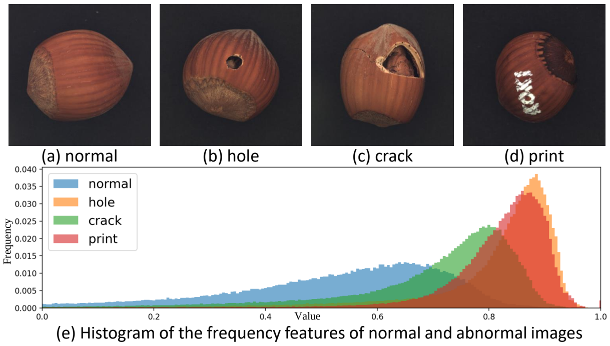
Supplementary Material for FE-CLIP: Frequency Enhanced CLIP Model for Zero-Shot Anomaly Detection and Segmentation

Tao Gong [1, 2, 3], Qi Chu [1, 2, 3*], Bin Liu[1, 2, 3], Wei Zhou[4], Nenghai Yu[1, 2, 3]
[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]Anhui Province Key Laboratory of Digital Security
[3]the CCCD Key Lab of Ministry of Culture and Tourism
[4]Ling Yang Industrial Internet Co., Ltd.
{tgong, qchu, flowice, ynh}@ustc.edu.cn, wezhou8@iflytek.com


(a) normal (b) hole (c) crack (d) print

(e) Histogram of the frequency features of normal and abnormal images

Figure 1. Histogram of the Frequency Features.

1. More Experiments

1.1. Histogram of the Frequency Features

We normalize the frequency features summed from FFE and LFS adapters, and then compute the histogram. As shown in the figure 1, the histogram of the frequency features of abnormal images is less uniform than normal images, which further validates the effectiveness of frequency-based cues from another view.

1.2. Compared with complex text prompts

WinCLIP [1] proposes a complex compositional ensemble on state words and prompt templates to build the anomaly text prompt and normal text prompt. Instead, FE-CLIP uses simple text prompts described in section 3.2 of the main manuscript. To validate that the simple text prompts also work well, we also attempt to use the complex text prompts of WinCLIP. As shown in Table 1, using complex text prompts of WinCLIP and simple text prompts achieves similar performance on both MVTec AD and VisA datasets. Therefore, we choose simple text prompt as the default setting due to its simplicity.

1.3. More Ablation Study on the LFS Adapter

In the proposed LFS adapter, we compute the mean of low-frequency and high-frequency responses from $Q \times Q$ ($Q = 3$) group frequency responses as the local frequency statistics. Here we explore the concatenation of low-frequency and high-frequency responses and then use a single conv to compress the channel to be consistent with the channel of visual features in the CLIP visual encoder. As shown in Table 2, the concatenation operation achieves inferior results than the mean operation.

We also explore the mean of the last 3 groups and the mean of the last 6 groups of the 3×3 group frequency responses, respectively. Table 2 shows that computing the mean of all 3×3 group frequency responses achieves the best results, which demonstrates that all frequency responses could be beneficial for the ZSAD and ZSAS tasks.

1.4. Experiments on Hyperparameters

We analyze the effect of three hyperparameters λ, P, and Q, standing for the ratio for injecting the frequency information from the FFE and LFS adapters, the size of local window DCT in FFE adapter, the size of sliding window DCT in LFS adapter, respectively.

Table 3 shows the effect of different λ. With the λ increasing from 0.05 to 0.1, the performance keeps increasing on MVTec AD and VisA datasets. This shows that more frequency information can benefit the ZSAD and ZSAS tasks. However, the performance has a downward trend when λ is larger than 0.1, leading to an inferior performance. The reasons are that injecting too much frequency information inevitably destroys the original feature distribution of the CLIP visual encoder and the frequency information of the boundary of the object is also enhanced, making the model

[*] Correponding Author

Methods	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
Text prompts of WinCLIP [1]	92.8 / 88.2	91.8 / 96.7	95.7 / 93.1	84.7 / 86.7
Simple text prompts	92.6 / 88.3	91.9 / 96.5	95.9 / 92.8	84.6 / 86.6

Table 1. Compared with complex combination of text prompts of WinCLIP. The metric of Pixel-level is AUROC / PRO for the ZSAS task, and the metric of Image-level is AUROC / AP for the ZSAD task. Best results are highlighted in red.

Methods	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
Mean of the last 3 groups from the 3 × 3 group frequency responses	91.2 / 86.9	90.4 / 95.7	95.2 / 91.7	83.4 / 85.5
Mean of the last 6 groups from the 3 × 3 group frequency responses	92.1 / 87.7	91.2 / 95.8	95.3 / 92.4	84.0 / 86.1
Concatenation of the 3 × 3 group frequency responses	92.3 / 87.9	91.6 / 95.9	95.2 / 92.4	83.6 / 86.3
Mean of the 3 × 3 group frequency responses (Ours)	92.6 / 88.3	91.9 / 96.5	95.9 / 92.8	84.6 / 86.6

Table 2. More exploration for the $Q \times Q$ ($Q = 3$) group frequency responses in the LFS adapter. The metric of Pixel-level is AUROC / PRO for ZSAS task, and the metric of Image-level is AUROC / AP for ZSAD task. Best results are highlighted in red.

λ	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
0.05	89.2 / 84.8	88.0 / 94.1	94.5 / 88.5	82.7 / 83.7
0.1	92.6 / 88.3	91.9 / 96.5	95.9 / 92.8	84.6 / 86.6
0.2	90.9 / 86.8	90.4 / 95.2	96.1 / 92.3	83.8 / 85.3
0.3	89.5 / 85.0	88.8 / 93.9	94.1 / 89.0	81.9 / 83.5

Table 3. The effect of λ, standing for the ratio for injecting the frequency information from the FFE and LFS adapters. The metric of Pixel-level is AUROC / PRO for ZSAS task, and the metric of Image-level is AUROC / AP for ZSAD task. Best results are highlighted in red.

Q	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
3	92.6 / 88.3	91.9 / 96.5	95.9 / 92.8	84.6 / 86.6
5	91.1 / 87.2	90.9 / 94.9	94.7 / 91.7	83.5 / 85.7
7	89.9 / 85.8	89.8 / 93.1	93.5 / 89.9	81.7 / 83.4

Table 4. The effect of Q, standing for the size of sliding window DCT in the LFS adapter. The metric of Pixel-level is AUROC / PRO for ZSAS task, and the metric of Image-level is AUROC / AP for ZSAD task. Best results are highlighted in red.

P	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
2	92.5 / 88.1	91.8 / 96.2	95.5 / 92.5	84.7 / 86.8
3	92.6 / 88.3	91.9 / 96.5	95.9 / 92.8	84.6 / 86.6
4	92.3 / 88.1	91.6 / 95.9	95.3 / 92.2	84.5 / 86.5
6	90.1 / 86.0	90.2 / 93.5	93.9 / 90.2	82.3 / 84.7

Table 5. The effect of P, standing for the size of local window DCT in the FFE adapter. The metric of Pixel-level is AUROC / PRO for ZSAS task, and the metric of Image-level is AUROC / AP for ZSAD task. Best results are highlighted in red.

confused to localize the abnormal regions. Therefore, λ is set to 0.1 as the default setting.

Table 5 and Table 4 show the effect of different P and Q, respectively. We can conclude that $P = 3$ and $Q = 3$ achieve the best performance on both datasets. It will lead to performance degradation if P or Q is set to a too-large value. The reason may be that a large window size will cover both the anomaly region and the normal region, therefore, the frequency patterns of the anomaly region inevitably are weakened due to the existence of the normal region.

2. Visualization

We also visualize the results of the proposed FE-CLIP. As shown in Figure 2, the proposed FE-CLIP successfully localizes the abnormal region of the input image from highly diverse class semantics (e.g. tile in the second column, pill in the fifth column, and toothbrush in the sixth column) from various defect inspections (e.g. crack in the second column, cut in the third column, and bent in the fourth column).

References

[1] 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*, pages 19606–19616, 2023. 1, 2

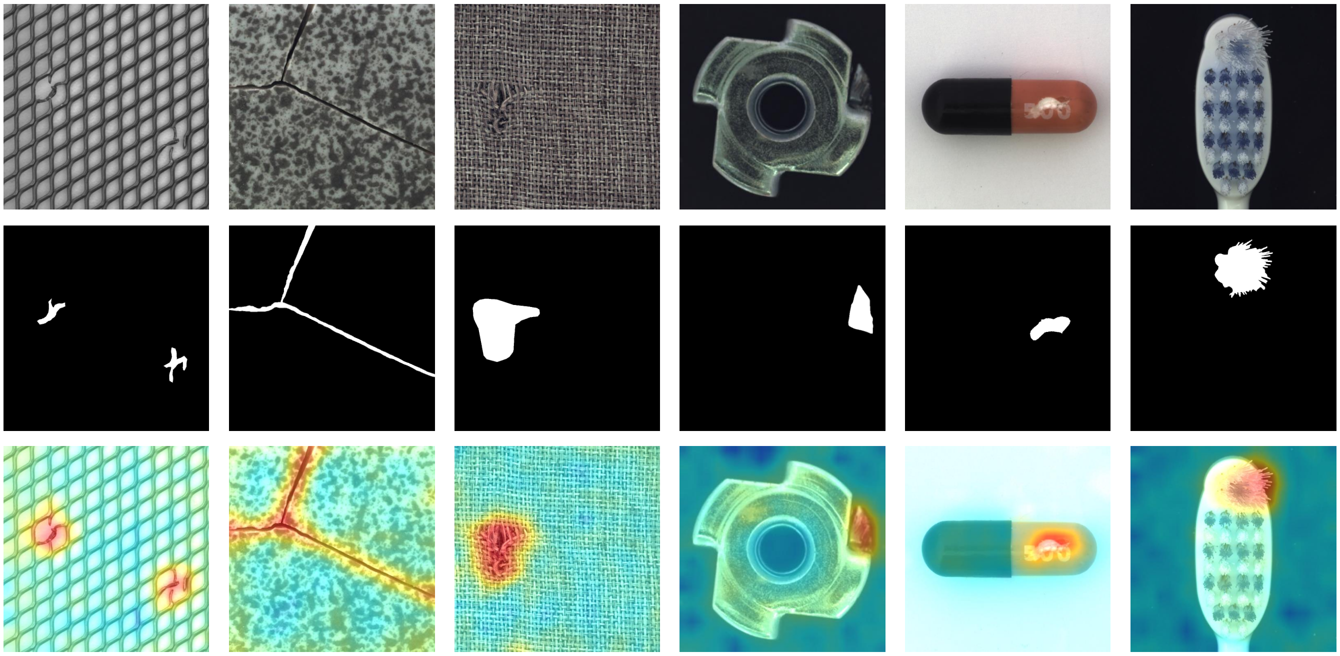


Figure 2. The visualization results of the proposed FE-CLIP. The first row represents the input abnormal image. The second row represents the ground-truth mask, where the white region denotes the abnormal region. The third row represents the prediction of abnormal score map, where the red color denotes the abnormal region and the blue-green color denotes the normal region.