Supplemental Material for PatchEAD: Unifying Industrial Visual Prompting Frameworks for Patch-Exclusive Anomaly Detection

This supplementary material provides additional details to support the findings of our paper on training-free industrial anomaly detection with a patch-exclusive approach. § 1 includes information on the datasets used, detailing the types of multimodal data, preprocessing steps, and anomaly labeling. § 2 covers the architecture, feature extraction methods, and experimental setup, including hardware, software, and hyperparameters. § 3 introduce a information theory to analyze the information relevant to the target task(anomaly detection) and the usage under mutli-modality scenario and show some experiments on text prompts compare to vision prompts. § 4 demonstrate the detail of our alignment and masking strategy in PatchEAD$^+$. § 5 present the evaluation process, including metrics and baseline comparisons. § 6 provides visualizations that demonstrate the effectiveness of our approach in detecting anomalies. Finally, § 7 presents a case study to showcase the practical application of our method in a real-world industrial scenario, highlighting its performance and potential challenges.

1. Dataset Details

The experiments described in the main paper utilize seven widely recognized industrial anomaly detection datasets: MVTec [2], VisA [13], MPDD [5], BTAD [7], KSDD [10], DAGM [12], and DTD-Synthetic [1]. For datasets that provide a predefined training and testing split, we exclusively use the test data to evaluate the model's performance. In cases where no such split is provided, we allocate 25% of the normal images and all the anomalous images as the test set. For the KSDD dataset, each image is divided into three parts to better align its aspect ratio with the other datasets. Images containing a defect are labeled as anomalous, while those without a defect are considered normal. A detailed summary of the datasets, including key characteristics and splits, is presented in Table 1.

Dataset	Type	# of Categiries	Samples Normal	Samples Abnormal
MVTec [2]	Obj & Texture	15	467	1,258
VisA [13]	Obj	12	962	1,200
MPDD [5]	Obj	6	176	282
BTAD [7]	Obj	3	451	290
KSDD [10]	Texture	1	286	54
DAGM [12]	Texture	10	6,996	1,054
DTD-Synthetic [1]	Texture	12	357	947

Table 1. Anomaly Detection Datasets in Industrial Domain.

2. Implementation Details

Across all experiments, we adopt base-size backbones to keep parameter counts comparable. Our CLIP variant uses `ViT-B-16-plus-240` pretrained on `laion400m_e32` from OpenCLIP [4], with inputs resized to 240×240 to match the checkpoint. The DINOv2 variant employs `dinov2_vitb14_reg` from the official codebase [8] at its default 448×448 resolution. For PE_spatial, we load `PE-Spatial-B16-512`; for DINOv3, we use the distilled `ViT-B/16`. These settings keep each method close to its standard configuration and enable a fair comparison across architectures.

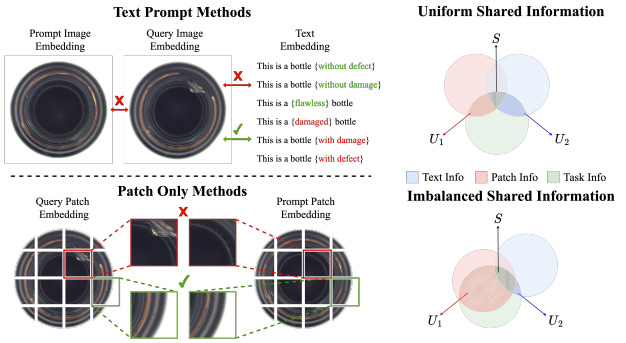
3. Multimodal Information Analysis

Following the formulation in [6] for handling multimodal information, we consider two input modalities, $X_i (i \in \{1, 2\})$, annotated with a label variable Y. The joint distribution of these inputs and Y can be expressed using conditional mutual information:

$$I(X_1, X_2, Y) = S + U_1 + U_2, \tag{1}$$

where $S = I(X_1; X_2; Y)$ denotes the task-relevant shared information; $U_1 = I(X_1; Y|X_2)$ and $U_2 = I(X_2; Y|X_1)$ denote the unique information of modalities X_1 and X_2, respectively. Figure 1 provides a concept plot of the imbalanced shared *v.s.* unique information provided in different modalities. Our hypothesis is that patch-level information should be more important over the text information for industrial defect detection. Our aim is to uncover the imbalanced distribution of U_1 and U_2 across different modalities of texts, patches, and whole images.

Loss functions in large multi-modality models, such as CLIP [9], are often optimized to capture shared information across modalities. Consequently, other anomaly detection



(a) Overview of Patch vs. Text (b) Assumption

Figure 1. (a) Existing anomaly detection methods using image–text similarity often assume clean separation via simple positive/negative prompts. Our **patch-only** method detects anomalies by comparing query patches with normal prompt patches, leveraging richer context. (b) Considering shared information S and modality-unique information U_1, U_2, task relevance varies; we argue patch-level information outweighs text information for industrial defect detection.

methods tend to add additional modules to enhance information. However, these additional modules may introduce overlapping information between modalities. Our focus is to examine the unique information provided by each modality (U_1 and U_2).

The unique information (U_1 and U_2) can be challenging to compute directly, so we approximate it using an upper-bound, I_{CLUB}, as proposed in contrastive learning [3]. When I_{CLUB} is small, the unique information in a modality is expected to be limited. For instance, we can estimate the unique information U_1 as follows:

$$U_1 = I(X_1; Y|X_2) \leq I_{CLUB}(X_1; Y|X_2), \tag{2}$$

where I_{CLUB} is given by:

$$I_{CLUB}(X_1; Y|X_2) = E[f^*(X_1, Y^+|X_2)] - E[f^*(X_1, Y^-|X_2)], \tag{3}$$

with f^* as the optimal critic from a contrastive pre-trained network. I_{CLUB} represents the expected difference between positive and negative samples, X^+ and X^- for a given X. By applying an optimal augmentation assumption [11], we replace Y with an augmented variable X' of X, assuming that only task-relevant information is shared between X and X'. I_{CLUB} can then be calculated via this adjustment as:

$$I_{CLUB}(X_1; Y|X_2) = E[f^*(X_1, X_1'^+|X_2)] - E[f^*(X_1, X'^-|X_2)]. \tag{4}$$

If the expected anomaly scores between positive and negative samples are not significantly different, this upper bound will be small, suggesting limited unique information. Minimal differences in anomaly scores may also result if X_2 provides sufficient information, thus reducing the reliance on $X_1'^+$ and $X_1'^-$.

Considering X_1 for the patch modality and X_2 as the text modality, we use Eq. (4) to derive the upper bounds for U_1 and U_2. The upper bound for U_2, $I_{CLUB}(X_2; Y|X_1)$, is expected to be small because augmenting text prompts results in minimal changes in anomaly scores once patches are already included. On the other hand, the upper bound for U_1, $I_{CLUB}(X_1; Y|X_2)$, can be large, as augmentations like rotation, cropping, or noise perturbations cause noticeable variations in the final anomaly scores. This larger upper bound for U_1 arises because the text prompts often fail to capture details about missing objects or defects, which are better represented in the patch modality.

3.1. Comparison of Text *vs.* Patch Information

Figure 2(a,b) compares the **normal *vs.* abnormal text prompts** on the MVTec dataset. These plots explore how different prompts affect the image-text similarity score histograms for normal and abnormal samples. The information upper bound of each modality, as described in Eq. (4), is estimated based on the expected score difference between positive and negative augmentations. By comparing the distributions of scores for positive and negative prompts in panels (a) and (b), we illustrate how the text information upper bound impacts the anomaly detection performance.

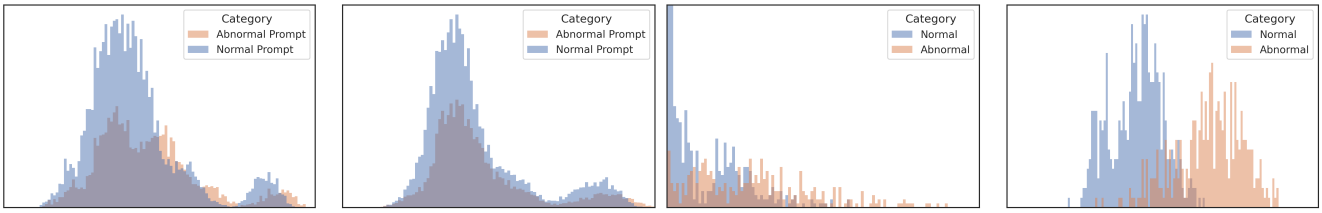
Figure 2(c,d) compares the **text *vs.* patch modalities** by using the histogram of anomaly scores on the MVTec dataset. (c) plots the anomaly score histogram using WinCLIP, a model featuring vision-text multimodal capability. The scores were derived from the similarity between the image and text prompt tokens. (d) plots the histogram using our PatchEAD, where only patch information is used. Observe that (d) exhibits a more separable space than (c), suggesting more discriminative results with only patch information.

Results show that the positive and negative prompts yield a highly overlapped distribution regardless of normal or abnormal images. The upper bound of text information is limited, leading to inferior discriminative power than patches.

4. Details of Alignment and Masking

We introduce both alignment and masking to MVTec, VisA, KSDD, DAGM, and DTD-Synthetic datasets, and only masking applied on MPDD, and BTAD due to the object is well-aligned(BTAD) or 3D rotation(MPDD).

Alignment: In the few-shot setting, we calculate the displacement and rotation between each prompt image and the query image. We then add the corrected images and their 180-degree rotations to the prompt list to ensure consistent



(a) Normal images (b) Abnormal images (c) Text prompts (d) Patch prompts

Figure 2. **Histogram plots** summarizing experimental results on the MVTec dataset. **(a)** and **(b)** depict the **image–text similarity scores** for normal and abnormal images under various text prompts. Ideally, normal images should exhibit high similarity with their corresponding normal prompts and lower similarity when paired with abnormal prompts. **(c)** and **(d)** show the **anomaly scores** computed using text and patch prompts, respectively.

positions and angles during patch matching. In the zero-shot setting, we use the first image as the query and apply the same process to all other images. For similarity calculation, we average the scores from the three images for each sample to determine the final anomaly score.

Masking: We extract the final attention layer and average all attention heads to get the [CLS] token's attention scores for each patch. After normalizing, patches with scores above 0.05 are set to 1, and those below to 0.5. This reduces the incorrect activations in background areas.

5. Experimental Details

Our experiments entail few-shot and zero-shot settings with various datasets and thus only present the average results in the main manuscript under the page limitation. We report the few-shot results with each category in the MVTec and ViSA datasets in Table 2 and Table 3. The zero-shot results are reported in Table 4 using seven datasets.

When we reveal the category-wise AUC in the few-shot setting, we observe the capsule, screw, and transistor in the MVTec dataset are harder categories with lower image-level and pixel-level AUC. Using DINOv2 significantly improves PatchEAD in these categories compared to CLIP, highlighting the benefits of vision purpose pretraining for the patch-only method, even when performed in an unsupervised manner. This improvement can consistently be observed in 1-, 2-, and 4-shot results, especially on image-level AUC. PatchEAD$^+$ strengthening alignment and masking mainly facilitates the screw category in different few-shot settings.

Similarly, the few-shot image-level AUC attains relatively significant improvements on specific categories such as capsules, fryum, macaroni1, and pcb2 when comparing DINOv2 and CLIP in the VisA dataset. The improvement using PatchEAD$^+$ is eminent in the fryum, macaroni1, and pcb2 categories on both image-level and pixel-level AUC over different few-shot settings. The enhanced categories are generally aligned between few-shot and zero-shot settings. Our proposed PatchEAD using only patch information complements these hard cases with delicate visual appearances.

6. Visualization Results

Additionally, to demonstrate the defect localization capability of our proposed PatchEAD framework, we conducted visualization on all test datasets. Specifically, we linearly interpolated the predicted patch anomaly scores to match the size of the original images, followed by normalization, resulting in a heatmap for each image. In the few-shot setting, we used the MVTec and VisA datasets for visualization, with results shown in Figures 3 and 4. In the zero-shot setting, we conducted experiments using seven datasets, with results presented in Figures 5, 6, 7, 8, 9, 10, 11. These test datasets encompass a wide variety of object types, and the visualization results indicate that our method, even without training, can accurately localize various defects under different settings, demonstrating the framework's strong flexibility and generalization capability.

7. Case Study

We conducted a visual comparison to evaluate the effectiveness of alignment and masking in our method. As shown in Figure 12, masking reduces the weight of background regions, thereby decreasing the likelihood of misclassification. Meanwhile, alignment addresses variations in object orientation and displacement by aligning objects, enabling the comparison of similar contours and significantly reducing misclassification at object edges.

Category	Image-Level AUC				Pixel-Level AUC			
	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B
1-shot								
bottle	$98.2_{\pm 0.7}$	$98.8_{\pm 1.6}$	$98.5_{\pm 0.7}$	$99.7_{\pm 0.4}$	$97.9_{\pm 0.1}$	$98.3_{\pm 0.1}$	$98.0_{\pm 0.1}$	$98.1_{\pm 0.1}$
cable	$90.4_{\pm 2.0}$	$90.8_{\pm 1.9}$	$88.8_{\pm 2.0}$	$92.5_{\pm 1.1}$	$93.9_{\pm 0.4}$	$93.2_{\pm 0.5}$	$93.9_{\pm 0.3}$	$93.4_{\pm 0.4}$
capsule	$80.8_{\pm 12.0}$	$82.4_{\pm 13.3}$	$78.2_{\pm 11.8}$	$85.9_{\pm 11.2}$	$90.6_{\pm 0.7}$	$97.4_{\pm 0.2}$	$93.4_{\pm 0.6}$	$98.0_{\pm 0.2}$
carpet	$99.9_{\pm 0.1}$	$98.5_{\pm 1.4}$	$100.0_{\pm 0.0}$	$99.6_{\pm 0.3}$	$99.3_{\pm 0.1}$	$98.7_{\pm 0.1}$	$99.3_{\pm 0.0}$	$98.7_{\pm 0.1}$
grid	$99.4_{\pm 0.5}$	$99.1_{\pm 0.6}$	$99.7_{\pm 0.2}$	$97.7_{\pm 0.6}$	$98.2_{\pm 0.1}$	$98.7_{\pm 0.1}$	$98.2_{\pm 0.1}$	$98.7_{\pm 0.2}$
hazelnut	$99.4_{\pm 0.2}$	$98.6_{\pm 0.8}$	$99.7_{\pm 0.2}$	$99.4_{\pm 0.3}$	$98.8_{\pm 0.2}$	$98.9_{\pm 0.1}$	$98.2_{\pm 0.1}$	$99.4_{\pm 0.0}$
leather	$100.0_{\pm 0.0}$	$99.8_{\pm 0.2}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.3_{\pm 0.0}$	$98.8_{\pm 0.1}$	$99.4_{\pm 0.0}$	$99.0_{\pm 0.1}$
metal_nut	$99.9_{\pm 0.1}$	$98.8_{\pm 0.9}$	$98.3_{\pm 0.3}$	$99.7_{\pm 0.2}$	$87.2_{\pm 0.5}$	$95.8_{\pm 0.4}$	$90.5_{\pm 0.3}$	$95.4_{\pm 0.5}$
pill	$96.3_{\pm 0.2}$	$91.4_{\pm 5.1}$	$93.8_{\pm 0.6}$	$93.7_{\pm 3.6}$	$93.8_{\pm 0.4}$	$93.8_{\pm 0.6}$	$93.0_{\pm 0.5}$	$94.1_{\pm 0.7}$
screw	$62.8_{\pm 5.5}$	$77.4_{\pm 4.2}$	$78.5_{\pm 1.3}$	$80.6_{\pm 3.7}$	$92.5_{\pm 0.9}$	$97.2_{\pm 0.2}$	$94.8_{\pm 0.6}$	$97.9_{\pm 0.1}$
tile	$99.5_{\pm 0.3}$	$99.8_{\pm 0.1}$	$98.4_{\pm 0.7}$	$100.0_{\pm 0.0}$	$95.4_{\pm 0.2}$	$94.4_{\pm 0.2}$	$95.4_{\pm 0.1}$	$95.0_{\pm 0.2}$
toothbrush	$95.2_{\pm 1.0}$	$94.0_{\pm 4.4}$	$94.9_{\pm 0.9}$	$94.4_{\pm 3.1}$	$97.7_{\pm 0.3}$	$98.8_{\pm 0.2}$	$98.1_{\pm 0.2}$	$98.2_{\pm 0.3}$
transistor	$75.6_{\pm 7.5}$	$76.3_{\pm 7.0}$	$81.5_{\pm 4.6}$	$82.0_{\pm 6.1}$	$86.7_{\pm 1.7}$	$86.7_{\pm 1.4}$	$84.5_{\pm 1.7}$	$85.8_{\pm 1.5}$
wood	$98.8_{\pm 0.4}$	$99.3_{\pm 0.2}$	$99.1_{\pm 0.4}$	$100.0_{\pm 0.0}$	$95.2_{\pm 0.2}$	$92.1_{\pm 0.9}$	$94.9_{\pm 0.2}$	$93.4_{\pm 1.0}$
zipper	$94.1_{\pm 1.0}$	$99.2_{\pm 0.2}$	$92.2_{\pm 1.5}$	$99.7_{\pm 0.1}$	$90.7_{\pm 0.8}$	$96.9_{\pm 0.4}$	$91.9_{\pm 0.4}$	$97.1_{\pm 0.3}$
Average	$92.8_{\pm 0.9}$	$\underline{93.6_{\pm 0.7}}$	$93.4_{\pm 0.8}$	$\mathbf{95.0_{\pm 0.6}}$	$94.6_{\pm 0.2}$	$\underline{96.0_{\pm 0.1}}$	$94.9_{\pm 0.2}$	$\mathbf{96.1_{\pm 0.2}}$
2-shot								
bottle	$98.7_{\pm 0.8}$	$99.2_{\pm 1.2}$	$98.9_{\pm 0.7}$	$99.8_{\pm 0.4}$	$97.9_{\pm 0.1}$	$98.4_{\pm 0.1}$	$98.0_{\pm 0.1}$	$98.1_{\pm 0.1}$
cable	$91.4_{\pm 2.2}$	$91.6_{\pm 2.0}$	$90.0_{\pm 2.5}$	$93.3_{\pm 1.7}$	$94.3_{\pm 0.5}$	$93.8_{\pm 0.7}$	$94.3_{\pm 0.6}$	$93.9_{\pm 0.5}$
capsule	$76.5_{\pm 11.8}$	$85.2_{\pm 10.4}$	$74.1_{\pm 11.3}$	$86.6_{\pm 8.3}$	$90.6_{\pm 0.6}$	$97.6_{\pm 0.2}$	$93.2_{\pm 0.7}$	$98.1_{\pm 0.2}$
carpet	$99.9_{\pm 0.1}$	$99.0_{\pm 1.1}$	$100.0_{\pm 0.0}$	$99.6_{\pm 0.4}$	$99.3_{\pm 0.1}$	$98.7_{\pm 0.0}$	$99.2_{\pm 0.0}$	$98.7_{\pm 0.1}$
grid	$99.1_{\pm 0.6}$	$99.4_{\pm 0.5}$	$99.4_{\pm 0.4}$	$98.0_{\pm 0.5}$	$98.1_{\pm 0.2}$	$99.0_{\pm 0.1}$	$98.2_{\pm 0.1}$	$98.9_{\pm 0.2}$
hazelnut	$99.3_{\pm 0.4}$	$97.6_{\pm 2.0}$	$99.7_{\pm 0.2}$	$98.7_{\pm 1.3}$	$98.8_{\pm 0.1}$	$99.5_{\pm 0.1}$	$98.5_{\pm 0.1}$	$99.4_{\pm 0.0}$
leather	$100.0_{\pm 0.0}$	$99.9_{\pm 0.2}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.3_{\pm 0.0}$	$98.7_{\pm 0.1}$	$99.4_{\pm 0.0}$	$98.9_{\pm 0.1}$
metal_nut	$100.0_{\pm 0.1}$	$98.8_{\pm 0.6}$	$99.0_{\pm 0.8}$	$99.9_{\pm 0.2}$	$88.5_{\pm 1.5}$	$96.4_{\pm 0.7}$	$91.6_{\pm 1.2}$	$96.0_{\pm 0.8}$
pill	$96.4_{\pm 0.4}$	$91.6_{\pm 3.7}$	$93.7_{\pm 0.6}$	$94.0_{\pm 2.6}$	$94.2_{\pm 0.4}$	$94.4_{\pm 0.8}$	$93.3_{\pm 0.5}$	$94.7_{\pm 0.8}$
screw	$67.3_{\pm 6.4}$	$79.4_{\pm 3.7}$	$79.5_{\pm 1.5}$	$84.3_{\pm 4.7}$	$92.9_{\pm 1.2}$	$97.5_{\pm 0.3}$	$95.2_{\pm 0.6}$	$98.0_{\pm 0.2}$
tile	$99.6_{\pm 0.3}$	$99.9_{\pm 0.1}$	$98.7_{\pm 0.7}$	$100.0_{\pm 0.0}$	$95.5_{\pm 0.1}$	$94.5_{\pm 0.2}$	$95.4_{\pm 0.1}$	$95.1_{\pm 0.1}$
toothbrush	$96.2_{\pm 1.3}$	$96.0_{\pm 3.7}$	$95.5_{\pm 1.0}$	$96.8_{\pm 3.3}$	$97.9_{\pm 0.3}$	$99.0_{\pm 0.3}$	$98.3_{\pm 0.3}$	$98.4_{\pm 0.4}$
transistor	$75.6_{\pm 7.0}$	$77.8_{\pm 8.0}$	$80.7_{\pm 5.3}$	$83.0_{\pm 7.8}$	$86.6_{\pm 1.8}$	$86.2_{\pm 2.2}$	$84.4_{\pm 2.0}$	$85.3_{\pm 2.3}$
wood	$98.6_{\pm 0.5}$	$99.5_{\pm 0.3}$	$99.0_{\pm 0.4}$	$100.0_{\pm 0.0}$	$95.3_{\pm 0.2}$	$92.4_{\pm 0.8}$	$95.0_{\pm 0.2}$	$93.7_{\pm 0.9}$
zipper	$92.7_{\pm 4.0}$	$99.4_{\pm 0.3}$	$90.7_{\pm 6.4}$	$99.7_{\pm 0.1}$	$90.6_{\pm 0.7}$	$96.7_{\pm 0.4}$	$92.0_{\pm 0.5}$	$97.0_{\pm 0.3}$
Average	$92.8_{\pm 0.5}$	$\underline{95.0_{\pm 0.0}}$	$93.1_{\pm 0.4}$	$\mathbf{96.2_{\pm 0.4}}$	$94.8_{\pm 0.1}$	$\underline{96.3_{\pm 0.1}}$	$95.2_{\pm 0.2}$	$\mathbf{96.4_{\pm 0.1}}$
4-shot								
bottle	$99.0_{\pm 0.8}$	$99.5_{\pm 1.1}$	$99.1_{\pm 0.7}$	$99.9_{\pm 0.3}$	$97.9_{\pm 0.1}$	$98.5_{\pm 0.1}$	$98.0_{\pm 0.1}$	$98.2_{\pm 0.1}$
cable	$92.0_{\pm 2.2}$	$92.1_{\pm 1.9}$	$90.7_{\pm 2.4}$	$93.9_{\pm 1.7}$	$94.5_{\pm 0.6}$	$94.0_{\pm 0.7}$	$94.5_{\pm 0.6}$	$94.1_{\pm 0.6}$
capsule	$78.6_{\pm 12.5}$	$84.2_{\pm 11.3}$	$76.0_{\pm 12.2}$	$86.1_{\pm 9.3}$	$90.8_{\pm 0.8}$	$97.7_{\pm 0.3}$	$93.3_{\pm 0.8}$	$98.2_{\pm 0.2}$
carpet	$99.9_{\pm 0.2}$	$99.1_{\pm 1.0}$	$100.0_{\pm 0.1}$	$99.6_{\pm 0.4}$	$99.3_{\pm 0.1}$	$98.8_{\pm 0.1}$	$99.3_{\pm 0.0}$	$98.8_{\pm 0.1}$
grid	$99.0_{\pm 0.5}$	$99.6_{\pm 0.5}$	$99.4_{\pm 0.2}$	$98.2_{\pm 0.5}$	$98.2_{\pm 0.2}$	$98.8_{\pm 0.3}$	$98.3_{\pm 0.1}$	$98.9_{\pm 0.2}$
hazelnut	$99.5_{\pm 0.5}$	$98.2_{\pm 1.9}$	$99.8_{\pm 0.2}$	$99.1_{\pm 1.2}$	$98.8_{\pm 0.1}$	$99.5_{\pm 0.1}$	$98.5_{\pm 0.1}$	$99.4_{\pm 0.1}$
leather	$100.0_{\pm 0.0}$	$99.9_{\pm 0.2}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.3_{\pm 0.1}$	$99.5_{\pm 0.1}$	$99.4_{\pm 0.0}$	$98.9_{\pm 0.1}$
metal_nut	$100.0_{\pm 0.0}$	$99.0_{\pm 0.6}$	$99.4_{\pm 0.8}$	$99.9_{\pm 0.2}$	$89.5_{\pm 2.0}$	$96.8_{\pm 0.8}$	$92.4_{\pm 1.5}$	$96.5_{\pm 0.9}$
pill	$96.4_{\pm 0.4}$	$92.5_{\pm 3.3}$	$93.9_{\pm 0.8}$	$94.7_{\pm 2.4}$	$94.5_{\pm 0.6}$	$94.7_{\pm 0.8}$	$93.7_{\pm 0.7}$	$95.0_{\pm 0.8}$
screw	$71.3_{\pm 7.7}$	$81.5_{\pm 4.5}$	$80.6_{\pm 2.4}$	$86.2_{\pm 5.0}$	$93.8_{\pm 1.6}$	$97.8_{\pm 0.5}$	$95.6_{\pm 0.7}$	$98.2_{\pm 0.3}$
tile	$99.7_{\pm 0.3}$	$99.9_{\pm 0.1}$	$99.0_{\pm 0.8}$	$100.0_{\pm 0.0}$	$95.5_{\pm 0.2}$	$94.5_{\pm 0.2}$	$95.5_{\pm 0.2}$	$95.1_{\pm 0.2}$
toothbrush	$96.3_{\pm 1.1}$	$96.2_{\pm 3.4}$	$95.7_{\pm 1.2}$	$97.0_{\pm 3.0}$	$97.9_{\pm 0.3}$	$99.0_{\pm 0.2}$	$98.3_{\pm 0.2}$	$98.4_{\pm 0.3}$
transistor	$80.0_{\pm 8.5}$	$82.1_{\pm 9.1}$	$83.7_{\pm 6.2}$	$86.4_{\pm 8.0}$	$87.9_{\pm 2.4}$	$87.4_{\pm 2.5}$	$85.9_{\pm 2.7}$	$86.6_{\pm 2.7}$
wood	$98.3_{\pm 0.8}$	$99.6_{\pm 0.3}$	$98.9_{\pm 0.3}$	$100.0_{\pm 0.0}$	$95.3_{\pm 0.2}$	$92.3_{\pm 0.7}$	$95.1_{\pm 0.2}$	$93.6_{\pm 0.8}$
zipper	$93.6_{\pm 3.5}$	$99.5_{\pm 0.3}$	$92.3_{\pm 5.7}$	$99.8_{\pm 0.1}$	$90.9_{\pm 0.9}$	$97.0_{\pm 0.5}$	$92.3_{\pm 0.7}$	$97.3_{\pm 0.4}$
Average	$95.2_{\pm 0.9}$	$\underline{96.0_{\pm 1.2}}$	$95.2_{\pm 0.8}$	$\mathbf{97.0_{\pm 0.8}}$	$95.6_{\pm 0.1}$	$\underline{96.8_{\pm 0.1}}$	$95.9_{\pm 0.1}$	$\mathbf{96.9_{\pm 0.1}}$

Table 2. Comparison of image-level and pixel-level AUC for each category in MVTec using training-free few-shot methods.

Category	Image-Level AUC				Pixel-Level AUC			
	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B
1-shot								
candle	$92.0_{\pm 1.6}$	$86.2_{\pm 3.6}$	$94.8_{\pm 1.0}$	$90.7_{\pm 1.1}$	$96.8_{\pm 0.0}$	$98.7_{\pm 0.1}$	$97.0_{\pm 0.1}$	$98.7_{\pm 0.1}$
capsules	$79.2_{\pm 1.1}$	$95.7_{\pm 0.6}$	$80.5_{\pm 2.4}$	$93.1_{\pm 0.3}$	$94.7_{\pm 0.3}$	$97.3_{\pm 0.1}$	$95.4_{\pm 0.1}$	$96.7_{\pm 0.1}$
cashew	$92.4_{\pm 1.5}$	$86.2_{\pm 1.8}$	$93.4_{\pm 1.4}$	$90.5_{\pm 1.1}$	$97.0_{\pm 0.3}$	$99.0_{\pm 0.1}$	$97.5_{\pm 0.1}$	$98.4_{\pm 0.1}$
chewinggum	$97.4_{\pm 0.0}$	$98.2_{\pm 0.7}$	$97.9_{\pm 0.1}$	$98.7_{\pm 0.1}$	$98.7_{\pm 0.1}$	$99.4_{\pm 0.0}$	$98.9_{\pm 0.1}$	$99.4_{\pm 0.0}$
fryum	$93.7_{\pm 2.3}$	$94.2_{\pm 1.0}$	$92.6_{\pm 1.5}$	$95.7_{\pm 1.0}$	$91.0_{\pm 0.7}$	$92.7_{\pm 0.2}$	$93.1_{\pm 0.4}$	$93.8_{\pm 0.3}$
macaroni1	$85.1_{\pm 3.1}$	$88.6_{\pm 2.2}$	$91.1_{\pm 0.8}$	$92.3_{\pm 1.1}$	$96.7_{\pm 0.4}$	$99.2_{\pm 0.0}$	$97.6_{\pm 0.3}$	$99.3_{\pm 0.0}$
macaroni2	$63.5_{\pm 4.1}$	$62.9_{\pm 3.9}$	$73.3_{\pm 2.4}$	$67.4_{\pm 2.9}$	$91.2_{\pm 0.9}$	$96.9_{\pm 0.2}$	$94.6_{\pm 0.1}$	$98.3_{\pm 0.1}$
pcb1	$66.3_{\pm 17.8}$	$75.6_{\pm 4.8}$	$79.7_{\pm 8.6}$	$82.1_{\pm 7.0}$	$95.6_{\pm 1.1}$	$99.0_{\pm 0.0}$	$98.0_{\pm 0.3}$	$99.0_{\pm 0.0}$
pcb2	$74.9_{\pm 2.9}$	$75.0_{\pm 3.5}$	$72.7_{\pm 3.4}$	$82.3_{\pm 3.7}$	$92.2_{\pm 0.3}$	$95.7_{\pm 0.4}$	$92.8_{\pm 0.3}$	$95.9_{\pm 0.4}$
pcb3	$79.7_{\pm 4.6}$	$77.5_{\pm 0.8}$	$73.5_{\pm 5.6}$	$86.8_{\pm 1.5}$	$93.3_{\pm 0.6}$	$96.7_{\pm 0.2}$	$93.4_{\pm 0.7}$	$96.2_{\pm 0.2}$
pcb4	$90.6_{\pm 4.3}$	$87.2_{\pm 1.4}$	$85.5_{\pm 4.3}$	$93.1_{\pm 1.3}$	$89.8_{\pm 1.0}$	$94.4_{\pm 0.6}$	$92.1_{\pm 0.8}$	$94.7_{\pm 0.6}$
pipe_fryum	$99.4_{\pm 0.2}$	$97.6_{\pm 0.9}$	$99.6_{\pm 0.2}$	$95.8_{\pm 0.7}$	$97.0_{\pm 0.4}$	$98.7_{\pm 0.1}$	$96.6_{\pm 0.6}$	$97.9_{\pm 0.2}$
Average	$84.5_{\pm 1.3}$	$85.4_{\pm 0.3}$	$\underline{86.2_{\pm 0.6}}$	$\mathbf{89.1_{\pm 0.4}}$	$94.5_{\pm 0.2}$	$\underline{97.3_{\pm 0.1}}$	$95.6_{\pm 0.1}$	$\mathbf{97.4_{\pm 0.1}}$
2-shot								
candle	$92.2_{\pm 1.4}$	$88.9_{\pm 3.7}$	$94.6_{\pm 1.1}$	$91.7_{\pm 1.4}$	$97.0_{\pm 0.2}$	$98.8_{\pm 0.2}$	$97.2_{\pm 0.2}$	$98.8_{\pm 0.2}$
capsules	$79.8_{\pm 1.8}$	$96.0_{\pm 1.3}$	$80.7_{\pm 2.6}$	$93.1_{\pm 0.6}$	$94.7_{\pm 0.3}$	$97.3_{\pm 0.3}$	$95.4_{\pm 0.2}$	$96.7_{\pm 0.3}$
cashew	$92.7_{\pm 1.7}$	$87.8_{\pm 2.8}$	$94.6_{\pm 1.7}$	$91.6_{\pm 1.7}$	$97.2_{\pm 0.3}$	$99.1_{\pm 0.1}$	$97.6_{\pm 0.2}$	$98.6_{\pm 0.2}$
chewinggum	$97.1_{\pm 0.5}$	$97.9_{\pm 0.7}$	$97.5_{\pm 0.5}$	$98.7_{\pm 0.2}$	$98.7_{\pm 0.1}$	$99.4_{\pm 0.1}$	$98.8_{\pm 0.1}$	$99.3_{\pm 0.0}$
fryum	$94.7_{\pm 1.9}$	$95.5_{\pm 1.5}$	$93.0_{\pm 1.2}$	$96.8_{\pm 1.3}$	$91.6_{\pm 0.8}$	$93.4_{\pm 0.7}$	$93.5_{\pm 0.5}$	$94.3_{\pm 0.6}$
macaroni1	$87.1_{\pm 3.6}$	$88.9_{\pm 1.9}$	$90.4_{\pm 1.7}$	$92.9_{\pm 1.2}$	$97.1_{\pm 0.8}$	$99.3_{\pm 0.1}$	$97.8_{\pm 0.6}$	$99.4_{\pm 0.1}$
macaroni2	$71.6_{\pm 8.7}$	$66.5_{\pm 4.9}$	$76.8_{\pm 3.9}$	$69.6_{\pm 3.2}$	$92.9_{\pm 1.9}$	$97.3_{\pm 0.5}$	$95.3_{\pm 0.8}$	$98.6_{\pm 0.3}$
pcb1	$73.8_{\pm 15.0}$	$76.7_{\pm 3.6}$	$81.7_{\pm 6.9}$	$83.5_{\pm 5.4}$	$96.8_{\pm 1.4}$	$99.1_{\pm 0.1}$	$98.3_{\pm 0.4}$	$99.1_{\pm 0.1}$
pcb2	$74.5_{\pm 2.7}$	$75.7_{\pm 4.1}$	$71.4_{\pm 3.3}$	$83.2_{\pm 3.2}$	$92.9_{\pm 0.8}$	$96.2_{\pm 0.6}$	$93.3_{\pm 0.7}$	$96.4_{\pm 0.6}$
pcb3	$83.3_{\pm 5.0}$	$80.1_{\pm 3.5}$	$77.1_{\pm 5.5}$	$87.5_{\pm 1.9}$	$93.7_{\pm 0.6}$	$96.8_{\pm 0.2}$	$93.9_{\pm 0.7}$	$96.3_{\pm 0.2}$
pcb4	$87.4_{\pm 12.4}$	$87.3_{\pm 1.2}$	$85.2_{\pm 11.1}$	$93.3_{\pm 1.2}$	$90.9_{\pm 1.4}$	$94.5_{\pm 0.5}$	$92.9_{\pm 1.1}$	$94.9_{\pm 0.5}$
pipe_fryum	$99.5_{\pm 0.2}$	$97.5_{\pm 0.7}$	$99.7_{\pm 0.2}$	$95.9_{\pm 0.5}$	$97.2_{\pm 0.5}$	$98.7_{\pm 0.1}$	$96.8_{\pm 0.6}$	$98.0_{\pm 0.2}$
Average	$\underline{87.8_{\pm 1.4}}$	$87.7_{\pm 0.3}$	$87.6_{\pm 1.2}$	$\mathbf{90.6_{\pm 0.5}}$	$95.6_{\pm 0.1}$	$\mathbf{97.7_{\pm 0.1}}$	$\underline{96.2_{\pm 0.1}}$	$\mathbf{97.7_{\pm 0.1}}$
4-shot								
candle	$92.9_{\pm 1.7}$	$90.2_{\pm 3.6}$	$94.9_{\pm 1.2}$	$92.3_{\pm 1.5}$	$97.2_{\pm 0.4}$	$98.9_{\pm 0.2}$	$97.4_{\pm 0.4}$	$98.9_{\pm 0.2}$
capsules	$80.5_{\pm 2.2}$	$96.2_{\pm 1.1}$	$81.6_{\pm 2.5}$	$93.4_{\pm 0.7}$	$95.0_{\pm 0.6}$	$97.4_{\pm 0.3}$	$95.7_{\pm 0.5}$	$96.8_{\pm 0.3}$
cashew	$92.8_{\pm 1.9}$	$88.5_{\pm 2.6}$	$94.7_{\pm 1.6}$	$92.3_{\pm 1.7}$	$97.2_{\pm 0.3}$	$99.1_{\pm 0.2}$	$97.6_{\pm 0.2}$	$98.6_{\pm 0.2}$
chewinggum	$97.5_{\pm 0.7}$	$98.2_{\pm 0.7}$	$97.6_{\pm 0.5}$	$98.7_{\pm 0.3}$	$98.7_{\pm 0.1}$	$99.4_{\pm 0.0}$	$98.8_{\pm 0.1}$	$99.3_{\pm 0.0}$
fryum	$95.1_{\pm 1.7}$	$95.9_{\pm 1.5}$	$93.6_{\pm 1.4}$	$97.0_{\pm 1.1}$	$92.2_{\pm 1.0}$	$93.8_{\pm 0.9}$	$94.0_{\pm 0.8}$	$94.7_{\pm 0.7}$
macaroni1	$87.3_{\pm 3.0}$	$89.1_{\pm 1.7}$	$90.3_{\pm 1.4}$	$93.2_{\pm 1.1}$	$97.2_{\pm 0.7}$	$99.4_{\pm 0.1}$	$97.9_{\pm 0.5}$	$99.4_{\pm 0.1}$
macaroni2	$73.0_{\pm 7.5}$	$67.6_{\pm 4.4}$	$77.0_{\pm 3.2}$	$70.8_{\pm 3.8}$	$93.6_{\pm 1.9}$	$97.5_{\pm 0.5}$	$95.4_{\pm 0.7}$	$98.7_{\pm 0.3}$
pcb1	$78.8_{\pm 14.2}$	$79.1_{\pm 4.8}$	$84.4_{\pm 7.0}$	$85.9_{\pm 5.8}$	$97.1_{\pm 1.3}$	$99.2_{\pm 0.1}$	$98.5_{\pm 0.4}$	$99.2_{\pm 0.1}$
pcb2	$76.6_{\pm 4.1}$	$77.9_{\pm 4.7}$	$74.5_{\pm 5.3}$	$84.7_{\pm 3.4}$	$93.5_{\pm 1.1}$	$96.5_{\pm 0.7}$	$93.9_{\pm 1.0}$	$96.7_{\pm 0.7}$
pcb3	$84.7_{\pm 4.6}$	$80.4_{\pm 2.9}$	$80.0_{\pm 6.1}$	$88.2_{\pm 1.9}$	$94.1_{\pm 0.7}$	$97.1_{\pm 0.4}$	$94.4_{\pm 0.8}$	$96.5_{\pm 0.4}$
pcb4	$87.3_{\pm 10.3}$	$87.6_{\pm 1.1}$	$84.5_{\pm 9.2}$	$93.6_{\pm 1.1}$	$91.8_{\pm 1.8}$	$94.9_{\pm 0.6}$	$93.6_{\pm 1.3}$	$95.2_{\pm 0.6}$
pipe_fryum	$99.6_{\pm 0.2}$	$97.5_{\pm 1.0}$	$99.8_{\pm 0.2}$	$96.2_{\pm 0.9}$	$97.3_{\pm 0.4}$	$98.8_{\pm 0.1}$	$96.9_{\pm 0.5}$	$98.1_{\pm 0.2}$
Average	$89.3_{\pm 0.5}$	$88.9_{\pm 0.0}$	$\underline{89.5_{\pm 0.3}}$	$\mathbf{91.9_{\pm 0.3}}$	$96.1_{\pm 0.1}$	$\mathbf{98.0_{\pm 0.0}}$	$\underline{96.7_{\pm 0.1}}$	$\mathbf{98.0_{\pm 0.0}}$

Table 3. Comparison of image-level and pixel-level AUC for each category in VisA using training-free few-shot methods.

Category	Image-Level AUC				Pixel-Level AUC			
	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B
MVTec								
bottle	98.7	99.5	97.7	99.8	98.2	98.7	98.2	97.9
cable	93.5	96.4	92.4	95.1	95.2	95.3	94.3	93.0
capsule	88.0	86.9	91.1	89.1	92.8	97.7	92.1	98.5
carpet	98.2	98.6	99.1	99.9	98.9	98.3	99.0	98.6
grid	98.5	90.1	93.9	99.7	97.9	98.6	97.8	99.0
hazelnut	98.2	90.2	98.2	95.1	99.1	99.4	99.1	99.3
leather	99.4	95.1	96.9	100.0	98.9	98.5	99.0	99.3
metal_nut	95.4	96.2	97.1	95.4	79.4	87.6	78.4	85.4
pill	95.2	95.2	95.2	95.6	95.0	93.0	94.9	95.2
screw	80.9	81.1	82.5	87.0	97.0	97.0	93.6	97.3
tile	100.0	100.0	99.8	99.8	96.0	96.7	95.6	95.9
toothbrush	100.0	100.0	99.2	98.9	98.4	98.8	98.4	97.4
transistor	89.4	91.9	90.0	92.7	91.3	92.5	89.7	88.9
wood	92.2	93.9	94.9	97.3	96.1	92.8	96.0	91.0
zipper	90.3	98.7	91.3	93.9	90.1	93.2	90.1	97.1
Average	94.5	94.3	<u>94.6</u>	**95.9**	95.0	**95.9**	94.4	<u>95.6</u>
VisA								
candle	93.1	90.8	93.3	89.8	97.7	98.2	97.7	99.0
capsules	80.7	92.5	81.7	94.6	96.2	96.0	95.7	97.0
cashew	86.0	88.7	84.3	93.7	97.9	99.4	97.7	98.5
chewinggum	97.7	97.7	98.1	98.8	98.5	98.8	98.6	99.0
fryum	95.6	92.7	92.3	96.0	94.1	92.7	93.1	94.6
macaroni1	87.5	85.3	90.8	96.7	98.0	98.5	98.0	99.5
macaroni2	75.3	67.2	76.2	82.4	95.6	97.5	95.6	98.9
pcb1	81.3	79.5	82.9	88.9	98.2	99.0	98.4	99.3
pcb2	81.4	81.6	81.7	87.3	95.0	95.8	95.0	97.2
pcb3	88.2	91.2	90.5	88.2	95.5	96.8	95.2	96.6
pcb4	78.3	87.9	75.0	93.8	93.3	90.9	92.6	94.8
pipe_fryum	96.8	94.5	96.4	96.4	97.9	99.0	97.8	98.0
Average	86.8	<u>87.5</u>	86.9	**92.2**	96.5	<u>96.9</u>	96.3	**97.7**
MPDD								
bracket_black	50.9	46.3	53.4	45.8	89.8	95.3	91.4	96.6
bracket_brown	50.5	50.7	51.7	59.7	93.8	94.0	93.8	94.3
bracket_white	43.1	35.7	45.0	43.2	92.2	97.6	92.8	98.6
connector	90.4	77.7	90.5	79.5	97.5	97.7	97.6	97.7
metal_plate	90.4	88.6	91.5	88.4	92.2	91.9	91.7	93.1
tubes	91.7	92.9	91.7	94.9	98.5	98.8	98.6	99.2
Average	<u>69.5</u>	65.3	**70.6**	68.6	94.0	<u>95.9</u>	94.3	**96.6**

Table 4. Comparison of image-level and pixel-level AUC for each category in MVTec, VisA, and MPDD using training-free zero-shot methods.

Category	Image-Level AUC				Pixel-Level AUC			
	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B	PatchEAD ViT-B/16+	PatchEAD DINOv2-B	PatchEAD+ ViT-B/16+	PatchEAD+ DINOv2-B
BTAD								
01	96.4	95.7	93.5	94.1	95.3	97.3	95.3	97.0
02	81.0	82.2	81.7	81.4	94.6	95.3	95.0	94.4
03	99.5	99.2	99.2	98.1	99.4	99.6	99.4	99.5
Average	**92.3**	**92.3**	91.5	91.2	96.4	**97.4**	96.5	<u>97.0</u>
KSDD								
KSDD	89.0	93.7	89.4	97.0	97.6	<u>99.2</u>	97.7	**99.3**
Average	89.0	<u>93.7</u>	89.4	**97.0**	97.6	<u>99.2</u>	97.7	**99.3**
DAGM								
Class1	78.2	85.1	83.6	88.3	86.7	80.8	87.5	76.2
Class2	99.4	100.0	99.6	99.8	99.2	99.8	99.2	99.6
Class3	97.1	99.6	95.3	99.9	95.3	97.8	95.7	97.3
Class4	100.0	100.0	100.0	100.0	99.2	97.1	99.0	95.8
Class5	97.7	99.9	98.8	100.0	96.0	99.8	96.1	99.7
Class6	91.6	100.0	95.3	100.0	91.6	99.6	92.1	99.5
Class7	99.9	100.0	100.0	100.0	95.9	97.8	95.7	96.8
Class8	82.4	99.8	83.7	99.0	93.1	99.7	93.5	99.7
Class9	78.5	91.6	87.3	95.9	92.2	99.6	95.0	99.9
Class10	100.0	100.0	99.8	100.0	99.6	99.7	99.6	99.3
Average	92.5	<u>97.6</u>	94.3	**98.3**	94.9	**97.2**	95.3	<u>96.4</u>
DTD-Synthetic								
Blotchy_099	93.7	94.1	97.6	99.3	99.0	97.8	99.0	99.0
Fibrous_183	97.6	90.4	98.1	96.7	98.7	98.7	98.6	99.1
Marbled_078	86.0	88.6	86.7	97.4	96.9	96.2	96.0	97.9
Matted_069	75.7	73.1	76.5	92.2	95.4	94.2	95.1	98.7
Mesh_114	88.2	95.6	87.4	95.9	97.7	98.4	97.1	97.1
Perforated_037	97.3	94.0	98.4	98.3	99.2	99.0	99.1	98.7
Stratified_154	90.6	98.5	84.9	98.8	99.2	99.3	99.2	99.8
Woven_001	98.9	97.5	98.5	99.6	99.4	99.8	99.2	99.8
Woven_068	94.4	91.3	94.5	95.4	98.7	98.7	98.3	98.1
Woven_104	96.5	85.5	90.9	98.9	97.2	97.4	97.6	98.6
Woven_125	94.9	91.8	94.2	99.2	99.0	99.0	98.8	98.5
Woven_127	95.1	95.9	94.7	99.0	94.2	95.7	94.3	94.9
Average	<u>92.4</u>	91.4	91.9	**97.6**	<u>97.9</u>	<u>97.9</u>	97.7	**98.3**

Table 5. Comparison of image-level and pixel-level AUC for each category in BTAD, KSDD, DAGM, and DTD-Synthetic using training-free zero-shot methods.

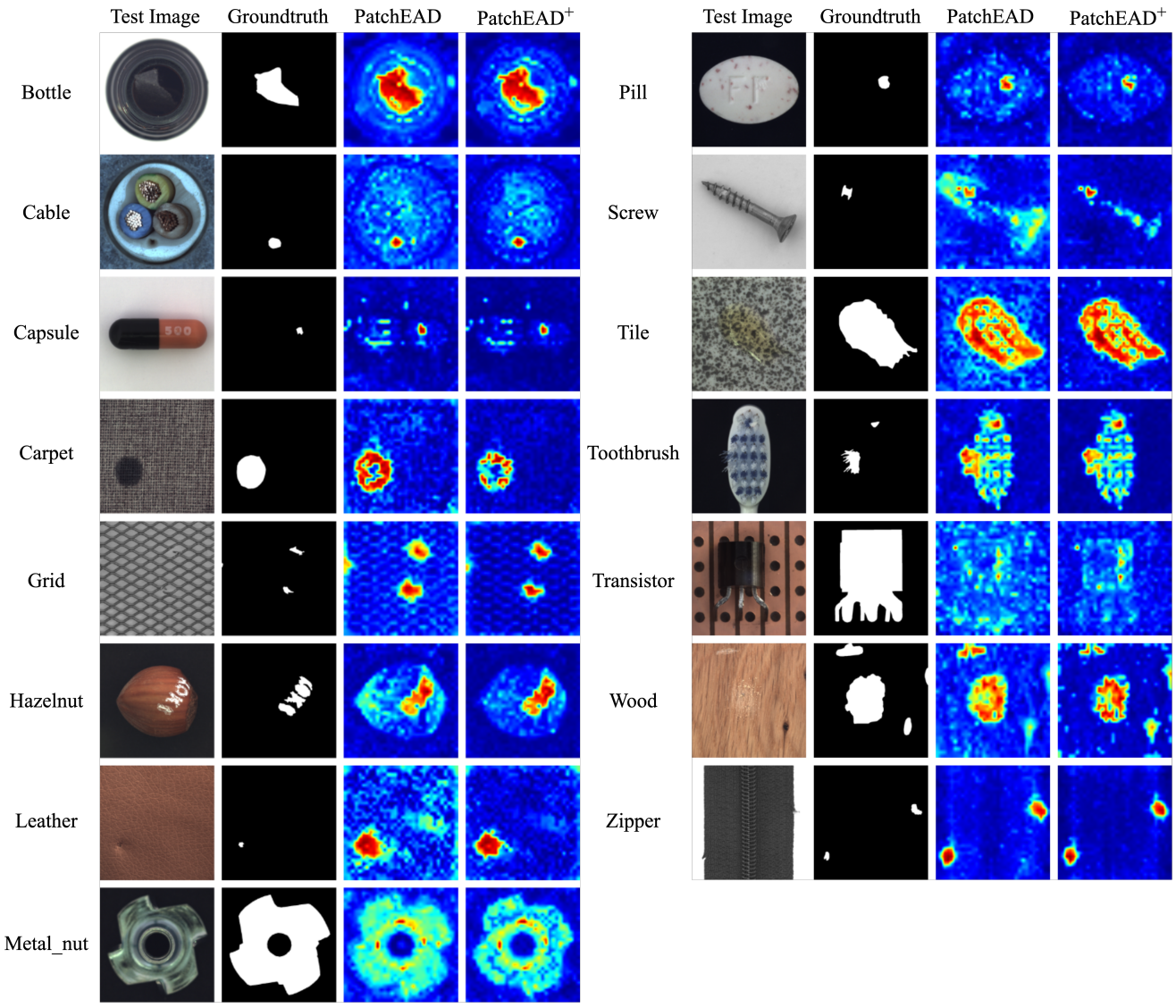


Figure 3. Additional visualization results from PatchEAD and PatchEAD+(4-shot), tested on MVTec.

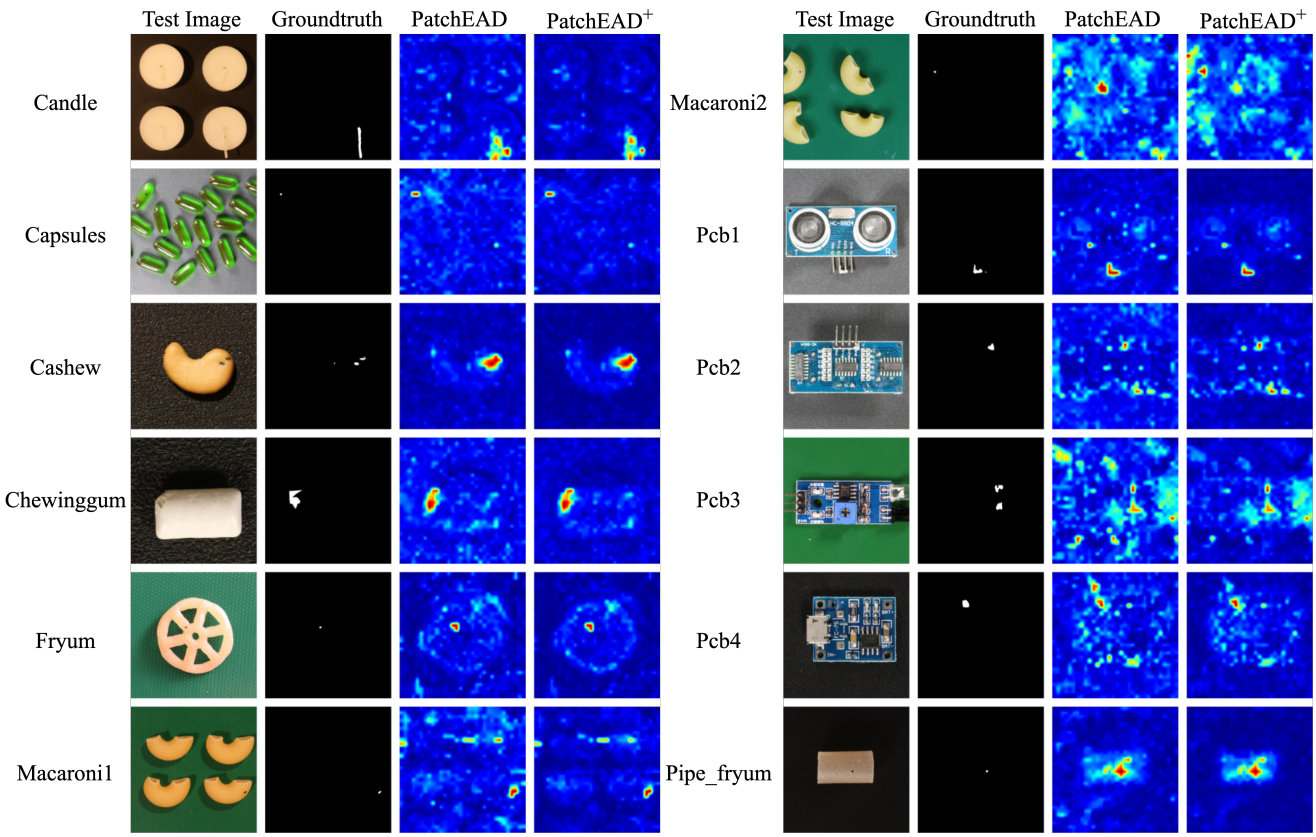


Figure 4. Additional visualization results from PatchEAD and PatchEAD+(4-shot), tested on VisA.

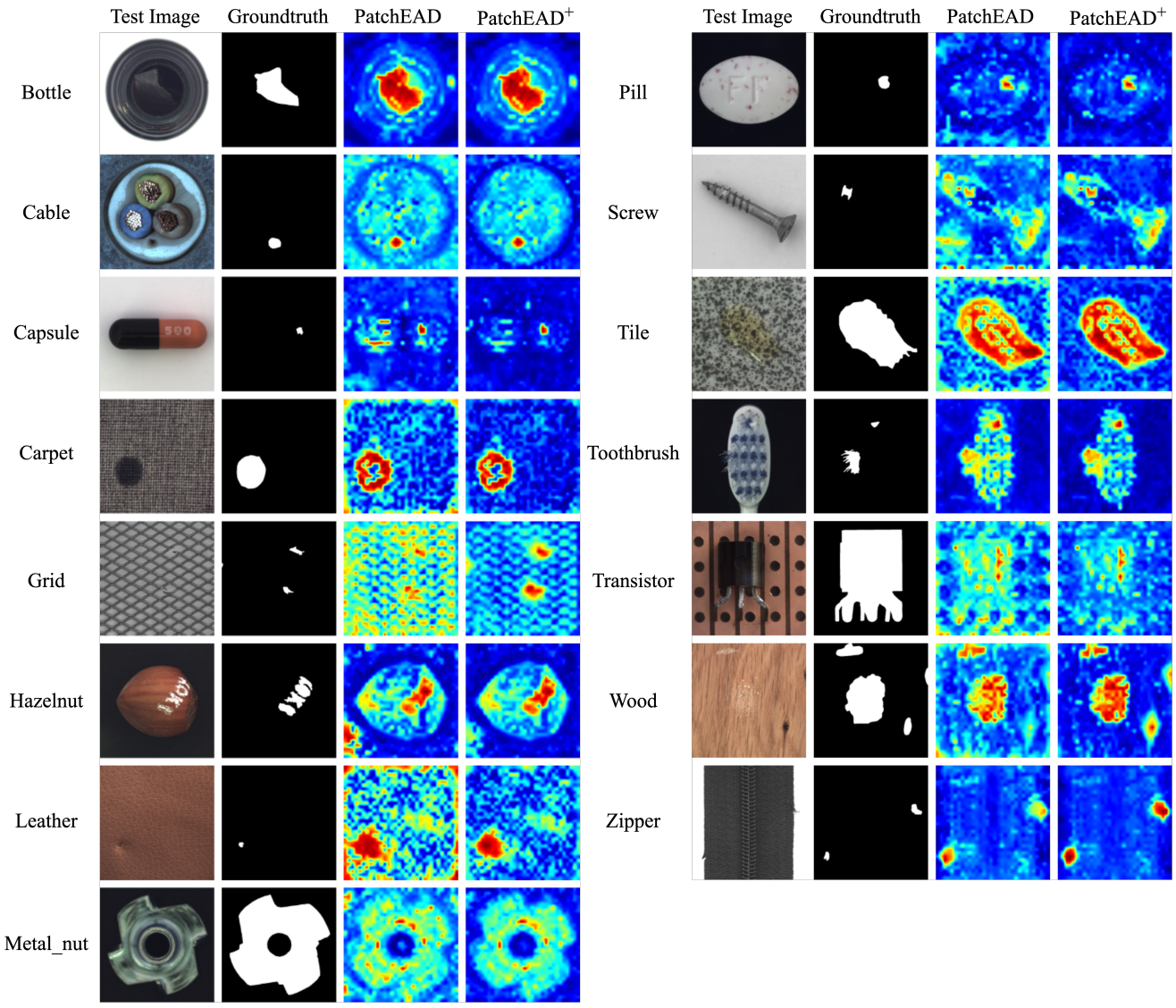


Figure 5. Additional visualization results from PatchEAD and PatchEAD+(0-shot), tested on MVTec.

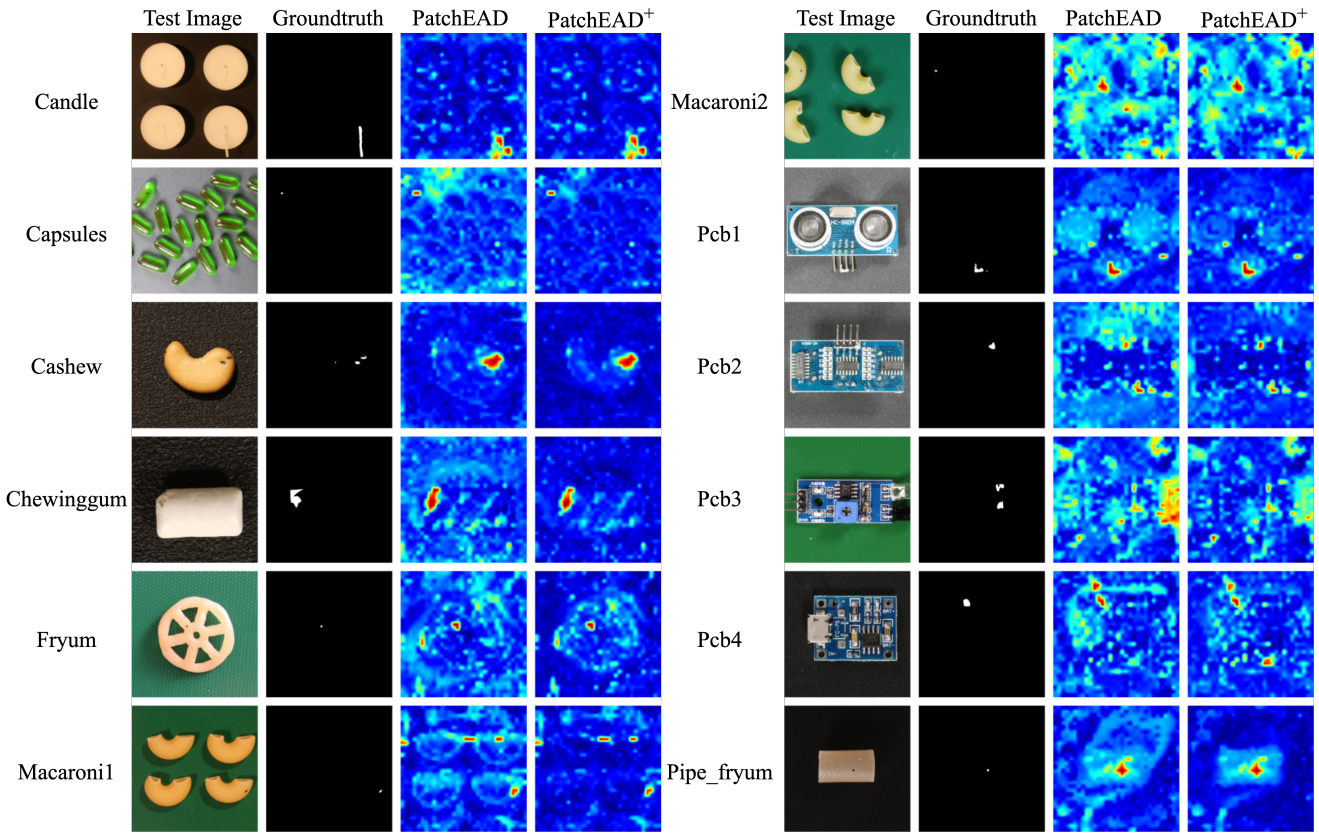


Figure 6. Additional visualization results from PatchEAD and PatchEAD+(0-shot), tested on VisA.

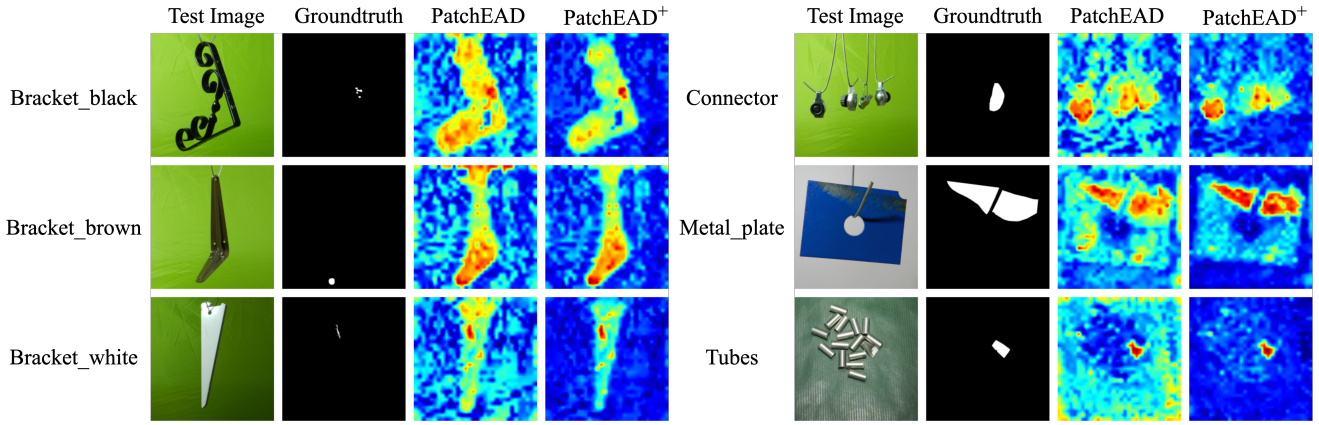


Figure 7. Additional visualization results from PatchEAD and PatchEAD+(0-shot), tested on MPDD.

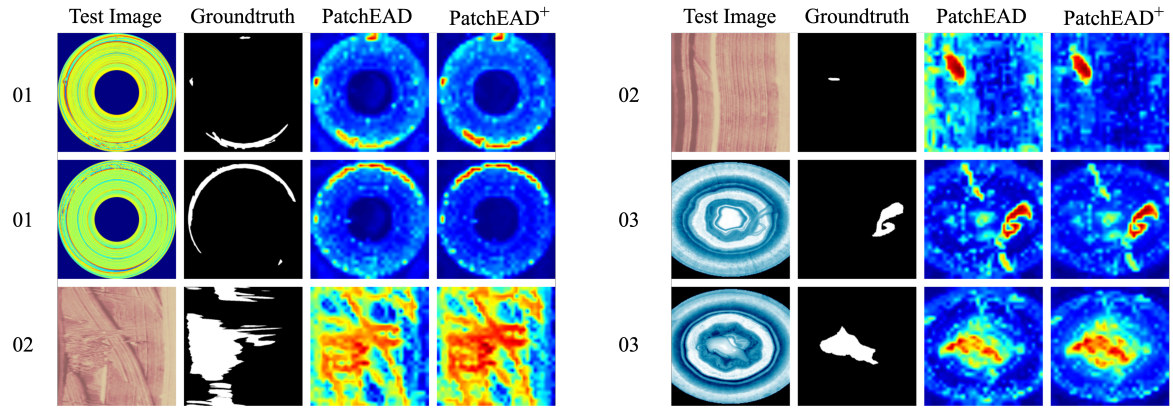


Figure 8. Additional visualization results from PatchEAD and PatchEAD+(0-shot), tested on BTAD.

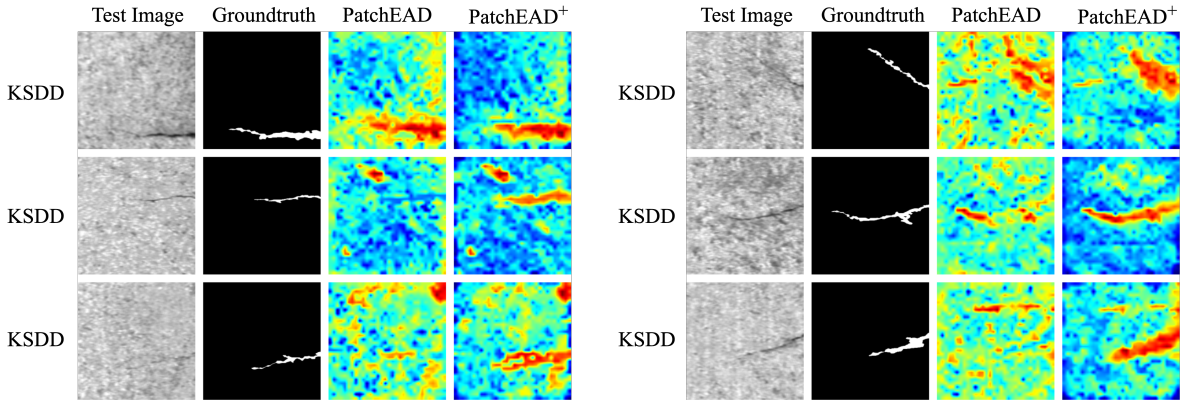


Figure 9. Additional visualization results from PatchEAD and PatchEAD+(0-shot), tested on KSDD.

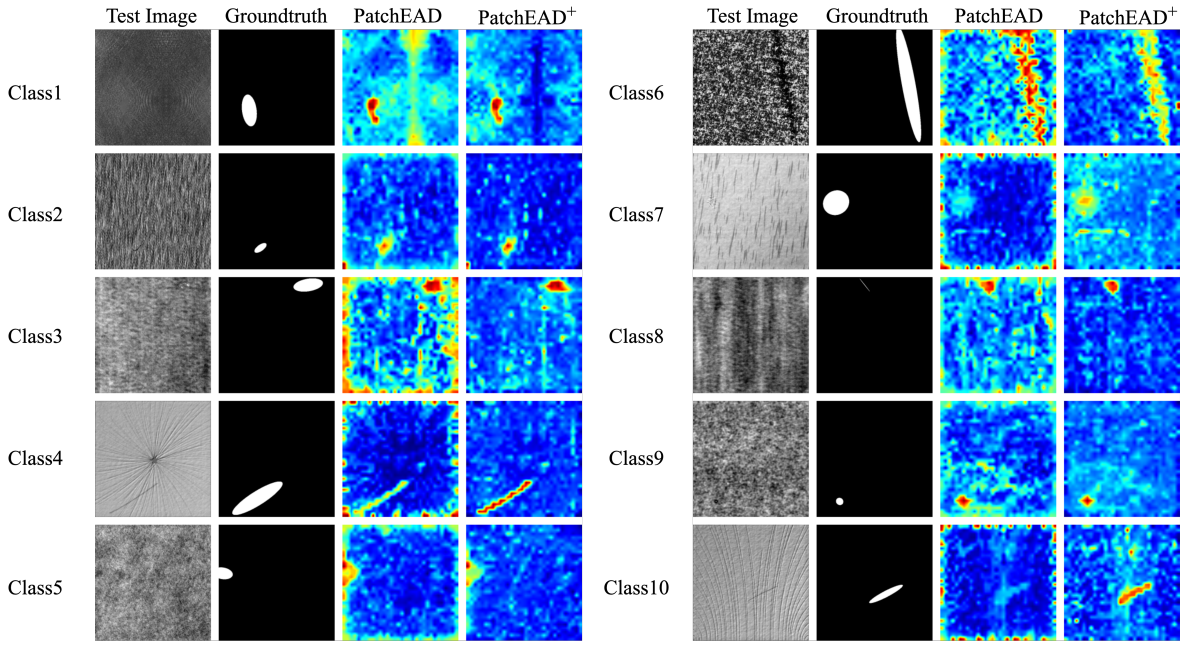


Figure 10. Additional visualization results from PatchEAD and PatchEAD+(0-shot), tested on DAGM.

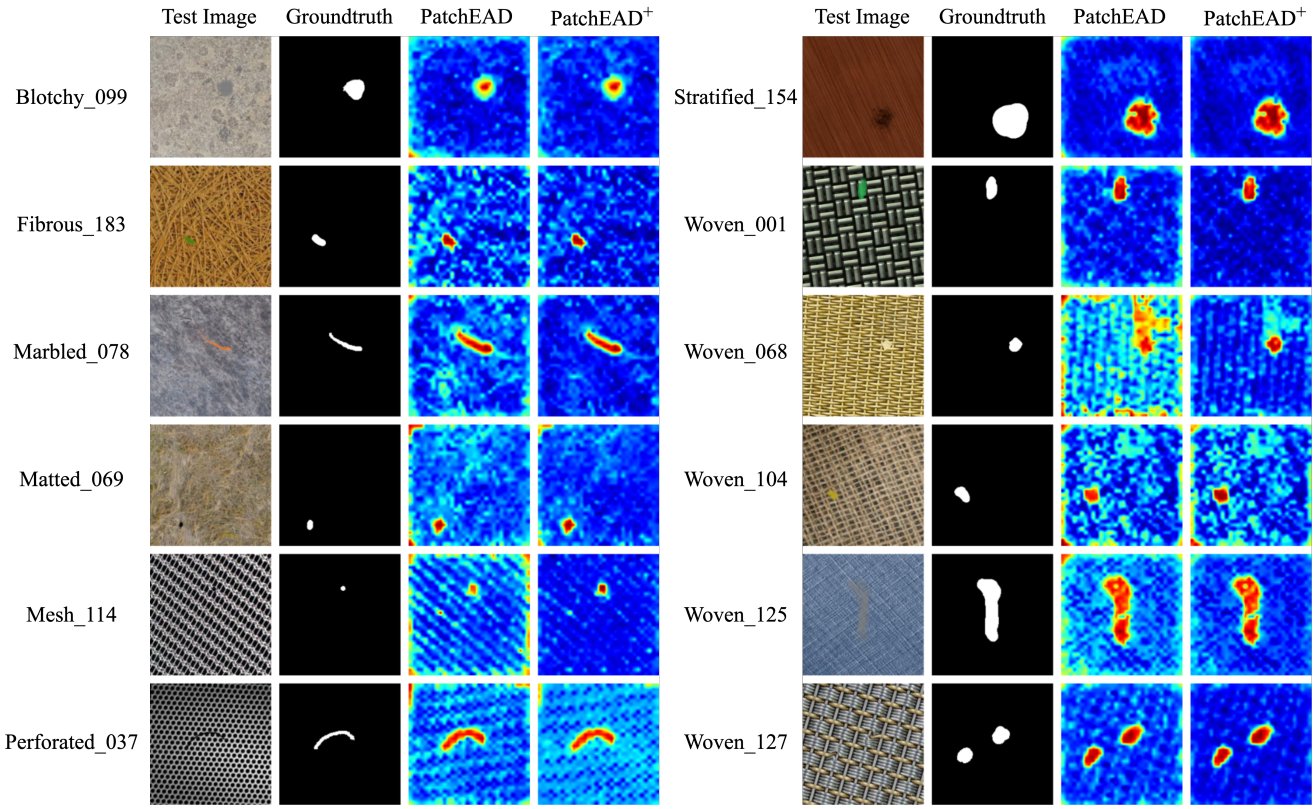


Figure 11. Additional visualization results from PatchEAD and PatchEAD+(0-shot), tested on DTD-Synthetic.

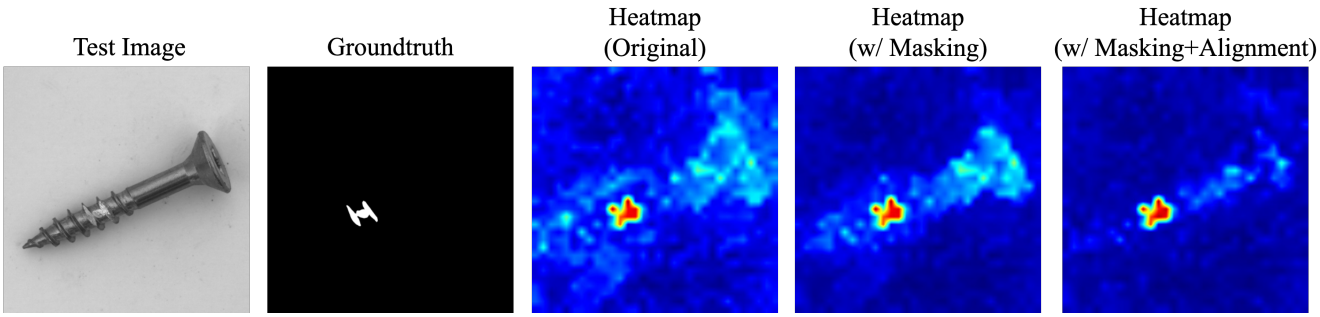


Figure 12. Additional visualization results for the comparison to evaluate the effectiveness of Alignment and Masking.

References

[1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023. 1

[2] 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*, pages 9592–9600, 2019. 1

[3] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020. 2

[4] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. If you use this software, please cite it as below. 1

[5] 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)*, pages 66–71, 2021. 1

[6] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: going beyond multi-view redundancy. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 32971–32998, 2023. 1

[7] 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 *30th IEEE/IES International Symposium on Industrial Electronics (ISIE)*, 2021. 1

[8] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. 1

[9] 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*, pages 8748–8763. PMLR, 2021. 1

[10] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020. 1

[11] Y-H Tsai, Y Wu, R Salakhutdinov, and L-P Morency. Self-supervised learning from a multi-view perspective. In *Proceedings of the International Conference on Learning Representations (ICLR), 2021*, 2021. 2

[12] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, page 11, 2007. 1

[13] 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*, pages 392–408. Springer, 2022. 1