Dynamic Addition of Noise in a Diffusion Model for Anomaly Detection

Supplementary Material

7. Appendix

7.1. Implementation Details

Training the Unet is conducted for 300 epochs using the AdamW optimizer [23]. We set a learning rate of 0.0001 and weight decay to 0.01. The noise schedule is from 0.0015 to 0.0195 and we set $T = 1000$. For the ResNet-34, we set the dynamic conditioning feature blocks $\mathbb{J}$ to 2 whereas for anomaly map computation, features are extracted from blocks 2 and 3. The guidance temperature is either 0 (indicating no guidance) or within the range 7-10. We set the number of epochs γ for fine-tuning the feature extractor in the range 0 to 3. The weighting parameter λ for the anomaly maps is set to 0.85 and the final anomaly map is smoothed with a Gaussian filter with $\sigma = 4$. The pretrained VAE from [31] is used without further training.

7.2. Additional Details and Ablations

7.2.1 Noiseless Reconstruction

We studied the influence of the 'noiseless' and only scaled input on performance of the VisA benchmark. In figure 13, we provide different fractions of noise influence and the corresponding metrics. We tested fractions $\omega \in \{0, 0.1, 0.2, ..., 1\}$ of the noise as follows:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \omega\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \quad (12)$$

We perceived best performance with our proposed noiseless scaling ($\omega = 0$) with a declining performance as ω increases. In addition, we conducted a qualitative analysis to compare the visual impact of image-level perturbations in the forward diffusion process (as outlined in Equation 1; refer to Figures 10 and 11). Our tests extend up to the 400th time step, revealing that introducing noise degrades the visual quality of the signal rapidly. Furthermore, Figure 12 illustrates the effect of noiseless scaling over an extended period, up to the 800th time step. To provide a more transparent comparison, we executed these disturbance analyses in the pixel space rather than in the latent space. Figure 14 shows the anomaly map construction and reconstruction with varying perturbance levels for noiseless scaling versus noising. The Figure shows similar segmentation performance with slightly less artifacts in the anomaly map created by noiseless scaling, while for high perturbance levels (T=240,320) the noising paradigm is prone to hallucinations in the reconstruction as highlighted by a red circle.

7.2.2 Additional quantitative analysis

We extend the analysis of MVTec provided in Table 3 with a more detailed Table 7. While showcasing decent performance on diverse categories, we got unexpectedly weak results for the Screw category. We don't think this results are inherently due to our proposed approach but could be solved with further hyperparameter tuning. Table 8 shows the localization performance measured by P-AUROC on the VisA benchmark.

7.2.3 Additional qualitative analysis

In Figure 8, we present a side-by-side comparison of our method's reconstruction capabilities against those of DRAEM [41]. This comparison underscores a notable improvement in reconstruction quality achieved by our approach. Moreover, Figure 9 provides additional instances of anomaly segmentation, further illustrating our method's efficacy. Notably, Figure 7 encompasses both reconstruction and segmentation outcomes. The remarkable segmentation results are attributed to our method's robust reconstruction abilities and the utilization of domain-adapted feature signals. Our method's strength in reconstruction is bolstered by initially estimating the anomaly size, which allows for the effective scaling of large anomalies, as illustrated in rows 3-5 of Figure 7. Additionally, our approach demonstrates impeccable reconstruction of smaller defects, as shown in rows 1, 2, and 6-9, thanks to the selection of appropriate scaling levels. This aspect is further corroborated by Table 5 in the main document. The implementation of a noiseless, scaled latents further enhances these effects, as detailed in Figure 13 and discussed in Appendix section 7.2.1. Furthermore, the domain-adapted feature extractor effectively learns the subtleties of the target domain, efficiently filtering out any artifacts that may arise during the reconstruction process.

7.2.4 Computational analysis

Lastly we present an evaluation on inference time and the frames per second (FPS) rate, as detailed in Table 9. We compare to various representation and reconstruction-based methods and achieve competitive performance. All experiments were carried out on one Nvidia Quadro 8000 graphics card, with a set batch size of 30. The evaluation for the baseline methods got performed with the Anomalib package [4].

Table 7. A detailed comparison of Anomaly Classification and localisation performance of various methods on MVTec benchmark [6] in the format of (I-AUROC, P-AUROC, PRO). Best results are highlighted in bold.

	Representation-based			Reconstruction-based		
Method	PatchCore [32]	SimpleNet [22]	RD++ [38]	SkipGANomaly [3]	DRAEM [41]	Ours
Carpet	(98.7,99.0,96.6)	(99.7,98.2,-)	(**100**,**99.2**,**97.7**)	(70.9,-)	(97.0,95.5,92.9)	(94.2,97.6,95.1)
Grid	(98.2,98.7,96.0)	(99.7,98.8,-)	(**100**,99.3,97.7)	(47.7,-)	(99.9,**99.7**,**98.4**)	(**100**,99.2,96.9)
Leather	(**100**,99.3,98.9)	(**100**,99.2,-)	(**100**,**99.4**,**99.2**)	(60.9,-)	(**100**,98.6,98.0)	(98.5,**99.4**,98.1)
Tile	(98.7,95.4,87.3)	(**99.8**,97.0,-)	(99.7,96.6,92.4)	(29.9,-)	(99.6,**99.2**,**98.9**)	(95.5,94.7,93.6)
Wood	(99.2,95.0,89.4)	(**100**,94.5,-)	(99.3,95.8,93.3,)	(19.9,-)	(99.1,**96.4**,**94.6**)	(99.7,95.9,91.0)
Bottle	(**100**,98.6,96.2)	(**100**,98.0,-)	(**100**,98.8,97.0)	(85.2,-)	(99.2,**99.1**,**97.2**)	(**100**,98.6,96.0)
Cable	(99.5,**98.4**,92.5)	(**99.9**,97.6,-)	(99.2,**98.4**,**93.9**)	(54.4,-)	(91.8,94.7,76.0)	(97.8,93.3,87.3)
Capsule	(98.1,98.8,95.5)	(97.7,**98.9**,-)	(**99.0**,98.8,96.4)	(54.3,-)	(98.5,94.3,91.7)	(96.6,97.9,90.7)
Hazelnut	(**100**,98.7,93.8)	(**100**,97.9,-)	(**100**,99.2,96.3)	(24-5,-)	(**100**,**99.7**,**98.1**)	(98.0,98.8,91.8)
Metal nut	(**100**,98.4,91.4)	(**100**,98.8,-)	(**100**,98.1,93.0)	(81.4,-)	(98.7,**99.5**,**94.1**)	(98.9,96.1,89.7)
Pill	(96.6,97.4,93.2)	(99.0,**98.6**,-)	(98.4,98.3,**97.0**)	(67.1,-)	(98.9,97.6,88.9)	(**99.2**,98.2,96.2)
Screw	(98.1,99.4,97.9)	(98.2,99.3,-)	(**98.9**,**99.7**,**98.6**)	(87.9,-)	(93.9,97.6,98.2)	(83.9,99.0,95.5)
Toothbrush	(**100**,98.7,91.5)	(99.7,98.5,-)	(**100**,**99.1**,94.2)	(58.6,-)	(**100**,98.1,90.3)	(**100**,99.0,**94.6**)
Transistor	(**100**,96.3,83.7)	(**100**,**97.6**,-)	(98.5,94.3,81.8)	(84.5,-)	(93.1,90.9,81.6)	(96.8,95.6,**86.9**)
Zipper	(99.4,98.8,**97.1**)	(99.9,**98.9**,-)	(98.6, 98.8,96.3)	(76.1,-)	(**100**,98.8,96.3)	(98.2,98.3,95.3)
Average	(99.1,98.1,93.4)	(**99.6**,98.1,-)	(99.4,**98.3**,**95.0**)	(60.2,-)	(98.0,97.3,93.0)	(97.2,97.4,93.3)

Table 8. Localization performance (P-AUROC) of various methods on VisA benchmark. The best results are highlighted in bold.

Method	SPADE	PaDiM	RD4AD	PatchCore	DRAEM	Ours
P-AUROC	85.6	98.1	96.5	**98.8**	93.5	97.9

Table 9. Inference time for one image in seconds and frames-per-second (FPS) of selected models on VisA benchmark.

	Representation-based		Reconstruction-based	
Method	RD4AD	PatchCore	DRAEM	Ours
FPS	(4.8)	(4.8)	(4.3)	(2.9)
Inference Time	(0.21)	(0.21)	(0.23)	(0.34)

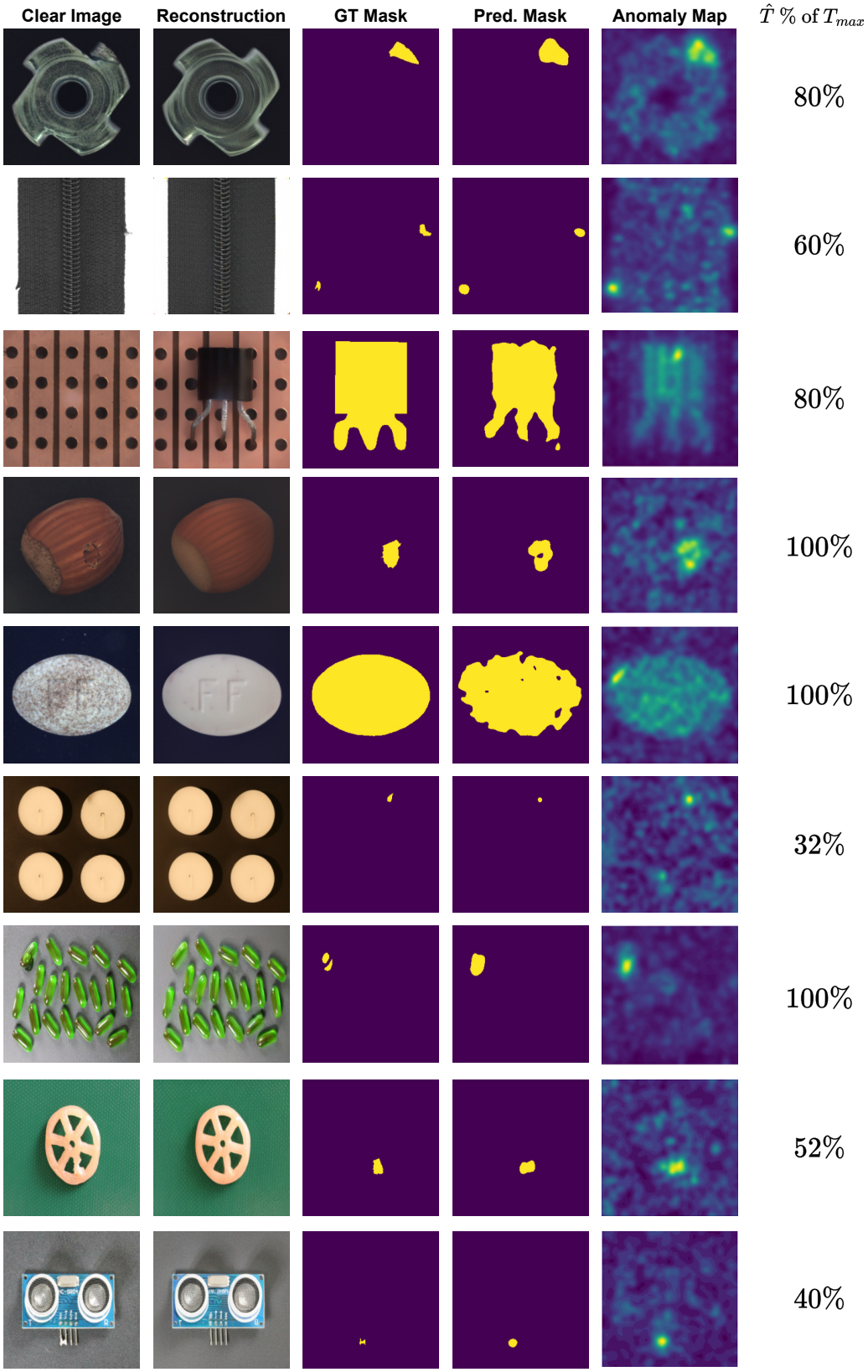

Figure 7. Reconstruction and segmentation performance of our approach of various categories of the VisA and MVTec benchmark.

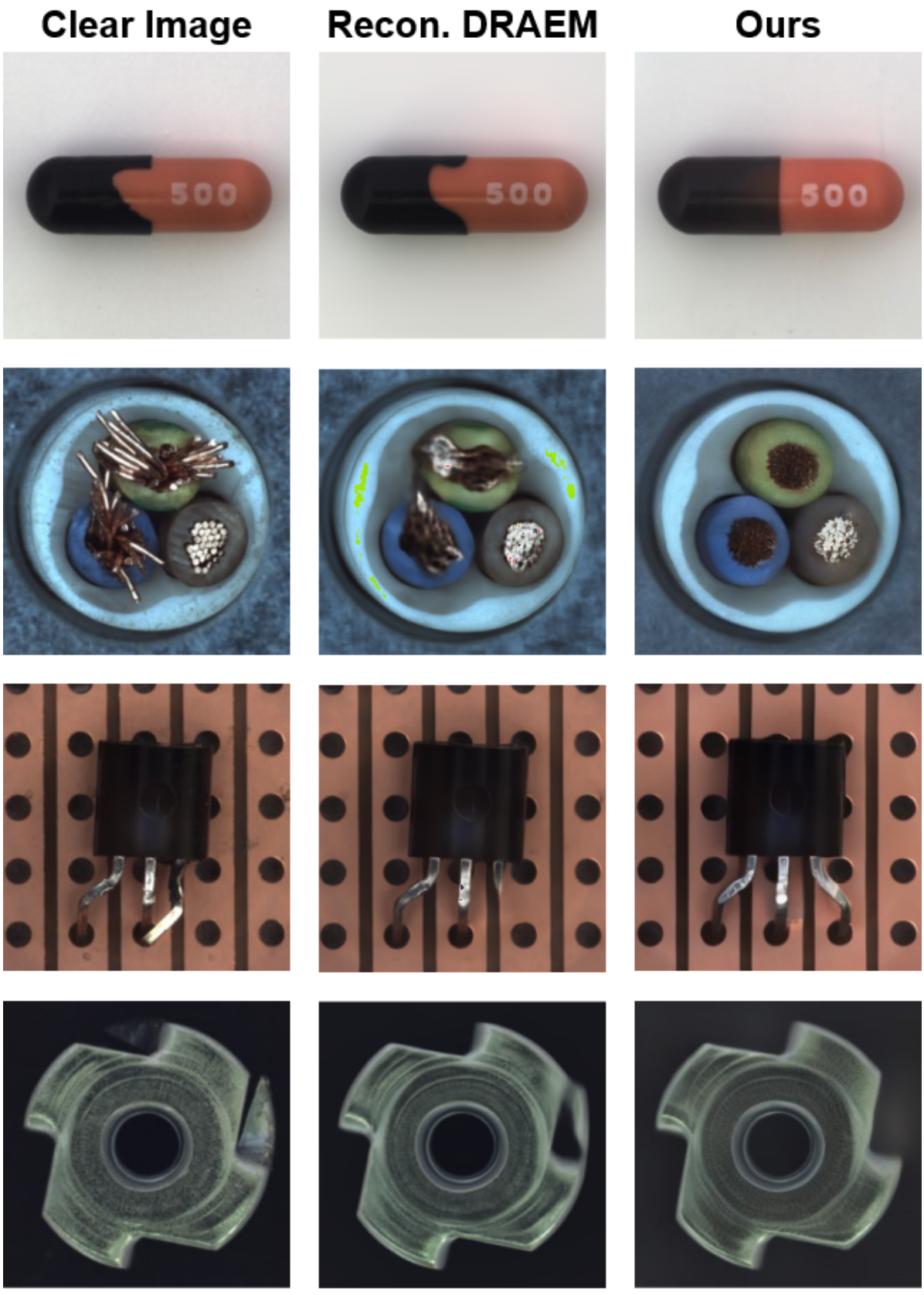

Figure 8. Reconstruction comparison with DRAEM [41] on various MVTec categories.

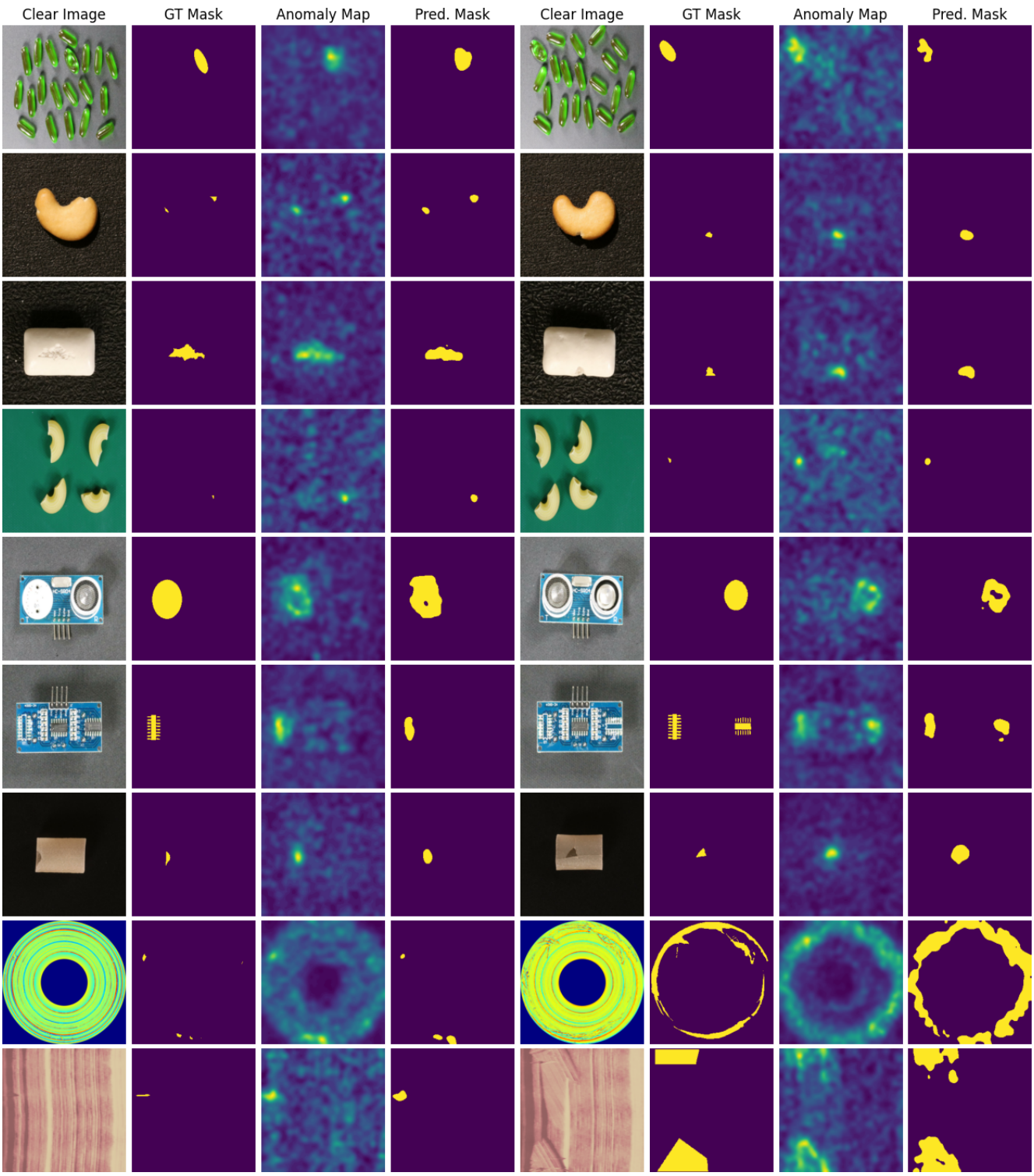

Figure 9. Additional examples from anomalies across all scales from the VisA and BTAD benchmark.

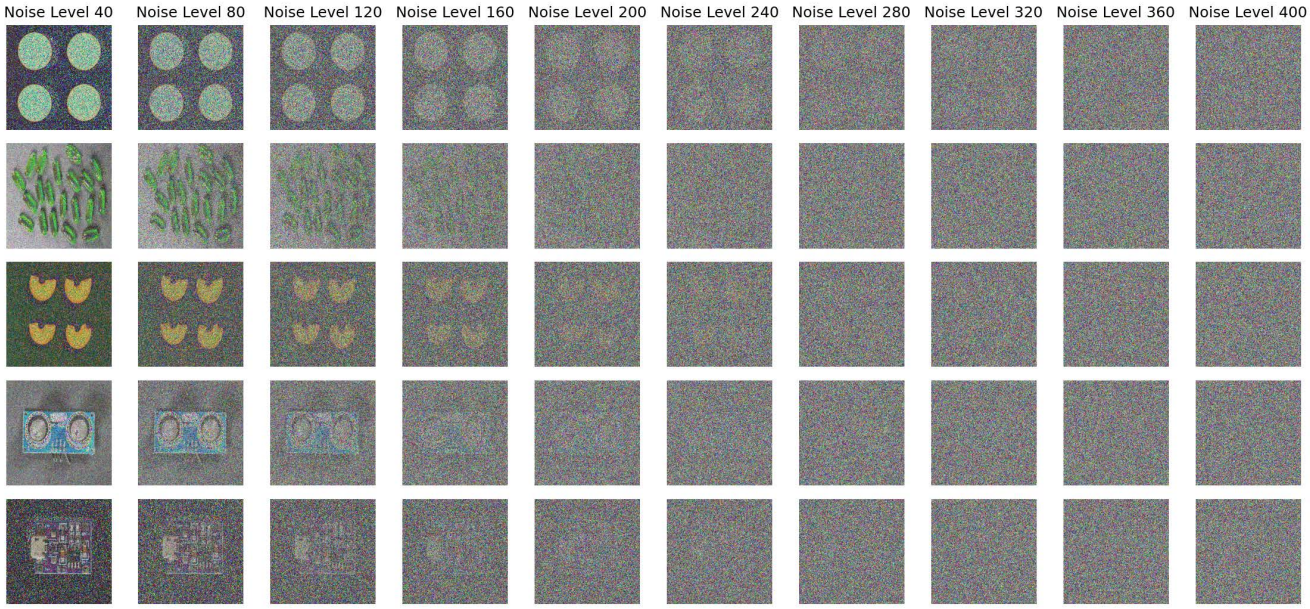

Figure 10. Visualization of the forward diffusion process in pixel space on various categories of the VisA benchmark.

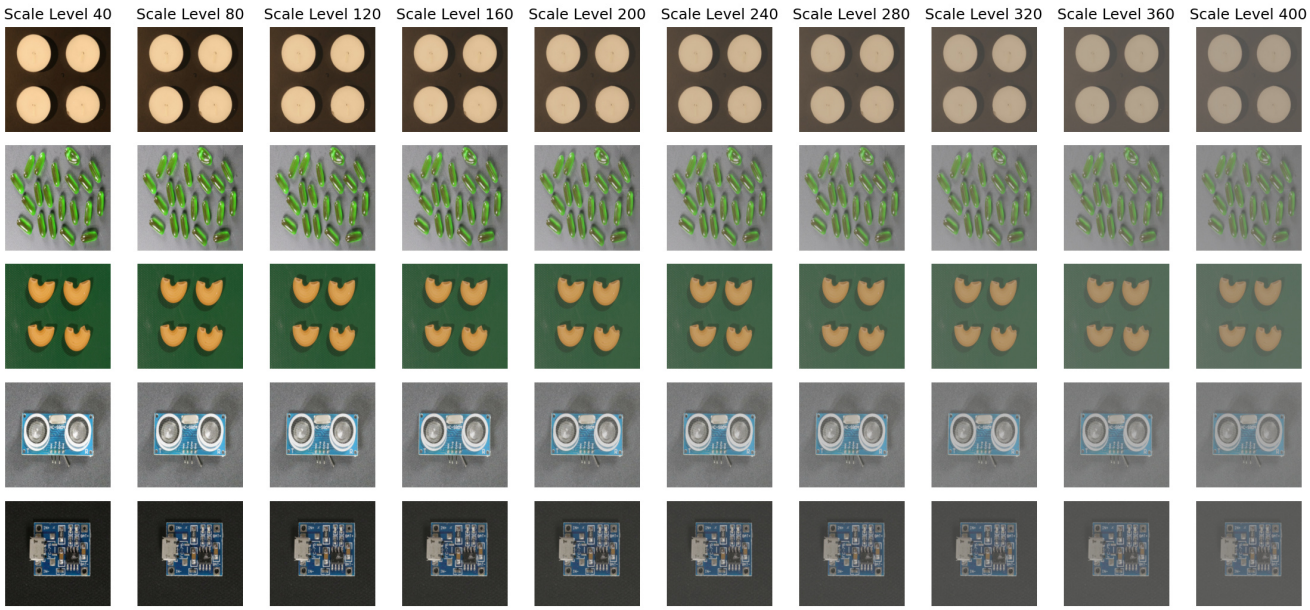

Figure 11. Visualization of the noiseless-forward scaling process in pixel space on various categories of the VisA benchmark.

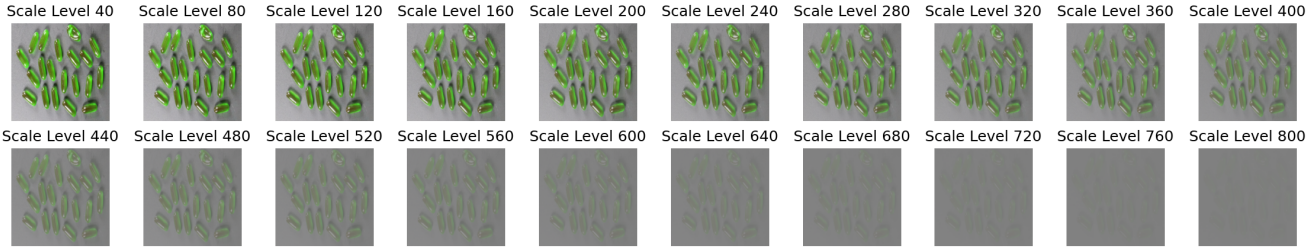

Figure 12. Visualization of the noiseless-forward scaling process in pixel space up to the time step $t = 800$ on the capsules category of the VisA benchmark.

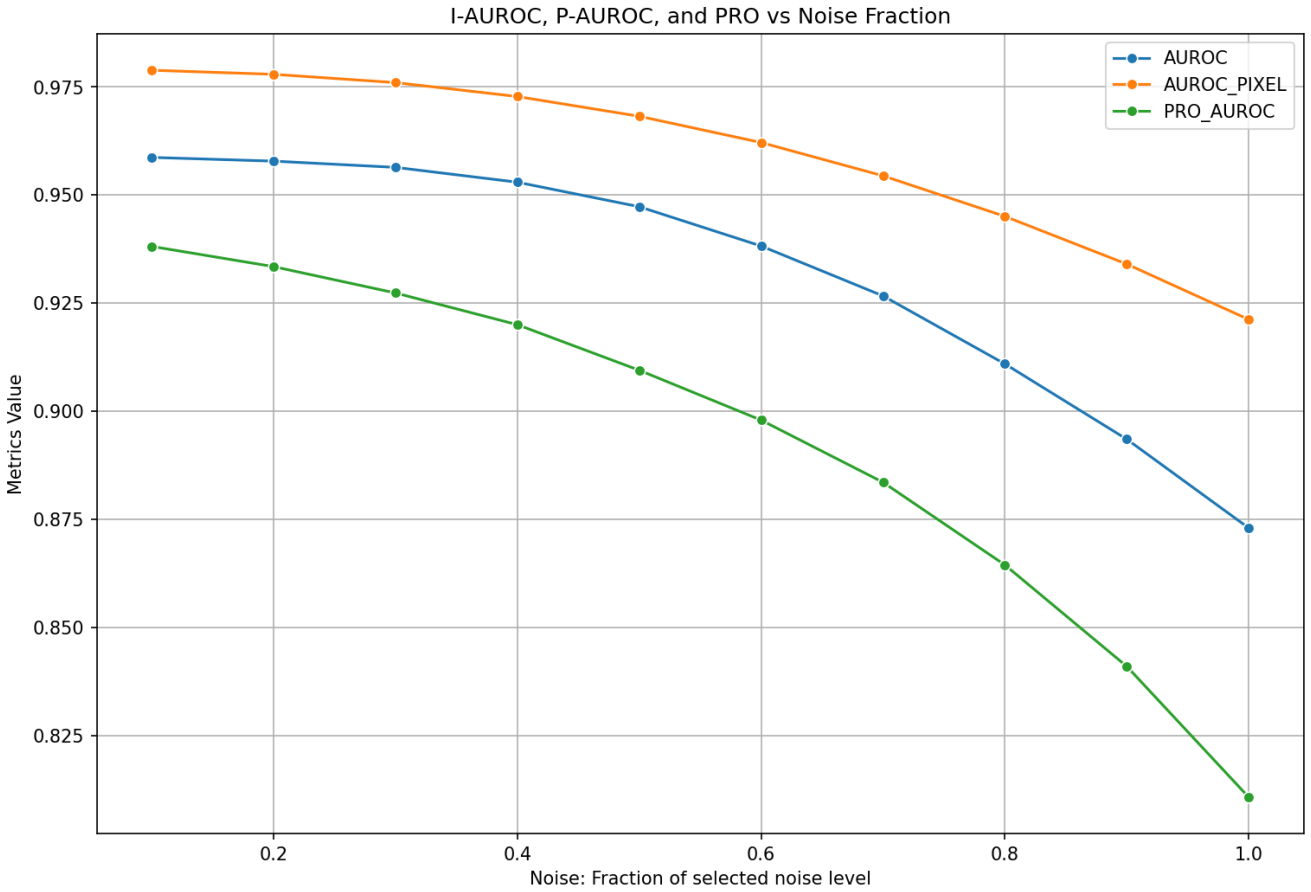

Figure 13. Impact of adding a fraction of the total noise on the VisA benchmark. Showcasing a decline in performance with increasing fraction of the noise.

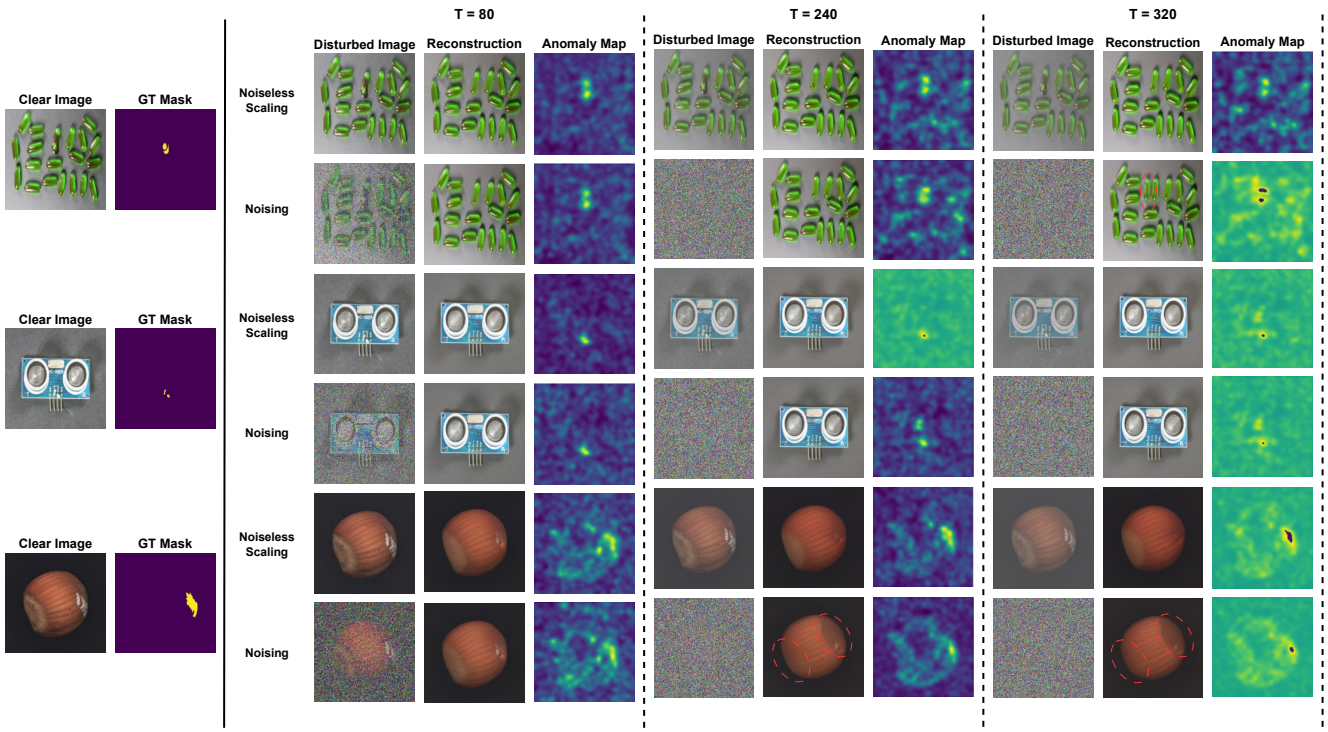

Figure 14. Impact of noiseless scaling versus noising on reconstruction and anomaly map construction on Categories Capsules and PCB1 of VisA and Hazelnut of MVTec. Failed reconstructions are circled in red. The disturbed image level columns are only added for visualization, our approach performs scaling/noising on the latent level.

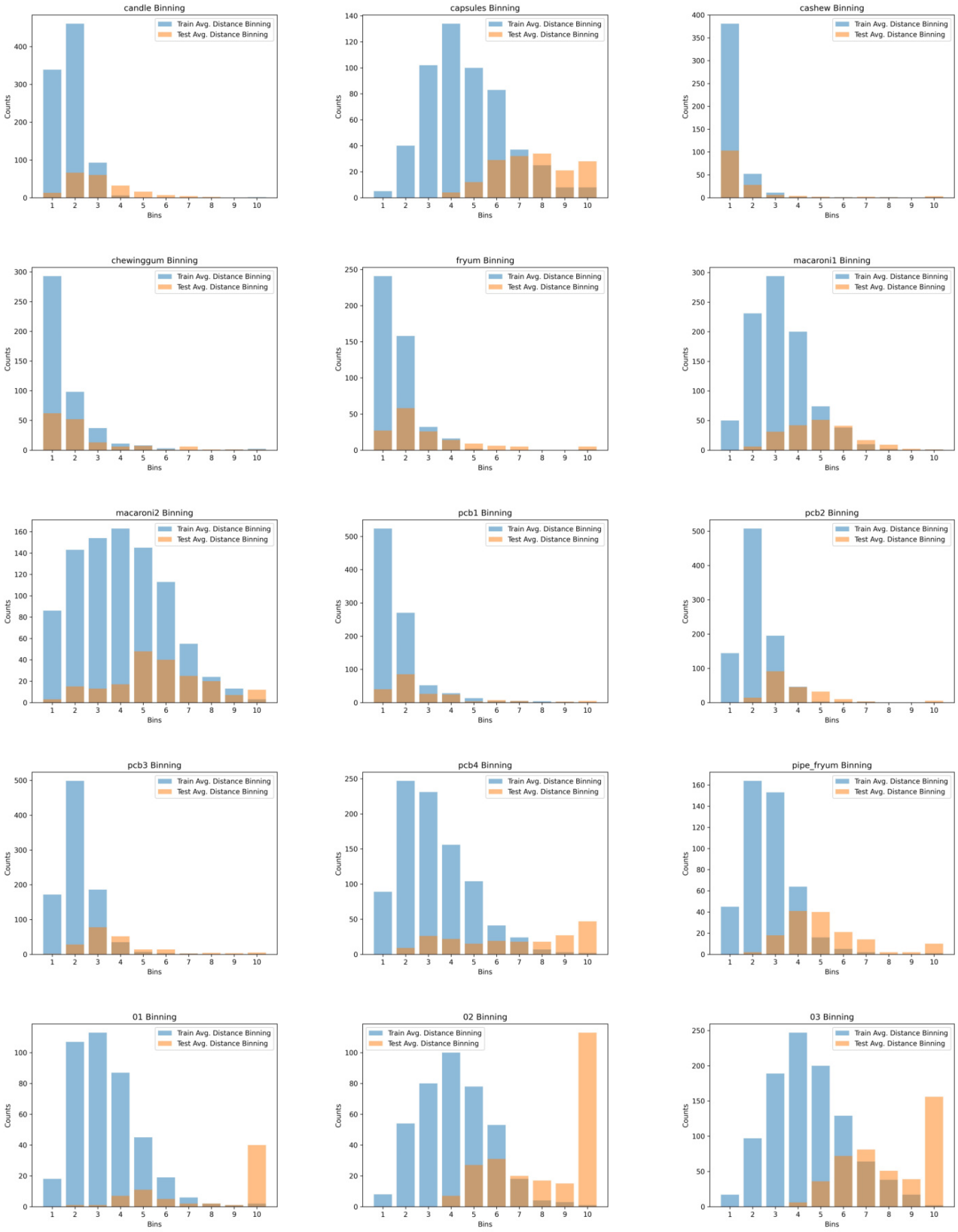

Figure 15. Binning distributions for the training and test set for all categories of the VisA and BTAD benchmark.