ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888834

KAnoCLIP: Zero-Shot Anomaly Detection through Knowledge-Driven Prompt Learning and Enhanced Cross-Modal Integration

Chengyuan Li[1,†]*, Suyang Zhou*[1]*, Jieping Kong*[1]*, Lei Qi*[2]*, Hui Xue*[2]

[1] College of Software Engineering, Southeast University, Nanjing, China
[2] School of Computer Science and Engineering, Southeast University, Nanjing, China
$\{chengyuanli, suyangzhou, jiepingk, qilei, hxue\}$ @$seu.edu.cn$

ABSTRACT

Zero-shot anomaly detection (ZSAD) identifies anomalies without needing training samples from the target dataset, essential for scenarios with privacy concerns or limited data. Vision-language models like CLIP show potential in ZSAD but have limitations: relying on manually crafted fixed textual descriptions or anomaly prompts is time-consuming and prone to semantic ambiguity, and CLIP struggles with pixel-level anomaly segmentation, focusing more on global semantics than local details. To address these limitations, We introduce KAnoCLIP, a novel ZSAD framework that leverages vision-language models. KAnoCLIP combines general knowledge from a Large Language Model (GPT-3.5) and fine-grained, image-specific knowledge from a Visual Question Answering system (Llama3) via Knowledge-Driven Prompt Learning (KnPL). KnPL uses a knowledge-driven (KD) loss function to create learnable anomaly prompts, removing the need for fixed text prompts and enhancing generalization. KAnoCLIP includes the CLIP visual encoder with V-V attention (CLIP-VV), Bi-Directional Cross-Attention for Multi-Level Cross-Modal Interaction (Bi-CMCI), and Conv-Adapter. These components preserve local visual semantics, improve local cross-modal fusion, and align global visual features with textual information, enhancing pixel-level anomaly detection. KAnoCLIP achieves state-of-the-art performance in ZSAD across 12 industrial and medical datasets, demonstrating superior generalization compared to existing methods.

Index Terms— Zero-shot Anomaly Detection, Vision-Language Models, Prompt Learning

1. INTRODUCTION

Anomaly detection (AD) [1, 2] is crucial in fields like industrial quality assurance, medical diagnostics, and video analysis, encompassing anomaly classification (AC) and anomaly segmentation (AS) for image-level and pixel-level anomalies. The main challenges in AD are the rarity and diversity of anomalies, making dataset collection costly and time-consuming. Traditional one-class or unsupervised methods [2] often fall short due to the diverse and long-tail distribution of anomalies across domains such as industrial defects and medical lesions, necessitating a generalizable model. Zero-shot anomaly detection (ZSAD) [3–5] is vital for identifying anomalies without extensive training samples, particularly when data privacy concerns or insufficient labeled data are issues. Vision-Language Models (VLMs), like CLIP [6], have advanced ZSAD by leveraging training on large datasets of image-text pairs.

However, CLIP's original model still faces challenges in anomaly detection due to the task's complexity. There are two main limitations in applying CLIP to zero-shot anomaly detection: (1) existing methods [3, 4, 6, 7] rely on fixed text descriptions or anomaly prompts. For zero-shot anomaly detection using CLIP, a commonly used text prompt template is "a photo of a [class] with holes." However, these handcrafted prompts require extensive expertise, are time-consuming, and suffer from semantic ambiguity. Inspired by prompt learning in NLP, CoOp [8] utilizes learnable vectors for prompts and requires only a few labeled images. Despite this, CoOp tends to overfit to base classes, thus diminishing its ability to generalize to unseen classes. (2) while CLIP aligns image-level semantics with anomaly prompts through cross-modal contrastive training, it performs poorly in precise anomaly segmentation (pixel-level detection) because it emphasizes global semantics and overlooks local details [3, 6]. A mechanism for refining CLIP models in the local visual space and integrating local pixel-level cross-modal features is necessary for superior anomaly segmentation.

To address the two main limitations, we propose the KAnoCLIP framework, which introduces KnPL. This approach leverages large language model (LLM) and a visual question answering (VQA) system to generate anomaly descriptions, forming a knowledge base that guides the development of Learnable Normal Prompts (LNPs) and Learnable Abnormal Prompts (LAPs) using KD loss. This Loss minimizes the Euclidean distance between LLM-VQA-generated abnormal prompts and LAPs while maximizing the distance to LNPs, effectively eliminating the reliance on fixed text prompts and enhancing generalization to new anomaly classes. Furthermore, the framework integrates CLIP-VV, Bi-CMCI, and Conv-Adapter to preserve local visual semantics, improve cross-modal fusion, and align global visual features with textual information. These enhancements collectively improve the detection of subtle anomalies. Our main contributions are summarized as follows:

- KAnoCLIP is a novel zero-shot anomaly detection solution that doesn't require training samples from the target dataset, making it ideal for applications with privacy concerns or limited data, especially in industrial and medical fields.
- We introduce knowledge-driven prompt learning to eliminate manual text prompting and alleviate overfitting, thereby enhancing generalization to new anomaly classes.
- KAnoCLIP integrates CLIP-VV, Bi-CMCI, and Con-Adapter to refine local visual spaces and enhance cross-modal interactions, achieving robust pixel-level anomaly segmentation.
- Extensive experiments on 12 industrial and medical datasets demonstrate that KAnoCLIP consistently outperforms state-of-the-art techniques, setting a new benchmark for ZSAD.

† Corresponding author

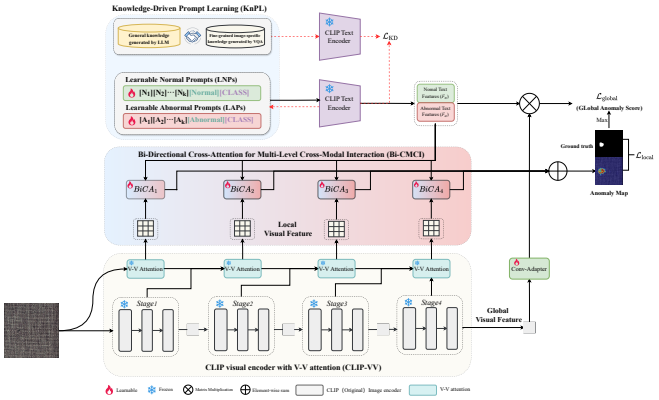

Fig. 1. Illustration of KAnoCLIP, which consists of four key components: KnPL, CLIP-VV, Bi-CMCI, and Conv-Adapter. KnPL uses an LLM and VQA system to form an LLM-VQA knowledge base, guiding the generation of learnable normal and abnormal prompts (LNPs and LAPs), reducing overfitting and enhancing generalization. The CLIP-VV visual encoder captures local visual details with V-V attention, while Conv-Adapter and Bi-CMCI provide comprehensive cross-modal fusion of global and local features. The red dashed line represents the $\mathcal{L}_{\text{KD}}$ loss function introduced by KnPL, guiding LNPs and LAPs learning during training.

2. THE METHOD

2.1. Overview

As shown in Figure 1, our KAnoCLIP framework introduces Knowledge-Driven Prompt Learning to remove the need for fixed text prompts and enhance generalization. By integrating CLIP-VV, Bi-CMCI, and Conv-Adapter, KAnoCLIP enhances local visual features, cross-modal interactions. These innovations significantly improve zero-shot anomaly detection performance.

During training, the KAnoCLIP framework minimizes the loss function $\mathcal{L}_{\text{total}}$ in Equation 13 using an auxiliary anomaly detection dataset. During inference, a test image x_i is processed through the CLIP-VV visual encoder to extract patch features F^i_{patch}. LNPs and LAPs, previously learned normal and abnormal text prompts, are inputted into the CLIP text encoder to extract text features F_{text}. These patch and text features are combined in the Bi-CMCI module to generate an anomaly map $M_i \in \mathbb{R}^{H\times W}$ for each layer. The maps are summed and normalized to create the final anomaly map M. After adjusting global visual features using the Conv-Adapter, the maximum value of M determines the global anomaly score.

2.2. Knowledge-Driven Prompt learning

CoOp-based prompt learning [8] uses the pre-trained CLIP model for downstream tasks, but overfits to base class objects, reducing performance on unseen classes. KgCoOp [9] found that performance drops are linked to the distance between learnable (CoOp) and fixed (CLIP) prompts. By reducing this distance, generalization improves. Inspired by KgCoOp, we initially used fixed CLIP prompts (e.g., "a photo of an abnormal [class]") to guide learnable anomaly prompts, but this failed due to insufficient anomaly knowledge. We propose combining LLMs' general knowledge with VLMs' detailed image descriptions to create a knowledge-driven prompt learning (KnPL) method for zero-shot anomaly detection.

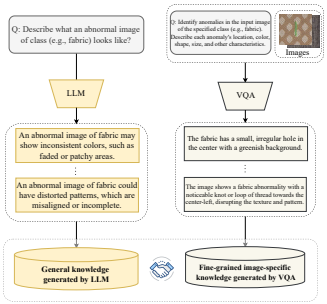

Fig. 2. Constructing the LLM-VQA Knowledge Base: Generating Potential Anomalies and Image-Specific Descriptions.

2.2.1. *Constructing the LLM-VQA Knowledge Base*

As shown in Figure 2, for each anomaly category, the LLM (GPT-3.5) generates n optimal anomaly descriptions and background information (default n=5). The prompt template is: "Q: Describe what an abnormal image of class (e.g., fabric) looks like?" LLM-generated descriptions such as: "An abnormal image of fabric may show inconsistent colors, such as faded or patchy areas". The VQA model (Llama3) generates m image-specific anomaly descriptions per image (default m=1). The prompt template is: "<IMAGE> + Q: Identify anomalies in the input image of the specified class (e.g., fabric). Describe each anomaly's location, color, shape, size, and other characteristics." VQA-generated descriptions such as: "The fabric has a small, irregular hole in the center with a greenish background". We refer to the anomaly descriptions generated by the LLM and VQA as the LLM-VQA knowledge base.

2.2.2. *Guiding Learnable Normal/Abnormal Prompts*

The knowledge-driven Learnable Normal Prompts (LNPs) and Learnable Abnormal Prompts (LAPs) are defined as follows:

$$\begin{aligned} P_n &= [N_1][N_2]\dots[N_K][Normal][class] \\ P_a &= [A_1][A_2]\dots[A_K][Abnormal][class], \end{aligned} \tag{1}$$

where N_i and A_i ($i \in \{1,\dots,K\}$) are learnable word embeddings for normal and abnormal text prompts, respectively. K denotes the length of the learnable prefix. [Normal] includes adjectives like "normal" and "perfect", while [Abnormal] includes "abnormal" and "defective". [class] represents the anomaly category to be detected.

We introduce a Knowledge-Driven (KD) loss function, $\mathcal{L}_{\text{KD}}$, to guide and constrain the generation of LNPs and LAPs. This loss minimizes the Euclidean distance between abnormal prompts generated by LLM-VQA and the LAPs, while simultaneously maximizing the distance to the LNPs.

$$\mathcal{L}_{KD} = \max\left(0, d(\frac{\bar{w}^k}{\|\bar{w}^k\|_2}, \frac{\bar{w}^a}{\|\bar{w}^a\|_2}) - d(\frac{\bar{w}^k}{\|\bar{w}^k\|_2}, \frac{\bar{w}^n}{\|\bar{w}^n\|_2})\right), \tag{2}$$

where $d(\cdot,\cdot)$ represent the Euclidean distance. $\bar{\mathbf{w}}^n$ and $\bar{\mathbf{w}}^a$ are the means of all LNPs/LAPs features, while $\bar{\mathbf{w}}^k$ denotes the mean of the abnormal prompts features generated by LLM-VQA:

$$\bar{\mathbf{w}}^k = \frac{\sum_{i=1}^{N} g(\mathbf{p}_i^{LLM}) + \sum_{j=1}^{M} g(\mathbf{p}_j^{VQA})}{N+M}, \tag{3}$$

where $g(\cdot)$ be the CLIP text encoder. $\mathbf{p}_i^{LLM}$ and $\mathbf{p}_i^{VQA}$ represent abnormal text prompts generated by LLM and VQA, respectively, with N and M being their respective counts.

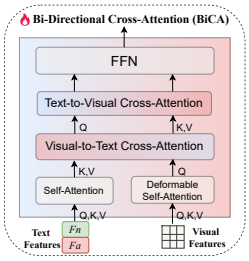

Fig. 3. BiCA.

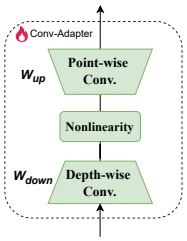

Fig. 4. Conv-Adapter.

2.3. Refining Local Visual Space with V-V Attention

The original CLIP visual encoder, pre-trained with contrastive loss, produces global embeddings and uses a Q-K self-attention mechanism that disrupts local visual semantics, impairing fine-grained anomaly detection. To address this, we introduce a V-V attention mechanism [10] into the CLIP visual encoder, enhancing the extraction of local visual features crucial for pixel-level anomaly segmentation without altering the original structure. V-V attention preserves local visual semantics by focusing on relationships between local features, resulting in attention maps with a distinct diagonal pattern. This enhanced visual encoder is termed CLIP-VV, and its feature extraction process is described as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{Q} \cdot \mathbf{K}^{\mathrm{T}} \cdot scale) \cdot \mathbf{V} \quad (4)$$

$$S_{l-1}^{ori} = [s_{cls}; s_1; s_2; \ldots; s_T] \quad (5)$$

$$S_{l-1} = [s'_{cls}; s'_1; s'_2; \ldots; s'_T] \quad (6)$$

$$[Q_l, K_l, V_l] = [W_q S_{l-1}^{ori}, W_k S_{l-1}^{ori}, W_v S_{l-1}^{ori}]. \quad (7)$$

using the V-V attention mechanism, the output of layer l is:

$$S_l = Project_l(Attention(V_l, V_l, V_l)) + S_{l-1}, \quad (8)$$

where S_{l-1}^{ori} is the original output and S_{l-1} is the local-aware output. $Project_l$ are linear projections. Final outputs are S_l. For anomaly detection, $S_l[0]$ is used for image-level detection, and $S_l[1:]$ for pixel-level detection.

2.4. Enhancing Cross-Modal Feature Integration

We propose the Bi-Directional Cross-Attention for Multi-Level Cross-Modal Interaction (Bi-CMCI) and Conv-Adapter to address the limitations of traditional cross-modal feature fusion methods, such as linear layer projections [7]. These traditional methods often struggle to capture complex interactions and dependencies between visual and textual modalities, which are crucial for effective anomaly detection. Bi-CMCI manages local cross-modal interactions, while Conv-Adapter aligns global features. This dual approach effectively captures both local details and global context, thereby enhancing the model's accuracy in detecting subtle anomalies.

Bi-CMCI. As illustrated in Figure 1 and 3, Bi-CMCI employs a bi-directional cross-attention (BiCA) [11] mechanism to align textual descriptions with visual parts through Text-to-Visual Cross-Attention and refine text with local visual features via Visual-to-Text Cross-Attention. Self-Attention ensures independent understanding within each modality, while Deformable Self-Attention [12] adapts to key visual parts, addressing variations in shape and position. Bi-CMCI operates across the four stages of the CLIP-VV visual encoder to capture multi-level detailed information for identifying subtle anomalies. Each stage extracts intermediate patch-level features $F_{patch}^i \in \mathbb{R}^{H_i W_i \times C_i}$. The LNPs and LAPs are input into the CLIP Text Encoder $g(\cdot)$, producing text features representing normal and abnormal cases, $F_{text} = [F_n, F_a] \in \mathbb{R}^{2 \times C}$. The Bi-CMCI module then deeply fuses the cross-modal features F_{patch}^i and F_{text}, resulting in the normal map $M_i^n \in \mathbb{R}^{H \times W}$, anomaly map $M_i^a \in \mathbb{R}^{H \times W}$, and the final localization result M.

$$M^n, M^a = Norm(\sum_{i=1}^{4} BiCA_i(F_{\text{patch}}^i, F_{\text{text}})) \quad (9)$$

$$M = \frac{G_\sigma(M^a + 1 - M^n)}{2}, \quad (10)$$

where $\mathrm{BiCA_i}(\cdot,\cdot)$ represents the Bi-Directional Cross-Attention operation between the local visual and textual features at the i-th stage. $\mathrm{Norm}(\cdot)$ normalizes the anomaly map values to the range of 0 to 1. G_σ is a Gaussian filter with σ controlling the smoothing.

Conv-Adapter. As illustrated in Figure 4, the Conv-Adapter uses a depthwise separable convolution bottleneck [13] to better align global visual and textual features. It includes three main components: depthwise convolution for capturing spatial features, activation functions for nonlinearity, and pointwise convolution for dimensionality reduction and efficiency. The Conv-Adapter and global anomaly score is calculated as:

$$\hat{I}_G = \mathrm{ReLU}(\mathrm{LN}(I_G)W_{\text{down}})W_{\text{up}} \quad (11)$$

$$S_{global} = \mathrm{softmax}(\hat{I}_G \cdot F_{text}^T) + \max(M), \quad (12)$$

where I_G denotes the global visual feature, LN represents layer normalization, $W_{\text{down}} \in \mathbb{R}^{d \times d_{\text{bottle}}}$ is the down-projection weight matrix, and $W_{\text{up}} \in \mathbb{R}^{d_{\text{bottle}} \times d}$ is the up-projection weight matrix. The ReLU activation function is denoted as ReLU. The output feature after the Conv-Adapter transformation is $\hat{I}_G$. M is the anomaly map calculated in equation 10, and $\max(\cdot)$ denotes the maximum operation.

2.5. Joint Optimization

We develop joint optimization, a method for effectively learning knowledge-driven text prompts from both global and local perspectives. The total loss function $\mathcal{L}_{\text{total}}$ is defined as follows:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{KD}} + \beta\mathcal{L}_{\text{global}} + \gamma\mathcal{L}_{\text{local}}, \quad (13)$$

where the hyperparameters α, β, and γ balance the three loss components, and we default all of them to 1.

Global Loss ($\mathcal{L}_{\text{global}}$). The global loss is a binary cross-entropy loss that matches the cosine similarity between text embeddings and visual embeddings from auxiliary data of normal/anomalous images.

$$\mathcal{L}_{\text{global}} = \mathrm{BCE}(S_{\text{global}}, \mathit{Label}), \quad (14)$$

where S_{global} represents the global anomaly score as calculated in equation 12, and $Label$ indicates whether the image is anomalous.

Local Loss ($\mathcal{L}_{\text{local}}$). The local loss combines focal loss [14] and dice loss [15] to handle fine-grained local anomalous regions in the intermediate layers of the visual encoder.

$$\mathcal{L}_{\text{local}} = \mathrm{Focal}(M^a, G) + \mathrm{Dice}(M^n, I - G) + \mathrm{Dice}(M^a, G), \quad (15)$$

where $G \in \mathbb{R}^{H \times W}$ is the ground truth segmentation mask, with $G_{ij} = 1$ for anomalous pixels and $G_{ij} = 0$ otherwise. I denotes a matrix of ones.

Table 1. ZSAD Performance comparison in industrial domain. The best-performing result is highlighted in bold and red, while the second-best is highlighted in blue.

Method	Public	MVTec-AD		VisA		MPDD		BTAD		SDD		DAGM	
		Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC
CLIP	ICML 2021	74.1	38.4	66.4	46.6	54.3	62.1	34.5	30.6	65.7	39	79.6	28.2
CLIP-AC	IMCL 2021	71.5	38.2	65	47.8	56.2	58.7	51	32.8	65.2	32.5	82.5	32.7
WinCLIP	CVPR 2023	91.8	85.1	78.1	79.6	63.6	76.4	68.2	72.7	84.3	68.8	91.8	87.6
April-GAN	CVPR 2023	86.1	87.6	78	94.2	73	94.1	73.6	60.8	79.8	79.8	94.4	92.4
AnomalyCLIP	ICLR 2024	91.5	91.1	82.1	95.5	77	96.5	88.3	94.2	84.7	90.6	**97.5**	95.6
MVFA-AD	CVPR 2024	88.9	89.3	82.3	94.8	76.4	96.6	80.1	81.4	82.5	91.7	95.8	95.1
Ours (KAnoCLIP)	—	**93.1**	**94.3**	**83.8**	**97.7**	**77.8**	**98.3**	**90.6**	**96.5**	**86.2**	**93.2**	97.3	**96.8**

Table 2. ZSAD Performance comparison in medical domain. Since the Br35H, COVID-19, and HeadCT datasets lack pixel-level anomaly segmentation ground truth, we only performed anomaly classification on these three datasets.

Method	Public	BrainMRI		LiverCT		RESC		Br35H		COVID-19		HeadCT	
		Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC
CLIP	ICML 2021	53.9	71.7	57.1	82.3	38.5	70.8	78.4	-	73.7	-	56.5	-
CLIP-AC	IMCL 2021	60.6	66.4	61.9	88.4	40.3	75.9	82.7	-	75	-	60.0	-
WinCLIP	CVPR 2023	66.5	86.0	64.2	92.2	42.5	80.6	80.5	-	66.4	-	81.8	-
April-GAN	CVPR 2023	76.4	91.8	70.6	94.1	75.6	85.2	93.1	-	15.5	-	89.1	-
AnomalyCLIP	ICLR 2024	78.3	92.2	73.7	96.2	83.3	91.5	**94.6**	-	80.1	-	93.4	-
MVFA-AD	CVPR 2024	78.6	90.3	74.2	95.9	84.2	92.0	90.5	-	83.5	-	91.3	-
Ours (KAnoCLIP)	—	**80.7**	**93.5**	**78.2**	**98.6**	**84.8**	**93.5**	94.2	-	**85.4**	-	**95.8**	-

3. EXPERIMENTS

3.1. EXPERIMENT SETUP

Datasets. We conducted experiments on 12 datasets, covering industrial and medical anomaly detection. Industrial datasets included MVTec-AD [16] , VisA [17], MPDD [18], BTAD [19], SDD [20], and DAGM [21]. Medical datasets were BrainMRI [22], LiverCT [23], RESC [24], Br35H [25], COVID-19 [26], and HeadCT [22].
Evaluation Metrics. We used the Area Under the Receiver Operating Characteristic Curve (AUC) for evaluation, applying image-level AUC for classification and pixel-level AUC for segmentation.
Baselines. Our KAnoCLIP was thoroughly compared with state-of-the-art methods, including CLIP [6], CLIP-AC [6], WinCLIP [4], APRIL-GAN [7], MVFA-AD [3], and AnomalyCLIP [5].
Implementation details. We used the CLIP model (VIT-L/14@336px) as our backbone, keeping all parameters frozen. Following APRIL-GAN and AnomalyCLIP, we trained our model on the MVTec-AD test data and evaluated the ZSAD performance on other datasets. Learnable word embeddings length was set to 12. Experiments were conducted in PyTorch 2.0.0 on a single NVIDIA RTX 3090 GPU.

3.2. MAIN RESULTS

3.2.1. ZSAD Performance Comparison

As illustrated in Tables 1 and 2, our KAnoCLIP model has achieved state-of-the-art results across 12 industrial and medical anomaly detection datasets, surpassing the performance of MVFA-AD and AnomalyCLIP. Notably, in the prominent MVTec-AD and VisA datasets, our method improved pixel-level AUC scores by **3.2** and **2.2**, respectively, and enhanced image-level AUC scores by **1.3** and **1.5**. This highlights our superior performance in both anomaly segmentation and anomaly classification. Additionally, across six medical datasets, KAnoCLIP exhibited significant improvements, particularly in BrainMRI and LiverCT, with image-level AUC increases of **2.1** and **4.0** and pixel-level AUC enhancements of **1.3** and **2.4**. Overall, KAnoCLIP has demonstrated robust effectiveness and generalization across a diverse range of anomaly detection datasets.

3.2.2. Results Analysis and Discussion

The original CLIP model performed poorly because its text prompts were designed for image classification, not anomaly detection. CLIP-AC made slight improvements by adding "normal" and "abnormal" descriptors but still fell short. WinCLIP and APRIL-GAN, with manual crafted anomaly detection prompts, performed better. AnomalyCLIP, using learnable anomaly text prompts, further improved zero-shot performance but overfit to base classes without general knowledge guidance. MVFA-AD, an adapter for WinCLIP, enhanced cross-domain performance but lacked learnable prompts and effective cross-modal interactions. Our KAnoCLIP method achieved the best results by improving cross-modal interactions and integrating global-local context, making it highly effective for zero-shot anomaly detection in industrial and medical applications.

Table 3. The ablation results for our KAnoCLIP framework demonstrate the incremental effects of adding each module.

KnPL	CLIP-VV	Bi-CMCI	Conv-Adapter	MVTec-AD		VisA	
				Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC
✗	✗	✗	✗	66.8	55.4	58.2	47.1
✓	✗	✗	✗	85.9	87.4	78.6	91.2
✓	✓	✗	✗	87.5	88.9	80.3	92.3
✓	✓	✓	✗	92.0	92.4	82.3	95.8
✓	✓	✓	✓	**93.1**	**94.3**	**83.8**	**97.7**

3.3. ABLATION STUDY

We conducted ablation experiments on the MVTec-AD and VisA datasets to validate the effectiveness of the four key modules in the KAnoCLIP framework: KnPL, CLIP-VV, Bi-CMCI, and Conv-Adapter. Each module significantly enhanced zero-shot anomaly detection, with KnPL providing the most substantial improvement. Detailed results are in Table 3.

4. CONCLUSION

In this paper, we introduce KAnoCLIP, a zero-shot anomaly detection framework using KnPL, which integrates LLM's general knowledge with VQA system's image-specific insights to create learnable anomaly prompts, eliminating fixed text prompts and enhancing generalization to new anomaly classes. KAnoCLIP employs three modules: CLIP-VV, Bi-CMCI, and Conv-Adapter, optimizing visual space, cross-modal interactions, and global-local context integration. Extensive experiments on 12 industrial and medical datasets show that KAnoCLIP consistently outperforms SOTA methods, demonstrating superior generalization capabilities.

5. REFERENCES

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

[2] Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro, "Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images," *Medical Image Analysis*, p. 102930, 2023.

[3] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang, "Adapting visual-language models for generalizable anomaly detection in medical images," *arXiv preprint arXiv:2403.12570*, 2024.

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

[5] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," *arXiv preprint arXiv:2310.18961*, 2023.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

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

[8] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

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

[10] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li, "Clip surgery for better explainability with enhancement in open-vocabulary tasks," *arXiv preprint arXiv:2304.05653*, 2023.

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen, "Cross attention network for few-shot classification," *Advances in neural information processing systems*, vol. 32, 2019.

[12] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.

[13] Lukasz Kaiser, Aidan N Gomez, and Francois Chollet, "Depthwise separable convolutions for neural machine translation," *arXiv preprint arXiv:1706.03059*, 2017.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[15] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li, "Dice loss for data-imbalanced nlp tasks," *arXiv preprint arXiv:1911.02855*, 2019.

[16] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600.

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

[18] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak, "Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions," in *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*. IEEE, 2021, pp. 66–71.

[19] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2021, pp. 01–06.

[20] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *Journal of Intelligent Manufacturing*, vol. 31, no. 3, pp. 759–776, 2020.

[21] Matthias Wieler and Tobias Hahn, "Weakly supervised learning for industrial optical inspection," in *DAGM symposium in*, 2007, vol. 6.

[22] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14902–14912.

[23] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al., "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, pp. 102680, 2023.

[24] Junjie Hu, Yuanyuan Chen, and Zhang Yi, "Automated segmentation of macular edema in oct using deep neural networks," *Medical image analysis*, vol. 55, pp. 216–227, 2019.

[25] Mohammed BOURENNANE and Hilal NAIMI, "Deep feature extraction with cubic-svm for classification brain tumor," *Rare Metal Materials and Engineering*, vol. 52, no. 9, pp. 54–64, 2023.

[26] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al., "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," *Computers in biology and medicine*, vol. 132, pp. 104319, 2021.