

Supplementary Materials: FiLo: Zero-Shot Anomaly Detection by Fine-Grained Description and High-Quality Localization

Anonymous Authors

1 FINE-GRAINED ZSAD PERFORMANCE

In the main paper, we have compared FiLo with existing ZSAD methods on anomaly detection and localization across the MVTec [1] and VisA [10] datasets. Our evaluation primarily utilizes Image-level AUC and Pixel-level AUC as metrics for detection and localization, respectively. Here, we provide detailed performance analysis of FiLo and other ZSAD methods at the fine-grained data subset level, including the methods we using for comparison: CLIP [6], CLIP-AC [6], WinCLIP [4], APRIL-GAN [2] and AnomalyCLIP [9].

Tables 1 and Tables 2 depict the anomaly localization performance of FiLo on the MVTec and VisA datasets, and the anomaly detection performance of FiLo on the VisA and MVTec datasets is showcased in Table 4 and Table 3 respectively. Across the 15 classes in the MVTec dataset, FiLo achieves the highest Pixel-level AUC in 12 classes, while in the VisA dataset comprising 12 classes, FiLo attains the highest Pixel-level AUC in 8 classes. Notably, FiLo surpasses the state-of-the-art method AnomalyCLIP [9] by 1.1% on Pixel-level AUC in the MVTec dataset and by 0.4% in the VisA dataset, demonstrating the efficacy of FiLo.

2 FINE-GRAINED ANOMALY DESCRIPTIONS

Table 5 and Table 6 present the detailed anomaly types generated by leveraging LLM for each category within the MVTec and VisA datasets. During the inference process with FiLo, we substitute these detailed anomaly descriptions generated by LLM for the "[ANOMALY CLASS]" portion in the text template to obtain the detailed anomaly description content for each category of items.

In Figure 1, we additionally display the similarity between each detailed anomaly description generated by LLM and the image features. We showcase the top 5 detailed anomaly descriptions with the highest similarity to the image, highlighting the most similar descriptions in red. By identifying the detailed anomaly description with the highest similarity, we can further discern the type of anomaly present in the sample.

3 ADDITIONAL ABLATIONS

In this section, we conducted further ablation studies on various detailed components of FiLo, including the backbone utilized, learning rate, employment of VV Attention, different treatments on QKV and VV Attention results, learning strategies for adaptive learning templates, number of learnable vectors, the structure and connectivity of Adapters, etc. Subsequently, detailed analyses for each aspect will be presented.

3.1 Different Backbones and Learning Rates

Previous anomaly detection methods based on CLIP have typically utilized different CLIP backbones. WinCLIP[4] employs ViT-B-16@240px, while methods like APRIL-GAN [2] and AnomalyCLIP [9] use ViT-L-14@336px. Existing methods have shown that

using a backbone with higher image resolution is more beneficial for pixel-level anomaly localization. However, these methods with higher resolutions have not surpassed WinCLIP, which uses a resolution of 240x240, in terms of image-level AUC. We also implemented our FiLo method on these two commonly used backbones, and the results are shown in Table 7.

In addition to the choice of backbone, the setting of learning rates also significantly influences experimental results. Table 7 further illustrates the experimental results of FiLo under different learning rates ranging from 1e-3 to 1e-5. It can be observed that FiLo achieves the best anomaly detection and localization performance on both datasets when using a learning rate of 1e-3 for the learnable text vectors and a learning rate of 1e-4 for the MMCI module.

3.2 Adaptively Learned Text Templates

CoOp [8] and CoCoOp [7] are two distinct methods that utilize learnable vectors to replace manually crafted text prompts. These methods exhibit some differences in their approaches. Specifically, the learnable vectors in CoOp are agnostic to image content and are directly embedded into the text prompt, emphasizing the universality and uniformity of the text prompt. On the other hand, CoCoOp builds upon the learnable vectors embedded in the text prompt by incorporating a lightweight meta-net to append image features to the text prompt. This approach emphasizes generating tailored text prompts for each image, aiming to better match the image content.

Table 8 presents the experimental results of FiLo under the respective usage of CoOp and CoCoOp. Inspired by AnomalyCLIP [9], we also explored the performance under the addition of class name information in the text content. The experimental results indicate that when using CoOp, omitting class name from the text yields better results, consistent with findings in AnomalyCLIP. This is because CoOp inherently emphasizes the generality and uniformity of the text prompt. Conversely, when employing CoCoOp for learning text templates, including class name information improves performance. This is attributed to the alignment of CoCoOp's approach, which incorporates image features into the text prompt via a meta-net, with the concept of FiLo, utilizing fine-grained anomaly description and position enhancement to obtain precise representations of each image's text content, aiming for a better match with image content.

The results in Table 8 further demonstrate that CoCoOp outperforms CoOp, highlighting the effectiveness of leveraging fine-grained anomaly descriptions to enhance anomaly detection.

We also examined the impact of varying the number of learnable vectors in adaptively learned text templates. The findings are illustrated in Figure 2. It can be observed that utilizing 12 learnable vectors yields the best performance in both anomaly detection and localization tasks.

Table 1: Fine-grained data-subset-wise performance comparison (AUROC) for anomaly localization on MVTec-AD. The best performance is in bold, and the second-best is underlined.

Object name	CLIP [6]	CLIP-AC [6]	WinCLIP [4]	APRIL-GAN [2]	AnomalyCLIP [9]	FiLo (ours)
Carpet	11.5	10.7	95.4	98.4	<u>98.8</u>	99.4
Bottle	17.5	23.3	89.5	83.4	<u>90.4</u>	92.6
Hazelnut	25.2	34.0	94.3	96.1	<u>97.1</u>	97.6
Leather	9.9	5.6	96.7	<u>99.1</u>	98.6	99.4
Cable	37.4	37.5	77.0	72.3	78.9	<u>78.4</u>
Capsule	50.9	49.1	86.9	92.0	<u>95.8</u>	96.9
Grid	8.7	11.9	82.2	95.8	<u>97.3</u>	97.8
Pill	55.8	60.8	80.0	76.2	92	<u>89.1</u>
Transistor	51.1	48.5	<u>74.7</u>	62.4	71	74.8
Metal_nut	43.9	53.6	61.0	65.4	74.4	<u>72.5</u>
Screw	80.1	76.4	89.6	<u>97.8</u>	97.5	98.1
Toothbrush	36.3	35.0	86.9	<u>95.8</u>	91.9	96.0
Zipper	51.5	44.7	<u>91.6</u>	91.1	91.4	96.6
Tile	49.9	39.1	<u>77.6</u>	92.7	<u>94.6</u>	97.4
Wood	45.7	42.4	93.4	95.8	<u>96.5</u>	98.3
Mean	38.4	38.2	85.1	87.6	<u>91.1</u>	92.3

Table 2: Fine-grained data-subset-wise performance comparison (AUROC) for anomaly localization on VisA. The best performance is in bold, and the second-best is underlined.

Object name	CLIP [6]	CLIP-AC [6]	WinCLIP [4]	APRIL-GAN [2]	AnomalyCLIP [9]	FiLo (ours)
Candle	33.6	50.0	88.9	97.8	98.8	<u>98.7</u>
Capsules	56.8	61.5	81.6	97.5	<u>95.0</u>	92.3
Cashew	64.5	62.5	84.7	86.0	<u>93.8</u>	95.1
Chewinggum	43.0	56.5	93.3	99.5	99.3	<u>99.4</u>
Fryum	45.6	62.7	88.5	92.0	<u>94.6</u>	95.2
Macaroni1	20.3	22.9	70.9	<u>98.8</u>	98.3	99.1
Macaroni2	37.7	28.8	59.3	<u>97.8</u>	97.6	98.1
Pcb1	57.8	51.6	61.2	92.7	<u>94.1</u>	94.4
Pcb2	34.7	38.4	71.6	89.7	<u>92.4</u>	93.7
Pcb3	54.6	44.6	85.3	<u>88.4</u>	<u>88.4</u>	90.8
Pcb4	52.1	49.9	94.4	94.6	<u>95.7</u>	95.8
Pipe_fryum	58.7	44.7	75.4	96.0	98.2	<u>97.7</u>
Mean	46.6	47.8	79.6	94.2	<u>95.5</u>	95.9

3.3 Utilization of V-V Attention

Pre-trained on large-scale datasets, CLIP exhibits excellent zero-shot performance on downstream image classification tasks. However, directly using the features extracted from the CLIP Image Encoder for each position in the feature map and measuring their similarity with textual features often results in significant noise activation outside of objects during fine-grained semantic segmentation or object detection tasks. CLIP Surgery [5] addresses this issue, identifying it as stemming from the QKV attention mechanism within CLIP, which leads to feature pooling from semantically disparate regions, consequently causing noise activation in erroneous areas. The proposed solution involves employing V-V self-attention to mitigate this problem.

Approaches such as AnoVL [3] and AnomalyCLIP [9] have also incorporated V-V attention into anomaly detection and localization tasks, resolving the issue of misalignment between patch-level features and textual features encountered in WinCLIP and APRIL-GAN, achieving remarkable zero-shot performance. However, V-V attention suffers from training difficulties, as slight mishandling may result in model outputs entirely comprised of zeros, causing the AUC to plummet to 50. To address this challenge, we simultaneously utilize the output results of both QKV attention and V-V attention, exploring the differential effects of applying distinct processing methods to the output results of QKV attention and V-V attention. The results, as shown in Table 9, indicate that employing a simple linear layer projection on the output results of QKV attention and



Figure 1: Illustration of similarities between images and different fine-grained anomaly descriptions.

inputting the output results of V-V attention into the MMCI module yields the best detection and localization performance for FiLo.

3.4 Ablations of Adapter

In this section, we compare the performance impact of the structure and connection methods of the adapter on FiLo. Regarding structure, we test the use of a simple linear layer and the bottleneck adapter actually employed. We also conduct experiments to assess the performance difference of the adapter when utilizing residual connection versus not utilizing it. Experimental results are shown in Table 10. It can be observed that when employing the bottleneck structure without residual connection, the adapter achieves the best performance.

3.5 Convolution Kernel's Shape of MMCI

In this section, we extensively experiment on the impact of different kernel shapes used in MMCI on the performance of FiLo. We start with the sole use of 1x1 convolutional kernels and gradually incorporate shapes such as 3x3, 5x5, 7x7, 1x5, 5x1, and 9x9. We then evaluate the various experimental results under these settings, as depicted in Figure 3. Based on the experimental findings, we ultimately select a combination of kernel shapes including 1x1, 3x3, 5x5, 7x7, 1x5, and 5x1. This combination harnesses the advantages of multi-scale and multi-shape kernels, enabling precise localization of anomalous regions of different sizes and shapes.

Table 3: Fine-grained data-subset-wise performance comparison (AUROC) for anomaly detection on MVTec AD. The best performance is in bold, and the second-best is underlined.

Object name	CLIP [6]	CLIP-AC [6]	WinCLIP [4]	APRIL-GAN [2]	AnomalyCLIP [9]	FiLo (ours)
Carpet	96	93.1	100.0	99.5	100.0	<u>99.9</u>
Bottle	45.9	46.1	99.2	92.0	89.3	<u>98.6</u>
Hazelnut	88.7	91.1	<u>93.9</u>	89.6	97.2	92.8
Leather	99.4	99.5	100.0	99.7	<u>99.8</u>	100
Cable	58.1	46.6	<u>86.5</u>	88.4	69.8	77.9
Capsule	71.4	68.8	72.9	79.9	89.9	<u>89.2</u>
Grid	72.5	63.7	98.8	86.3	97.0	<u>97.4</u>
Pill	73.6	73.8	79.1	80.5	<u>81.8</u>	87.8
Transistor	48.8	51.2	<u>88.0</u>	80.8	92.8	80.5
Metal_nut	62.8	63.4	97.1	68.4	<u>93.6</u>	77
Screw	78.2	66.7	<u>83.3</u>	84.9	81.1	74.5
Toothbrush	73.3	<u>89.2</u>	88.0	53.8	84.7	94.2
Zipper	60.1	36.1	91.5	89.6	98.5	<u>98.1</u>
Tile	88.5	89.0	100.0	<u>99.9</u>	100.0	100
Wood	94	94.9	<u>99.4</u>	99.0	96.8	99.7
Mean	74.1	71.5	91.8	86.1	<u>91.5</u>	91.2

Table 4: Fine-grained data-subset-wise performance comparison (AUROC) for anomaly detection on VisA. The best performance is in bold, and the second-best is underlined.

Object name	CLIP [6]	CLIP-AC [6]	WinCLIP [4]	APRIL-GAN [2]	AnomalyCLIP [9]	FiLo (ours)
Candle	37.9	33.0	95.4	<u>83.8</u>	79.3	79.3
Capsules	69.7	75.3	85.0	61.2	81.5	80.9
Cashew	69.1	72.7	92.1	87.3	76.3	<u>90</u>
Chewinggum	77.5	76.9	96.5	96.4	<u>97.4</u>	98.4
Fryum	67.2	60.9	80.3	94.3	<u>93.0</u>	88.3
Macaroni1	64.4	67.4	76.2	71.6	<u>87.2</u>	88.3
Macaroni2	65	65.7	63.7	64.6	73.4	<u>68.5</u>
Pcb1	54.9	43.9	73.6	53.4	<u>85.4</u>	87
Pcb2	62.6	59.5	51.2	<u>71.8</u>	62.2	77.6
Pcb3	52.2	49.0	73.4	66.8	62.7	<u>69.5</u>
Pcb4	87.7	89.0	79.6	<u>95.0</u>	93.9	95.7
Pipe_fryum	88.8	86.4	69.7	<u>89.9</u>	92.4	83.8
Mean	66.4	65.0	78.1	78.0	<u>82.1</u>	83.9

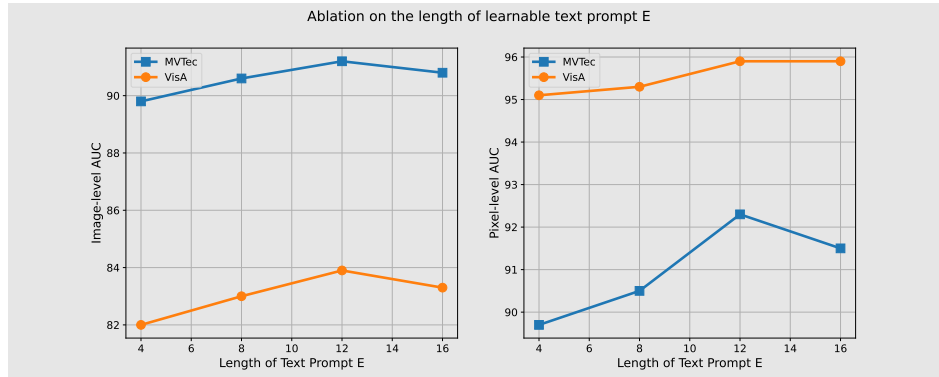


Figure 2: Comparison of FiLo on MVTec and VisA datasets with different numbers of learnable vectors.

Table 5: Fine-Grained anomaly description of every object within MVTec dataset.

Class name	Descriptions
Carpet	discoloration in a specific area, irregular patch or section with a different texture, frayed edges or unraveling fibers, burn mark or scorching
Grid	crooked, cracks, excessive gaps, discoloration, deformation, missing, inconsistent spacing between grid elements, corrosion, visible signs, chipping
Leather	scratches, discoloration, creases, uneven texture, tears, brittleness, damage, seams, heat damage, mold
Tile	chipped, irregularities, discoloration, efflorescence, warping, missing, depressions, lippage, fungus, damage
Wood	knots, warping, cracks along the grain, mold growth on the surface, staining from water damage, wood rot, woodworm holes, rough patches, protruding knots
Bottle	cracked large, cracked small, dented large, dented small, leaking, discolored, deformed, missing cap, excessive condensation, unusual odor
Cable	twisted, knotted cable strands, detached connectors, excessive stretching, dents, corrosion, scorching along the cable, exposed conductive material
Capsule	irregular shape, discoloration coloring, crinkled, uneven seam, condensation inside the capsule, foreign particles, unusually soft or hard
Hazelnut	fungal growth, Unusual discoloration, rotten or foul odor emanating, insect infestation, wetness, misshapen shell, unusually thin, contaminants, unusual texture
Metal_nut	cracks, irregular threading, corrosion, missing, distortion, signs of discoloration, excessive wear on contact surfaces, inconsistent texture
Pill	irregular shape, crumbling texture, excessive powder, uneven coating, presence of air bubbles, disintegration, abnormal specks
Screw	rust on the surface, bent, damaged threads, stripped threads, deformed top, coating damage, uneven grooves, inconsistent size
Toothbrush	loose bristles, uneven bristle distribution, excessive shedding of bristles, staining on the bristles, abrasive texture, irregularities in the shape
Transistor	burn marks, detached leads, signs of corrosion, irregularities in the shape, presence of cracks or fractures, signs of physical trauma, irregularities in the surface texture
Zipper	bent, frayed, misaligned, excessive stiffness, corroded, detaches, loose, warped

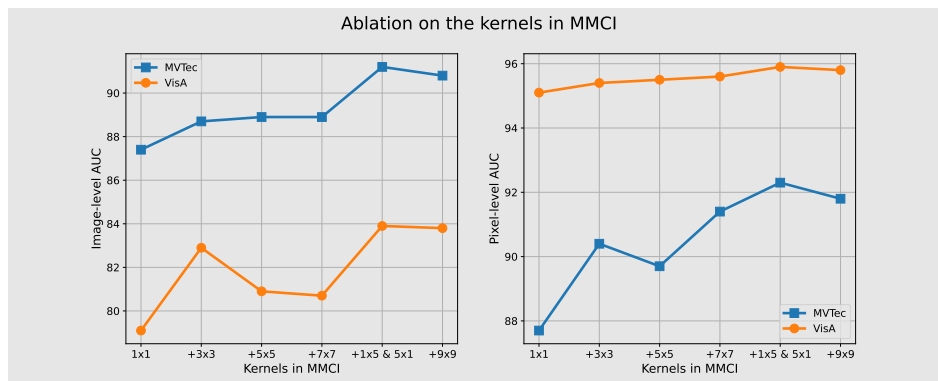
**Figure 3: Comparison of FiLo on MVTec and VisA datasets with different convolution kernels.**

Table 6: Fine-Grained anomaly description of every object within VisA dataset.

Class name	Descriptions
Candle	cracks or fissures in the wax, Wax pooling unevenly around the wick, tunneling, incomplete wax melt pool, irregular or flickering flame, other, extra wax in candle, wax melded out of the candle
Capsules	uneven capsule size, capsule shell appears brittle, excessively soft, dents, condensation, irregular seams or joints, specks
Cashew	uneven coloring, fungal growth, presence of foreign objects, unusual texture, empty shells, signs of moisture, stuck together
Chewinggum	consistency, presence of foreign objects, uneven coloring, excessive hardness, similar colour spot
Fryum	irregular shape, unusual odor, uneven coloring, unusual texture, small scratches, different colour spot, fryum stuck together, other
Macaroni1	uneven shape , small scratches, small cracks, uneven coloring, signs of insect infestation, uneven texture, Unusual consistency
Macaroni2	irregular shape, small scratches, presence of foreign particles, excessive moisture, signs of infestation, small cracks, unusual texture
Pcb1	oxidation on the copper traces, separation of layers, presence of solder bridges, excessive solder residue, discoloration, Uneven solder joints, bowing of the board, missing vias
Pcb2	oxidation on the copper traces, separation of layers, presence of solder bridges, excessive solder residue, discoloration, Uneven solder joints, bowing of the board, missing vias
Pcb3	oxidation on the copper traces, separation of layers, presence of solder bridges, excessive solder residue, discoloration, Uneven solder joints, bowing of the board, missing vias
Pcb4	oxidation on the copper traces, separation of layers, presence of solder bridges, excessive solder residue, discoloration, Uneven solder joints, bowing of the board, missing vias
Pipe_fryum	uneven shape, presence of foreign objects, different colour spot, unusual odor, empty interior, unusual texture, similar colour spot, stuck together

Table 7: Experimental results of FiLo on MVTec and VisA datasets under different backbones and learning rates.

Backbone	learnable vec's lr	MMCI's lr	VisA		MVTec-AD	
			Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC
ViT-B-16@240	1e-3	1e-4	78.1	93.5	77.9	88.2
ViT-L-14@336	1e-3	1e-4	83.9	95.9	91.2	92.3
ViT-L-14@336	1e-3	1e-3	80.3	95.7	86.2	89.7
ViT-L-14@336	1e-4	1e-4	82.4	95.7	88	91.2
ViT-L-14@336	1e-4	1e-5	78.2	95.1	83.5	89
ViT-L-14@336	1e-5	1e-5	80.4	95.2	85.8	90.7

Table 8: Comparison of different learning methods for learnable vectors and whether to use class name.

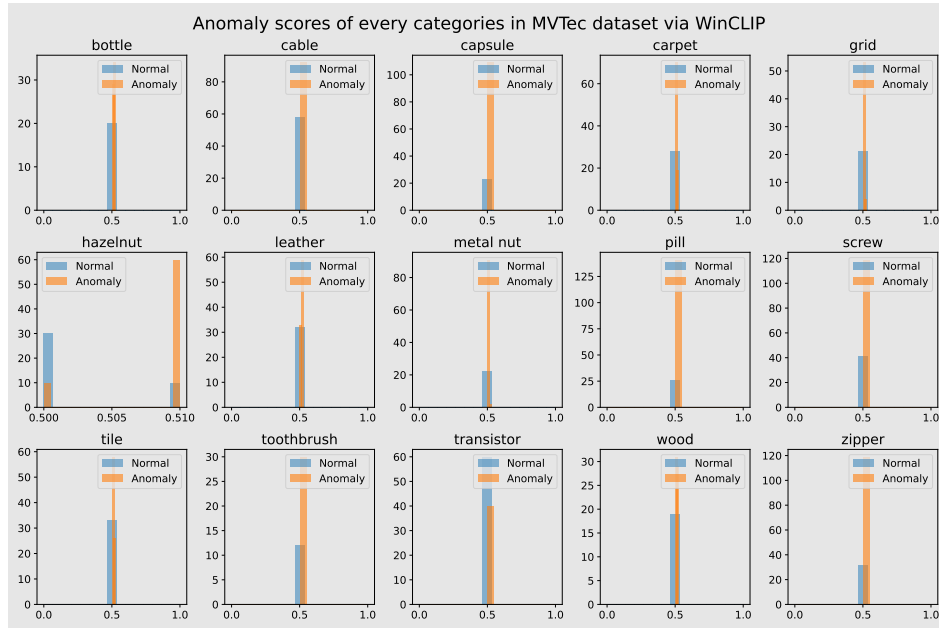
learning method	with class name	VisA		MVTec-AD	
		Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC
CoOp		81.7	95	90.8	89.5
CoOp	✓	80.9	95.5	89.9	90.4
CoCoOp		82.3	95.4	91	90.5
CoCoOp	✓	83.9	95.9	91.2	92.3

Table 9: Comparison of results of different processing methods for the output results of QKV and VV Attention.

QKV results	VV results	VV's first layer	VisA		MVTec-AD	
			Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC
Linear	MMCI	1	81.9	95.3	87.8	89.2
Linear	MMCI	7	83.9	95.9	91.2	92.3
Linear	Linear	7	82.7	95.1	88.5	88.8
MMCI	MMCI	7	83.2	95.5	50	50
MMCI	Linear	7	82.5	95.7	56.9	57.6

Table 10: Comparison of different adapter structures and connection types.

Adapter's arch	residual	VisA		MVTec-AD	
		Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC
Bottleneck		83.9	95.9	91.2	92.3
Bottleneck	✓	83.6	95.8	89.9	91.4
Linear		83.9	95.9	90.2	92.3
Linear	✓	83.8	95.9	88.6	91.1

**Figure 4: Anomaly scores of WinCLIP on the MVTec dataset. Each sub-figure represents the visualization of one object.**

4 VISUALIZATION

4.1 Anomaly Scores for Every Categories

In this section, we present the statistical analysis of anomaly scores generated by WinCLIP [4] and FiLo for each class object in the MVTec and VisA datasets. These visualizations aim to illustrate the effectiveness of FiLo's detailed anomaly descriptions and adaptively learned text templates compared to WinCLIP's manually crafted two-class text adjustment. As depicted in Figure 4 and Figure 5, WinCLIP's scores for both normal and abnormal samples heavily overlap and are concentrated around 0.5, indicating its failure to

effectively distinguish between normal and abnormal samples. In contrast, Figure 6 and Figure 7 illustrate FiLo's visualization results on these two datasets, where it can be observed that the scores for normal samples significantly decrease while those for abnormal samples notably increase, resulting in a significant reduction in the overlapping area.

4.2 Anomaly Maps

Figure 8 showcase the Anomaly Maps generated by FiLo on a broader set of samples from the MVTec and VisA datasets, demonstrating FiLo's robust anomaly localization capability.

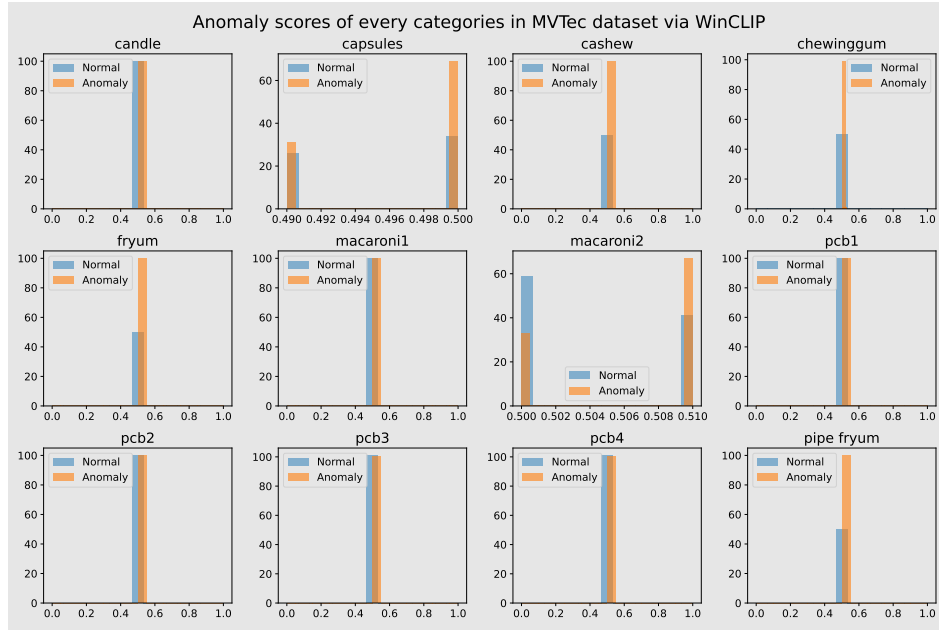


Figure 5: Anomaly scores of WinCLIP on the VisA dataset. Each sub-figure represents the visualization of one object.

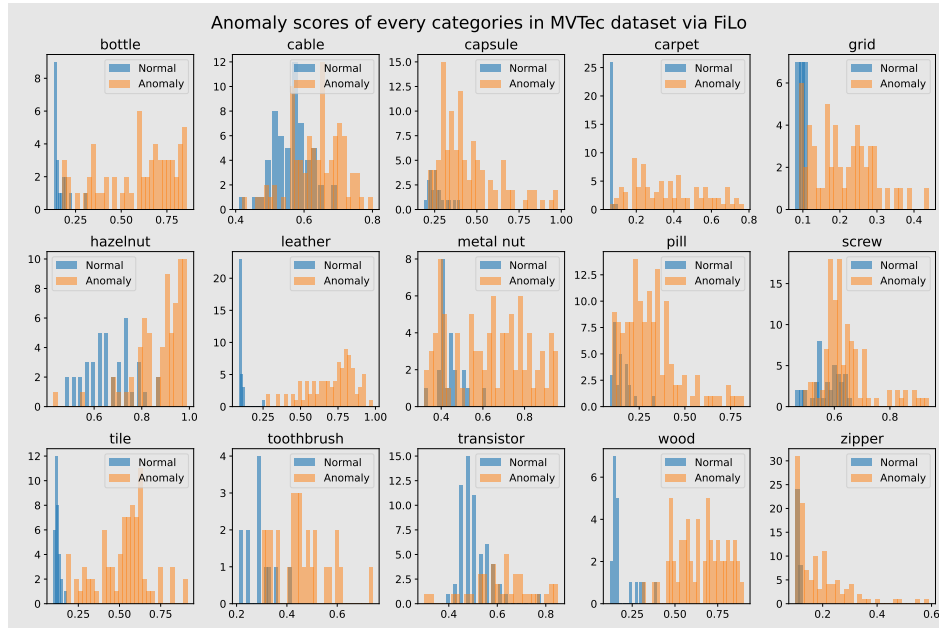


Figure 6: Anomaly scores of FiLo on the MVTec dataset. Each sub-figure represents the visualization of one object.

5 LIMITATION AND FUTURE WORK

Compared to previous works like WinCLIP [4], FiLo has made advancements in anomaly detection, localization, and interpretability through the use of Fine-Grained Description and High-Quality Localization methods. However, despite these strides forward, certain limitations still persist, warranting further investigation and

refinement. As illustrated in Figure 6 and Figure 7, while the differentiation between normal and abnormal samples is more distinct compared to previous methods, significant overlap still exists in certain categories such as zipper and metal nut. In the future, we plan to further improve the differentiation between normal and abnormal sample scores through approaches such as metric learning.

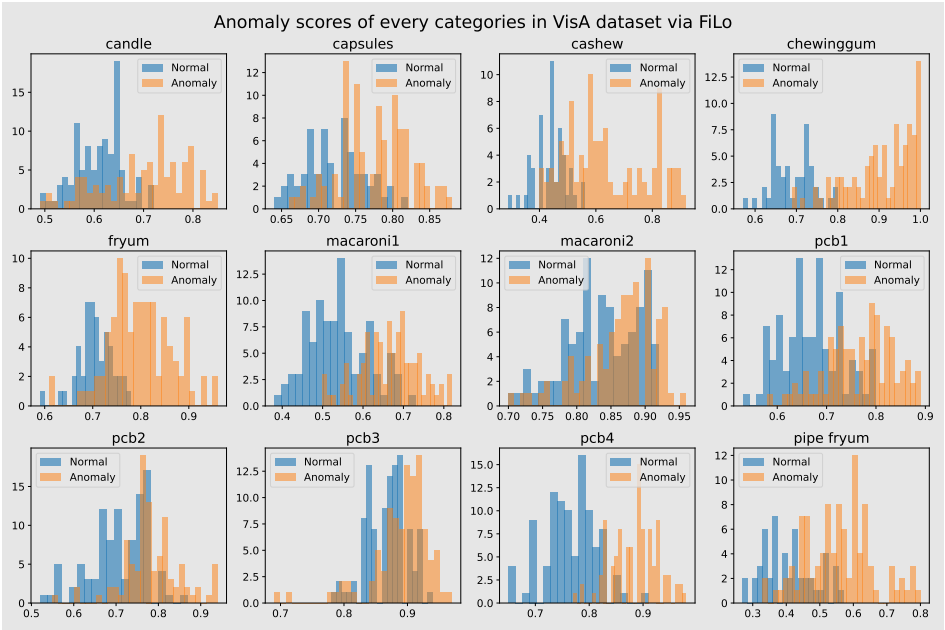


Figure 7: Anomaly scores of FiLo on the VisA dataset. Each sub-figure represents the visualization of one object.

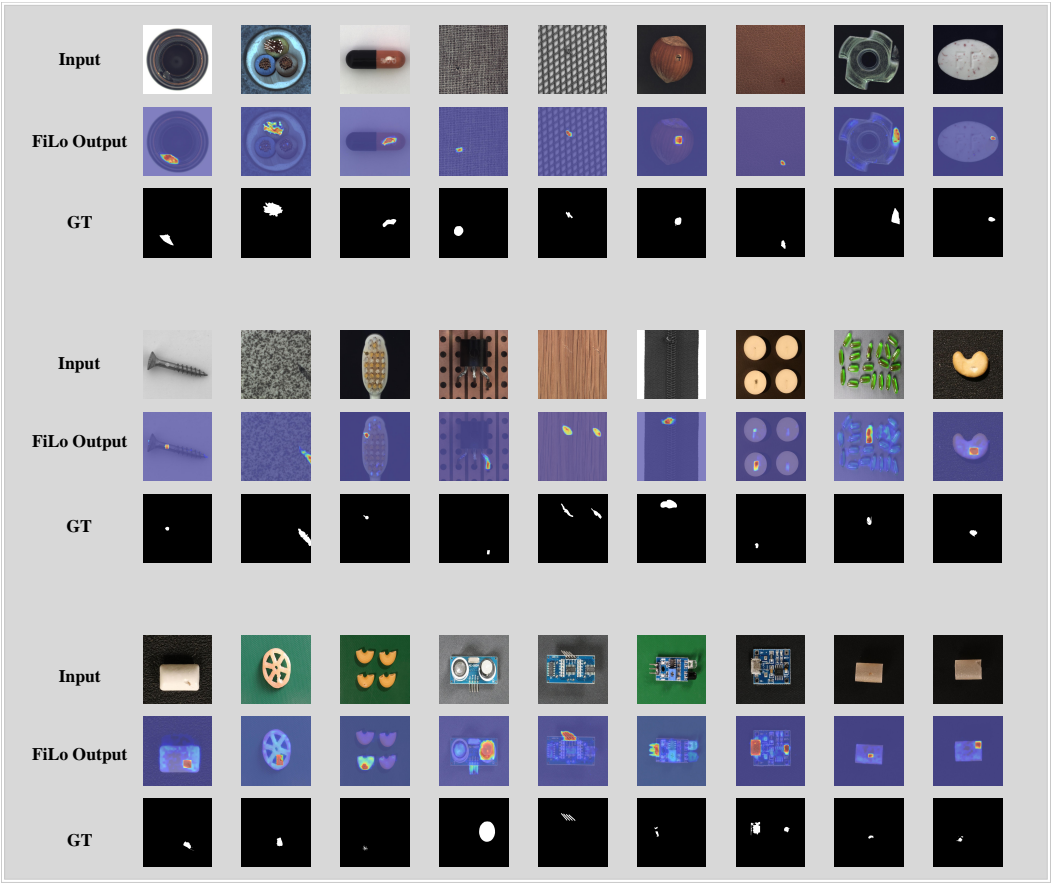


Figure 8: More visualization results on MVTec and VisA datasets.

REFERENCES

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9592–9600.

[2] Xuhai Chen, Yue Han, and Jiangning Zhang. 2023. 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).

[3] Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. 2023. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939* (2023).

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

[5] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653* (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. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16816–16825.

[8] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.

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

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