# DHR-CLIP: Dynamic High-Resolution Object-agnostic Prompt Learning for Zero-shot Anomaly Segmentation

Jiyul Ham
*Department of Industrial and Management Engineering*
*Korea University*
Seoul, South Korea
jiyulham@korea.ac.kr

Jun-Geol Baek*
*Department of Industrial and Management Engineering*
*Korea University*
Seoul, South Korea
jungeol@korea.ac.kr

***Abstract*—Zero-shot anomaly segmentation (ZSAS) is crucial for detecting and localizing defects in target datasets without need for training samples. This approach is particularly valuable in industrial quality control, where there are distributional shifts between training and operational environments or when data access is restricted. Recent vision-language models have demonstrated strong zero-shot performance across various visual tasks. However, the variations in the granularity of local anomaly regions due to resolution changes and their focus on class semantics make it challenging to directly apply them to ZSAS. To address these issues, we propose DHR-CLIP, a novel approach that incorporates dynamic high-resolution processing to enhance ZSAS in industrial inspection tasks. Additionally, we adapt object-agnostic prompt design to detect normal and anomalous patterns without relying on specific object semantics. Finally, we implement deep-text prompt tuning in the text encoder for refined textual representations and employ V-V attention layers in the vision encoder to capture detailed local features. Our integrated framework enables effective identification of fine-grained anomalies through refinement of image and text prompt design, providing precise localization of defects. The effectiveness of DHR-CLIP has been demonstrated through comprehensive experiments on real-world industrial datasets, MVTecAD and VisA, achieving strong performance and generalization capabilities across diverse industrial scenarios.**

***Keywords*—Prompt Learning, Dynamic High-Resolution, Object-agnostic prompt, Zero-shot Anomaly Segmentation, Multimodal Learning**

## I. Introduction

The advancement of smart factories has significantly increased the importance of automated visual inspection systems [1]. Anomaly detection (AD), which identifies products that deviate from normal patterns, is fundamental technology for industrial quality control [2]. Anomaly segmentation (AS), specifically, detects and localizes defects in particular regions, playing a vital role in analyzing and addressing quality issues [3]. However, conventional AS approaches have several limitations. First, they rely on one-class classification with only normal samples, requiring model retraining for new classes. Second, they struggle with unexpected defect patterns and the scarcity of defective samples for supervised learning. Lastly, their performance may vary due to differences between training and operational environments [4]. To address these issues, zero-shot anomaly segmentation (ZSAS) has emerged as an effective approach, enabling anomaly detection without requiring training data from the target domain. Recent advances in vision-language models (VLMs) have made ZSAS feasible, with CLIP [5] demonstrating strong generalization capabilities across domains by effectively learning class semantics from image-text pairs. CoOp [6] further enhanced CLIP by introducing learnable prompts to overcome the limitations of static prompt design. WinCLIP [7] advanced CLIP by introducing multi-scale window-based filters and a compositional prompt ensemble (CPE), highlighting the importance of multi-scale image analysis for AS. This approach emphasizes that effective AD requires both focus on specific regions and comprehensive object characteristic analysis, as illustrated in Fig. 1(b). The CPE framework offered a systematic method for optimizing prompts. AnoVL [8] advanced the field by incorporating local-aware features Value-Value (V-V) attention mechanism, along with domain-aware prompts. AnoVL+ subsequently improved this approach with test-time adaptation to better accommodate input image distributions. Despite these advancements, existing studies have yet to fully address the importance of resolution in AS for local defect detection, and maintaining consistent ZSAS performance remains challenging due to domain-dependent prompt designs.

Therefore, we propose DHR-CLIP, which introduces dynamic high-resolution (DHR) processing to enhance object recognition capabilities, as shown in Fig. 1(c). Furthermore, we adapted an object-agnostic prompt design to learn normal and anomalous patterns without object semantics. Finally, we improved the CLIP architecture for ZSAS by enhancing local feature extraction through V-V attention in the vision encoder and refining textual embeddings via deep-text prompt tuning in the text encoder. The key contributions are as follows:

- We introduce a novel approach, DHR-CLIP, for handling high-resolution images through DHR, while enabling prompt learning for generalized normal and anomalous patterns through object-agnostic prompts.
- We developed a framework that integrates deep-text prompt tuning and V-V attention mechanisms to achieve generalized ZSAS. This architecture optimizes the refinement of textual representations as well as the extraction of local visual features.
- We demonstrate the effectiveness and generalization capabilities of DHR-CLIP through experiments on real-world industrial datasets, MVTecAD and VisA, showing superior performance in ZSAS.

## II. Methodology

The proposed DHR-CLIP framework consists of three main components to enhance ZSAS: DHR processing, a vision encoder with V-V attention, and a text encoder with deep-text prompt tuning. Fig. 2 shows the overall architecture of DHR-CLIP. The framework starts by processing input images

*Corresponding author—Tel: 82-2-3290-3396; Fax: +82-2-3290-4550

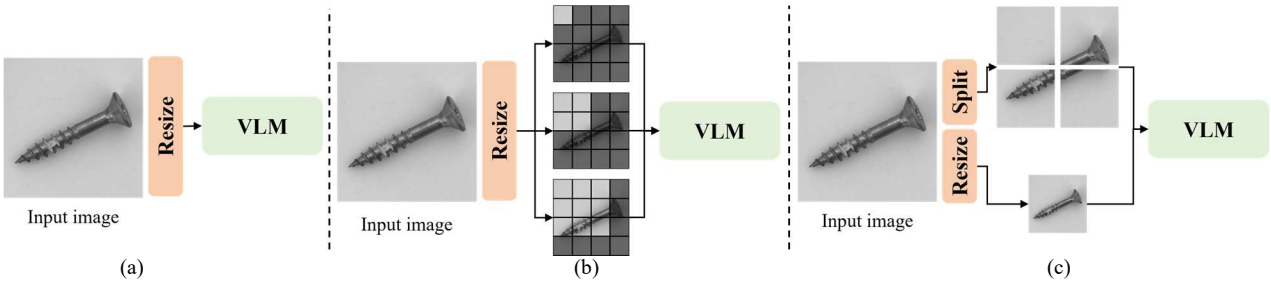


Figure 1. Comparison of image processing methods for anomaly segmentation. (a) Shows the conventional processing method, where the image is resized and normalized before being input to the model. (b) Demonstrates a window-based image processing method for fine-grained detection of local features. Since the image is divided using a window-based approach at a fixed image size, the size of each patch varies. (c) Illustrates the DHR processing method, which maintains the resolution of each patch while enabling fine-grained detection of local regions. This involves resizing the image and creating patches by splitting it into sections of the same size as the resized image.

through DHR to capture fine-grained details. The vision encoder uses V-V attention mechanisms across multiple blocks to extract detailed local features, while the text encoder applies object-agnostic prompts with deep-text tuning to generate refined textual representations. These components work together to enable precise anomaly detection and localization without requiring domain-specific training data.

### A. Problem definition

We first formulate the ZSAS problem. Given an input image $X = \{x_1, x_2, \dots, x_i\}$ where image $x_i \in \mathbb{R}^{H \times W \times C}$ represents the input with height $H$, width $W$, and channels $C$. Our goal is to learn prompts that effectively detect and segment anomalies in the test domain.

### B. Dynamic high-resolution

DHR designed to efficiently handle high-resolution images while preserving fine-grained details for precise AS. The method involves adaptive resolution handling through a two main branches: one focusing on patch-based processing and the other on glocal context reconstruction, as outlined below. Given the input image $x_i$, the DHR process divides the input into smaller patches while retaining the glocal context by resizing the entire image. This ensures efficient processing and high-resolution reconstruction suitable for AS. The input image $x_i$ is resized into $H', W'$ and divided into four non-overlapping sub-patches, each of size $(\frac{H'}{2}, \frac{W'}{2})$. These patches are denoted as $\{P_1, P_2, P_3, P_4\}$, where:

$$x_i = Concat(P_1, P_2, P_3, P_4),\ P_i \in \mathbb{R}^{\frac{H'}{2} \times \frac{W'}{2}},\ i = 1,2,3,4. \quad (1)$$

Additionally, the entire image $x_i$ is resized to $(\frac{H'}{2}, \frac{W'}{2})$ to form a glocal context patch $P_g$. After this patch-based processing, the features passed through the encoder undergo glocal context reconstruction to restore them to the original $H, W$.

### C. Vision encoder

In vision encoder, ViT [9] is used to extract local visual features. Each sub-patch $P_i$ and the resized global patch $P_g$ are passed through the shared vision encoder $f(\cdot)$, which extracts feature representations:

$$F_i = f(P_i),\ F_g = f(P_g),\ where\ F_i, F_g \in \mathbb{R}^d. \quad (2)$$

Instead of using conventional QKV attention mechanism, V-V attention is utilized to detect fine-grained defects by focusing on local regions [10]. This approach replaces both the query and key with the same value, which intensifies correlation among local features. As a result, the model can effectively capture both localized and holistic features that are essential for identifying anomalies at different scales.

### D. Object-agnostic prompt design

In DHR-CLIP, we applied an object-agnostic prompt design aimed at capturing normal and anomalous patterns without relying on specific class semantics [11]. Normal and anomaly text prompts, $t_n$ and $t_a$ are defined as follows:

$$\begin{aligned} t_n &= [W_i][object] \\ t_a &= [V_j][damaged][object]. \end{aligned} \quad (3)$$

Normal and anomaly prompts are crafted in a binary format, following an object-agnostic prompt structure. Specifically, $[W_i]$ $(i \in [1, E])$ denotes a learnable token for normal conditions, representing the general state of each object. While $[A_j]$ $(j \in [1, L])$ serves as a learnable token for anomalous conditions, indicating defects or damage specific to each object. By using the term `object` in a generalized manner, the design allows the prompts to learn object-agnostic representations, and the term `damaged` is manually integrated into the anomaly prompt to explicitly represent anomalous conditions.

### E. Text encoder

We utilized deep-text prompt tuning by inserting learnable tokens into each layer of the text encoder, refining text embeddings and enhancing their interaction with visual embeddings [10]. Specifically, at the $i$-th layer, the learnable token $t_i^{learnable}$ is concatenated with the text embeddings $t_i$, to adjust the text embedding. This process is represented as follows:

$$t_i' = [t_i^{learnable}, t_i]. \quad (4)$$

The updated text embeddings in each layer are passed to the next layer, allowing more detailed text information can be learned using a new $t_i^{learnable}$ token in each layer. Then the text embeddings are aligned with visual features, enabling a more accurate detection of normal and anomalous patterns.

### F. Training and Inference

The training process aims to optimize ZSAS through a specialized loss function. The total loss is represented as:

$$L_{total} = \sum_{M_k \in \mathcal{M}} L_{local}^{M_k}, \quad (5)$$

where $\mathcal{M}$ denotes a set of intermediate layers used to extract local features. For each spatial location $(j, k)$ in the feature map, where $j \in [1, H']$ and $k \in [1, W']$, we compute similarity using the function $A(\cdot)$, which calculates cosine

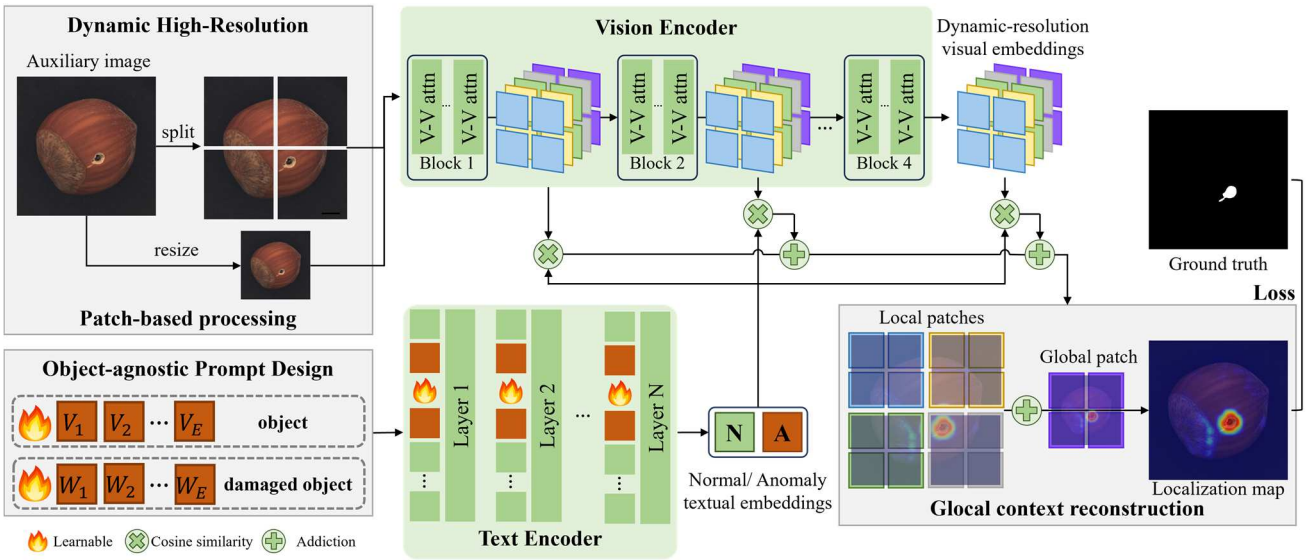


Figure 2. Overall architecture of DHR-CLIP

similarity between the extracted features for both normal and anomalous conditions. There similarity scores are represented as:

$$S_{n,i,M_k}^{(j,k)} = A\left(t_n, F_{i,M_k}^{m(j,k)}\right), S_{a,i,M_k}^{(j,k)} = A\left(t_a, F_{i,M_k}^{m(j,k)}\right), \quad (6)$$

where $F_{i,M_k}^{m(j,k)}$ is $i$-th sub-patch visual features in the $M_k$. Each sub-patches has similarity scores for both normal and anomaly conditions, denoted as $S_{i,M_k}^{n}$ and $S_{i,M_k}^{a}$, respectively. Simultaneously, the same process is applied for the global similarity scores. Finally local patches are reconstructed back into a high resolution representation as follows:

$$\begin{aligned} S_{l,M_k}^{n} &= R\left(S_{1,M_k}^{n}, S_{2,M_k}^{n}, S_{3,M_k}^{n}, S_{4,M_k}^{n}, (H', W')\right) \\ S_{l,M_k}^{a} &= R\left(S_{1,M_k}^{a}, S_{2,M_k}^{a}, S_{3,M_k}^{a}, S_{4,M_k}^{a}, (H', W')\right), \end{aligned} \quad (7)$$

where $R(\cdot)$ is reconstruction process. The global patch is resized to match the original resolution, denoted as $S_{g,M_k}^{n}$ and $S_{g,M_k}^{a}$. The final representation is obtained by $S_{DHR,M_k}^{n} = \frac{1}{2}\left(S_{l,M_k}^{n} + S_{g,M_k}^{n}\right)$, $S_{DHR,M_k}^{a} = \frac{1}{2}\left(S_{g,M_k}^{a} + S_{g,M_k}^{n}\right)$, adding element-wise. Given a ground-truth mask $S \in R^{H' \times W'}$, where $S_{j,k} = 1$ indicates anomalous regions and $S_{j,k} = 0$ indicates normal areas, the local loss is formulated as:

$$\begin{aligned} L_{local} = Focal\left(\left([S_{DHR,M_k}^{n}, S_{DHR,M_k}^{a}]\right), S\right) \\ + Dice\left(S_{DHR,M_k}^{n}, I - S\right) \\ + Dice\left(S_{DHR,M_k}^{n}, S\right), \end{aligned} \quad (8)$$

where $Focal(\cdot,\cdot)$ [12] and $Dice(\cdot,\cdot)$ [13] represent the loss functions. Focal loss assigns higher weights to important samples in imbalanced data, while dice loss is used to reduce the difference between the predicted and actual anomaly regions. $[\cdot,\cdot]$ represent channel-wise concatenation, and $I$ denotes a matrix with all elements equal to 1. During inference, anomaly segmentation are performed based on the anomaly localization map. The anomaly localization map, denoted as $Map \in \mathbb{R}^{H \times W}$, is computed as follows:

$$Map = G_\sigma\left(\sum_{M_k}\left(\frac{1}{2}\left(I - S_{DHR,M_k}^{n}\right) + \frac{1}{2}\left(S_{DHR,M_k}^{a}\right)\right)\right), \quad (9)$$

where $G_\sigma$ represents a gaussian filter and the parameter $\sigma$ controls the smoothing effect.

## III. Experiment

### A. Datasets and Implementation details

To evaluate the performance of the proposed DHR-CLIP, we conducted experiments on MVTecAD [14] and VisA [15], both from the industrial domain. MVTecAD consists of 15 classes, each containing object and texture defect categories. The dataset includes 467 normal samples and 1258 anomaly samples. VisA consists of 12 classes, each with object defect categories, and contains 962 normal samples and 1200 anomaly samples. Both datasets use photography as their modality. We adopted the `VIT-L/14@336px` CLIP model as the backbone, keeping all parameters of CLIP model frozen. The input image dimensions were set to $(672, 672)$ and was divided into $(336,336)$ patches. For the MVTecAD, we trained the model using the VisA test data, and for VisA we used MVTecAD test data. All experiments were conducted on a single NVIDIA RTX 4090 24 GB GPU.

### B. Comparison Methods and Evaluation Metrics

We compared our model against several state-of-the-art (SOTA) methods, including CLIP, CoOp, WinCLIP, AnoVL, and AnoVL+. To assess the effectiveness of anomaly detection, we employed the area under the receiver operating characteristic curve (AUROC) as the primary metric. For a more comprehensive analysis of anomaly segmentation performance, we utilized two additional metrics: AUPRO, which measures the accuracy of anomalous region localization, and F1-MAX, which evaluates the confidence level of model localization predictions. To ensure a thorough assessment of overall performance, we calculated mean values across all class categories.

### C. Quantitative Results

Table 1 presents the quantitative results comparing our proposed DHR-CLIP against existing SOTA methods across two major industrial datasets: MVTec AD and VisA. The experimental results demonstrate that DHR-CLIP consistently outperforms previous approaches across all evaluation metrics. On the MVTec AD dataset, DHR-CLIP achieves 91.7% AUROC, 85.9% PRO, and 39.7% F1Max, surpassing the previous best performance AnoVL+ by significant margins of 1.3%, 8.6%, and 26.4% respectively. The performance gap is

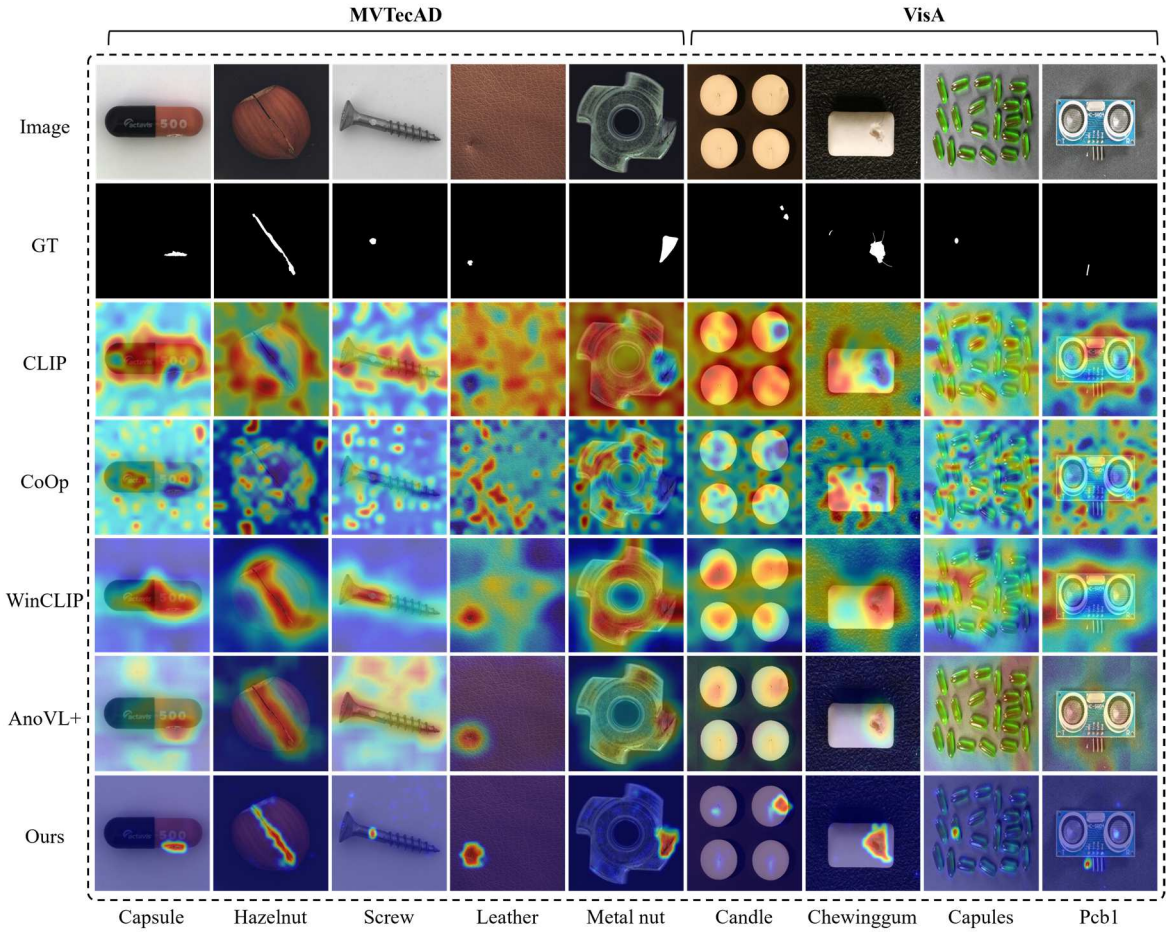


Figure 3. Qualitative results of experiments

TABLE I
QUANTITATIVE RESULTS OF EXPERIMENTS

| Dataset | MVTecAD | | | VisA | | |
|---|---|---|---|---|---|---|
| Method | AUROC | PRO | F1Max | AUROC | PRO | F1Max |
| CLIP | 38.2 | 8.8 | 6.8 | 47.9 | 16.1 | 3.2 |
| CoOp | 44.4 | 11.1 | 11.1 | 42.1 | 12.2 | 2.5 |
| WinCLIP | 82.3 | 61.9 | 24.8 | 73.2 | 51.1 | 9 |
| AnoVL | 86.6 | 70.4 | 30.1 | 85.9 | 65.1 | 14.7 |
| AnoVL+ | 90.5 | 79.1 | 31.4 | 89.8 | 70.9 | 16.2 |
| **Ours** | **91.7** | **85.9** | **39.7** | **93.8** | **83.7** | **22.1** |

even more pronounced on the VisA dataset, where DHR-CLIP attains 93.6% AUROC, 83.8% PRO, and 21.9% F1Max, showing substantial improvements over existing methods. These results validate that our dynamic high-resolution approach, combined with refined prompt learning and attention mechanisms, enables more precise ZSAS across different industrial scenarios. The consistent superior performance demonstrates its robust generalization capability in real-world industrial applications.

### *D. Qualitative Results*

Fig. 3 illustrates the qualitative results comparing DHR-CLIP with baseline methods across various industrial objects from the MVTec AD and VisA datasets. The visualization demonstrates that DHR-CLIP achieves more precise anomaly localization compared to CLIP, CoOp, WinCLIP and AnoVL+. While baseline methods generate diffuse or imprecise activation maps, DHR-CLIP accurately identifies defect regions across different scenarios. For instance, in the capsule and hazelnut examples, DHR-CLIP precisely localizes small and linear defects that align closely with ground truth masks. The method shows effectiveness in challenging cases such as subtle surface anomalies in leather and metal nuts, and complex geometric patterns in capsules and Pcb1 samples, where it successfully distinguishes true anomalies while maintaining clear object boundaries. These visual results validate that the integration of dynamic high-resolution processing and refined attention mechanisms enables more accurate and reliable ZSAS across diverse industrial inspection scenarios.

## IV. CONCLUSION

In this paper, we proposed DHR-CLIP, a refined framework for ZSAS that enhances industrial AD through DHR processing and object-agnostic prompt learning. By incorporating dynamic resolution handling for precise localization, along with deep-text prompt tuning and V-V attention mechanisms, our framework achieves improved feature extraction at both global and local levels. Through comprehensive experiments on MVTecAD and VisA datasets, we demonstrated the superior performance of DHR-CLIP across diverse industrial applications without requiring

domain-specific training data, highlighting its potential as a foundation for advancing multi-scale ZSAS methods. Future research could focus on developing adaptive strategies for efficient multi-scale feature integration and leveraging self-supervised learning to further enhance its applicability to industrial visual inspection systems.

## ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MSIT) (2022R1A2C2004457, RS-2024-00393801) And this research was also funded by Brain Korea 21 FOUR and Samsung Electronics Co., Ltd. (IO201210-07929-01).

## REFERENCES

[1] P. Osterrieder, L. Budde, and T. Friedli, "The smart factory as a key construct of industry 4.0: A systematic literature review," *International Journal of Production Economics*, vol. 221, p. 107476, 2020.

[2] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[3] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024.

[4] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

[5] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

[6] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6757–6767, 2023.

[7] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19606–19616, 2023.

[8] H. Deng, Z. Zhang, J. Bao, and X. Li, "Anovl: Adapting vision-language models for unified zero-shot anomaly localization," *arXiv preprint arXiv:2308.15939*, 2023.

[9] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] J. Ham, Y. Jung, and J.-G. Baek, "GlocalCLIP: Object-agnostic global-local prompt learning for zero-shot anomaly detection," *arXiv preprint arXiv:2411.06071*, 2024.

[11] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," *arXiv preprint arXiv:2310.18961*, 2023.

[12] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2980–2988, 2017.

[13] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," *arXiv preprint arXiv:1911.02855*, 2019.

[14] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019

[15] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," *European Conference on Computer Vision*, pp. 392–408, 2022.