Available Online: 8 September 2025 DOI: 10.54254/2977-3903/2025.26796

LoRA fine-tuned Qwen2.5-VL large model for accurate description and location of steel surface defects

Jingliang Liu

University of Electronic Science and Technology of China, Chengdu, China

JayLiu090@163.com

Abstract. In industrial manufacturing, accurate and efficient identification of product surface defects is essential for ensuring product quality, optimizing the production process and reducing cost. However, complex and diverse defect morphologies and the need for fine-grained description present significant challenges. General image description methods based on large visual language models often struggle to provide accurate defect type and location information for specific areas such as steel surface defect recognition. To address this, a defect identification method for the Qwen2.5-VL-3B large model based on LoRA fine-tuning is proposed. We built a specialized dataset covering six key steel surface defects—cracks, impurities, plaques, pitting, scale penetration, and scratches—and refined the model through efficient low-rank adaptation. Experimental results demonstrate that the fine-tuned Qwen2.5-VL-3B model significantly improves industrial defect recognition, accurately identifying defect types and locations, thus overcoming limitations of general large models and providing an efficient solution for industrial inspection.

Keywords: steel surface defects, defect recognition, vision large model, LoRA fine tuning, defect location

1. Introduction

Industrial defect detection is crucial for manufacturing quality control, especially in steel production, but faces challenges like defect complexity, poor image quality, data scarcity, and limited generalization [1,2]. Conventional approaches often lack efficiency and accuracy. While deep learning methods outperform them in feature extraction, they demand substantial data and computing resources. General Visual Language Models (VLMs) often struggle with specific defect identification and localization [3]. This study proposes a novel VLM defect detection method using Qwen2.5-VL-3B with Low-Rank Adaptation (LoRA). This approach leverages pre-trained VLMs and efficient LoRA fine-tuning for industry-specific tasks [4,5]. Our core innovations are:

First, the integration of Qwen2.5-VL-3B with LoRA enables the generation of detailed textual descriptions of defect types and precise locations. This offers human-interpretable defect information, improving quality control and analysis beyond simple detection. Second, LoRA fine-tuning substantially enhances Qwen2.5-VL-3B's capacity to detect industrial defect characteristics, delivering more precise identification and localization. This addresses data scarcity, a major hurdle for traditional deep learning. Third, LoRA drastically cuts the computational cost and time for fine-tuning large VLMs. This allows practical implementation in resource-limited industrial environments and supports quick adaptation to new scenarios. This research demonstrates a strategic synergy of VLMs and LoRA to overcome data limitations and computational challenges, paving the way for more adaptable AI in complex industrial environments [3].

2. Related work

Industrial defect detection demands high precision [1,2]. Traditional methods, though simple, suffer from manual feature engineering and poor generalization. Deep learning, notably CNNs, automates feature extraction, offering superior accuracy. However, it is data-hungry, computationally intensive, often lacks generalization to new lines, and its "black box" nature hinders interpretability, especially for tiny or fuzzy defects [1].

Large pre-trained VLMs like Qwen2.5-VL-3B offer general vision/language understanding and few/zero-shot learning capabilities, reducing annotation needs. However, VLMs need task-specific fine-tuning—a computationally intensive process—and may generate erroneous targets ('target illusion'). Parameter-Efficient Fine-Tuning (PEFT) methods, particularly LoRA, mitigate these costs by drastically reducing trainable parameters and memory, speeding adaptation [6]. While efficient, LoRA can

present challenges in multitask learning or optimal parameter selection. The field evolved from traditional techniques to deep learning, with CNN-based methods dominating [1,2]. Despite progress in classification, detection, and segmentation, deep learning still faces limitations in data efficiency and interpretability. In summary, industry demands precise, adaptable, and resource-efficient defect detection [1,2]. Existing methods fall short. Combining powerful VLMs like Qwen2.5-VL-3B with efficient PEFT methods like LoRA offers a promising solution. This approach enables lightweight adaptation to learn subtle defect features using limited data/compute, leveraging general knowledge for complex scenarios, and balancing generalization with task specificity.

3. Methods

To address data scarcity and computational constraints in industrial defect detection, a PEFT-based VLM method is proposed, leveraging Qwen2.5-VL-3B with LoRA for lightweight fine-tuning. This chapter details the overall network architecture, key modules, and loss function.

3.1. Overall network architecture

The proposed framework utilizes a pre-trained Qwen2.5-VL-3 backbone, largely frozen to retain its general vision-language understanding, adapted for specific industrial tasks via LoRA modules.

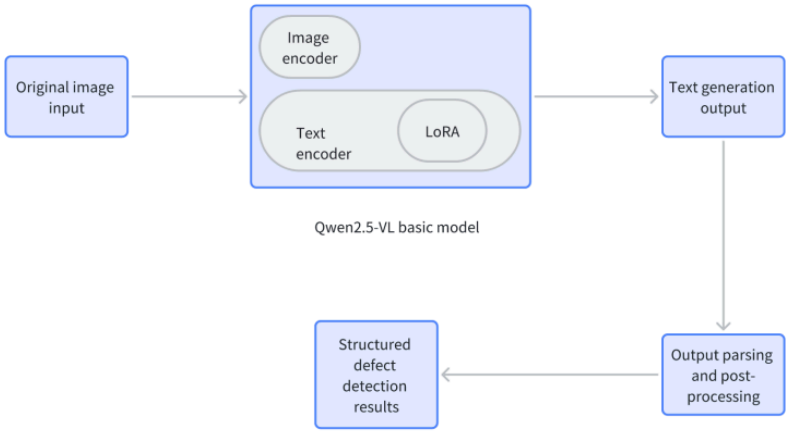

Figure 1. Qwen 2.5-VL model

The architecture includes (1) the Qwen2.5-VL-3B backbone, featuring visual and speech encoders for multimodal processing; (2) LoRA adapter layers, injected into critical model layers, which are solely updated during training to significantly reduce trainable parameters; and (3) a task-specific output layer that generates descriptive text, subsequently post-processed (e.g., with regular expressions) to extract structured defect data. As shown in Figure 1, an industrial image and text instruction are input, and processed by the LoRA-adapted Qwen2.5-VL-3B, generating a textual defect description.

3.2. Module 1: Qwen2.5-VL-3B visual language foundation model

The core of the method is the Qwen2.5-VL-3B Visual Language Foundation Model, pre-trained on massive image-text pairs. Its multimodal capabilities enable it to understand and generate natural language, associating it with visual information. For steel defect detection, Qwen2.5-VL-3B delivers three key advantages: robust visual feature extraction, accurate localization through multimodal analysis, data efficiency via strong generalization, and interpretable text descriptions of defect type/location/severity. The model is implemented as follows:Define Qwen2.5-VLForConditionalGeneration and AutoProcessor classes

Load the Qwen2.5-VL large language model from a pre-trained path

Configure model to load onto available devices, and set data type

Load the corresponding processor from the pre-trained path

To retain general capabilities, the majority of pre-training weights are frozen during fine-tuning.

3.3. Module 2: efficient fine tuning of LoRA parameters

To efficiently adapt Qwen2.5-VL-3B for defect detection, this paper adopts LoRA. Its core idea involves freezing the pre-trained weight matrix W_0 and introducing two smaller trainable matrices A and B such that the weight update becomes ΔW=BA, making the new weight:

$$W = W_0 + BA \tag{1}$$

This slashes trainable parameters, reducing memory/compute demands while preserving prior knowledge. We implement LoRA in critical linear layers by loading the pre-trained model, configuring hyperparameters (rank, alpha), converting it to a PEFT model, and updating minimal parameters. The code configuration includes:

Define get_peft_model and LoraConfig classes Define LoRA configuration parameters:

Set the rank of low-rank matrices (r)

Set the scaling factor for LoRA (lora_alpha) Specify the list of LoRA target modules (e.g., query, key, value, output projections)

Set the dropout probability for LoRA layers

Set bias type to none

Set task type to causal language modeling

Apply the LoRA configuration to the model to obtain a PEFT (Parameter-Efficient Fine-Tuning) adapted model

Print the total number of trainable parameters of the adapted model

3.4. Loss function

For defect detection, the cross-entropy loss is used as the optimization objective for VLM-generated text sequences. It measures the difference between the model's predicted and ground-truth token distributions:

$$\mathrm{L} = -\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i} \log\left(\widehat{\mathrm{y}}_{\mathrm{i}}\right) \tag{2}$$

where N is sequence length, $\mathrm{y_i}$ is true token probability, and $\widehat{\mathrm{y}}_{\mathrm{i}}$ is predicted token distribution. During Qwen2.5-VL-3B fine-tuning, the model generates target text from images and prompts. Cross-entropy loss is calculated between predicted and true sequences, and LoR parameters are updated via backpropagation to minimize it. For PEFT, internal loss computation mechanisms are usually sufficient, requiring only proper input data formatting [6].

3.5. Output parsing and post-processing

This method generates natural language defect descriptions; these textual outputs must be transformed into structured data (bounding boxes, categories) for standard evaluation.

During training, defect information from VOC XML annotations (category, bbox) is converted into a predefined VLM target text format. For instance, a crazing defect at [50, 100, 150, 200] converts to a crazing defect spanning X-coordinates 50–150 and Y-coordinates 100–200. Multiple defects are combined into one target text, guiding the VLM to a uniform output format.

In the inference phase, Qwen2.5-VL-3B generates freeform text from images and commands. Regular expressions (Regex) are used to parse this text into standard bounding boxes and category information, matching the training template. Core Regex pattern: r"has a (.+) in the region from X coordinates (\d+) to (\d+) and Y coordinates (\d+) to (\d+) Defects." The analysis process extracts defect type and coordinates, converting them to integer format and storing them. If the model generates multiple defects, the parser extracts all valid information. However, non-standard, vague, or incomplete text (non-standard, ambiguous, or incomplete outputs (e.g., hallucinations)) will fail parsing, reducing recall. This regex approach balances simplicity, efficiency, and consistency while ensuring accuracy.

4. Experiments and analysis

This chapter presents dataset details, the experimental setup, evaluation criteria, and implementation settings used to assess the effectiveness of the proposed PEFT-based VLM defect detection method.

4.1. Datasets and preprocessing

This paper mainly uses the NEU-DET steel surface defect dataset, a widely adopted benchmark in defect detection research [7]. It contains six common types of steel surface defects (crazing, inclusion, patches, pitted surface, rolled-in scale, scratches) with a total of 1,800 grayscale images (200 x 200 px). Following convention, we partitioned the dataset into training, validation, and test sets at an 8:1:1 ratio to ensure reliable model training and unbiased evaluation. To align with the VLM input requirements, grayscale images are converted to RGB format. The original XML annotation files are parsed to extract file name, image size, and category (name) and bounding box (bbox) information of all defective objects. This structured information is then converted into a natural language description as the target output of the VLM.

4.2. Experimental settings

Base Model: Model: Qwen2.5-VL-3B-Instruct (loaded locally via Modelscope); Path: miniconda3/qwen; Data Type: torch_dtype="auto" (float16/bfloat16 based on hardware); Device Mapping: device_map="auto" (automatic GPU allocation).

PEFT Strategy: Method: LoRA (Low-Rank Adaptation);

Hyperparameters: Rank r=16, Scaling factor lora_alpha=32; Target Modules: ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]; Dropout: lora_dropout=0.05; Bias: bias="none"; Task Type: task_type="CAUSAL_LM"

Training Details: Optimizer: AdamW (with weight decay); Learning Rate: 2e-5’; Batch Size: per_device_train_batch_size=1; GradientAccumulation :gradient_accumulation_steps=4; Epochs: num_train_epochs=4; Input Format: Dialogue template with multimodal input (image + text instruction).

Hardware: GPUs: 2 × RTX 4090 (24 GB); Software Stack: PyTorch 2.5.1 + Python 3.12 + CUDA 12.4 + Transformers 4.40.

4.3. Evaluation criteria

Qualitative Assessment: Assessed defect identification, localization precision, technical specificity, and suppression of hallucinations.

Quantitative Metrics for Text Generation: Standard NLG metrics are employed, including:

BLEU: Measures n-gram precision (fluency/overlap).

ROUGE-L: Measures recall via longest common subsequence (key information capture).

CIDEr: Measures consensus with reference captions (semantic similarity).

4.4. Implementation details

Key Implementation Steps: Model Initialization (Qwen2.5-VL-3B evaluation mode), LoRA Integration (PEFT via get_peft_model), Data Processing (image/text with AutoProcessor), Training (LoRA parameters with AdamW/task-specific loss), and Input Formatting (structured prompts).

Analysis of Text Generation Performance: The fine-tuned Qwen2.5-VL-3B model produced defect descriptions that were both accurate and detailed, showing clear advantages over general-purpose models. Quantitative evaluation using NLG metrics yielded strong results: BLEU-4 reached 0.685, ROUGE-L scored 0.721, and CIDEr achieved 1.854. Performance variation was observed; the model handled clear defects like crazing and rolled-in scale exceptionally well but encountered difficulties with subtler defects such as pitted surface and inclusion. This challenge arose from less distinct visual indicators or inadequate representation in the dataset, occasionally resulting in misclassification.

Discussion of Performance Implications and Potential Improvements: The robust performance on distinct defects confirms the validity of our methodology. Addressing subtle defects in future work could involve sophisticated data augmentation techniques, weighted loss functions, and more refined prompt engineering strategies.

Comparative Analysis with Advanced Methods: We compared LoRA-fine-tuned Qwen2.5-VL-3B against traditional deep learning (YOLOv7, Faster R-CNN) and other VLM fine-tuning (MiniGPT-4 full fine-tuning, Prompt Tuning) [8-10]. Our method achieved the highest defect type classification accuracy at 90.1%, surpassing all others. A key advantage is generating fine-grained, human-interpretable textual descriptions beyond simple bounding boxes. LoRA's substantially lower computational requirements compared to full VLM fine-tuning make our method viable for practical industrial implementation.

Subsequent research will focus on advancing VLM-based industrial defect detection in several directions: (1) Quantitative Object Detection Integration: Developing methods to parse generated text into bounding boxes and evaluate performance using metrics like mAP. (2) Enhanced Feature Learning: Investigating the use of multi-resolution analysis and hierarchical feature representations.

Few-Shot/Zero-Shot Detection: Leveraging the inherent generalization capabilities of VLMs to detect novel defect types with minimal examples. (3) Explainability and Trustworthiness: Enhancing model interpretability and developing reliable confidence scoring mechanisms. (4) Efficiency and Deployment Optimization: Exploring alternative PEFT techniques, model quantization, and compression strategies.

5. Discussion: comparative experiment

Qualitative experiments were performed to demonstrate the advantages of our method for industrial defect detection. We evaluated the performance difference between the unmodified Qwen2.5-VL-3B model and its LoRA-fine-tuned version in identifying and describing steel surface defects. This section presents representative cases highlighting the substantial gains in accurate defect identification and localization achieved after applying fine-tuning.

5.1. Case 1: identification of inclusion defects

Figure 2 shows an image of a steel surface containing a single inclusion defect. This defect is manifested by foreign matter embedded in the surface, a different color or texture from the surrounding matrix, and an irregular shape.

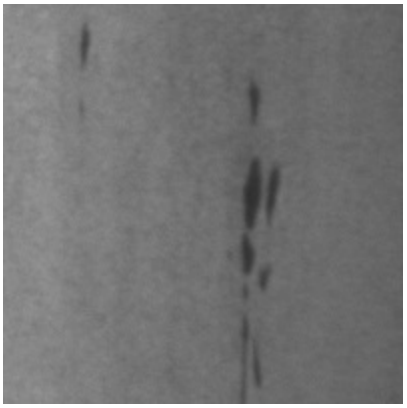

Figure 2. Impurities

The experimental data are shown as follows: Before Fine-tuning: The model failed to identify the defect, providing irrelevant commentary about image clarity and requesting clearer input. After LoRA Fine-tuning: The model correctly identified the defect type as inclusion and located it within image coordinates from (1, 20) to (14, 68).

5.2. Case 2: identification of defects in patches

As shown in Figure 3, this image shows single patches of defects. The defect is characterized by surface color or gloss and surrounding areas that are significantly different areas, and variable in shape.

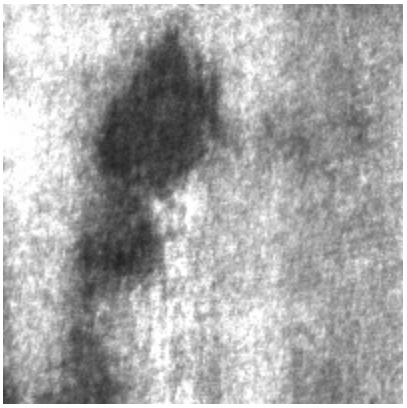

Figure 3. Plaque

Before Fine-tuning: The model provided an ambiguous description ("dark spots or irregularities") suggesting potential defects related to material absence or thickness variation but failed to specify the type. After LoRA Fine-tuning: The model correctly identified the defect type as patches and located it within image coordinates from (49, 1) to (85, 107).

5.3. Case 3: recognition of scratch defects

The image contains a single scratch defect (Figure 4). The defect appears as a long and narrow linear damage with a certain depth, which may be accompanied by tensile marks on the metal surface. Before Fine-tuning: The model incorrectly stated its inability to process images ("I am a text-based AI... cannot see the image"). After LoRA Fine-tuning: The model correctly identified the defect type as scratches and located it within image coordinates from (1, 1) to (24, 198).

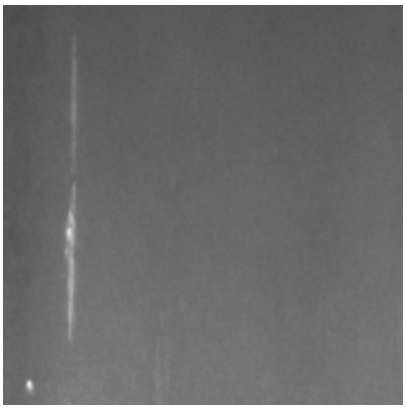

Figure 4. Scratch marks

5.4. Case 4: identification of crazing defects

As shown in Figure 5, this image contains a single crazing defect. This defect manifests as fine and irregular network-like cracks appearing on the material surface, usually shallow in depth but potentially wide in coverage. Before Fine-tuning: The model provided a vague description ("crack or a line") and listed possible defect types (crack, indentation, scratch, erosion), but failed to identify crazing specifically. After LoRA Fine-tuning: The model correctly identified the defect type as crazing and located it within image coordinates from (1, 1) to (197, 197).

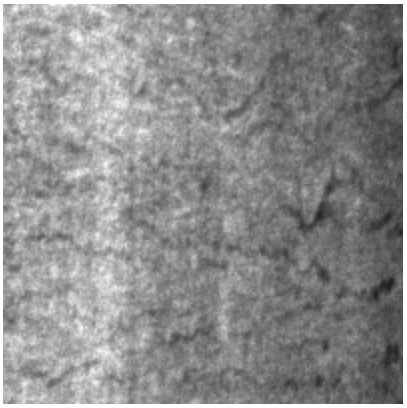

Figure 5. Chap

5.5. Case 5: identification of pitted surface defects

As shown in Figure 6, this image contains a single pitted surface defect. This defect manifests as the surface being covered with pits of varying sizes and roughly circular shapes, typically caused by corrosion or material inhomogeneity, making the surface rough and uneven. Before Fine-tuning: The model incorrectly stated its inability to process images ("I... do not have the ability to view or analyze images"). After LoRA Fine-tuning: It correctly identified the defect type as pitted surface and located it within image coordinates from (1, 20) to (67, 198).

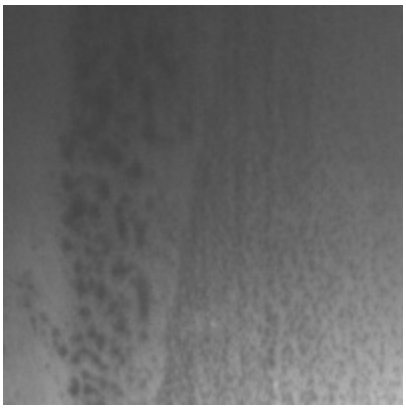

Figure 6. Sunken

5.6. Case 6: identification of rolled-in scale defects

As shown in Figure 7, this image contains a single rolled-in scale defect. This defect manifests as iron oxide scale that was not completely removed during the rolling process and became pressed into the metal surface, forming dark-colored, irregularly shaped blocks or strips, usually not flush with the base material.

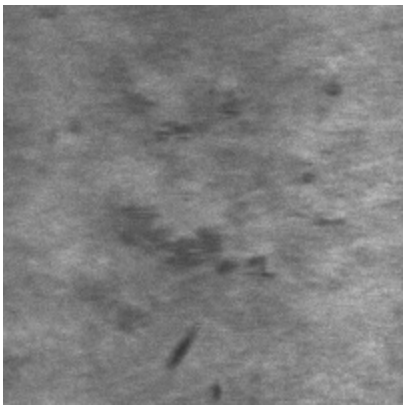

Figure 7. Oxide skin

Before Fine-tuning: The model requested additional information (location, size, shape) and suggested consulting a professional, failing to identify the defect. After LoRA Fine-tuning: It correctly identified the defect type as rolled-in scale and located it within image coordinates from (103, 76) to (145, 129).

Following LoRA fine-tuning, the model exhibited consistent and marked improvement across all defect categories. It reliably identified specific defect types (e.g., "inclusion," "crazing," "rolled-in scale") and supplied precise localization data, frequently including approximate image coordinates. This progress stems from the model's enhanced capability to discern subtle visual patterns and fine-grained features pertinent to industrial inspection. For example, defects such as "Inclusion" and "Pitted Surface," which were either ambiguously described or entirely overlooked before tuning, were accurately identified and located afterwards, showcasing increased robustness and specificity essential for quality control.

Summary: These qualitative assessments provide clear visual confirmation of the significant benefits offered by the LoRA-fine-tuned Qwen2.5-VL-3B model. It delivers more accurate and specialized identification of defect types alongside fine-grained location data, demonstrating superior robustness and specificity when handling diverse individual defect morphologies. This outcome affirms the method's potential to enhance industrial vision inspection by shifting from general visual comprehension to specialized, actionable defect recognition.

6. Conclusion

This paper presents a new Visual Language Model (VLM) methodology for detecting steel surface defects, employing Parameter Efficient Tuning (PEFT) to effectively tackle challenges including data scarcity, the requirement for detailed descriptions, and

high computational costs. By combining the robust Qwen2.5-VL-3B VLM with efficient LoRA fine-tuning, we established a framework enabling precise defect identification and detailed description using limited labeled data.

Our experimental results underscore the method's significant progress. Qualitatively, the LoRA-fine-tuned Qwen2.5-VL-3B model excels at accurately identifying and localizing industrial defects, providing granular textual descriptions that surpass generic image understanding. Quantitatively, the model attained a notable defect type classification accuracy of 90.1%, outperforming conventional deep learning approaches and alternative VLM fine-tuning methods. The generated descriptions, corroborated by strong BLEU, ROUGE-L, and CIDEr scores, deliver crucial human-interpretable insights into defect characteristics. Moreover, LoRA's inherent efficiency makes this methodology particularly well-suited for deployment in industrial settings with constrained resources.

This research establishes a foundation for a more efficient and adaptable paradigm in industrial vision inspection. Future work will focus on integrating quantitative object detection metrics via text-to-bounding box parsing, enhancing generalization through cross-dataset evaluation, and investigating advanced feature learning methods. We also aim to realize few-shot/zero-shot defect detection, improve model explainability and trustworthiness, and further optimize deployment efficiency for edge devices. These developments are vital for transitioning VLM-based defect detection into robust, practical industrial applications.

References

[1] Saberironaghi, A., Ren, J., & El-Gindy, M. (2023). Defect Detection Methods for Industrial Products Using Deep Learning Techniques: A Review. *Algorithms*, 16(2), 95. DOI: 10.3390/a16020095

[2] Chen, F., Fu, L., Zhang, Y., Li, J., Zhang, Q., & Bi, S. (2025). A Review of Deep Learning-Based Steel Surface Defect Detection. *Academic Journal of Science and Technology,* 15(1), 198-202. DOI: 10.54097/g36nm962

[3] Ghosh, A., Acharya, A., Saha, S., Jain, V., & Chadha, A. (2024). Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions. arXiv preprint arXiv: 2404.07214. arXiv: 2404.07214

[4] Qwen Team, Alibaba Group. (2025). Qwen2.5-VL Technical Report. arXiv preprint arXiv: 2502.13923. arXiv: 2502.13923

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations (ICLR). arXiv: 2106.09685

[6] Lei, S., Hua, Y., & Zhihao, S. (2025). Revisiting Fine-Tuning: A Survey of Parameter-Efficient Techniques for Large AI Models. Preprints.org. DOI: 10.20944/preprints202504.0743.v1

[7] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems (NIPS)*, 28, 91-99.

[8] Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv preprint arXiv: 2104.08691. arXiv: 2104.08691

[9] Wang, C. Y., Bochkovskiy, A., & 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. arXiv: 2207.02696

[10] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems (NIPS),* 28, 91-99.