

A. Appendix

A.1. Overview

This supplementary material consists of:

- Analysis on decoupled anomaly alignment loss and multiple tokens (Sec. A.2).
- More implementation details (Sec. A.3).
- More details of downstream supervised segmentation model implementation and usage (Sec. A.4).
- More ablation studies (Sec. A.5), including ablation studies on the Unbalanced Abnormal Text Prompt design, the Separation and Sharing Fine-tuning loss, the minimum size requirement for training images, the training strategy of SeaS, the cross-attention maps for Decoupled Anomaly Alignment, the features for Coarse Feature Extraction, the features of VAE for Refined Mask Prediction, the normal image supervision for Refined Mask Prediction, the Mask Refinement Module, and the threshold for mask binarization.
- More qualitative and quantitative results of anomaly image generation (Sec. A.6).
- Qualitative comparison results of supervised segmentation models trained on image-mask pairs generated by different anomaly generation methods (Sec. A.7).
- Qualitative comparison results of different supervised segmentation models trained on image-mask pairs generated by SeaS (Sec. A.8).
- Comparison with the Textual Inversion (Sec. A.9).
- More experiments on lighting conditions (Sec. A.10).
- More results on generation of small defects (Sec. A.11).
- More analysis on generation of unseen anomaly types (Sec. A.12).
- More experiments on comparison with DRAEM (Sec. A.13.)

A.2. Analysis on decoupled Anomaly alignment loss and multiple tokens

Here we give a more detailed analysis of the learning process of the DA loss. According to Eq. 3, intuitively, the DA loss may pull the anomaly tokens similar to each other. However, the U-Net in Stable Diffusion uses multi-head attention, which ensures that different anomaly tokens cover different attributes of the anomalies. In Eq. 3, the cross-attention map is the product of the feature map of U-Net and the anomaly tokens. In the implementation of multi-head attention, both the learnable embedding of the anomaly token and the U-Net feature are decomposed into several groups along the channel dimension. E.g., the conditioning vector $e_a \in \mathbb{R}^{1 \times C_1}$, which is corresponding to anomaly token, is divided into $\{e_{a,i} \in \mathbb{R}^{1 \times \frac{C_1}{q}} | i \in [1, q]\}$, and the image feature $v \in \mathbb{R}^{r \times r \times C_2}$ is divided into $\{v_i \in \mathbb{R}^{1 \times \frac{C_2}{q}} | i \in [1, q]\}$, where q is the number of heads in the multi-head attention. Then the corresponding groups are multiplied, and the out-

puts of all the heads are averaged. The attention map A of e_a is calculated by:

$$A = \frac{1}{q} \sum_{i=1}^q \text{softmax}\left(\frac{Q_i K_{a,i}^\top}{\sqrt{d}}\right), Q_i = \phi_q(v_i), K_{a,i} = \phi_k(e_{a,i}). \quad (8)$$

Therefore, in the defect region, the DA loss only ensures the average of each head tends to 1, but does not require the anomaly tokens to be the same as each other. In addition, each e_a is different from each other, and is combined by $e_{a,i}$. **The update direction of each $e_{a,i}$ is related to v_i and covers some features of the defect, it encompasses the attributes of anomalies from various perspectives, thereby providing diversified information.**

We provide more examples in Fig. 8, where new anomalies are generated that significantly differ from the training samples in terms of color and shape. For example, we showcase *bottle_contamination*, *hazelnut_print*, and *tile_gray_stroke* with a novel shape, *wood_color* and *metal_nut_scratch* with a novel color, and *pill_crack* with a new shape, featuring multiple cracks where the training samples only exhibit a single crack. These examples demonstrate the model’s ability to create unseen anomalies based on recombining the decoupled attributes.

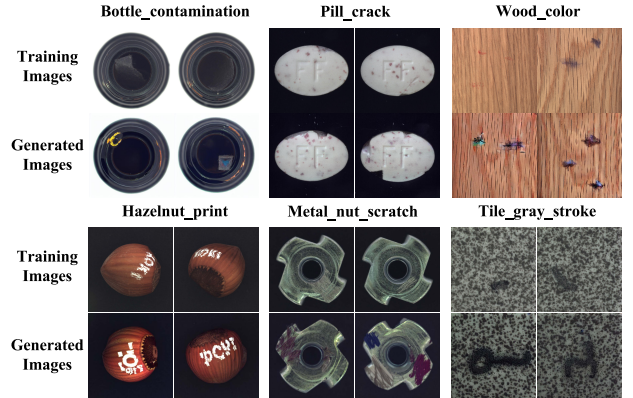


Figure 8. Visualization of the generation results for unseen anomalies on MVTec AD.

A.3. More implementation details

More training details. For the Unbalanced Abnormal Text Prompt, we set the number N of multiple $\langle \text{df}_n \rangle$ to 4 and the number N' of $\langle \text{ob} \rangle$ to 1, these parameters are fixed across all product classes. For a particular type of anomaly, we use the Unbalanced Abnormal (UA) Text Prompt \mathcal{P}_n with different sets of anomaly tokens as the condition to generate the specified type of anomaly.

$$\mathcal{P}_n = a \ \langle \text{ob} \rangle \ \text{with} \ \langle \text{df}_{4 \times n-3} \rangle, \ \langle \text{df}_{4 \times n-2} \rangle, \ \langle \text{df}_{4 \times n-1} \rangle, \ \langle \text{df}_{4 \times n} \rangle$$

where n represents the index of the anomaly types in the product. To generate normal images, we use the embedding \mathbf{e}_{ob} corresponding to the normal tokens of \mathcal{P} , i.e., “a <ob>”, to guide the U-Net in predicting noise. For example, for the normal token <ob>, given the lookup $\mathcal{U} \in \mathbb{R}^{b \times 768}$, where b is the number of text embeddings stored by the pre-trained text encoder, we use a placeholder string “ob1” as the input. Firstly, “ob1” is converted to a token ID $s_{\text{ob1}} \in \mathbb{R}^{1 \times 1}$ in the tokenizer. Secondly, $s_{\text{ob1}} \in \mathbb{R}^{1 \times 1}$ is converted to a one-hot vector $\mathcal{S}_{\text{ob1}} \in \mathbb{R}^{1 \times (b+1)}$. Thirdly, one learnable new embedding $g \in \mathbb{R}^{1 \times 768}$ corresponding to s_{ob1} is inserted to the lookup \mathcal{U} , resulting in $\mathcal{U}' \in \mathbb{R}^{(b+1) \times 768}$. Here, $g \in \mathbb{R}^{1 \times 768}$ is the learnable embedding of <ob>. **These embeddings and U-Net are learnable during the fine-tuning process.**

Training image generation model. For each product, we perform $800 \times G$ steps for fine-tuning, where G represents the number of anomaly categories of the product. The batch size of the training image generation model is set to 4. During each step of our fine-tuning process, we sample 2 images from the abnormal training set X_{df} , and 2 images from the normal training set X_{ob} . We utilize the AdamW [25] optimizer with a learning rate of U-Net is 4×10^{-6} . The learning rate of the text embedding is 4×10^{-5} .

Training Refined Mask Prediction branch. We design a cascaded Refined Mask Prediction (RMP) branch, which is grafted onto the U-Net trained according to SeaS. For each product, we perform $800 \times G$ steps for the RMP model, where G represents the number of anomaly types for the product. The batch size of training the RMP branch is set to 4. During each step of our fine-tuning process, we sample 2 images with their corresponding masks from the abnormal training set X_{df} , and 2 images from the normal training set X_{ob} . The masks used to suppress noise in normal images have each pixel value set to 0. The learning rate of the RMP model is 5×10^{-4} .

More inference details. For all experiments, we use $t = 1500$ to perform diffusion forward on normal images to get the initial noise. We employ $T = 25$ steps for sampling.

Metrics. For anomaly image generation, we report 4 metrics: the Inception Score (IS) and Intra-cluster pairwise LPIPS Distance (IC-LPIPS) to evaluate the anomaly images, KID [5] to assess the authenticity of normal images, and IC-LPIPS calculated only on anomaly regions (short for IC-LPIPS(a)), to evaluate diversity. The Inception Score (IS), proposed in [2], serves as an independent metric to evaluate the fidelity and diversity of generated images, by measuring the mutual information between input samples and their predicted classes. The IC-LPIPS [29] is used to evaluate the diversity of generated images, which quantifies the perceptual similarity between image patches in the same cluster. For pixel-level anomaly segmentation and image-level anomaly detection, we report 3 metrics: Area

Table 7. Comparison on resource requirement and time consumption.

| Methods | Training Overall Time | Inference Time (per image) |
|----------------------|--------------------------|-------------------------------|
| DFMGAN[11] | 414 hours | 48 ms |
| AnomalyDiffusion[17] | 249 hours | 3830 ms |
| SeaS | 73 hours | 720 ms |

Under Receiver Operator Characteristic curve (AUROC), Average Precision (AP), and F_1 -score at the optimal threshold (F_1 -max). **All of these metric are calculated using the scikit-learn library.** In addition, we calculate the Intersection over Union (IoU) to more accurately evaluate the anomaly segmentation result.

More training details on anomaly detection methods. In this section, we provide more training details of the comparative anomaly detection methods in Tab .2 and Tab .3 in the main text. For DRAEM [41], GLASS [9], and HVQ-Trans [26], we use the official checkpoints on the MVTec AD dataset, while the others are self-trained due to the lack of official checkpoints. For GLASS, the official foreground masks for the VisA and MVTec 3D AD datasets are not available, so this operation was not used. For PatchCore [32], we use the image size of 256 without center cropping, as some anomalies appear at the edges. For MambaAD [15], we use the provided official checkpoints.

Resource requirement and time consumption. We conduct our training on a NVIDIA Tesla A100 40G GPU sequentially for each product category, which may use about 20G memory. The comparison on time consumption is shown in Tab. 7. For the MVTec AD datasets, our training takes 73 hours, which is shorter than the 249 hours required by AnomalyDiffusion and the 414 hours required by DFMGAN. In terms of inference time, SeaS costs 720 ms per image, which is shorter than the 3830 ms per image required by the Diffusion-based method AnomalyDiffusion. The inference time of the GAN-based method DFMGAN is 48ms per image.

A.4. More details of the supervised segmentation models

As mentioned in the experiment part, we choose three supervised segmentation models (BiSeNet V2 [40], UPerNet [38], LFD [45]) to verify the validity of the generated image-mask pairs on the downstream supervised anomaly segmentation as well as detection tasks. **For BiSeNet V2 and UPerNet, we generally follow the implementation provided by MMsegmentation. For LFD, we also use the official implementation.**

Specifically, for BiSeNet V2, we choose a backbone

structure of a detail branch of three stages with 64, 64 and 128 channels and a semantic branch of four stages with 16, 32, 64 and 128 channels respectively, with a decode head and four auxiliary heads (corresponding to the number of stages in the semantic branch). As for UPerNet, we choose ResNet-50 as the backbone, with a decode head and an auxiliary head.

In training supervised segmentation models for downstream tasks, we adopt a training strategy of training a unified supervised segmentation model for all classes of products, rather than training separate supervised segmentation models for each class. Experimental results are shown in Tab. 8, which indicate that the performance of the unified supervised segmentation model surpasses that of multiple individual supervised segmentation models.

Table 8. Ablation on the training strategy of supervised segmentation models.

| Models | Multiple Models | | | | Unified Model | | | |
|------------|-----------------|-------|------------|-------|---------------|--------------|--------------|--------------|
| | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | IoU |
| BiSeNet V2 | 96.00 | 67.68 | 65.87 | 54.11 | 97.21 | 69.21 | 66.37 | 55.28 |
| UPerNet | 96.77 | 73.88 | 70.49 | 60.37 | 97.87 | 74.42 | 70.70 | 61.24 |
| LFD | 93.02 | 72.97 | 71.56 | 55.88 | 98.09 | 77.15 | 72.52 | 56.47 |
| Average | 95.26 | 71.51 | 69.31 | 56.79 | 97.72 | 73.59 | 69.86 | 57.66 |

A.5. More ablation studies

Ablation on the Unbalanced Abnormal Text Prompt design

In the design of the prompt for industrial anomaly image generation, we conduct experiments to validate the effectiveness of our Unbalanced Abnormal (UA) Text Prompt for each anomaly type of each product. We set the number of learnable $\langle df_n \rangle$ to N , and the number of learnable $\langle ob_j \rangle$ to N' . As shown in Tab. 9, by utilizing the UA Text Prompt, i.e.,

$$\mathcal{P} = a \langle ob \rangle \text{ with } \langle df_1 \rangle, \langle df_2 \rangle, \langle df_3 \rangle, \langle df_4 \rangle$$

we are able to provide high-fidelity and diverse images for downstream supervised anomaly segmentation tasks, resulting in the best performance in segmentation metrics.

Ablation on the Separation and Sharing Fine-tuning loss

In the design of the DA loss and NA loss for the Separation and Sharing Fine-tuning, we conduct two sets of experiments: (a) We remove the second term in the DA loss (short for w/o ST in Tab. 10); (b) We replace the second term in DA loss with another term in the NA loss (short for with AT in Tab. 10), which aligns the background area with the token $\langle ob \rangle$ according to the mask:

$$\mathcal{L}_{ob} = \sum_{l=1}^L (||A_{ob}^l - (1 - M^l)||^2) + ||\epsilon_{ob} - \epsilon_{\theta}(\hat{z}_{ob}, t_{ob}, \mathbf{e}_{ob})||_2^2 \quad (9)$$

where $A_{ob}^l \in \mathbb{R}^{r \times r \times 1}$ is the cross-attention map corresponding to the normal token $\langle ob \rangle$. As shown in Tab. 10, the experimental results demonstrate that, our adopted loss design achieves the best performance in downstream supervised segmentation tasks.

Table 9. Ablation on the Unbalanced Abnormal Text Prompt design.

| Prompt | AUROC | AP | F_1 -max | IoU |
|---------------------------------|--------------|--------------|--------------|--------------|
| $N' = 1, N = 1$ | 96.48 | 63.69 | 62.50 | 52.02 |
| $N' = 1, N = 4$ (Ours) | 97.21 | 69.21 | 66.37 | 55.28 |
| $N' = 4, N = 4$ | 96.55 | 66.28 | 63.95 | 54.07 |

Table 10. Ablation on the Separation and Sharing Fine-tuning loss.

| Loss | AUROC | AP | F_1 -max | IoU |
|-------------|--------------|--------------|--------------|--------------|
| w/o ST | 96.44 | 67.73 | 65.23 | 54.99 |
| with AT | 96.42 | 63.99 | 62.43 | 53.36 |
| Ours | 97.21 | 69.21 | 66.37 | 55.28 |

Ablation on the minimum size requirement for training images

In the few-shot setting, for a fair comparison, we follow the common setting in DFMGAN [11] and AnomalyDiffusion [17], i.e., using one-third abnormal image-mask pairs for each anomaly type in training. In this setting, the minimum number of abnormal training images is 2. Once we adopt a 3-shot setting, we need to reorganize the test set. To ensure that the test set is not reorganized for fair comparison, we take 1-shot and 2-shot settings for all anomaly types during training, i.e., $H = 1$ and $H = 2$, where H is the image number. The results are shown in Tab. 11 and Fig. 9. Observably, the models trained by 1-shot and 2-shot settings still generate anomaly images with decent diversity and authenticity.

Table 11. Ablation on the minimum size requirement for training images.

| Size | IS | IC-L |
|------------------------------|--------------|--------------|
| $H = 1$ | 1.790 | 0.311 |
| $H = 2$ | 1.794 | 0.314 |
| $H = \frac{1}{3} \times H_0$ | 1.876 | 0.339 |

Ablation on the training strategy of SeaS

During each step of the fine-tuning process, we sample the same number of images from the abnormal training set X_{df} and the normal training set X_{ob} . To investigate the efficacy of this strategy, we conduct three distinct sets of ex-

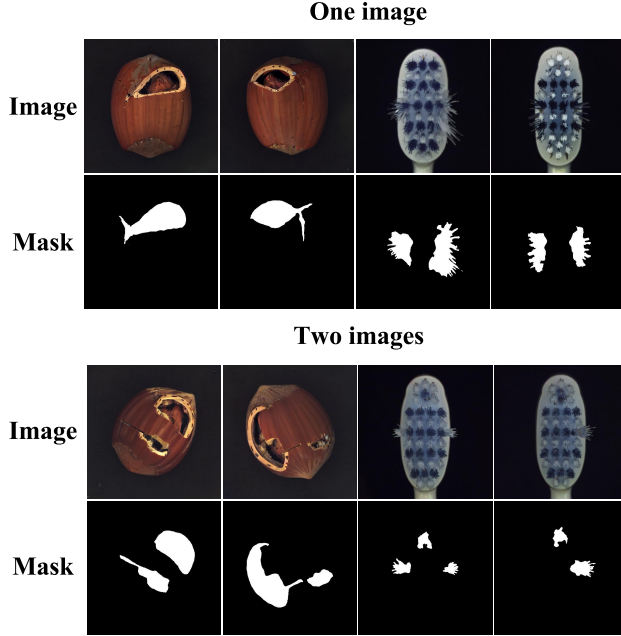


Figure 9. Visualization of the ablation study on the minimum size requirement for training images. In the figure, the first row is for generated images, and the second row is for generated masks.

periments: (a) prioritizing training with abnormal images followed by normal images (short for Abnormal-Normal in Tab. 12); (b) prioritizing training with abnormal images followed by anomaly images (short for Normal-Abnormal in Tab. 12); (c) training with a mix of both normal and abnormal images in each batch (short for Abnormal&Normal in Tab. 12). As shown in Tab. 12, SeaS yields superior performance in anomaly image generation, characterized by both high fidelity and diversity in the generated images.

Table 12. Ablation on training strategy of SeaS.

| Strategy | IS | IC-L |
|------------------------|-------------|-------------|
| Abnormal-Normal | 1.53 | 0.28 |
| Normal-Abnormal | 1.70 | 0.32 |
| Abnormal&Normal (Ours) | 1.88 | 0.34 |

Ablation on the cross-attention maps for Decoupled Anomaly Alignment

In Decoupled Anomaly Alignment (DA) loss, we leverage cross-attention maps from various layers of the U-Net encoder. Specifically, we investigate the impact of integrating different cross-attention maps, denoted as $A^1 \in \mathbb{R}^{64 \times 64}$, $A^2 \in \mathbb{R}^{32 \times 32}$, $A^3 \in \mathbb{R}^{16 \times 16}$ and $A^4 \in \mathbb{R}^{8 \times 8}$. These correspond to the cross-attention maps of the “down-1”, “down-2”, “down-3”, and “down-4” lay-

ers of the encoder in U-Net respectively. As shown in Tab. 13, the experimental results demonstrate that, employing a combination of $\{A^2, A^3\}$ for DA loss, achieves the best performance in downstream supervised segmentation tasks.

Table 13. Ablation on the cross-attention maps for Decoupled Anomaly Alignment.

| A^l | AUROC | AP | F_1 -max | IoU |
|-------------------|--------------|--------------|--------------|--------------|
| $l = 1, 2, 3$ | 96.42 | 68.92 | 66.24 | 54.52 |
| $l = 2, 3, 4$ | 95.71 | 64.51 | 62.33 | 52.46 |
| $l = 2, 3$ (Ours) | 97.21 | 69.21 | 66.37 | 55.28 |

Ablation on the features for Coarse Feature Extraction

In the coarse feature extraction process, we extract coarse but highly-discriminative features for anomalies from U-Net decoder. Specifically, we investigate the impact of integrating different features, denoted as $F_1 \in \mathbb{R}^{16 \times 16 \times 1280}$, $F_2 \in \mathbb{R}^{32 \times 32 \times 1280}$, $F_3 \in \mathbb{R}^{64 \times 64 \times 640}$ and $F_4 \in \mathbb{R}^{64 \times 64 \times 320}$. These correspond to the output feature “up-1”, “up-2”, “up-3”, and “up-4” layers of the encoder in U-Net respectively.

As shown in Fig. 10, we use the output features of the “up-2” and “up-3” layers of the decoder in U-Net, and apply convolution blocks and concatenation operations, then we can obtain the unified coarse feature $\hat{F} \in \mathbb{R}^{64 \times 64 \times 192}$, which can be used to predict masks corresponding to anomaly images. As shown in Tab. 14, the experimental results demonstrate that, employing a combina-

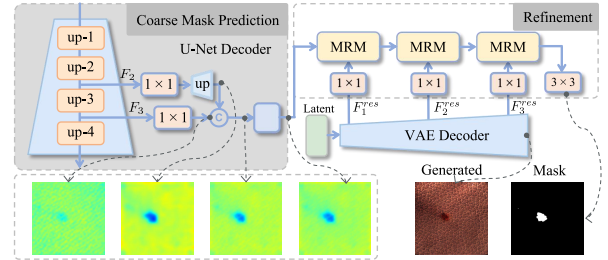


Figure 10. Visualization of the U-Net decoder features in mask prediction process.

Table 14. Ablation on the features for Coarse Feature Extraction.

| F_y | AUROC | AP | F_1 -max | IoU |
|-------------------|--------------|--------------|--------------|--------------|
| $y = 1, 2, 3$ | 94.35 | 63.58 | 60.54 | 52.36 |
| $y = 2, 3, 4$ | 96.93 | 67.42 | 64.26 | 55.31 |
| $y = 2, 3$ (Ours) | 97.21 | 69.21 | 66.37 | 55.28 |

Ablation on the features of VAE for Refined Mask Prediction

In the Refined Mask Prediction, we combine the high-resolution features of VAE decoder with discriminative features from U-Net, to generate accurately aligned anomaly image-mask pairs. In addition, we can also use the VAE encoder features as high-resolution features. As shown in Tab. 15, the experimental results show that, using VAE decoder features achieves better performance in downstream supervised segmentation tasks.

Table 15. Ablation on the features of VAE for Refined Mask Prediction.

| F^{res} | AUROC | AP | F_1 -max | IoU |
|--------------------|--------------|--------------|--------------|--------------|
| VAE encoder | 96.14 | 66.26 | 63.48 | 54.99 |
| VAE decoder | 97.21 | 69.21 | 66.37 | 55.28 |

Ablation on the normal image supervision for Refined Mask Prediction

In the Refined Mask Prediction branch, we predict masks for normal images as the supervision for the mask prediction. We conduct two sets of experiments: (a) We remove the second and the fourth term in the loss for RMP, i.e., the normal image supervision (short for NIA in Tab. 16); (b) We use the complete form in RMP branch loss, i.e., we use the normal image for supervision, as in Eq. (10):

$$\mathcal{L}_{\mathcal{M}} = \mathcal{F}(\hat{M}_{df}, \mathbf{M}_{df}) + \mathcal{F}(\hat{M}_{ob}, \mathbf{M}_{ob}) + \mathcal{F}(\hat{M}'_{df}, \mathbf{M}'_{df}) + \mathcal{F}(\hat{M}'_{ob}, \mathbf{M}'_{ob}) \quad (10)$$

As shown in Tab. 16, the experimental results show that, using normal images for supervision achieves better performance in downstream supervised segmentation tasks. We also provide further qualitative results of the effect of normal image supervision (short for NIA in Fig. 11) on MVTec AD.

Table 16. Ablation on the normal image supervision for Refined Mask Prediction.

| F^{res} | AUROC | AP | F_1 -max | IoU |
|------------------------|--------------|--------------|--------------|--------------|
| w/o NIA | 96.20 | 66.03 | 64.09 | 53.97 |
| with NIA (Ours) | 97.21 | 69.21 | 66.37 | 55.28 |

Ablation on the Mask Refinement Module

In the Refined Mask Prediction branch, the Mask Refinement Module (MRM) is utilized to generate refined masks. We devise different structures for MRM, as shown in Fig.

12, including Case a): those without conv blocks, Case b): with one conv block, and Case c): with chained conv blocks. As shown in Fig. 13, we find that using the conv blocks in Case b), which consists of two 1×1 convolutions and one 3×3 convolution, helps the model learn the features of the defect area more accurately, rather than focusing on the background area for using one convolution alone in Case a). Based on this observation, we further designed a chained conv blocks structure in Case c), and the acquired features better reflect the defect area. This one-level-by-one level of residual learning helps the model achieve better residual correction results for the defect area features. As shown in Tab. 17 in the Appendix, Case c) improves the performance by + 0.28% on AUROC, + 2.29% on AP and + 2.29% on F_1 -max, + 0.32% on IoU compared with Case b). We substantiate the superiority of the MRM structures that we design, through the results of downstream supervised segmentation experiments.

Table 17. Ablation on the Mask Refinement Module.

| Model | AUROC | AP | F_1 -max | IoU |
|---------------------|--------------|--------------|--------------|--------------|
| with MRM (a) | 96.75 | 68.18 | 64.96 | 55.51 |
| with MRM (b) | 96.93 | 66.92 | 64.08 | 54.96 |
| with MRM (c) | 97.21 | 69.21 | 66.37 | 55.28 |

Ablation on the threshold for mask binarization

In the Refined Mask Prediction branch, we take the threshold τ for the second channel of refined anomaly masks \hat{M}'_{df} to segment the final anomaly mask. We train supervised segmentation models using anomaly masks with τ settings ranging from 0.1 to 0.5. As shown in Tab. 18, results indicate that setting $\tau = 0.2$ yields the best model performance.

A.6. More qualitative and quantitative anomaly image generation results

More detailed quantitative results In this section, we report the detailed generation results for each category on the MVTec AD dataset, VisA dataset, and MVTec 3D AD dataset, which are presented in Tab. 19, Tab. 20, and Tab. 21.

More qualitative generation results

We provide further qualitative results of every category on the MVTec AD dataset, from Fig. 15 to Fig. 16. We report the anomaly image generation results of SeaS for varying types of anomalies. The first column represents the generated anomaly images, the second column represents the corresponding generated masks, and the third column represents the masks generated without using the Mask Refinement Module.

We provide further qualitative results of every category

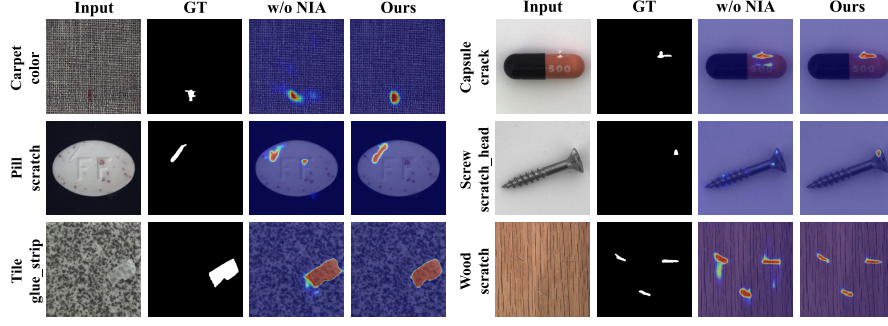


Figure 11. Qualitative results of the effect of normal image supervision on MVTEC AD.

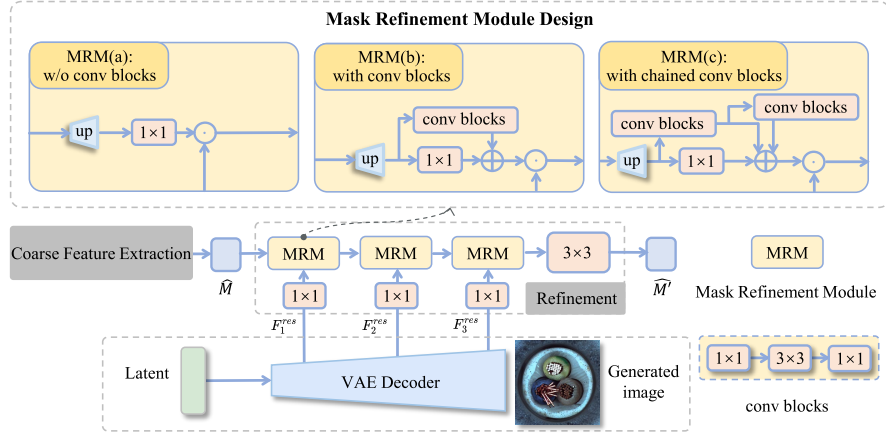


Figure 12. Different structure designs for the mask refinement module in the mask prediction branch.

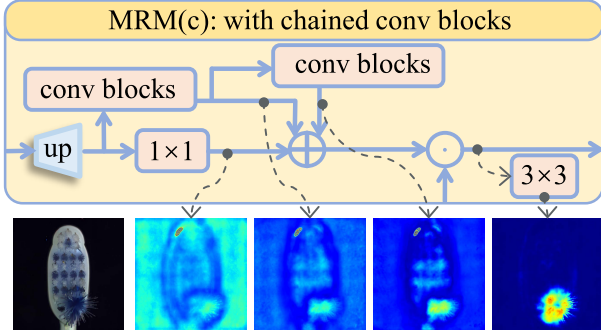


Figure 13. Visualization of the MRM module intermediate results. The top is for the MRM structure diagram, and the bottom is sequentially for the input image, feature maps of the MRM intermediate process and the predicted mask.

on the MVTEC 3D AD dataset in Fig. 17. We report the anomaly image generation results of SeaS for varying types of anomalies. The first column represents the generated anomaly images, and the second column represents the corresponding generated masks.

Table 18. Ablation on the threshold for mask binarization.

| threshold | AUROC | AP | F_1 -max | IoU |
|---------------------|--------------|--------------|--------------|--------------|
| $\tau = 0.1$ | 97.56 | 65.33 | 63.38 | 52.40 |
| $\tau = 0.2$ (Ours) | 97.21 | 69.21 | 66.37 | 55.28 |
| $\tau = 0.3$ | 97.20 | 66.92 | 64.35 | 54.68 |
| $\tau = 0.4$ | 95.31 | 63.55 | 61.97 | 53.03 |
| $\tau = 0.5$ | 94.11 | 60.85 | 59.92 | 50.87 |

A.7. More qualitative and quantitative comparison results of supervised segmentation models trained on image-mask pairs generated by different anomaly generation methods

We provide further qualitative results with different anomaly generation methods on the MVTEC AD dataset. We report the generation results of SeaS for varying types of anomalies in each category. Results are from Fig. 18 to Fig. 21.

We provide further qualitative comparisons on downstream supervised segmentation trained by the generated

Table 19. Comparison on IS and IC-LPIPS on MVTec AD. Bold indicates the best performance, while underlined denotes the second-best result.

| Category | Crop& Paste [23] | | SDGAN [28] | | Defect-GAN [42] | | DFMGAN [11] | | Anomaly Diffusion [17] | | Ours | |
|------------|------------------|--------|-------------|--------|-----------------|--------|-------------|-------------|------------------------|-------------|-------------|-------------|
| | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ |
| bottle | 1.43 | 0.04 | 1.57 | 0.06 | 1.39 | 0.07 | <u>1.62</u> | 0.12 | 1.58 | <u>0.19</u> | 1.78 | 0.21 |
| cable | 1.74 | 0.25 | 1.89 | 0.19 | 1.70 | 0.22 | 1.96 | 0.25 | 2.13 | <u>0.41</u> | <u>2.09</u> | 0.42 |
| capsule | 1.23 | 0.05 | 1.49 | 0.03 | 1.59 | 0.04 | 1.59 | 0.11 | 1.59 | <u>0.21</u> | <u>1.56</u> | 0.26 |
| carpet | 1.17 | 0.11 | 1.18 | 0.11 | 1.24 | 0.12 | <u>1.23</u> | 0.13 | 1.16 | <u>0.24</u> | 1.13 | 0.25 |
| grid | 2.00 | 0.12 | 1.95 | 0.10 | 2.01 | 0.12 | 1.97 | <u>0.13</u> | <u>2.04</u> | 0.44 | 2.43 | 0.44 |
| hazelnut | 1.74 | 0.21 | 1.85 | 0.16 | 1.87 | 0.19 | <u>1.93</u> | <u>0.24</u> | 2.13 | 0.31 | 1.87 | 0.31 |
| leather | 1.47 | 0.14 | 2.04 | 0.12 | 2.12 | 0.14 | <u>2.06</u> | 0.17 | 1.94 | 0.41 | 2.03 | <u>0.40</u> |
| metal_nut | 1.56 | 0.15 | 1.45 | 0.28 | 1.47 | 0.30 | 1.49 | 0.32 | 1.96 | 0.30 | <u>1.64</u> | <u>0.31</u> |
| pill | 1.49 | 0.11 | 1.61 | 0.07 | 1.61 | 0.10 | 1.63 | 0.16 | 1.61 | <u>0.26</u> | <u>1.62</u> | 0.33 |
| screw | 1.12 | 0.16 | 1.17 | 0.10 | 1.19 | 0.12 | 1.12 | 0.14 | <u>1.28</u> | <u>0.30</u> | 1.52 | 0.31 |
| tile | 1.83 | 0.20 | 2.53 | 0.21 | 2.35 | 0.22 | 2.39 | 0.22 | <u>2.54</u> | 0.55 | 2.60 | <u>0.50</u> |
| toothbrush | 1.30 | 0.08 | 1.78 | 0.03 | <u>1.85</u> | 0.03 | 1.82 | 0.18 | 1.68 | <u>0.21</u> | 1.96 | 0.25 |
| transistor | 1.39 | 0.15 | 1.76 | 0.13 | 1.47 | 0.13 | <u>1.64</u> | <u>0.25</u> | 1.57 | 0.34 | 1.51 | 0.34 |
| wood | 1.95 | 0.23 | 2.12 | 0.25 | 2.19 | 0.29 | 2.12 | 0.35 | <u>2.33</u> | <u>0.37</u> | 2.77 | 0.46 |
| zipper | 1.23 | 0.11 | 1.25 | 0.10 | 1.25 | 0.10 | 1.29 | <u>0.27</u> | <u>1.39</u> | 0.25 | 1.63 | 0.30 |
| Average | 1.51 | 0.14 | 1.71 | 0.13 | 1.69 | 0.15 | 1.72 | 0.20 | <u>1.80</u> | <u>0.32</u> | 1.88 | 0.34 |

Table 20. Comparison on IS and IC-LPIPS on VisA. Bold indicates the best performance.

| Category | DFMGAN [11] | | AnomalyDiffusion [17] | | Ours | |
|------------|-------------|-------------|-----------------------|-------------|-------------|-------------|
| | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ |
| candle | 1.19 | 0.23 | 1.28 | 0.17 | 1.20 | 0.12 |
| capsules | 1.25 | 0.22 | 1.39 | 0.50 | 1.58 | 0.60 |
| cashew | 1.25 | 0.24 | 1.27 | 0.26 | 1.21 | 0.28 |
| chewinggum | 1.33 | 0.24 | 1.15 | 0.19 | 1.29 | 0.27 |
| fryum | 1.28 | 0.20 | 1.20 | 0.14 | 1.14 | 0.21 |
| macaroni1 | 1.14 | 0.24 | 1.15 | 0.14 | 1.15 | 0.18 |
| macaroni2 | 1.47 | 0.38 | 1.56 | 0.38 | 1.57 | 0.39 |
| pcb1 | 1.12 | 0.16 | 1.18 | 0.35 | 1.18 | 0.26 |
| pcb2 | 1.12 | 0.26 | 1.26 | 0.21 | 1.25 | 0.27 |
| pcb3 | 1.19 | 0.18 | 1.21 | 0.24 | 1.22 | 0.21 |
| pcb4 | 1.21 | 0.28 | 1.14 | 0.25 | 1.15 | 0.22 |
| pipe_fryum | 1.43 | 0.32 | 1.29 | 0.17 | 1.31 | 0.16 |
| Average | 1.25 | 0.25 | 1.26 | 0.25 | 1.27 | 0.26 |

Table 21. Comparison on IS and IC-LPIPS on MVTec 3D AD. Bold indicates the best performance.

| Category | DFMGAN [11] | | AnomalyDiffusion [17] | | Ours | |
|-------------|-------------|-------------|-----------------------|--------|-------------|-------------|
| | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ | IS ↑ | IC-L ↑ |
| bagel | 1.07 | 0.26 | 1.02 | 0.22 | 1.28 | 0.29 |
| cable_gland | 1.59 | 0.25 | 1.79 | 0.19 | 2.21 | 0.19 |
| carrot | 1.94 | 0.29 | 1.66 | 0.17 | 2.07 | 0.22 |
| cookie | 1.80 | 0.31 | 1.77 | 0.29 | 2.07 | 0.38 |
| dowel | 1.96 | 0.37 | 1.60 | 0.20 | 1.95 | 0.26 |
| foam | 1.50 | 0.17 | 1.77 | 0.30 | 2.20 | 0.39 |
| peach | 2.11 | 0.34 | 1.91 | 0.23 | 2.40 | 0.28 |
| potato | 3.05 | 0.35 | 1.92 | 0.17 | 1.98 | 0.22 |
| rope | 1.46 | 0.29 | 1.28 | 0.25 | 1.53 | 0.41 |
| tire | 1.53 | 0.25 | 1.35 | 0.20 | 1.81 | 0.31 |
| Average | 1.80 | 0.29 | 1.61 | 0.22 | 1.95 | 0.30 |

images. The segmentation anomaly maps are shown in Fig. 22. There are fewer false positives (e.g., *potato_combined*) and fewer false negatives (e.g., *bagel_contamination*), when the BiSeNet V2 is trained on the image-mask pairs generated by our method.

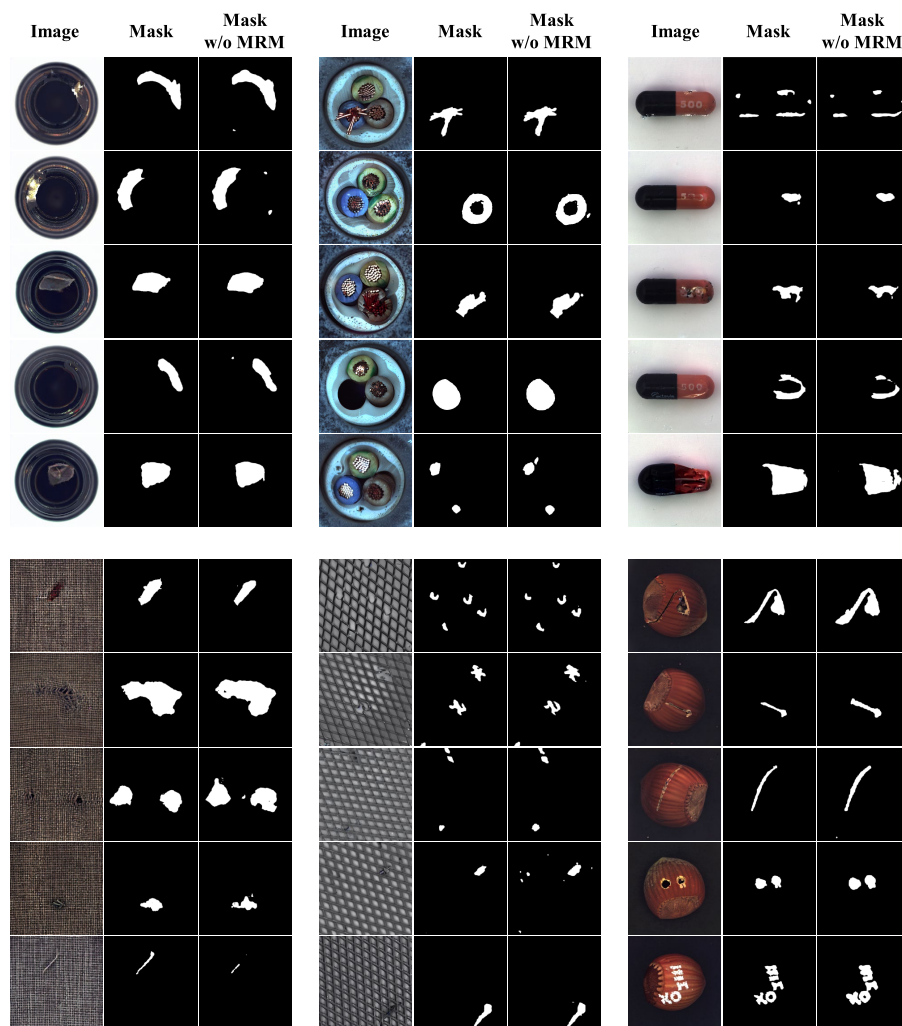
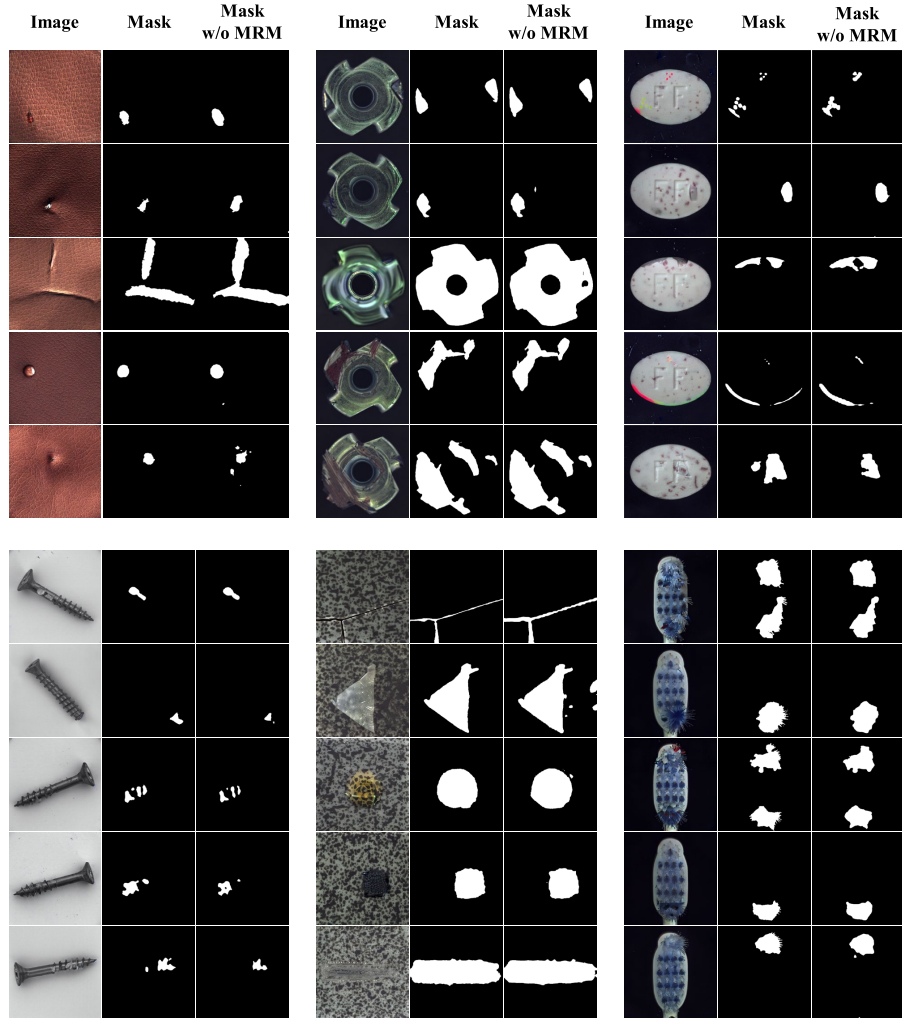


Figure 14. Qualitative results of our anomaly image generation results on MVTec AD. In the first row, from left to right, are the results for *bottle*, *cable*, and *capsule* categories. In the second row, from left to right, are the results for *carpet*, *grid*, and *hazelnut* categories.



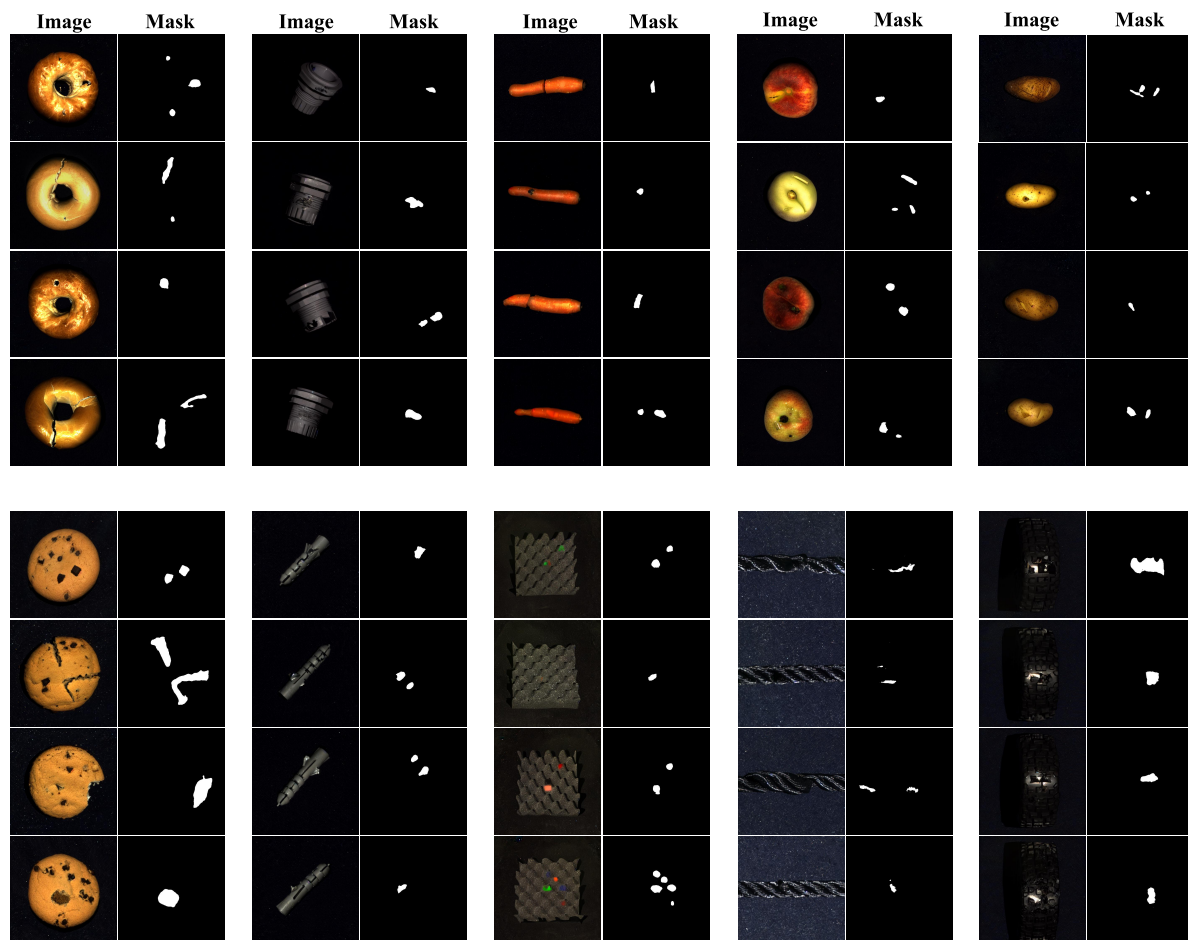


Figure 17. Qualitative results of our anomaly image generation results on MVTec 3D AD. In the first row, from left to right, are the results for *bagel*, *cable_gland*, *carrot*, *peach*, and *potato* categories. In the second row, from left to right, are the results for *cookie*, *dowel*, *foam*, *rope*, and *tire* categories.

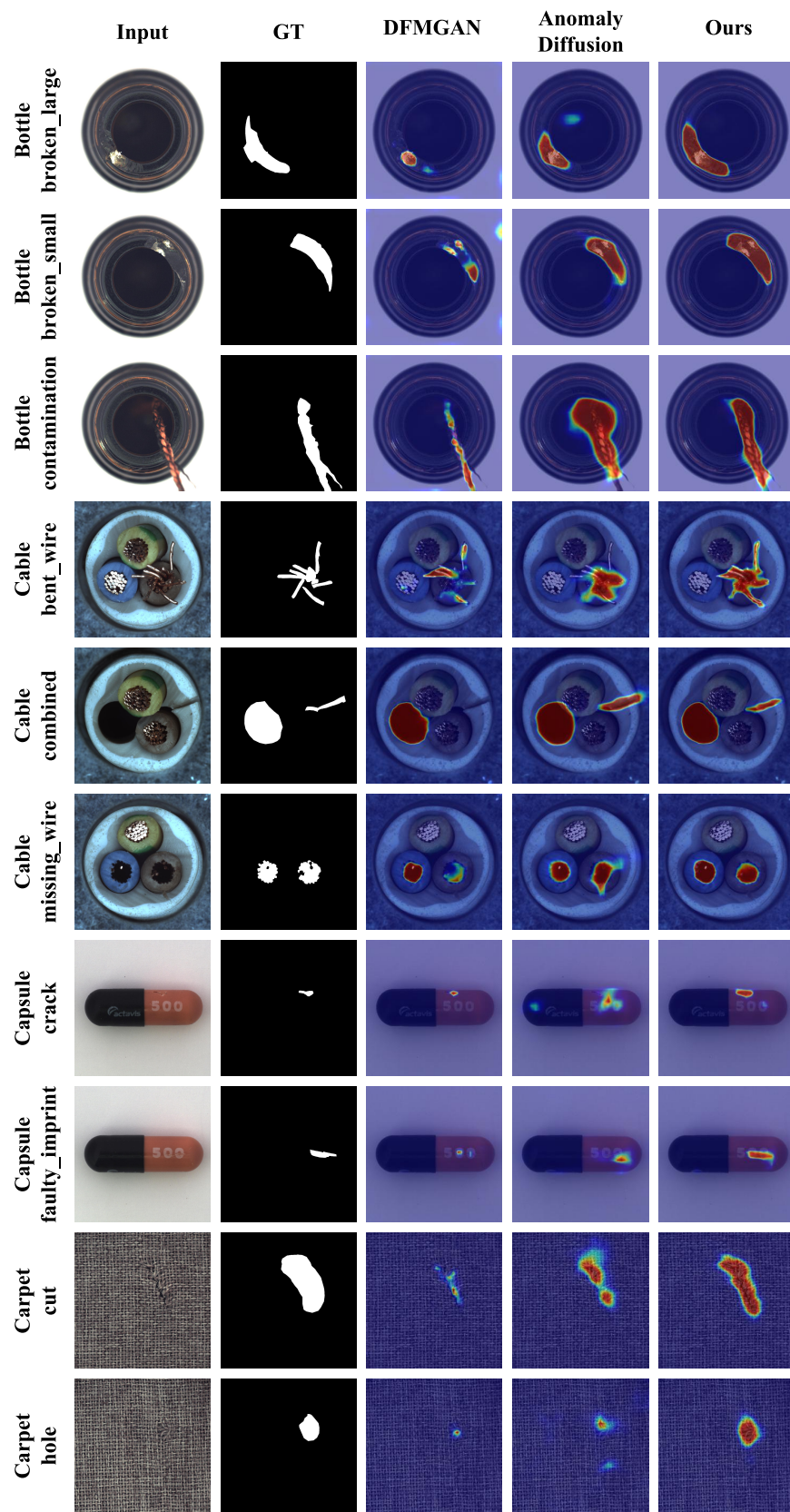


Figure 18. Comparison results with the anomaly supervised segmentation model BiSeNet V2 on MVTec AD. In the figure, from top to bottom are the results for *bottle*, *cable*, *capsule* and *carpet* categories.

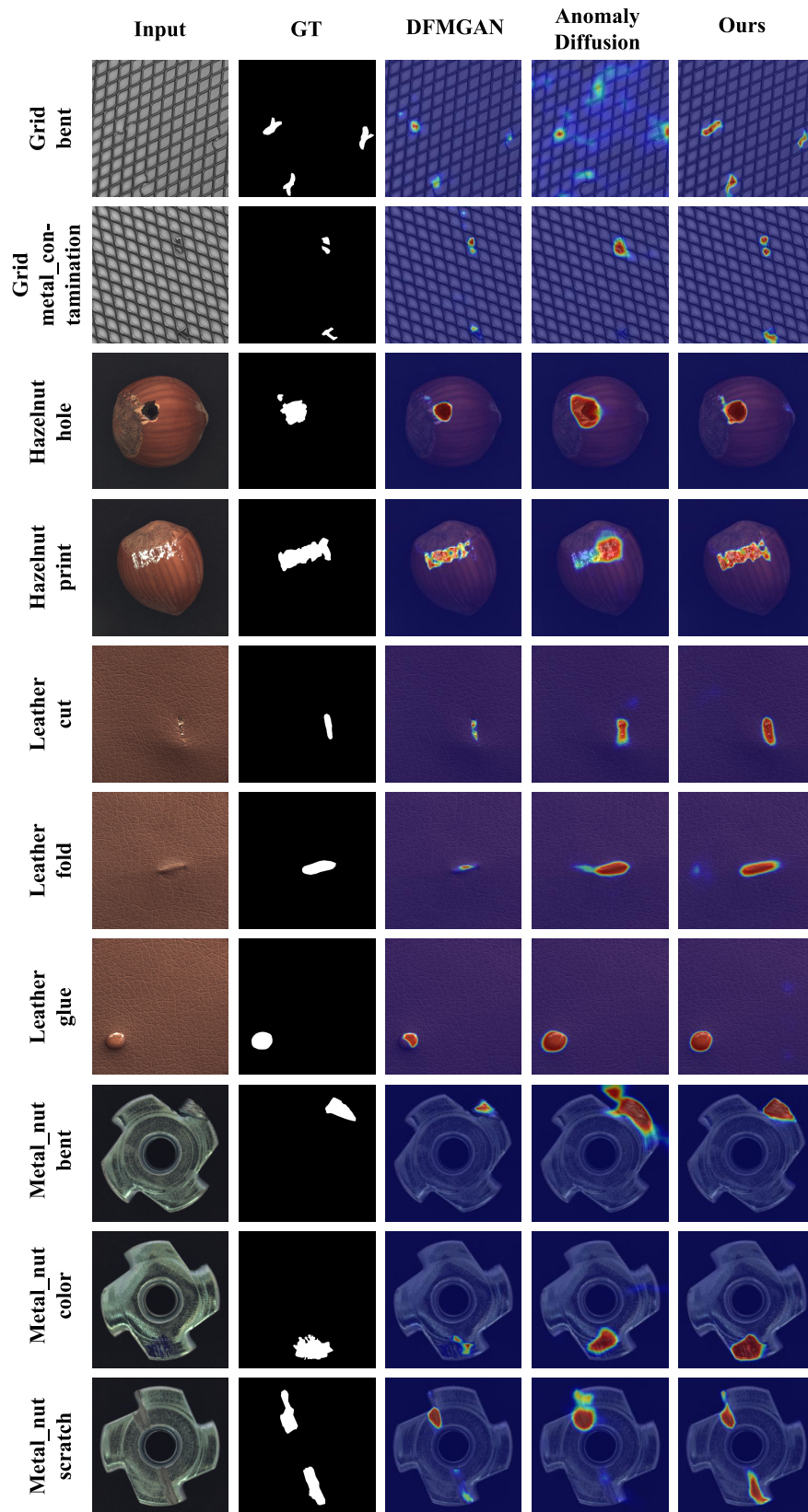


Figure 19. Comparison results with the anomaly supervised segmentation model BiSeNet V2 on MVTec AD. In the figure, from top to bottom are the results for *grid*, *hazelnut*, *leather* and *metal_nut* categories.

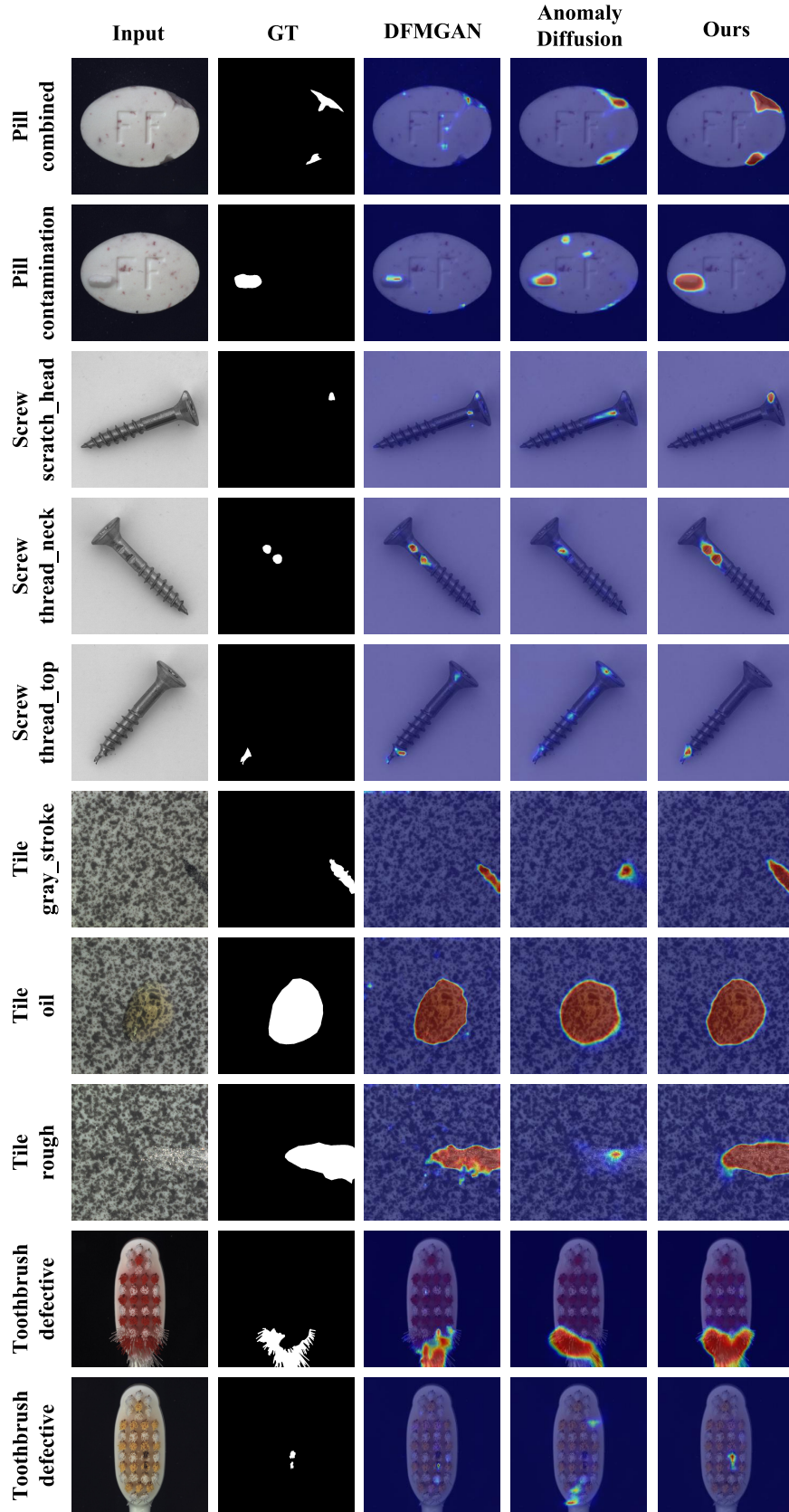


Figure 20. Comparison results with the anomaly supervised segmentation model BiSeNet V2 on MVTec AD. In the figure, from top to bottom are the results for *pill*, *screw*, *tile* and *toothbrush* categories.

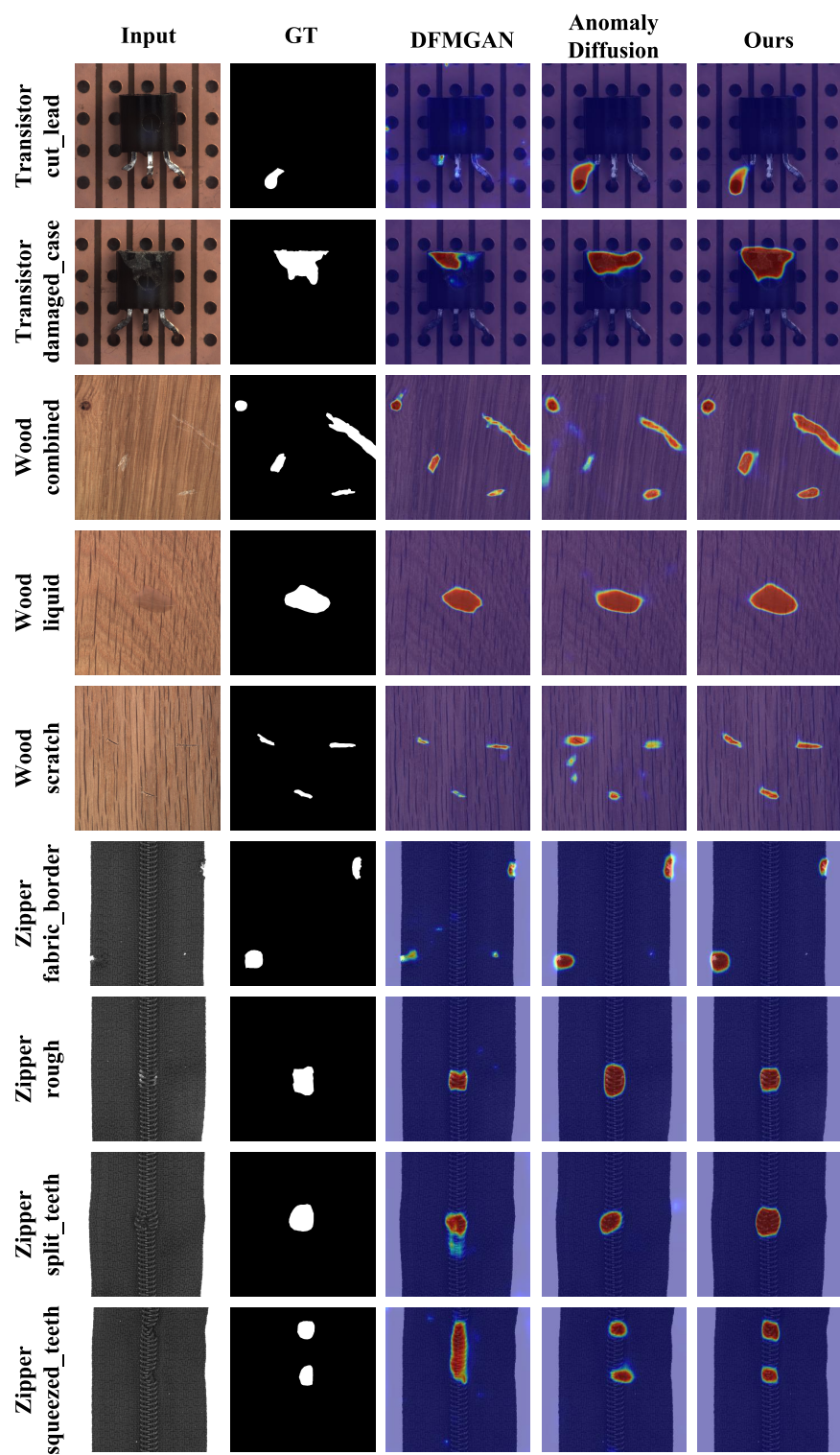


Figure 21. Comparison results with the anomaly supervised segmentation model BiSeNet V2 on MVTec AD. In the figure, from top to bottom are the results for *transistor*, *wood* and *zipper* categories.

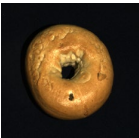

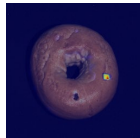
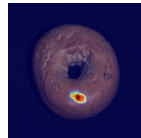
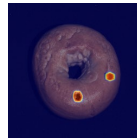






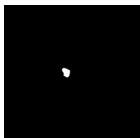

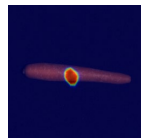




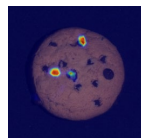
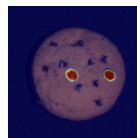

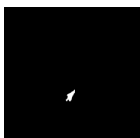
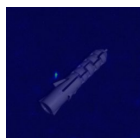
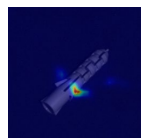
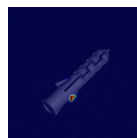

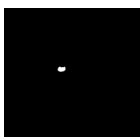
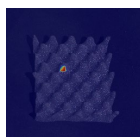
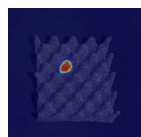
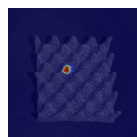
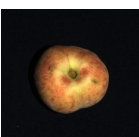

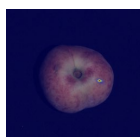
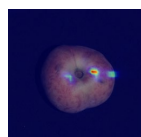
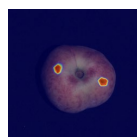
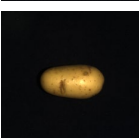
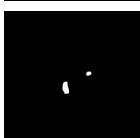

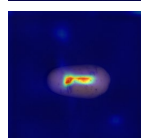
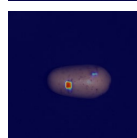
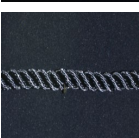
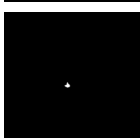
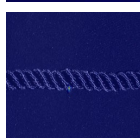
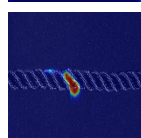
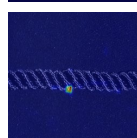
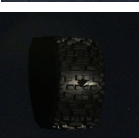
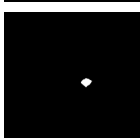
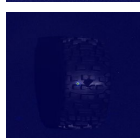
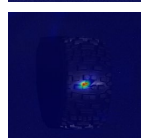

| | Input | GT | DFMGAN | Anomaly Diffusion | Ours |
|-------------------------|---|---|---|--|---|
| Bagel combined |  |  |  |  |  |
| Cable_gland Cut |  |  |  |  |  |
| Carrot contamination |  |  |  |  |  |
| Cookie combined |  |  |  |  |  |
| Dowel bent |  |  |  |  |  |
| Foam color |  |  |  |  |  |
| Peach combined |  |  |  |  |  |
| Potato contamination |  |  |  |  |  |
| Rope contamination |  |  |  |  |  |
| Tire contamination |  |  |  |  |  |

Figure 22. Qualitative supervised anomaly segmentation results with BiSeNet V2 on MVTEC 3D AD.

We report the detailed segmentation results of SeaS for each category on the MVTec AD datasets, compared with DFMGAN [11] and AnomalyDiffusion [17], which are presented from Tab. 22 to Tab. 27

A.8. More qualitative comparison results of different supervised segmentation models trained on image-mask pairs generated by SeaS

In this section, we provide further qualitative results with different supervised segmentation models on the MVTec AD and MVTec 3D AD datasets. We choose three models with different parameter quantity scopes (BiSeNet V2 [40]: 3.341M, UPerNet [38]: 64.042M, LFD [45]: 0.936M). We report the segmentation results of SeaS for varying types of anomalies in each category. Results are from Fig. 23 to Fig. 27.

Table 22. Comparison on supervised anomaly segmentation on BiSeNet V2.

| Category | DFMGAN | | | | AnomalyDiffusion | | | | Ours | | | |
|------------|--------------|--------------|--------------|--------------|------------------|--------------|--------------|-------|--------------|--------------|--------------|--------------|
| | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | IoU |
| bottle | 89.34 | 64.67 | 62.78 | 44.71 | 99.00 | 88.02 | 80.53 | 68.25 | 99.46 | 93.43 | 85.59 | 75.86 |
| cable | 93.87 | 67.98 | 64.74 | 44.02 | 92.84 | 69.86 | 66.32 | 46.49 | 89.85 | 72.07 | 71.58 | 53.24 |
| capsule | 74.88 | 16.43 | 23.01 | 29.97 | 92.71 | 38.11 | 40.67 | 19.44 | 86.33 | 24.64 | 30.54 | 39.70 |
| carpet | 94.53 | 42.53 | 47.44 | 39.88 | 98.65 | 73.10 | 65.83 | 43.25 | 99.61 | 82.30 | 72.94 | 55.52 |
| grid | 96.86 | 24.40 | 37.40 | 29.93 | 80.59 | 8.08 | 16.79 | 14.26 | 99.36 | 37.91 | 42.50 | 39.80 |
| hazelnut | 99.87 | 96.75 | 90.07 | 71.68 | 97.71 | 63.34 | 59.87 | 43.12 | 97.82 | 78.55 | 73.09 | 68.47 |
| leather | 97.50 | 51.10 | 52.26 | 50.67 | 99.30 | 57.49 | 59.62 | 43.94 | 98.91 | 59.84 | 58.62 | 45.82 |
| metal_nut | 99.39 | 97.59 | 92.52 | 70.40 | 99.03 | 95.67 | 88.69 | 58.8 | 99.69 | 98.29 | 93.23 | 74.40 |
| pill | 97.09 | 83.98 | 79.26 | 36.39 | 99.44 | 93.16 | 86.62 | 41.18 | 98.31 | 76.97 | 68.00 | 55.43 |
| screw | 97.94 | 37.10 | 41.01 | 31.63 | 94.08 | 17.95 | 25.90 | 20.00 | 97.64 | 40.20 | 45.35 | 38.43 |
| tile | 99.65 | 97.08 | 91.16 | 75.94 | 97.79 | 85.58 | 78.28 | 60.46 | 99.67 | 97.29 | 91.48 | 75.75 |
| toothbrush | 97.70 | 51.32 | 54.05 | 23.38 | 98.43 | 49.64 | 54.08 | 26.53 | 97.15 | 46.09 | 49.02 | 28.56 |
| transistor | 84.31 | 45.34 | 46.07 | 30.00 | 98.85 | 85.27 | 77.95 | 49.83 | 96.75 | 69.52 | 66.11 | 57.24 |
| wood | 98.32 | 64.82 | 63.11 | 58.99 | 96.78 | 63.38 | 60.31 | 45.73 | 98.38 | 80.81 | 74.03 | 56.22 |
| zipper | 97.29 | 65.18 | 63.24 | 49.93 | 98.81 | 78.89 | 72.66 | 62.03 | 99.23 | 80.27 | 73.41 | 64.80 |
| Average | 94.57 | 60.42 | 60.54 | 45.83 | 96.27 | 64.5 | 62.27 | 42.89 | 97.21 | 69.21 | 66.37 | 55.28 |

Table 23. Comparison on image-level anomaly detection on BiSeNet V2.

| Category | DFMGAN | | | AnomalyDiffusion | | | Ours | | |
|------------|---------------|---------------|---------------|------------------|---------------|---------------|---------------|---------------|---------------|
| | AUROC | AP | F_1 -max | AUROC | AP | F_1 -max | AUROC | AP | F_1 -max |
| bottle | 96.74 | 98.75 | 95.35 | 98.14 | 99.34 | 97.67 | 100.00 | 100.00 | 100.00 |
| cable | 79.47 | 85.00 | 74.13 | 95.37 | 96.71 | 92.91 | 94.61 | 96.39 | 89.83 |
| capsule | 85.51 | 95.16 | 89.82 | 84.06 | 95.01 | 89.74 | 88.81 | 96.92 | 89.21 |
| carpet | 91.42 | 96.29 | 88.89 | 90.55 | 96.41 | 90.32 | 98.16 | 99.31 | 97.56 |
| grid | 99.64 | 99.82 | 97.56 | 81.19 | 89.92 | 83.95 | 99.17 | 99.63 | 98.73 |
| hazelnut | 100.00 | 100.00 | 100.00 | 93.39 | 95.74 | 90.91 | 100.00 | 100.00 | 100.00 |
| leather | 98.31 | 99.23 | 95.24 | 100.00 | 100.00 | 100.00 | 95.83 | 98.38 | 95.93 |
| metal_nut | 97.37 | 99.16 | 94.66 | 99.01 | 99.66 | 97.71 | 100.00 | 100.00 | 100.00 |
| pill | 84.86 | 95.27 | 91.00 | 90.38 | 97.43 | 91.35 | 96.59 | 99.12 | 95.24 |
| screw | 74.95 | 85.50 | 80.72 | 58.18 | 75.32 | 81.25 | 77.24 | 89.55 | 80.60 |
| tile | 99.47 | 99.74 | 99.12 | 98.78 | 99.44 | 97.39 | 100.00 | 100.00 | 100.00 |
| toothbrush | 78.33 | 87.73 | 83.72 | 78.33 | 89.26 | 79.17 | 90.42 | 94.49 | 89.47 |
| transistor | 79.52 | 75.77 | 69.57 | 94.40 | 94.68 | 94.34 | 99.23 | 98.39 | 94.92 |
| wood | 98.87 | 99.46 | 97.67 | 90.48 | 94.12 | 93.33 | 100.00 | 100.00 | 100.00 |
| zipper | 98.97 | 99.64 | 97.56 | 98.89 | 99.62 | 97.56 | 100.00 | 100.00 | 100.00 |
| Average | 90.90 | 94.43 | 90.33 | 90.08 | 94.84 | 91.84 | 96.00 | 98.14 | 95.43 |

Table 24. Comparison on supervised anomaly segmentation on UPerNet.

| Category | DFMGAN | | | | AnomalyDiffusion | | | | Ours | | | |
|------------|--------------|--------------|--------------|-------|------------------|--------------|--------------|-------|--------------|--------------|--------------|--------------|
| | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | IoU |
| bottle | 87.94 | 56.89 | 56.56 | 45.41 | 99.54 | 93.01 | 85.94 | 75.31 | 99.28 | 91.73 | 84.53 | 78.73 |
| cable | 87.52 | 64.30 | 65.61 | 41.02 | 91.00 | 68.12 | 67.49 | 51.84 | 91.08 | 76.25 | 74.63 | 59.00 |
| capsule | 67.92 | 12.31 | 20.32 | 30.47 | 97.64 | 51.90 | 51.66 | 37.00 | 92.09 | 39.60 | 43.89 | 50.18 |
| carpet | 95.85 | 36.05 | 34.52 | 48.10 | 99.45 | 82.13 | 72.55 | 53.17 | 99.67 | 82.01 | 73.53 | 60.60 |
| grid | 97.49 | 29.67 | 36.15 | 31.37 | 94.22 | 28.97 | 38.50 | 32.93 | 99.18 | 44.94 | 48.28 | 44.21 |
| hazelnut | 99.36 | 79.76 | 71.10 | 72.90 | 97.77 | 70.48 | 67.93 | 54.47 | 99.54 | 81.84 | 75.48 | 73.30 |
| leather | 80.97 | 17.60 | 26.21 | 30.17 | 99.48 | 63.46 | 60.54 | 48.70 | 99.42 | 68.26 | 65.52 | 57.01 |
| metal_nut | 98.44 | 95.64 | 91.48 | 64.92 | 98.62 | 95.11 | 88.62 | 61.31 | 99.70 | 98.33 | 92.90 | 76.07 |
| pill | 97.58 | 83.74 | 80.02 | 42.33 | 99.33 | 95.04 | 88.77 | 49.18 | 98.59 | 81.16 | 74.26 | 62.62 |
| screw | 97.49 | 53.83 | 53.02 | 42.05 | 93.89 | 36.60 | 42.68 | 34.08 | 98.97 | 52.02 | 51.65 | 46.61 |
| tile | 99.79 | 97.29 | 91.11 | 77.46 | 94.70 | 73.34 | 67.79 | 58.54 | 99.67 | 95.89 | 90.71 | 77.89 |
| toothbrush | 97.42 | 51.09 | 59.23 | 28.33 | 97.52 | 60.67 | 59.46 | 33.98 | 98.50 | 63.62 | 63.07 | 42.09 |
| transistor | 82.07 | 36.31 | 39.48 | 27.44 | 94.26 | 73.68 | 69.50 | 53.64 | 93.88 | 70.37 | 68.12 | 56.98 |
| wood | 97.90 | 69.02 | 62.21 | 63.10 | 96.09 | 70.10 | 64.38 | 51.44 | 99.28 | 85.28 | 76.28 | 65.09 |
| zipper | 97.28 | 71.60 | 66.64 | 54.54 | 99.54 | 86.18 | 78.50 | 66.47 | 99.17 | 85.01 | 77.57 | 68.21 |
| Average | 92.33 | 57.01 | 56.91 | 46.64 | 96.87 | 69.92 | 66.95 | 50.80 | 97.87 | 74.42 | 70.70 | 61.24 |

Table 25. Comparison on image-level anomaly detection on UPerNet.

| Category | DFMGAN | | | AnomalyDiffusion | | | Ours | | |
|------------|--------|-------|------------|------------------|---------------|---------------|---------------|---------------|---------------|
| | AUROC | AP | F_1 -max | AUROC | AP | F_1 -max | AUROC | AP | F_1 -max |
| bottle | 94.19 | 97.86 | 93.18 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| cable | 85.64 | 90.03 | 80.33 | 95.58 | 97.06 | 92.56 | 94.40 | 96.38 | 92.44 |
| capsule | 81.04 | 94.26 | 87.01 | 96.00 | 98.77 | 95.48 | 94.43 | 98.44 | 92.21 |
| carpet | 96.72 | 98.58 | 93.75 | 98.68 | 99.53 | 98.36 | 99.94 | 99.97 | 99.20 |
| grid | 98.33 | 99.13 | 96.30 | 96.67 | 98.73 | 97.44 | 99.76 | 99.88 | 98.73 |
| hazelnut | 99.84 | 99.87 | 97.96 | 99.17 | 99.43 | 97.87 | 100.00 | 100.00 | 100.00 |
| leather | 79.91 | 90.70 | 81.75 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| metal_nut | 98.30 | 99.38 | 97.71 | 98.65 | 99.62 | 98.41 | 99.72 | 99.91 | 99.21 |
| pill | 88.54 | 96.56 | 92.39 | 91.23 | 97.78 | 90.91 | 98.28 | 99.58 | 97.92 |
| screw | 89.01 | 94.54 | 88.24 | 85.06 | 93.87 | 85.33 | 93.47 | 97.07 | 90.45 |
| tile | 99.68 | 99.81 | 99.13 | 99.68 | 99.81 | 99.13 | 100.00 | 100.00 | 100.00 |
| toothbrush | 75.00 | 86.99 | 80.00 | 90.00 | 95.13 | 90.00 | 95.00 | 97.65 | 94.74 |
| transistor | 83.04 | 73.59 | 74.19 | 100.00 | 100.00 | 100.00 | 99.52 | 99.16 | 96.43 |
| wood | 93.36 | 95.60 | 95.45 | 98.62 | 99.49 | 97.62 | 99.87 | 99.94 | 98.82 |
| zipper | 98.48 | 99.51 | 98.14 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Average | 90.74 | 94.43 | 90.37 | 96.62 | 98.61 | 96.21 | 98.29 | 99.20 | 97.34 |

Table 26. Comparison on supervised anomaly segmentation on LFD.

| Category | DFMGAN | | | | AnomalyDiffusion | | | | Ours | | | |
|------------|--------------|--------------|--------------|--------------|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | IoU |
| bottle | 90.41 | 61.51 | 58.49 | 40.19 | 98.71 | 89.64 | 81.55 | 67.10 | 99.28 | 92.65 | 84.86 | 73.82 |
| cable | 96.49 | 79.40 | 75.25 | 53.47 | 97.89 | 79.85 | 72.75 | 53.69 | 94.53 | 75.41 | 72.70 | 55.98 |
| capsule | 91.82 | 56.11 | 58.56 | 32.50 | 95.80 | 38.17 | 48.92 | 32.04 | 91.80 | 49.76 | 53.69 | 41.14 |
| carpet | 89.10 | 48.04 | 49.89 | 39.46 | 94.83 | 53.15 | 51.79 | 42.21 | 99.10 | 82.74 | 74.51 | 57.56 |
| grid | 89.18 | 34.89 | 41.21 | 19.21 | 85.19 | 24.32 | 34.76 | 18.22 | 98.78 | 62.24 | 58.44 | 41.69 |
| hazelnut | 99.36 | 95.16 | 89.80 | 76.43 | 98.54 | 77.39 | 70.42 | 45.97 | 98.97 | 88.00 | 81.77 | 73.39 |
| leather | 97.82 | 51.86 | 52.25 | 48.09 | 98.99 | 65.73 | 62.85 | 42.65 | 99.11 | 76.49 | 69.30 | 56.51 |
| metal_nut | 98.16 | 95.16 | 90.99 | 63.02 | 99.38 | 97.34 | 91.63 | 64.59 | 99.23 | 96.66 | 91.42 | 75.15 |
| pill | 95.80 | 75.90 | 70.31 | 31.73 | 98.96 | 92.51 | 85.35 | 50.04 | 98.11 | 79.63 | 72.54 | 56.73 |
| screw | 93.96 | 38.00 | 41.69 | 30.88 | 92.68 | 44.64 | 49.17 | 34.08 | 98.27 | 52.40 | 52.32 | 41.02 |
| tile | 97.37 | 88.79 | 82.05 | 66.30 | 92.98 | 79.59 | 73.52 | 55.08 | 99.38 | 96.24 | 89.90 | 75.50 |
| toothbrush | 95.17 | 55.21 | 53.95 | 28.83 | 98.31 | 68.60 | 66.14 | 29.67 | 96.97 | 54.84 | 53.19 | 27.91 |
| transistor | 97.68 | 89.68 | 84.18 | 46.98 | 98.20 | 83.97 | 75.84 | 44.22 | 98.80 | 84.32 | 77.02 | 55.57 |
| wood | 97.47 | 77.72 | 70.91 | 58.77 | 95.68 | 67.54 | 63.06 | 42.78 | 98.60 | 88.57 | 81.46 | 62.94 |
| zipper | 93.80 | 58.43 | 56.82 | 46.44 | 98.42 | 84.05 | 77.08 | 64.14 | 99.15 | 86.67 | 79.09 | 69.37 |
| Average | 94.91 | 67.06 | 65.09 | 45.49 | 96.30 | 69.77 | 66.99 | 45.77 | 98.01 | 77.77 | 72.81 | 57.62 |

Table 27. Comparison on image-level anomaly detection on LFD.

| Category | DFMGAN | | | AnomalyDiffusion | | | Ours | | |
|------------|--------|-------|------------|------------------|---------------|---------------|---------------|---------------|---------------|
| | AUROC | AP | F_1 -max | AUROC | AP | F_1 -max | AUROC | AP | F_1 -max |
| bottle | 96.98 | 98.76 | 95.35 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| cable | 90.98 | 94.21 | 88.14 | 99.52 | 99.55 | 97.71 | 92.05 | 94.95 | 88.70 |
| capsule | 86.32 | 95.99 | 88.46 | 83.25 | 94.62 | 89.44 | 93.80 | 98.19 | 93.42 |
| carpet | 88.02 | 95.33 | 87.60 | 86.00 | 93.42 | 87.22 | 97.98 | 99.22 | 96.67 |
| grid | 85.48 | 92.61 | 85.71 | 93.69 | 97.08 | 91.14 | 96.79 | 98.76 | 96.10 |
| hazelnut | 99.90 | 99.91 | 98.97 | 98.28 | 98.60 | 95.83 | 100.00 | 100.00 | 100.00 |
| leather | 95.93 | 98.15 | 93.65 | 99.90 | 99.95 | 99.20 | 100.00 | 100.00 | 100.00 |
| metal_nut | 96.16 | 98.57 | 96.18 | 99.01 | 99.65 | 98.46 | 98.58 | 99.54 | 97.64 |
| pill | 82.85 | 94.40 | 92.00 | 94.15 | 98.42 | 94.47 | 98.16 | 99.50 | 96.84 |
| screw | 82.60 | 92.15 | 82.22 | 81.54 | 91.32 | 82.05 | 87.83 | 94.39 | 85.54 |
| tile | 98.94 | 99.43 | 96.55 | 98.25 | 99.13 | 95.65 | 99.36 | 99.69 | 99.12 |
| toothbrush | 77.08 | 87.68 | 80.95 | 100.00 | 100.00 | 100.00 | 87.92 | 94.08 | 87.80 |
| transistor | 88.04 | 85.06 | 77.78 | 97.38 | 96.57 | 92.86 | 98.10 | 96.90 | 94.55 |
| wood | 99.87 | 99.94 | 98.82 | 97.24 | 98.70 | 96.47 | 100.00 | 100.00 | 100.00 |
| zipper | 97.07 | 98.78 | 96.25 | 99.01 | 99.71 | 99.39 | 100.00 | 100.00 | 100.00 |
| Average | 91.08 | 95.40 | 90.58 | 95.15 | 97.78 | 94.66 | 96.70 | 98.35 | 95.76 |

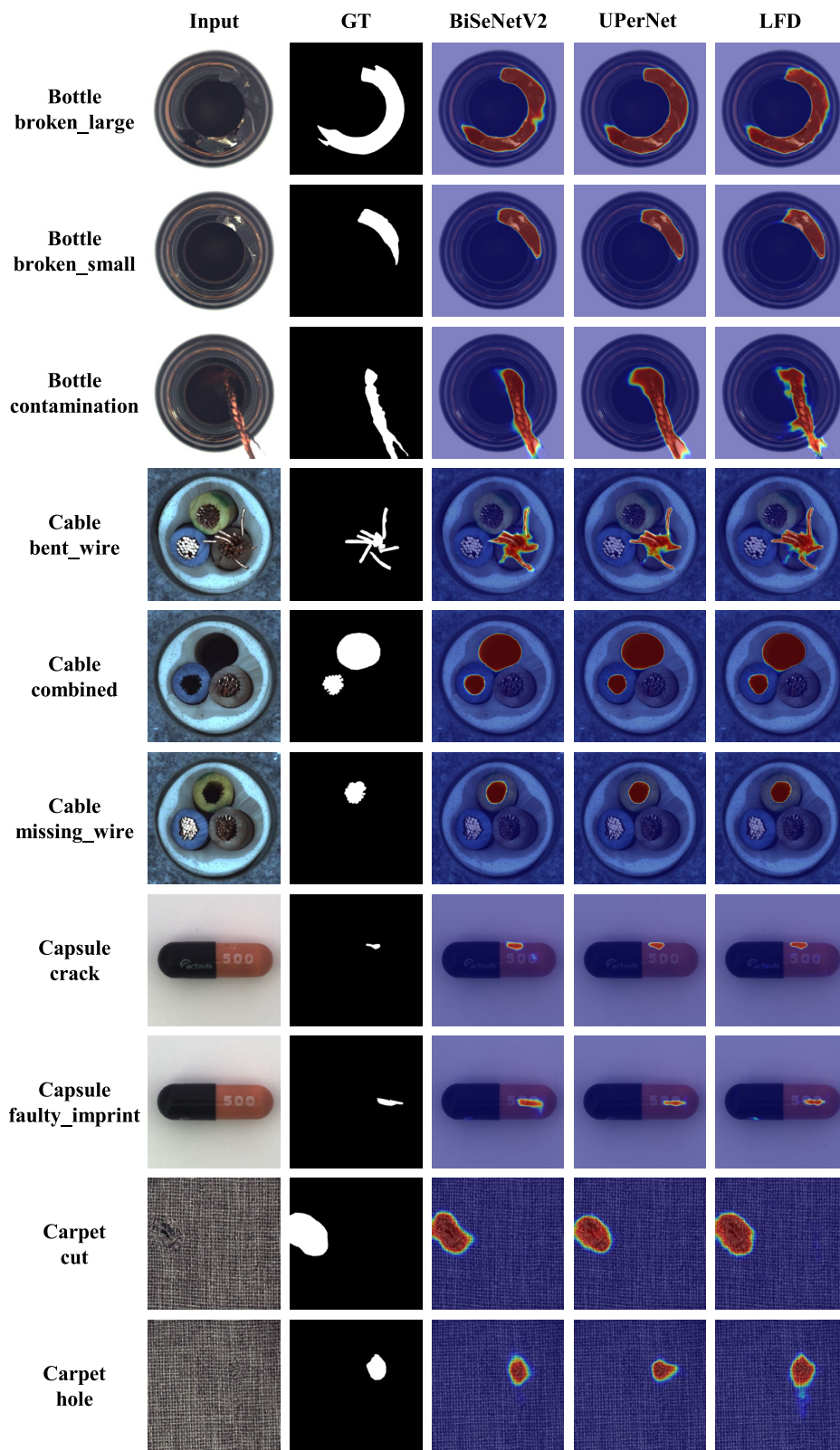


Figure 23. Qualitative comparison results with the supervised segmentation models on MVTec AD. In the figure, from top to bottom are the results for *bottle*, *cable*, *capsule* and *carpet* categories.

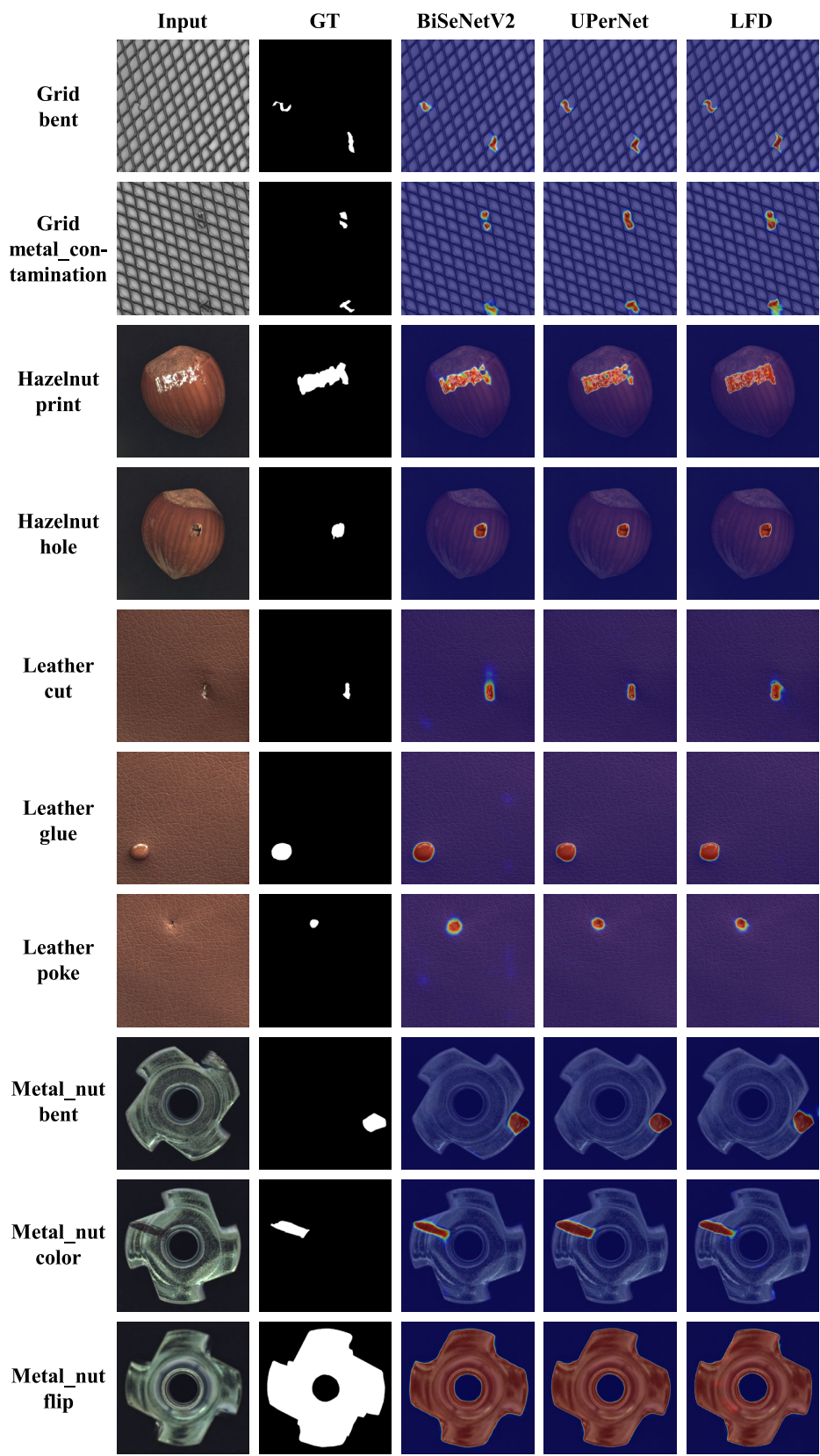


Figure 24. Qualitative comparison results with the supervised segmentation models on MVTec AD. In the figure, from top to bottom are the results for *grid*, *hazelnut*, *leather* and *metal_nut* categories.

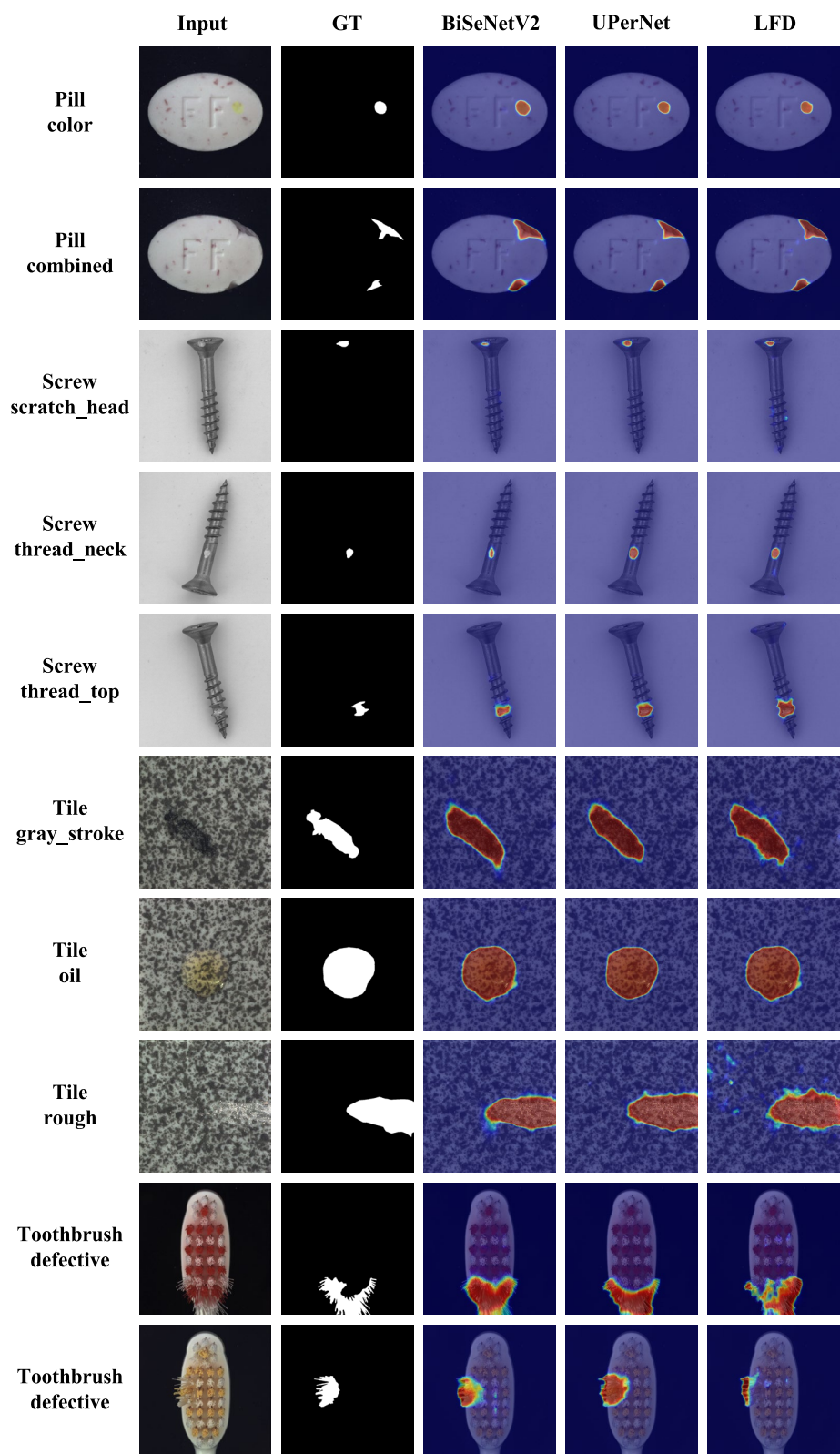


Figure 25. Qualitative comparison results with the supervised segmentation models on MVTec AD. In the figure, from top to bottom are the results for *pill*, *screw*, *tile* and *toothbrush* categories.

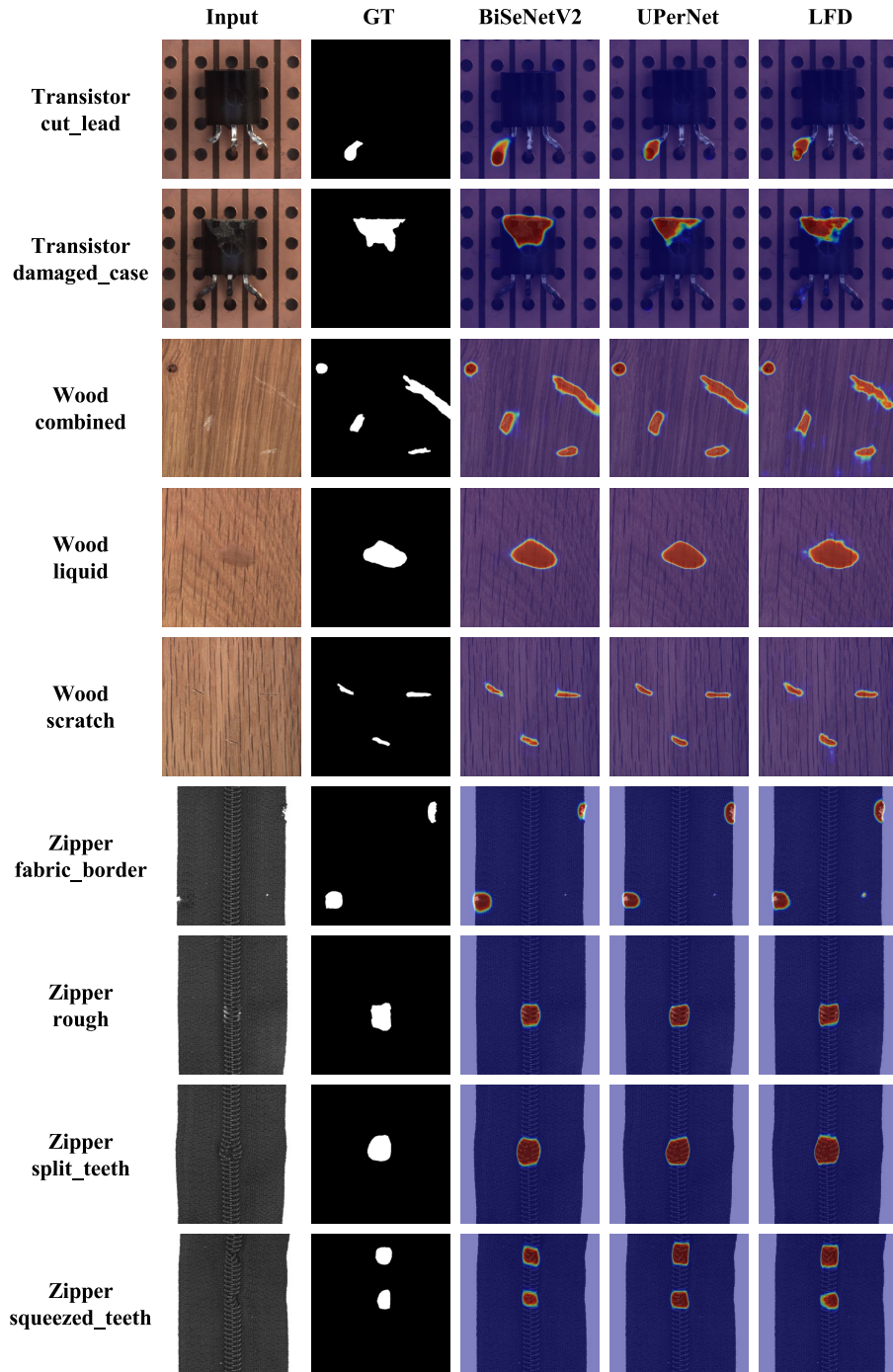


Figure 26. Qualitative comparison results with the supervised segmentation models on MVTec AD. In the figure, from top to bottom are the results for *transistor*, *wood*, and *zipper* categories.

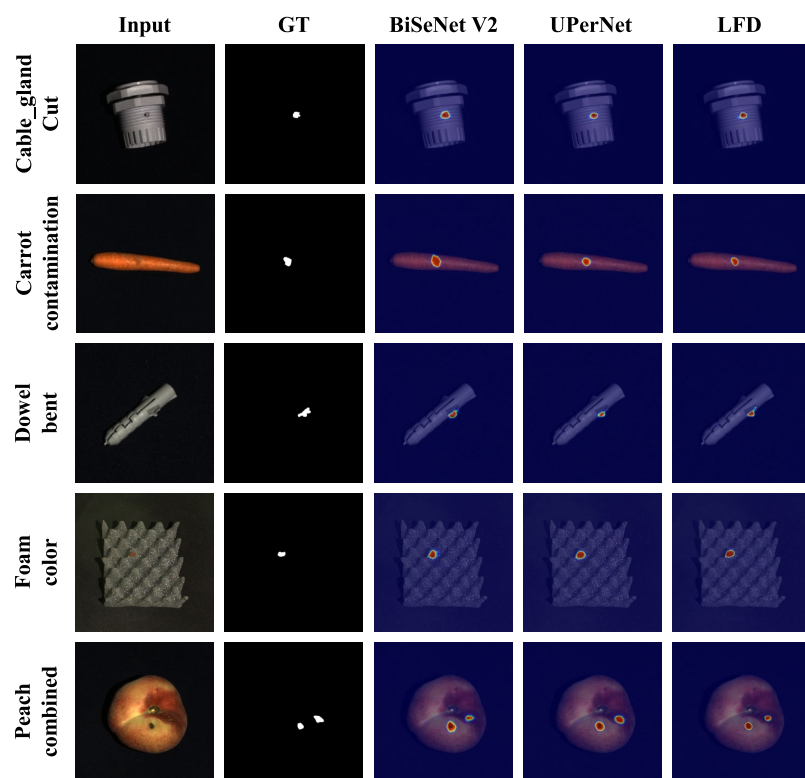


Figure 27. Qualitative comparison results with the supervised anomaly segmentation models on MVTec 3D AD. In the figure, from top to bottom are the results for *cable_gland*, *carrot*, *dowel*, *foam* and *peach* categories.

A.9. Comparison with the Textual Inversion

We conduct the experiment of only using the Textual Inversion (TI) [12] method to learn the product, and the generated images are shown in Fig. 28. The TI method struggles to generate images similar to the real product due to the limited number of learnable parameters. In contrast, for the AIG method, the products satisfy global consistency with minor variations in local details, while the anomalies have randomness, so the generated products should be globally consistent with the real products. Therefore, unlike the method AnomalyDiffusion [17], where the TI method alone is sufficient to meet the anomaly generation needs, we fine-tune the U-Net to ensure the global consistency of the generated products.

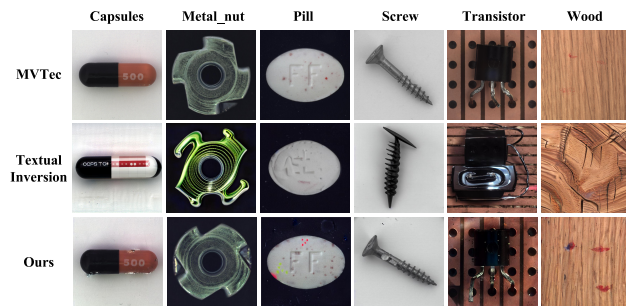


Figure 28. Qualitative comparison on the generation results with Textual Inversion.

A.10. More experiments on lighting conditions

We choose one defect class from peach, a product in the MVTec3D dataset, that has significant variations in lighting conditions and backgrounds, to conduct experiments. Images with strong lighting conditions depict the top side of the peach, whereas those with weak lighting conditions show the bottom side. Consequently, the background in the images, whether the top or bottom of the peach, also differs. We selected three training sets with different lighting conditions for experiments: 1) only images from the top side with strong lighting condition, 2) only images from the bottom side with weak lighting condition, 3) half of the images from the top side with strong lighting condition, and a half from the bottom side with weak lighting condition. The generated images of different settings are shown in Fig. 29. It can be seen that SeaS is robust against lighting conditions and background variations.

A.11. More results on generation of small defects.

SeaS is capable of preserving fine-grained details in small-scale anomalies, as shown in Fig. 30. However, generating extremely subtle anomalies may be challenging due to the limited resolution of the latent space. We will explore this point in our future work.

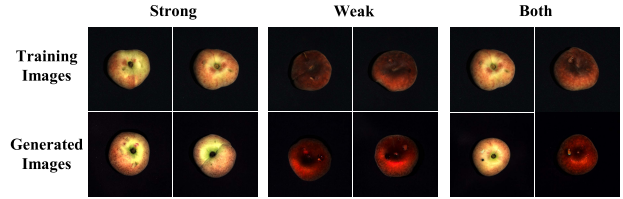


Figure 29. Visualization of the generation results on MVTec3D AD on different lighting conditions and backgrounds. In the figure

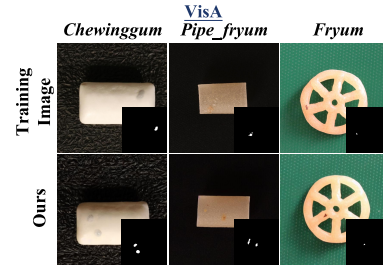


Figure 30. Generation results of small-scale anomalies.

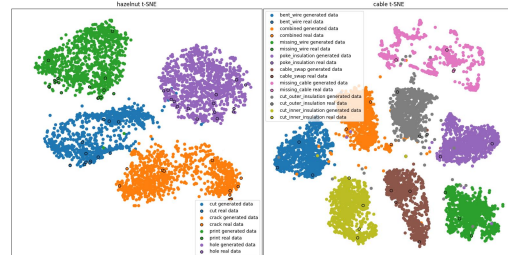


Figure 31. T-SNE visualization of different anomaly types of the same product in real and generated data.

A.12. More analysis on generation of unseen anomaly types.

SeaScan can generate diverse unseen anomalies within known anomaly types as analyzed in Appendix A.2. However, generating truly unseen anomaly types remains challenging. The t-SNE visualizations in Fig. 31 show that different types of anomalies of the same product form compact clusters. Intra-cluster variation is achievable, but cross-cluster generalization is limited by the lack of prior knowledge. We believe that generalizing to unseen anomaly types is important and will explore this in future work.

A.13. More experiments on comparison with DRAEM.

As shown in Tab. 28, training DRAEM[41] with the same anomaly images used in SeaS leads to better results than using only anomaly-free images. However, DRAEM + SeaS achieves further improvements, demonstrating that the gain is not only from real anomalies but also from the diverse and realistic anomalies generated by SeaS.

Table 28. Comparison on combining generated anomalies with synthesis-based anomaly detection method across multiple datasets.

| Segmentation Models | MVTec AD | | | | | | | | VisA | | | | | | | | MVTec 3D AD | | | | | | | |
|------------------------|--------------|--------------|--------------|--|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--|--------------|--------------|--------------|--------------|
| | Image-level | | | | Pixel-level | | | | Image-level | | | | Pixel-level | | | | Image-level | | | | Pixel-level | | | |
| | AUROC | AP | F_1 -max | | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | | AUROC | AP | F_1 -max | IoU | AUROC | AP | F_1 -max | | AUROC | AP | F_1 -max | IoU |
| DRAEM | 98.00 | 98.45 | 96.34 | | 97.90 | 67.89 | 66.04 | 60.30 | 86.28 | 85.30 | 81.66 | | 92.92 | 17.15 | 22.95 | 13.57 | 79.16 | 90.90 | 89.78 | | 86.73 | 14.02 | 17.00 | 12.42 |
| DRAEM + training data | 97.43 | 98.84 | 97.84 | | 96.41 | 74.42 | 71.86 | 59.84 | 83.74 | 86.00 | 82.75 | | 94.63 | 39.22 | 43.06 | 29.02 | 73.86 | 88.46 | 86.69 | | 82.43 | 19.36 | 25.05 | 17.01 |
| DRAEM + SeaS | 98.64 | 99.40 | 97.89 | | 98.11 | 76.55 | 72.70 | 58.87 | 88.12 | 87.04 | 83.04 | | 98.45 | 49.05 | 48.62 | 35.00 | 85.45 | 93.58 | 90.85 | | 95.43 | 20.09 | 26.10 | 17.07 |