PromptAD: Learning Prompts with only Normal Samples for Few-Shot Anomaly Detection

Supplementary Material

A. Experimental details

Data pre-processing. Referring to WinCLIP [7], we employ the data pre-processing pipeline specified in OpenCLIP [6] for both the MVTec [1] and VisA [14] datasets to mitigate potential train-test discrepancies. This involves channel-wise standardization using precomputed mean [0.48145466, 0.4578275, 0.40821073] and standard deviation [0.26862954, 0.26130258, 0.27577711] after normalizing each RGB image to [0, 1]. Subsequently, bicubic resizing is performed based on the `Pillow` implementation. As a default, we set the input resolution to 240 for the shorter edge resulting from resizing, aligning with ViT-B/16$\pm$ in our experiments.

Hyper-parameter. The length of trainable tokens in normal prompts (E_N) is set to 4, and the length of trainable tokens in learnable anomaly prompts (E_A) is set to 1. For each detection object, the number of normal prompts (N) is set to 1, the number of learnable anomaly suffixes (L) is set to 4, and the number of manual anomaly suffixes (M) depends on the number of anomaly labels in the dataset. λ is set to 0.001. Referring to CoOp [13], the optimizer parameters of prompt learning: learning rate, momentum, and weight decay are set to 0.002, 0.9, and 0.0005, respectively.

Manual anomaly suffixes. We used two kinds of manual anomaly suffixes: generic anomaly suffixes and object-customized anomaly suffixes. Generic anomaly suffixes are manually designed and object-customized anomaly suffixes are generated through anomaly labels in the datasets [1, 14]. The specific details are shown in Figure 1.

Evaluation metrics. In addition to the results for the Area Under the Receiver Operator Curve documented in the body of the paper, We also supplement the image-level Precision-Recall (AUPR) results and pixel-level Per-region-overlap (PRO) [1, 2] results.

Other details. Since model performance in the few-shot setting is affected by random sampling, we report the mean and standard deviation over 5 random seeds for each measurement. In addition, the few-shot results of SPADE [4], PaDiM [5], and PatchCore [11] in the experiments adopt the results recorded in WinCLIP [7].

1. Generic anomaly suffixes

```
'damaged {}', 'broken {}', '{} with flaw', '{} with defect', '{} with damage'
```

2. Object-customized anomaly suffixes (MVTec)

```
'bottle': ['{} with large breakage', '{} with small breakage', '{} with contamination'],
'toothbrush': ['{} with defect', '{} with anomaly'],
'carpet': ['{} with hole', '{} with color stain', '{} with metal contamination', '{} with thread residue', '{} with thread', '{} with cut'],
'hazelnut': ['{} with crack', '{} with cut', '{} with hole', '{} with print'],
'leather': ['{} with color stain', '{} with cut', '{} with fold', '{} with glue', '{} with poke'],
'cable': ['{} with bent wire', '{} with missing part', '{} with missing wire', '{} with cut', '{} with poke'],
'capsule': ['{} with crack', '{} with faulty imprint', '{} with poke', '{} with scratch', '{} squeezed with compression'],
'grid': ['{} with breakage',  '{} with thread residue', '{} with thread', '{} with metal contamination', '{} with glue', '{} with a bent shape'],
'pill': ['{} with color stain', '{} with contamination', '{} with crack', '{} with faulty imprint', '{} with scratch', '{} with abnormal type'],
'transistor': ['{} with bent lead', '{} with cut lead', '{} with damage', '{} with misplaced transistor'],
'metal_nut': ['{} with a bent shape ', '{} with color stain', '{} with a flipped orientation', '{} with scratch'],
'screw': ['{} with manipulated front',  '{} with scratch neck', '{} with scratch head'],
'zipper': ['{} with broken teeth', '{} with fabric border', '{} with defect fabric', '{} with broken fabric', '{} with split teeth',
           '{} with squeezed teeth'],
'tile': ['{} with crack', '{} with glue strip', '{} with gray stroke', '{} with oil', '{} with rough surface'],
'wood': ['{} with color stain', '{} with hole', '{} with scratch', '{} with liquid'],
```

3. Object-customized anomaly suffixes (VisA)

```
'candle': ['{} with melded wax', '{} with foreign particals', '{} with extra wax', '{} with chunk of wax missing', '{} with weird candle wick',
           '{} with damaged corner of packaging', '{} with different colour spot'],
'capsules': ['{} with scratch', '{} with discolor', '{} with misshape', '{} with leak', '{} with bubble'],
'cashew': ['{} with breakage', '{} with small scratches', '{} with burnt', '{} with stuck together', '{} with spot'],
'chewinggum': ['{} with corner missing', '{} with scratches', '{} with chunk of gum missing', '{} with colour spot', '{} with cracks'],
'fryum': ['{} with breakage', '{} with scratches', '{} with burnt', '{} with colour spot', '{} with fryum stuck together', '{} with colour spot'],
'macaroni1': ['{} with color spot', '{} with small chip around edge', '{} with small scratches', '{} with breakage', '{} with cracks'],
'macaroni2': ['{} with color spot', '{} with small chip around edge', '{} with small scratches', '{} with breakage', '{} with cracks'],
'pcb1': ['{} with bent', '{} with scratch', '{} with missing', '{} with melt'],
'pcb2': ['{} with bent', '{} with scratch', '{} with missing', '{} with melt'],
'pcb3': ['{} with bent', '{} with scratch', '{} with missing', '{} with melt'],
'pcb4': ['{} with scratch', '{} with extra', '{} with missing', '{} with wrong place', '{} with damage', '{} with burnt', '{} with dirt'],
'pipe_fryum': ['{} with breakage', '{} with small scratches', '{} with burnt', '{} with stuck together', '{} with colour spot', '{} with cracks']
```

Figure 1. Illustration of manual anomaly suffixes.

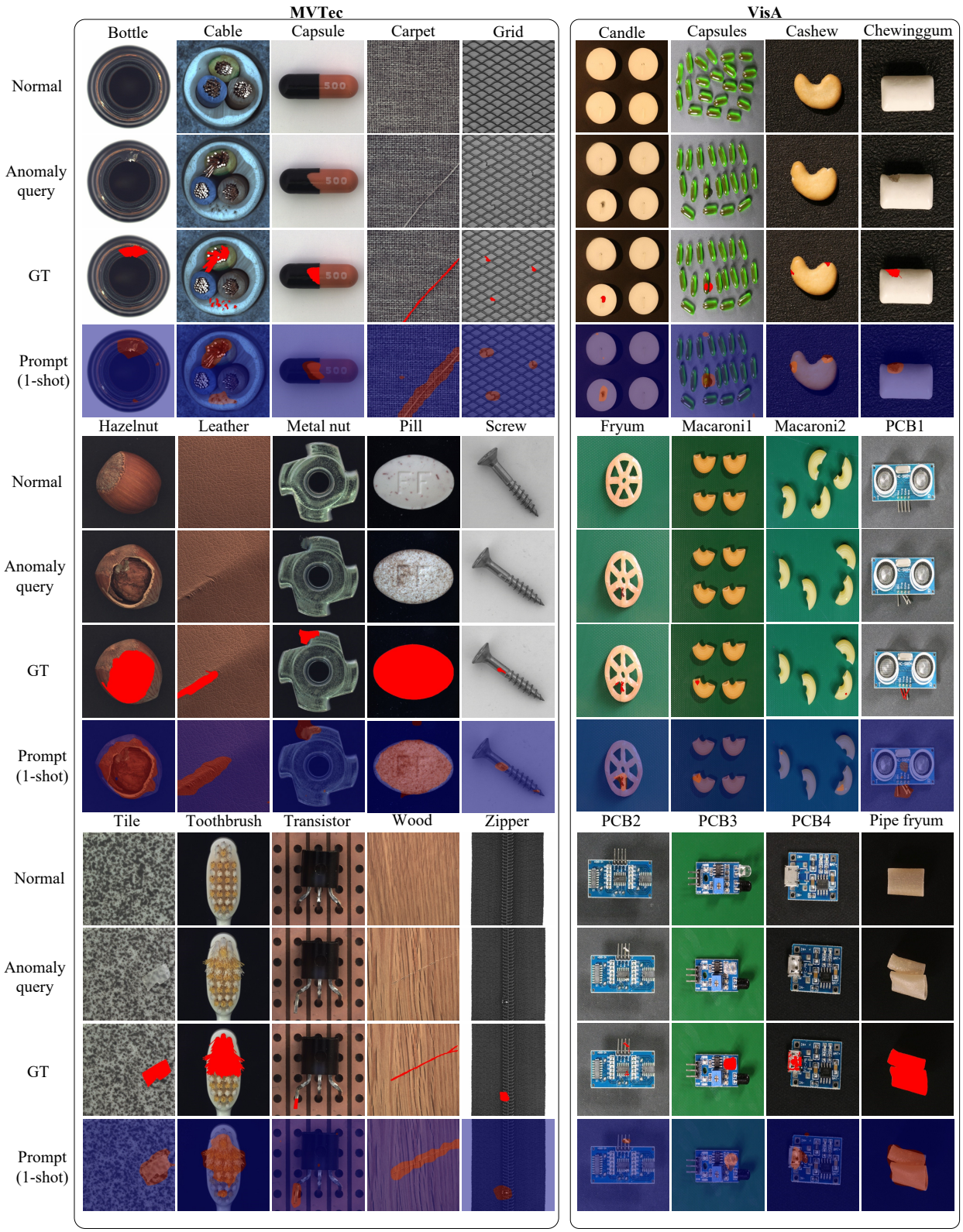

Figure 2. Additional qualitative results from PromptAD (1-shot), tested on MVTec [1] and VisA [14].

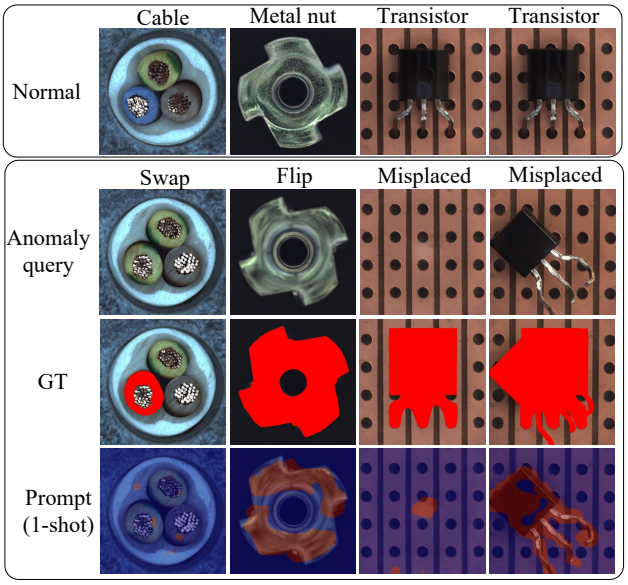

Figure 3. Qualitative results of logical anomaly detection.

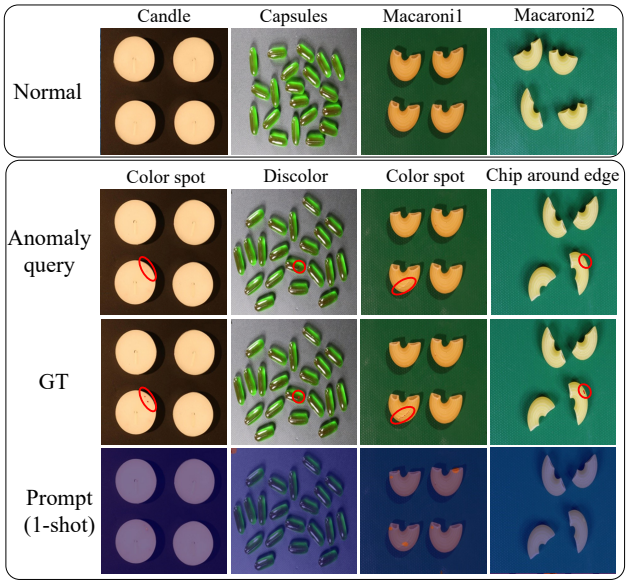

Figure 4. Qualitative results of extremely small anomaly detection.

B. Additional Qualitative Results

In Figure 2, we provide further qualitative results obtained from our (1-shot) PromtAD for pixel-level anomaly detection in MVTec [1] and VisA [14]. It can be seen that PromptAD can accurately locate both large-area surface defects and small-area surface defects. In addition, as shown in Figure 3, we also provide quantitative results of PromptAD on some logical anomalies. Logical anomalies are mainly found in some industrial components in MVTec. It can be seen that PromptAD has poor detection results on the swap anomaly of "cable" and missing anomaly of "transistor", while PromptAD has good localization results on the flip anomaly of "Metal nut" and misplaced anomaly of "Transistor". Figure 4 shows PromptAD's detection results for extremely small anomalies, which are usually hard to detect by humans. For the convenience of viewing, we circle the anomaly positions in red circles, and it can be seen that PromptAD has difficulty in completing the accurate localization of extremely small anomalies.

C. Comparison with Other Prompt Learning Methods

PromptAD is the first prompt design paradigm for one class classification (OCC) promblem, which overcomes the poor performance of other prompt learning methods in anomaly detection. See Table 1, the classical prompt learning methods (1-shot) perform relatively poorly in anomaly detection tasks, and except for the pixel-level results of CoOp [13] on MVTec and VisA, the results of CoOp [13], CoCoOp [12] and Maple [9] are even worse than contextual prompt engineering (CPE). PromptAD not only outperforms the classical prompt learning method in anomaly detection tasks, but also, compared with CPE, brings improvements of 0.8%/5.8% and 5.0%/8.9% on MVTec and VisA respectively (image-level/pixel-level).

	Method	MVTec		VisA	
		image	pixel	image	pixel
0-shot	CPE+WinCLIP	91.8	85.1	78.1	79.6
	CPE+VV-CLIP	90.5	86.7	77.2	82.9
1-shot	CoOp+VV-CLIP	81.5	87.8	72.6	85.5
	CoCoOp+VV-CLIP	60.6	52.1	61.6	72.3
	Maple	66.8	64.9	60.5	61.5
	PromptAD	**91.3**	**92.5**	**83.2**	**91.8**

Table 1. Results (AUROC) of different methods (w/o VAD).

D. Ablation Study

We further evaluate the impact of MAPs and LAPs on PromptAD, respectively. For better comparison, we removed the effect of vision-guided anomaly detection (VAD). Table 2 (second row) shows that without MAPs, the performance degradation is more pronounced, but it is still better than CPE, indicating that PromptAD can still automatically learn effective anomaly prompts even without using any manually annotated anomaly description suffix. Table 2 (third row) shows that there is also a slight decrease in performance without LAPs, which indicates that LAPs can expand the set of exception hints and thus improve the

	Method	MVTec		VisA	
		image	pixel	image	pixel
1	CPE+VV-CLIP	90.5	86.7	77.2	82.9
2	Prompt w/o MAPs	87.6	89.8	78.8	86.2
3	Prompt w/o LAPs	90.5	91.3	82.5	90.3
4	PromptAD	**91.3**	**92.5**	**83.2**	**91.8**

Table 2. Effects (AUROC) of MAPs and LAPs (1-shot & w/o VAD).

Backbone	Dataset	1-shot	2-shot	4-shot
ResNet101	MVTec	85.8/93.0	87.6/94.3	90.3/94.8
ViT-L/14	MVTec	92.4/95.5	93.8/95.8	94.9/96.2
ViT-B/16+	MVTec	**94.6/95.9**	**95.7/96.2**	**96.6/96.5**
ResNet101	VisA	80.4/95.1	84.5/96.3	85.3/96.9
ViT-L/14	VisA	85.2/**96.8**	86.3/**97.2**	86.7/**97.4**
ViT-B/16+	VisA	**86.9**/96.7	**88.3**/97.1	**89.1/97.4**

Table 3. Image-level/pixel-level results (AUROC) with other backbones.

Method	Dataset	1-shot	2-shot	4-shot
PatchCore	MPDD	59.2/78.5	59.6/79.2	79.9/79.8
WinCLIP+	MPDD	68.2/92.6	69.3/94.7	75.2/96.0
PromptAD	MPDD	**80.7/96.2**	**85.3/97.2**	**87.2/97.3**
PatchCore	LOCO	64.9/70.3	65.4/71.5	68.7/72.2
WinCLIP+	LOCO	68.0/71.2	69.7/71.9	71.3/72.8
PromptAD	LOCO	**71.2/73.0**	**72.6/74.1**	**73.5/74.5**

Table 4. Image-level/pixel-level results (AUROC) on MPDD and LOCO.

model performance.

In addition, we explore the impact of CLIP's different visual backbones on PromptAD. The results of using different backbones are recorded in Table 3, where the self-attention modules of VIT add the vv-attention branches, and the attention pooling of ResNet101 adopts V-V attention. All of the backbones require no additional training. The overall performance of ViT is better than ResNet, VIT-L/14 shows better pixel-wise anomaly detection than VIT-B/16+.

E. Results on Other Benchmarks

In addition to the two datasets MVTec [1] and VisA [14], we also evaluate PromptAD in the few-shot setting on MPDD [8] and LOCO [3]. We reproduce the results of PatchCore and WinCLIP+. As shown in Table 4, compared with PatchCore and WinCLIP+, PromptAD achieves the first place in few-shot settings of both datasets

	screw	grid	transistor	mean
6^{th}+9^{th}	91.7	58.8	84.9	92.0
3^{th}+8^{th}	89.6	81.8	87.2	93.2

Table 5. Pixel-level results (AUROC) of using different layer features as the feature memory.

F. Detailed Comparison Results

In this section, we report the detailed subset-level results of PromptAD. In addition, we evaluate PromptAD's results on Image-level AUPR and pixel-level PRO. Specifically, the results for MVTec are recorded in Table 6,7,10,11 and the results for VisA are recorded in Table 8,9,12,13 with the first place marked in bold and the second place mean underlined.

G. Visualization Results of Attention Map

To further analyze the working mechanism of CLIP [10], we provide visualization results of the attention map in the CLIP visual encoder (ViT-B/16+). Figure 5 is the visualization result of the original QK attention map. It can be seen that QK attention in the shallow layers focuses more on local information (main diagonal activations). From layer 5, QK attention starts to focus on more global information. Both global and local information play an important role in anomaly detection. Therefore, to preserve both local and global information, PromptAD uses the features of the 3^{th} and 8^{th} layers instead of the features of the 6^{th} and 9^{th} layers used by PatchCore [11] when storing visual features. The comparison results are reported in Table 5, compared with 6^{th}+9^{th} features, there is a 1.2% improvement using the 3^{th}+8^{th} features. Figure 6 shows the visualization results of the VV attention map, it can be seen that compared with QK attention, VV attention focuses on local information from the first layer to the last layer, which is more conducive to completing the localization task. While, most layers of QK attention focus on global information, which is more conducive to classification tasks.

References

[1] 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.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[3] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sat-

MVTec Image-AUROC	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Bottle	98.7±0.6	97.4±0.7	99.4±0.4	98.2±0.9	99.8±0.3	99.5±0.1	98.5±1.0	99.2±0.3	99.3±0.3	100.0±0.1	99.5±0.2	98.8±0.2	99.2±0.3	99.3±0.4	100.0±0.2
Cable	71.2±3.3	57.7±4.6	88.8±4.2	88.9±1.9	94.2±1.2	76.2±5.2	62.3±5.9	91.0±2.7	88.4±0.7	99.9±1.5	83.4±3.1	70.0±6.1	91.0±2.7	90.9±0.9	98.8±0.9
Capsule	70.2±3.0	57.7±7.3	67.8±2.9	72.3±6.8	84.6±6.7	70.9±6.1	64.3±3.0	72.8±7.0	77.3±8.8	100.0±7.7	78.9±5.5	65.2±2.5	72.8±7.0	82.3±8.9	100.0±1.5
Garpet	98.1±0.2	96.6±1.0	95.3±0.8	99.8±0.3	100.0±0.0	98.3±0.4	97.8±0.5	96.6±0.5	99.8±0.3	100.0±0.0	98.6±0.2	97.9±0.4	96.6±0.5	100.0±0.0	100.0±0.0
Grid	40.0±6.8	54.2±6.7	63.6±10.3	99.5±0.3	99.8±0.9	41.3±3.6	67.2±4.2	67.7±8.3	99.4±0.2	98.7±1.2	44.6±6.6	68.1±3.8	67.7±8.3	99.6±0.1	98.6±0.5
Hazelnut	95.8±1.3	88.3±2.6	88.3±2.7	97.5±1.4	99.8±0.8	96.2±2.1	90.8±0.8	93.2±3.8	98.3±0.7	99.4±0.6	98.4±1.3	91.9±1.2	93.2±3.8	98.4±0.4	99.8±0.2
Leather	100.0±0.0	97.5±0.7	97.3±0.7	99.9±0.0	100.0±0.0	100.0±0.0	97.5±0.9	97.9±0.7	99.9±0.0	91.8±0.1	100.0±0.0	98.5±0.2	97.9±0.7	100.0±0.0	95.4±0.1
Metal nut	71.0±2.2	53.0±3.8	73.4±2.9	98.7±0.8	99.1±0.5	77.0±7.9	54.8±3.8	77.7±8.5	99.4±0.2	91.3±0.7	77.8±5.7	60.7±5.2	77.7±8.5	99.5±0.2	91.5±0.1
Pill	86.5±3.1	61.3±3.8	81.9±2.8	91.2±2.1	92.6±1.5	84.8±0.9	59.1±6.4	82.9±2.9	92.3±0.7	100.0±0.8	86.7±0.3	54.9±2.7	82.9±2.9	92.8±1.0	99.8±0.8
Screw	46.7±2.5	55.0±2.5	44.4±4.6	86.4±0.9	65.0±2.9	46.6±2.2	54.0±4.4	49.0±3.8	86.0±2.1	98.6±2.0	50.5±5.4	50.0±4.1	49.0±3.8	87.9±1.2	100.0±2.3
Tile	99.9±0.1	92.2±2.2	99.0±0.9	99.9±0.0	100.0±0.0	99.9±0.1	93.3±1.1	98.5±1.0	99.9±0.2	93.6±0.1	100.0±0.0	93.1±0.6	98.5±1.0	99.9±0.1	92.9±0.1
Toothbursh	71.7±2.6	82.5±1.2	83.3±3.8	92.2±4.9	98.9±1.0	78.6±3.2	87.6±4.2	85.9±3.5	97.5±1.6	71.0±1.5	78.8±5.2	89.2±2.5	85.9±3.5	96.7±2.6	83.6±0.6
Transistor	77.2±2.0	73.3±6.0	78.1±6.9	83.4±3.8	94.0±6.5	81.3±3.7	72.8±6.3	90.0±4.3	85.3±1.7	97.5±4.0	81.4±2.1	82.4±6.5	90.0±4.3	85.7±2.5	98.1±3.8
Wood	98.8±0.3	96.1±1.2	97.8±0.3	99.9±0.1	97.9±0.3	99.2±0.4	96.9±0.5	98.3±0.6	99.9±0.1	97.4±0.1	98.9±0.6	97.0±0.2	98.3±0.6	99.8±0.3	95.6±0.3
Zipper	89.3±1.9	85.8±2.7	92.3±0.5	88.8±5.9	93.9±3.2	93.3±2.9	86.3±2.6	94.0±2.1	94.0±1.4	95.8±1.7	95.1±1.3	88.3±2.0	94.0±2.1	94.5±0.5	95.0±2.3
Mean	81.0±2.0	76.6±3.1	83.4±3.0	93.1±2.0	**94.6±1.7**	82.9±2.6	78.9±3.1	86.3±3.3	94.4±1.3	**95.7±1.5**	84.8±2.5	80.4±2.5	88.8±2.6	95.2±1.3	**96.6±0.9**

Table 6. Comparison of image-level anomaly detection in terms of subset-wise AUROC on MVTec.

MVTec Pixel-AUROC	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Bottle	95.3±0.2	96.1±0.5	97.9±0.1	97.5±0.2	99.6±0.2	95.7±0.2	96.9±0.1	98.1±0.0	97.7±0.1	99.5±0.1	96.1±0.0	97.1±0.1	98.2±0.0	97.8±0.0	99.5±0.1
Cable	86.4±0.2	88.4±1.2	95.5±0.8	93.8±0.6	98.4±0.5	87.4±0.3	90.0±0.8	96.4±0.3	94.3±0.4	97.6±0.3	88.2±0.2	92.1±0.4	97.5±0.3	94.9±0.1	98.2±0.2
Capsule	96.3±0.2	94.5±0.6	95.6±0.4	94.6±0.8	99.5±0.7	96.7±0.1	95.2±0.5	96.5±0.4	96.4±0.3	99.3±0.5	97.0±0.2	96.2±0.4	96.8±0.6	96.2±0.5	99.3±0.3
Garpet	98.2±0.0	97.8±0.2	98.4±0.1	99.4±0.0	95.9±0.0	98.3±0.0	98.2±0.0	98.5±0.1	99.3±0.0	96.1±0.0	98.4±0.0	98.4±0.0	98.6±0.1	99.3±0.0	96.2±0.0
Grid	80.7±1.3	70.2±2.8	58.8±4.9	96.8±1.0	95.1±0.5	83.5±1.0	70.8±2.0	62.6±3.2	97.7±0.8	95.5±0.4	87.2±1.1	77.0±1.8	69.4±1.3	98.0±0.2	95.2±0.3
Hazelnut	97.2±0.1	95.4±0.7	95.8±0.6	98.5±0.2	97.3±0.4	97.6±0.1	96.8±0.3	96.3±0.6	98.7±0.1	97.6±0.3	97.7±0.1	97.2±0.2	97.6±0.1	98.8±0.0	97.9±0.2
Leather	99.1±0.0	98.5±0.1	98.8±0.2	99.3±0.0	93.3±0.0	99.1±0.0	98.7±0.1	99.0±0.1	99.3±0.0	93.2±0.0	99.1±0.0	98.8±0.0	99.1±0.0	99.3±0.0	93.9±0.0
Metal nut	83.8±0.7	74.6±1.1	89.3±1.4	90.0±0.6	97.3±0.7	85.8±1.1	80.3±2.1	94.6±1.4	91.4±0.4	97.4±0.7	87.1±0.7	82.7±3.9	95.9±1.8	92.9±0.4	97.7±0.5
Pill	89.4±0.4	84.8±1.0	93.1±1.1	96.4±0.3	98.4±0.4	89.9±0.2	87.3±0.7	94.2±0.3	97.0±0.2	98.5±0.3	90.7±0.2	88.9±0.5	94.8±0.4	97.1±0.0	98.5±0.2
Screw	94.8±0.2	83.3±0.7	89.6±0.5	94.5±0.4	91.4±0.5	95.6±0.4	89.8±0.8	90.0±0.7	95.2±0.3	95.1±0.7	96.4±0.4	90.8±0.2	91.3±1.0	96.0±0.5	94.9±0.8
Tile	91.7±0.3	84.1±1.1	94.1±0.5	96.3±0.2	92.8±0.1	92.0±0.1	87.7±0.2	94.4±0.2	96.5±0.1	94.1±0.1	92.2±0.1	88.9±0.3	94.6±0.1	96.6±0.1	94.1±0.1
Toothbursh	94.6±0.6	97.3±0.3	97.3±0.4	97.8±0.1	94.0±0.2	96.2±0.3	97.7±0.3	97.5±0.2	98.1±0.1	95.5±0.1	97.0±0.6	98.4±0.2	98.4±0.4	98.4±0.5	96.2±0.0
Transistor	71.4±1.3	90.2±2.8	84.9±2.7	85.0±1.8	99.1±1.9	72.8±0.9	92.3±2.1	89.6±0.9	88.3±1.0	99.0±1.1	73.4±0.7	94.0±2.7	90.7±1.4	88.5±1.2	99.0±0.6
Wood	93.4±0.1	90.7±0.4	92.7±0.9	94.6±1.0	89.4±0.3	93.8±0.1	91.9±0.1	93.2±0.7	95.3±0.4	89.1±0.2	93.9±0.1	92.2±0.1	93.5±0.3	95.4±0.2	90.6±0.1
Zipper	94.9±0.3	93.9±0.8	97.4±0.4	93.9±0.8	96.6±0.5	95.8±0.2	95.4±0.3	98.0±0.1	94.1±0.7	95.5±0.4	96.2±0.1	96.1±0.2	98.1±0.1	94.2±0.4	96.6±0.3
Mean	91.2±0.4	89.3±0.9	92.0±1.0	95.2±0.5	**95.9±0.5**	92.0±0.3	91.3±0.7	93.3±0.6	96.0±0.3	**96.2±0.3**	92.7±0.3	92.6±0.7	94.3±0.5	96.2±0.3	**96.5±0.2**

Table 7. Comparison of pixel-level anomaly detection in terms of subset-wise AUROC on MVTec.

VisA Image-AUROC	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Candle	86.1±5.6	70.8±4.1	85.1±1.4	93.4±1.4	90.3±2.7	91.3±3.3	75.8±2.1	85.3±1.5	94.8±1.0	91.0±1.2	92.8±2.1	77.5±1.6	87.8±0.8	95.1±0.3	93.0±1.2
Capsules	73.3±7.5	51.0±7.8	60.0±7.6	85.0±3.1	84.5±3.5	71.7±11.2	51.7±4.6	57.8±5.4	84.9±0.8	84.9±3.7	73.4±7.1	52.7±3.4	63.4±5.4	86.8±1.7	80.6±2.1
Cashew	95.9±1.1	62.3±9.9	89.5±4.4	94.0±0.4	95.6±1.0	97.3±1.4	74.6±3.6	93.6±0.6	94.3±0.5	94.7±1.4	96.4±1.3	77.7±3.2	93.0±1.5	95.2±0.8	93.6±2.2
Chewinggum	92.1±2.0	69.9±4.9	97.3±0.3	97.6±0.8	96.4±1.2	93.4±1.0	82.7±2.1	97.8±0.6	97.3±0.8	96.6±0.6	93.5±1.4	83.5±3.7	98.3±0.3	97.7±0.3	96.8±0.4
Fryum	81.1±4.0	58.3±5.9	75.0±4.8	88.5±1.9	90.3±1.8	90.5±3.9	69.2±9.0	83.4±2.4	90.5±0.4	89.2±0.9	92.9±1.6	71.2±5.9	88.6±1.3	90.8±0.5	89.0±2.3
Macaroni1	66.0±10.5	62.1±4.6	68.0±3.4	82.9±1.5	88.6±3.1	69.1±8.2	62.2±5.0	75.6±4.6	83.3±1.9	84.2±2.5	65.8±1.2	65.9±3.9	82.9±2.7	85.2±0.9	88.2±2.5
Macaroni2	55.8±6.1	47.5±5.9	55.6±4.6	70.2±0.9	69.1±3.0	58.3±4.4	50.8±2.9	57.3±5.6	71.8±2.0	82.6±0.9	56.7±3.2	55.0±2.9	61.7±1.8	70.9±2.2	81.2±1.8
PCB1	87.2±2.3	76.2±1.2	78.9±1.1	75.6±23.0	88.7±0.7	86.7±1.1	62.4±10.8	71.5±20.0	76.7±5.2	90.9±5.4	83.4±8.5	82.6±1.5	84.7±6.7	88.3±1.7	90.9±2.5
PCB2	73.5±3.7	61.2±2.0	81.5±0.8	62.2±3.9	71.6±4.1	70.3±8.1	66.8±2.0	84.3±1.7	62.6±3.7	73.0±3.0	71.7±7.0	73.5±2.4	84.3±1.0	67.5±2.6	78.6±2.5
PCB3	72.2±1.0	51.4±12.2	82.7±2.3	74.1±1.1	79.1±3.6	75.8±5.7	67.3±3.8	84.8±1.2	78.8±1.9	76.2±2.2	79.0±4.1	65.9±1.9	87.0±1.1	83.3±1.7	80.3±1.7
PCB4	93.4±1.3	76.1±3.6	93.9±2.8	85.2±8.9	91.4±3.2	86.1±8.2	69.3±13.7	94.3±3.2	82.3±9.9	97.5±2.4	95.4±2.3	85.4±2.0	95.6±1.6	87.6±8.0	97.8±1.4
Pipe fryum	77.9±3.2	66.7±2.2	90.7±1.7	97.2±1.1	96.9±0.2	78.1±3.0	75.3±1.8	93.5±1.3	98.0±0.6	98.9±0.3	79.3±0.9	82.9±2.2	96.4±0.7	98.5±0.4	98.6±0.2
Mean	79.5±4.0	62.8±5.4	79.9±2.9	83.8±4.0	**86.9±2.3**	80.7±5.0	67.4±5.1	81.6±4.0	84.6±2.4	**88.3±2.0**	81.7±3.4	72.8±2.9	85.3±2.1	87.3±1.8	**89.1±1.7**

Table 8. Comparison of image-level anomaly detection in terms of subset-wise AUROC on VisA.

VisA Pixel-AUROC	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Candle	97.9±0.3	91.7±2.2	97.2±0.2	97.4±0.2	95.8±0.2	98.1±0.2	94.9±0.8	97.7±0.3	97.7±0.1	95.9±0.1	98.2±0.1	95.4±0.2	97.9±0.1	97.8±0.2	96.0±0.1
Capsules	95.5±0.5	70.9±1.1	93.2±0.9	96.4±0.6	95.4±0.9	96.5±0.9	75.7±1.7	94.0±0.2	96.8±0.3	96.1±0.7	97.7±0.1	79.1±0.7	94.8±0.5	97.1±0.2	96.8±0.6
Cashew	95.9±0.5	95.5±0.6	98.1±0.1	98.5±0.2	99.1±0.2	95.9±0.4	96.4±0.4	98.2±0.2	98.5±0.1	99.2±0.1	95.9±0.3	97.2±0.3	98.3±0.2	98.7±0.0	99.2±0.1
Chewinggum	96.0±0.4	90.1±0.4	96.9±0.3	98.6±0.1	99.1±0.1	96.0±0.3	93.1±0.7	96.6±0.1	98.6±0.1	99.2±0.1	95.7±0.3	94.4±0.5	96.8±0.1	98.5±0.1	99.2±0.2
Fryum	93.5±0.3	93.3±0.6	93.3±0.5	96.4±0.3	95.4±0.3	93.9±0.2	94.1±0.6	94.0±0.3	97.0±0.2	96.4±0.2	94.4±0.1	95.0±0.4	94.2±0.2	97.1±0.1	96.6±0.2
Macaroni1	97.9±0.2	89.4±0.9	95.2±0.4	96.4±0.6	97.8±0.1	98.5±0.2	91.7±0.3	96.0±1.3	96.5±0.7	98.3±0.1	98.8±0.1	93.5±0.5	97.0±0.3	97.0±0.2	98.2±0.1
Macaroni2	94.1±1.0	86.4±1.1	89.1±1.6	96.8±0.4	96.6±0.5	95.2±0.4	90.1±0.8	90.2±1.9	96.8±0.6	97.2±0.3	96.4±0.2	90.2±0.3	93.9±0.3	97.3±0.3	97.0±0.3
PCB1	94.7±0.4	89.9±0.3	96.1±1.5	96.6±0.6	96.6±0.8	96.5±1.5	90.6±0.6	97.6±0.9	97.0±0.9	96.9±0.4	96.8±1.5	93.2±1.5	98.1±1.0	98.1±0.9	98.2±0.3
PCB2	95.1±0.2	90.9±1.4	95.4±0.2	93.0±0.4	93.5±0.9	95.7±0.1	93.9±0.9	96.0±0.3	93.9±0.2	94.8±0.8	96.3±0.0	93.7±1.0	96.6±0.2	94.6±0.4	95.3±0.5
PCB3	96.0±0.1	93.9±0.3	96.2±0.3	94.3±0.3	95.9±0.5	96.6±0.1	95.1±0.5	97.1±0.1	95.1±0.2	96.1±0.4	96.9±0.0	95.7±0.1	97.4±0.2	95.8±0.1	96.8±0.2
PCB4	92.0±0.6	89.6±0.6	95.6±0.6	94.0±0.9	95.5±0.5	92.8±0.3	90.7±0.9	96.2±0.4	95.6±0.3	95.6±0.3	94.1±0.2	92.1±0.5	97.0±0.2	96.1±0.3	96.2±0.4
Pipe fryum	98.4±0.2	97.2±0.6	98.8±0.2	98.3±0.2	99.1±0.2	98.7±0.1	98.1±0.4	99.1±0.1	98.5±0.2	99.4±0.2	98.8±0.0	98.5±0.1	99.1±0.0	98.7±0.1	99.3±0.3
Mean	95.6±0.4	89.9±0.8	95.4±0.6	96.4±0.4	**96.7±0.4**	96.2±0.4	92.0±0.7	96.1±0.5	96.8±0.3	**97.1±0.3**	96.6±0.3	93.2±0.5	96.8±0.3	97.2±0.2	**97.4±0.3**

Table 9. Comparison of pixel-level anomaly detection in terms of subset-wise AUROC on VisA.

MVTec Image-level AUPR	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Bottle	99.6±0.1	99.2±0.2	99.8±0.1	99.4±0.3	99.8±0.1	99.8±0.0	99.6±0.3	99.8±0.1	99.8±0.1	99.9±0.1	99.9±0.0	99.7±0.0	99.8±0.1	99.8±0.1	100.0±0.1
Cable	79.6±2.3	64.9±3.8	93.8±2.2	93.2±1.1	95.5±0.7	84.5±3.1	69.6±6.6	95.1±1.3	92.9±0.6	96.9±0.7	88.8±1.9	76.1±5.6	97.1±0.7	94.4±0.3	97.4±0.5
Capsule	91.2±0.9	86.9±2.2	89.4±2.0	91.6±2.7	97.8±3.1	91.6±2.1	88.4±0.8	91.0±2.9	93.3±3.6	97.0±3.0	94.4±1.9	87.8±0.8	94.9±1.1	95.1±3.3	98.6±2.2
Carpet	99.4±0.0	99.0±0.2	98.7±0.2	99.9±0.1	100.0±0.0	99.5±0.1	99.4±0.1	99.0±0.1	99.9±0.1	100.0±0.0	99.6±0.1	99.4±0.1	98.8±0.2	100.0±0.0	100.0±0.0
Grid	66.9±2.1	75.0±3.3	81.1±4.9	99.9±0.1	98.8±0.3	68.3±2.1	82.5±2.3	84.1±4.0	99.8±0.1	99.9±0.3	68.8±4.2	83.0±1.8	86.4±4.0	99.9±0.0	99.7±0.1
Hazelnut	97.9±0.6	93.3±1.7	92.9±2.2	98.6±0.7	99.7±0.3	98.0±1.1	94.1±0.5	96.0±2.0	99.1±0.4	99.8±0.2	99.1±0.7	94.8±0.6	97.0±1.2	99.1±0.2	99.9±0.2
Leather	100.0±0.0	99.2±0.2	99.1±0.2	100.0±0.0	100.0±0.0	100.0±0.0	99.2±0.3	99.3±0.2	100.0±0.0	100.0±0.0	100.0±0.0	99.6±0.1	99.6±0.1	100.0±0.0	100.0±0.0
Metal nut	91.7±0.8	82.0±2.7	91.0±1.1	99.7±0.2	99.6±0.1	93.7±2.4	82.2±1.4	92.3±4.0	99.9±0.0	100.0±0.1	94.1±1.8	85.5±1.7	97.0±2.6	99.9±0.1	99.9±0.0
Pill	97.0±0.8	88.3±1.3	96.5±0.6	98.3±0.5	98.5±0.3	96.5±0.4	87.9±2.6	96.6±0.7	98.6±0.1	97.8±0.2	97.0±0.2	87.0±1.2	96.9±0.4	98.6±0.2	98.5±0.1
Screw	71.3±1.8	78.1±1.0	71.4±2.3	94.2±0.6	78.5±3.0	71.0±1.4	77.3±1.3	72.9±3.4	94.1±1.5	86.7±1.9	73.7±2.4	75.7±2.8	71.8±1.9	94.9±0.8	93.8±2.1
Tile	100.0±0.0	97.2±0.7	99.6±0.3	100.0±0.0	100.0±0.0	100.0±0.0	97.6±0.4	99.4±0.4	100.0±0.1	100.0±0.0	100.0±0.0	97.6±0.2	99.6±0.1	100.0±0.0	100.0±0.0
Toothbursh	88.3±0.6	93.7±0.5	93.5±1.4	96.7±2.0	98.9±0.4	90.8±1.3	95.2±1.6	94.1±1.4	99.0±0.6	99.3±0.4	91.3±2.6	95.8±0.7	94.8±0.7	98.7±1.1	99.7±0.1
Transistor	76.2±1.7	66.2±7.5	77.7±5.5	79.0±4.0	91.2±5.9	81.6±3.4	69.0±6.5	89.3±3.9	80.7±2.3	92.2±2.9	80.3±2.6	77.6±8.4	84.5±9.0	80.7±3.2	92.2±1.2
Wood	99.6±0.1	98.8±0.3	99.3±0.1	100.0±0.0	99.6±0.3	99.7±0.1	99.0±0.1	99.5±0.2	100.0±0.0	99.7±0.1	99.7±0.2	99.1±0.0	99.5±0.2	99.9±0.1	99.5±0.1
Zipper	96.9±0.5	95.5±0.9	97.2±0.3	96.8±1.8	99.0±0.6	98.2±0.8	95.4±1.0	97.8±1.0	98.3±0.4	99.3±0.5	98.6±0.4	96.2±0.8	99.1±0.7	98.5±0.2	98.5±0.3
Mean	90.6±0.8	88.1±1.7	92.2±1.5	96.5±0.9	**97.1±1.0**	91.7±1.2	89.3±1.7	93.8±1.7	97.0±0.7	**97.9±0.7**	92.5±1.2	90.5±1.6	94.5±1.5	97.3±0.6	**98.5±0.5**

Table 10. Comparison of image-level anomaly detection in terms of subset-wise AUPR on MVTec.

MVTec Pixel-PRO	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Bottle	91.1±0.4	89.8±0.8	93.5±0.3	91.2±0.4	93.6±0.1	91.8±0.5	91.7±0.2	93.9±0.3	91.8±0.3	93.9±0.2	92.5±0.1	92.2±0.2	94.0±0.2	91.6±0.2	94.5±0.2
Cable	63.5±0.7	59.1±3.2	84.7±1.0	72.5±2.3	87.3±1.2	66.7±0.9	66.5±2.8	88.5±0.9	74.7±2.3	87.8±0.7	69.5±0.4	74.2±1.8	91.7±0.6	77.0±1.1	88.9±0.3
Capsule	92.7±0.4	80.0±2.0	83.9±0.9	85.6±2.7	80.1±1.7	93.4±0.3	82.3±2.1	86.6±1.0	90.6±0.6	79.2±2.2	94.1±0.6	85.7±1.3	87.8±1.9	90.1±1.5	88.7±2.0
Garpet	96.1±0.0	92.9±0.3	93.3±0.3	97.4±0.4	98.3±0.3	96.2±0.0	93.9±0.2	93.7±0.4	97.3±0.3	98.2±0.3	96.3±0.0	94.4±0.2	93.9±0.4	97.0±0.2	98.2±0.1
Grid	67.7±1.9	41.2±4.6	21.7±9.5	90.5±2.7	94.3±1.0	72.1±1.5	45.1±3.6	23.7±3.8	92.8±2.5	95.0±0.8	78.0±1.5	55.5±3.4	30.4±4.6	93.6±0.6	93.8±0.7
Hazelnut	94.9±0.3	85.7±1.9	88.3±1.3	93.7±0.9	92.9±0.5	95.6±0.2	89.4±0.9	89.8±1.3	94.2±0.3	93.4±0.5	95.6±0.1	90.4±0.7	92.0±0.3	94.2±0.3	95.2±0.5
Leather	98.7±0.0	95.6±0.2	95.2±1.0	98.6±0.0	98.7±0.5	98.8±0.0	96.2±0.2	95.9±0.3	98.3±0.4	98.7±0.4	98.8±0.0	96.3±0.1	96.4±0.1	98.0±0.4	98.4±0.5
Metal nut	73.4±1.1	38.1±1.6	66.7±2.9	84.7±1.1	83.1±0.6	78.1±1.8	48.2±5.0	79.6±4.2	86.7±0.8	87.7±1.1	81.2±1.4	54.0±8.8	83.8±5.5	89.4±0.1	87.6±0.3
Pill	92.8±0.3	78.9±0.6	89.5±1.6	93.5±0.2	90.8±0.4	93.3±0.2	84.3±0.4	91.6±0.5	94.5±0.2	90.5±0.4	93.9±0.2	86.6±0.4	92.5±0.4	94.6±0.3	92.0±0.1
Screw	85.0±0.8	51.6±1.7	68.1±1.3	82.3±1.1	78.1±1.7	87.2±1.2	69.5±2.1	69.0±2.1	84.1±0.5	74.7±1.4	89.5±1.3	72.3±0.8	72.4±3.1	86.3±1.8	86.7±2.5
Tile	84.2±0.4	66.7±1.5	82.5±1.1	89.4±0.4	90.7±0.3	84.6±0.2	71.9±0.5	82.5±0.5	89.6±0.4	90.9±0.2	84.9±0.1	73.6±0.9	83.0±0.1	89.9±0.3	90.9±0.2
Toothbursh	83.5±1.3	82.1±1.5	79.0±2.4	85.3±1.0	90.1±0.5	87.4±1.1	83.3±2.6	81.0±0.7	84.7±1.4	91.6±0.5	89.0±1.1	87.1±1.7	85.5±3.0	86.0±3.3	91.3±0.2
Transistor	55.3±2.0	70.3±7.0	70.9±4.6	65.0±1.8	67.5±3.7	57.6±1.4	76.5±5.5	78.8±1.5	68.6±1.1	68.1±2.1	58.5±0.7	82.2±7.4	79.5±2.8	69.0±1.1	73.0±1.2
Wood	92.9±0.1	86.5±0.6	87.1±1.0	91.0±0.6	92.4±0.8	93.1±0.1	88.0±0.2	86.8±1.4	91.8±0.6	91.6±0.6	93.2±0.1	88.4±0.2	87.7±0.4	91.7±0.3	91.4±0.5
Zipper	86.8±0.6	81.7±2.0	91.2±1.1	86.0±1.7	81.0±1.4	89.0±0.4	85.6±0.7	92.8±0.4	86.4±1.6	86.4±0.5	90.1±0.2	87.2±0.8	93.4±0.2	86.9±0.7	87.5±0.6
Mean	83.9±0.7	73.3±2.0	79.7±2.0	87.1±1.2	**87.9±1.0**	85.7±0.7	78.2±1.8	82.3±1.3	88.4±0.9	**88.5±0.8**	87.0±0.5	81.3±1.9	84.3±1.6	89.0±0.8	**90.5±0.7**

Table 11. Comparison of pixel-level anomaly detection in terms of subset-wise PRO on MVTec.

VisA Image-level AUPR	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Candle	86.5±4.3	69.2±3.9	86.6±2.3	93.6±1.5	93.7±2.9	90.7±3.2	72.8±1.0	86.8±1.7	95.1±1.1	93.6±1.1	92.6±1.9	72.5±1.1	88.9±1.1	95.3±0.4	92.9±1.1
Capsules	79.4±4.9	63.4±5.7	72.3±5.3	89.9±2.5	90.1±1.9	79.9±5.8	63.4±2.0	73.6±4.7	88.9±0.7	88.3±2.6	81.1±4.5	63.0±2.3	78.4±3.1	91.5±1.4	89.8±1.0
Cashew	97.9±0.4	78.2±5.7	94.6±2.0	97.2±0.2	97.6±0.7	98.6±0.6	86.1±2.2	96.9±0.3	97.3±0.2	97.4±0.5	98.3±0.6	88.4±2.0	96.5±0.7	97.7±0.4	97.0±1.1
Chewinggum	96.4±0.9	79.8±3.6	98.9±0.1	99.0±0.3	99.1±0.7	97.1±0.4	89.5±1.9	99.1±0.2	98.9±0.3	98.4±0.2	97.1±0.6	88.5±3.2	99.3±0.1	99.0±0.1	98.5±0.3
Fryum	89.8±1.8	74.5±2.9	87.6±2.4	94.7±1.0	93.8±1.0	94.5±2.3	81.0±5.4	92.1±1.3	95.8±0.2	96.0±0.7	95.8±1.0	81.5±3.0	95.0±0.6	96.0±0.3	93.6±0.5
Macaroni1	61.9±11.2	60.4±2.9	67.8±3.4	84.9±1.2	86.3±1.9	64.5±9.5	63.1±4.3	74.9±5.2	84.7±1.5	91.1±1.7	60.2±2.7	64.9±2.1	82.1±3.5	86.5±0.6	89.2±1.3
Macaroni2	52.7±4.2	51.7±5.0	54.9±3.2	68.4±1.8	72.5±2.5	55.9±3.1	52.7±1.5	57.2±2.6	70.4±1.8	84.7±1.6	51.9±2.3	54.9±2.5	60.2±3.0	69.6±2.8	82.2±1.0
PCB1	84.9±3.7	68.6±2.4	72.1±2.5	76.5±19.0	88.0±11.3	83.8±2.1	60.4±7.7	72.6±16.4	78.3±4.3	80.9±6.3	83.2±7.2	77.4±2.9	81.0±9.2	87.7±1.7	90.1±3.6
PCB2	74.9±2.9	63.3±1.2	84.4±0.4	64.9±3.3	75.4±2.7	71.7±6.6	68.9±2.6	86.6±1.1	65.8±4.0	73.0±4.8	74.2±5.0	75.0±1.7	86.2±1.0	71.3±3.4	75.3±2.5
PCB3	75.5±2.1	52.3±10.8	84.6±1.5	73.5±1.6	75.2±3.8	78.3±5.2	65.2±3.8	86.1±0.5	80.9±1.6	82.8±2.2	81.0±3.6	64.5±2.4	88.3±1.1	84.8±1.8	83.5±1.6
PCB4	92.9±1.6	74.7±2.6	92.8±3.1	78.5±15.5	90.5±1.2	81.9±11.2	67.6±11.9	93.2±3.4	72.5±16.2	94.5±2.9	94.8±2.9	84.0±2.0	94.9±1.2	85.6±8.9	97.5±1.3
Pipe fryum	88.3±2.0	79.2±1.5	95.4±0.6	98.6±0.5	98.3±0.1	88.1±1.7	84.5±1.7	96.8±0.7	99.0±0.3	99.1±0.1	88.8±1.0	89.8±1.7	98.3±0.3	99.2±0.2	99.3±0.1
Mean	82.0±3.3	68.3±4.0	82.8±2.3	85.1±4.0	**88.4±2.6**	82.3±4.3	71.6±3.8	84.8±3.2	85.8±2.7	**90.0±2.1**	83.4±2.7	75.6±2.2	87.5±2.1	88.8±1.8	**90.8±1.3**

Table 12. Comparison of image-level anomaly detection in terms of subset-wise AUPR on VisA.

VisA Pixel-PRO	1-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	2-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD	4-shot SPADE	PaDiM	PatchCore	WinCLIP+	PromptAD
Candle	95.6±0.5	81.5±5.3	92.6±0.4	94.0±0.4	91.8±1.2	95.6±0.4	87.3±1.2	93.4±0.6	94.2±0.2	91.6±0.7	95.7±0.1	88.3±0.7	94.1±0.4	94.4±0.2	90.6±0.5
Capsules	83.1±1.1	30.6±1.1	66.6±4.5	73.6±3.5	70.0±1.6	85.4±3.1	38.4±3.7	67.9±2.3	75.9±1.9	70.8±2.5	89.0±1.2	43.3±2.0	69.0±3.2	77.0±1.4	72.4±3.2
Cashew	89.8±1.1	73.4±2.1	90.8±0.2	91.1±0.8	92.3±0.5	90.4±0.5	78.4±2.7	91.4±1.0	90.4±0.6	92.7±1.8	90.4±0.6	81.2±2.8	92.1±0.3	91.3±0.9	92.8±2.0
Chewinggum	73.9±1.2	58.1±0.6	78.2±1.3	91.0±0.5	89.8±1.3	73.8±1.1	63.7±2.4	78.0±0.4	90.9±0.7	87.8±1.3	72.7±0.9	67.2±1.8	79.3±0.8	91.0±0.4	89.4±0.6
Fryum	83.7±1.2	71.1±1.6	78.7±2.3	89.1±1.0	83.5±3.1	84.5±0.9	71.2±0.8	81.4±2.8	89.3±0.2	86.2±3.1	86.2±0.9	73.2±1.3	81.0±1.2	89.7±0.5	80.3±0.7
Macaroni1	92.0±0.6	62.2±4.4	83.4±1.3	84.6±2.3	87.5±1.9	93.9±0.8	71.8±2.4	86.2±4.6	85.2±1.4	90.6±1.5	95.1±0.4	76.6±2.1	89.6±0.7	86.8±0.8	91.5±1.3
Macaroni2	80.0±3.3	54.9±3.6	66.0±3.0	89.3±2.4	80.6±1.5	81.7±1.5	65.6±3.4	67.2±6.5	88.6±1.7	82.7±1.0	86.0±0.8	65.9±1.5	78.3±0.9	90.5±1.3	87.2±0.6
PCB1	81.3±5.7	63.9±1.8	79.0±10.7	82.5±6.0	89.2±13.1	87.2±2.3	68.4±4.1	86.1±1.7	83.8±5.0	90.2±7.5	88.0±2.7	70.2±3.3	88.1±2.6	87.9±2.1	90.2±6.0
PCB2	83.7±0.6	64.4±3.8	80.9±0.5	73.6±1.5	79.3±2.0	85.5±1.0	72.9±3.4	82.9±1.8	76.2±0.9	79.3±1.9	87.0±0.5	71.9±2.6	83.7±1.0	78.0±1.3	76.3±1.6
PCB3	84.3±1.0	69.0±1.2	78.1±2.0	79.5±2.5	84.0±1.4	86.1±0.6	74.0±2.3	82.2±1.1	82.3±1.8	84.7±1.1	87.7±0.6	77.2±0.8	84.4±1.9	84.2±1.0	85.0±1.3
PCB4	66.9±2.0	59.1±1.8	77.9±3.1	76.6±4.1	78.8±2.3	69.3±1.1	62.6±3.6	79.5±4.8	81.7±1.2	78.3±2.6	74.7±1.0	67.9±2.6	83.5±2.5	84.2±0.7	83.4±2.4
Pipe fryum	94.3±0.5	83.9±0.8	93.6±0.5	96.1±0.6	95.2±0.4	95.0±0.2	86.9±0.9	94.5±0.4	96.2±0.6	94.8±0.5	95.0±0.3	88.7±1.3	95.0±0.5	96.6±0.2	95.3±0.3
Mean	84.1±1.6	64.3±2.4	80.5±2.5	**85.1±2.1**	**85.1±2.5**	85.7±1.1	70.1±2.6	82.6±2.3	**86.2±1.4**	85.8±2.1	87.3±0.8	72.6±1.9	84.9±1.4	**87.6±0.9**	86.2±1.7

Table 13. Comparison of pixel-level anomaly detection in terms of subset-wise PRO on VisA.

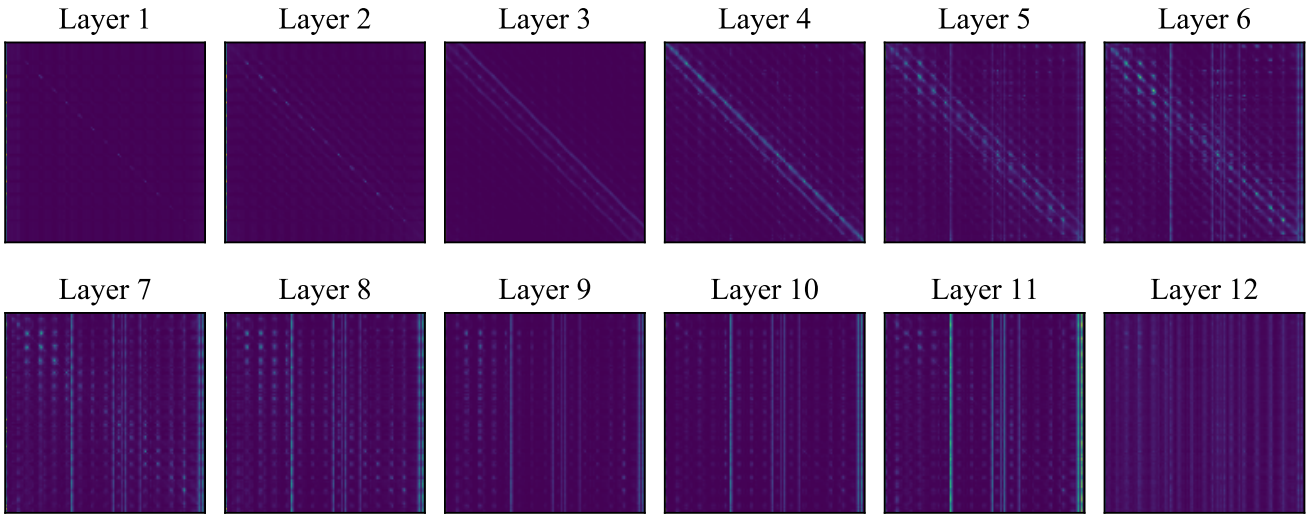

Figure 5. Visualization of the QK attention map in the vision encoder.

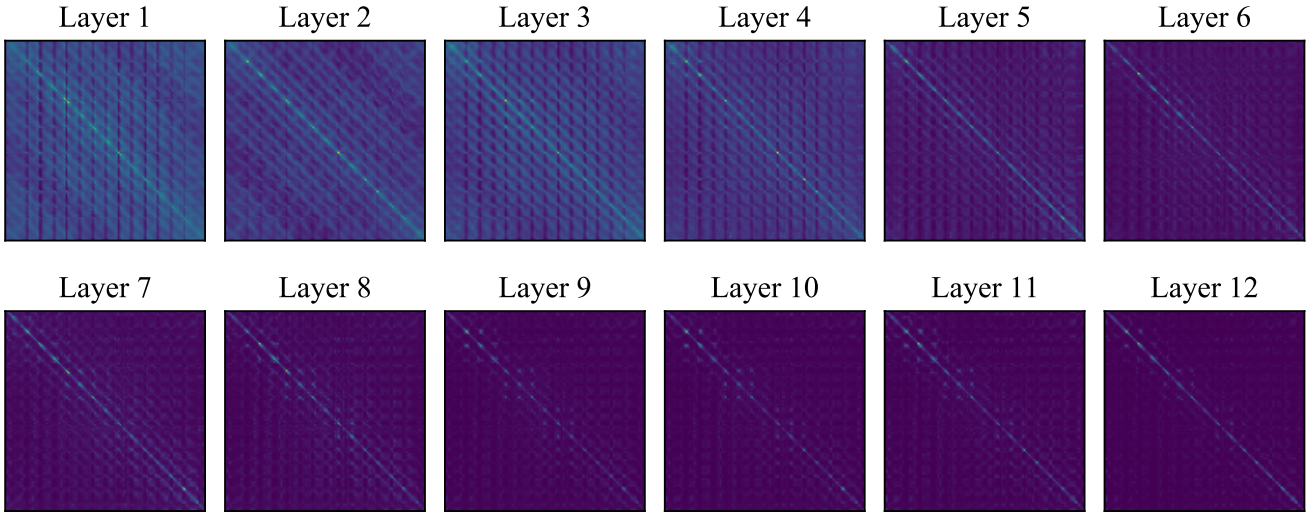

Figure 6. Visualization of the VV attention map in the vision encoder.

tlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130 (4):947–969, 2022.

[4] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[5] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021.

[6] 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.

[7] 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.

[8] 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. IEEE, 2021.

[9] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19113–19122. IEEE, 2023.

[10] 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.

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

[12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

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

[14] 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.