南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

GLAD: Towards Better Reconstruction with Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection

Hang Yao[1], Ming Liu[1,2](✉), Haolin Wang[1], Zhicun Yin[1], Zifei Yan[1,2], Xiaopeng Hong[1], and Wangmeng Zuo[1,2]
yaohang_1@outlook.com, csmliu@outlook.com, cswmzuo@gmail.com

[1]Harbin Institute of Technology, Harbin, China
[2]Pazhou Lab Huangpu, Guangzhou, China

ECCV 2024

Background

Anomaly detection (AD) aims to detect and locate abnormal patterns of objects:

- it challenging to collect enough abnormal samples for all anomaly types in situations.
- ever-changing product design and production processes, it is impossible to collect all anomalies in advance.

unsupervised anomaly detection (UAD) has drawn much attention with only normal samples required.

Related works

Embedding-based methods: extract feature of images to evaluate abnormal.

1. Knowledge distillation-based methods: train student network with normal samples, features from the pre-trained teacher network are compared with features from the student network to detect and locate anomalies.
2. PaDiM builds multivariate Gaussian distributions for patch features of normal samples and uses Mahalanobis distance as the anomaly score.
3. PatchCore proposes a memory bank to save features of normal images, which are compared with feature maps of test images to distinguish the difference between normal and abnormal features.

Reconstruction-based methods: detected and located via the comparison between the given sample and its normal counterpart.

Based on the hypothesis that models trained on normal samples only can reconstruct normal images well. Anomalies can be detected by comparing the samples before and after reconstruction.

AE(early), GAN, transformer, UNet architecture

Motivation

$$\mathbf{x}^a \overset{diff}{\rightarrow} \mathbf{x}_t^a \overset{gen}{\rightarrow} \hat{\mathbf{x}}^a$$

Diffusion models have prominent modeling ability.

During the training process , the diffusion model captures the distribution of normal samples only.

The same denoising step:

- Different anomalies is uneven.
- Less preserved details of the original

The anomaly noise inevitably deviates from the standard Gaussian distribution.

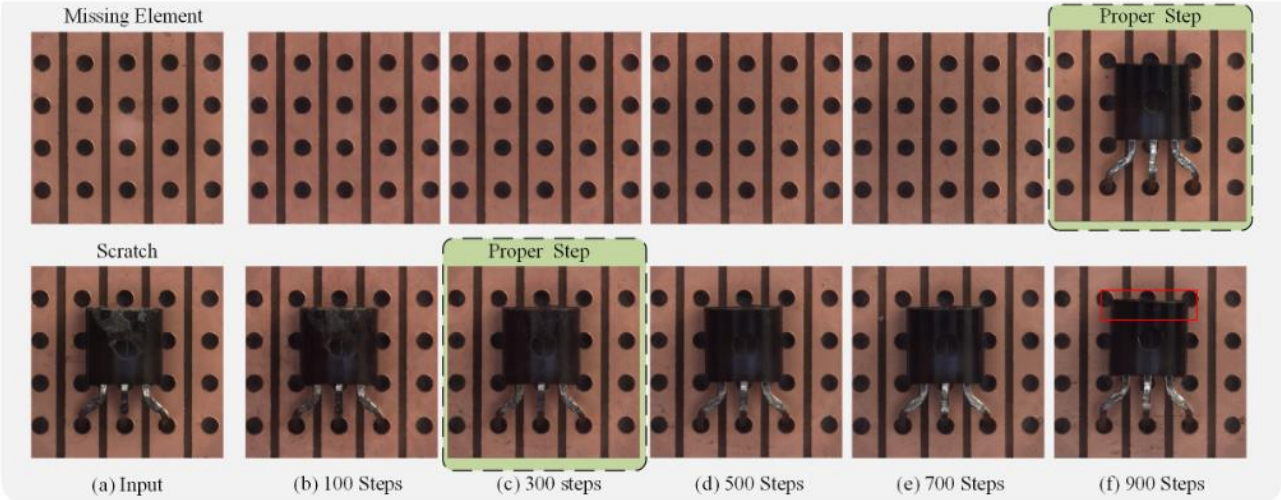


Fig. 1: Illustration of adaptive denoising process. For severe anomalies like missing elements, it requires a large number of denoising steps (900) to add the element back, while for small anomalies like scratch, 300 steps are already enough. Besides, setting a large enough denoising step (*e.g.*, 900) for all samples will affect the detail preservation. For example, in the area bounded by red lines, the position of the element is changed, which will be marked as anomalies during the comparison process.

Method

Inference:
Adaptive Denoising Step (ADS) : achieves a better trade-off between reconstruction quality and detail preservation ability.
Spatial-Adaptive Feature Fusion (SAFF): avoid reconstruction of normal regions.

Training:
Anomaly-oriented Training Paradigm (ATP): allow diffusion model to predict non-Gaussian noise at abnormal regions.

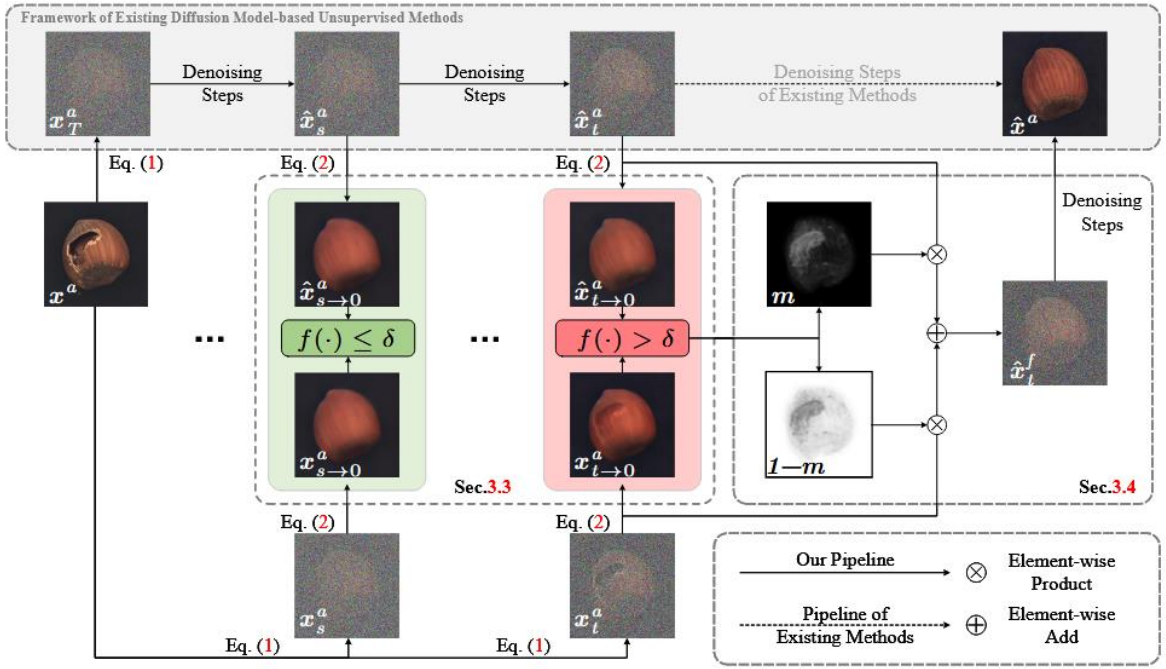


Fig. 2: The reconstruction pipeline of the proposed GLAD, including the Adaptive Denoising Steps (Sec. 3.3) and the Spatial-Adaptive Feature Fusion Scheme (Sec. 3.4).

Method / Preliminary

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Diffusion Process $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \qquad \bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$

Intermediate Result $\boldsymbol{x}_{t\to 0} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\boldsymbol{x}_t, t)),$

Generation Process $\hat{\boldsymbol{x}}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}_{t\to 0} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\hat{\boldsymbol{x}}_t, t).$

$\boldsymbol{x}^a \xrightarrow{diff} \boldsymbol{x}_t^a \xrightarrow{gen} \hat{\boldsymbol{x}}^a$

$\|\hat{\boldsymbol{x}}^a - \boldsymbol{x}\|_\infty < \tau$

detected and located by comparison between $\hat{\boldsymbol{x}}^a$ and $\boldsymbol{x}^a$

$\boldsymbol{x}^a = \boldsymbol{x} + \boldsymbol{n}$

$$\begin{aligned}\boldsymbol{x}_t^a &= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^a + \sqrt{1-\bar{\alpha}_t}\epsilon^a \\ &= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon^a + \sqrt{\bar{\alpha}_t}\boldsymbol{n},\end{aligned}$$

$$\begin{aligned}\hat{\boldsymbol{x}}^a - \boldsymbol{x} &= g_t(\boldsymbol{x}_t^a) - g_t(\boldsymbol{x}_t) \\ &= g_t(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon^a + \sqrt{\bar{\alpha}_t}\boldsymbol{n}) - g_t(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon) \\ &\overset{\propto}{\sim} \sqrt{1-\bar{\alpha}_t}(\epsilon^a - \epsilon) + \sqrt{\bar{\alpha}_t}\boldsymbol{n},\end{aligned}$$

we make our efforts to reduce the errors for a better reconstruction quality.

Method / Adaptive Denoising Steps (ADS)

$$\boldsymbol{x}_{t\to 0} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\boldsymbol{x}_t, t)),$$

x^a —add T noises→ x_T^a —generate→ $\hat{x}_t^a$ → $\hat{x}_{t\to 0}^a$

x^a —add t noises→ x_t^a → $x_{t\to 0}^a$

the anomalies will be preserved to some extent

(t + n) step of denoising,
normal regions are best preserved;
the anomalies can be reconstructed

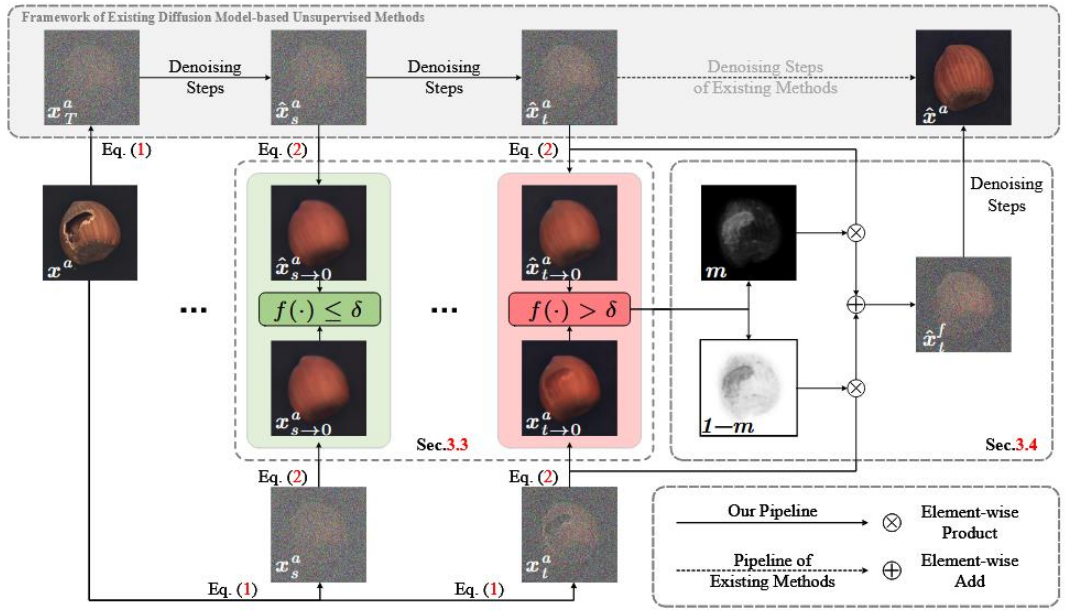


Fig. 2: The reconstruction pipeline of the proposed GLAD, including the Adaptive Denoising Steps (Sec. 3.3) and the Spatial-Adaptive Feature Fusion Scheme (Sec. 3.4).

Method / Spatial-Adaptive Feature Fusion (SAFF)

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

the whole image is sub-optimal

we need only to set a larger step for the abnormal regions.

mask m, which means the possibility for the pixels to be part of the anomalies

$$\hat{\boldsymbol{x}}_t^f = \boldsymbol{m} \cdot \hat{\boldsymbol{x}}_t^a + (1 - \boldsymbol{m}) \cdot \boldsymbol{x}_t^a.$$

$$\begin{aligned}\hat{\boldsymbol{x}}_t^f &= \sqrt{\bar{\alpha}_t}\hat{\boldsymbol{x}}_{t\to 0}^f + \sqrt{1-\bar{\alpha}_t}\epsilon \\ &= \sqrt{\bar{\alpha}_t}(\boldsymbol{m} \cdot \hat{\boldsymbol{x}}_{t\to 0}^a + (1-\boldsymbol{m}) \cdot \boldsymbol{x}_{t\to 0}^a) + \sqrt{1-\bar{\alpha}_t}\epsilon.\end{aligned}$$

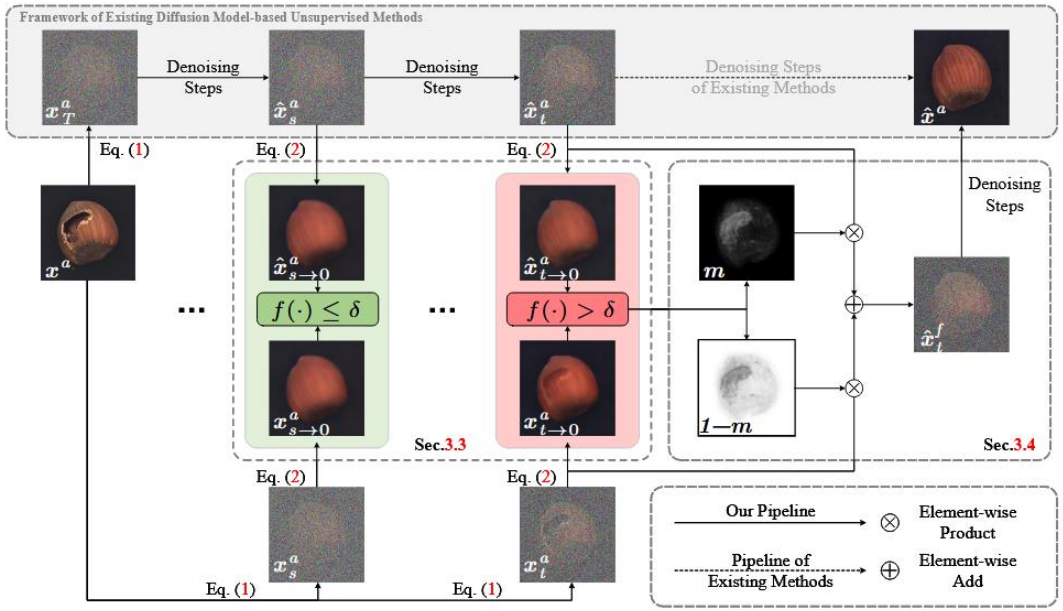


Fig. 2: The reconstruction pipeline of the proposed GLAD, including the Adaptive Denoising Steps (Sec. 3.3) and the Spatial-Adaptive Feature Fusion Scheme (Sec. 3.4).

Method / Anomaly-oriented Training Paradigm (ATP)

$$\hat{\boldsymbol{x}}^a - \boldsymbol{x} \overset{\propto}{\simeq} \sqrt{1-\bar{\alpha}_t}(\epsilon^a - \epsilon) + \sqrt{\bar{\alpha}_t}\boldsymbol{n} \to 0.$$

$$\epsilon^a \to \epsilon - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{n}.$$

$$\begin{aligned} L_{ATP} &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{x}^a)\sim p_{data},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}[\|(\epsilon - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{n}) - \epsilon^a\|_2] \\ &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{x}^a)\sim p_{data},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}[\|(\epsilon - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}(\boldsymbol{x}^a - \boldsymbol{x})) - \epsilon_\theta(\boldsymbol{x}_t^a, t)\|_2]. \end{aligned}$$

MemSeg [29] to synthesize abnormal samples with normal ones, which enables the training to proceed in an unsupervised manner

Method / Anomaly Scoring and Map Construction

pre-trained model: DINO

$F_t \in \mathbb{R}^{c \times u \times v}$ features of test images

$F_r \in \mathbb{R}^{c \times u \times v}$ features of reconstructed images

$$M_l^{(i,j)}(F_t^l, F_r^l) = \min(1 - \langle F_t^{l^{(i,j)}}, F_r^l \rangle),$$

$$M = \sum\nolimits_l M_l(F_t^l, F_r^l),$$

Anomaly score: the whole image is the average of the top K maximum values of M .

Experiment

We use the pre-trained latent diffusion model (LDM) [21] and fine-turn the UNet to adapt data for reconstruction.

DINO [5] with ViT-B/8 architectures is utilized as a feature extraction model

Experiment

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 1: Comparison with SOTA methods on MVTec-AD dataset. I-AUROC and P-AUROC are displayed in each entry. The best results among all methods are shown in bold, and the underlined results denote the best results among reconstruction-based methods.

Category	Embedding-based methods			Reconstruction-based methods					
	PatchCore [22]	RD4AD [8]	SimpleNet [17]	DRAEM [32]	OCR-GAN [14]	Lu *et al.* [18]	DiffAD [33]	DDAD [19]	Ours
Carpet	98.7/**99.0**	98.9/98.8	**99.7**/98.2	97.0/95.5	99.4/-	-/97.7	98.3/98.1	99.3/98.7	99.0/98.5
Grid	98.2/98.7	**100**/97.0	99.7/98.8	99.9/**99.7**	99.6/-	-/95.6	**100**/**99.7**	**100**/99.4	**100**/99.6
Leather	**100**/99.3	**100**/98.6	**100**/99.2	**100**/98.5	97.1/-	-/97.5	**100**/99.1	**100**/99.4	**100**/**99.8**
Tile	98.7/95.6	99.3/98.9	99.8/97.0	99.6/99.2	95.5/-	-/98.9	**100**/**99.4**	**100**/98.2	**100**/98.7
Wood	99.2/95.0	99.2/**99.3**	**100**/94.5	99.1/96.4	95.7/-	-/99.1	**100**/96.7	**100**/95.0	99.4/98.4
Bottle	**100**/98.6	**100**/99.0	**100**/98.0	99.2/**99.1**	99.6/-	-/97.3	**100**/98.8	**100**/98.7	**100**/98.9
Cable	99.5/98.4	95.0/99.4	**99.9**/97.6	91.8/94.7	99.1/-	-/**99.5**	94.6/96.8	99.4/98.1	**99.9**/98.1
Capsule	98.1/98.8	96.3/97.3	97.7/**98.9**	98.5/94.3	96.2/-	-/96.8	97.5/98.2	99.4/95.7	**99.5**/98.5
Hazelnut	**100**/98.7	99.9/98.2	**100**/97.9	**100**/**99.7**	98.5/-	-/92.5	**100**/99.4	**100**/98.4	**100**/99.5
Metal nut	**100**/98.4	**100**/**99.6**	**100**/98.8	98.7/99.5	99.5/-	-/99.0	**100**/99.4	**100**/99.0	**100**/98.8
Pill	96.6/97.4	**99.6**/95.7	99.0/**98.6**	98.9/97.6	98.3/-	-/92.1	97.7/97.7	**100**/99.1	98.1/97.9
Screw	98.1/**99.4**	97.0/99.1	98.2/99.3	93.9/97.6	**100**/-	-/98.6	97.2/99.0	99.0/99.3	96.9/99.1
Toothbrush	**100**/98.7	99.5/93.0	99.7/98.5	**100**/98.1	98.7/-	-/93.1	**100**/99.2	**100**/98.7	**100**/**99.4**
Transistor	**100**/96.3	96.7/95.4	**100**/**97.6**	93.1/90.9	98.3/-	-/94.5	96.1/93.7	**100**/95.3	98.3/96.2
Zipper	99.4/98.8	98.5/98.2	99.9/98.9	**100**/98.8	99.0/-	-/97.6	**100**/**99.0**	**100**/98.2	98.5/97.9
Average	99.1/98.1	98.5/97.8	99.6/98.1	98.0/97.3	98.3/-	-/96.7	98.7/98.3	**99.8**/98.1	99.3/**98.6**

Table 2: Comparison with SOTA on MPDD dataset. I-AUROC and P-AUROC are displayed in each entry. The best results among all methods are shown in bold, and the underlined results denote the best results among reconstruction-based methods.

Category	Embedding-based methods			Reconstruction-based methods			
	PatchCore [22]	RD4AD [8]	SimpleNet [17]	DRAEM [32]	OCR-GAN [14]	DDAD [19]	Ours
Bracket Black	85.3/97.6	90.2/98.0	85.1/96.0	91.8/98.2	99.9/-	**98.7**/96.7	98.0/**99.4**
Bracket Brown	92.5/**98.1**	94.2/97.0	**98.3**/94.4	90.3/63.7	89.4/-	92.7/97.2	90.7/97.5
Bracket White	92.3/99.7	90.1/99.3	98.0/96.7	88.8/98.9	88.1/-	96.6/91.8	**98.3**/**99.7**
Connector	100/99.4	99.5/99.4	**100**/**99.5**	**100**/91.2	100/-	96.2/98.6	**100**/98.2
Metal Plate	**100**/98.8	**100**/99.1	**100**/98.5	**100**/96.6	100/-	100/98.1	99.9/**99.4**
Tubes	77.4/97.2	97.6/99.1	97.9/**99.2**	94.7/95.9	98.1/-	99.2/99.0	**98.1**/97.8
Average	91.3/98.5	95.3/**98.7**	96.6/97.4	94.3/90.7	95.9/-	97.2/96.9	**97.5**/**98.7**

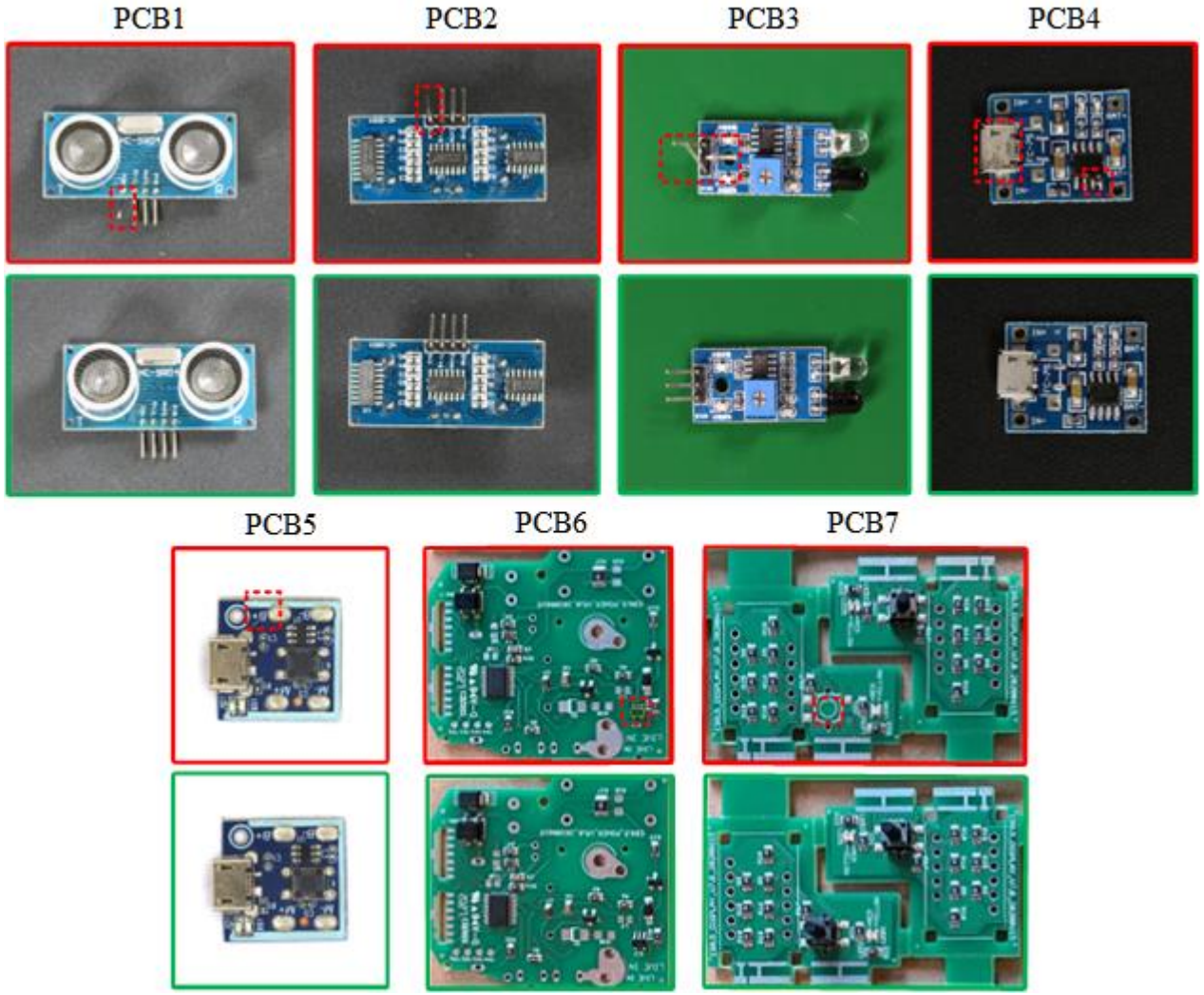
Table 3: Comparison with SOTA methods on VisA dataset. I-AUROC and P-AUROC are displayed in each entry. The best results among all methods are shown in bold, and the underlined results denote the best results among reconstruction-based methods.

Category	Embedding-based methods			Reconstruction-based methods			
	PatchCore [22]	RD4AD [8]	SimpleNet [17]	DRAEM [32]	OCR-GAN [14]	DDAD [19]	Ours
Candle	98.7/**99.2**	96.2/98.9	96.9/98.6	89.6/91.0	98.9/-	**99.9**/98.7	**99.9**/94.8
Capsules	68.8/96.5	91.8/99.4	89.5/99.2	89.2/99.0	98.8/-	**100**/99.5	99.1/**99.6**
Cashew	97.7/99.2	**98.7**/94.4	94.8/**99.0**	88.3/85.0	97.4/-	94.5/97.4	98.4/97.0
Chewinggum	99.1/98.9	99.3/97.6	**100**/98.5	96.4/97.7	99.4/-	98.1/96.5	99.6/**99.1**
Fryum	91.6/95.9	96.9/96.4	96.6/94.5	94.7/82.5	96.3/-	99.0/**96.9**	**99.4**/**96.9**
Macaroni1	90.1/98.5	98.7/99.3	97.6/99.6	93.9/99.4	97.2/-	99.2/98.7	**99.9**/**99.8**
Macaroni2	63.4/93.5	91.4/99.1	83.4/96.4	88.3/99.7	95.1/-	**99.2**/98.2	98.9/**99.8**
Pcb1	96.0/**99.8**	96.7/99.6	99.2/**99.8**	84.7/98.4	96.1/-	**100**/93.4	99.6/99.6
Pcb2	95.1/98.4	97.2/98.3	99.2/**98.8**	96.2/94.0	98.3/-	99.7/97.4	**100**/ 98.6
Pcb3	93.0/98.9	96.5/**99.3**	98.6/99.2	97.4/94.3	98.1/-	97.2/96.3	**99.9**/98.9
Pcb4	99.5/98.3	99.4/98.2	98.9/98.6	98.9/97.6	99.7/-	**100**/98.5	99.9/**99.5**
Pipe fryum	99.0/99.3	99.6/99.1	99.2/99.3	94.7/65.8	99.7/-	**100**/**99.5**	98.9/99.4
Average	91.0/98.1	96.9/98.3	96.2/98.5	92.4/92.0	97.9/-	98.9/97.6	**99.5**/**98.6**

Table 4: Comparison with SOTA methods on PCB-Bank dataset. I-AUROC and P-AUROC are displayed in each entry. The best results among all methods are shown in bold, and the underlined results denote the best results among reconstruction-based methods.

Category	Embedding-based methods			Reconstruction-based methods			
	PatchCore [22]	RD4AD [8]	SimpleNet [17]	DRAEM [32]	OCR-GAN [14]	DDAD [19]	Ours
Pcb1	96.0/**99.8**	96.7/99.6	99.2/**99.8**	84.7/98.4	96.1/-	**100**/93.4	99.6/99.6
Pcb2	95.1/98.4	97.2/98.3	99.2/**98.8**	96.2/94.0	98.3/-	99.7/97.4	**100**/ 98.6
Pcb3	93.0/98.9	96.5/**99.3**	98.6/99.2	97.4/94.3	98.1/-	97.2/96.3	**99.9**/98.9
Pcb4	99.5/98.3	99.4/98.2	98.9/98.6	98.9/97.6	99.7/-	**100**/98.5	99.9/**99.5**
Pcb5	94.6/**99.8**	94.1/99.5	94.5/99.4	97.2/97.8	85.9/-	**99.7**/96.0	99.6/99.1
Pcb6	82.2/98.9	89.4/98.9	91.7/97.5	72.4/94.6	75.1/-	87.8/98.5	**92.2**/**99.7**
Pcb7	93.7/98.8	99.0/99.6	**100**/**99.9**	97.7/98.3	85.7/-	94.4/98.7	99.6/99.8
Average	94.2/99.1	96.0/99.1	96.2/98.5	91.5/96.4	91.3/-	97.4/96.5	**98.7**/**99.3**

Experiment



Experiment

other methods usually fail to reconstruct **large-scale** anomalies into normal regions

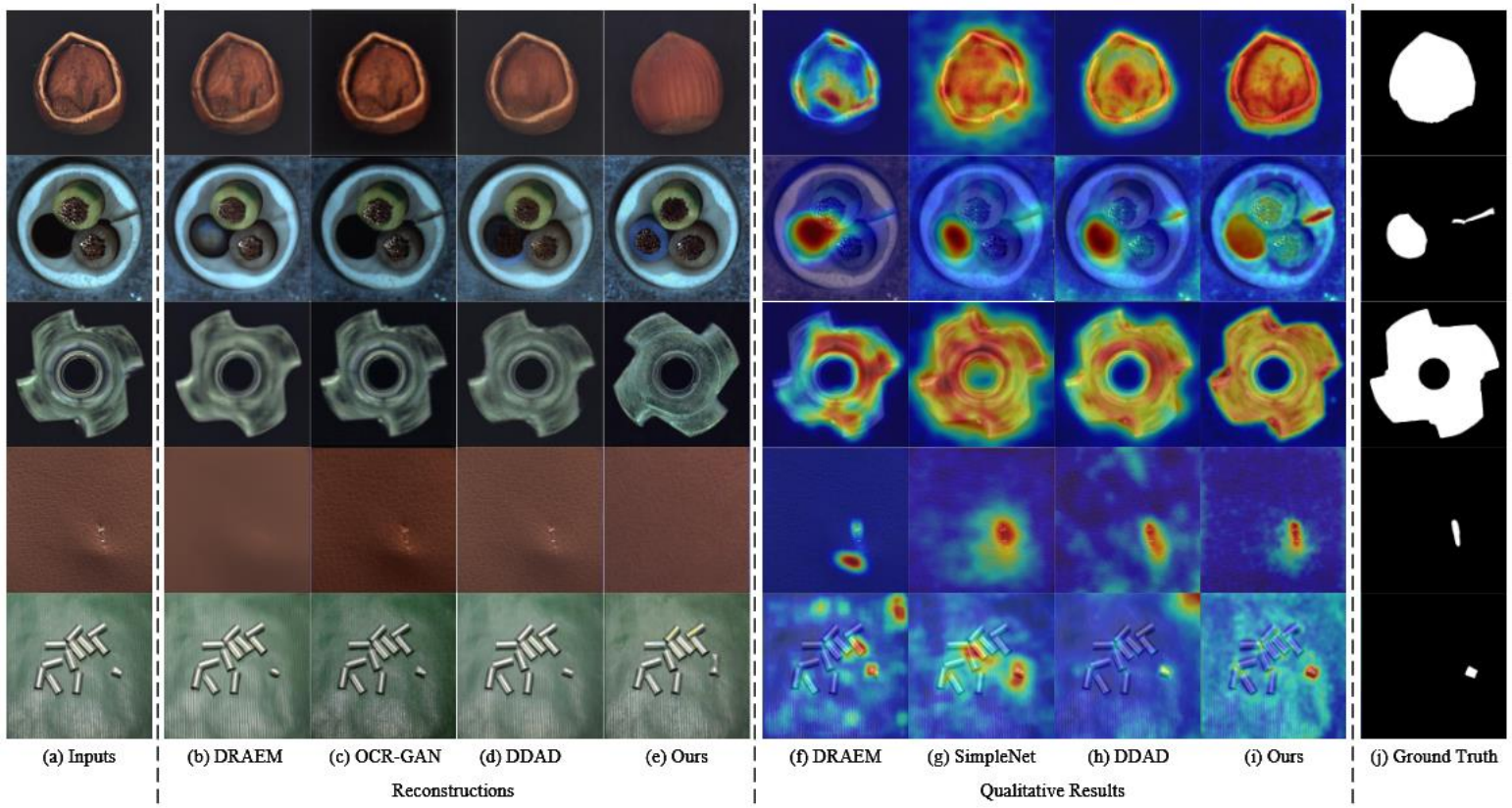


Fig. 3: Reconstructions and qualitative comparisons with other methods. The first four rows display examples of the MVTec-AD dataset, and the last row is for the MPDD dataset. OCR-GAN only produces anomaly scores, and there is no anomaly map. SimpleNet is the embedding-based method.

Experiment / Ablation Study

Adaptive Denoising Step (ADS)
Anomaly-oriented Training Paradigm (ATP)
Spatial-Adaptive Feature Fusion (SAFF)

Compare to fix steps
ADS: ensures anomaly-free reconstruction and preserves as much information as possible about normal regions.

Table 6: Performance of each component on MVTec-AD dataset. The best results are shown in bold.

Method	I-AUROC	P-AUROC
Baseline	98.3	98.0
Baseline + ADS	99.0	98.5
Baseline + ATP	98.7	98.5
Baseline + ADS + ATP w/o SAFF	99.2	98.3
Baseline + ADS + ATP with SAFF (Ours)	**99.3**	**98.6**

Table 7: Comparison of adaptive steps and different fixed steps on MVTec-AD dataset. The best results are shown in bold.

Denoising steps	fixed steps			Adaptive steps
	350 step	550 step	750 step	
I-AUROC	98.8	98.8	98.7	**99.3**
P-AUROC	97.6	98.1	98.5	**98.6**

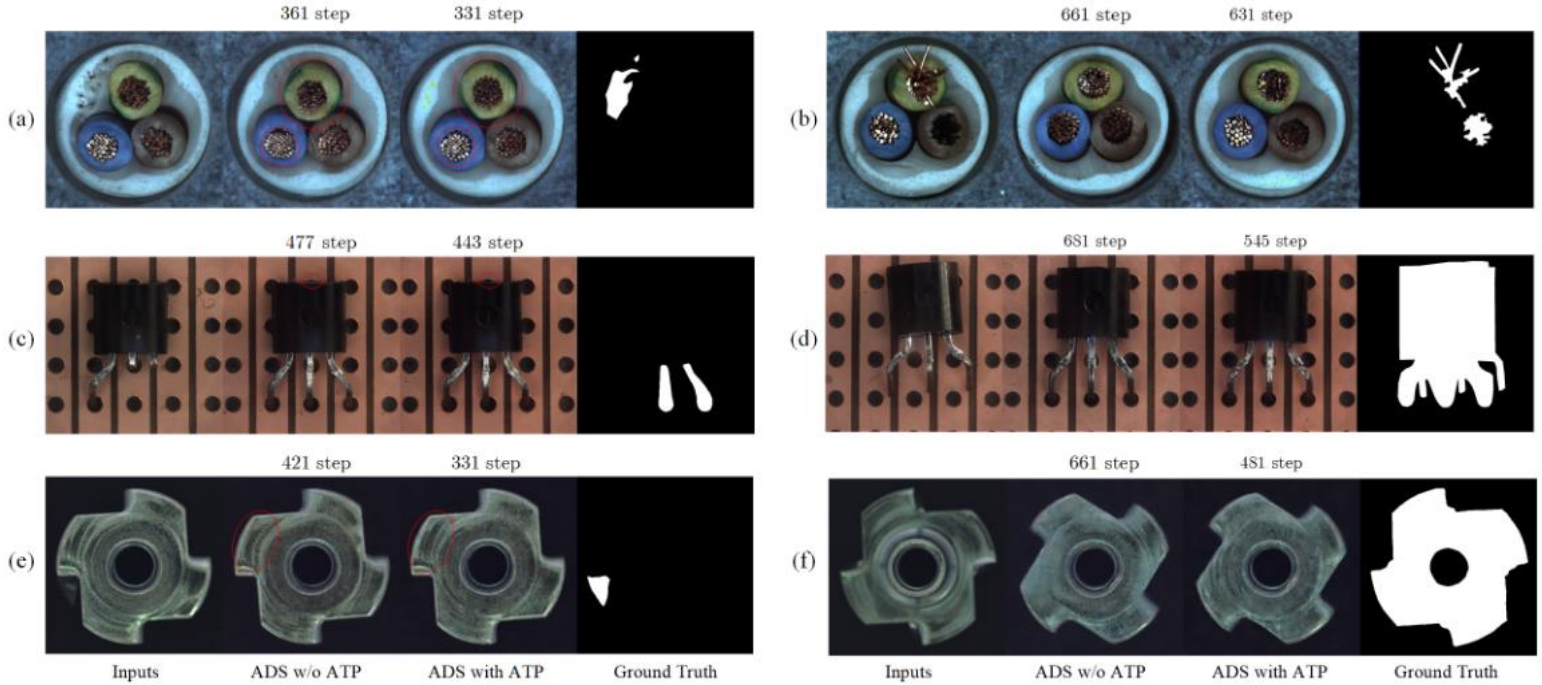
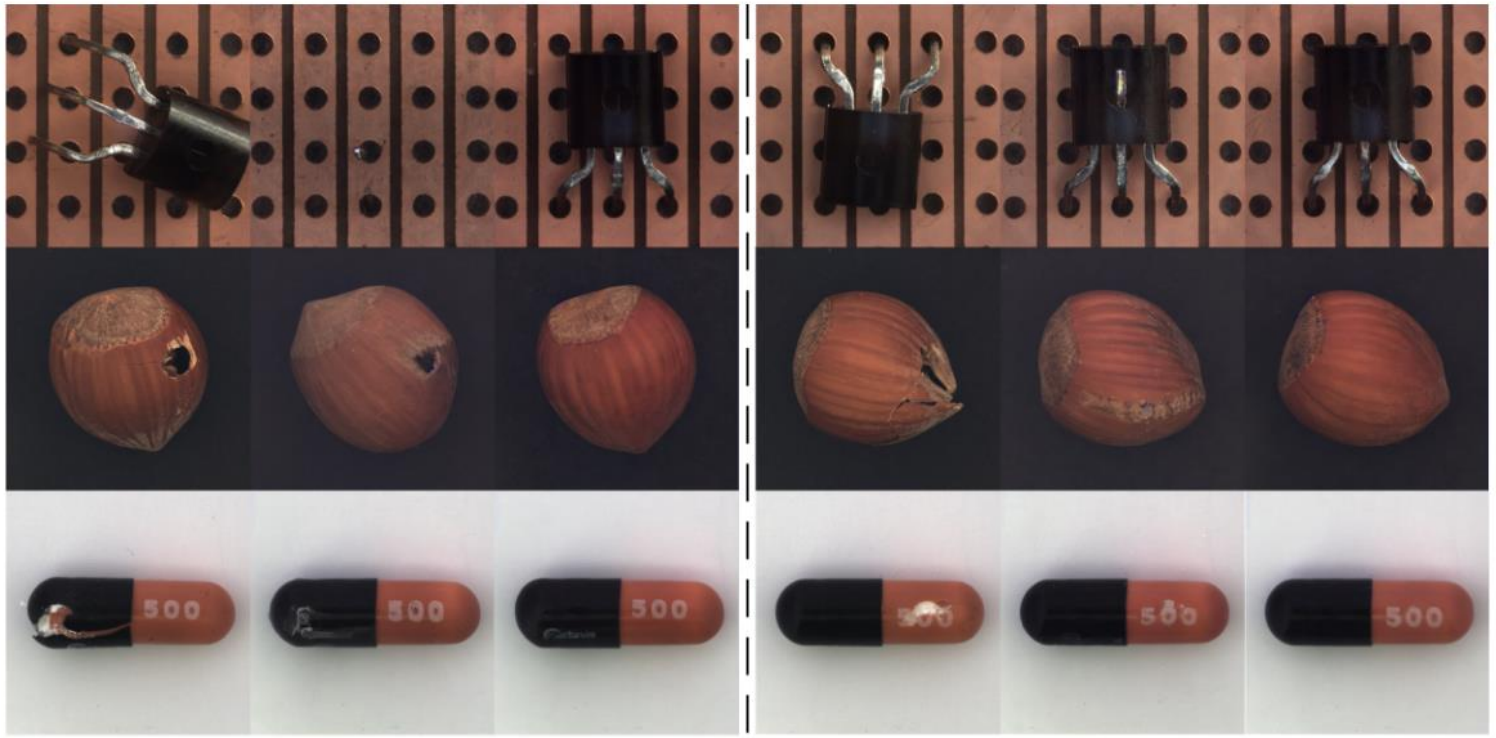


Fig. 4: Reconstructions of different types of anomaly and proper steps. Examples (a), (c), and (e) contain small-scale anomalies, and (b), (d), and (f) are large-scale anomalies. The numbers above the reconstructed images represent the proper steps. Differences in details of normal areas are marked in red circles.



(a) Inputs (b) Baseline (c) ATP (d) Inputs (e) Baseline (f) ATP

Fig. 5: Qualitative comparisons between baseline and proposed ATP on MVTec-AD. The same denoising steps are used for the two methods.

Thank you