual prompts. (c) The interaction between text and refined visual prompts deepens VLM's semantic understanding of relevant areas of the image.

sparse modelling sample selection strategy to obtain more refined visual prompts. Refined Visual Prompt can more accurately highlight target instances, reduce background interference, and retain global knowledge. Specifically, based on the coarse-grained visual prompt F_i^R obtained by expert-assisted domain adaptation, the sparse modelling sample selection strategy activates the coarse-grained visual prompt through a learnable scorer to obtain an uncertainty score (M_k) for each pixel. The formula is as follows:

$$M_k = \varphi_{\mathrm{i}}^{\mathrm{s}}\left(\varphi_{\mathrm{i}}^{\mathrm{c}}\left(F_i^R\right)\right) \tag{4}$$

Among them, $\varphi_{\mathrm{i}}^{\mathrm{c}}$ represents channel activation and $\varphi_{\mathrm{i}}^{\mathrm{s}}$ represents spatial activation. i represents different features. The formula of $\varphi_{\mathrm{i}}^{\mathrm{c}}$ and $\varphi_{\mathrm{i}}^{\mathrm{s}}$ are as follows:

$$\begin{aligned}\varphi_{\mathrm{i}}^{\mathrm{c}} &= \left(\sigma \mathrm{Q_k}\left(\mathrm{W_j}\left(\mathrm{X_C}^{mean}\left(F_i^R\right)\right)\right) + \sigma \mathrm{Q_k}\left(\mathrm{W_j}\left(\mathrm{X_C}^{\max}\left(F_i^R\right)\right)\right)\right.\\ \varphi_{\mathrm{i}}^{\mathrm{s}} &= \sigma\left(Cat\left[\mathrm{X_s}^{mean}\left(F_i^R\right); \mathrm{X_s}^{\max}\left(F_i^R\right)\right]\right)\end{aligned} \tag{5}$$

Among them, σ represents the activation function sigmoid, $\mathrm{Q_k} \in [\mathrm{Q_1}, \mathrm{Q_2}], \mathrm{W_j} \in [\mathrm{W_1}, \mathrm{W_2}]$ belong to the complementary multi-layer activation perception weights,X_c^{mean} represents the channel direction average, X_c^{max} represents the channel direction maximum, X_s^{mean} represents the spatial direction average, X_c^{max} represents the spatial direction maximum, and Cat represents channel superposition. The uncertainty score reflects the area with the most information. Traditional uncertainty methods usually select the samples with the highest scores. However, samples with significant uncertainty scores are not the optimal feature. Therefore, this paper creates a sparse sampling mechanism. Specifically, industrial images have significant sparse characteristics, and the sparse sampling mechanism selects the most high-frequency areas in the image to reduce the redundancy of irrelevant features. This paper uses patch segmentation with different neighbourhood sizes to retrieve the uncertainty score M_k. Specifically, given a set of segment patches $N_{h,w}^p$, a sparse selection mechanism is used to select high-frequency pixels. This paper performs pixel-level mean on the feature space inside the $N_{h,w}^p$. Then, completes sparse sample selection by selecting the top-k pixels, which pixel values inside the patch are more significant than the mean. The process can be described as the following formula:

$$\mathrm{M}_{prompt}^{\mathrm{i,j}} = \left(Argmax\left(\mathrm{N}_{h,\mathrm{w}}^{\mathrm{p}}.mean, \mathrm{D}_{\mathrm{i,j}}^{\mathrm{cs\ s}}\right)\right), \left(\mathrm{i,j} \in \mathrm{N}_{h,\mathrm{w}}^{\mathrm{p}}\right) \tag{6}$$

Among them, $D_{i,j}^{c,s}$ is any pixel in $N_{h,w}^p$, $N_{h,w}^p.mean$ represents the pixel-level mean in the neighbourhood space, and $M_{prompt}^{i,j}$ is the selected sparse sample. After multiple selections, $M_{prompt}^{i,j}$ can be considered as a refined visual prompt. The sparse sampling mechanism can more accurately describe the contours and other details of the object. However, the lack of adequate supervision in the sparse selection mechanism causes a small number of selected feature points to deviate from the object. Therefore, this paper introduces CIoU-optimized feature activation to make the selected pixels fall within the ground truth. The optimization mechanism uses an additional detection head for regression prediction and continuously optimizes the visual prompt through CIoU.

Cross-modal Interactive Visual-Text Prompt: To further refine the semantic features related to the object, this paper interacts the text prompt with the visual prompt to refine the visual prompt in the multi-scale feature extraction. As shown in **Figure 3-c**, given the text embedding T_i from text encoder and the image feature $\mathrm{F_L} \in \mathrm{R^{C\times H\times W}}(\mathrm{L} \in \{1,2,3,4\})$. This paper adopts multi-scale image features and aggregates the text features into the image features using the maximum Sigmoid attention query text-image matching semantic features. The formula is as Eq-7, $F_{img-text}$ represent the RTVP.

$$\mathrm{F}_{img-text} = \mathrm{F_L} \times Sigmoid\left(Argmax\left(\mathrm{F_L} \times \mathrm{T_i^T}\right)\right)^{\mathrm{T}} \tag{7}$$

Experiments

In this section, ablation studies and comparative experiments are performed to demonstrate the effectiveness of RTVP. For more experimental analysis, please refer to the **supplementary materials**.

Dataset and Evaluation Metrics

This paper conducts experiments on MMIO-80K, MSCOCO, and LVIS datasets. For the MMIO zero-

(a) MMIO Closed Scenarios				
Method(general detection)	AP ↑	AP_{50} ↑	Precision ↑	Recall ↑
YOLOv8-S (ultralytics 2023)	39.9	64.5	66.9	67.2
YOLOv8-M	40.7	67.5	71.4	64.4
YOLOv9 (Wang, Yeh, and Liao 2024)	41.1	72.1	71.7	69.2
YOLOv10-S (Wang et al. 2024)	41.7	69.2	70.8	67.0
Method(defect detection)	AP ↑	AP_{50} ↑	Precision ↑	Recall ↑
SSGD (Han et al. 2023)	38.9	69.3	68.7	66.4
LiteYOLO-ID (Li et al. 2024)	40.0	67.9	69.2	70.6
LF-YOLO (Liu et al. 2023a)	41.1	72.6	66.9	70.6
Steel det (Božič, Tabernik, and Skočaj 2021)	37.8	70.7	71.6	69.3
RTVP	**42.4**	**76.1**	**74.8**	**72.6**
(b) MMIO Zero-shot				
Method	Params ↓	AP ↑	AP_{50} ↑	AP_{75} ↑
MDETR (Kamath et al. 2021)	169M	18.1	17.7	20.3
GLIP-T (Li et al. 2022b)	232M	19.6	30.1	22.2
GLIPv2-T	232M	22.4	31.8	25.4
DetCLIP-T (Yao et al. 2022)	155M	22.9	24.7	26.6
YOLO-W-S (Cheng et al. 2024)	**77M**	21.2	32.1	24.6
Grounding DINO-T (Liu et al. 2023b)	172M	20.2	29.4	23.9
RTVP	131M	**24.7**	**35.7**	**27.3**

Table 2: Comparison with expert models on MMIO closed scenarios and Zero-shots.

shot task, this paper split MMIO into 65 base classes for training and 35 novel classes for testing. For the MMIO closed task, this paper split MMIO into a training-test set ratio of 80% and 20%. To evaluate generalization, this paper performs closed-scene validation on COCO and zero-shot validation on LVIS. This paper uses the COCO and LVIS metrics to measure the model's accuracy.

Implementation Details

The experimental model is built using PyTorch 2.0.1, and the hardware environment is 4 Nvidia RTX 4090 GPUs. The model is trained 200 epochs using AdamW and a cosine learning rate scheduler with a batch size of 32. The input image size is 640. The initial learning rate is 2e-3, the weight decay is 5e-4, the momentum is 0.9, the text encoder (CLIP-T-Text) and Mobile-SAM-T are frozen during pre-training.

Comparison with the State-of-the-art

MMIO closed scenario: Table 2-a compares general and expert defect models in MMIO closed scenarios. RTVP achieves SOTA in multi-scale AP and Recall, which indicates that RTVP is highly sensitive to multi-scale and long-tail distribution data sets. Compared with the traditional expert model's user training-detection mode, RTVP allows users to customize vocabulary and automatically generate refined prompts, providing accuracy that traditional expert models cannot provide. Compared with the recently proposed YOLOv10, AP of RTVP is improved by 1%, and Recall is improved by 5%. Experiments show that RTVP can also enhance the accuracy in closed scenarios. Beacuse of RTVP introduces expert model and industrial expert text labels to convey richer knowledge to VLM.

MMIO Zero-shot: Table 2-b compares the common zero-shot methods. Compared with Grounding DINO-T, AP and AP_{50} are improved by 4.5% and 6.3%, respectively. For the latest YOLO-World-s, AP and AP_{50} are improved by 3.5% and 3.6%, respectively. Experiments show that RTVP can achieve the best performance with the litter parameters. This is because RTVP considers the sparse characteristics of industrial scenarios, which is helpful for industrial zero-shot tasks. RTVP uses Mobile-SAM and uncertainty sparse modelling to obtain refined visual prompt, and the interaction between text and visual prompt helps VLM achieve better zero-shot performance. It is worth noting that the AP of all methods is very low, indicating that MMIO's tasks are more challenging and valuable and will promote subsequent industrial scenario zero-shot tasks.

Method	AP ↑	AP_s ↑	AP_m ↑	AP_l ↑
PP-YOLOE-S (Xu et al. 2022)	43.0	23.2	46.4	56.9
PP-YOLOE-M	**49.0**	28.6	52.9	63.8
YOLOv8-S (ultralytics 2023)	44.9	-	-	-
YOLOv9-S (Wang, Yeh, and Liao 2024)	46.7	26.6	56.0	64.5
YOLOv9-M	51.1	33.6	57.0	**68.0**
YOLOv10-S (Wang et al. 2024)	46.3	-	-	-
RTVP	47.2	**35.2**	**58.3**	65.4

Table 3: Comparative experiment on MS-COCO. This paper uses YOLOv8 as the baseline to test the RTVP.

Method	Pre-trained Data	AP ↑	AP_r ↑	AP_c ↑	AP_f ↑
GLIP-T (Li et al. 2022b)	O365	17.8	13.5	12.8	22.2
YOLO-W (Cheng et al. 2024)	O365	23.5	16.2	21.1	27.0
ViLD (Gu et al. 2021)	O365	-	-	20.0	**28.3**
RTVP	O365	**24.1**	**17.3**	**22.4**	27.9
Grounding-Dino-T (Liu et al. 2023b)	O365,GlodG	25.6	14.4	19.6	**32.2**
GLIP-T (Li et al. 2022b)	O365,GlodG	24.9	17.7	19.5	31.0
YOLO-W-s (Cheng et al. 2024)	O365,GlodG	24.2	16.4	21.7	27.8
RTVP	O365,GlodG	**26.8**	**19.5**	**23.4**	30.7

Table 4: Zero-shot experiments of LVIS. This paper uses different datasets for pre-training.

Comparison with COCO: This paper verifies the generalization ability of RTVP in closed scenes on COCO (Lin et al. 2014). To achieve a fairer fully supervised comparison, this paper uses YOLOv8 as the baseline for retesting. The results of MS-COCO are shown in **Table 3**. Compared with YOLOv8s, AP increased from 44.9% to 47.2%. RTVP achieved the best in AP_s and AP_m, demonstrating that RTVP has the generalization ability to improve the accuracy of expert models in closed environments.

Comparison with LVIS: This paper evaluates the zero-shot generalization ability of RTVP on LVIS (**Table 4**). Specifically, RTVP is pre-trained on the Object365 (Shao et al. 2019) and GOLDG datasets with YOLO-world as the baseline and fine-tuned on LVIS base. Compared with the baseline, RTVP improves by 0.6% in AP, proving that RTVP's refined visual prompt improves VLM's zero-shot scene understanding ability. Compared with other models, RTVP's AP is still improved. With the increase of pre-training data, the performance of RTVP has improved, indicating that pre-training with a large amount of data improves accurate prompt expression.

Ablation Study

Component ablation study: Table 5 shows the results of different component ablation study on MMIO. Expert domain adaptation can effectively improve the accuracy of zero-shot and closed scenes. Because the expert model can provide expert knowledge supervision to Mobile-SAM, it enhances the migration effect in the industrial field. The refined visual prompt is more conducive to zero-shot industrial detection. Because it uses sparse modelling of industrial images to help focus on key features. Cross-modal text-visual interaction is conducive to further refinement of semantic

Method(Closed)	AP ↑	AP_{50} ↑	AP_{75} ↑
-Expert Domain Adapter	30.2	61.8	29.1
-Refined Visual Prompt	38.6	72.9	35.4
-Text-Visual Interation	40.0	73.6	36.8
RTVP	42.4	76.1	38.2
Method(Zero-shot)	AP ↑	AP_{50} ↑	AP_{75} ↑
-Expert Domain Adapter	18.5	27.4	18.8
-Refined Visual Prompt	20.0	33.7	23.9
-Text-Visual Interation	22.8	31.6	25.1
RTVP	24.7	35.7	27.3

Table 5: Ablation study on MMIO closed scenes and zero-shots.

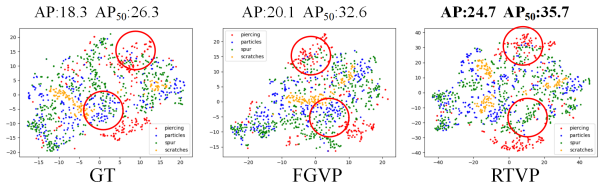

Figure 4: t-SNE visualization in the MMIO zero-shot scenario. RTVP can generate more compact feature representations.

features related to objects, which is benefit to VLM identifying invisible class features accurately.

Different prompts: This paper tests the zero-shot detection effect of the ground truth as the visual prompt and FGVP (Fine-Grained Visual Prompting) on zero-shot task. Among them, FGVP (Yang et al. 2024) is reproduced with YOLO-World as the baseline. As shown in **Figure 4**, this paper uses t-SNE (Van der Maaten and Hinton 2008) to visualize the image features of four industrial invisible categories on MMIO and the output features before the detector. Compared with using the ground truth and FGVP (only visual prompt), the features of RTVP show clear clusters. Experiments demonstrate the importance of refined visual prompt and text prompt interactions. As shown in AP, RTVP has the highest AP on MMIO, indicating that RTVP produces more obvious features, generates well-separated clusters for different classes, and promotes the learning of invisible classes.

Different patch sizes: The sparse modelling mechanism uses patches of different sizes to select high-frequency features. As shown in **Table 6**, the patch size significantly affects the zero-shot task's accuracy. This is because the sparsity of industrial defects leads to obvious high-frequency features of defects. The smaller the patch, the more conducive it is to narrow the retrieval range, and it is easier to find high-frequency features.

Patches Size	AP ↑	AP_{50} ↑	AP_{75} ↑
32	24.7	35.7	27.3
64	21.3	34.1	22.5
128	19.9	33.0	21.4

Table 6: Comparison of different patch sizes in MMIO zero-shot scenario.

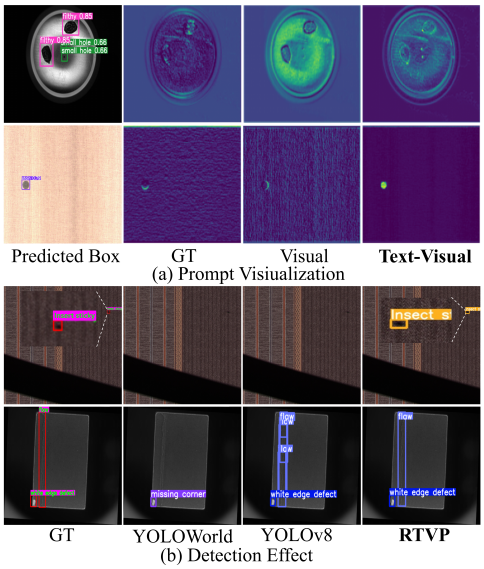

Figure 5: RTVP Visualization. (a) The effect of different prompts on the feature space of VLM. The yellow area has high attention. (b) Detection effect.

Qualitative Results

Prompt visualization: This paper maps and visualizes the VLM feature space after prompting. **Figure 5-a** shows the impact of different prompts on the VLM feature space. Directly using the ground truth will introduce noise, which can make it difficult for VLM to understand the key areas of defects. The Visual Prompt feature saliency map shows that RTVP can significantly activate the high-frequency areas of defects, proving the effectiveness of the sparse modelling mechanism. The feature saliency map of the Text-Visual Prompt shows that introducing text features can further refine the semantic information related to the defect.

Detection result visualization: Figure 5-b visualizes the detection results of YOLOv8s, YOLOWorlds, and RTVP. The visualization results show that the other methods are prone to missed and imprecise detection. In contrast, RTVP effectively avoids the above problems.

Conclusion

The application of LVLMs in the industrial field is challenging due to the significant domain differences. To address the lack of professional data, this paper constructs the Multi-Modal Industrial Open Dataset (MMIO). MMIO contains diverse product defect data from major industrial categories, including 6 super categories and 18 subcategories. Based on MMIO, this paper provides Refined Text-Visual prompt (RTVP) for zero-shot tasks in industrial open scenarios. Using expert-guided domain transfer, RTVP enhances the generalization ability of Mobile-Segment Anything (Mobile-SAM) in industrial scenarios. Secondly, RTVP proposes a text-visual interaction method to promote cross-modal mutual matching and understanding, which improves the ability to understand visual and textual content. Experiments

on MMIO show the strong zero-shot capability of RTVP in industrial scenarios. Future work will develop in multiple directions, such as anomaly detection, segmentation, and knowledge question answering.

Acknowledgements

This work was supported in part by the Key Research and Development Program of Shandong Province of China under Grant 2023CXGC010112, in part by the National Key Research and Development Program of China under Grant 2022YFB4500602, in part by the Distinguished Young Scholar of Shandong Province under Grant ZR2023JQ025, in part by the Taishan Scholars Program under Grant tsqn202211290, , in part by the National Natural Science Foundation of China under Grant No.62076215, and the Jiangsu University Qing Lan Project, and in part by the Major Basic Research Projects of Shandong Province under Grant ZR2022ZD32.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmadi, M.; Lonbar, A. G.; Sharifi, A.; Beris, A. T.; Nouri, M.; and Javidi, A. S. 2023. Application of segment anything model for civil infrastructure defect assessment. *arXiv preprint arXiv:2304.12600*.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023. Sam-med2d. *arXiv preprint arXiv:2308.16184*.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16911.

Everingham, M.; Eslami, S. M. A.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1): 98–136.

Fernando, T.; Gammulle, H.; Denman, S.; Sridharan, S.; and Fookes, C. 2020. Deep Learning for Medical Anomaly Detection – A Survey.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding YOLO Series in 2021. *arXiv preprint arXiv:2107.08430*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Han, H.; Yang, R.; Li, S.; Hu, R.; and Li, X. 2023. SSGD: A smartphone screen glass dataset for defect detection. In *ICASSP*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *International Conference on Computer Vision*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, X.; Xu, X.; and Shi, Y. 2023. How to efficiently adapt large segmentation model (sam) to medical images. *arXiv preprint arXiv:2306.13731*.

Jie, L.; and Zhang, H. 2023. AdapterShadow: Adapting Segment Anything Model for Shadow Detection. *arXiv preprint arXiv:2311.08891*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1780–1790.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. 2022a. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*.

Li, D.; Lu, Y.; Gao, Q.; Li, X.; Yu, X.; and Song, Y. 2024. LiteYOLO-ID: A Lightweight Object Detection Network for Insulator Defect Detection. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–12.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *COMPUTER VISION - ECCV 2014, PT V*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. ISBN 978-3-319-10602-1; 978-3-319-10601-4. 13th European Conference on Computer Vision (ECCV), Zurich, SWITZERLAND, SEP 06-12, 2014.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, M.; Chen, Y.; Xie, J.; He, L.; and Zhang, Y. 2023a. LF-YOLO: A lighter and faster yolo for weld defect detection of X-ray image. *IEEE Sensors Journal*, 23(7): 7430–7439.

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

Liu, Y.; Zhu, M.; Li, H.; Chen, H.; Wang, X.; and Shen, C. 2023c. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*.

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.

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.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Subramanian, S.; Merrill, W.; Darrell, T.; Gardner, M.; Singh, S.; and Rohrbach, A. 2022. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*.

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.

ultralytics. 2023. YOLOv8. *[Online]. Available: https://github.com/ultralytics/yolov8*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; and Ding, G. 2024. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Wang, C.-Y.; Yeh, I.-H.; and Liao, H.-Y. M. 2024. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*.

Wu, J.; Ji, W.; Liu, Y.; Fu, H.; Xu, M.; Xu, Y.; and Jin, Y. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, S.; Wang, X.; Lv, W.; Chang, Q.; Cui, C.; Deng, K.; Wang, G.; Dang, Q.; Wei, S.; Du, Y.; and Lai, B. 2022. PP-YOLOE: An evolved version of YOLO.

Xu, Y.; Tang, J.; Men, A.; and Chen, Q. 2023. Eviprompt: A training-free evidential prompt generation method for segment anything model in medical images. *arXiv preprint arXiv:2311.06400*.

Yan, J.; Xie, Y.; Guo, Y.; Wei, Y.; Zhang, X.; and Luan, X. 2023. CoCoOpter: Pre-train, prompt, and fine-tune the vision-language model for few-shot image classification. *International Journal of Multimedia Information Retrieval*, 12(2): 27.

Yang, L.; Wang, Y.; Li, X.; Wang, X.; and Yang, J. 2024. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36.

Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; and Xu, H. 2022. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35: 9125–9138.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2024. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5: 30–38.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023a. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

Zhang, R.; Jiang, Z.; Guo, Z.; Yan, S.; Pan, J.; Ma, X.; Dong, H.; Gao, P.; and Li, H. 2023b. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*.

Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; and Wang, J. 2023. Fast Segment Anything. arXiv:2306.12156.

Zhu, B.; Chen, Y.; Tang, M.; and Wang, J. 2024. Pixel-Level Contrastive Pretrainer for Industrial Image Representation. *IEEE Transactions on Instrumentation and Measurement*, 73.