RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection

Supplementary Material

A. Overview

We organize this supplementary material into the following sections: Appendix B provides additional implementation details for RealNet. Appendix C provides detailed results on the BTAD [12] and VisA [28] datasets, supplementary ablation study results, an analysis of RealNet's computational efficiency, anomaly detection results in multi-class setting, as well as synthetic anomaly image quality assessment results. Appendix D offers additional visualization results, including qualitative results of RealNet in anomaly localization, images generated by SDAS, and a straightforward visualization result of AFS. Appendix E discusses the limitations of our method.

B. More details

In SDAS, we use the learnable reverse diffusion variance [13] as $\Sigma_\theta(x_t, t)$, given by:

$$\Sigma_\theta(x_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t) \tag{S1}$$

Here, β_t represents the variance of the diffusion process, while $\tilde{\beta}_t$ represents the variance of the conditional posterior distribution $q(x_{t-1}|x_t, x_0)$, and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. The vector v is predicted by the model and weighted with β_t and $\tilde{\beta}_t$ in the log space. We optimize $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ with the loss $\mathcal{L}_{hybrid}$:

$$\mathcal{L}_{hybrid} = \mathcal{L}_{simple} + \gamma \mathcal{L}_{vlb} \tag{S2}$$

where

$$\begin{aligned} \mathcal{L}_{vlb} &= \mathcal{L}_0 + \mathcal{L}_1 + ... + \mathcal{L}_{T-1} + \mathcal{L}_T \\ \mathcal{L}_0 &= -\log p_\theta(x_0|x_1) \\ \mathcal{L}_{t-1} &= D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \\ \mathcal{L}_T &= D_{KL}(q(x_T|x_0)||p(x_T)) \end{aligned} \tag{S3}$$

We set γ to 0.001 in Eq. (S2), and stop the gradient of $\mu_\theta(x_t, t)$ in $\mathcal{L}_{vlb}$ during the training phase. To accelerate the convergence of the diffusion model, we initialize it with weights pre-trained on ImageNet [5]. We set the reverse diffusion step T of 20, and generating 10,000 images at a resolution of 256×256 takes 6 hours using a single NVIDIA GeForce RTX 3090.

The SDAS with DDIM [17] is described in Algorithm S1, which provides three options for applying perturbation variance in the deterministic reverse diffusion process: $\Sigma = \beta_t$, $\Sigma = \tilde{\beta}_t$, and $\Sigma = \Sigma_\theta(x_t, t)$. Experimental observations show that the anomaly images obtained by IDDPM [13] are slightly better than those obtained by DDIM [17], and therefore, we use IDDPM [13] for SDAS. Some examples can be found in Appendix D.

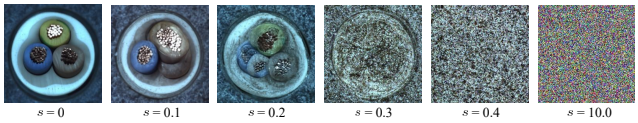

Figure S1. Sample anomaly images generated by SDAS with different anomaly strengths s.

Fig. S1 presents examples of images generated by SDAS with a broader range of anomaly strengths. As the anomaly strength increases, the generated anomalous images contain more noise, reducing their authenticity. In the experiments, we set the anomaly strength between 0.1 and 0.2, allowing SDAS to encompass a wider range of real-world anomalies.

In RRS, the global reconstruction residual $E(A_n)$ originates from distinct reconstruction networks, leading to disparate distributions across its dimensions. We apply a `BatchNorm` [9] layer (without Affine) to $E(A_n)$ and then perform reconstruction residuals selection to ensure a consistent distribution across the dimensions of $E(A_n)$.

The discriminator is implemented using a basic MLP with upsampling layers to map anomaly scores from feature resolution to image resolution. During the training phase of RealNet, we do not use any data augmentation for the synthesis of anomalous images, and maintain an equal ratio between normal images and synthetic anomalous images. In the process of image blending, we uniformly sample the opacity δ from 0.5 to 1.0 in Eq. (3). The training of RealNet is performed on a single NVIDIA GeForce RTX 3090, with an approximate average training time of 2 hours.

C. More results

C.1. Experimental results on BTAD

We evaluate the anomaly detection and localization performance of RealNet and alternative methods on the BTAD dataset [12], with the results shown in Tab. S1. Although SIA does not show a significant performance improvement compared to DTD [2] due to the absence of complex structural anomalies in the three industrial products of the BTAD dataset [12], RealNet demonstrates state-of-the-art performance in anomaly detection and localization when compared to other methods, without any structural or hyperparameter tuning.

Algorithm S1 SDAS with DDIM [17]

Input: diffusion model $\epsilon_\theta(x_t, t)$, perturbation variance Σ, anomaly strength s
$x_T \sim \mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
$x_{t-1} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t), s\Sigma)$
end for
return x_0

Table S1. Comparison of RealNet with alternative anomaly detection methods on the BTAD dataset [12], employing Image AUROC (%) and Pixel AUROC (%) as evaluation metrics.

Category	VT-ADL [12]	P-SVDD [21]	FastFlow [23]	SPADE [3]	RD++ [20]	RealNet (SIA)	RealNet (DTD [2])
01	(-, **99**)	(95.7, 91.6)	(-, 95)	(91.4, 97.3)	(96.8, 96.2)	(**100.0**, 98.2)	(**100.0**, 98.1)
02	(-, 94)	(72.1, 93.6)	(-, 96)	(71.4, 94.4)	(**90.1**, **96.4**)	(88.6, 96.3)	(87.5, 96.3)
03	(-, 77)	(82.1, 91.0)	(-, 99)	(99.9, 99.1)	(**100.0**, **99.7**)	(99.6, 99.4)	(99.4, 99.6)
AVG	(-, 90.0)	(83.3, 92.1)	(-, 96.7)	(87.6, 96.9)	(95.6,97.4)	(**96.1**, 97.9)	(95.7, **98.0**)

Table S2. Comparison of RealNet with alternative anomaly detection methods on the VisA dataset [28], employing Image AUROC (%) and Pixel AUROC (%) as evaluation metrics.

Category	SPADE [3]	FastFlow [23]	DRAEM [25]	PatchCore [14]	RealNet (SIA)	RealNet (DTD [2])
Candle	(91.0, 97.9)	(92.8, 94.9)	(91.8, 96.6)	(**98.6**, **99.5**)	(96.1, 99.1)	(95.0, 99.0)
Capsules	(61.4, 60.7)	(71.2, 75.3)	(74.7, 98.5)	(81.6, **99.5**)	(**93.2**, 98.7)	(88.1, 97.6)
Cashew	(**97.8**, 86.4)	(91.0, 91.4)	(95.1, 83.5)	(97.3, **98.9**)	(**97.8**, 98.3)	(95.9, 97.6)
Chewing gum	(85.8, 98.6)	(91.4, 98.6)	(94.8, 96.8)	(99.1, 99.1)	(99.9, **99.8**)	(**100.0**, **99.8**)
Fryum	(88.6, 96.7)	(88.6, **97.3**)	(**97.4**, 87.2)	(96.2, 93.8)	(97.1, 96.2)	(95.3, 95.2)
Macaroni1	(95.2, 96.2)	(98.3, 97.3)	(97.2, **99.9**)	(97.5, 99.8)	(**99.8**, **99.9**)	(98.2, 99.7)
Macaroni2	(87.9, 87.5)	(86.3, 89.2)	(85.0, 99.2)	(78.1, 99.1)	(**95.2**, **99.6**)	(91.8, 99.3)
PCB1	(72.1, 66.9)	(77.4, 75.2)	(47.6, 88.7)	(**98.5**, **99.9**)	(**98.5**, 99.7)	(97.1, 99.4)
PCB2	(50.7, 71.1)	(61.9, 67.3)	(89.8, 91.3)	(97.3, **99.0**)	(**97.6**, 98.0)	(97.5, 97.8)
PCB3	(90.5, 95.1)	(74.3, 94.8)	(92.0, 98.0)	(97.9, **99.2**)	(**99.1**, 98.8)	(97.6, 98.4)
PCB4	(83.1, 89.0)	(80.9, 89.9)	(98.6, 96.8)	(99.6, **98.6**)	(**99.7**, **98.6**)	(99.2, **98.6**)
Pipe fryum	(81.1, 81.8)	(72.0, 87.3)	(**100.0**, 85.8)	(99.8, 99.1)	(99.9, **99.2**)	(99.9, 98.6)
AVG	(82.1, 85.6)	(82.2, 88.2)	(88.7, 93.5)	(95.1, **98.8**)	(**97.8**, **98.8**)	(96.3, 98.4)

C.2. Experimental results on VisA

We present the performance of RealNet and alternative methods on the VisA dataset under the one-class protocol [28] in Tab. S2. RealNet achieves the best performance in both anomaly detection and localization. Compared to DTD [2], the RealNet trained using SIA shows an improvement of 1.5% in Image AUROC and 0.4% in Pixel AUROC.

C.3. Supplementary ablation studies

To further investigate RealNet's anomaly detection performance on the MVTec-AD dataset [1], we examine various backbones and reconstruction feature dimension settings. As shown in Tab. S3, when WideResNet50 [24] is employed as the backbone and the reconstruction feature dimensions $\{m_1, ..., m_K\}$ are reduced from {256, 512, 512, 256} to {128, 256, 256, 128}, there is a slight decrease of 0.16% in Image AUROC. Despite this reduction, RealNet maintains its competitive performance compared to other methods. Additionally, the adoption of EfficientNetB4 [18] and ResNet34 [7] as backbones also results in competitive performance, demonstrating the effectiveness of RealNet across various settings.

C.4. Computational efficiency analysis

We investigate the computational efficiency and detection performance of three different multi-scale feature reconstruction architectures on the MVTec-AD dataset [1], as illustrated in Fig. S2. To provide a comprehensive analysis, Tab. S4 presents the inference speed, model size (including backbone), and anomaly detection performance of these architectures. The inference is performed using a single Nvidia GeForce RTX 3090, with all other settings adhering to the specifications detailed in Sec. 4.1.

We utilize a consistent reconstruction network based on the U-Net model with skip connections across three distinct architectures. The employed U-Net model initiates with a

Table S3. Performance evaluation of RealNet with varying backbones and reconstruction feature dimension settings on the MVTec-AD dataset [1], employing Image AUROC (%), Pixel AUROC (%), and PRO (%) as evaluation metrics.

Backbone	EfficientNetB4 [18]	ResNet34 [7]	WideResNet50 [24]	
$\{m_1,...,m_K\}$	{24, 32, 56, 160}	{64, 128, 256, 128}	{128, 256, 256, 128}	{256, 512, 512, 256}
Bottle	(**100.0**, 98.83, **95.96**)	(**100.0**, 98.56, 95.91)	(**100.0**, **99.41**, 94.37)	(**100.0**, 99.30, 95.62)
Cable	(96.36, 96.33, 88.61)	(96.31, 96.32, 88.68)	(98.35, 98.01, 92.99)	(**99.19**, **98.10**, **93.38**)
Capsule	(97.97, 99.16, **91.46**)	(96.81, 98.78, 87.87)	(99.44, **99.39**, 79.76)	(**99.56**, 99.32, 84.48)
Carpet	(**100.0**, 98.27, 96.35)	(99.76, 98.37, 94.45)	(99.80, 98.91, 96.32)	(99.84, **99.19**, **96.41**)
Grid	(99.92, 99.31, 97.35)	(**100.0**, 99.26, **97.39**)	(**100.0**, **99.55**, 96.38)	(**100.0**, 99.51, 97.28)
Hazelnut	(99.89, 98.45, **94.98**)	(99.93, 99.35, 94.36)	(**100.0**, 99.67, 93.06)	(**100.0**, **99.68**, 93.14)
Leather	(**100.0**, 99.34, 97.75)	(99.97, 99.40, **98.28**)	(**100.0**, **99.81**, 96.99)	(**100.0**, 99.76, 96.22)
Metal Nut	(99.07, 96.90, 92.65)	(99.17, 96.68, 93.34)	(**99.90**, **98.75**, **95.10**)	(99.76, 98.58, 94.39)
Pill	(96.10, 94.86, 86.60)	(97.55, 98.23, **93.17**)	(97.85, **99.19**, 80.73)	(**99.13**, 99.02, 91.04)
Screw	(92.95, 99.05, **92.68**)	(96.99, 99.09, 89.57)	(97.99, 99.28, 88.60)	(**98.83**, **99.45**, 87.90)
Tile	(99.49, 95.69, 92.10)	(99.93, 97.40, 91.65)	(**100.0**, 99.27, 97.20)	(99.96, **99.44**, **97.70**)
Toothbrush	(99.44, 98.90, **92.39**)	(**100.0**, 98.26, 91.74)	(**100.0**, **99.26**, 91.22)	(99.44, 98.71, 91.57)
Transistor	(99.58, **98.57**, **93.63**)	(99.33, 97.70, 88.53)	(99.79, 98.26, 83.34)	(**100.0**, 98.00, 92.92)
Wood	(98.77, 94.47,**92.67**)	(98.16, 96.35, 91.46)	(**99.56**, **98.22**, 90.76)	(99.21, **98.22**, 90.54)
Zipper	(99.71, 98.01, 91.68)	(**99.90**, 98.55, **93.91**)	(99.74, **99.20**, 90.73)	(99.82, 99.17, 93.43)
AVG	(98.62, 97.74, **93.12**)	(98.92, 98.15, 92.69)	(99.49, **99.07**, 91.17)	(**99.65**, 99.03, 93.07)

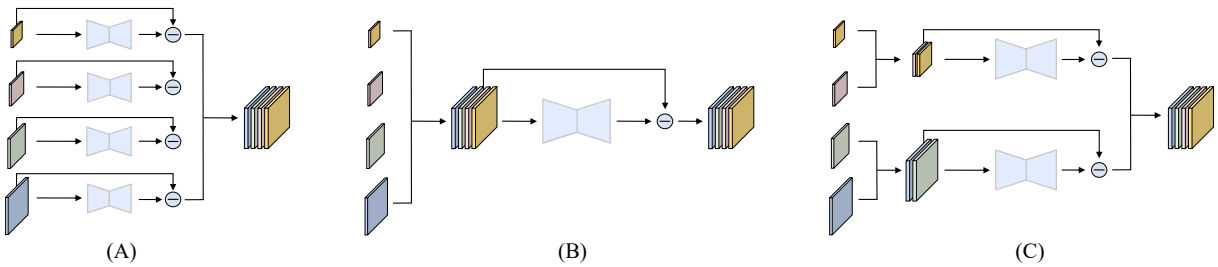

Figure S2. Various architectures of multi-scale feature reconstruction for anomaly detection. (A) Independent Reconstruction Architecture uses separate networks for multi-scale feature reconstruction. (B) Fully Aligned Feature Reconstruction Architecture aligns all features for reconstruction. (C) Neighboring Aligned Feature Reconstruction Architecture aligns and reconstructs neighboring resolution features.

stack of residual layers and down-sampling layers, gradually decreasing the spatial dimensions while increasing the number of channels. Subsequently, the model utilizes a stack of residual layers and up-sampling layers to inversely reconstruct features. Throughout this process, skip connections are incorporated at equivalent spatial resolutions to ensure a smooth and logical flow.

Specifically, architecture **A** adopts separate reconstruction networks to reconstruct multi-scale features without the need for feature interpolation or alignment. This method ensures outstanding anomaly detection performance while maintaining high computational efficiency. With a resolution of 256×256 and reconstruction feature dimensions of $\{256, 512, 512, 256\}$, architecture **A** with model size of 2.2 GB achieves a rapid inference speed of 31.93 FPS. And it can perform inference using only 4GB of GPU memory. Concurrently, it attains an Image AUROC of 99.65% and a Pixel AUROC of 99.03%. By decreasing the reconstruction feature dimensions to $\{128, 256, 256, 128\}$, architecture **A** reduces the model size to 0.74 GB and achieves a higher inference speed of 40.42 FPS, while preserving an Image AUROC of 99.49% and a Pixel AUROC of 99.07%. Furthermore, at a high resolution of 512×512, it delivers an inference speed of 13.53 FPS, along with an Image AUROC of 99.40% and a Pixel AUROC of 98.71%. These inference speeds indicate that architecture **A** satisfies the real-time requirements for industrial inspection applications.

Regarding architecture **B**, as referenced in [16, 19, 22], it is used to align the multi-scale features of a small pre-trained network. As aligning down-sampled features will reduce the resolution of model detection and cause predictable performance loss, the experiment only discusses

Table S4. Performance evaluation of various reconstruction architectures on the MVTec-AD dataset [1]. The metrics include Image AUROC (%), Pixel AUROC (%), and PRO (%).

	Speed (FPS) ↑	Model Size (GB) ↓	Metrics ↑
$\{m_1,...,m_K\}$ is {128, 256, 256, 128} and image size is 256×256			
A	**40.42**	**0.74**	(99.49, **99.07**, 91.17)
$\{m_1,...,m_K\}$ is {256, 512, 512, 256} and image size is 256×256			
A	31.93	2.20	(**99.65**, 99.03, 93.07)
B	10.83	7.22	(98.44, 98.17, 94.27)
C	22.39	3.75	(99.62, 98.90, **94.71**)
$\{m_1,...,m_K\}$ is {256, 512, 512, 256} and image size is 512×512			
A	13.53	2.20	(99.40, 98.71, 94.01)

up-sampling alignment. Compared to architecture **A**, architecture **B** reconstructs the interpolated features, significantly reducing computational efficiency and increasing model size. Moreover, due to the limited number of normal images, the overly large reconstruction network in architecture **B** is prone to overfitting, resulting in reduced detection performance. Consequently, for large-scale pre-trained networks with high-dimensional features, aligning and reconstructing all features is suboptimal.

Moreover, we observe that utilizing multiple reconstruction networks for feature reconstruction in architecture **A** causes minor deviations in localizing small-area anomalies, resulting in a reduced PRO. To address this, we propose architecture **C**, which aligns and reconstructs features from two neighboring resolution, thereby reducing the number of reconstruction networks, controlling the model size, and striking a balance between computational efficiency and localization accuracy. At a 256×256 resolution, with reconstruction feature dimensions of {256, 512, 512, 256}, architecture **C** has a 3.75 GB model size and achieves an inference speed of 22.39 FPS, while attaining an Image AUROC of 99.62%, a Pixel AUROC of 98.90%, and a PRO of 94.71%.

In summary, the design of RealNet balances both anomaly detection performance and computational efficiency. The introduction of AFS allows us to flexibly customize models of various sizes to accommodate different usage scenarios. Furthermore, among our three key innovations, both AFS and RRS introduce no additional learnable parameters, ensuring strong interpretability. As for SDAS, it only introduces perturbation during the reverse diffusion process, without requiring any prior knowledge about the distribution of real anomaly images.

C.5. Anomaly detection in multi-class setting

In the multi-class setting [22, 27], anomaly detection is performed across multiple target classes concurrently, without access to sample class labels during both training and inference phases. Learning the data distributions of multiple classes jointly makes the reconstruction more complex.

Table S5. Comparison of RealNet with alternative methods in multi-class anomaly detection on the MVTec-AD dataset [1].

Methods	Image AUROC	Pixel AUROC
DRAEM [25]	88.1	87.2
PaDiM [4]	84.2	89.5
UniAD [22]	96.5	96.8
OmniAL [27]	97.2	98.3
RealNet	**97.3**	**98.4**

Table S6. Image quality comparison of SIA with alternative anomaly synthesis approaches on the MVTec-AD dataset [1].

Methods	FID [8] ↓	LPIPS [26] ↑
DTD [2]	120.52±0.63	0.16±0.00
CutPaste [11]	77.34±0.09	0.11±0.00
NSA [15]	68.76±0.16	0.09±0.01
SIA	**60.39±1.26**	**0.18±0.01**

In such settings, previous reconstruction methods tend to output copies of the input images instead of performing selective reconstruction, which leads to a significant decrease in performance. We evaluate the performance of RealNet in multi-class anomaly detection on the MVTec-AD dataset [1] and compare it with alternative state-of-the-art methods. We use DTD [2] for anomaly synthesis as class labels are unavailable during training. The remaining settings are consistent with Sec. 4.1.

The results are shown in Tab. S5. When detecting anomalies across 15 categories of the MVTec-AD dataset [1] concurrently, RealNet achieves an Image AUROC of 97.3% and a Pixel AUROC of 98.4% using a ResNet50 [7] pre-trained on ImageNet [5], surpassing state-of-the-art multi-class anomaly detection methods [22, 27]. To ensure that normal regions can be reconstructed correctly, we do not explicitly constrain the generalization ability of the reconstructed network in RealNet. Instead, we implicitly constrain the reconstruction network to ensure that anomalous regions can be correctly detected by discarding a part of the reconstruction residuals.

C.6. Synthetic anomaly image quality assessment

In this section, we evaluate the quality of anomaly images generated by various anomaly synthesis methods on the MVTec-AD dataset [1]. Specifically, we use the following evaluation metrics:

- FID (Fréchet Inception Distance) [8]: FID measures the distance between the distribution of synthetic anomaly images and real anomaly images, evaluating both the realism and diversity of the synthetic anomaly images. A lower value indicates better performance.

- LPIPS (Learned Perceptual Image Patch Similarity) [26]: We employ cluster-based LPIPS [6] to evaluate the diversity of synthetic anomaly images. Supposing a category contains N real anomaly images, we partition the synthesized anomaly images into N groups by finding the lowest LPIPS, then we compute the mean pairwise LPIPS within each group and compute the average of all groups. A higher cluster LPIPS indicates greater diversity.

We employ various anomaly synthesis methods to generate 1,000 anomaly images for evaluation, with each method independently assessed three times. The experimental results are shown in Tab. S6. In comparison to other anomaly synthesis methods, SIA achieves the best FID and LPIPS metrics, highlighting the outstanding performance of SDAS in generating both realistic and diverse anomaly images, and demonstrating the effectiveness of SDAS in improving anomaly detection performance.

D. Visualization

We conduct a comprehensive visual analysis of RealNet on the four datasets. Fig. S3 shows the qualitative results of RealNet in anomaly localization, showcasing its outstanding performance in pixel-level anomaly localization. Figs. S4 and S5 display the anomaly images and normal images generated by SDAS, respectively. Fig. S6 illustrates images synthesized using SIA with localized anomalous regions. Fig. S7 provides an intuitive explanation of pre-training bias, indicating that not all feature maps contribute equally to anomaly detection and localization, which validates the efficacy of AFS.

E. Limitations

In some categories with more texture anomalies, such as the texture categories in MVTec-AD dataset [1], SIA's performance may slightly underperform when compared to DTD [2]. Given that DTD dataset [2] includes a diverse range of real-world texture images, it effectively simulates common anomaly types in the textural category, such as color, oil, and glue. Nonetheless, SIA excels in the majority of scenarios, outperforming DTD [2] and offering superior capability in synthesizing anomalies in images with intricate structures.

Compared to anomaly synthesis methods based on data augmentation [11, 15] or external data [25], SDAS increases additional offline training time. For instance, we generate 10,000 anomaly images at a resolution of 256×256 for each category, and it will take 6 hours using a single NVIDIA GeForce RTX 3090. However, it is pivotal to clarify that RealNet omits SDAS without any additional computational cost during inference and real-world applications. Therefore, we believe that the slight increase in training time to enhance performance is necessary and worthwhile.

In order to achieve higher computational efficiency, we do not upsample multi-scale features. Instead, we employ multiple reconstruction networks for feature reconstruction, which reduce the resolution of anomaly detection. The lower feature reconstruction resolution may introduce minor deviations in localizing small anomalous areas, leading to a decrease in PRO. However, we found that increasing the resolution of anomaly detection by reducing the number of reconstruction networks can improve PRO. For instance, architecture **C** in Fig. S2 achieved a higher PRO score of 94.71%. Furthermore, increasing the resolution of images can also lead to an improvement in PRO, as detailed in Tab. S4.

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, 3, 4, 5, 7, 8, 9, 10

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1, 2, 4, 5

[3] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 2

[4] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2021. 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4

[6] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 571–578, 2023. 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 4

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 1

[10] 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. 7, 8, 9, 10

[11] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 4, 5

[12] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 1, 2, 7, 8, 9, 10

[13] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1

[14] 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. 2

[15] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 474–489. Springer, 2022. 4, 5

[16] Yong Shi, Jie Yang, and Zhiquan Qi. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing*, 424:9–22, 2021. 3

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 2

[18] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2, 3

[19] Xian Tao, Dapeng Zhang, Wenzhi Ma, Zhanxin Hou, Zhenfeng Lu, and Chandranath Adak. Unsupervised anomaly detection for surface defects with dual-siamese network. *IEEE Transactions on Industrial Informatics*, 18(11):7707–7717, 2022. 3

[20] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24511–24520, 2023. 2

[21] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[22] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, 2022. 3, 4

[23] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 2

[24] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2016. 2, 3, 11

[25] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem: A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 2, 4, 5

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 5

[27] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3924–3933, 2023. 4

[28] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 392–408. Springer, 2022. 1, 2, 7, 8, 9, 10

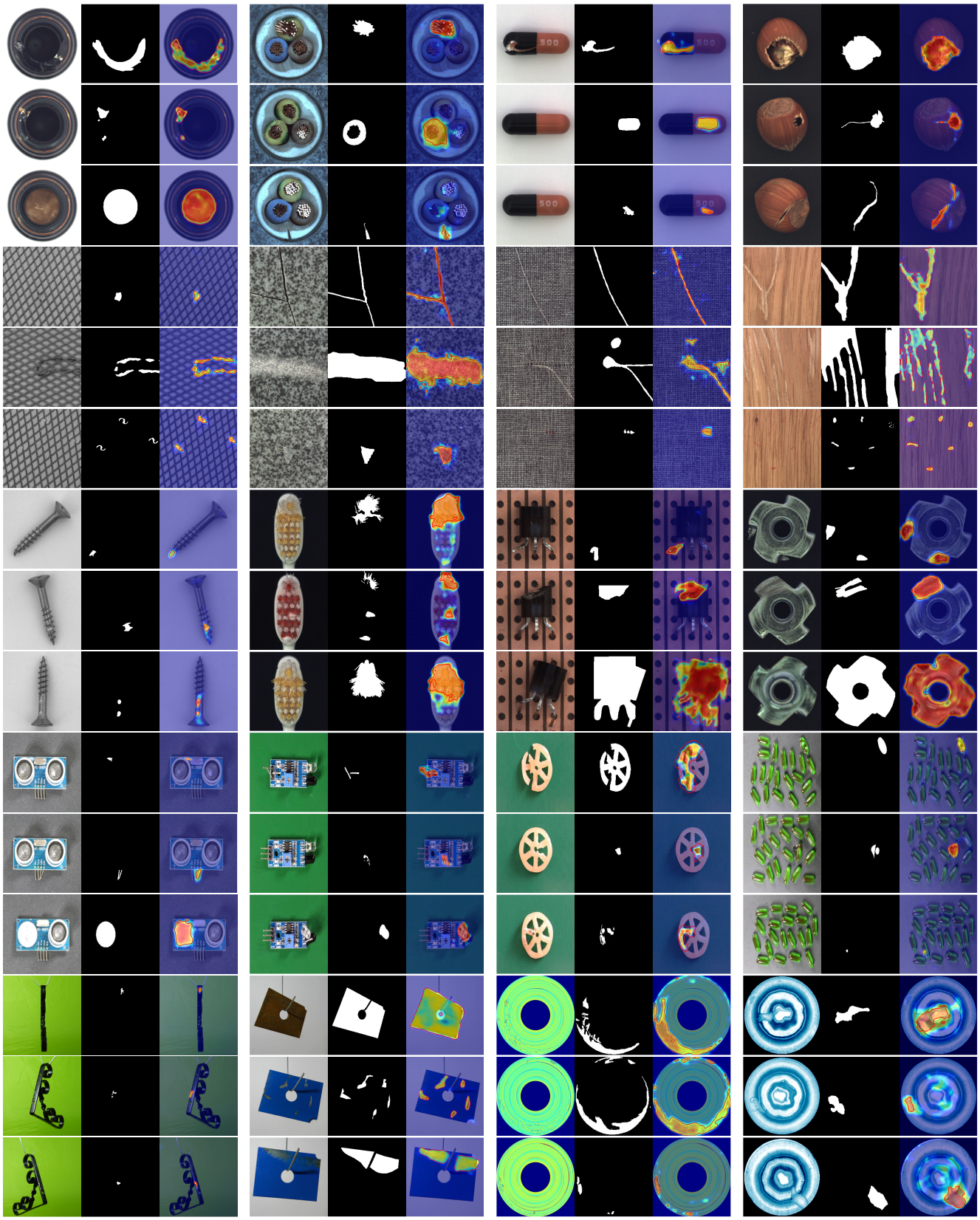

Figure S3. Qualitative results of RealNet. Within each group, from left to right, are the anomaly image, ground-truth, and predicted anomaly score. The examples are from the MVTec-AD [1], MPDD [10], BTAD [12], and VisA [28] datasets.

Figure S4. Anomaly images generated by SDAS. The examples are from the MVTec-AD [1], MPDD [10], BTAD [12], and VisA [28] datasets. Within each group, from top to bottom, the anomaly strength gradually increases.

Figure S5. Normal images generated by SDAS (when $s = 0$). The examples are from the MVTec-AD [1], MPDD [10], BTAD [12], and VisA [28] datasets.

Figure S6. Local anomaly images synthesized by SIA. The examples are from the MVTec-AD [1], MPDD [10], BTAD [12], and VisA [28] datasets. Within each group, from top to bottom, the anomaly strength gradually increases.

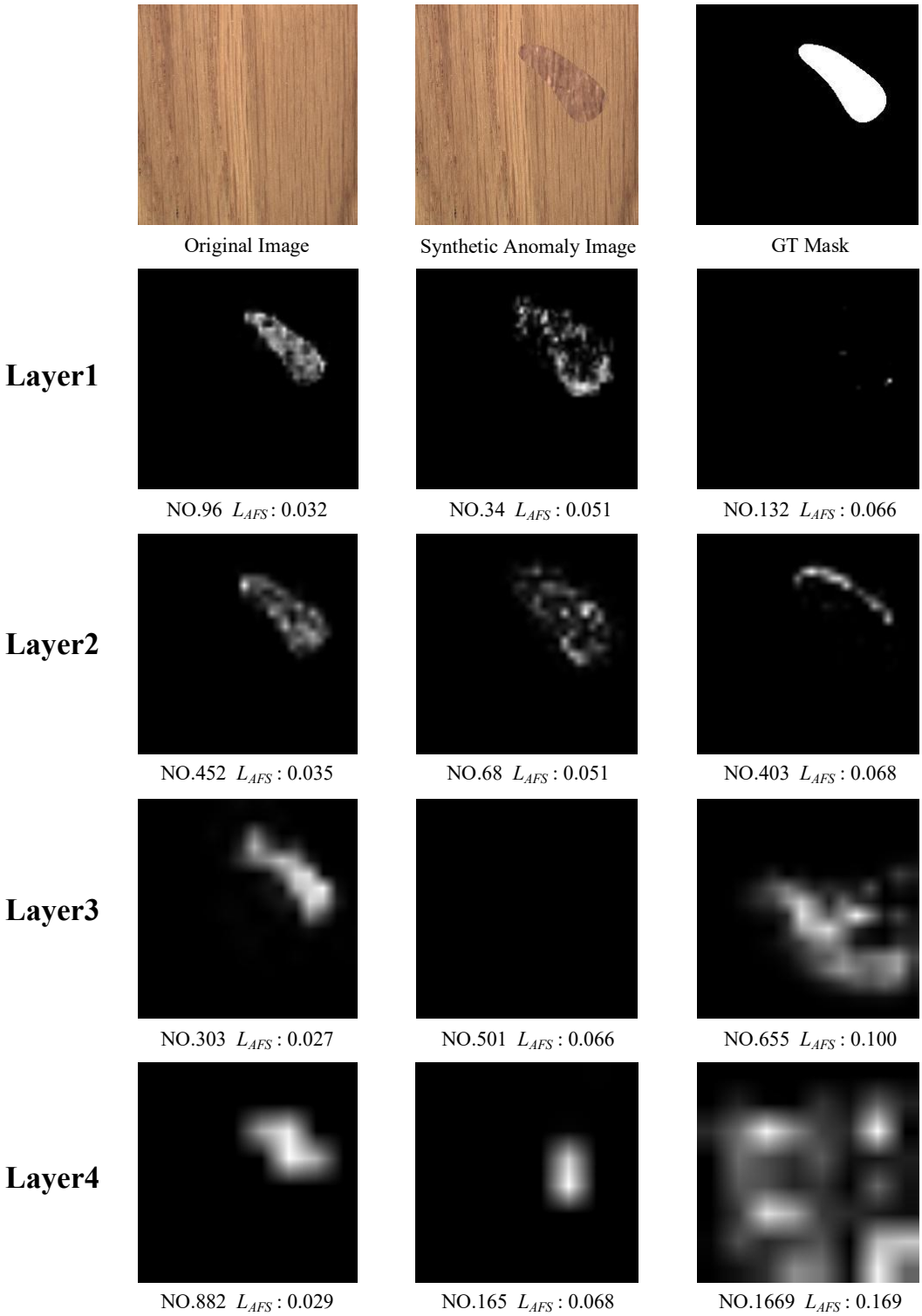

Figure S7. Visualization of AFS. For an original image and a synthetic anomaly image, we visualize the normalized difference between their corresponding feature maps across different layers of a pre-trained WideResNet50 [24]. From top to bottom, the feature map respectively come from the first layer to the fourth layer. Each feature map is labelled with its index in the layer and the corresponding AFS loss. From left to right, the localization performance of the feature maps gradually decreases. Our visualization intuitively demonstrates the localization bias caused by pre-training, indicating that not all feature maps contribute equally to anomaly detection and localization, as well as emphasizing the effectiveness of AFS.