

Image Size		128		256								
Category/Method		MKD [33]	Ours	GT [10]	GN [2]	US [4]	PSVDD [43]	DAAD [16]	MF [40]	PaDiM [8]	CutPaste [23]	Ours
Textures	Carpet	79.3	99.2	43.7	69.9	91.6	92.9	86.6	94.0	99.8	93.9	98.9
	Grid	78.0	95.7	61.9	70.8	81.0	94.6	95.7	85.9	96.7	100	100
	Leather	95.1	100	84.1	84.2	88.2	90.9	86.2	99.2	100	100	100
	Tile	91.6	99.4	41.7	79.4	99.1	97.8	88.2	99.0	98.1	94.6	99.3
	Wood	94.3	98.8	61.1	83.4	97.7	96.5	98.2	99.2	99.2	99.1	99.2
	<i>Average</i>	87.7	98.6	58.5	77.5	91.5	94.5	91.0	95.5	98.8	97.5	99.5
Objects	Bottle	99.4	100	74.4	89.2	99.0	98.6	97.6	99.1	99.9	98.2	100
	Cable	89.2	97.1	78.3	75.7	86.2	90.3	84.4	97.1	92.7	81.2	95.0
	Capsule	80.5	89.5	67.0	73.2	86.1	76.7	76.7	87.5	91.3	98.2	96.3
	Hazelnut	98.4	99.8	35.9	78.5	93.1	92.0	92.1	99.4	92.0	98.3	99.9
	Metal Nut	73.6	99.2	81.3	70.0	82.0	94.0	75.8	96.2	98.7	99.9	100
	Pill	82.7	93.3	63.0	74.3	87.9	86.1	90.0	90.1	93.3	94.9	96.6
	Screw	83.3	91.1	50.0	74.6	54.9	81.3	98.7	97.5	85.8	88.7	97.0
	Toothbrush	92.2	90.3	97.2	65.3	95.3	100	99.2	100	96.1	99.4	99.5
	Transistor	85.6	99.5	86.9	79.2	81.8	91.5	87.6	94.4	97.4	96.1	96.7
	Zipper	93.2	94.3	82.0	74.5	91.9	97.9	85.9	98.6	90.3	99.9	98.5
	<i>Average</i>	87.8	95.4	71.6	75.5	85.8	90.8	88.8	96.0	93.8	95.5	98.0
<i>Total Average</i>		87.8	96.5	67.2	76.2	87.7	92.1	89.5	95.8	95.5	96.1	98.5

Table 1. *Anomaly Detection* results on MVTec [3]. For each category with images of 256×256 resolution, methods achieved for the top two AUROC (%) are highlighted in bold. Our method ranks first according to the average scores of **textures**, **objects** and overall.

anomalous region. Hence, we define the maximum value in S_{AL} as sample-level anomaly score S_{AD} . The intuition is that no significant response exists in their anomaly score map for normal samples.

4. Experiments and Discussions

Empirical evaluations are carried on both the MVTec anomaly detection and localization benchmark and unsupervised one-class novelty detection datasets. In addition, we perform ablation study on the MVTec benchmark, investigating the effects of different modules/blocks on the final results.

4.1. Anomaly Detection and Localization

Dataset. MVTec [3] contains 15 real-world datasets for *anomaly detection*, with 5 classes of **textures** and 10 classes of **objects**. The training set comprises a total of 3,629 anomaly-free images. The test set has both anomalous and anomaly-free images, totaling 1,725. Each class has multiple defects for testing. In addition, pixel-level annotations are available in the test dataset for *anomaly localization* evaluation.

Experimental settings. All images in MVTec are resized to a specific resolution (e.g. 128×128 , 256×256 etc.). Following convention in prior works, anomaly detection and localization are performed on one category at a time. In this experiment, we adopt WideResNet50 as Backbone E in our T-S model. We also report the AD results with ResNet18 and ResNet50 in ablation study. To train our reserve distillation model, we utilize Adam optimizer [18] with $\beta = (0.5, 0.999)$. The learning rate is set to 0.005. We train 200 epochs with a batch size of 16. A Gaussian filter

with $\sigma = 4$ is used to smooth the anomaly score map (as described in Sec. 3.3).

For *Anomaly detection*, we take area under the receiver operating characteristic (AUROC) as the evaluation metric. We include prior arts in this experiments, including MKD [33], GT [10], GANomaly (GN) [2], Uninformed Student (US) [4], PSVDD [43], DAAD [16], MetaFormer (MF) [40], PaDiM (WResNet50) [8] and CutPaste [23].

For *Anomaly Localization*, we report both AUROC and per-region-overlap (PRO) [4]. Different from AUROC, which is used for per-pixel measurement, the PRO score treats anomaly regions with any size equally. The comparison baselines includes MKD [33], US [4], MF [40], SPADE (WResNet50) [7, 29], PaDiM (WResNet50) [8], RIAD [46] and CutPaste [23].

Experimental results and discussions. Anomaly detection results on MVTec are shown in Tab. 1. The average outcome shows that our method exceeds SOTA by **2.5%**. For **textures** and **objects**, our model reaches new SOTA of **99.5%** and **98.0%** of AUROC, respectively. The statistics of the anomaly scores are shown in Fig. 5. The non-overlap distribution of normal (blue) and anomalies (red) indicates the strong AD capability in our T-S model.

Quantitative results on anomaly localization are summarized in Tab. 2. For both AUROC and PRO average scores over all categories, our approach surpasses state-of-the-art with **97.8%** and **93.9%**. To investigate the robustness of our method to various anomalies, we classify the defect types into two categories: large defects or structural anomalies and tiny or inconspicuous defects, and qualitative evaluate the performance by visualization in Fig. 6 and Fig. 7. Compared to the runner-up (i.e. CutPaste [23]) in Tab. 1, our method produces a significant response to the whole