FFAD: Fixed-Position Few-Shot Anomaly Detection for Wire Harness Utilizing Vision-Language Models

Powei Liao[1], Pei-Chun Chien[2], Hiroki Tsukida[2], Yoichi Kato[3] and Jun Ohya[1]
[1]*Department of Modern Mechanical and Engineering, Waseda University, Tokyo, Japan*
[2]*AI Digital Division, Yazaki Corporation, Tokyo, Japan*
[3]*Global Center for Science and Engineering, Waseda University, Tokyo, Japan*

Keywords: Few-Shot Learning, Anomaly Detection, Vision-Language Model.

Abstract: Anomaly detection in wire harness assembly for automobiles is a challenging task due to the deformable nature of cables and the diverse assembly environments. Traditional deep learning methods require large datasets, which are difficult to obtain in manufacturing settings. To address these challenges, we propose Fixed-Position Few-Shot Anomaly Detection (FFAD), a method that leverages pre-trained vision-language models, specifically CLIP, to perform anomaly detection with minimal data. By capturing images from fixed positions and using position-based learnable prompts and visual augmentation, FFAD can detect anomalies in complex wire harness situations without the need for extensive data collection. Our experiments demonstrated that FFAD achieves over 90% accuracy with fewer than 16 shots per class, outperforming existing few-shot learning methods.

1 INTRODUCTION

With the rapid advancement of deep learning, manufacturing around the world is becoming increasingly intelligent through integrating complex sensors and the use of Internet of Things (IoT) (Wang et al., 2018). The manufacturing sector has also seen a steady increase in the adoption of intelligent industrial robots, where their performance and problem-solving ability have been substantially improved year-by-year (Benotsmane et al., 2020). One can confidently state that automation technologies, especially robotics and digital solutions, are reshaping the production landscape as we know it, making these processes more flexible and efficient than ever before (Karabegović et al., 2018; Papulová et al., 2022). However, despite the significant advancement in automation technologies, the "nerves and blood vessels" of automobiles – wire harness – is still mostly handled and manufactured by humans due to inherent difficulties (Heisler et al., 2021).

Broadly speaking, a wire harness functions as a bundle of cables that transmit power and electrical signals between various parts of an automobile. As automobiles become more sensor-reliant and with electric vehicles (EVs) receive more attention, the demand for wire harnesses of greater complexity continues to rise. However, given the declining availability of labour in many regions of the world, the automation of wire harness manufacturing is rapidly emerging as a prominent area of research (Navas-Reascos et al., 2022). One particular interest is in the ability to perform anomaly detection and quality assurance with deep-learning models. Such a system would require a vision system capable of detecting various anomalies during the manufacturing process. With effective anomaly detection, the system can intervene to halt an operation when necessary, thereby preventing damage such as wire entanglement or foreign object intrusion. This also opens the possibility for auto rectification to be performed.

Currently, anomaly detection for wire harnesses remains challenging for several reasons. Firstly, cables – also referred to as deformable linear objects (DLOs) – are non-rigid and highly deformable, meaning their shape can change drastically during the manufacturing process, complicating detection efforts (Zhou et al., 2020). Secondly, the definition of an anomaly can often be ambiguous as only parts of wire harness are anomalous while the overall appearance remains satisfactory. Thirdly, the layout of a wire har-

*The work described by this paper is a collaborative research between Yazaki Corp. and Waseda University.

Liao, P., Chien, P.-C., Tsukida, H., Kato, Y. and Ohya, J.
FFAD: Fixed-Position Few-Shot Anomaly Detection for Wire Harness Utilizing Vision-Language Models.
DOI: 10.5220/0013164400003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 647-656
ISBN: 978-989-758-730-6; ISSN: 2184-4313

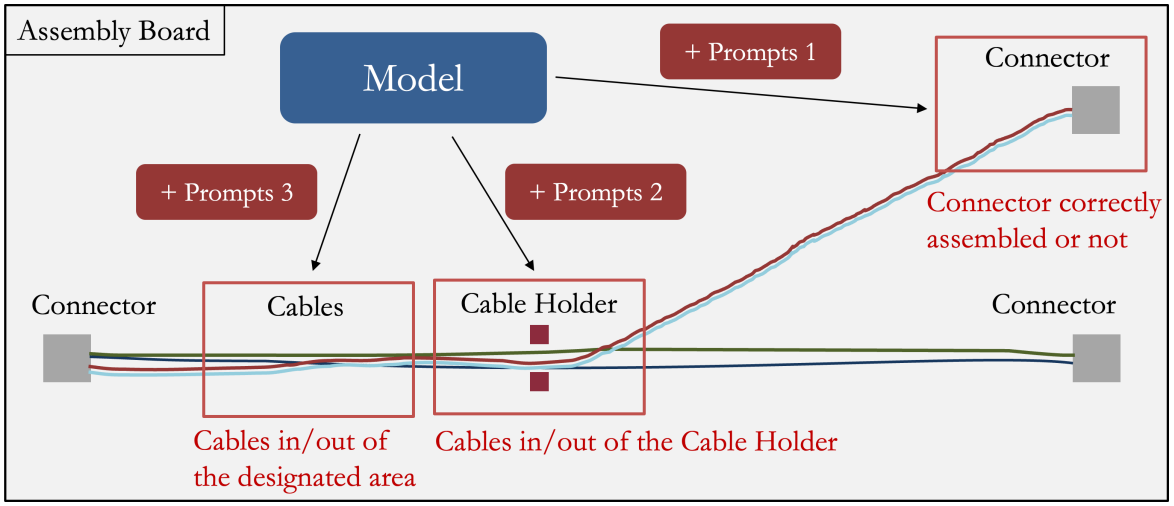

Figure 1: **Overview of our application scenario.** The model will not be trained throughout the entire process; instead, a pair of prompts will be trained at each position to detect anomalies.

ness can vary significantly for each scenario, leading to diverse and unpredictable assembly environments. Finally, as with other industrial anomaly detection tasks, most of the deep learning methods heavily depend on available datasets. The high demand for both the quantity and quality of datasets makes it challenging to reproduce the same results on varying production lines.

To address these challenges, we propose a method – Fixed-Position Few-Shot Anomaly Detection (FFAD) – which aims to detect user-defined anomalies in complex wire harness situations without requiring a large amount of data for learning and can be easily adapted to different environments by leveraging the power of vision-language models.

Few-shot learning, which enables a model to perform downstream tasks with only a small amount of data (typically fewer than 16 shots), has been gaining increasing attention (Wang et al., 2020). CLIP (Contrastive Language-Image Pre-Training) (Radford et al., 2021), a vision-language model pre-trained on 400 million image-text pairs, aligns images and texts in a shared embedding space, enabling it to associate visual and textual concepts. This architecture allows CLIP to demonstrate remarkable zero-shot learning capabilities. The extensive pre-trained knowledge of CLIP can be easily adapted to downstream tasks; for instance, CoOp (Zhou et al., 2022b) adapts CLIP for downstream classification tasks by making the text input prompt learnable, then performs training on a few samples from the target tasks. Building on a similar concept as CoOp, we propose our method FFAD that leverages CLIP's pre-trained knowledge to perform few-shot learning for anomaly detection in wire harnesses.

Figure 1 shows an overview of our application scenario. In traditional deep learning methods, the entire pre-trained model is fine-tuned with a large amount of additional data to detect all possible anomalies, which requires substantial labor and time. Furthermore, if an additional anomaly class is to be added, the whole model needs to be fine-tuned again. To address this, the FFAD approach trains a specific prompt for each position on the assembly board, allowing for detection of different anomalies without the need of full model fine-tuning. Additionally, we introduce Position-based Visual Augmentation to further enhance the model's anomaly detection capabilities. Our experiments show that FFAD outperformed other few-shot learning methods, achieving over 90% accuracy with **16 shots** per class, less than **2 minutes** of training time, and requiring less than **200 KB** to store each prompt.

The main contributions of this paper are as follows:

- Proposed a novel wire harness anomaly detection method – FFAD, and introduced Position-Based Visual Augmentation to enhance model accuracy.
- Development of a wire harness anomaly dataset for detection tasks and demonstrated FFAD's superiority over existing methods through extensive experiments.
- Ablation studies were carried out to investigate the impact of visual prompt engineering and model scaling on FFAD's performance.

2 RELATED WORK

2.1 Anomaly Detection for Wire Harness

Wire harness is an essential component of automobiles, connecting various parts throughout the vehicle. It transmits power and information between multiple components, ranging from assistive systems to safety-critical systems. Due to its importance, up to 90% of the wire harness manufacturing process still requires human involvement (Nguyen et al., 2021). Therefore, automating wire harness production while maintaining quality and reliability has become a significant focus in recent years (Navas-Reascos et al., 2022; Trommnau et al., 2019; Hermansson et al., 2013).

For such process to be automated, a vision system capable of understanding the assembly scene and detecting anomalies within said scene is crucial (Wang et al., 2024). Li et al. (2024) developed a mobile-based visual imaging system using YOLOv5 for real-time detection and recognition, which is used to error-proof the installation process of automotive wiring harness relays. Nguyen et al. (2022) proposed a deep learning-based data processing pipeline that utilizes both real and synthetic point clouds for automated inspection. Chien et al. (2024) introduced a method for classifying different cable tendencies, enabling further anomaly detection through semantic segmentation by utilizing real and simulated RGB data.

While these methods can detect anomalies in the wire harness process using deep learning techniques, they often rely on large datasets and may require synthetic data to achieve optimal performance. Our proposed method requires only a small amount of data by leveraging the pre-trained knowledge of vision-language models.

2.2 Few-Shot Learning Utilizing Vision-Language Models

Following the remarkable success of CLIP (Radford et al., 2021), there has been a rising interest in research on vision-language models. Numerous downstream tasks based on CLIP have emerged in recent years, including multi-label classification (Sun et al., 2022; Liao and Nakano, 2025), object detection (Gu et al., 2021; Wang et al., 2022), image segmentation (Liang et al., 2023; Lüddecke and Ecker, 2022; Xie et al., 2022), image editing (Kwon and Ye, 2022; Wei et al., 2022), and image captioning (Hessel et al., 2021; Mokady et al., 2021).

Among the various studies, significant research has been devoted to utilizing CLIP for few-shot learning. CoOp (Zhou et al., 2022b) treats the prompt in the input text as a set of learnable vectors, enabling few-shot learning by directly training the prompt. Co-CoOp (Zhou et al., 2022a), on the other hand, takes a different approach by training a lightweight network that uses the input image to adjust the prompt adaptively, rather than training it directly. CLIP-Adapter (Gao et al., 2024) enhances the pre-trained CLIP image and text encoders by adding three fully connected layers, performing few-shot learning by training only the additional layers. Shtedritski et al. (2023) demonstrates that visual prompt engineering, such as simple image edits like drawing a red circle, can improve CLIP's performance on complex tasks. Building on the concepts of CoOp and visual prompt engineering, our proposed method adapts the pre-trained CLIP model for wire harness assembly anomaly detection using few-shot learning. Additionally, we incorporate Position-Based Visual Augmentation to further enhance the model's performance.

Figure 2: RealSense D405 camera mounted on UR5e robot.

Table 1: The number of images from different position.

Position	Normal	Anomaly
Connector	80	51
Cable Holder	80	51
Cable 1	65	66
Cable 2	63	65

3 DATASET

We constructed a wire harness assembly dataset to demonstrate the effectiveness of our method. As shown in Figure 2, we mounted a camera (Intel RealSense D405) onto a robot arm (Universal Robots UR5e) to capture fixed-position images. For each capture, the wire harness is automatically assembled using the robot arm by traversing through registered

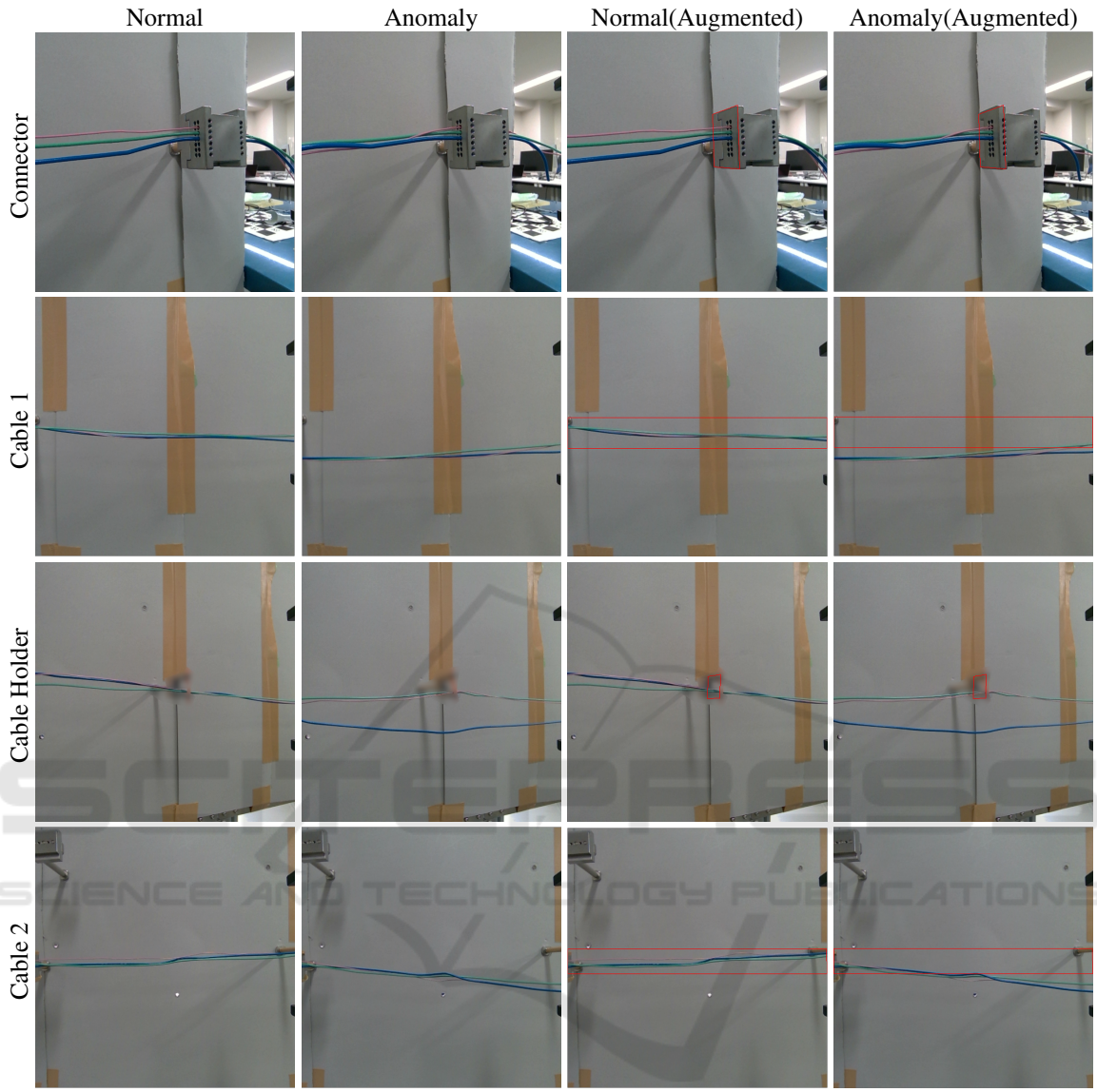

Figure 3: Examples of each position from the dataset before and after visual augmentation. (The cable holder is blurred for confidentiality reasons).

coordinates. Specifically, the robot arm first grabs the connector from the starting base, attaches the cables to the cable holder – a partially enclosed C-shaped cable supporter, and places the connector onto the goal base. Then, images of the assembled wire harness are captured from different fixed positions. As illustrated in Figure 1, different positions have different anomalies that need to be detected. For wire harnesses, some common anomalies to be identified include, but are not limited to:

- if the connector is appropriately fixed to its base
- if the cable is routed through the cable holder
- if the cable is situated within a certain region

In this dataset, images are taken from four different positions: the connector, the cable holder, and two positions along the cables. Table 1 shows the number of normal and anomalous images for each position. Figure 3 provides examples of both normal and anomalous situations for each position.

4 METHODOLOGY

The structure of the proposed Fixed-position Few-shot Anomaly Detection (FFAD) is depicted in Figure 4. In our setup, the camera is mounted on the robot arm and moves in conjunction with the robot's

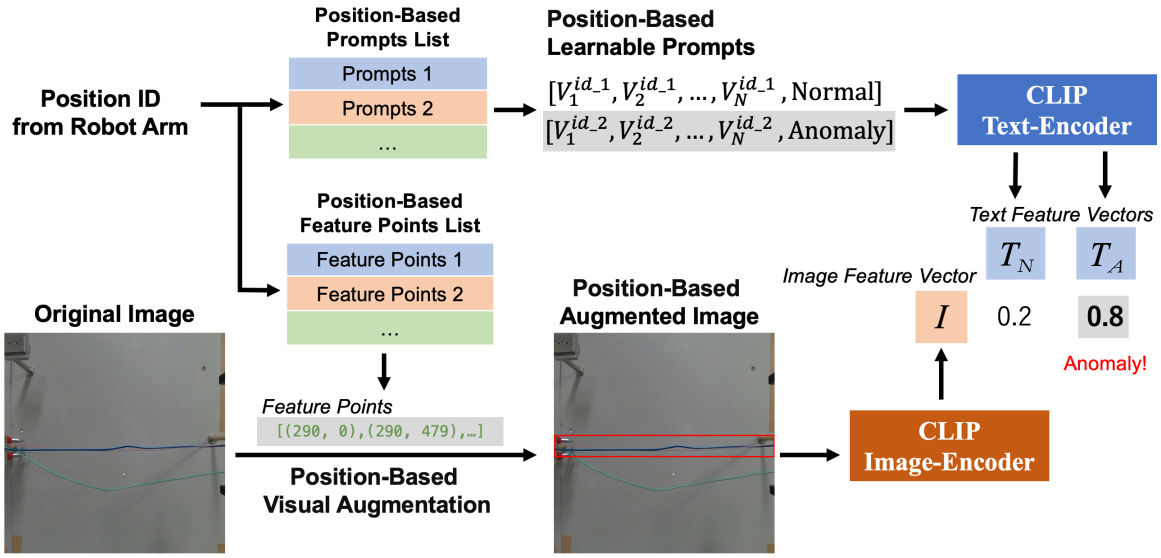

Figure 4: Illustration of our proposed approach FFAD.

movements. This arrangement ensures that the "fixed-position" requirement of FFAD can be effectively met by registering robot coordinates. By capturing images at the same coordinates, the camera's position and angle are guaranteed to be identical. FFAD takes two inputs: the original image and a *Position ID*, which identifies the location where the image was captured.

Based on the *Position ID*, we create two lists, Position-Based Prompts List and Position-Based Feature Points List. In Position-Based Feature Points List, four feature points are stored for each position. These feature points are manually crafted for use in the Position-Based Visual Augmentation.

The visual augmentation process uses OpenCV to draw a quadrilateral on the original image based on these feature points, allowing the model to better focus on the relevant area of the quadrilaterals. As shown in Figure 3, each position has a specific visual augmentation policy. As for the connector, the goal is to detect whether it is properly set on the base. Thus, the feature points of quadrilateral are instead defined along the edges of the connector and the base. In the anomaly example of connector shown in Figure 3, distinguishing between normal and anomalous situations is difficult when the visual augmentation has not been applied. However, after applying visual augmentation, it is noticeable that the connector is slightly misaligned to the left of the base.

Similarly for the cable position, we define a quadrilateral around the cables as the appropriate region of which cables are to exist within. Any loose cable that falls out of this region is to be classified as lack of tension. Lastly, for the cable holder position, a quadrilateral is drawn around the holder to highlight the key area for detecting whether the cables are properly routed through the holder.

In Position-Based Prompts List, each position is associated with a pair of learnable prompts: prompt for normal, Prompt^N, and prompt for anomaly, Prompt^A. There are N learnable vectors $V_1, \cdots, V_N$ in each prompt with the word "Normal" and "Anomaly", these prompts can be written as:

$$\text{Prompt}_{id}^N = [V_1^{id\_N}, V_2^{id\_N}, \ldots, V_N^{id\_N}, \text{Normal}] \quad (1)$$

$$\text{Prompt}_{id}^A = [V_1^{id\_A}, V_2^{id\_A}, \ldots, V_N^{id\_A}, \text{Anomaly}] \quad (2)$$

These vectors are input into the pre-trained CLIP text encoder E_t, producing two Text-Feature Vectors T_N and T_A.

$$T_N = E_t(\text{Prompt}_{id}^N) \quad (3)$$

$$T_A = E_t(\text{Prompt}_{id}^A) \quad (4)$$

On the other side, the Position-Based Augmented image is input into the pre-trained CLIP image encoder E_i, yielding the Image Feature Vector I:

$$I = E_i(\text{Augmented Image}) \quad (5)$$

Subsequently, the Image Feature Vector I is compared to both Text-Feature Vectors, T_N and T_A, using cosine similarity $S_{\cos}$ to produce normal and anomaly logits:

$$S_N = S_{\cos}(I, T_N),\ S_A = S_{\cos}(I, T_A) \quad (6)$$

Finally, with the pair of logits, the binary classification output p is given by:

$$p = \frac{\exp(S_A/\tau)}{\exp(S_N/\tau) + \exp(S_A/\tau)} \quad (7)$$

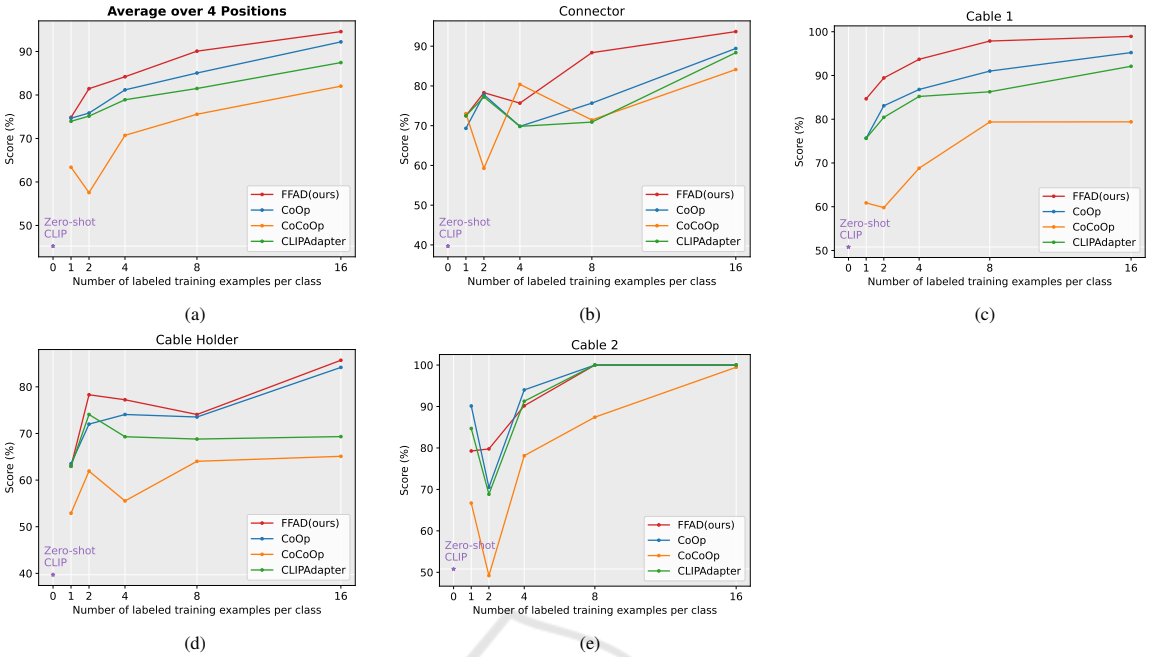

Figure 5: Results of few-shot learning of the defined four positions.

5 EXPERIMENTS

5.1 Few-Shot Anomaly Detection for Wire Harness

5.1.1 Training Setting

To evaluate the effectiveness of our method for anomaly detection in wire harnesses, we trained and tested it on the dataset described in Section 3. We trained FFAD under the few-shot setups of 1, 2, 4, 8, 16 shots. These training examples were randomly selected from the dataset and served as training and validation sets, and the remaining data was used for testing. In the Position-Based Learnable Prompts, we use 8 learnable vectors for each prompt, suffixed with the class name "Normal" or "Anomaly" for each position. We employed ViT-B/16 (Dosovitskiy, 2020) as our choice of the pre-trained CLIP image encoder, as well as the use of pre-trained models provided by OpenAI.[1] The entire pre-trained model is kept untrained, while only the learnable prompts are updated during training. We did not use data augmentation technique but used only normalization to preserve fixed-position information as much as possible. We used cross-entropy as the loss function and SGD as the optimizer. The model was trained for 200 epochs with a learning rate of 0.002. The overall implementation is based on CoOp's open-source code.[2]

5.1.2 Baselines

For our task, every single position can be regarded as a binary classification task. We compare our method against one zero-shot method, CLIP, and three few-shot methods: CoOp, CoCoOp, and CLIP-Adapter. Both CoOp and CoCoOp utilize 8 learnable vectors, which is consistent with our approach. All few-shot methods follow the same training setup as FFAD and utilize the ViT-B/16 pre-trained model from OpenAI.

5.1.3 Experiments Result & Analysis

The plot in Figure 5 shows the accuracy for four different positions as well as the average accuracy across these positions. For few-shot learning methods, the result is the average over three separate runs. As depicted in Figure 5(a), FFAD outperforms the other three methods when averaged over 4 positions in every few-shot setup.

Figure 5(b) shows that FFAD outperforms the other methods at the "Connector" position when the few-shot setup is of 8 or 16 shots. The visual difference between normal and anomaly is relatively subtle when no adequate visual augmentation is applied, as seen in Figure 3, which may be challenging for the

[1] https://github.com/openai/CLIP

[2] https://github.com/KaiyangZhou/CoOp

Table 2: Result of different visual prompt engineering.

Color	Thickness	Connector	Cable 1	Cable Holder	Cable 2	Average
Red	1	81.70	92.92	75.65	89.84	85.03
Red	2	82.85	90.69	76.30	90.27	85.00
Green	1	81.39	94.71	78.93	87.42	85.61
Blue	1	79.92	91.54	77.13	90.16	84.69

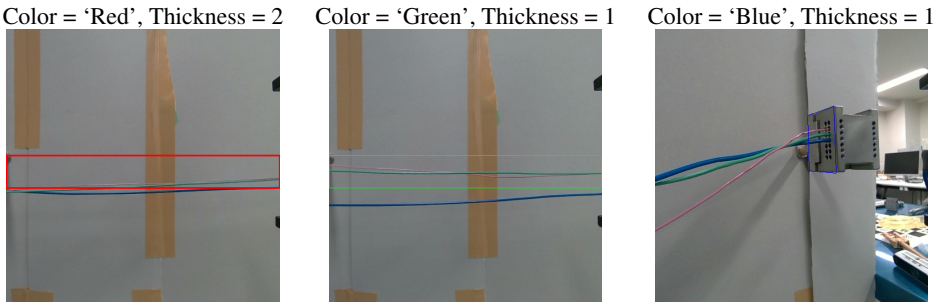

Figure 6: Examples of visual prompt engineering with different settings.

model to distinguish differences. CoCoOp performs well with fewer shots but is unstable and does not perform as effectively when the few-shot setups increase to 8 or 16 shots.

Figure 5(c) plot demonstrates that FFAD significantly outperforms the other methods at the "Cable 1" position, highlighting the effectiveness of visual augmentation.

In Figure 5(d), FFAD shows a slight advantage over the other methods at the "Cable Holder" position, indicating that visual augmentation aids the model focusing on the relationship between the cable and the cable holder, even though it is not as intuitive as in other positions.

However, Figure 5(e) presents unexpected results where FFAD does not surpass the other methods at the "Cable 2" position. This is likely due to the relative simplicity of this task, as methods without visual augmentation were already achieving 100% accuracy with 8 and 16 shots. As shown in Figure 3, the background of "Cable 2" is relatively clean, the application of visual augmentation may have introduced noise, leading to a negative effect in setups with fewer shots.

Zero-shot CLIP did not perform well across all four positions, likely due to the specific examples in our task not being available in the web-scraped data that CLIP was originally trained on. Additionally, using the terms 'Normal' and 'Anomaly' to describe the two classes may have posed a challenge for the model to comprehend when no additional training is done. Meanwhile, all few-shot methods demonstrated that with proper instruction, the pre-trained knowledge can be effectively adapted to the downstream task. This suggests that a small amount of task-specific guidance can significantly enhance model performance, leveraging the underlying capabilities of the pre-trained model.

Overall, the results from Figure 5 show that integrating visual augmentation allows FFAD to improve anomaly detection performance across different positions and few-shot setups.

5.2 Ablation Study

To further assess the impact of the FFAD components, an ablation study is carried out to examine the roles of "Visual Prompt Engineering" and "Model Scaling", aiming to gain deeper insights into their influence on the performance of our few-shot anomaly detection method.

5.2.1 Visual Prompt Engineering

Inspired by visual prompt engineering (Shtedritski et al., 2023), we propose Position-Based Visual Augmentation in FFAD, where a red quadrilateral is drawn on the original image to emphasize region of interest thus enhance the performance. To investigate how different colors or thicknesses of the quadrilateral in visual augmentation affect detection results, we conducted experiments with various settings. Table 2 presents the results of our model under different visual augmentation configurations, averaged over few-shot settings of 1, 2, 4, 8, and 16 shots, with 3 runs for each setting. Figure 6 shows examples of visual prompt engineering with different settings.

By adjusting the thickness to 2 pixels while keep-

Table 3: Average result of few-shot learning on different model scaling.

Models	Average Accuracy
ViT-B/32@224px	85.32
ViT-B/16@224px	85.03
ViT-L/14@224px	87.70
ViT-L/14@336px	86.05

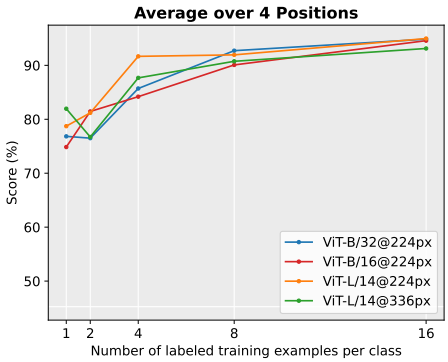

Figure 7: Result of few-shot learning on different model scaling.

ing the color as "red", there is no significant change in performance. However, a slight decrease in accuracy at position "Cable 2" was observed. This may be due to the thicker red quadrilateral resembling the yellow tape in the background, and the thicker line may occasionally occlude the cable, as is shown in Figure 6(a), causing confusion for the model.

Changing the color to "green" while keeping the thickness at 1 pixel, we observed an increase in performance at the "Cable 1" and "Cable Holder" positions but again a decrease at the "Cable 2" position. This may be caused by the green quadrilateral having more contrast with the backgrounds of "Cable 1" and "Cable Holder" compared to the red quadrilateral, as is shown in Figure 6(b), while there is less contrast at "Cable 2" position.

A similar effect was observed with the color blue, where it blended into the metallic texture of the connector, as is shown in Figure 6(c), resulting in a performance drop at the "Connector" position. However, due to its higher contrast with the cable holder, the performance improved at the "Cable Holder" position.

In summary, the color of the quadrilateral and its contrast with the background may have some impact on performance. Choosing suitable colors for different positions that improve contrast could help enhance model performance.

5.2.2 Model Scaling

Another important factor to performance is the model scaling of the pre-trained CLIP model. To understand how different model sizes impact our task, four different sizes of Vision Transformer (ViT) obtained from OpenAI's official CLIP implementation were evaluated. Table 3 presents the results for these models. Each result is averaged across four different positions, where each position is averaged over five different few-shot setups: 1, 2, 4, 8, and 16 shots. Each few-shot setup result is again averaged over three runs. Table 3 indicates that increasing the size of the vision transformer model from B (Base) to L (Large) results in a slight improvement in accuracy. However, changing the patch size of the ViT-B model from 32 to 16 did not significantly affect performance.

Interestingly, for the ViT-L/14 model, increasing the input image resolution from 224×224 to 336×336 unexpectedly decreases the performance. This could be due to that, in our task, the region of interest for anomaly detection has been zoomed in. As a result, the anomalies already occupy a significant portion of the input image, and a resolution of 224×224 is sufficient to capture the necessary details. Increasing the input resolution further might introduce unnecessary complexity, making the model harder to train and demanding more computational resources without yielding much improvement.

Figure 7 shows the accuracy plot under different few-shot configurations. The ViT model with a Large size outperforms the Base size when trained with fewer shots, such as 1 and 4. However, the performance difference diminishes as the number of shots increases to 8 or 16, where both models achieve nearly the same accuracy.

6 CONCLUSIONS

In this paper, we proposed FFAD, a Fixed-Position Few-Shot Anomaly Detection method for wire harness assembly, which leverages pre-trained vision-language models, specifically CLIP, to perform anomaly detection with minimal data. By capturing images from fixed positions and employing position-based learnable prompts along with visual augmentation, FFAD can effectively detect anomalies in complex wire harness situations without the need for extensive data collection. Our experimental results demonstrated that FFAD outperforms other few-shot learning methods, and achieves over 90% accuracy with fewer than 16 shots per class. The ablation studies indicated that visual prompt engineering, such as

adjusting the color and thickness of the augmentation, can impact the model's performance. Additionally, model scaling shows that larger pre-trained models can improve accuracy, especially in low-shot scenarios. We hope that the ideas and empirical findings presented in this paper provide valuable insights for unlocking potential industrial applications that only have access to limited data. In the future, we plan to enable the model to learn the visual augmentation policy from the data itself, rather than relying on manual decisions, in order to simplify the overall process.

REFERENCES

Benotsmane, R., Dudás, L., and Kovács, G. (2020). Survey on new trends of robotic tools in the automotive industry. In *Vehicle and automotive engineering*, pages 443–457. Springer.

Chien, P.-C., Liao, P., Fukuzawa, E., and Ohya, J. (2024). Classifying cable tendency with semantic segmentation by utilizing real and simulated rgb data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8430–8438.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. (2024). Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.

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*.

Heisler, P., Utsch, D., Kuhn, M., and Franke, J. (2021). Optimization of wire harness assembly using human–robot-collaboration. *Procedia CIRP*, 97:260–265.

Hermansson, T., Bohlin, R., Carlson, J. S., and Söderberg, R. (2013). Automatic assembly path planning for wiring harness installations. *Journal of manufacturing systems*, 32(3):417–422.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Karabegović, I., Karabegović, E., Mahmić, M., and Husak, E. (2018). Innovative automation of production processes in the automotive industry. *International Journal of Engineering*.

Kwon, G. and Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071.

Li, S., Yuan, M., Wang, W., Cao, F., Shi, H., Zhang, Y., and Meng, X. (2024). Enhanced yolo-and wearable-based inspection system for automotive wire harness assembly. *Applied Sciences*, 14(7):2942.

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., and Marculescu, D. (2023). Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070.

Liao, P. and Nakano, G. (2025). Bridgeclip: Automatic bridge inspection by utilizing vision-language model. In *International Conference on Pattern Recognition*, pages 61–76. Springer.

Lüddecke, T. and Ecker, A. (2022). Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096.

Mokady, R., Hertz, A., and Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Navas-Reascos, G. E., Romero, D., Stahre, J., and Caballero-Ruiz, A. (2022). Wire harness assembly process supported by collaborative robots: Literature review and call for r&d. *Robotics*, 11(3):65.

Nguyen, H. G., Habiboglu, R., and Franke, J. (2022). Enabling deep learning using synthetic data: A case study for the automotive wiring harness manufacturing. *Procedia CIRP*, 107:1263–1268.

Nguyen, H. G., Kuhn, M., and Franke, J. (2021). Manufacturing automation for automotive wiring harnesses. *Procedia Cirp*, 97:379–384.

Papulová, Z., Gažová, A., and Šufliarskỳ, L. (2022). Implementation of automation technologies of industry 4.0 in automotive manufacturing companies. *Procedia Computer Science*, 200:1488–1497.

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*, pages 8748–8763. PMLR.

Shtedritski, A., Rupprecht, C., and Vedaldi, A. (2023). What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997.

Sun, X., Hu, P., and Saenko, K. (2022). Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582.

Trommnau, J., Kühnle, J., Siegert, J., Inderka, R., and Bauernhansl, T. (2019). Overview of the state of the art in the production process of automotive wire harnesses, current research and future trends. *Procedia CIRP*, 81:387–392.

Wang, H., Salunkhe, O., Quadrini, W., Lämkull, D., Ore, F., Despeisse, M., Fumagalli, L., Stahre, J., and Johansson, B. (2024). A systematic literature review of computer vision applications in robotized wire harness assembly. *Advanced Engineering Informatics*, 62:102596.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., and Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of manufacturing systems*, 48:144–156.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Wang, Z., Codella, N., Chen, Y.-C., Zhou, L., Yang, J., Dai, X., Xiao, B., You, H., Chang, S.-F., and Yuan, L. (2022). Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*.

Wei, T., Chen, D., Zhou, W., Liao, J., Tan, Z., Yuan, L., Zhang, W., and Yu, N. (2022). Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081.

Xie, J., Hou, X., Ye, K., and Shen, L. (2022). Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4492.

Zhou, H., Li, S., Lu, Q., and Qian, J. (2020). A practical solution to deformable linear object manipulation: A case study on cable harness connection. In *2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 329–333. IEEE.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.