

Supplementary Materials

SimCLIP: Refining Image-Text Alignment with Simple Prompts for Zero-/Few-shot Anomaly Detection

Anonymous Authors

1 MORE EXPERIMENTAL DETAILS

In this paper, we conduct experiments using the pre-trained CLIP with ViT-L/14@336px as our foundational framework. The hierarchy indices L for the multi-hierarchy vision adapter are specified as $\{6, 12, 18, 24\}$. Input images are resized into 518×518 . Our primary focus is to investigate the model’s generalization performance in anomaly detection tasks involving unseen classes. Therefore, we finetune the pre-trained CLIP using the test set from a publicly available dataset (e.g., MVTec-AD) and then evaluate the performance using another dataset[1–3, 6]. The finetune process is conducted on a single NVIDIA-3080Ti GPU over 5 epochs. The Adam optimizer is employed with a learning rate of $1e-3$ for updating the model parameters, coupled with a batch size of 8.

2 ADDITIONAL QUALITATIVE RESULTS

We conduct a thorough qualitative comparison with the current state-of-the-art methods to evaluate the performance of our model. Details regarding reproducibility are provided as follows:

- (1) CLIP-AC[5]. Unlike the original pre-trained CLIP model, which utilizes a single text prompt, CLIP-AC ensemble 80 different context prompts to tackle zero-shot anomaly detection tasks. All parameters are maintained consistent with those specified in their paper.
- (2) APRIL-GAN[1]. It not only employs additional linear layers to map image features into a joint embedding space but also carefully selects text prompts suitable for anomaly detection tasks, thereby enhancing its zero-shot detection capability. All parameters are maintained consistent with those specified in their paper.
- (3) AnomalyCLIP[6]. AnomalyCLIP stands as the state-of-the-art model in recent developments. It achieves precise recognition of various types of anomalies by innovatively learning an object-agnostic text prompt. All parameters are maintained consistent with those specified in their paper.

Figure 1, 2, 3 and 4 reports the visualization comparison between SimCLIP and other SOTA methods mentioned above on MVTec-AD and VisA benchmarks. We observe that CLIP-AC does not accurately localize anomalies within images. This limitation is attributed to the fact that during pre-training, CLIP primarily focuses on aligning high-level global semantic information extracted from images and text, leading to a notable gap between the original CLIP model and downstream anomaly segmentation tasks. While APRIL-GAN partially addresses the challenge of CLIP’s inability to precisely localize anomalies in images through the integration of ensemble prompts and linear mappings, unfortunately, this approach also results in misidentifying normal regions as anomalies. AnomalyCLIP incorporates object-agnostic textual prompts at the input side of the language branch to enhance its ability to identify anomalous

Table 1: Comparison with Object-agnostic Prompt Learning (OPL) on both computation and memory overhead.

| Method | Source | FLOPs(G) | Params(M) |
|-----------|-------------|----------|-----------|
| OPL | AnomalyCLIP | 520.35 | 428.99 |
| IPT(ours) | SimCLIP | 513.75 | 428.77 |

regions, but it still mistakenly identifies normal areas in images as anomalies due to the inherent bias of the language branch towards focusing on global semantic information. The qualitative results demonstrate that the segmentation produced by our SimCLIP aligns more closely with the ground truth compared to the segmentation produced by the state-of-the-art method AnomalyCLIP. This illustrates the effectiveness of the proposed SimCLIP, which achieves realignment of vision and language through bidirectional interaction adjustments within both branches.

3 IPT VS. OBJECT-AGNOSTIC PROMPT LEARNING

In this section, we conduct a comparative analysis between Implicit Prompt Tuning (IPT) and Object-agnostic Prompt Learning (OPL)[6], highlighting the advantages of IPT over OPL.

Firstly, IPT eliminates inherent bias in the language branch. The OPL focuses on optimizing learnable word embeddings at the input side of the language branch. Hence, it is more susceptible to the inherent bias of the language branch, which only concentrates on global semantics information extracted from text. This is evident in the resulting visualizations(e.g., Figure 3,4). However, IPT refines the original text features by combining a task-specific prompt embedding on the output side of the language branch. This strategy eliminates inherent biases on the language branch and enables a focus on local semantic information, facilitating accurate anomaly segmentation in images.

Secondly, efficient computation. Table 1 presents a comparison of computation and memory overhead between IPT and OPL. The results indicate that under similar parameter conditions, IPT requires significantly fewer FLOPs, demonstrating its superior computational efficiency.

Thirdly, stable optimization. OPL optimizes learnable word embedding to capture the generic normality and abnormality in an image. This process is similar to prefix-tuning, where directly updating the word embeddings can lead to unstable optimization[4]. While IPT also encounters similar challenges, we mitigate this issue by using a smaller matrix consisting of a large feedforward neural network (MLP) to improve optimization stability.

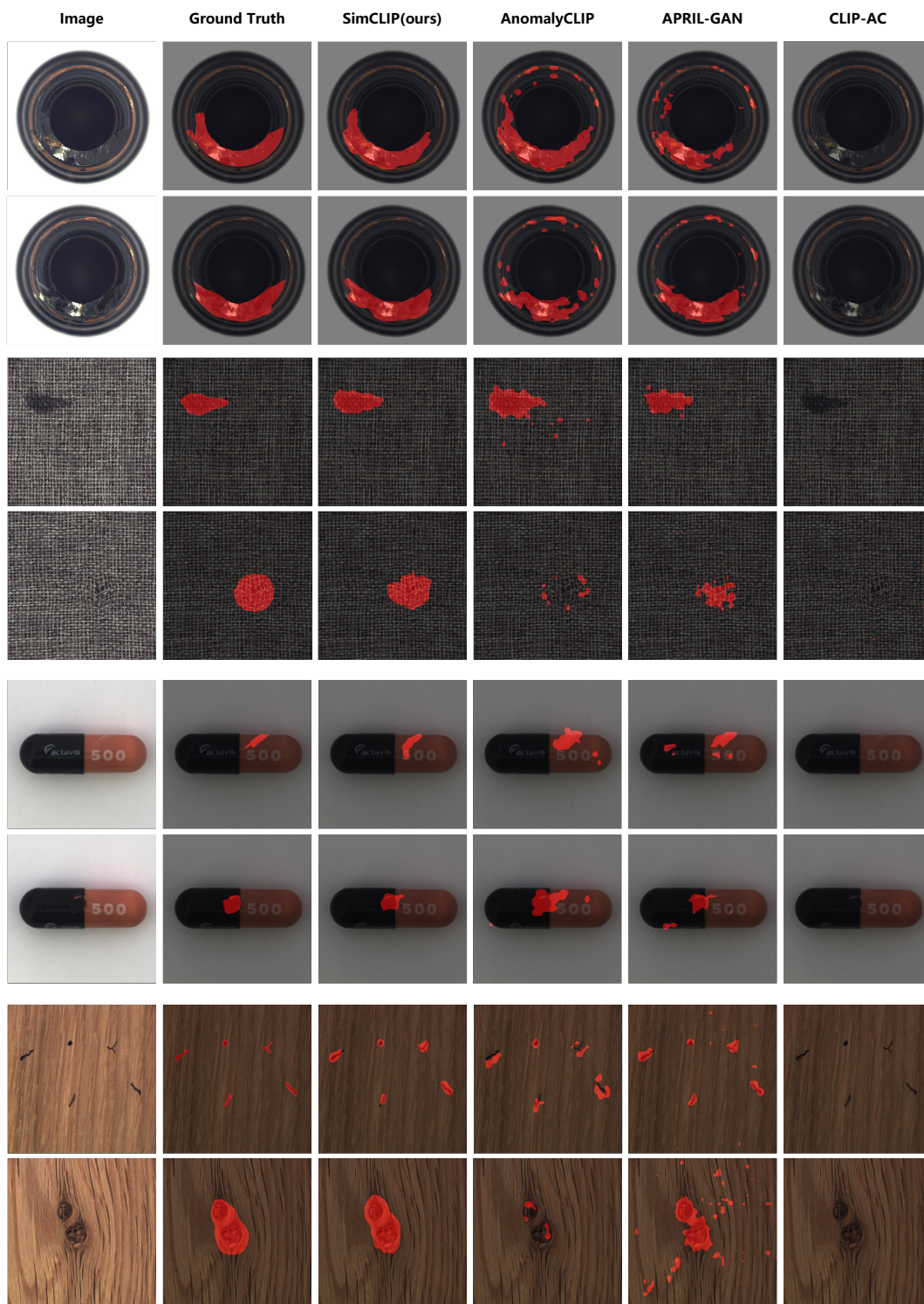
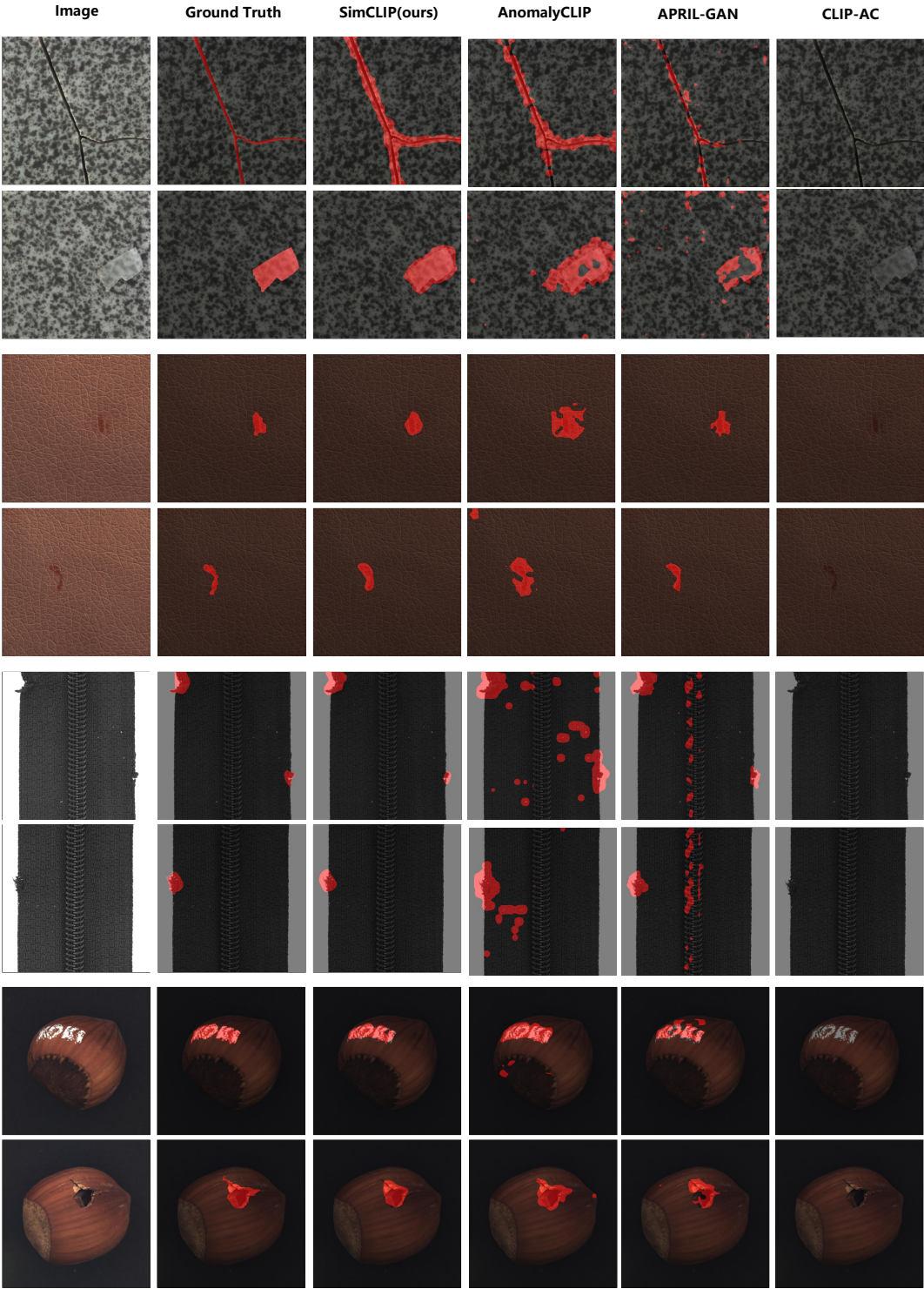


Figure 1: Qualitative comparison of zero-shot anomaly segmentation results on MVTec-AD benchmark.



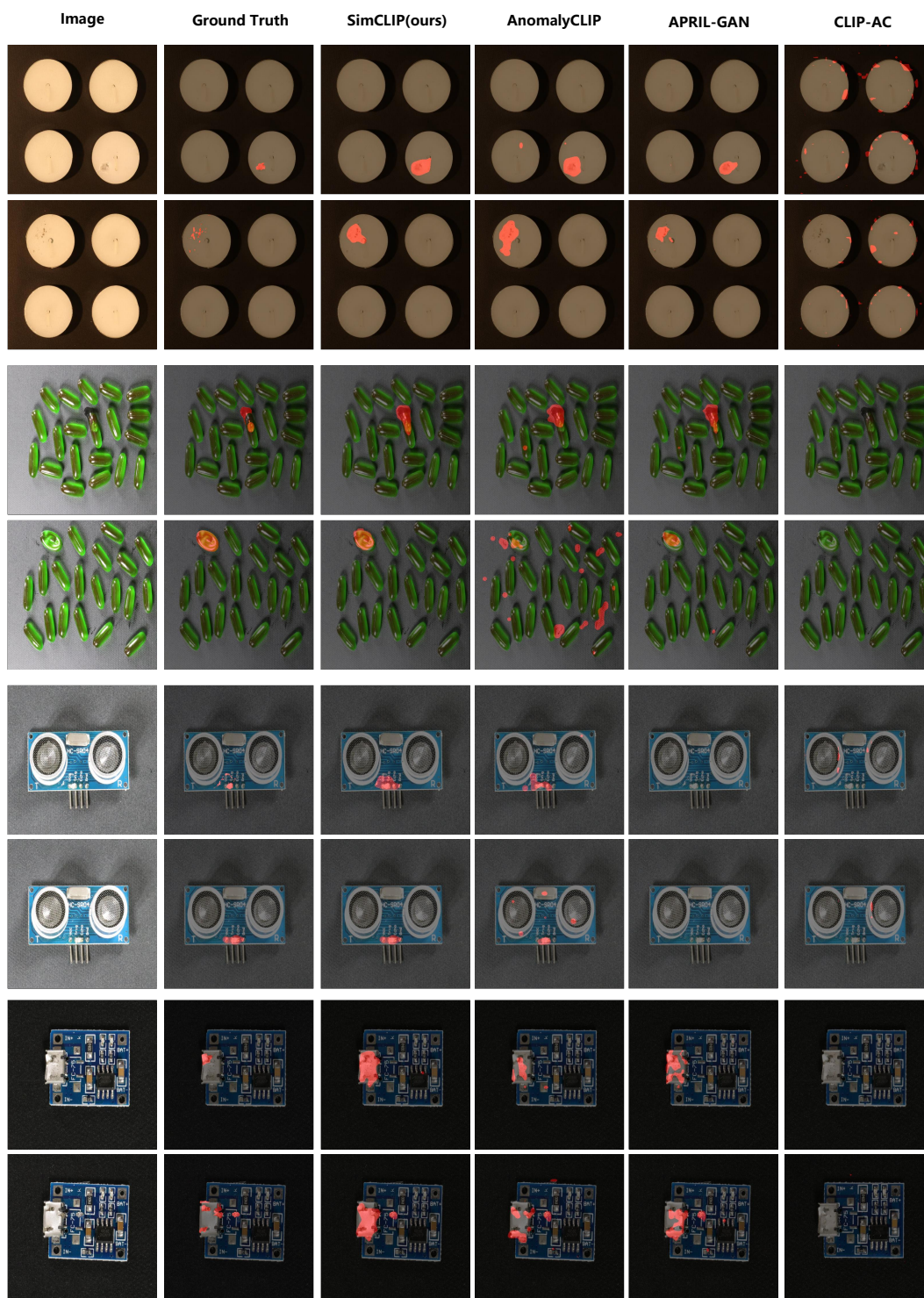


Figure 3: Qualitative comparison of zero-shot anomaly segmentation results on VisA benchmark.



Figure 4: Qualitative comparison of zero-shot anomaly segmentation results on VisA benchmark.

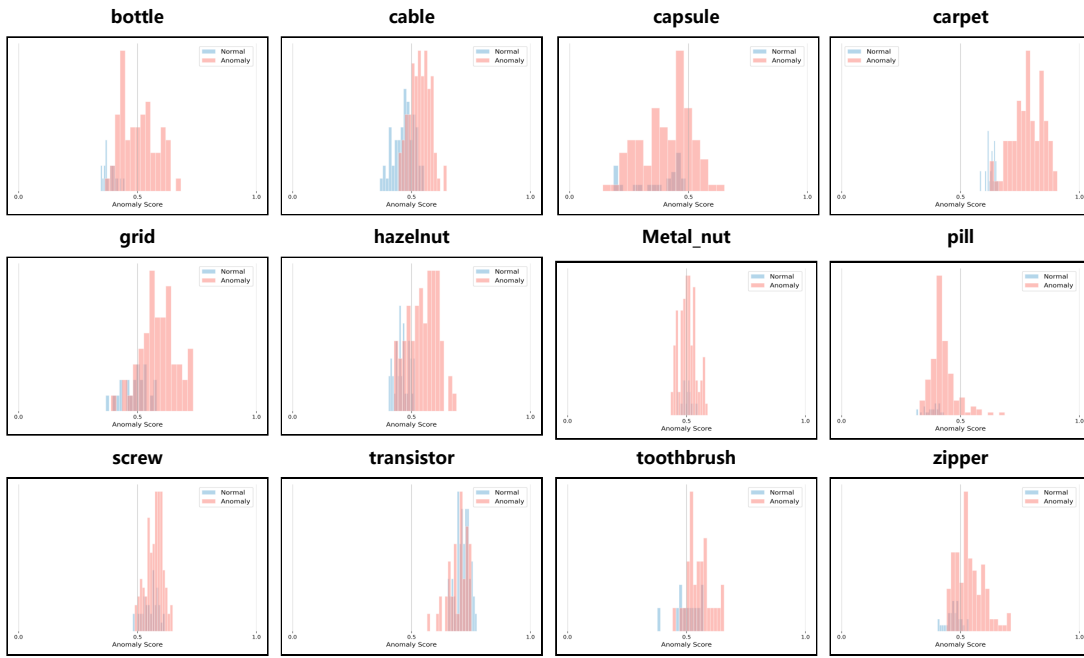


Figure 5: Statistical histograms of anomaly scores for each category obtained using the initial CLIP on the MVTec-AD benchmark.

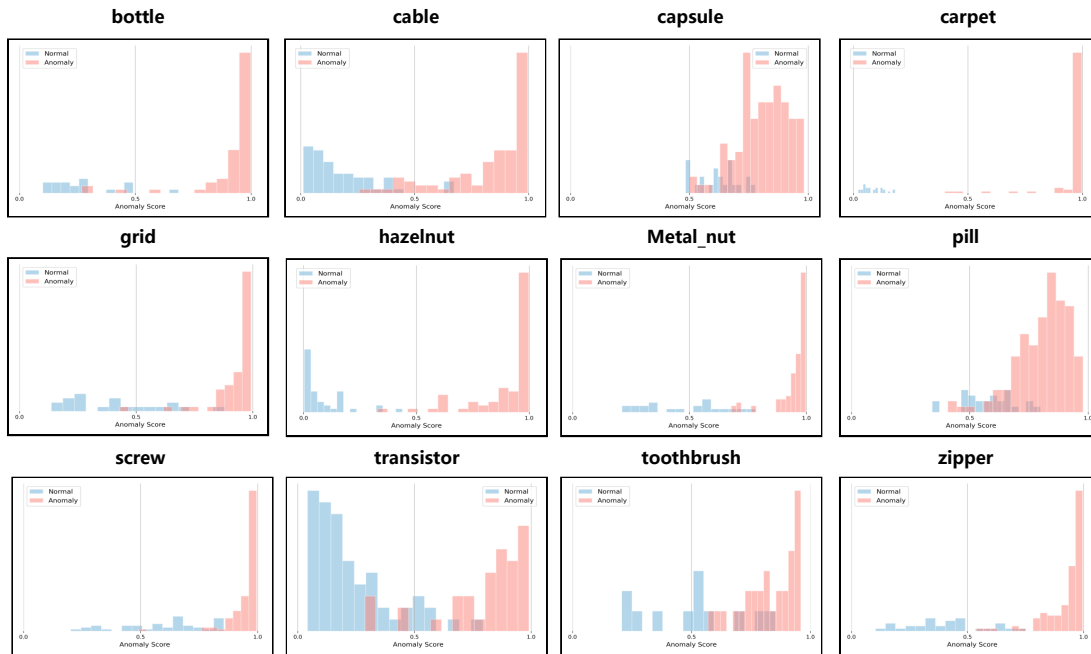


Figure 6: Statistical histograms of anomaly scores for each category obtained using the SimCLIP on the MVTec-AD benchmark.

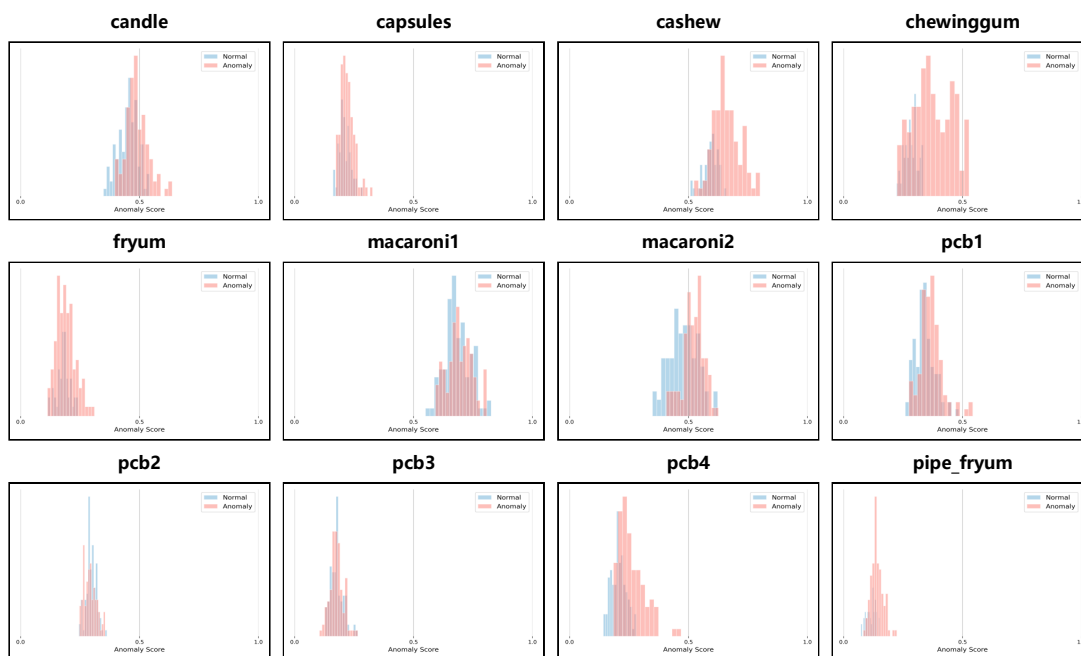


Figure 7: Statistical histograms of anomaly scores for each category obtained using the initial CLIP on the VisA benchmark.

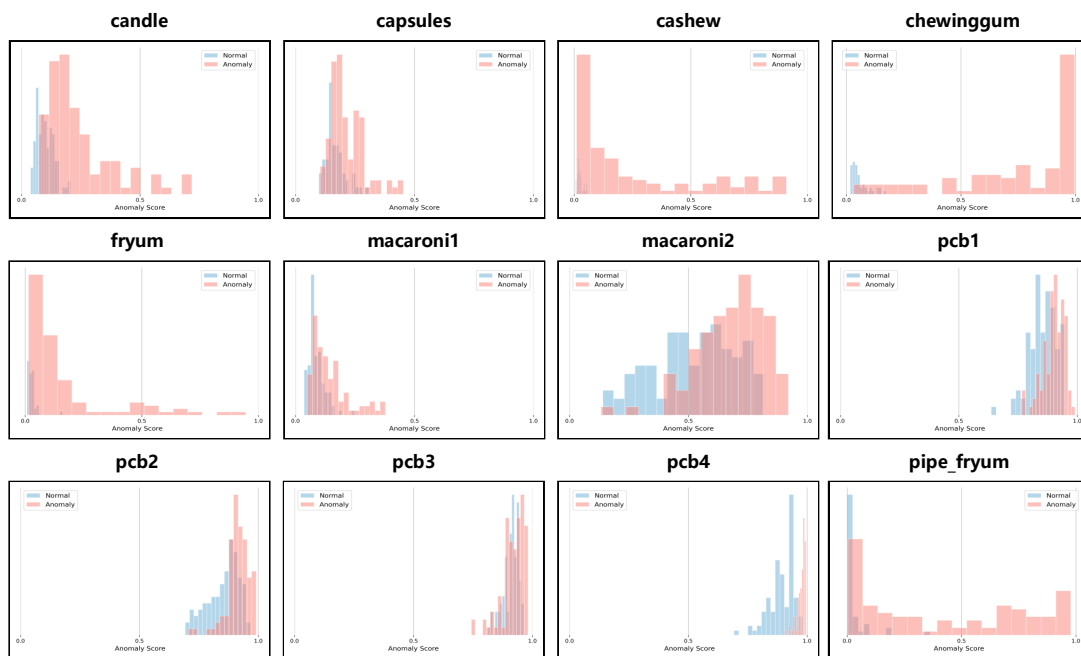


Figure 8: Statistical histograms of anomaly scores for each category obtained using the SimCLIP on the VisA benchmark.

Table 2: Quantitative results of the 0-shot setting on MVTec-AD benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|-------|--------|------------------------|-------|--------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| bottle | 93.2 | 87.2 | 60.6 | 93.3 | 98.1 | 93.0 |
| cable | 79.2 | 60.5 | 17.9 | 73.7 | 83.9 | 78.0 |
| capsule | 95.5 | 91.9 | 27.1 | 79.4 | 95.0 | 91.5 |
| carpet | 99.5 | 98.3 | 77.0 | 100.0 | 100.0 | 100.0 |
| grid | 97.4 | 91.9 | 41.8 | 92.4 | 96.9 | 91.8 |
| hazelnut | 97.3 | 90.1 | 54.3 | 93.6 | 96.4 | 90.7 |
| leather | 99.6 | 99.2 | 57.6 | 99.9 | 100.0 | 99.5 |
| metal_nut | 77.2 | 75.4 | 35.0 | 93.4 | 98.4 | 93.8 |
| pill | 86.2 | 94.4 | 35.1 | 84.4 | 96.4 | 92.5 |
| screw | 98.5 | 93.2 | 43.8 | 75.6 | 89.6 | 87.0 |
| tile | 95.2 | 93.0 | 72.9 | 100.0 | 100.0 | 100.0 |
| toothbrush | 93.6 | 89.3 | 35.8 | 85.6 | 94.2 | 88.2 |
| transistor | 70.4 | 53.0 | 17.2 | 82.6 | 81.5 | 74.4 |
| wood | 97.1 | 95.7 | 67.2 | 98.9 | 99.7 | 97.4 |
| zipper | 97.1 | 88.3 | 61.1 | 97.1 | 99.3 | 96.3 |
| Mean | 91.8 | 86.8 | 47.0 | 90.0 | 95.3 | 91.6 |

Table 3: Quantitative results of the 1-shot setting on MVTec-AD benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|------------|------------|------------------------|-------------|------------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| bottle | 97.4 ± 0.3 | 94.5 ± 0.4 | 73.4 ± 1.5 | 98.1 ± 0.2 | 99.4 ± 0.0 | 96.0 ± 0.1 |
| cable | 89.7 ± 0.2 | 84.9 ± 0.7 | 44.8 ± 0.5 | 78.1 ± 0.7 | 86.3 ± 1.2 | 79.8 ± 0.7 |
| capsule | 97.9 ± 0.1 | 96.9 ± 0.4 | 43.3 ± 4.3 | 93.7 ± 1.1 | 98.7 ± 0.2 | 94.3 ± 1.2 |
| carpet | 99.4 ± 0.0 | 98.1 ± 0.0 | 75.0 ± 0.2 | 100.0 ± 0.0 | 100.0 ± 0.0 | 99.4 ± 0.0 |
| grid | 98.1 ± 0.1 | 93.0 ± 0.5 | 47.9 ± 0.2 | 99.2 ± 0.2 | 99.7 ± 0.0 | 97.7 ± 0.4 |
| hazelnut | 99.0 ± 0.1 | 95.8 ± 1.1 | 68.7 ± 2.2 | 99.6 ± 0.2 | 99.8 ± 0.1 | 98.1 ± 0.7 |
| leather | 99.7 ± 0.0 | 99.2 ± 0.1 | 64.7 ± 0.4 | 100.0 ± 0.0 | 100.0 ± 0.0 | 99.5 ± 0.0 |
| metal_nut | 88.6 ± 0.6 | 89.3 ± 0.8 | 49.9 ± 1.5 | 96.3 ± 0.5 | 99.1 ± 0.1 | 96.5 ± 0.8 |
| pill | 95.0 ± 0.5 | 97.8 ± 0.2 | 56.5 ± 1.7 | 94.4 ± 0.3 | 98.9 ± 0.1 | 95.4 ± 0.3 |
| screw | 98.6 ± 0.1 | 93.8 ± 0.2 | 43.3 ± 2.5 | 83.7 ± 1.6 | 94.0 ± 0.7 | 87.9 ± 0.5 |
| tile | 96.5 ± 0.1 | 93.7 ± 0.1 | 72.2 ± 0.1 | 99.9 ± 0.0 | 100.0 ± 0.0 | 99.4 ± 0.0 |
| toothbrush | 98.4 ± 0.2 | 92.9 ± 1.3 | 60.0 ± 2.4 | 97.2 ± 2.6 | 98.9 ± 1.0 | 96.1 ± 3.3 |
| transistor | 79.9 ± 0.7 | 64.9 ± 1.0 | 35.3 ± 2.3 | 91.9 ± 3.2 | 90.4 ± 3.4 | 82.3 ± 4.8 |
| wood | 97.6 ± 0.0 | 96.9 ± 0.0 | 70.6 ± 0.3 | 99.1 ± 0.1 | 99.7 ± 0.0 | 97.2 ± 0.4 |
| zipper | 98.1 ± 0.1 | 93.8 ± 0.4 | 64.7 ± 1.1 | 98.9 ± 0.2 | 99.7 ± 0.1 | 98.0 ± 0.2 |
| Mean | 95.6 ± 0.2 | 92.4 ± 0.2 | 58.0 ± 0.5 | 95.3 ± 0.3 | 97.7 ± 0.3 | 94.5 ± 0.5 |

Table 4: Quantitative results of the 2-shot setting on MVTec-AD benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|------------|------------|------------------------|-------------|------------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| bottle | 97.6 ± 0.1 | 94.3 ± 0.5 | 73.6 ± 0.6 | 98.3 ± 0.2 | 99.5 ± 0.0 | 96.1 ± 0.1 |
| cable | 90.5 ± 0.4 | 86.1 ± 1.2 | 46.7 ± 1.3 | 81.2 ± 2.2 | 89.0 ± 1.3 | 80.2 ± 1.0 |
| capsule | 98.2 ± 0.0 | 97.4 ± 0.1 | 45.7 ± 1.1 | 95.3 ± 0.3 | 99.1 ± 0.0 | 95.5 ± 0.4 |
| carpet | 99.5 ± 0.0 | 98.0 ± 0.1 | 75.5 ± 0.4 | 99.9 ± 0.0 | 100.0 ± 0.0 | 99.4 ± 0.0 |
| grid | 98.4 ± 0.2 | 94.8 ± 0.2 | 48.8 ± 0.6 | 99.4 ± 0.1 | 99.8 ± 0.0 | 98.0 ± 0.4 |
| hazelnut | 99.2 ± 0.1 | 96.4 ± 0.8 | 72.7 ± 1.5 | 99.7 ± 0.1 | 99.9 ± 0.0 | 99.1 ± 0.3 |
| leather | 99.7 ± 0.0 | 99.1 ± 0.1 | 63.9 ± 0.4 | 100.0 ± 0.0 | 100.0 ± 0.0 | 99.5 ± 0.0 |
| metal_nut | 90.3 ± 0.2 | 90.6 ± 0.1 | 53.8 ± 0.5 | 96.8 ± 0.5 | 99.2 ± 0.1 | 96.7 ± 0.8 |
| pill | 95.5 ± 0.1 | 97.9 ± 0.1 | 58.5 ± 0.2 | 94.3 ± 0.2 | 98.8 ± 0.0 | 95.5 ± 0.2 |
| screw | 98.5 ± 0.2 | 93.8 ± 0.9 | 40.7 ± 5.0 | 85.5 ± 3.7 | 94.5 ± 1.8 | 89.6 ± 1.1 |
| tile | 96.8 ± 0.0 | 93.9 ± 0.2 | 72.7 ± 0.1 | 99.9 ± 0.0 | 100.0 ± 0.0 | 99.6 ± 0.3 |
| toothbrush | 98.5 ± 0.0 | 94.0 ± 0.9 | 59.4 ± 0.6 | 98.6 ± 0.8 | 99.4 ± 0.3 | 97.9 ± 0.8 |
| transistor | 80.8 ± 0.0 | 65.8 ± 0.6 | 36.6 ± 1.0 | 93.1 ± 1.5 | 92.5 ± 1.5 | 85.7 ± 3.5 |
| wood | 97.6 ± 0.2 | 96.9 ± 0.1 | 69.8 ± 1.0 | 99.1 ± 0.0 | 99.7 ± 0.0 | 97.5 ± 0.0 |
| zipper | 98.3 ± 0.1 | 94.3 ± 0.4 | 65.5 ± 0.9 | 99.0 ± 0.2 | 99.7 ± 0.0 | 98.0 ± 0.4 |
| Mean | 96.0 ± 0.2 | 92.9 ± 0.1 | 58.9 ± 0.1 | 96.0 ± 0.2 | 98.1 ± 0.1 | 95.2 ± 0.1 |

Table 5: Quantitative results of the 4-shot setting on MVTec-AD benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|------------|------------|------------------------|-------------|------------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| bottle | 97.8 ± 0.1 | 94.8 ± 0.6 | 74.5 ± 1.0 | 98.5 ± 0.2 | 99.5 ± 0.0 | 96.3 ± 0.4 |
| cable | 91.0 ± 0.4 | 86.9 ± 0.4 | 48.0 ± 1.5 | 82.7 ± 1.6 | 89.5 ± 1.4 | 81.6 ± 1.2 |
| capsule | 98.3 ± 0.1 | 97.2 ± 0.7 | 44.5 ± 5.3 | 94.9 ± 1.5 | 98.9 ± 0.3 | 95.1 ± 1.1 |
| carpet | 99.5 ± 0.0 | 97.9 ± 0.2 | 75.4 ± 0.1 | 100.0 ± 0.0 | 100.0 ± 0.0 | 99.4 ± 0.0 |
| grid | 98.5 ± 0.2 | 94.6 ± 0.1 | 49.1 ± 1.0 | 99.4 ± 0.2 | 99.8 ± 0.1 | 97.6 ± 0.4 |
| hazelnut | 99.4 ± 0.0 | 96.7 ± 0.4 | 75.1 ± 1.8 | 99.8 ± 0.0 | 99.9 ± 0.0 | 99.1 ± 0.3 |
| leather | 99.7 ± 0.0 | 98.7 ± 0.2 | 62.2 ± 0.4 | 100.0 ± 0.0 | 100.0 ± 0.0 | 99.5 ± 0.0 |
| metal_nut | 90.4 ± 0.2 | 90.8 ± 0.1 | 54.8 ± 0.4 | 97.8 ± 0.5 | 99.4 ± 0.1 | 97.4 ± 0.4 |
| pill | 95.3 ± 0.2 | 97.9 ± 0.0 | 57.6 ± 1.2 | 94.7 ± 0.2 | 99.0 ± 0.1 | 95.4 ± 0.0 |
| screw | 98.9 ± 0.1 | 95.5 ± 0.4 | 44.9 ± 4.3 | 87.6 ± 2.0 | 95.5 ± 1.0 | 89.6 ± 0.7 |
| tile | 96.7 ± 0.3 | 93.5 ± 0.8 | 72.0 ± 0.9 | 100.0 ± 0.0 | 100.0 ± 0.0 | 99.4 ± 0.0 |
| toothbrush | 98.8 ± 0.3 | 94.8 ± 0.8 | 62.1 ± 2.5 | 98.3 ± 1.4 | 99.4 ± 0.5 | 96.7 ± 2.4 |
| transistor | 81.9 ± 0.5 | 65.8 ± 0.7 | 36.4 ± 0.7 | 93.7 ± 0.4 | 92.5 ± 0.5 | 86.7 ± 0.8 |
| wood | 97.7 ± 0.0 | 96.9 ± 0.0 | 70.3 ± 0.3 | 99.1 ± 0.0 | 99.7 ± 0.0 | 97.5 ± 0.0 |
| zipper | 98.4 ± 0.1 | 94.6 ± 0.5 | 66.4 ± 1.3 | 99.7 ± 0.1 | 99.9 ± 0.0 | 99.2 ± 0.0 |
| Mean | 96.2 ± 0.1 | 93.1 ± 0.1 | 59.6 ± 0.2 | 96.4 ± 0.2 | 98.0 ± 0.2 | 95.4 ± 0.2 |

Table 6: Quantitative results of the 0-shot setting on VisA benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|-------|--------|------------------------|------|--------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| candle | 98.7 | 95.5 | 40.8 | 91.0 | 92.0 | 83.6 |
| capsules | 96.5 | 89.4 | 43.1 | 74.8 | 82.8 | 80.6 |
| cashew | 92.5 | 95.3 | 29.2 | 86.4 | 94.4 | 86.3 |
| chewinggum | 99.5 | 91.3 | 78.0 | 98.6 | 99.4 | 96.9 |
| fryum | 93.8 | 92.0 | 30.6 | 89.8 | 94.8 | 88.4 |
| macaroni1 | 99.2 | 96.1 | 31.8 | 75.4 | 77.1 | 72.4 |
| macaroni2 | 98.2 | 89.0 | 6.20 | 75.2 | 75.7 | 73.8 |
| pcb1 | 93.8 | 88.9 | 12.3 | 75.0 | 75.3 | 73.6 |
| pcb2 | 92.2 | 77.1 | 22.1 | 78.8 | 78.6 | 75.0 |
| pcb3 | 90.6 | 77.7 | 21.8 | 62.4 | 69.5 | 66.40 |
| pcb4 | 95.2 | 86.9 | 29.1 | 96.2 | 96.2 | 90.2 |
| pipe_fryum | 97.0 | 97.0 | 40.6 | 93.0 | 96.4 | 91.0 |
| Mean | 95.6 | 89.7 | 32.1 | 83.1 | 86.0 | 81.5 |

Table 7: Quantitative results of the 1-shot setting on VisA benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|------------|------------|------------------------|------------|------------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| candle | 99.0 ± 0.2 | 97.8 ± 0.3 | 38.9 ± 4.0 | 93.4 ± 1.1 | 94.3 ± 0.9 | 86.8 ± 0.5 |
| capsules | 98.2 ± 0.3 | 91.4 ± 0.9 | 47.6 ± 1.3 | 95.9 ± 0.1 | 97.7 ± 0.0 | 91.6 ± 0.5 |
| cashew | 95.3 ± 0.3 | 96.3 ± 0.4 | 50.7 ± 0.6 | 95.3 ± 0.5 | 97.9 ± 0.1 | 92.3 ± 0.6 |
| chewinggum | 99.7 ± 0.0 | 95.8 ± 0.1 | 72.1 ± 0.6 | 98.8 ± 0.1 | 99.6 ± 0.0 | 98.2 ± 0.2 |
| fryum | 96.4 ± 0.1 | 93.8 ± 0.3 | 42.3 ± 0.4 | 97.3 ± 0.3 | 98.6 ± 0.1 | 95.5 ± 0.4 |
| macaroni1 | 99.6 ± 0.0 | 97.0 ± 0.4 | 22.6 ± 3.3 | 93.0 ± 1.3 | 94.4 ± 0.8 | 86.8 ± 1.4 |
| macaroni2 | 97.9 ± 0.8 | 90.2 ± 1.9 | 18.7 ± 0.6 | 86.3 ± 1.7 | 88.6 ± 1.2 | 78.4 ± 1.6 |
| pcb1 | 97.9 ± 0.1 | 93.5 ± 0.4 | 41.6 ± 0.8 | 96.2 ± 0.2 | 96.4 ± 0.2 | 90.1 ± 0.3 |
| pcb2 | 96.3 ± 0.4 | 84.7 ± 1.1 | 33.4 ± 1.1 | 85.3 ± 1.4 | 87.7 ± 1.7 | 77.9 ± 1.4 |
| pcb3 | 94.8 ± 0.1 | 87.1 ± 1.5 | 41.5 ± 0.5 | 79.0 ± 8.3 | 81.3 ± 8.2 | 74.9 ± 4.2 |
| pcb4 | 96.3 ± 0.2 | 87.2 ± 1.1 | 29.8 ± 2.0 | 97.0 ± 3.2 | 97.5 ± 2.3 | 92.2 ± 5.3 |
| pipe_fryum | 97.9 ± 0.2 | 97.5 ± 0.2 | 45.5 ± 1.1 | 98.8 ± 0.1 | 99.4 ± 0.1 | 98.0 ± 0.7 |
| Mean | 97.4 ± 0.1 | 92.7 ± 0.2 | 40.4 ± 0.5 | 93.0 ± 1.1 | 94.5 ± 0.9 | 88.5 ± 0.7 |

Table 8: Quantitative results of the 2-shot setting on VisA benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|------------|------------|------------------------|------------|------------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| candle | 99.1 ± 0.1 | 97.8 ± 0.1 | 39.0 ± 3.3 | 92.7 ± 1.5 | 93.7 ± 1.1 | 86.0 ± 1.0 |
| capsules | 98.5 ± 0.1 | 90.1 ± 0.2 | 47.4 ± 2.4 | 96.0 ± 0.4 | 97.8 ± 0.2 | 91.9 ± 0.8 |
| cashew | 95.3 ± 0.1 | 95.2 ± 0.6 | 53.9 ± 0.4 | 95.3 ± 0.6 | 97.8 ± 0.2 | 91.9 ± 0.6 |
| chewinggum | 99.7 ± 0.0 | 95.7 ± 0.2 | 73.1 ± 1.3 | 99.0 ± 0.1 | 99.6 ± 0.0 | 98.1 ± 0.5 |
| fryum | 96.6 ± 0.4 | 93.7 ± 0.3 | 44.2 ± 0.2 | 97.6 ± 0.4 | 98.8 ± 0.2 | 95.6 ± 0.7 |
| macaroni1 | 99.6 ± 0.0 | 97.4 ± 0.2 | 27.4 ± 3.5 | 94.0 ± 1.3 | 95.0 ± 1.0 | 86.8 ± 1.4 |
| macaroni2 | 98.1 ± 0.3 | 91.4 ± 0.9 | 21.9 ± 1.3 | 85.7 ± 1.1 | 88.1 ± 1.1 | 77.5 ± 0.7 |
| pcb1 | 98.4 ± 0.4 | 94.3 ± 0.4 | 49.0 ± 8.2 | 95.7 ± 1.6 | 96.0 ± 1.3 | 91.1 ± 1.8 |
| pcb2 | 96.8 ± 0.1 | 86.6 ± 0.6 | 33.4 ± 1.6 | 86.1 ± 0.9 | 88.5 ± 0.7 | 78.2 ± 1.3 |
| pcb3 | 95.4 ± 0.2 | 89.9 ± 0.5 | 47.7 ± 0.7 | 85.4 ± 1.7 | 87.6 ± 1.3 | 79.1 ± 2.0 |
| pcb4 | 97.0 ± 0.1 | 91.0 ± 0.5 | 33.5 ± 2.7 | 98.4 ± 1.1 | 98.4 ± 1.1 | 94.3 ± 2.3 |
| pipe_fryum | 98.1 ± 0.0 | 97.5 ± 0.0 | 48.3 ± 1.2 | 98.9 ± 0.2 | 99.4 ± 0.1 | 98.3 ± 0.2 |
| Mean | 97.7 ± 0.1 | 93.4 ± 0.0 | 43.2 ± 0.8 | 93.7 ± 0.2 | 94.9 ± 0.2 | 89.1 ± 0.2 |

Table 9: Quantitative results of the 4-shot setting on VisA benchmark.

| Object | Anomaly Segmentation | | | Anomaly Classification | | |
|------------|----------------------|------------|------------|------------------------|------------|------------|
| | AUROC | AUPRO | F1-max | AUROC | AP | F1-max |
| candle | 99.3 ± 0.0 | 98.2 ± 0.0 | 40.7 ± 2.8 | 94.1 ± 1.5 | 94.8 ± 1.0 | 87.9 ± 1.5 |
| capsules | 98.9 ± 0.0 | 92.9 ± 0.7 | 50.6 ± 0.4 | 96.6 ± 0.4 | 98.2 ± 0.2 | 93.7 ± 0.6 |
| cashew | 95.2 ± 0.2 | 95.9 ± 0.4 | 54.6 ± 1.5 | 95.5 ± 0.7 | 97.8 ± 0.4 | 92.3 ± 0.9 |
| chewinggum | 99.7 ± 0.0 | 95.8 ± 0.4 | 70.7 ± 0.4 | 99.1 ± 0.0 | 99.6 ± 0.0 | 98.3 ± 0.2 |
| fryum | 96.7 ± 0.1 | 94.0 ± 0.1 | 45.5 ± 0.8 | 97.6 ± 0.4 | 98.8 ± 0.2 | 95.4 ± 0.4 |
| macaroni1 | 99.6 ± 0.0 | 97.6 ± 0.1 | 26.8 ± 2.8 | 94.2 ± 0.5 | 95.2 ± 0.4 | 88.0 ± 1.4 |
| macaroni2 | 98.2 ± 0.2 | 91.5 ± 0.7 | 20.4 ± 0.8 | 87.4 ± 0.7 | 89.4 ± 0.5 | 79.2 ± 0.9 |
| pcb1 | 98.8 ± 0.4 | 94.9 ± 0.8 | 58.8 ± 9.1 | 95.3 ± 1.0 | 96.0 ± 0.6 | 90.5 ± 1.3 |
| pcb2 | 97.1 ± 0.1 | 87.8 ± 0.5 | 35.7 ± 0.0 | 88.4 ± 0.2 | 90.3 ± 0.2 | 82.0 ± 0.4 |
| pcb3 | 95.9 ± 0.1 | 91.4 ± 0.7 | 49.8 ± 1.2 | 86.9 ± 1.2 | 88.7 ± 1.5 | 81.6 ± 2.4 |
| pcb4 | 97.3 ± 0.1 | 91.8 ± 0.6 | 31.9 ± 0.6 | 99.3 ± 0.1 | 99.2 ± 0.1 | 96.9 ± 0.2 |
| pipe_fryum | 98.3 ± 0.0 | 97.6 ± 0.3 | 49.5 ± 1.0 | 99.0 ± 0.2 | 99.4 ± 0.2 | 97.8 ± 0.2 |
| Mean | 98.0 ± 0.2 | 94.1 ± 0.1 | 44.6 ± 1.1 | 94.4 ± 0.1 | 95.6 ± 0.1 | 90.3 ± 0.2 |

REFERENCES

[1] 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).

[2] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, Yunsheng Wu, and Yong Liu. 2023. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. *arXiv preprint arXiv:2311.00453* (2023).

[3] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. 2024. Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9.

[4] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4582–4597. <https://doi.org/10.18653/v1/2021.acl-long.353>

[5] 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.

[6] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. 2024. AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. In *The Twelfth International Conference on Learning Representations*.