Correcting Deviations from Normality: A Reformulated Diffusion Model for Multi-Class Unsupervised Anomaly Detection

Supplementary Material

A. Additional information about the datasets

MVTec-AD [1]: The training set contains 3,629 images with only anomaly-free samples. The test set consists of 1,725 images, comprising 467 normal samples and 1,258 abnormal ones. The anomalous samples exhibit diverse defects, including surface imperfections (e.g., scratches, dents), structural anomalies such as deformed object parts, and defects characterized by missing object components. Pixel-level annotations are provided for the anomaly localization evaluation.

VisA [41]: Including 9,621 normal images and 1,200 anomaly images with 78 types of anomalies. The VisA dataset comprises 12 subsets, each corresponding to a distinct object. 12 objects could be categorized into three object types: Complex structure, Multiple instances, and Single instance. The anomalous images exhibit a range of flaws, including surface defects such as scratches, dents, color spots, and cracks, as well as structural defects like misalignments or missing components.

Additional datasets: In addition to the MVTec-AD and Visa datasets, we have used two additional multi-class anomaly detection datasets to evaluate the generalizability of our model: the Metal Parts Defect Detection (**MPDD** [11]) dataset and the Real-Industrial Anomaly Detection (**Real-IAD** [29]) dataset. The MPDD dataset contains 1,346 multi-view images with pixel-precise defect annotations for 6 distinct industrial metal products. Real-IAD presents a more challenging scenario, encompassing objects from 30 categories with a total of 150K high-resolution, multi-view images, including 99,721 normal images and 51,329 anomalous images. The anomalies in Real-IAD span a broad spectrum, including pits, deformations, abrasions, scratches, damages, missing parts, foreign objects, and contamination.

B. Detailed per-category results

Detailed image-level and pixel-level Mc-UAD anomaly detection results for all categories of MVTec-AD of the proposed method, as well as comparing methods are presented in Tab. 3 and Tab. 4, respectively. Similarly, detailed per-category results for the VisA dataset are presented in Tab. 5 and Tab. 6. These results highlight the effectiveness of our approach, demonstrating its superiority over various state-of-the-art (SoTA) methods across most of the object categories.

C. Quantitaive results on additional datasets

To further validate the effectiveness of the proposed method on Mc-UAD we have conducted additional experiments on new datasets, i.e. MPDD and Real-IAD. For the MPDD dataset, as there are less data samples and categories we have use Medium size model, and for Real-IAD dataset, which includes a greater number of categories with higher complexity, we utilized the X-Large model size. As depicted in Tab. 7[3], the proposed DeCo-Diff demonstrates superior performance, achieving improved results across both image-level and pixel-level metrics for both additional datasets.

D. Additional ablations

D.1. Architecture design of the model.

In this section, we investigate the impact of model size on anomaly detection and localization performance. As mentioned before, we have employed a UNet with attention as the backbone for our deviation correction model η_θ. By default, we used a Large (L) configuration with 256 channels, attention resolution of {4,2,1}, channel-multiplication factor of {1,2,4} across consecutive scales, dropout of 0.4, and 2 residual blocks. We also define model sizes XS, S, M, and XL to have 64, 128, 192, and 320 channels respectively. As indicated in Tab. 1, the larger model sizes usually yield better image-level performance, while pixel-level performance seems to be consistent for all model sizes greater than XS.

Table 1. **Ablation studies on the impact of different model sizes.**

Model size	# params	Image-level AUROC / AUPRC / $f1_{mask}$	Pixel-level AUROC / AUPRC / $f1_{mask}$ / AUPRO
UNet-XS	25M	98.1 / 99.3 / 97.9	98.6 / 73.7 / 68.6 / 93.5
UNet-S	99M	98.7 / 99.5 / 98.3	**98.7** / **75.8** / **70.5** / 94.5
UNet-M	222M	98.8 / 99.6 / 98.4	98.6 / 75.5 / 70.3 / 94.7
UNet-L	395M	99.3 / **99.8** / 98.5	98.4 / 74.9 / 69.7 / **94.9**
UNet-XL	614M	**99.3** / 99.7 / **98.7**	98.5 / 75.5 / 70.2 / 94.5

D.2. Impact of different hyper-parameters.

In this section, the impact of five different hyper-parameters on the Mc-UAD performance of the proposed method has been evaluated, and their results are depicted in Tab. 2.

Patch size: During each iteration of the training phase, a patch-size is randomly selected to create a random mask and

[3]These results are either drawn from the original paper or papers referenced to them

reshuffle the patches. We experimented with three different sets of patch-sizes, each containing 3, 4, and 5 scales, where the scales increase exponentially (powers of 2) starting from a patch size of 1. As detailed in Appendix D, using a larger set of patch sizes, leads to improved performance.

Masking ratio (r_{mask}): In order to create a random mask for each input image, the ratio of masking (r_{mask}) is sampled uniformly from $U[0, R_{mask}]$. As depicted in Tab. 2, we have explored the effect of masking ratio, with three different uniform ranges with the maximum masking ratio (R_{mask}) set to 0.3, 0.5, and 0.7. The results indicated that when the model is exposed to higher masking ratios during training, it could better identify out-of-distribution regions, and therefore achieve better image-level and pixel-level performance.

Shuffled patches ratio: We have investigated the effect of incorporating reshuffled patches within the batch as noise to expose the model to more structured deviations. It should be mentioned that this ratio, is the ratio of shuffled patches to non-masked patches (specified by $r_{shuffle}$), and not the whole patches. As reflected in the Tab. 2, introducing reshuffled patches as noise at a low ratio results in a slight improvement in image-level metrics, albeit at the cost of a decline in pixel-level metrics. On the other hand, replacing a high portion of noise with shuffled patches could decrease both image-level and pixel-level metrics.

γ_p ***and*** γ_l: γ_p and γ_l are devised to prevent assigning excessive weights to locations with large pixel-level and latent-level deviations. We have assessed the impact of these thresholds and the results are shown in Tab. 2. It is worth mentioning that ✗ in the table indicates no threshold or scaling is applied to the deviations. As it can reasonably be anticipated, the image-level metrics did not have much variation. On the other hand, results indicate that applying these thresholds could improve pixel-level metrics, where $\gamma_p = 0.4$ and $\gamma_l = 0.2$ yield the best results.

Table 2. **Ablation studies on the impact of different hyper-parameters**. for each hyper-parameter, the best value is highlighted in bold. Also, the default setting for reported results in the main context is marked with "*".

Hyper-parameter value	Image-level AUROC / AUPRC / $f1_{mask}$	Pixel-level AUROC / AUPRC / $f1_{mask}$ / AUPRO
Patch Size set		
$[1, 2, 4]$	98.7 / 99.5 / 98.0	98.3 / 74.2 / 69.2 / 94.1
*$[1, 2, 4, 8]$	99.3 / **99.8** / 98.5	98.4 / 74.9 / 69.7 / **94.9**
$[1, 2, 4, 8, 16]$	**99.3** / 99.7 / **98.7**	**98.7** / **75.6** / **70.2** / 94.3
Masking Ratio		
R_{mask} : 0.3	98.7 / 99.5 / 98.3	97.6 / 72.4 / 68.0 / 93.7
R_{mask} : 0.5	99.1 / 99.6 / 98.5	98.4 / 74.7 / 69.6 / 94.7
*R_{mask} : 0.7	**99.3** / **99.8** / **98.5**	**98.4** / **74.9** / **69.7** / **94.9**
Shuffled Patches Ratio		
$R_{shuffle}$: 0.0	99.1 / 99.7 / 98.5	**98.6** / **75.9** / **70.5** / 94.8
*$R_{shuffle}$: 0.3	**99.3** / **99.8** / **98.5**	98.4 / 74.9 / 69.7 / **94.9**
$R_{shuffle}$: 0.6	98.9 / 99.6 / 98.1	98.6 / 75.6 / 70.3 / 94.6
γ_p and γ_l		
γ_p : 0.2 - γ_l : 0.4	99.1 / 99.7 / 98.5	98.4 / 75.0 / 69.8 / 94.8
γ_p : 0.4 - γ_l : 0.2	99.1 / 99.7 / 98.3	**98.4** / **75.8** / **70.4** / **94.9**
*γ_p : 0.4 - γ_l : 0.4	99.3 / 99.8 / 98.5	98.4 / 74.9 / 69.7 / 94.9
γ_p : 0.4 - γ_l : 0.6	99.3 / 99.8 / 98.5	98.4 / 74.2 / 69.4 / 94.4
γ_p : 0.4 - γ_l : ✗	99.3 / 99.8 / 98.5	98.4 / 73.2 / 69.0 / 93.9
γ_p : 0.6 - γ_l : 0.4	99.3 / 99.8 / 98.5	98.4 / 74.7 / 69.7 / 94.4
γ_p : ✗ - γ_l : 0.4	99.3 / 99.8 / 98.5	98.4 / 74.3 / 69.6 / 93.7
γ_p : ✗ - γ_l : ✗	**99.3** / **99.8** / **98.5**	98.3 / 72.8 / 68.9 / 93.5

E. Additional Qualitative Results.

We have visualized additional qualitative results for datasets, on 12 classes of MVTec-AD and VisA dataset, respectively in Fig. 1 and Fig. 2 to further support the effectiveness and superiority of the proposed DeCo-Diff model. Also, we have depicted the results for MPDD and Real-IAD datasets respectively in Fig. 3.

F. Limitations.

In this section, we further explore the limitations of the proposed method, as well as analyzing the failures. For this purpose, we have visualized the failures of the methods in Fig. 4, and categorized them into three different subsets, i.e. *"anomaly not detected"*, *"anomaly detected but not fully recovered"*, and *"normal detected as anomaly"*, First of all, we have considered the anomalies in the latent space as noise, while it might be too optimistic. This limitation becomes particularly pronounced when dealing with very large anomalies, which could result in not detecting the large displacements (as "Transistor: in the third column), or not fully recovering the normal counterpart of the image (as "bottle" in the sixth column). Exposing the model to more structured anomalies like synthetic anomalies could serve as a solution for this limitation. Also, as we have used VAE to map the images into the latent space, very small anomalies, like scratches that are barely visible, could be misinterpreted as the variation caused by variance of the VAE model, and therefore not detected. The first, second, and, fourth columns in Fig. 4, are failure examples due to this limitation. Training a better VAE model that is sensitive to these kinds of variations would probably improve the failures caused by this limitation. Furthermore, as DeCo-Diff model corrects the deviation progressively upon a point that the model considers them as in-distribution, in a few cases, there might still be the footprints of the anomaly, as is the case for the fifth, seventh, and eighth columns. This problem could be addressed by directly applying $\tilde{z}_0$ in each reverse time-step as proposed in Sec. 5.3.2 of the main paper, albeit at the cost of a slight decrease in image-level metrics.

Table 3. **Image-level performance on MVTec-AD.** Comparison to state-of-the-art methods on multi-class anomaly detection on the MVTec-AD dataset. AUROC / AUPRC / $F1_{max}$ are reported.

	Category	UniAD [33] *NeurIPS'22*	SimpleNet [16] *CVPR'23*	DeSTSeg [39] *CVPR'23*	DiAD [8] *AAAI'24*	MambaAD [7] *NeurIPS'24*	MoEAD [17] *ECCV'24*	GLAD [32] *ECCV'24*	Deviation Correction *Ours*
Objects	Bottle	99.7 / **100.** / **100.**	**100.** / **100.** / **100.**	98.7 / 99.6 / 96.8	99.7 / 96.5 / 91.8	**100.** / **100.** / **100.**	**100.** / **100.** / **100.**	**100.** / **100.** / **100.**	**100.** / **100.** / **100.**
	Cable	95.2 / 95.9 / 88.0	97.5 / 98.5 / 94.7	89.5 / 94.6 / 85.9	94.8 / 98.8 / 95.2	**98.8** / **99.2** / **95.7**	98.7 / **99.2** / **95.8**	97.4 / 98.6 / 94.2	**98.9** / **99.3** / 95.6
	Capsule	86.9 / 97.8 / 94.4	90.7 / 97.9 / 93.5	82.8 / 95.9 / 92.6	89.0 / 97.5 / 95.5	94.4 / 98.7 / 94.9	93.7 / 98.5 / 96.4	**96.6** / **99.3** / **96.8**	**96.3** / **99.2** / **96.9**
	Hazelnut	99.8 / **100.** / 99.3	99.9 / 99.9 / 99.3	98.8 / 99.2 / 98.6	99.5 / 99.7 / 97.3	**100.** / **100.** / **100.**	**100.** / **100.** / **100.**	97.1 / 98.4 / 94.5	99.6 / 99.8 / 97.8
	MetalNut	99.2 / 99.9 / **99.5**	96.9 / 99.3 / 96.1	92.9 / 98.4 / 92.2	99.1 / 96.0 / 91.6	**99.9** / **100.** / **99.5**	99.8 / **100.** / 98.9	**100.** / **100.** / **99.5**	99.7 / 99.9 / 98.9
	Pill	93.7 / 98.7 / 95.7	88.2 / 97.7 / 92.5	77.1 / 94.4 / 91.7	95.7 / 98.5 / 94.5	**97.0** / **99.5** / **96.2**	94.5 / 98.9 / 95.6	95.5 / 99.2 / 95.0	**99.6** / **99.9** / **99.6**
	Screw	87.5 / 96.5 / 89.0	76.7 / 90.6 / 87.7	69.9 / 88.4 / 85.4	90.7 / **99.7** / **97.9**	94.7 / 97.9 / 94.0	92.8 / 97.4 / 91.4	**94.9** / 98.3 / 93.7	**99.1** / **99.7** / **97.9**
	Toothbrush	94.2 / 97.4 / 95.2	89.7 / 95.7 / 92.3	71.7 / 89.3 / 84.5	**99.7** / **99.9** / **99.2**	98.3 / 99.3 / 98.4	95.0 / 97.8 / 96.8	**99.7** / **99.9** / 98.4	**100.** / **100.** / **100.**
	Transistor	99.8 / 98.0 / 93.8	99.2 / 98.7 / 97.6	78.2 / 79.5 / 68.8	99.8 / 99.6 / 97.4	**100.** / **100.** / **100.**	99.8 / 99.7 / 97.5	99.7 / 99.6 / 97.5	**100.** / **100.** / **100.**
	Zipper	95.8 / 99.5 / 97.1	99.0 / 99.7 / **98.3**	88.4 / 96.3 / 93.1	95.1 / 99.1 / 94.4	**99.3** / **99.8** / 97.5	98.3 / 99.5 / 97.5	97.9 / 99.4 / 96.3	**99.8** / **100.** / **99.2**
Textures	Carpet	**99.8** / **99.9** / **99.4**	95.7 / 98.7 / 93.2	95.9 / 98.8 / 94.9	99.4 / **99.9** / 98.3	**99.8** / **99.9** / **99.4**	**99.8** / **99.9** / **99.4**	96.8 / 99.0 / 95.6	**99.8** / **99.9** / 98.9
	Grid	98.2 / 99.5 / 97.3	97.6 / 99.2 / 96.4	97.9 / 99.2 / 96.6	98.5 / 99.8 / 97.7	**100.** / **100.** / **100.**	99.1 / 99.7 / 98.2	99.8 / 99.9 / 99.1	**100.** / **100.** / **100.**
	Leather	**100.** / **100.** / **100.**	**100.** / **100.** / **100.**	99.2 / 99.8 / 98.9	99.8 / 99.7 / 97.6	**100.** / **100.** / **100.**	**100.** / **100.** / **100.**	99.1 / 99.7 / 97.8	99.7 / 99.9 / 98.9
	Tile	99.3 / 99.8 / 98.2	99.3 / 99.8 / **98.8**	97.0 / 98.9 / 95.3	96.8 / **99.9** / 98.4	98.2 / 99.3 / 95.4	**99.4** / 99.8 / 97.6	**99.9** / **100.** / **99.4**	99.1 / 99.7 / 97.6
	Wood	98.6 / 99.6 / 96.6	98.4 / 99.5 / 96.7	**99.9** / **100.** / **99.2**	**99.7** / **100.** / **100.**	98.8 / 99.6 / 96.6	98.8 / 99.6 / 96.7	94.3 / 98.2 / 93.6	97.3 / 99.1 / 95.9
	Average	96.5 / 98.8 / 96.2	95.3 / 98.4 / 95.8	89.2 / 95.5 / 91.6	97.2 / 99.0 / 96.5	**98.6** / **99.6** / **97.8**	98.0 / 99.3 / 97.5	97.5 / 99.1 / 96.6	**99.3** / **99.8** / **98.5**

Table 4. **Pixel-level performance.** Comparison to state-of-the-art methods on multi-class anomaly detection on the MVTec-AD dataset. The following metrics are reported: AUROC / AUPRC / $f1_{max}$ / AUPRO. For each category, the best method (per metric) is highlighted in **blue**, whereas **red** is used to denote the second-best method.

	Category	UniAD [33] *NeurIPS'22*	SimpleNet [16] *CVPR'23*	DeSTSeg [39] *CVPR'23*	DiAD [8] *AAAI'24*	MambaAD [7] *NeurIPS'24*	MoEAD [17] *ECCV'24*	GLAD [32] *ECCV'24*	Deviation Correction *Ours*
Objects	Bottle	98.1 / 66.0 / 69.2 / 93.1	97.2 / 53.8 / 62.4 / 89.0	93.3 / 61.7 / 56.0 / 67.5	98.4 / 52.2 / 54.8 / 86.6	**98.8** / 79.7 / **76.7** / **95.2**	98.0 / 69.4 / 67.0 / 93.6	98.3 / **80.3** / 74.7 / **96.0**	**98.7** / **86.5** / **79.0** / 95.0
	Cable	97.3 / 39.9 / 45.2 / 86.1	96.7 / 42.4 / 51.2 / 85.4	89.3 / 37.5 / 40.5 / 49.4	96.8 / 50.1 / **57.8** / 80.5	95.8 / 42.2 / 48.1 / **90.3**	**97.7** / **56.8** / 49.1 / 89.6	94.1 / 52.9 / 54.4 / 89.4	**98.4** / **77.8** / **71.1** / **92.9**
	Capsule	98.5 / 42.7 / 46.5 / 92.1	98.5 / 35.4 / 44.3 / 84.5	95.8 / 47.9 / 48.9 / 62.1	97.1 / 42.0 / 45.3 / 87.2	98.4 / 43.9 / 47.7 / 92.6	**98.6** / 48.4 / 44.1 / 90.2	**99.1** / **49.8** / **52.2** / **96.3**	97.8 / **56.5** / **55.9** / **93.4**
	Hazelnut	98.1 / 55.2 / 56.8 / 94.1	98.4 / 44.6 / 51.4 / 87.4	98.2 / 65.8 / 61.6 / 84.5	98.3 / **79.2** / **80.4** / 91.5	**99.0** / 63.6 / 64.4 / **95.7**	97.8 / 54.4 / 52.3 / 92.3	**99.0** / 71.2 / 66.7 / 91.9	98.6 / **75.9** / **68.3** / **95.8**
	MetalNut	62.7 / 14.6 / 29.2 / 81.8	**98.0** / **83.1** / 79.4 / 85.2	84.2 / 42.0 / 22.8 / 53.0	97.3 / 30.0 / 38.3 / 90.6	96.7 / 74.5 / 79.1 / 93.7	94.8 / 68.0 / 58.4 / 88.5	97.3 / 81.2 / **82.3** / **94.2**	**98.3** / **89.0** / **84.9** / **94.4**
	Pill	95.0 / 44.0 / 53.9 / 95.3	96.5 / 72.4 / 67.7 / 81.9	96.2 / 61.7 / 41.8 / 27.9	95.7 / 46.0 / 51.4 / 89.0	97.4 / 64.0 / 66.5 / **95.7**	95.8 / 49.9 / 40.8 / 95.1	**97.8** / **73.9** / **69.4** / 94.7	**99.1** / **81.1** / **78.3** / **96.9**
	Screw	98.3 / 28.7 / 37.6 / 95.2	96.5 / 15.9 / 23.2 / 84.0	93.8 / 19.9 / 25.3 / 47.3	97.9 / **60.6** / **59.6** / 95.0	99.5 / 49.8 / 50.9 / **97.1**	98.8 / 37.1 / 28.5 / 95.1	**99.6** / 47.9 / 48.3 / 96.6	**99.8** / **69.7** / **65.5** / **98.4**
	Toothbrush	98.4 / 34.9 / 45.7 / 87.9	98.4 / 46.9 / 52.5 / 87.4	96.2 / 52.9 / 58.8 / 30.9	99.0 / **78.7** / **72.8** / 95.0	99.0 / 48.5 / 59.2 / 91.7	98.4 / 49.6 / 39.3 / 87.7	**99.2** / 47.1 / 60.0 / **96.0**	**99.4** / **77.6** / **72.3** / **95.5**
	Transistor	**97.9** / 59.5 / **64.6** / **93.5**	95.8 / 58.2 / 56.0 / 83.2	73.6 / 38.4 / 39.2 / 43.9	95.1 / 15.6 / 31.7 / 90.0	96.5 / **69.4** / **67.1** / 87.0	**97.6** / **63.7** / 56.5 / **93.9**	89.5 / 55.9 / 56.6 / 86.1	92.4 / 57.8 / 54.9 / 87.4
	Zipper	96.8 / 40.1 / 49.9 / 92.6	97.9 / 53.4 / 54.6 / 90.7	97.3 / **64.7** / 59.2 / 66.9	96.2 / 60.7 / 60.0 / 91.6	**98.4** / 60.4 / **61.7** / **94.3**	97.7 / 49.4 / 39.2 / 93.0	92.9 / 41.7 / 47.1 / 83.8	**99.5** / **84.1** / **76.8** / **97.9**
Textures	Carpet	98.5 / 49.9 / 51.1 / 94.4	97.4 / 38.7 / 43.2 / 90.6	93.6 / 59.9 / 58.9 / 89.3	98.6 / 42.2 / 46.4 / 90.6	**99.2** / 60.0 / 63.3 / **96.7**	98.2 / 50.1 / 46.6 / 94.0	98.8 / **71.9** / **68.3** / 95.0	**99.3** / **82.0** / **74.6** / **97.2**
	Grid	63.1 / 10.7 / 11.9 / 92.9	96.8 / 20.5 / 27.6 / 88.6	97.0 / 42.1 / 46.9 / 86.8	96.6 / **66.0** / **64.1** / 94.0	99.2 / 47.4 / 47.7 / 97.0	97.4 / 27.4 / 22.3 / 91.7	**99.4** / 40.8 / 45.3 / **97.6**	**99.5** / **63.9** / **61.0** / **97.9**
	Leather	98.8 / 32.9 / 34.4 / 96.8	98.7 / 28.5 / 32.9 / 92.7	**99.5** / **71.5** / **66.5** / 91.1	98.8 / 56.1 / **62.3** / 91.3	99.4 / 50.3 / 53.3 / **98.7**	98.6 / 31.7 / 30.1 / 96.7	**99.7** / **62.2** / 61.2 / **97.0**	99.3 / 58.8 / 58.7 / 95.4
	Tile	91.8 / 42.1 / 50.6 / 78.4	95.7 / 60.5 / 59.9 / 90.6	93.0 / 71.0 / 66.2 / 87.1	92.4 / 65.7 / 64.1 / 90.7	93.8 / 45.1 / 54.8 / 80.0	91.6 / 50.4 / 42.6 / 78.8	**97.9** / **72.8** / **75.1** / **96.6**	**98.2** / **86.4** / **76.5** / **91.4**
	Wood	93.2 / 37.2 / 41.5 / 86.7	91.4 / 39.7 / 34.8 / 76.3	95.9 / **77.3** / **71.3** / 83.4	93.3 / 43.3 / 43.5 / **97.5**	94.4 / 46.2 / 48.2 / 91.2	92.8 / 39.9 / 35.1 / 85.1	**96.8** / 68.6 / 63.1 / 86.7	**97.6** / **75.9** / **68.4** / **93.5**
	Average	96.8 / 43.4 / 49.5 / 90.7	96.9 / 45.9 / 49.7 / 86.5	93.1 / 54.3 / 50.9 / 64.8	96.8 / 52.6 / 55.5 / 90.7	**97.7** / 56.3 / 59.2 / **93.1**	96.9 / 49.8 / 43.5 / 91.0	97.4 / **60.8** / **60.7** / 93.0	**98.4** / **74.9** / **69.7** / **94.9**

Table 5. **Image-level performance on VisA.** Comparison to state-of-the-art methods on multi-class anomaly detection on the VisA dataset. AUROC / AUPRC / $f1_{max}$ are reported.

	Category	UniAD [33] *NeurIPS'22*	SimpleNet [16] *CVPR'23*	DeSTSeg [39] *CVPR'23*	DiAD [8] *AAAI'24*	MambaAD [7] *NeurIPS'24*	MoEAD [17] *ECCV'24*	GLAD [32] *ECCV'24*	Deviation Correction *Ours*
Complex	PCB1	92.8 / 92.7 / 87.8	91.6 / 91.9 / 86.0	87.6 / 83.1 / 83.7	88.1 / 88.7 / 80.7	95.4 / 93.0 / 91.6	**97.7** / **97.3** / **95.2**	78.1 / 79.8 / 74.0	**96.7** / **95.9** / **94.6**
	PCB2	87.8 / 87.7 / 83.1	92.4 / 93.3 / 84.5	86.5 / 85.8 / 82.6	91.4 / 91.4 / 84.7	94.2 / 93.7 / 89.3	**95.1** / **95.6** / **90.1**	88.0 / 86.8 / 82.8	**97.2** / **96.5** / **93.3**
	PCB3	78.6 / 78.6 / 76.1	89.1 / 91.1 / 82.6	93.7 / 95.1 / 87.0	86.2 / 87.6 / 77.6	93.7 / 94.1 / 86.7	92.2 / 92.6 / 85.2	**95.9** / **96.1** / **87.7**	**97.7** / **97.9** / **94.1**
	PCB4	98.8 / 98.8 / 94.3	97.0 / 97.0 / 93.5	97.8 / 97.8 / 92.7	99.6 / 99.5 / 97.0	**99.9** / **99.9** / **98.5**	**99.7** / **99.7** / 97.0	99.3 / 99.1 / **97.5**	98.5 / 97.3 / 94.6
Multiple	Macaroni1	79.9 / 79.8 / 72.7	85.9 / 82.5 / 73.1	76.6 / 69.0 / 71.0	85.7 / 85.2 / 78.8	91.6 / 89.8 / 81.6	**93.0** / **93.2** / 85.8	91.5 / 91.7 / **86.2**	**95.0** / **95.4** / **88.4**
	Macaroni2	71.6 / 71.6 / 69.9	68.3 / 54.3 / 59.7	68.9 / 62.1 / 67.7	62.5 / 57.4 / 69.6	81.6 / 78.0 / 73.8	**86.3** / **88.7** / **80.4**	73.8 / 71.2 / 71.8	**88.4** / **90.2** / **80.4**
	Capsules	55.6 / 55.6 / 76.9	74.1 / 82.8 / 74.6	87.1 / 93.0 / 84.2	58.2 / 69.0 / 78.5	91.8 / 95.0 / **88.8**	77.6 / 87.8 / 79.7	**92.4** / **95.9** / 88.0	**95.8** / **97.7** / **91.2**
	Candle	94.1 / 94.0 / 86.1	84.1 / 73.3 / 76.6	94.9 / 94.8 / 89.2	92.8 / 92.0 / 87.6	**96.8** / **96.9** / **90.1**	**97.2** / **97.3** / **92.8**	88.1 / 88.8 / 81.8	95.6 / 95.5 / 88.1
Single	Cashew	92.8 / 92.8 / 91.4	88.0 / 91.3 / 84.7	92.0 / 96.1 / 88.1	91.5 / 95.7 / 89.7	94.5 / 97.3 / 91.1	90.7 / 95.3 / 89.2	**96.6** / **98.5** / **94.6**	**97.8** / **98.5** / **95.6**
	ChewingGum	96.3 / 96.2 / 95.2	96.4 / 98.2 / 93.8	95.8 / 98.3 / 94.7	**99.1** / 99.5 / 95.9	97.7 / 98.9 / 94.2	98.9 / **99.6** / **98.5**	**99.3** / **99.7** / **97.0**	97.6 / 98.7 / 93.3
	Fryum	83.0 / 83.0 / 85.0	88.4 / 93.0 / 83.3	92.1 / 96.1 / 89.5	89.8 / 95.0 / 87.2	95.2 / 97.7 / 90.5	90.8 / 95.8 / 88.4	**98.8** / **99.4** / **96.6**	**99.3** / **99.6** / **98.7**
	PipeFryum	94.7 / 94.7 / 93.9	90.8 / 95.5 / 88.6	94.1 / 97.1 / 91.9	96.2 / 98.1 / 93.7	**98.7** / **99.3** / **97.0**	96.7 / 98.4 / 95.0	**99.7** / **99.9** / **98.0**	97.7 / 98.6 / 94.5
	Average	85.5 / 85.5 / 84.4	87.2 / 87.0 / 81.8	88.9 / 89.0 / 85.2	86.8 / 88.3 / 85.1	**94.3** / 94.5 / 89.4	93.0 / **95.1** / **89.8**	91.8 / 92.2 / 88.0	**96.4** / **96.8** / **92.2**

Table 6. **Pixel-level performance.** Comparison to state-of-the-art methods on multi-class anomaly detection on the VisA dataset. The following metrics are reported: AUROC / AUPRC / $f1_{max}$ / AUPRO. For each category, the best method (per metric) is highlighted in **blue**, whereas **red** is used to denote the second-best method.

	Category	UniAD [33] *NeurIPS'22*	SimpleNet [16] *CVPR'23*	DeSTSeg [39] *CVPR'23*	DiAD [8] *AAAI'24*	MambaAD [7] *NeurIPS'24*	MoEAD [17] *ECCV'24*	GLAD [32] *ECCV'24*	Deviation Correction *Ours*
Complex	PCB1	93.3 / 3.9 / 8.3 / 64.1	99.2 / **86.1** / **78.8** / 83.6	95.8 / 46.4 / 49.0 / 83.2	98.7 / 49.6 / 52.8 / 80.2	**99.8** / **77.1** / **72.4** / **92.8**	**99.6** / 64.1 / 68.2 / 92.0	97.5 / 38.1 / 45.9 / 91.9	99.5 / 66.0 / 69.8 / **94.0**
	PCB2	93.9 / 4.2 / 9.2 / 66.9	96.6 / 8.9 / 18.6 / 85.7	97.3 / 14.6 / **28.2** / 79.9	95.2 / 7.5 / 16.7 / 67.0	**98.9** / 13.3 / 23.4 / 89.6	98.4 / **19.0** / 11.7 / 86.0	97.5 / 5.4 / 12.5 / **90.8**	**99.1** / **56.2** / **55.3** / **90.8**
	PCB3	97.3 / 13.8 / 21.9 / 70.6	97.2 / **31.0** / **36.1** / 85.1	97.7 / 28.1 / 33.4 / 62.4	96.7 / 8.0 / 18.8 / 68.9	**99.1** / 18.3 / 27.4 / 89.1	**98.9** / 26.0 / 25.0 / 84.3	97.0 / 24.9 / 27.6 / **95.3**	98.7 / **49.1** / **52.0** / **90.1**
	PCB4	94.9 / 14.7 / 22.9 / 72.3	93.9 / 23.9 / 32.9 / 61.1	95.8 / **53.0** / **53.2** / 76.9	97.0 / 17.6 / 27.2 / 85.0	**98.6** / 47.0 / 46.9 / **87.6**	97.8 / 34.9 / 29.4 / 85.0	**99.4** / **52.2** / **53.2** / **94.6**	96.3 / 46.5 / 44.2 / 84.0
Multiple	Macaroni1	97.4 / 3.7 / 9.7 / 84.0	98.9 / 3.5 / 8.4 / 92.0	99.1 / 5.8 / 13.4 / 62.4	94.1 / 10.2 / 16.7 / 68.5	99.5 / 17.5 / 27.6 / 95.2	99.5 / **21.5** / 11.9 / **96.5**	**99.9** / 18.4 / **32.6** / **99.2**	**99.6** / **42.0** / **36.8** / 96.3
	Macaroni2	95.2 / 0.9 / 4.3 / 76.6	93.2 / 0.6 / 3.9 / 77.8	98.5 / 6.3 / 14.4 / 70.0	93.6 / 0.9 / 2.8 / 73.1	**99.5** / 9.2 / **16.1** / **96.2**	98.5 / **14.6** / 6.6 / 91.4	**99.6** / 5.7 / 12.2 / **98.0**	98.6 / **28.1** / **24.7** / **96.2**
	Capsules	88.7 / 3.0 / 7.4 / 43.7	97.1 / 52.9 / 53.3 / 73.7	96.9 / 33.2 / 9.1 / 76.7	97.3 / 10.0 / 21.0 / 77.9	99.1 / **61.3** / **59.8** / 91.8	98.9 / 58.4 / 59.4 / 80.6	**99.3** / 48.4 / 52.0 / **92.1**	**99.8** / **70.9** / **71.0** / **94.4**
	Candle	98.5 / 17.6 / 27.9 / 91.6	97.6 / 8.4 / 16.5 / 87.6	98.7 / **39.9** / **45.8** / 69.0	97.3 / 12.8 / 22.8 / 89.4	99.0 / 23.2 / 32.4 / **95.5**	**99.3** / 34.8 / 25.7 / **94.6**	98.9 / 26.5 / 34.2 / 94.0	**99.1** / **37.0** / **36.3** / **94.6**
Single	Cashew	98.6 / 51.7 / 58.3 / 87.9	**98.9** / **68.9** / **66.0** / 84.1	87.9 / 47.6 / 52.1 / 66.3	90.9 / 53.1 / **60.9** / 61.8	94.3 / 46.8 / 51.4 / 87.8	98.2 / 50.3 / 45.9 / **90.2**	84.9 / 24.1 / 34.3 / 60.3	**99.0** / **54.6** / 57.0 / **94.2**
	ChewingGum	98.8 / 54.9 / 56.1 / 81.3	97.9 / 26.8 / 29.8 / 78.3	98.8 / **86.9** / **81.0** / 68.3	94.7 / 11.9 / 25.8 / 59.5	98.1 / 57.5 / 59.9 / 79.7	99.3 / 59.6 / 59.3 / **84.1**	**99.7** / **78.5** / 73.1 / **93.3**	**99.4** / 73.3 / **79.9** / 81.6
	Fryum	95.9 / 34.0 / 40.6 / 76.2	93.0 / 39.1 / 45.4 / 85.1	88.1 / 35.2 / 38.5 / 47.7	**97.6** / **58.6** / **60.1** / 81.3	96.9 / 47.8 / **51.9** / 91.6	**97.4** / **53.0** / 44.9 / 84.1	97.2 / 39.8 / 47.1 / **96.6**	93.9 / 45.9 / 42.0 / **92.5**
	PipeFryum	98.9 / 50.2 / 57.7 / 91.5	98.5 / 65.6 / 63.4 / 83.0	98.9 / **78.8** / **72.7** / 45.9	**99.4** / **72.7** / **69.9** / 89.9	99.1 / 53.5 / 58.5 / 95.1	99.0 / 55.3 / 51.3 / 94.7	99.1 / 53.8 / 59.1 / **98.4**	**99.4** / 46.0 / 45.0 / **96.3**
	Average	95.9 / 21.0 / 27.0 / 75.6	96.8 / 34.7 / 37.8 / 81.4	96.1 / **39.6** / 43.4 / 67.4	96.0 / 26.1 / 33.0 / 75.2	**98.5** / 39.4 / **44.0** / 91.0	**98.7** / 36.6 / 41.0 / 88.6	97.5 / 34.6 / 40.3 / **92.0**	**98.5** / **51.3** / **51.2** / **92.1**

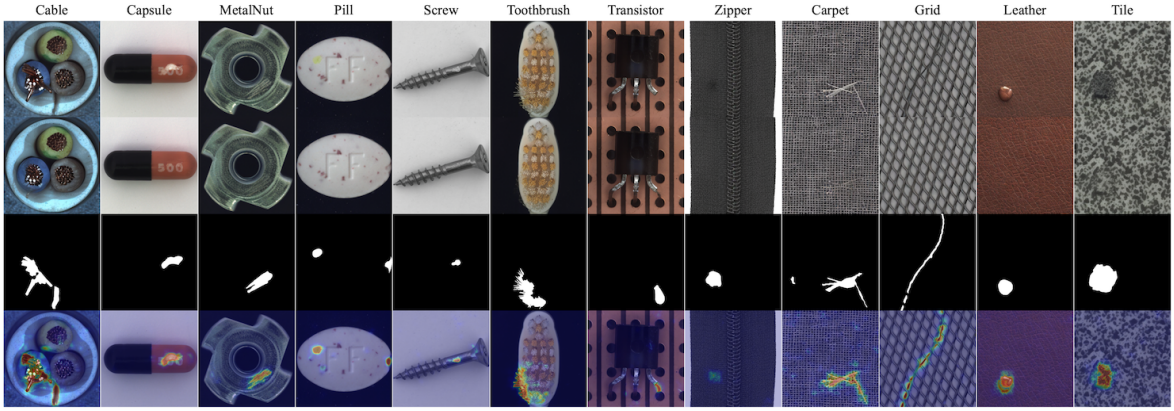

Figure 1. **Additional qualitative results on MVTec-AD dataset.** From *top* to *bottom*: the original input image (with anomalies), DeCo-Diff reconstruction, the ground truth mask, and the predicted anomaly mask across different objects of MVTec-AD dataset.

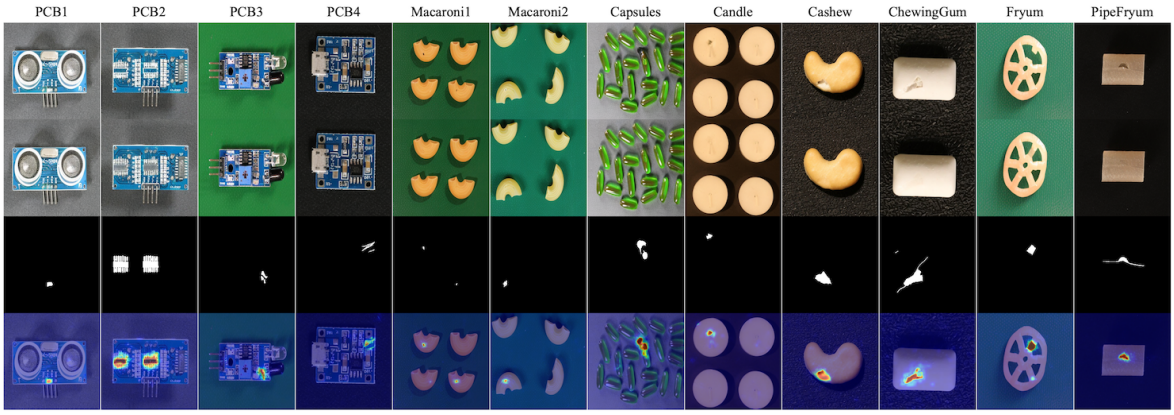

Figure 2. **Additional qualitative results on VisA dataset.** From *top* to *bottom*: the original input image (with anomalies), DeCo-Diff reconstruction, the ground truth mask, and the predicted anomaly mask across different objects of VisA dataset.

Table 7. **Quantitatve evaluation on additional datasets.**. Image and Pixel-level results on **MPDD** and **Real-IAD** datasets in *multi-class UAD*. The best method (per metric) is highlighted in **blue**, whereas **red** is used to denote the second-best approach.

Dataset	Method	Image-level			Pixel-level			
		AUROC	AUPRC	$f1_{max}$	AUROC	AUPRC	$f1_{max}$	AUPRO
MPDD [11]	RD4AD [3] *CVPR'22*	90.3	92.8	90.5	**98.3**	39.6	40.6	95.2
	UniAD [33] *NeurIPS'22*	80.1	83.2	85.1	95.4	19.0	25.6	83.8
	SimpleNet [16] *CVPR'23*	90.6	94.1	89.7	97.1	33.6	35.7	90.0
	DeSTSeg [39] *CVPR'23*	92.6	91.8	92.8	90.8	30.6	32.9	78.3
	DiAD [8] *AAAI'24*	85.8	89.2	86.5	91.4	15.3	19.2	66.1
	GLAD [32] *ECCV'24*	**97.5**	**97.1**	**96.8**	**98.0**	**40.9**	**41.5**	**93.0**
	MambaAD [7] *NeurIPS'24*	89.2	93.1	90.3	97.7	33.5	38.6	**92.8**
	DeCo-Diff (*Ours*)	**97.7**	**97.3**	**95.3**	95.1	**45.3**	**46.6**	79.5
Real-IAD [29]	RD4AD [3] *CVPR'22*	82.4	79.0	73.9	97.3	25.0	32.7	**89.6**
	UniAD [33] *NeurIPS'22*	83.0	80.9	74.3	97.3	21.1	29.2	86.7
	SimpleNet [16] *CVPR'23*	57.2	53.4	61.5	75.7	2.8	6.5	39.0
	DeSTSeg [39] *CVPR'23*	82.3	79.2	73.2	94.6	**37.9**	**41.7**	40.6
	DiAD [8] *AAAI'24*	75.6	66.4	69.9	88.0	2.9	7.1	58.1
	MambaAD [7] *NeurIPS'24*	**86.3**	**84.6**	**77.0**	**98.5**	33.0	38.7	**90.5**
	DeCo-Diff (*Ours*)	**87.0**	**86.1**	**79.2**	**97.4**	**46.4**	**48.6**	88.8

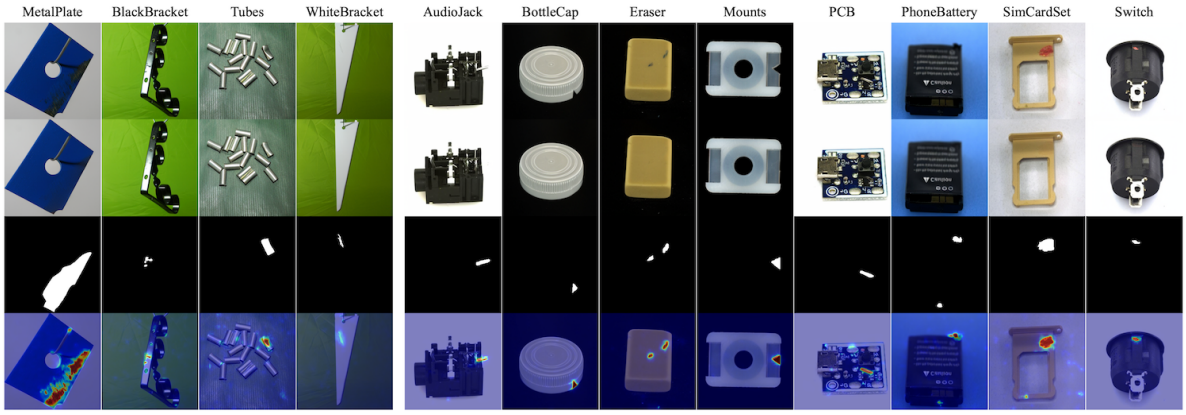

Figure 3. **Qualitative results on Additional datasets.** From *top* to *bottom*: the original input image (with anomalies), DeCo-Diff reconstruction, the ground truth mask, and the predicted anomaly mask for MPDD dataset (left side), and Real-IAD dataset(right side).

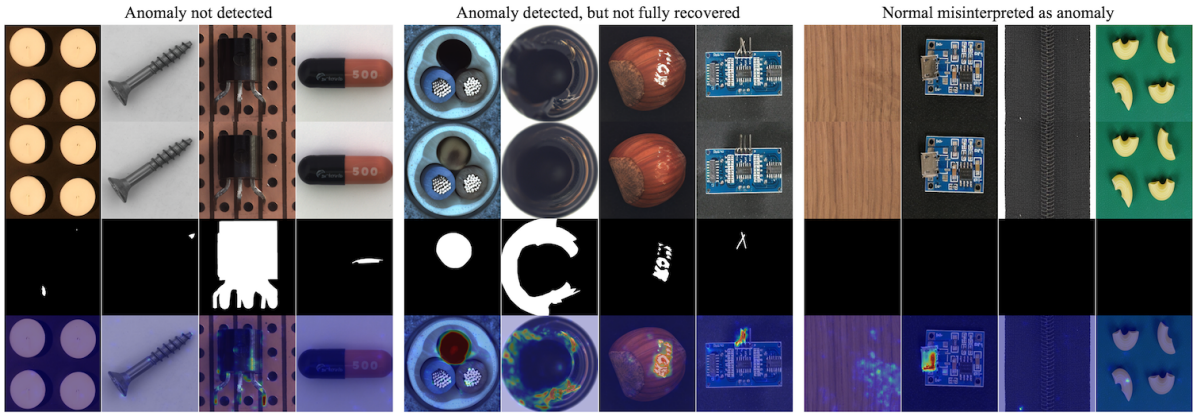

Figure 4. **Example of Failures on MVTec-AD and Visa datasets.** Failures are categorized into three subsets, i.e. "Anomaly not detected", "Anomaly detected but not fully recovered", and "normal misinterpreted as an anomaly." from right to left respectively. For each image, from *top* to *bottom*: the original input image, DeCo-Diff reconstruction, the ground truth mask, and the predicted anomaly are depicted.