MAGIC: Few-Shot Mask-Guided Anomaly Inpainting with Prompt Perturbation, Spatially Adaptive Guidance, and Context Awareness

Supplementary Material

1. Additional details on datasets and experimental settings

1.1. Details on experimental datasets.

MVTec-AD, VisA, and MVTec 3D-AD datasets. For these datasets, we followed a unified experimental setup based on the protocol proposed in AnomalyDiffusion [3]. All three datasets were trained under the same configuration to ensure consistency across benchmarks. For the VisA dataset, we used only the single-anomaly subset.

DAGM dataset. To follow the few-shot setting, we used five anomaly images per class as training images and treated the remaining images as test samples. Since DAGM primarily consists of texture-type defects, all categories were trained under identical settings without CAMA.

1.2. Implementation details

Our framework is based on the Stable Diffusion 2 inpainting model, using a DDIM scheduler with 50 denoising steps during inference. Training is performed for 5000 steps using AdamW (learning rate 5×10^{-6}) with a batch size of 4 on a single NVIDIA RTX A6000 GPU, requiring approximately 1.5 hours per anomaly class. Inference takes about 1 second per image without CAMA and roughly 5 seconds with CAMA. For semantic correspondence extraction in CAMA, we use the pretrained GeoAware-SC model [8].

2. Additional results

2.1. Ablation on hyperparameters

We investigate the impact of three key hyperparameters: the texture perturbation scale σ used in Gaussian Prompt Perturbation, the mask perturbation scale σ_m, and the minimum CFG scale w_m used in our Spatially Adaptive Guidance (SAG) mechanism. A summary of these hyperparameters is provided in Table 1. The texture perturbation scale σ, applied during both training and inference, controls the degree of global variation in the generated anomalies. We observe that $\sigma = 1.0$ strikes the most effective balance between diversity and fidelity, resulting in the best downstream anomaly classification performance. Smaller values reduce anomaly diversity, whereas larger values introduce excessive variation that can slightly degrade reconstruction fidelity and classification accuracy. Similarly, for the mask perturbation scale σ_m, we find that $\sigma_m = 0.1$ achieves the optimal downstream anomaly classification performance, offering a stable balance between perturbation strength and feature consistency. As shown in Table 2, We also examine the effect of the minimum CFG scale w_m that governs the initial guidance strength within anomaly regions in SAG. Lower values encourage diversity in early diffusion steps, while higher values restrict variability. Among the tested configurations, we observe that $w_m = 4.0$ provides the best overall performance. Although the model remains relatively robust to moderate changes in these hyperparameters, we report these ablation results in the main paper [1] to support reproducibility and to highlight empirically effective settings.

2.2. CAMA ablation study

IoU evaluation. To directly evaluate how much of the generated anomaly mask M_A falls onto the object's foreground mask M_O, we compute an alignment score defined as $|M_A \cap M_O|/|M_A|$. A higher score indicates better alignment. We evaluated this metric on the object-centric classes from MVTec-AD, MVTec 3D-AD, and VisA.

As illustrated in Table 3, CAMA provides a consistent improvement in mask alignment across all three tested datasets, suggesting it as an effective module.

Qualitative results. As illustrated in Fig. 1, incorporating multiple spatial keypoints can improve the accuracy of anomaly placement in Context-Aware Mask Alignment (CAMA). In particular, relying only on the central keypoint p_c may lead to suboptimal alignment, especially for anomaly classes that involve missing or removed structural parts, such as *missing_cable* in the *cable* category or *cut_lead* in *transistor*. In such cases, the central region often lacks distinctive features, making correspondence based solely on p_c less reliable. To address this, we supplement the alignment process with additional keypoints—namely, the upper and lower keypoints p_u and p_ℓ—to provide better spatial guidance. Fig. 1 qualitatively shows that using all three keypoints together consistently helps to guide the anomaly to more semantically appropriate regions.

3. Implementation details and reproduction settings for baselines

DualAnoDiff. We reproduced the DualAnoDiff baseline using the official implementation provided by the authors. Nonetheless, we found that our reproduced results did not fully match the quantitative values reported in the original paper. For the MVTecAD dataset, we applied the Background Compensation Module (BCM) to the same cate-

σ	KID ↓	Cls. Acc. ↑	AUC-P ↑	AP-P ↑
0.3	41.65	76.16	98.5	80.0
0.5	41.08	76.32	99.1	80.7
0.7	41.08	77.54	99.1	82.1
0.9	40.89	77.73	**99.2**	83.0
1.0	**40.27**	**78.06**	**99.2**	82.0
1.2	41.32	76.40	**99.2**	**83.6**

σ_m	KID ↓	IC-LPIPS ↑
0.5	55.13	**0.111**
0.1	**46.30**	0.098
0.05	49.82	0.097

Table 1. Classification and pixel-level detection performance under varying hyperparameters. The left subtable presents the effect of the texture-related perturbation scale σ used for texture anomaly generation, while the right subtable shows the effect of the mask-related perturbation scale σ_m used for mask generation.

w_m	KID ↓	Cls. Acc. ↑	AUC-P ↑	AP-P ↑
2.5	40.28	77.34	99.2	83.1
3.0	42.04	77.15	99.1	82.8
3.5	**39.39**	76.32	99.1	82.6
4.0	40.72	**78.06**	99.1	82.0
4.5	40.32	76.89	99.2	83.2
5.0	40.39	76.62	**99.3**	**83.9**

Table 2. Effect of the minimum guidance scale w_m used in Spatially Adaptive Guidance (SAG).

Dataset	Alignment Score (%) w/o CAMA	w/ CAMA
MVTec-AD	69.71	**91.06** (+ 21.35 pp)
MVTec 3D-AD	79.71	**86.86** (+ 7.15 pp)
VisA	50.11	**73.06** (+ 22.95 pp)

Table 3. Evaluation on CAMA. It measures how much of the input mask lies on the object's foreground mask.

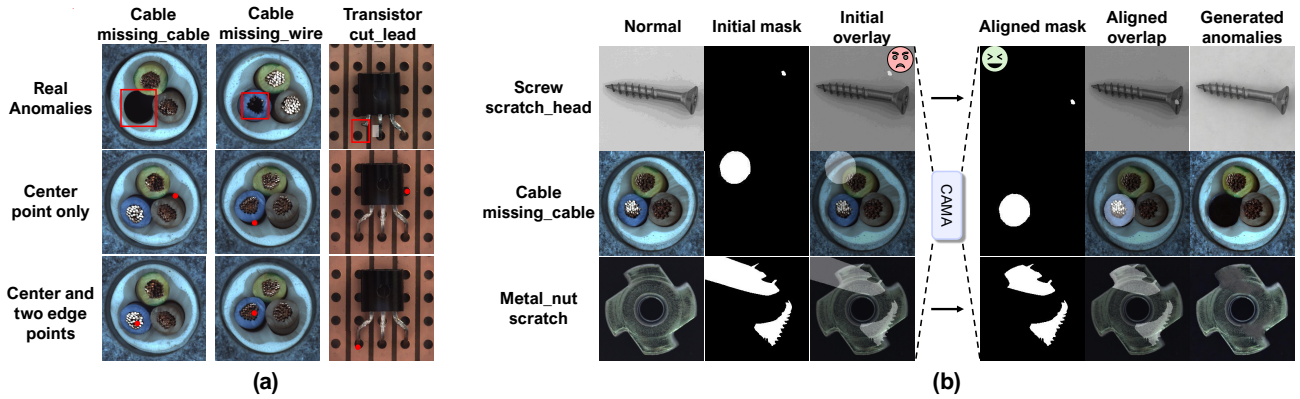

Figure 1. Qualitative results of keypoint-based anomaly localization. (a) Comparison between using the center point only (p_c) and using the center point with edge points (p_u, p_ℓ) in CAMA. Red dots denote the matched positions q_c^* on the normal image. When edge points are included, q_c^* is more accurately localized to semantically meaningful regions. (b) With CAMA, even misaligned masks are refined to align with the normal context, enabling precise anomaly generation.

gories reported in their paper: *bottle*, *toothbrush*, and *pill*. For the VisA and MVTec 3D-AD datasets, BCM was applied to all categories following our unified evaluation protocol. In contrast, for the DAGM dataset—which primarily consists of texture-type anomalies—we did not apply BCM, as foreground–background separation is not meaningful in this setting. Finally, for all datasets (MVTecAD, VisA, and MVTec 3D-AD), foreground masks were extracted using U^2-Net with the default threshold of 0.5, as specified in the original codebase.

Table 4 in this document summarizes the quantitative results in the following order: our reproduced model based on the official implementation, the results obtained from the official pre-trained checkpoint released by the authors, and the original performance reported in the paper. Additionally, we report classification performance and our proposed method's results under the same evaluation setup.

Furthermore, according to the authors of the original DualAnoDiff paper [4], the number of defective samples was divided by three and, in some cases, selectively rounded up when a remainder was present. However, it is not clearly specified in the paper which categories this rounding was applied to, nor which specific image indices or labels were used for training. To avoid reproduction variation due to such ambiguity, we adopted a deterministic protocol that is identical to the setting used in *AnomalyDiffusion* [3] by

Category	DualAnoDiff (official code)				DualAnoDiff* (pretrained weights)				DualAnoDiff† (paper values)				MAGIC (ours)			
	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I	AUC-P	AP-P	F1-P	AP-I
bottle	**99.7**	**98.7**	88.4	81.4	99.3	92.7	84.1	99.9	99.5	93.4	85.7	**100.0**	**99.7**	95.4	**88.5**	**100**
cable	**98.6**	**91.3**	72.9	69.9	91.2	54.4	52.6	89.4	97.5	82.6	**76.9**	98.3	96.5	81.5	75.4	**99.3**
capsule	98.8	47.8	52.8	99.0	97.8	45.7	46.0	96.2	**99.5**	**73.2**	**67.0**	99.2	98.3	61.7	60.8	**99.5**
carpet	99.1	87.1	78.5	99.6	**99.6**	87.8	79.3	99.7	99.4	**89.1**	**80.2**	**99.9**	99.1	86.5	78.4	98.7
grid	95.4	54.9	55.0	99.3	98.6	53.3	52.1	**100.0**	98.5	57.2	54.9	99.7	**99.5**	**59.3**	**59.0**	99.9
hazelnut	99.5	88.7	81.0	99.9	99.7	96.3	91.1	99.9	**99.8**	**97.7**	**92.8**	**100.0**	**99.8**	95.9	90.2	**100.0**
leather	**100.0**	**99.8**	**82.8**	73.6	99.9	86.5	77.3	**100.0**	99.9	88.8	78.8	73.6	99.6	82.2	74.4	**100.0**
metal_nut	99.5	96.7	91.8	99.8	99.5	97.5	91.0	99.7	99.6	98.0	93.0	99.9	**99.8**	**99.0**	**95.3**	**100.0**
pill	97.1	83.0	77.9	97.4	96.3	80.0	75.3	98.3	**99.6**	**95.8**	**89.2**	99.0	**99.6**	90.2	82.4	**99.6**
screw	98.2	50.4	51.2	97.6	95.2	9.4	14.0	95.0	98.1	**57.1**	**56.1**	95.0	**99.2**	51.4	52.1	99.4
tile	99.6	95.4	89.2	**100.0**	99.6	95.3	88.8	**100.0**	99.7	97.1	91.0	**100.0**	**99.8**	**97.9**	**92.5**	**100.0**
toothbrush	93.8	56.2	60.9	96.3	98.3	53.4	59.1	99.0	98.2	68.3	68.6	99.7	**99.2**	**71.2**	**74.5**	**100.0**
transistor	92.5	73.9	71.1	98.1	98.0	77.4	71.0	93.6	98.0	**86.7**	**79.6**	93.7	**98.9**	84.9	78.3	**100.0**
wood	98.3	84.6	77.2	99.9	99.0	87.8	80.0	**100.0**	**99.4**	**91.6**	**83.8**	99.9	98.8	85.1	77.4	99.6
zipper	99.5	89.0	82.0	**100.0**	99.5	86.0	78.6	**100.0**	**99.6**	**90.7**	**82.7**	**100.0**	99.5	87.6	80.8	**100.0**
Average	97.4	76.8	72.9	98.6	98.1	73.6	69.4	98.0	99.1	**84.5**	**78.7**	99.0	**99.2**	82.0	77.3	**99.7**

Table 4. Detection & localization results (AUC-P, AP-P, F1-P, AP-I). DualAnoDiff denotes our reproduced results using the official implementation; DualAnoDiff* indicates results obtained using the author's GitHub pretrained weights; DualAnoDiff† shows the values reported in the original paper. **Note that we have not been able to reproduce the paper values to this date.** For MAGIC (ours), AUC-P, AP-P, F1-P, and AP-I correspond to AUROC-P, AP-P, f1_max-P, and PRO-P from Table 2. Bold = best, underline = second-best.

Category	DualAnoDiff (official code)	DualAnoDiff* (pretrained weights)	DualAnoDiff† (paper values)	MAGIC (ours)
bottle	72.09	58.14	**79.07**	76.74
cable	56.25	56.25	**78.12**	68.75
capsule	48.00	53.33	**70.67**	58.67
carpet	70.97	67.74	**79.03**	62.90
grid	60.00	70.00	**80.00**	60.00
hazelnut	85.42	83.33	89.58	**97.92**
leather	84.13	**95.24**	90.48	85.71
metal_nut	76.56	70.31	89.06	**90.62**
pill	33.33	50.00	56.25	**67.71**
screw	58.02	48.15	70.37	**82.72**
tile	98.25	**100.00**	**100.00**	**100.00**
transistor	**71.43**	67.86	**71.43**	**89.29**
wood	71.43	**92.86**	85.71	73.81
zipper	73.17	26.83	**75.61**	**78.05**
Average	68.50	67.15	**79.67**	78.06

Table 5. Quantitative comparison of anomaly classification accuracy (%) across different generation methods using ResNet-34 [2]. DualAnoDiff denotes our reproduced results using the official implementation; DualAnoDiff* indicates results obtained using the author's GitHub pretrained weights; DualAnoDiff† shows the values reported in the original paper. **Note that we have not been able to reproduce the paper values to this date.** Bold = best, underline = second-best.

sorting test anomaly images within each anomaly class and using the first third as the training set without rounding up. Accordingly, all experiments in our main paper [1] were conducted using this consistent setup.

In Table 4 and Table 5, both our method and the reproduced DualAnoDiff baseline (denoted as DualAnoDiff) generate anomaly images following the same data selection protocol as AnomalyDiffusion [3] to ensure consistency in comparison. This reproduction was performed using the official codebase released by the authors. DualAnoDiff* refers to the results obtained using the official checkpoint provided by the original authors, while DualAnoDiff† denotes the original performance reported in their paper. While our reproduced results are slightly lower than the original DualAnoDiff† scores, they generally achieve higher performance than those obtained using the publicly

Category	DefectFill (paper values) KID ↓	DefectFill (paper values) IC-LPIPS ↑	MAGIC (ours) KID ↓	MAGIC (ours) IC-LPIPS ↑
bottle	**30.99**	0.12	47.16	**0.13**
capsule	**5.60**	0.18	22.61	**0.20**
carpet	50.37	**0.22**	**33.85**	**0.22**
hazelnut	**1.13**	0.31	5.75	**0.32**
leather	**74.66**	0.30	113.50	**0.31**
pill	**8.76**	0.23	59.52	**0.26**
tile	**45.14**	**0.44**	51.47	0.43
toothbrush	**3.19**	0.15	30.62	**0.25**
wood	**4.72**	**0.35**	28.53	0.34
zipper	**34.91**	0.20	58.61	**0.20**
Average	**25.95**	0.25	45.16	**0.27**

Category	DefectFill (paper values)	MAGIC (ours)
bottle	97.56	**97.67**
capsule	87.50	**98.67**
carpet	87.72	**90.32**
hazelnut	**100**	**100**
leather	93.22	**100**
pill	**97.53**	95.83
tile	**100**	**100**
wood	**100**	92.86
zipper	90.91	**98.78**
Average	94.94	**97.13**

Table 6. Quantitative results of DefectFill and MAGIC. The left table reports KID and IC-LPIPS scores, and the right table shows anomaly classification accuracy using ResNet-34 [2]. All results are obtained following the same evaluation protocol described in the DefectFill paper [7], where test-set anomaly masks are used during generation. Since this setting relies on test annotations and does not explicitly account for variation in object poses (e.g., rotation or translation), we respectfully present these results in the supplementary material rather than the main paper [1], in order to maintain alignment with our assumption of training-only supervision.

Category	DefectFill (paper values) AUC-P	AP-P	F1-P	PRO	MAGIC (ours) AUC-P	AP-P	F1-P	PRO
bottle	**1.00**	0.96	0.90	0.97	**1.00**	**0.97**	**0.92**	**0.98**
capsule	**1.00**	**0.75**	0.69	0.96	0.99	0.71	**0.70**	**0.96**
carpet	**0.99**	**0.92**	**0.86**	**0.96**	**0.99**	0.89	0.82	0.94
hazelnut	**1.00**	**0.99**	0.94	**0.99**	**1.00**	**0.99**	**0.95**	0.98
leather	**1.00**	**0.91**	**0.83**	**0.98**	**1.00**	0.89	0.81	**0.98**
pill	**1.00**	**0.98**	**0.93**	**0.98**	**1.00**	0.93	0.84	0.97
tile	**1.00**	0.97	0.90	0.98	**1.00**	**0.99**	**0.94**	**0.99**
toothbrush	0.99	0.89	0.82	0.94	**1.00**	**0.98**	**0.93**	**0.98**
wood	0.99	0.89	0.82	0.94	**1.00**	**0.92**	**0.88**	**0.98**
zipper	**1.00**	0.93	0.86	0.98	**1.00**	**0.95**	**0.88**	**0.98**
Average	**1.00**	**0.92**	**0.86**	**0.97**	**1.00**	**0.92**	**0.86**	**0.97**

Table 7. Comparison of pixel-level anomaly detection and localization performance between DefectFill (from paper) and MAGIC.

available DualAnoDiff* model across most categories in both detection and classification tasks.

DefectFill. To assess MAGIC under the same evaluation conditions as DefectFill [7], we additionally conduct experiments following the protocol described in their paper. Specifically, we adopt the same data split, where the generator is trained using one-third of the available anomaly images and evaluated on the remaining two-thirds. We follow the experimental procedure exactly as outlined in the DefectFill paper to ensure comparability under identical conditions. While this setup enables a fair comparison under identical conditions, it is worth noting that DefectFill makes use of **ground-truth anomaly masks** from the test set, which may not be available in practical deployment scenarios. Moreover, the paper does not report results for categories that exhibit significant variation in object poses (e.g., *screw* and *metal_nut*), which are included in our evaluation. To ensure consistency, we follow the same category selection as described in their work and disable CAMA during this evaluation.

In addition, the reported performance of DefectFill is based on images selected using a Low-Fidelity Selection (LFS) strategy, which filters generated samples based on their LPIPS similarity to real anomalies. In contrast, we evaluate MAGIC on all generated images without applying any filtering. As shown in Table 6, DefectFill yields lower KID scores, suggesting slightly higher fidelity. Meanwhile, MAGIC consistently achieves higher IC-LPIPS scores—approximately 0.02 higher—indicating greater diversity. This diversity may contribute to improved downstream performance: Table 6 shows a 2.19% improvement in classification accuracy. In contrast, the detection results in Table 7 show that DefectFill and MAGIC achieve identical scores across AUP-P, AP-P, and all other reported metrics. It is important to note that the evaluation protocol used in DefectFill directly leverages ground-truth anomaly masks from the test set during generation. As a result, the generated anomalies are inherently well-aligned with real defects, which can naturally lead to lower KID scores and improved classification and detection metrics. In contrast, the setting adopted in our main paper avoids reliance on test annotations and instead uses only training data to guide generation, which we believe provides a fairer and more practical framework for evaluating generalization to unseen anomalies.

SeaS. Unlike other baselines and our method, SeaS applies different manually tuned CFG scales per dataset for anomaly image generation. For SeaS, we reproduced the baseline using the official implementation released by the authors and followed the recommended guidance scale (CFG_scale) values exactly as described in the main paper [1]. Specifically, we used a CFG_scale of 8 for MVTecAD, 2 for VisA, and 5 for MVTec 3D-AD. Since the DAGM dataset was not addressed in the original SeaS paper, we applied the same CFG_scale used for MVTecAD (i.e., 8). In contrast, all other baselines, including MAGIC,

were evaluated using the **same hyperparameters** to ensure a fair and consistent comparison.

4. Analysis on artifact boundary

We examined the visual continuity of synthesized defects by measuring the mean gradient magnitude ($|\nabla I|$) along mask boundaries. As shown in Tab. 8, the average gradient of our generated images (5.59) remains lower than that of real images (10.41). These results suggest that our framework does not introduce systematic boundary artifacts, maintaining a natural transition.

	Boundary $\|\nabla I\|$
Real	10.41
Generated	5.59

Table 8. Mean image gradient magnitude ($|\nabla I|$) in mask boundary pixels on MVTec-AD.

Method	KID↓	Cls. acc.↑
CADS(s=0.1)	74.41	55.52
CADS(s=0.05)	74.76	53.65
CADS(s=0.025)	73.98	54.67
DIAGEN	58.95	64.87
Ours (GPP)	**42.59**	**74.36**

Table 9. Comparison vs other prompt augmentation baselines on MVTec-AD.

5. Comparison with other prompt augmentation baselines

While natural-language prompt diversification (noun/adjective swaps) is less applicable in our setting because the anomaly concept is represented by a fixed learned token ("sks"), we nevertheless compare against prompt augmentation baselines CADS [6] and DIAGEN [5] and show GPP yields substantially better KID and downstream accuracy (Tab. 9).

6. Ablation study on SAG

We evaluated various **SAG** scheduling functions—*Linear*, *Exp*, *Poly*, and *Cosine*. As shown in Tab. 10, the **Cosine** scheduler yields the best classification accuracy (78.06%), suggesting that a gradual decay in guidance strength is beneficial for our framework.

	Linear	Exp	Poly	Cosine (ours)
Cls acc.	76.73	75.71	75.51	**78.06**

Table 10. Ablation of schedulers in SAG.

7. Ablation of CAMA design choices.

CAMA uses two boundary points because single-point correspondence is unstable when the center region is weakly distinctive (e.g., "missing part"). The upper and lower keypoints are obtained by drawing a vertical line crossing the mask centroid. The "vertical-line" rule is just for selecting two separated boundary keypoints. Horizontal vs vertical is comparable (Tab. 11), but center-only is worse (Tab. 12), validating the need for edge points.

	Horizontal	Vertical (ours)
Cls acc.	78.02	**78.06**

Table 11. Ablation on candidate line direction.

	Center-only	Ours
Cls acc.	75.68	**78.06**

Table 12. Ablation on edge points.

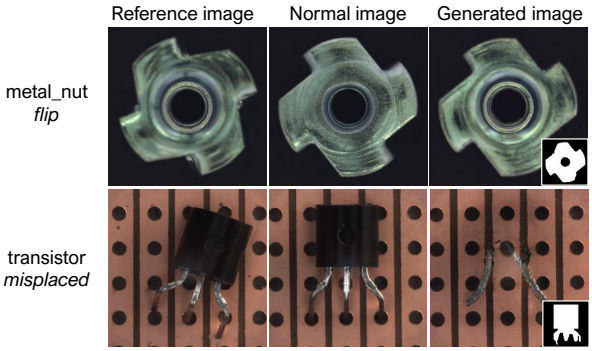

Figure 2. Failure cases of our method. The examples illustrate situations involving large structural changes or global object displacements, where MAG-based generation fails to produce plausible anomalies.

8. Ablation study on the VisA dataset

We add a module ablation on VisA (Tab. 13), showing consistent gains.

GPP	SAG	CAMA	KID↓	IC-L↑	Acc↑
×	×	×	81.76	0.326	61.34
✓	×	×	81.11	0.333	63.37
✓	✓	×	80.11	**0.336**	65.34
✓	✓	✓	**79.81**	0.323	**68.51**

Table 13. Ablation of MAGIC components on VisA.

9. Ablation of downstream task model

We evaluated the generalization of **MAGIC** using ResNet-18 as a downstream classifier. As shown in Tab. 14, the performance gains observed with ResNet-34 persist even with this smaller architecture. This consistent improvement suggests that our method captures robust anomaly representations regardless of the specific backbone.

	AnoDiff	AnoGen	DualAnoDiff	SeaS	**MAGIC**
Cls acc. (%)	66.14	57.98	69.20	58.68	**77.31**

Table 14. Mean anomaly classification accuracies using ResNet-18 on MVTec-AD.

10. Failure cases

As illustrated in Fig. 2, MAGIC is a mask-guided anomaly generation (MAG) method that synthesizes anomalies by inpainting defect regions on top of normal images. Due to this design choice, MAG inherently struggles with defects that involve large structural changes or global object displacements. For instance, categories such as *metal_nut (flip)* and *transistor (misplaced)* exhibit object-level transformations that cannot be captured as localized inpainting regions, making them particularly challenging for MAG-based approaches to generate.

11. Category-wise detection results

We provide category-wise detection results for all benchmark datasets, including MVTecAD, VisA, DAGM, and MVTec 3D-AD. Detailed per-category performance is summarized in Tab. 15, 16 (MVTec-AD), 17, 18 (VisA), 19, 20 (DAGM), 21, 22 (MVTec 3D-AD), offering a comprehensive view of the model's detection capability across diverse anomaly types and object categories.

12. Qualitative results of generated anomalies

Fig. 3 to 40 present qualitative results of the anomalies generated by our method.

Category	AnomalyDiffusion			AnoGen			DualAnoDiff			SeaS			MAGIC (ours)		
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max
bottle	99.7	99.8	98.9	99.3	99.7	97.6	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
cable	**99.8**	**99.9**	**98.4**	97.9	98.6	94.4	98.2	99	96.8	97.2	97.9	94.4	99.1	99.3	96.1
capsule	96.5	98.9	95.2	97.5	99.3	95.5	96.5	99	94.6	96.9	99.2	95.9	**98.4**	**99.5**	**96.6**
carpet	92.9	97.5	93.2	95.2	98.1	93.4	99	99.6	**98.4**	**99.4**	**99.7**	96.9	96.8	98.7	94.4
grid	98.1	99.3	98.7	97.6	99.2	98.7	98.3	99.3	97.4	99.3	99.7	97.5	**99.9**	**99.9**	**98.8**
hazelnut	99.9	**100**	99	99.6	99.7	97.9	99.9	99.9	98.9	99.8	99.9	98	**100**	**100**	**100**
leather	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.2	**100**	**100**	**100**
metal_nut	**100**	**100**	**100**	99.7	99.9	98.4	99.3	**100**	97.7	99.9	**100**	99.2	**100**	**100**	**100**
pill	97.9	99.5	96.8	96.4	99.1	95.3	89.7	97.4	90.2	**98.7**	**99.7**	96.9	98.3	99.6	**97.4**
screw	93.5	96.9	91.7	92.4	96.1	90.7	81.9	91.3	82.9	94.6	97.6	90.8	**98.8**	**99.4**	**96.4**
tile	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.8	99.9	99.1	**100**	**100**	**100**
toothbrush	98.8	99.3	97.4	**100**	**100**	**100**	92.9	96.3	88.9	**100**	**100**	**100**	**100**	**100**	**100**
transistor	**100**	**100**	**100**	99.4	98.8	94.7	99.1	98.1	94.9	97.2	96.4	94.3	**100**	**100**	**100**
wood	**99.9**	**99.9**	**98.8**	99	99.5	**98.8**	**99.9**	**99.9**	**98.8**	98.1	99.3	97.6	99.1	99.6	97.6
zipper	**100**	**100**	99.4	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Average	98.47	99.40	97.25	98.27	99.20	96.89	97.09	98.64	95.29	98.73	99.29	97.32	**99.36**	**99.73**	**98.49**

Table 15. Category-wise quantitative comparison of image-level anomaly localization on the MVTec-AD dataset.

Category	AnomalyDiffusion				AnoGen				DualAnoDiff				SeaS				MAGIC (ours)			
	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO
bottle	99.3	92.6	84.6	95.8	98	83	74.5	89.9	98.7	88.4	81.4	91.8	99.5	92.6	86.8	97	**99.7**	**95.4**	**88.5**	**97.6**
cable	**98.4**	**84.3**	**76.3**	93.6	94.8	68.3	64.7	88.6	91.3	72.9	69.9	80.9	96	76.7	73	90.1	96.5	81.5	75.4	**93.9**
capsule	97.9	41.6	43.7	89.7	96.3	34	37.5	92.4	**98.8**	47.8	52.8	**95.5**	96.8	44	46	94.7	98.3	**61.7**	**60.8**	95.2
carpet	96.4	74.8	71.8	83.6	98.8	81.4	74	89.6	99.1	**87.1**	**78.5**	94.5	**99.3**	85.3	76.4	**95.1**	99.1	86.5	78.4	94.6
grid	97.3	43.6	46	93.6	96.1	33.1	41.8	93.9	95.4	54.9	55	93.2	98.9	**63.9**	**61**	97.2	**99.5**	59.3	59	**97.8**
hazelnut	99.3	89.4	81.9	94.5	97.2	58.3	56.7	93.1	99.5	88.7	81	95.4	99.6	90.1	82.7	**98.1**	**99.8**	**95.9**	**90.2**	97.3
leather	**99.8**	78.4	71	98.3	99.4	77.7	70.4	98	**99.8**	**82.8**	73.6	**98.4**	99.4	77.1	69	98.1	99.6	82.2	**74.4**	98
metal_nut	99.6	97.8	92.5	95.2	95.2	77	72.5	89.5	99.5	96.7	91.8	92.5	99.7	98.5	94.1	95.6	**99.8**	**99**	**95.3**	**96.4**
pill	**99.6**	**95**	**88.5**	94.4	99.2	89.1	80.7	93.2	97.1	83	77.9	78.4	99.5	90.5	82.9	97.1	**99.6**	90.2	82.4	**97.5**
screw	95.6	11.8	21.6	83.8	92.7	21.9	27.2	77.5	98.2	50.4	51.2	91.5	96.3	48.3	49.8	90.5	**99.2**	**51.4**	**52.1**	**94.9**
tile	99.5	96	88.9	96.9	99.2	93.1	85	97	99.6	95.4	89.2	97.5	99.7	96.8	91.4	97.5	**99.8**	**97.9**	**92.5**	**98.3**
toothbrush	97.9	56.6	58.5	84.9	**99.2**	**93.1**	**85**	72.2	93.8	56.2	60.9	71.1	97.1	66	65.8	86	**99.2**	71.2	74.5	**93.5**
transistor	**99.2**	**89.1**	**81.2**	**96.4**	95	63.2	61.2	88.8	92.5	73.9	71.1	89.8	97.8	83.7	77	92.6	98.9	84.9	78.3	**96.4**
wood	96.7	74.6	67.4	91.3	98.5	78.2	71.5	94	98.3	84.6	77.2	94.5	96.7	80.4	74.8	90.8	**98.8**	**85.1**	**77.4**	**94.7**
zipper	99.4	84.7	78.2	96.6	99.1	78.9	71.8	95.6	**99.5**	**89**	**82**	96.6	98.8	81.1	74.9	95.9	**99.5**	87.6	80.8	**97.2**
Average	98.39	74.02	70.14	92.57	96.25	64.20	61.60	90.65	97.41	76.79	72.90	91.32	98.34	78.33	73.71	94.42	**99.15**	**81.99**	**77.33**	**96.22**

Table 16. Category-wise quantitative comparison of pixel-level anomaly localization on the MVTec-AD dataset.

Category	AnomalyDiffusion			AnoGen			DualAnoDiff			SeaS			MAGIC (ours)		
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max
candle	93.6	**93.1**	**87**	87.1	84.4	76	86.6	80.9	74.5	78.4	76.6	66	**94.5**	91.3	84
capsules	63.9	44.8	46.3	63.6	37.5	52.2	89.8	75.8	71.1	**97.3**	**96.3**	**92.7**	93.5	86.1	78.3
cashew	96.4	97.1	93	**98.7**	**99**	**95.3**	94.4	95.7	87.9	85.6	88.6	82.2	96	96.4	93.8
chewinggum	94.6	95.3	89.7	95.1	95.6	90.1	93.3	91.5	84.6	96.1	96.5	93.2	**99.3**	**99.1**	**96.3**
fryum	**91.5**	**93.7**	**86.8**	88	92.4	83.3	87.6	90.7	82.7	86.5	89.5	84.8	82.8	86.8	81.2
macaroni1	82.6	72.2	69.4	98.9	98.1	93.6	87.4	82.8	71.1	94.5	89.2	86	**99.5**	**99.3**	**96.4**
macaroni2	45.8	30.9	52.8	78.4	71.5	66.7	74.3	55.8	63	75.1	70.7	62.3	**91.8**	**90.4**	**82.4**
pcb1	**95.7**	**94.3**	84.2	94.9	93.7	**85.7**	78.6	80.3	71.3	92.7	90.5	79.7	91.3	86.6	82.2
pcb2	94.6	94.1	88.3	97	96.2	89.8	88.3	88.6	81.1	93.1	92.8	84.4	**97.3**	**97**	**94.2**
pcb3	85.3	85.7	78.9	91.7	88.3	82.7	77.5	74.9	65.4	85.4	84.8	76.5	**94.3**	**92.4**	**87.3**
pcb4	97.6	89.2	84.7	94.7	79.8	77.6	88.6	73.8	66.7	95.9	88.8	87.5	**98.8**	**93.3**	**94.3**
pipe_fryum	78.7	84.5	75.5	87.6	91.2	82.3	91	93.8	86.2	82.7	85	82.8	**92.3**	**94.9**	**87**
Average	85.03	81.24	78.05	89.64	85.64	81.28	86.45	82.05	75.47	88.61	87.44	81.51	**94.28**	**92.8**	**88.12**

Table 17. Category-wise quantitative comparison of image-level anomaly localization on the VisA dataset.

Category	AnomalyDiffusion				AnoGen				DualAnoDiff				SeaS				MAGIC (ours)			
	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO
candle	98.4	33.8	39.4	90.3	96.8	40.2	43.6	88.1	92.1	18.3	28.3	85	93.7	**55.8**	**59.9**	82.9	**99.3**	47.1	46.7	**96.4**
capsules	87.6	8	12.6	56.4	92.3	3.8	11.4	68.8	97.5	51.2	50.6	80.7	**99.9**	72.9	65.1	93.3	**99.9**	**81.5**	**74.1**	**96**
cashew	99	83	78.1	81.5	97.8	68.4	66	**94.9**	**99.8**	**95**	**92.9**	92.4	98.1	85	81.5	83.9	99.5	92	88.7	86.6
chewinggum	99.4	69.2	64.8	92	99.4	80.8	73.9	90.1	99.4	65.5	59.5	90.4	**99.8**	**88.8**	**80**	**96.7**	99.5	30	55.9	96.5
fryum	96.8	53.3	51.8	**94**	97.1	55.5	53.5	89.6	97.6	55.3	55.9	91.1	**98.8**	**83.9**	**77.7**	77.7	97.2	54.3	52.4	92.9
macaroni1	92.6	3.4	9.6	83.8	97.8	35.6	41.1	95.7	94.5	5.2	11.1	90.2	**99**	54.8	**56.3**	89.3	98.8	**57**	54.4	**96.3**
macaroni2	93	0.2	0.6	78	97.8	8.5	13.3	92.4	96.1	9.5	12.2	87.7	98.2	14.2	23.6	90.3	**99.3**	**40.4**	**44**	**95.7**
pcb1	94.9	71.4	70	**88.9**	97	69.9	66.8	87.3	96.1	70.9	68.5	76.8	**98.4**	**86**	**84.5**	84.2	97.9	78.7	76.7	87.7
pcb2	94.9	25.6	33.5	82.5	**97.3**	31.3	39.4	**90**	96.4	34.4	43.3	85.1	93.3	**54.7**	**55.9**	79.8	93.4	24.9	32.1	82.1
pcb3	97.5	28.9	31.9	81.5	**97.9**	33.1	34.9	**86.3**	96.5	25.9	35	80.6	96.8	**68.2**	**65.8**	74.1	96.1	35.4	42.2	81.9
pcb4	98.3	66.4	62.1	92.3	96.4	37.9	46.2	79.5	98	68.5	68.9	85.4	99.4	**82.9**	**78.4**	91.9	**99.8**	71.3	66.4	**96.9**
pipe_fryum	99.7	90.6	84.4	88.3	99.4	79.9	72.1	86.6	99.7	93.8	86.2	95.4	99.2	88	82	86.7	**99.9**	**96.3**	**90**	**95.7**
Average	96.01	44.48	44.9	84.13	97.25	45.41	46.85	87.44	96.98	49.46	51.03	86.73	97.88	**69.60**	**67.56**	85.90	**98.38**	59.08	60.30	**92.06**

Table 18. Category-wise quantitative comparison of pixel-level anomaly localization on the VisA dataset.

Category	AnomalyDiffusion			AnoGen			DualAnoDiff			SeaS			MAGIC (ours)		
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max
Class1	**100**	**100**	**100**	**100**	**100**	99.7	**100**	**100**	99.7	**100**	**100**	**100**	99.9	99.6	99.3
Class2	99.9	99.8	97.9	99.9	99.7	99.3	99.9	99.7	98.6	**100**	**100**	**100**	**100**	**100**	**100**
Class3	**100**	**100**	99.7	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Class4	99	98.2	94.7	99.4	98.8	96.9	95	94.1	91.4	**100**	**100**	**99.7**	99.8	99.5	99
Class5	98.6	97.3	94.6	86.5	80.3	73.6	**100**	**100**	**100**	**100**	**100**	**100**	99.5	99	96.9
Class6	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Class7	99.9	99.9	99.8	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.8	**100**	**100**	**100**
Class8	72.2	51.9	49.9	82.7	71.8	63.7	93.4	89.5	83.1	53.1	25.6	37.2	**99.4**	**98.7**	**95.9**
Class9	**100**	99.9	99	99.9	99.7	97.5	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Class10	99.9	99.8	99.2	**100**	99.9	99	**100**	99.9	99.7	**100**	**100**	**100**	**100**	**100**	99.8
Average	96.95	94.68	93.48	96.84	95.02	92.97	98.83	98.32	97.25	95.31	92.56	93.67	**99.86**	**99.68**	**99.09**

Table 19. Category-wise quantitative comparison of image-level anomaly localization on the DAGM dataset.

Category	AnomalyDiffusion				AnoGen				DualAnoDiff				SeaS				MAGIC (ours)			
	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO
Class1	**99.8**	83.1	72.7	**98.4**	99.6	86	76.1	98.2	99.2	**88**	**80.5**	97.5	98.9	86.1	79.4	96.9	99.5	87.2	79.1	97.7
Class2	**99.6**	70.9	66.2	**97.9**	98.4	64.4	61.9	97.5	99.3	69	64.7	97.8	99.3	71.2	66.4	**97.9**	99	**77.7**	**71.7**	97.1
Class3	99	74.4	68.3	96.7	98.6	77.2	70.9	96.4	98.9	**83.2**	**75.7**	97.1	98.6	76	71.5	96.6	**99.5**	80.3	73.2	**97.7**
Class4	98.1	69.3	65.3	94	97.3	69.7	65.7	93.9	95.4	71.9	69.7	86.8	**99.6**	**84.5**	**77.4**	**97.6**	99.3	79.4	73.8	96.8
Class5	97.8	71	68.9	92.5	92.7	47.7	51.8	80.4	99.1	**87.3**	**81.3**	96.7	**99.5**	85	77.1	**97.4**	99.2	76.5	70.4	96.7
Class6	99.5	90.8	82.4	97.4	99.6	90.7	81.8	97.4	99.8	95.1	88.1	97.9	**99.9**	**97.5**	**90.9**	**98.4**	**99.9**	96.5	89.5	98.2
Class7	98.8	85.3	78.9	96.5	96.8	80.3	77	93.5	**99.7**	**91.3**	**82.2**	**98.6**	99.4	85.9	77.9	97.7	**99.7**	89.3	80.6	98
Class8	91	19.5	26.7	72.1	91.2	32.1	40.7	77.8	89.2	42.2	48.4	82.5	50.5	0.1	0.2	16.3	**98.7**	**59.1**	**58.8**	**97.1**
Class9	**100**	92.2	83.6	98.5	**100**	85.6	75.7	98.2	**100**	97.5	**90.8**	99	**100**	96	88.1	98.8	**100**	**97.6**	**90.8**	**99.1**
Class10	96.1	66.8	63.7	94.3	**98.2**	65.9	62.8	**96.5**	97.1	**77.4**	**71.7**	95.8	97	66.5	65.2	95.2	96.7	75.8	70.5	95.4
Average	97.97	72.33	67.67	93.83	97.24	69.96	66.44	92.98	97.77	80.29	75.31	94.97	94.27	74.88	69.41	89.28	**99.15**	**81.94**	**75.84**	**97.38**

Table 20. Category-wise quantitative comparison of pixel-level anomaly localization on the DAGM dataset.

Category	AnomalyDiffusion			AnoGen			DualAnoDiff			SeaS			MAGIC (ours)		
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUROC	AP	F_1-max
bagel	94.5	98.1	93.4	93.5	97.9	94.1	**98.7**	**99.6**	**97.5**	97.9	99.3	96.7	97.0	98.7	96.8
cable_gland	**98.9**	**99.6**	**97.5**	88.5	94.7	91.8	61.8	84.8	84.9	91.2	96.8	92.1	98.2	99.4	96.6
carrot	**97.2**	99.2	**96.7**	94.1	98	94.1	90.7	97.1	91.3	92.6	97.2	94.4	97.0	**99.4**	95.2
cookie	88.9	94.7	91.8	89.9	95.6	91.9	84.3	93.7	85.7	92.0	97.2	90.2	**95.8**	**98.4**	**92.7**
dowel	64.4	86.4	84.8	83.7	94.3	86.3	82.7	93.5	88	79.8	92.8	85.2	**99.0**	**99.7**	**97.8**
foam	**93.5**	**97.9**	91.6	92.9	97.5	**92.7**	82.1	93.7	86.6	71.9	90.1	85.9	91.2	97.0	90.4
peach	79.7	90.3	85.7	90.2	96.7	88.7	88	95.3	89.7	**95.7**	**98.4**	**93.3**	91.8	97.3	90.2
potato	71.4	87.8	87	74.1	88.3	87.1	62.9	83.1	85.3	61.6	83.5	85.9	**86.9**	**95.8**	**87.8**
rope	99.5	99.7	97.9	98.5	99.2	96.8	98.1	99	96.7	**100.0**	**100.0**	**100.0**	91.2	95.1	86.5
tire	72.7	88.3	83.2	75.6	88.6	85.5	58.3	75.9	83.1	**89.7**	**96.0**	**88.7**	85.6	94.7	88.1
Average	86.07	94.2	90.96	88.1	95.08	90.9	80.76	91.57	88.88	87.24	95.13	91.24	**93.37**	**97.52**	**92.21**

Table 21. Category-wise quantitative comparison of image-level anomaly localization on the MVTec 3D-AD dataset.

Category	AnomalyDiffusion				AnoGen				DualAnoDiff				SeaS				MAGIC (ours)			
	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO	AUROC	AP	F_1-max	PRO
bagel	98.4	3.4	8.8	94.7	98.5	4.7	11.7	95.6	99.3	7.8	15.2	95.5	99.3	5.3	12.2	97.1	**99.7**	**11.0**	**19.4**	**98.2**
cable_gland	95.7	1.5	4.4	86.2	**99.3**	**9.5**	**18.9**	**97**	95.3	1.6	5.7	83.6	93.4	4.3	12.2	89.4	98.7	5.2	12.0	95.4
carrot	98.9	7.2	17.1	95.2	99.4	6.3	14	97.8	99.2	14.2	23.7	97.1	99.8	24.1	31.6	**98.1**	**99.9**	**29.9**	**34.7**	97.8
cookie	88.4	2.3	6.2	87.8	91.2	6.8	17.4	85.1	94.9	**18.9**	**33.5**	93.3	94.7	8.3	19.2	92.2	**97.5**	13.2	23.4	**95.9**
dowel	99.3	33.8	38.4	87.2	99	25.7	36.8	89.4	98.4	30.5	39.7	89.1	**99.8**	**57.3**	**57.8**	86.9	99.7	54.0	51.7	**97.9**
foam	99.9	74.6	68.7	90.2	99.9	65.3	61.1	82.2	96.4	58	63	73	99.8	70.0	66.5	83.4	**100.0**	**78.5**	**71.0**	**96.5**
peach	97.2	3.4	9.3	91.8	99.3	15.5	24.3	96.1	99.2	21.3	29.6	96.1	99.3	**37.0**	**40.3**	96.8	**99.8**	20.0	31.1	**97.3**
potato	97.5	1.9	5.1	90.7	99.5	14.8	25.4	90.4	96.9	7.5	14	90.8	96.4	10.1	14.1	95.1	**99.6**	**17.7**	**25.9**	**97.1**
rope	**99.2**	7.3	12.9	**97.1**	98.5	**11**	11.5	96.5	98.8	5.2	11.7	96.6	94.7	10.3	**17.0**	93.5	98.3	8.6	10.0	95.2
tire	98.6	16.2	27.6	92.3	98.4	**40**	**43.5**	83.2	95	2.9	10.3	81.7	**99.5**	31.9	35.8	**93.1**	98.6	22.0	28.0	92.5
Average	97.31	15.16	19.85	91.32	98.3	19.96	26.46	91.33	97.34	16.79	24.64	89.68	97.67	25.86	30.67	92.56	**99.18**	**26.01**	**30.72**	**96.38**

Table 22. Category-wise quantitative comparison of pixel-level anomaly localization on the MVTec 3D-AD dataset.

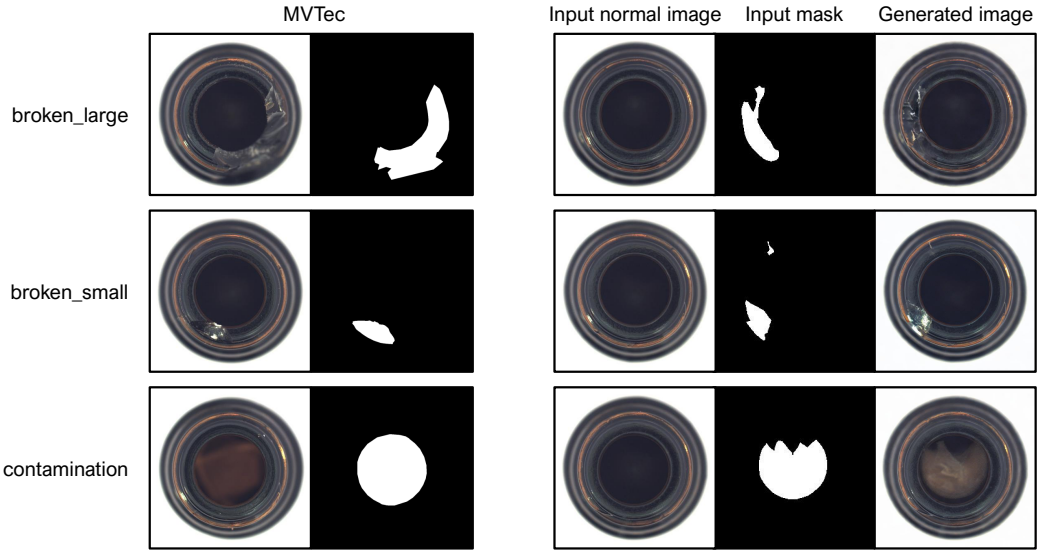

Figure 3. Generated images on *bottle*

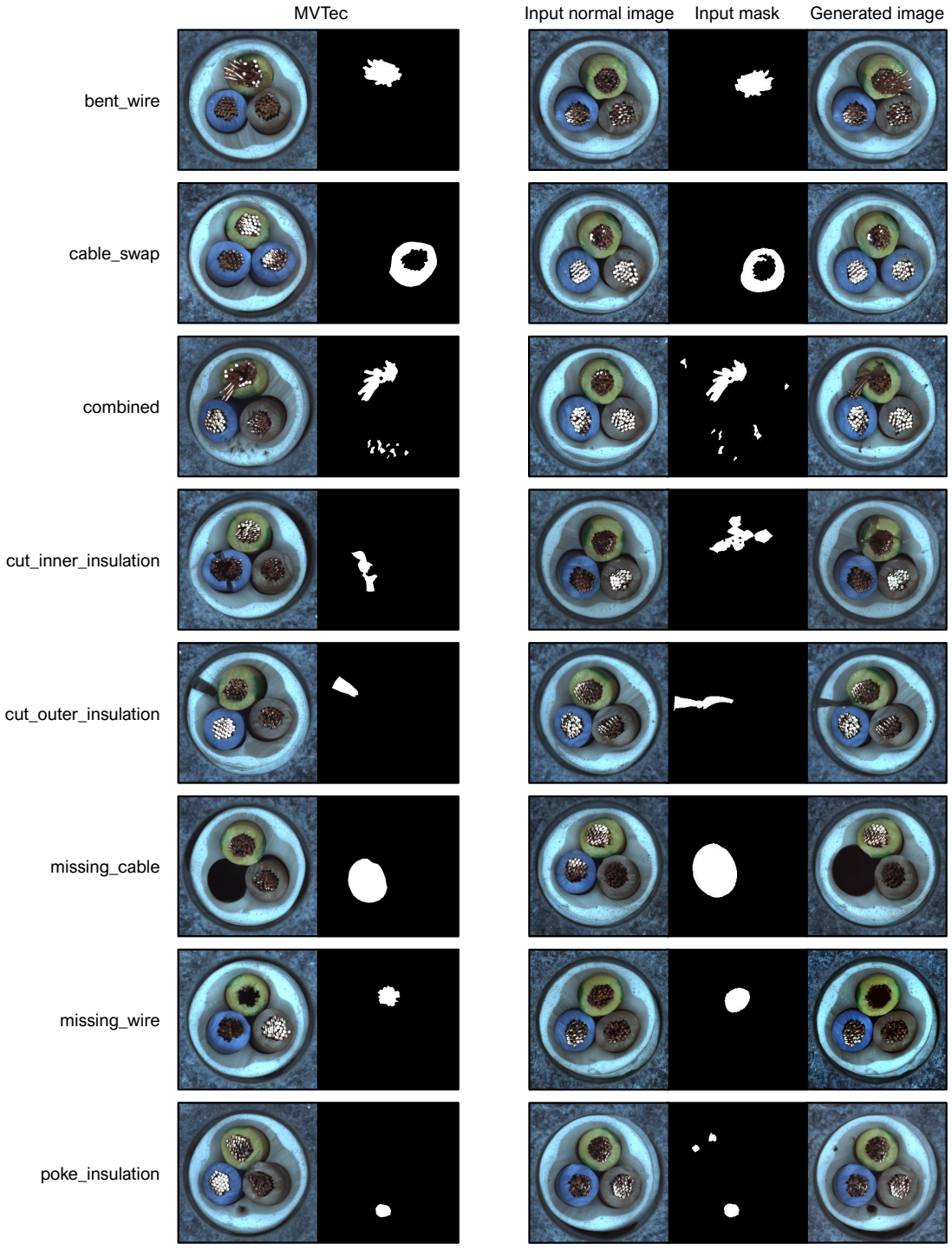

Figure 4. Generated images on *cable*

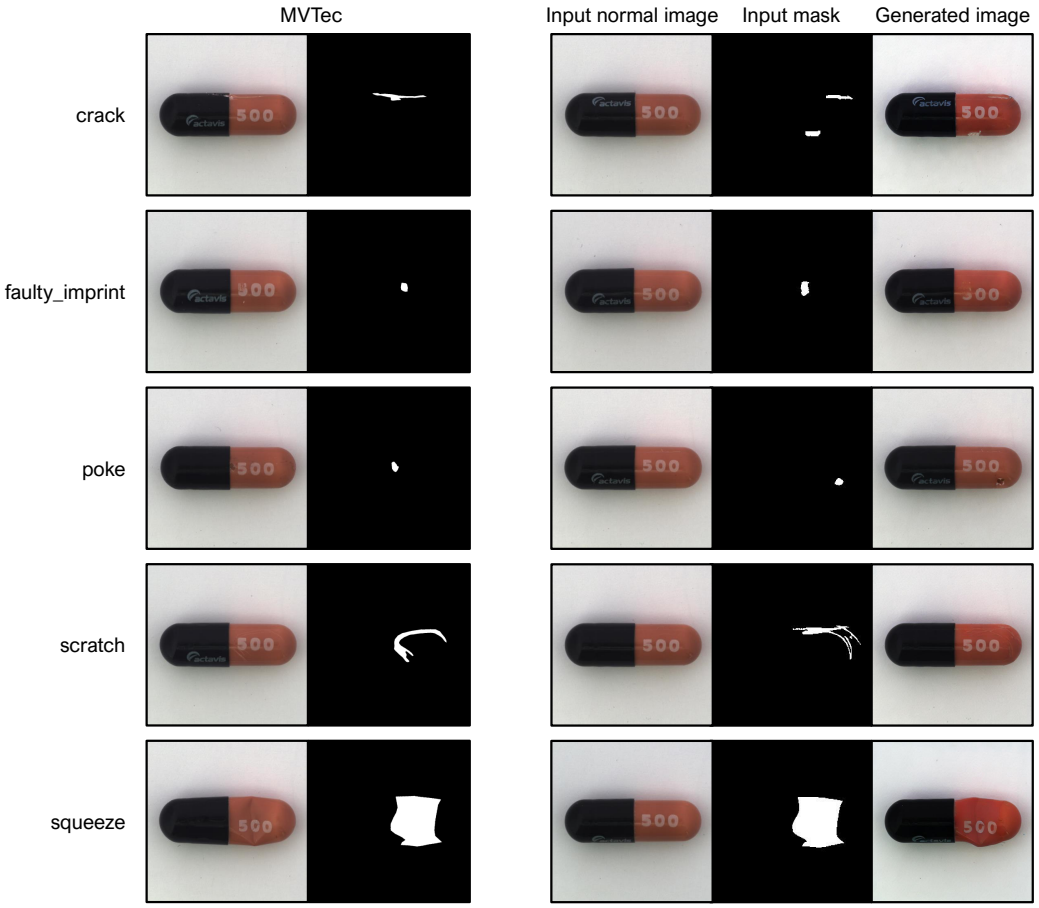

Figure 5. Generated images on *capsule*

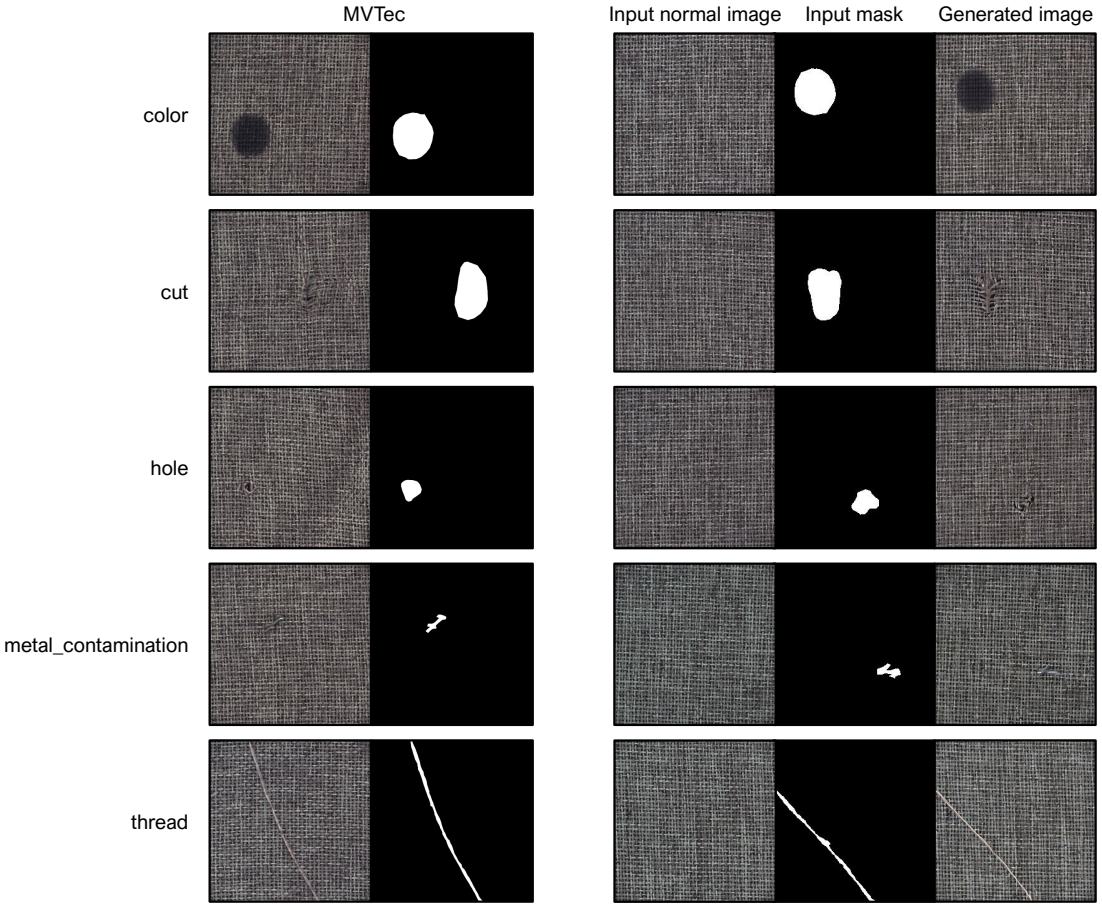

Figure 6. Generated images on *carpet*

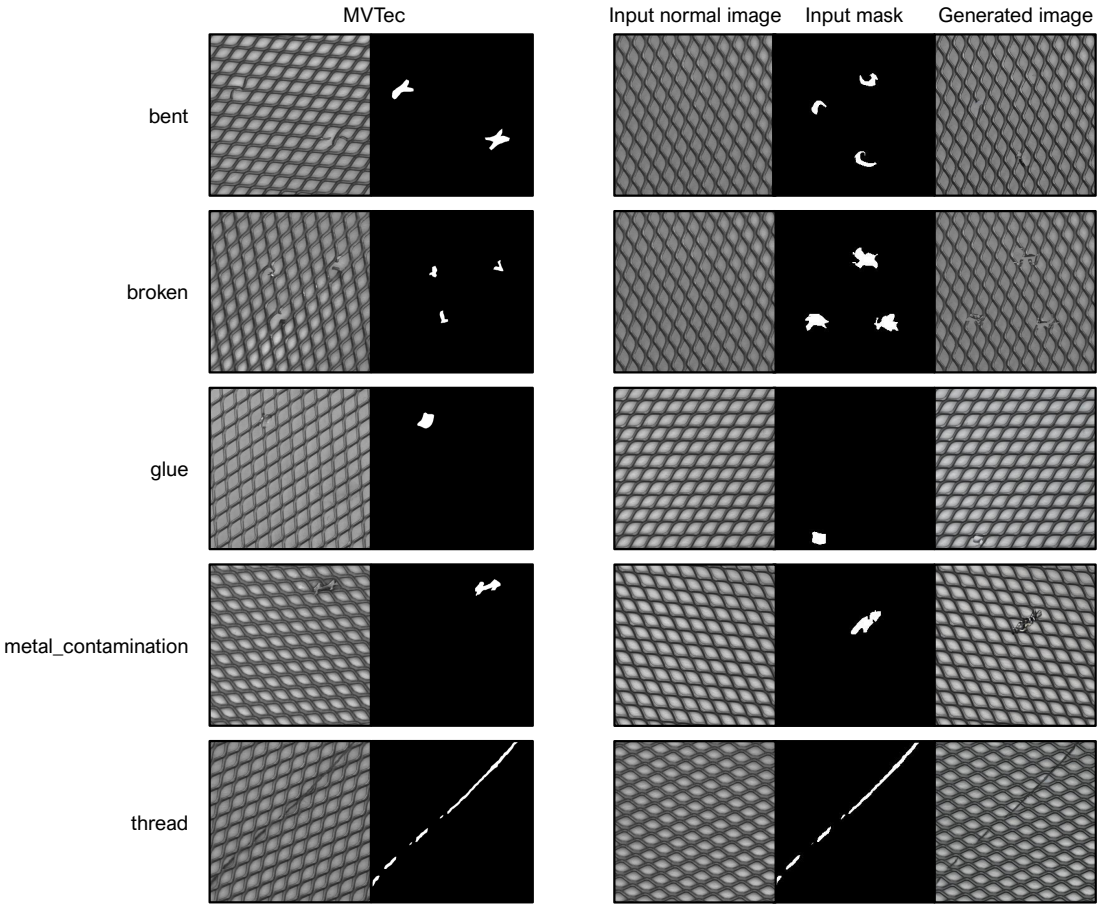

Figure 7. Generated images on *grid*

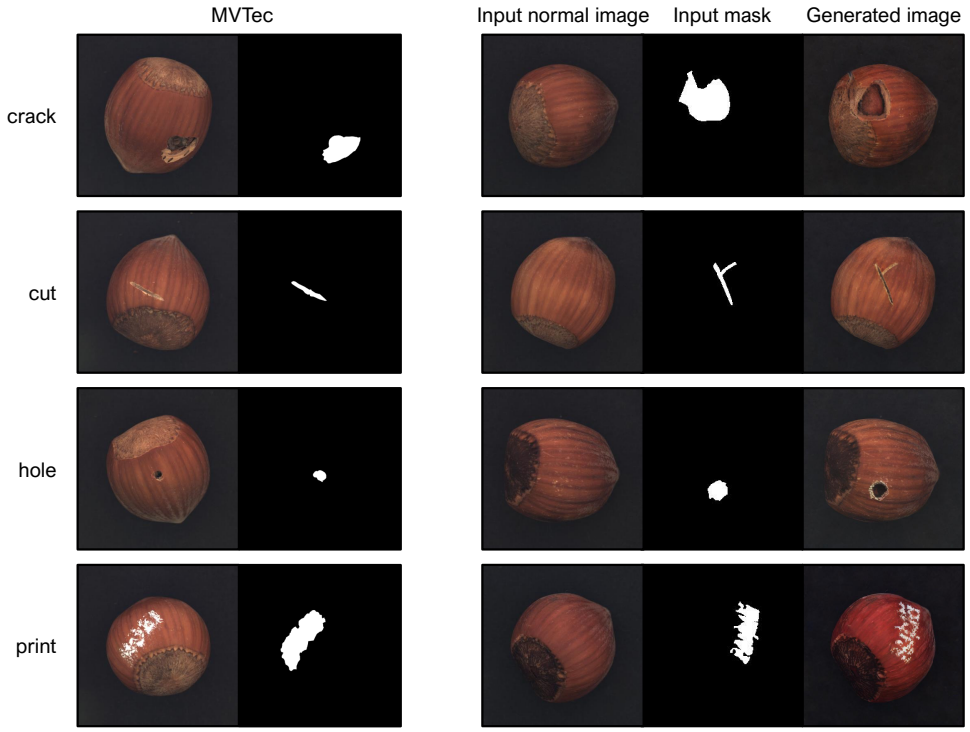

Figure 8. Generated images on *hazelnut*

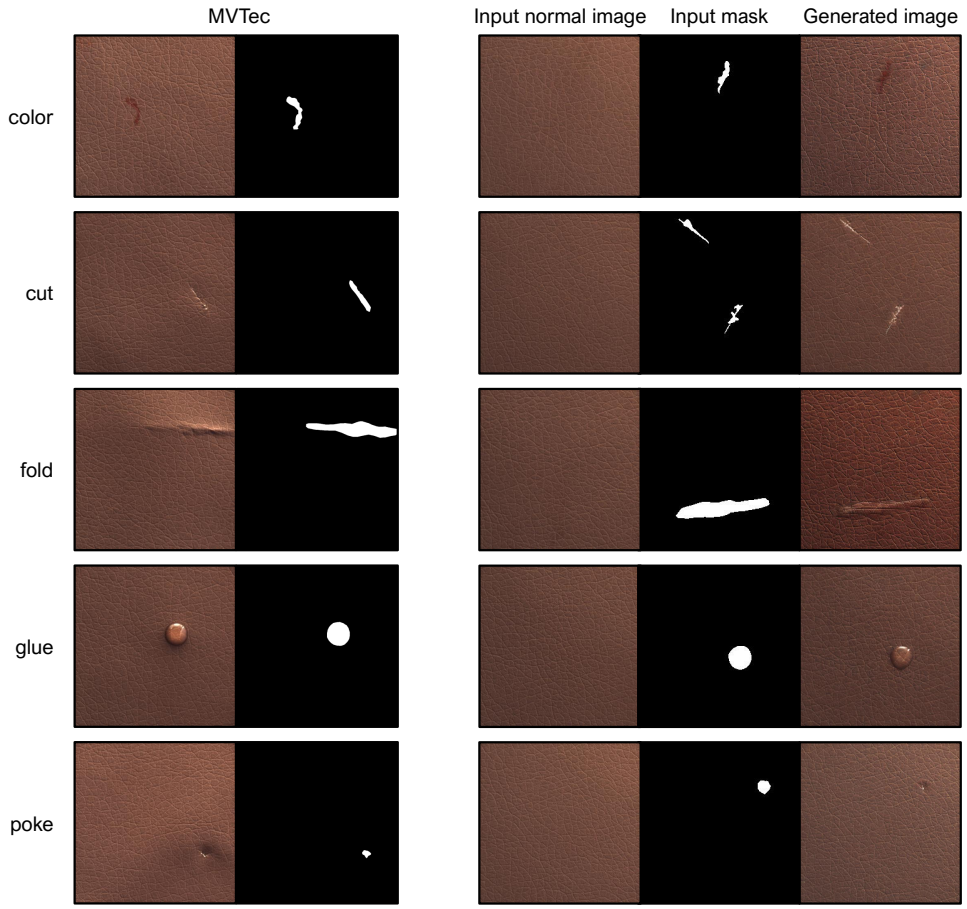

Figure 9. Generated images on *leather*

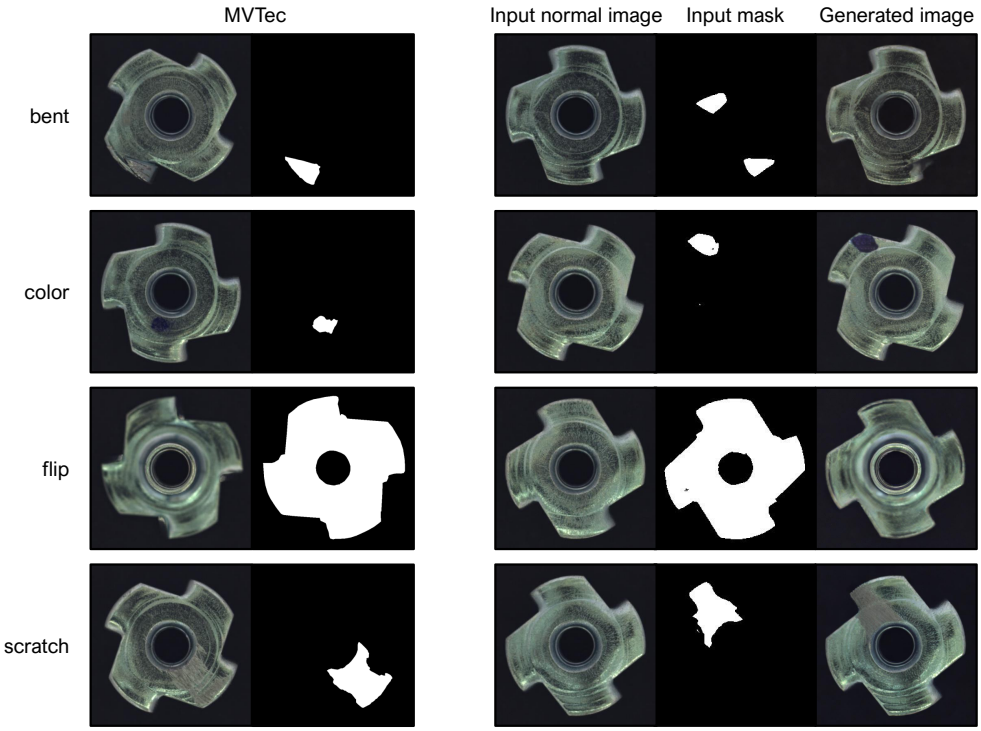

Figure 10. Generated images on *metal_nut*

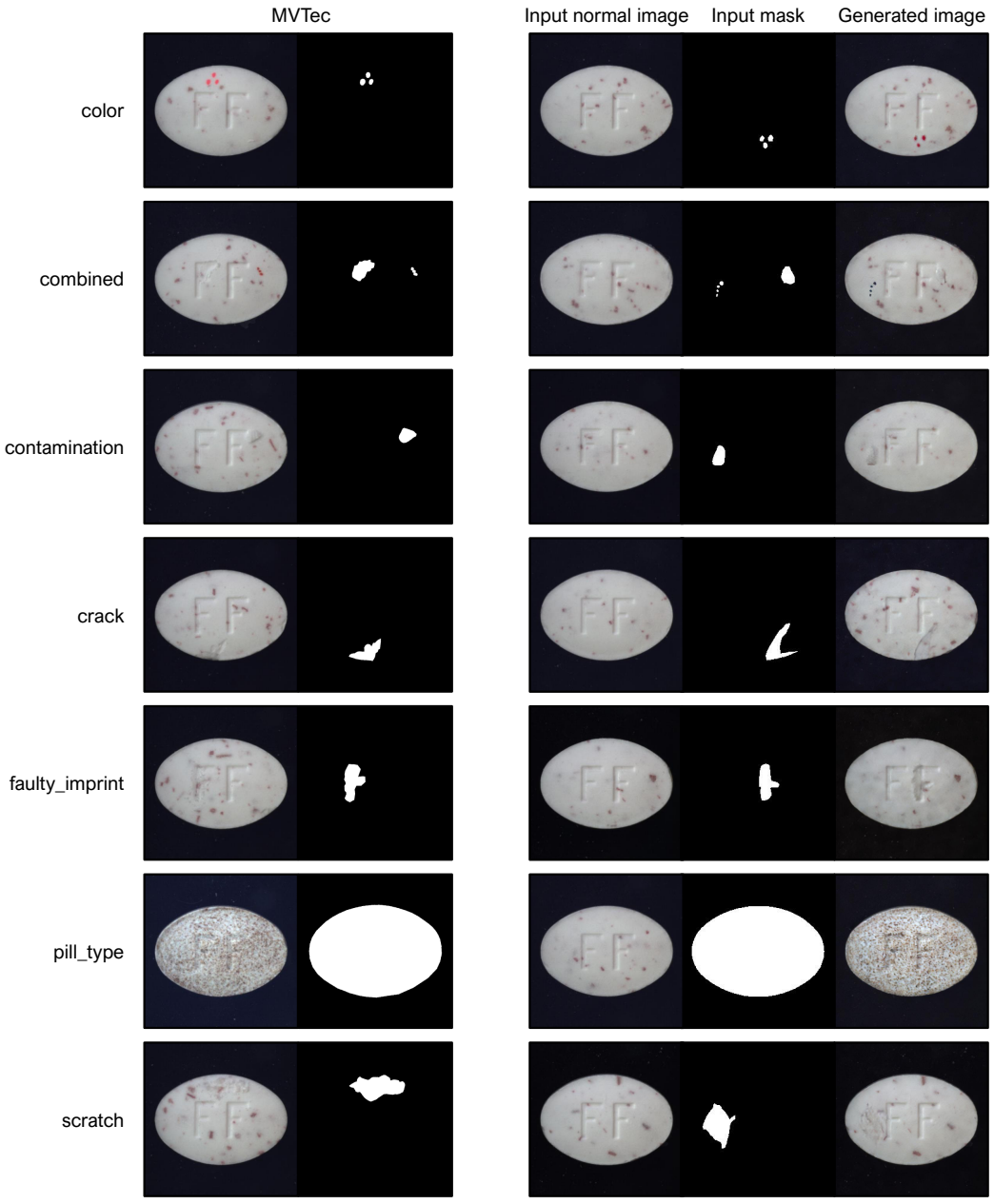

Figure 11. Generated images on *pill*

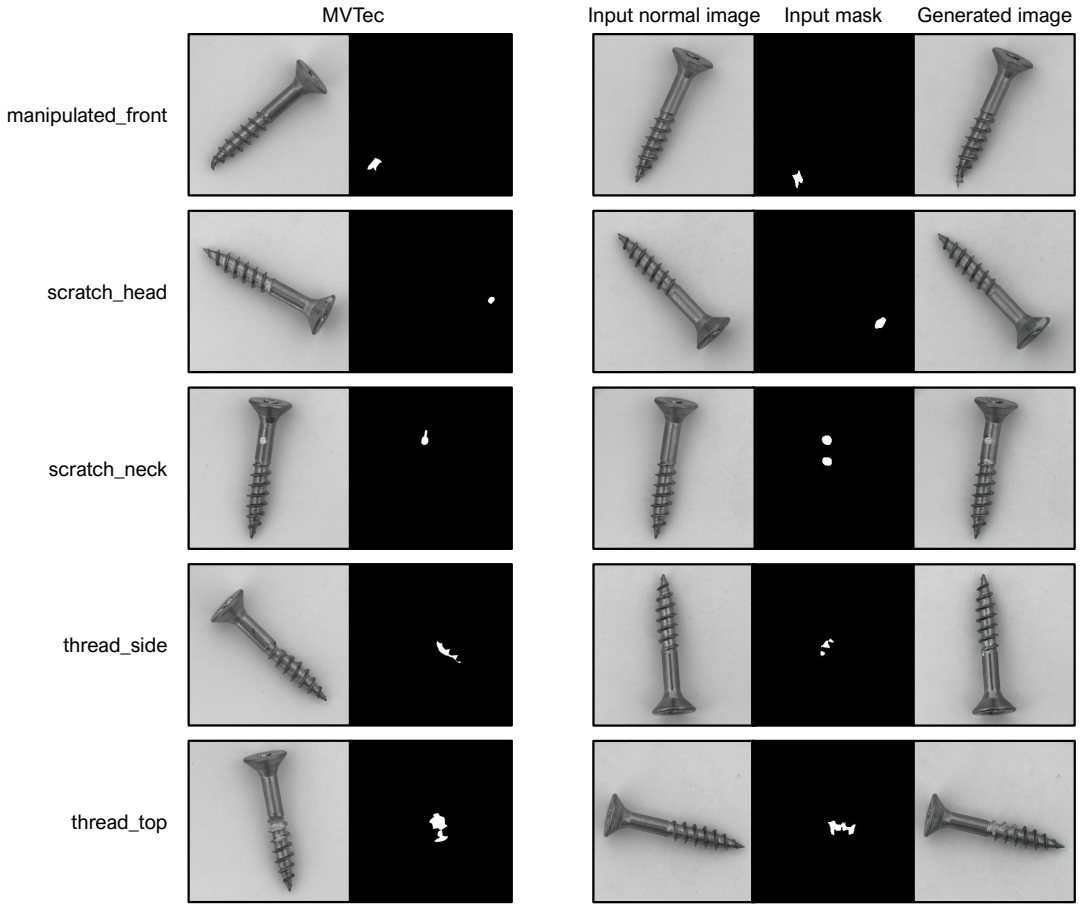

Figure 12. Generated images on *screw*

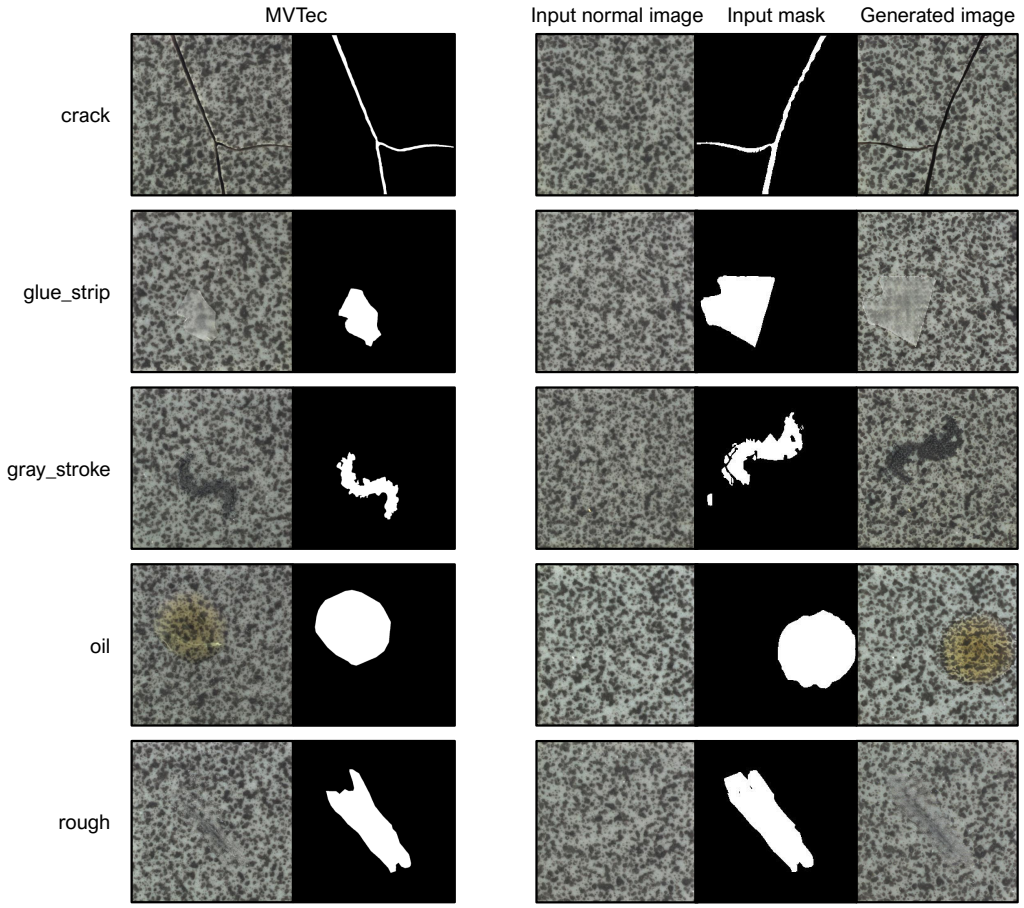

Figure 13. Generated images on *tile*

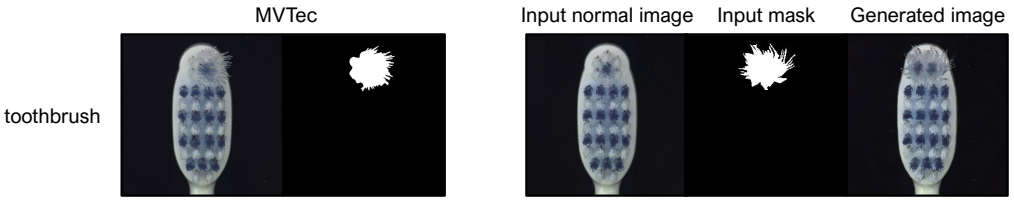

Figure 14. Generated images on *toothbrush*

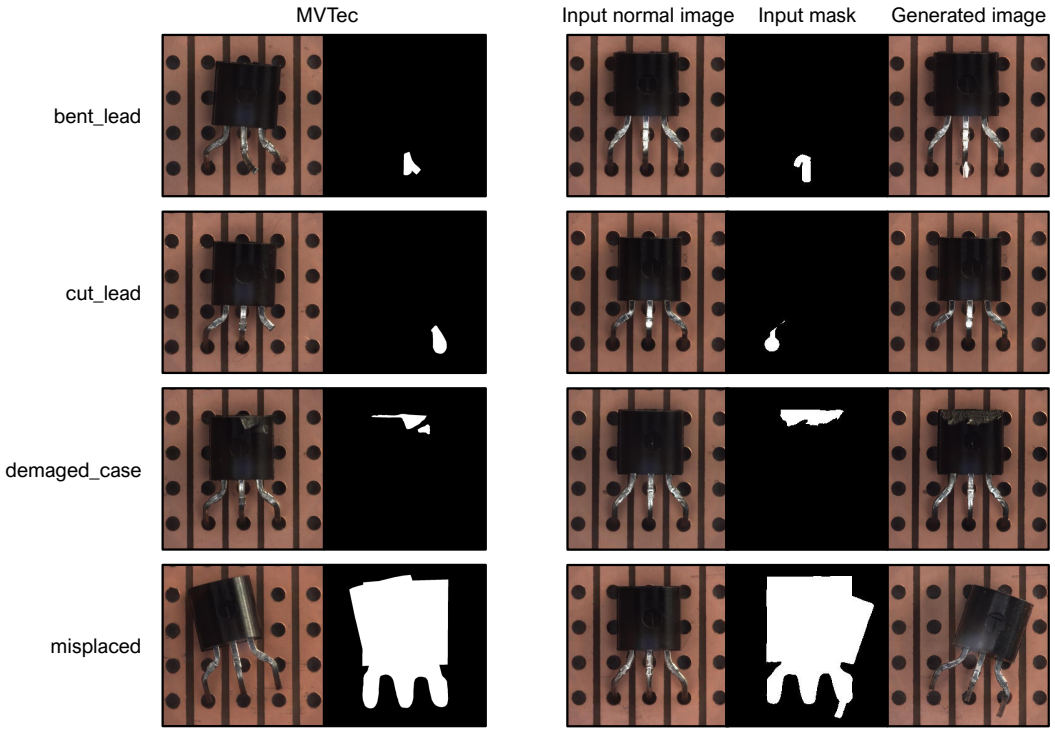

Figure 15. Generated images on *transistor*

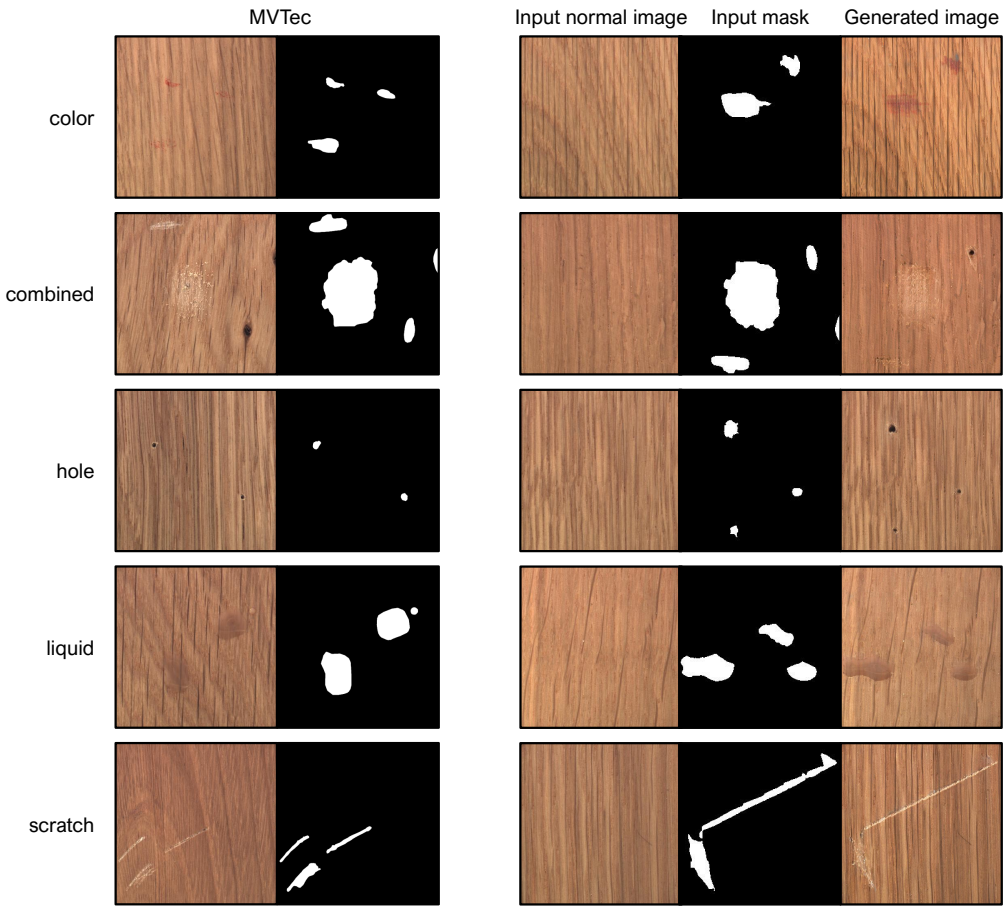

Figure 16. Generated images on *wood*

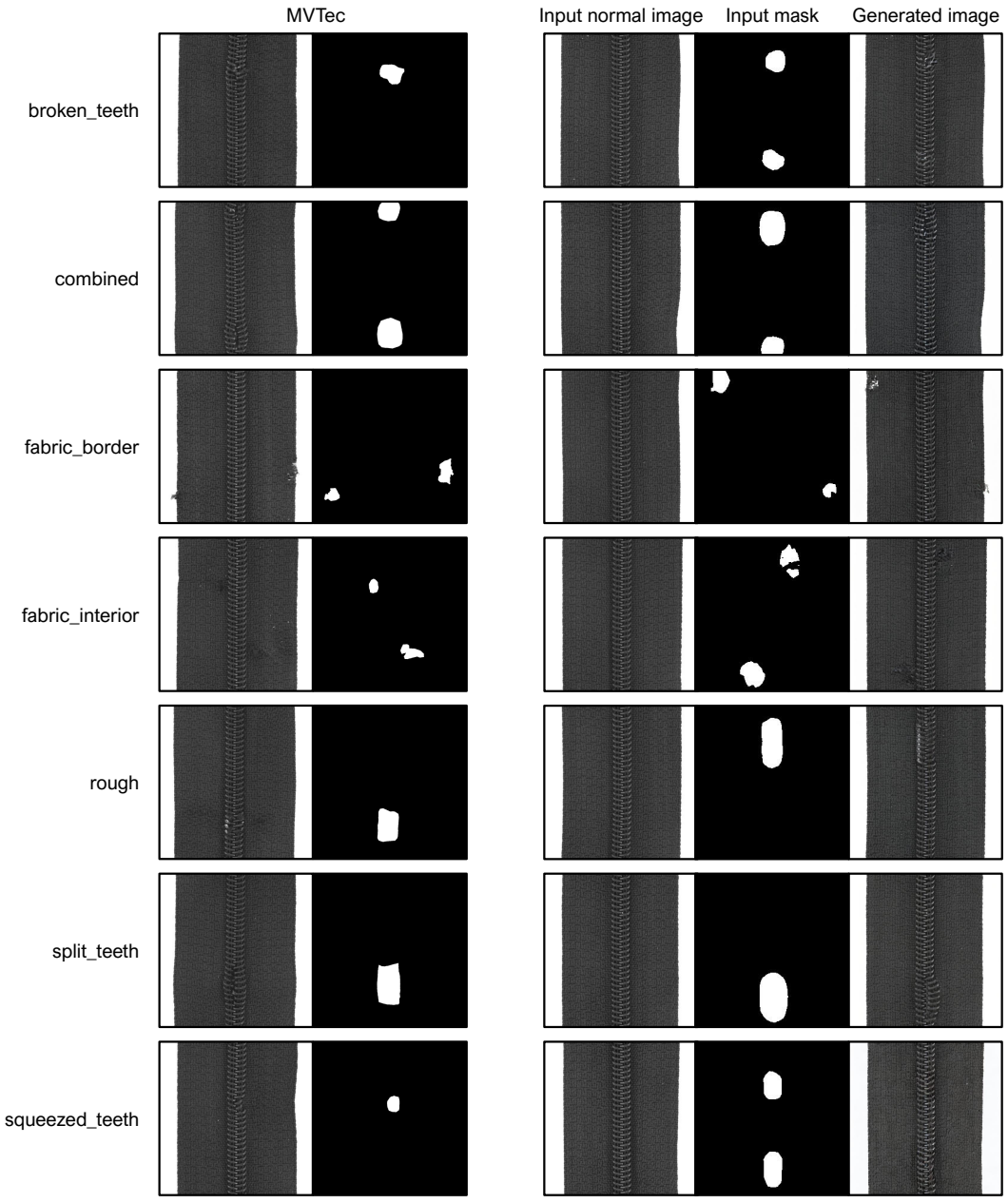

Figure 17. Generated images on *zipper*

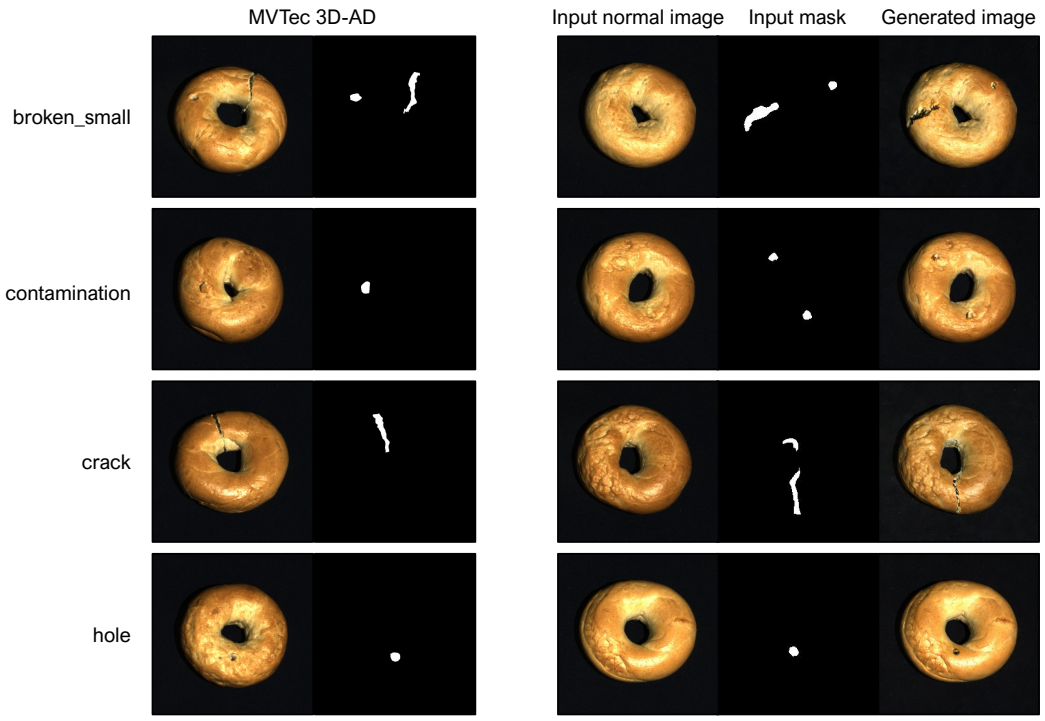

Figure 18. Generated images on *bagel*

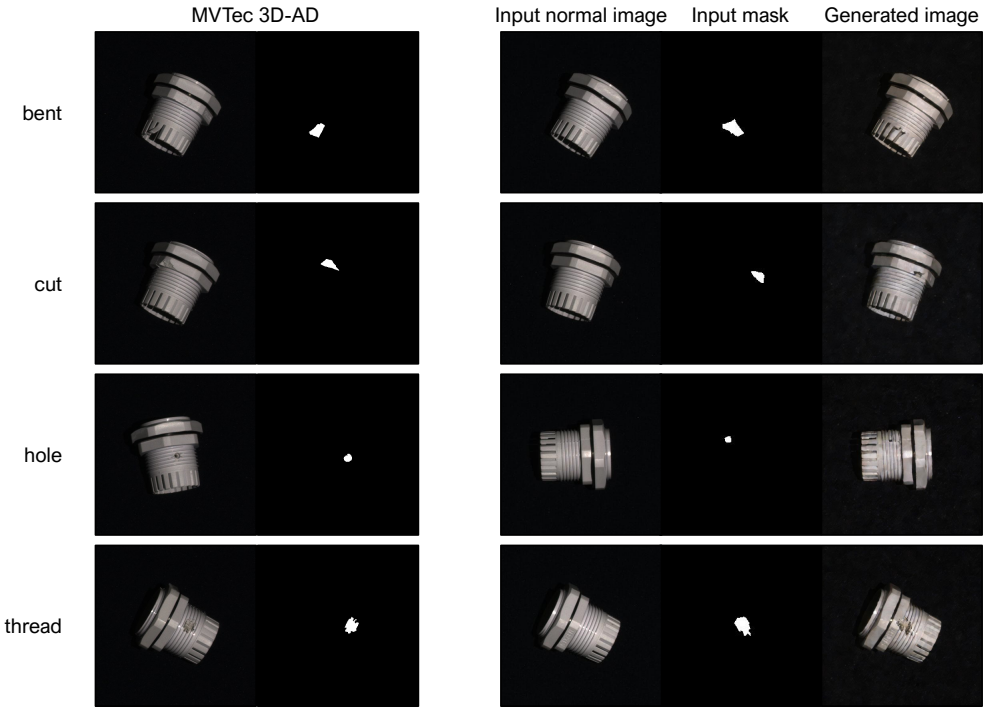

Figure 19. Generated images on *cable_gland*

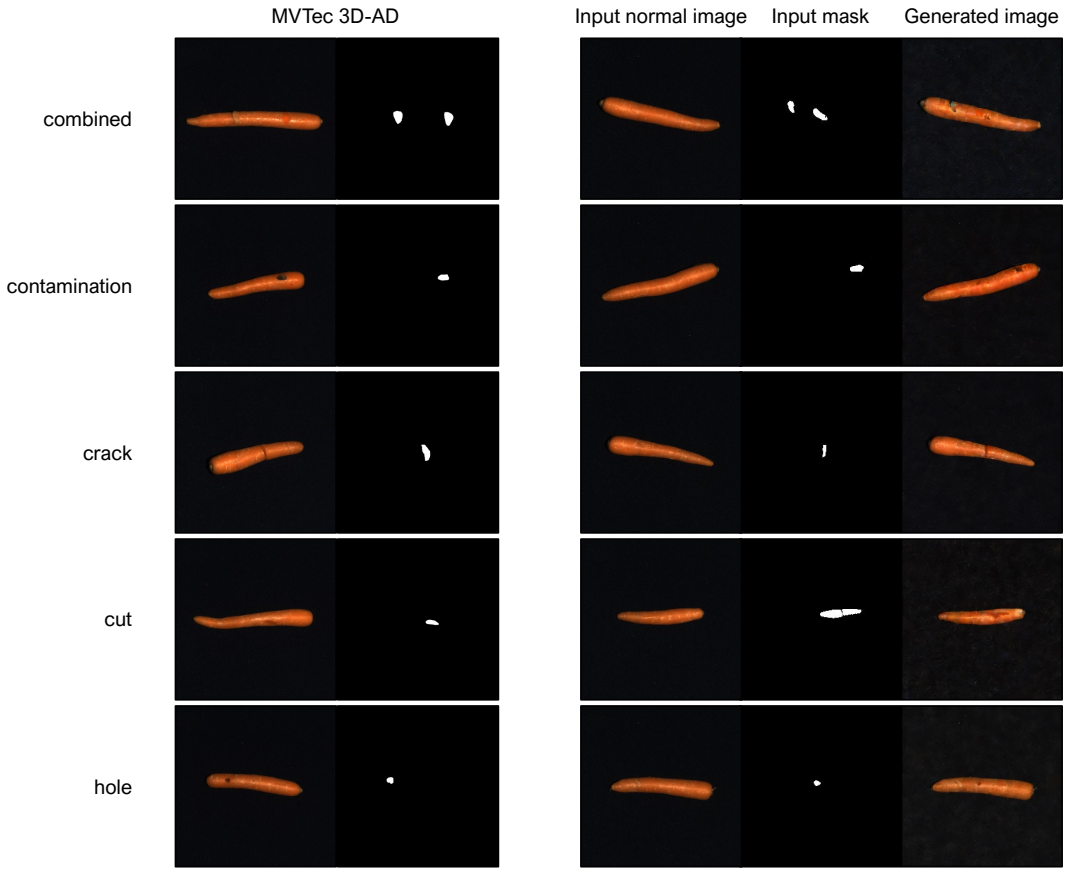

Figure 20. Generated images on *carrot*

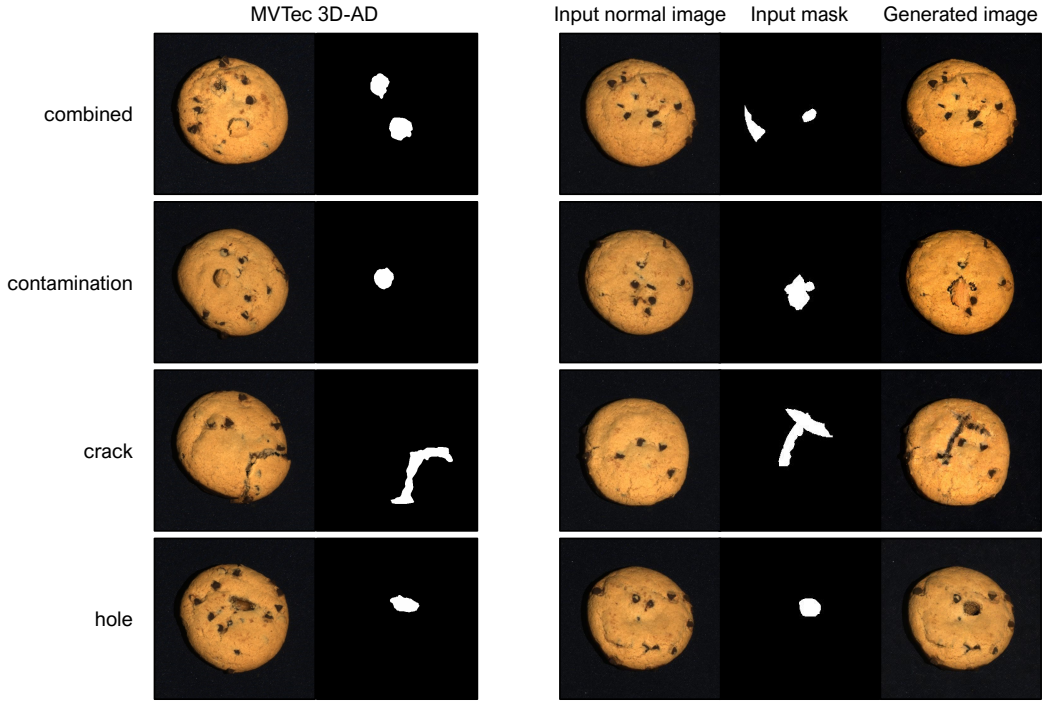

Figure 21. Generated images on *cookie*

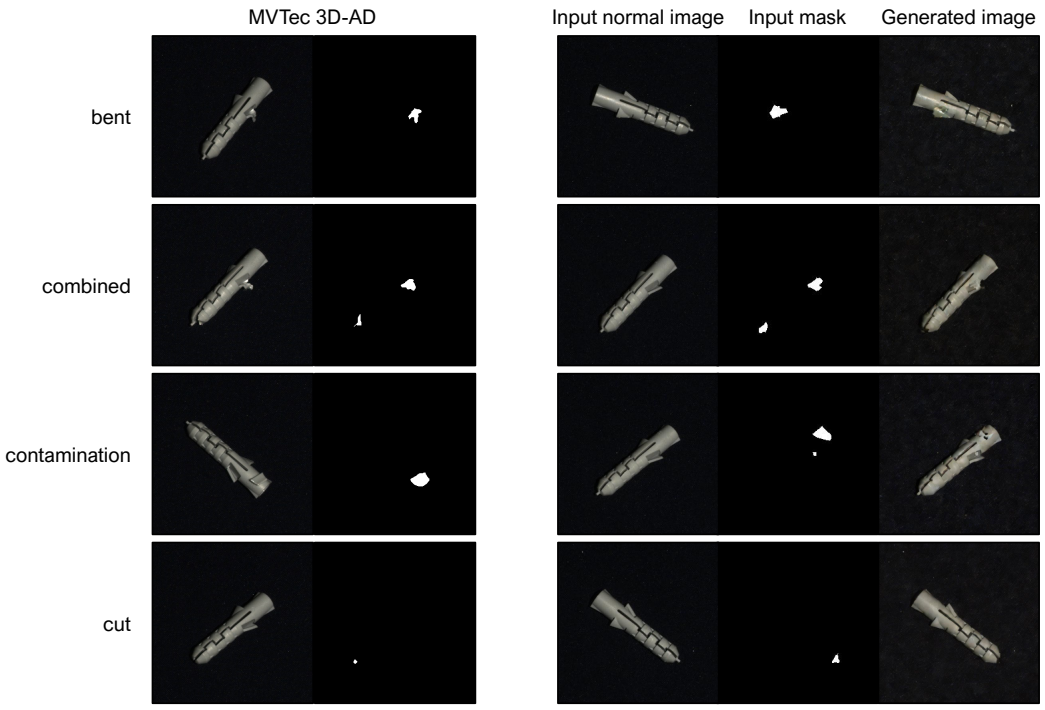

Figure 22. Generated images on *dowel*

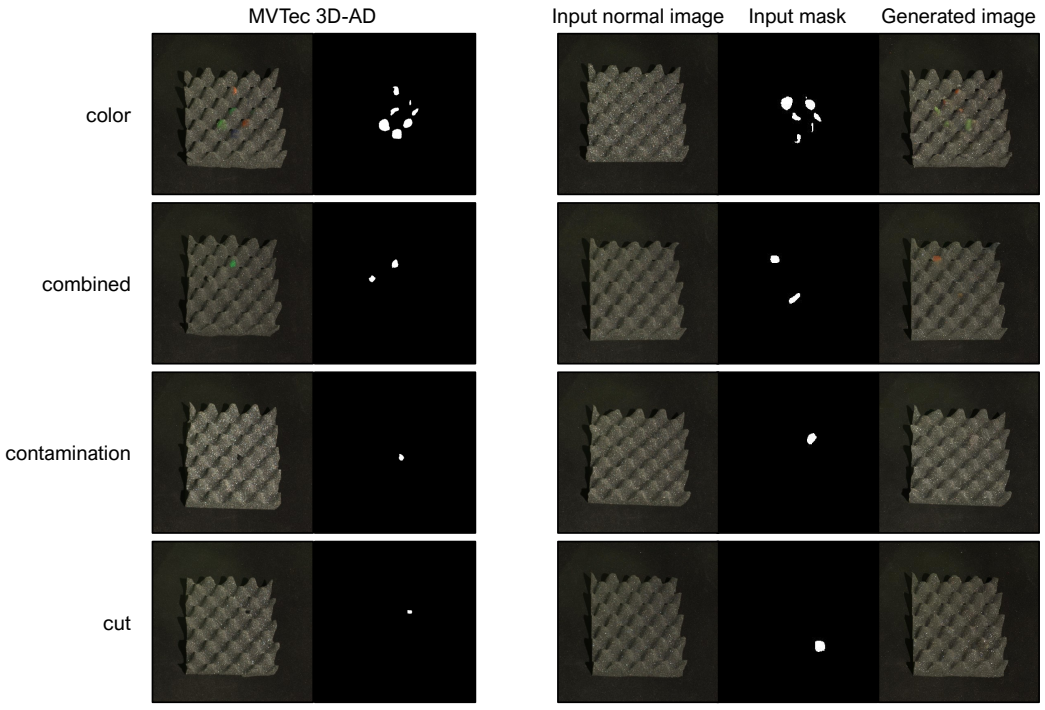

Figure 23. Generated images on *foam*

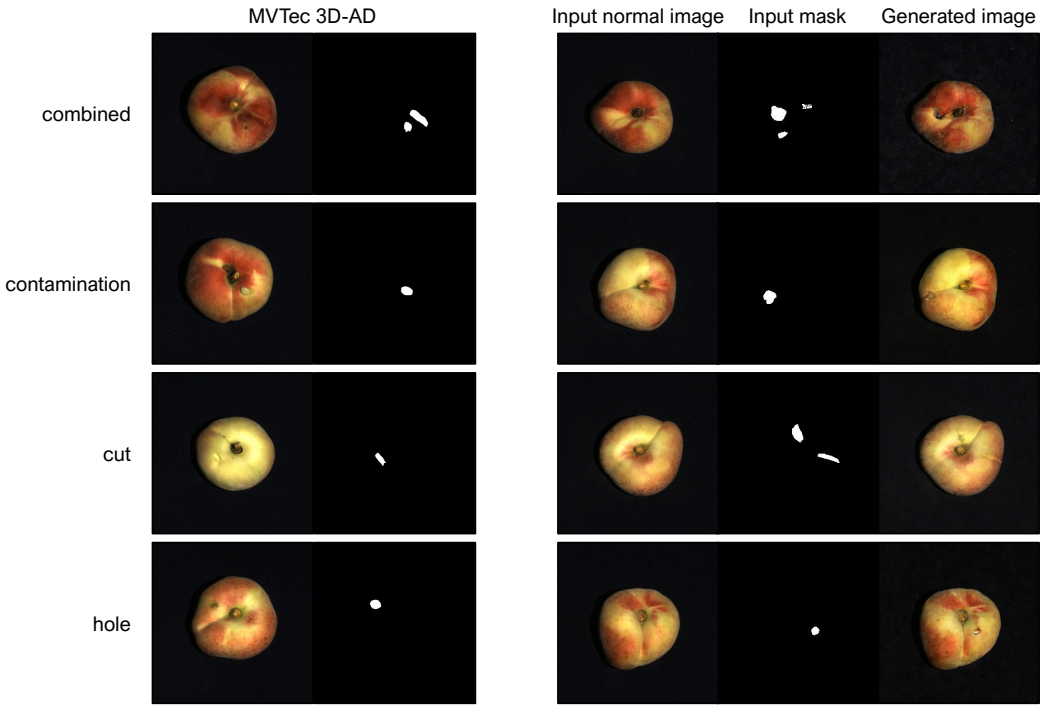

Figure 24. Generated images on *peach*

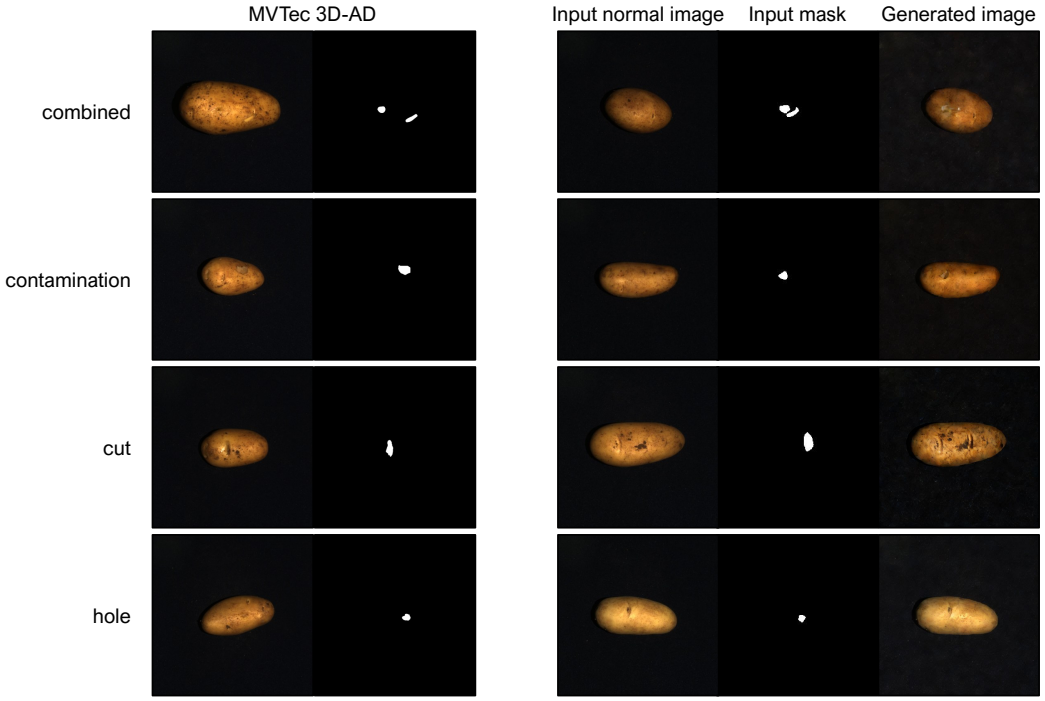

Figure 25. Generated images on *potato*

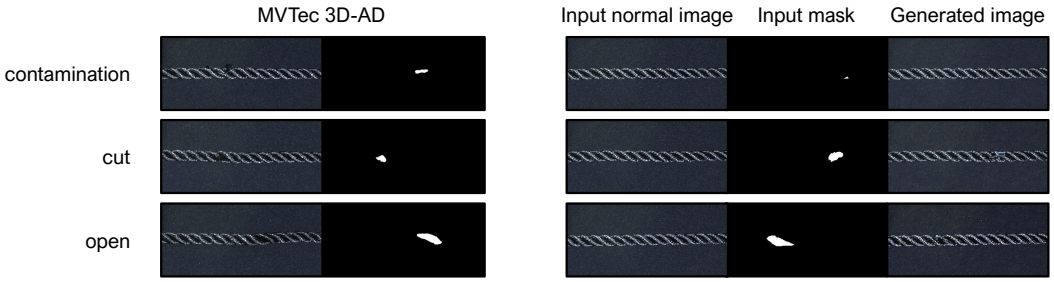

Figure 26. Generated images on *rope*

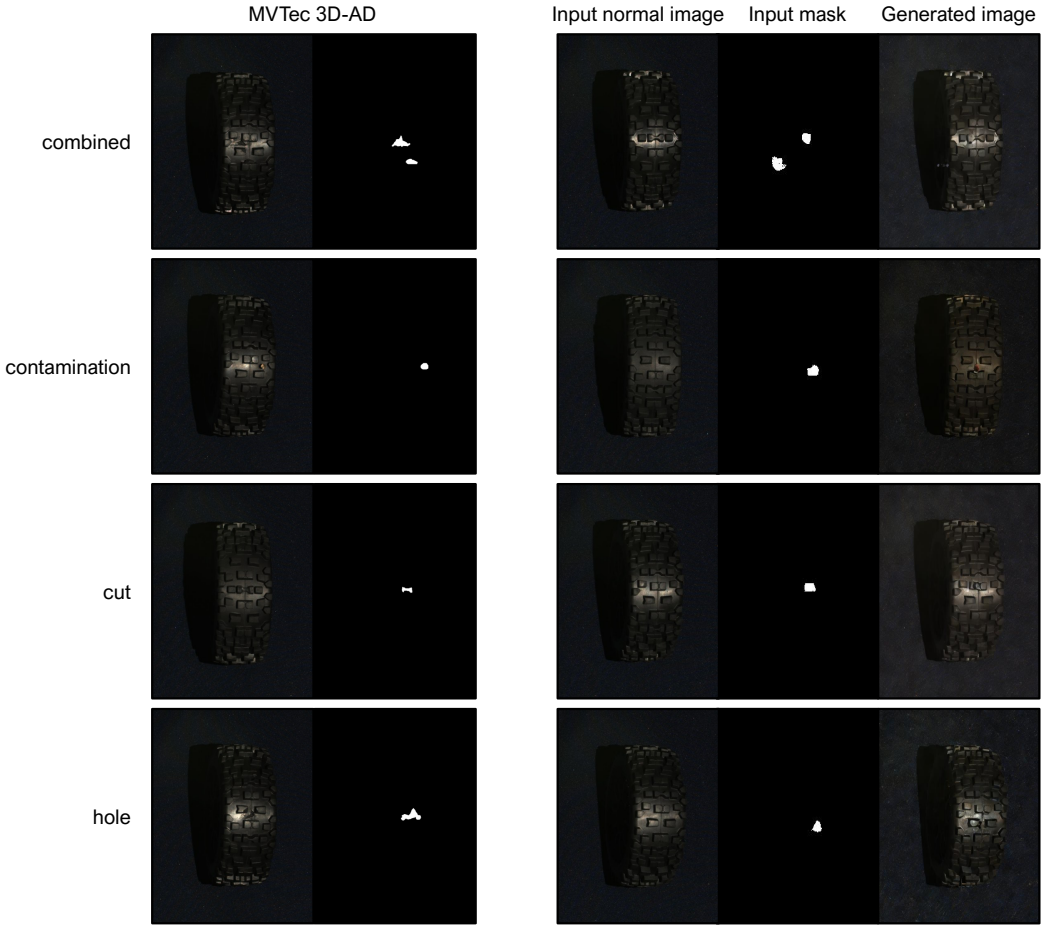

Figure 27. Generated images on *tire*

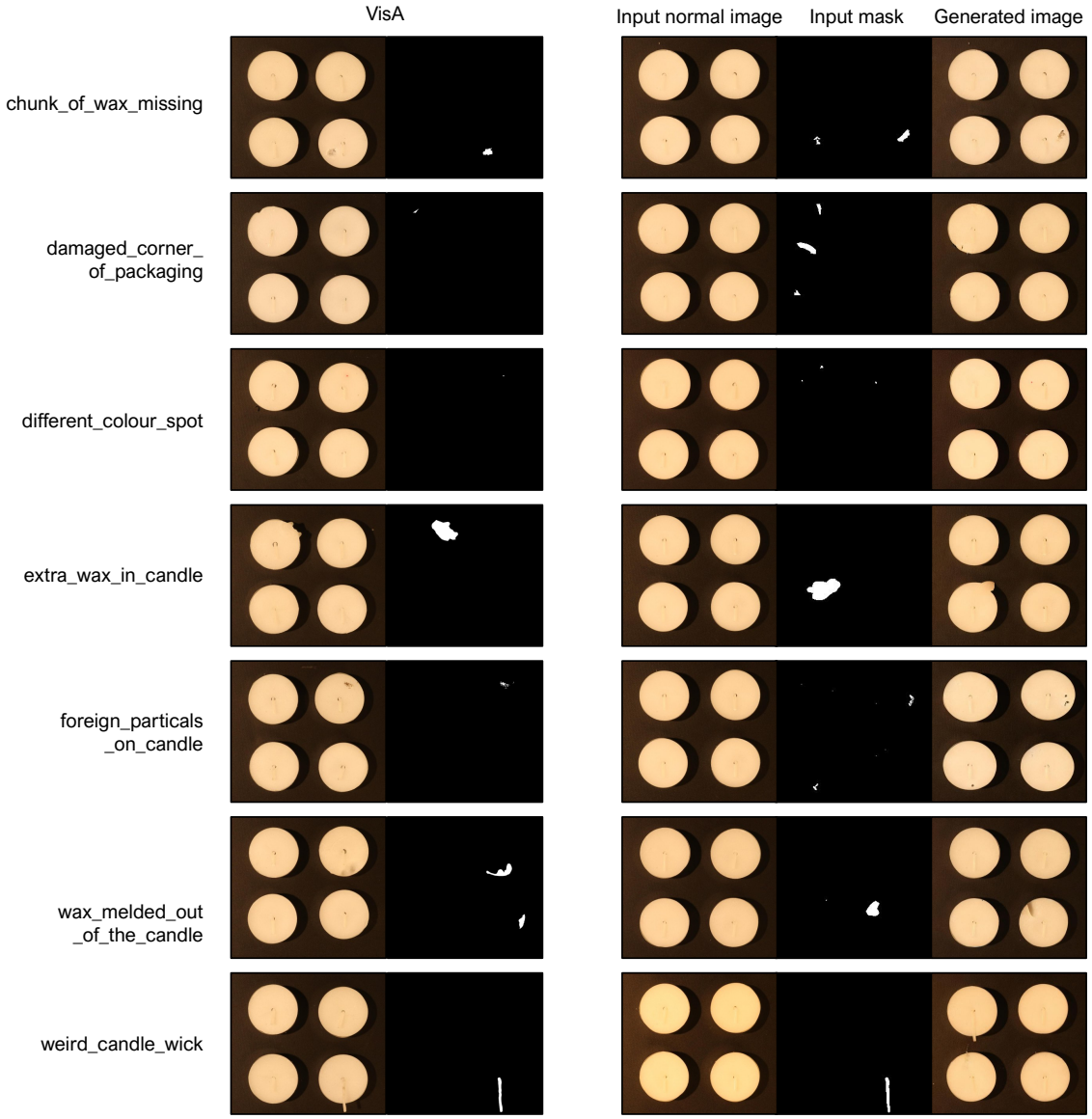

Figure 28. Generated images on *candle*

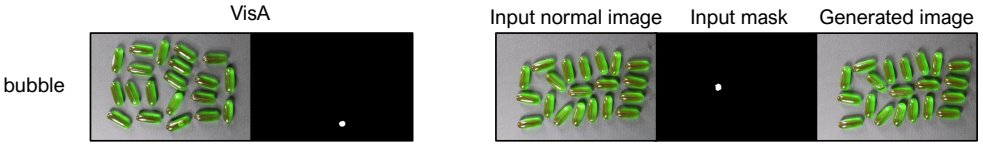

Figure 29. Generated images on *capsule*

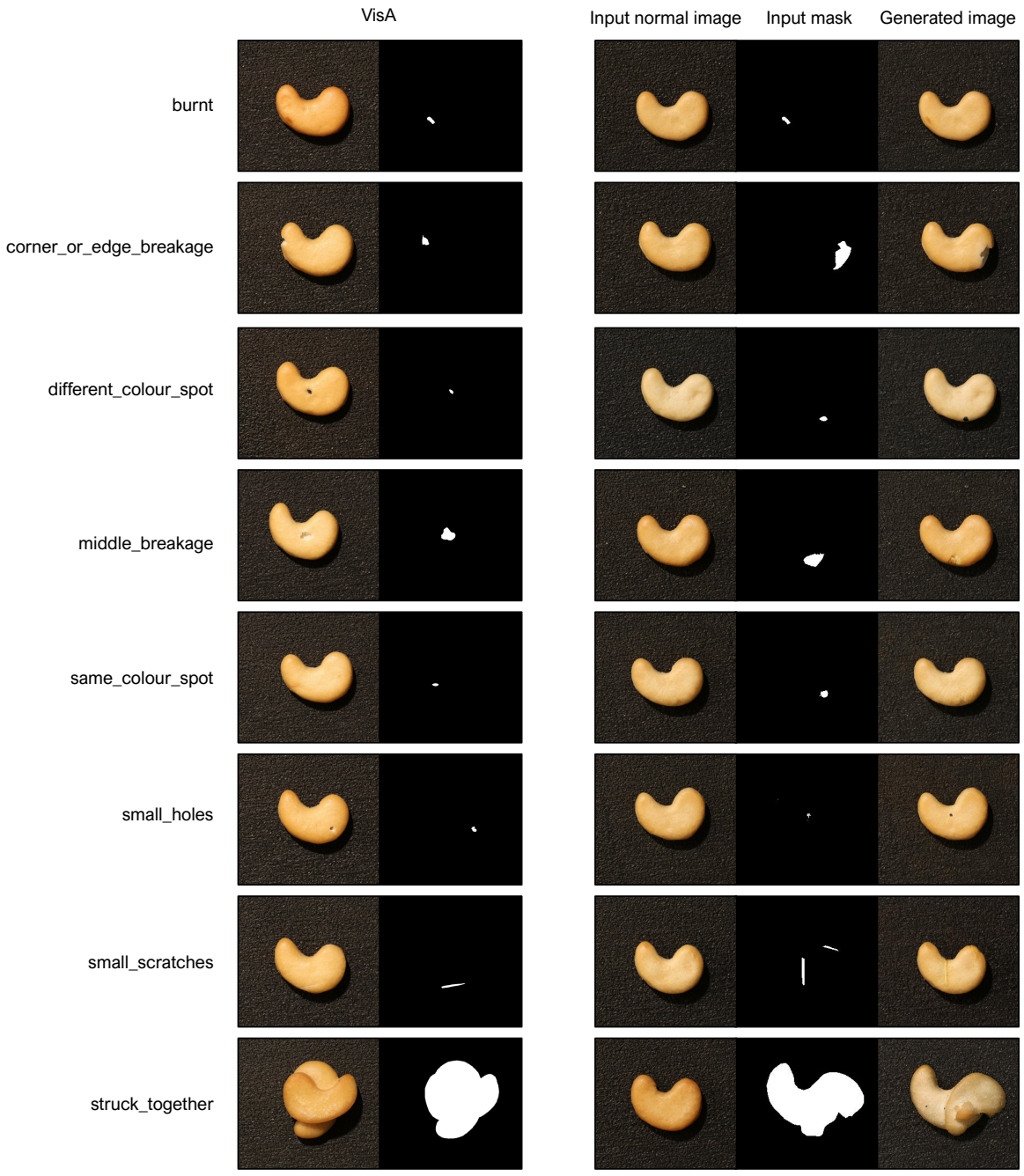

Figure 30. Generated images on *cashew*

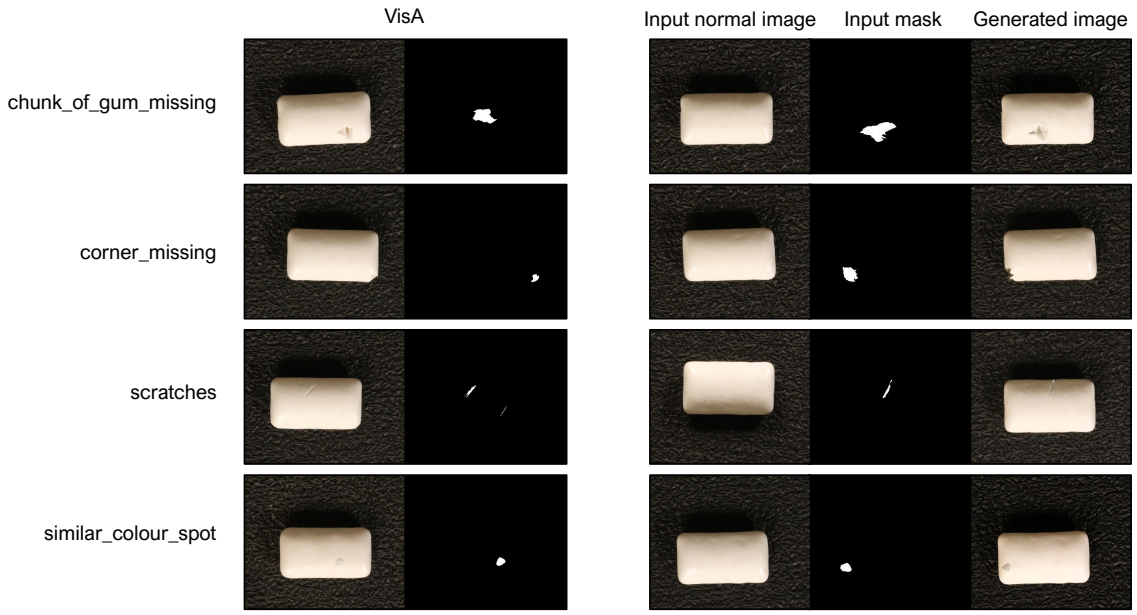

Figure 31. Generated images on *chewinggum*

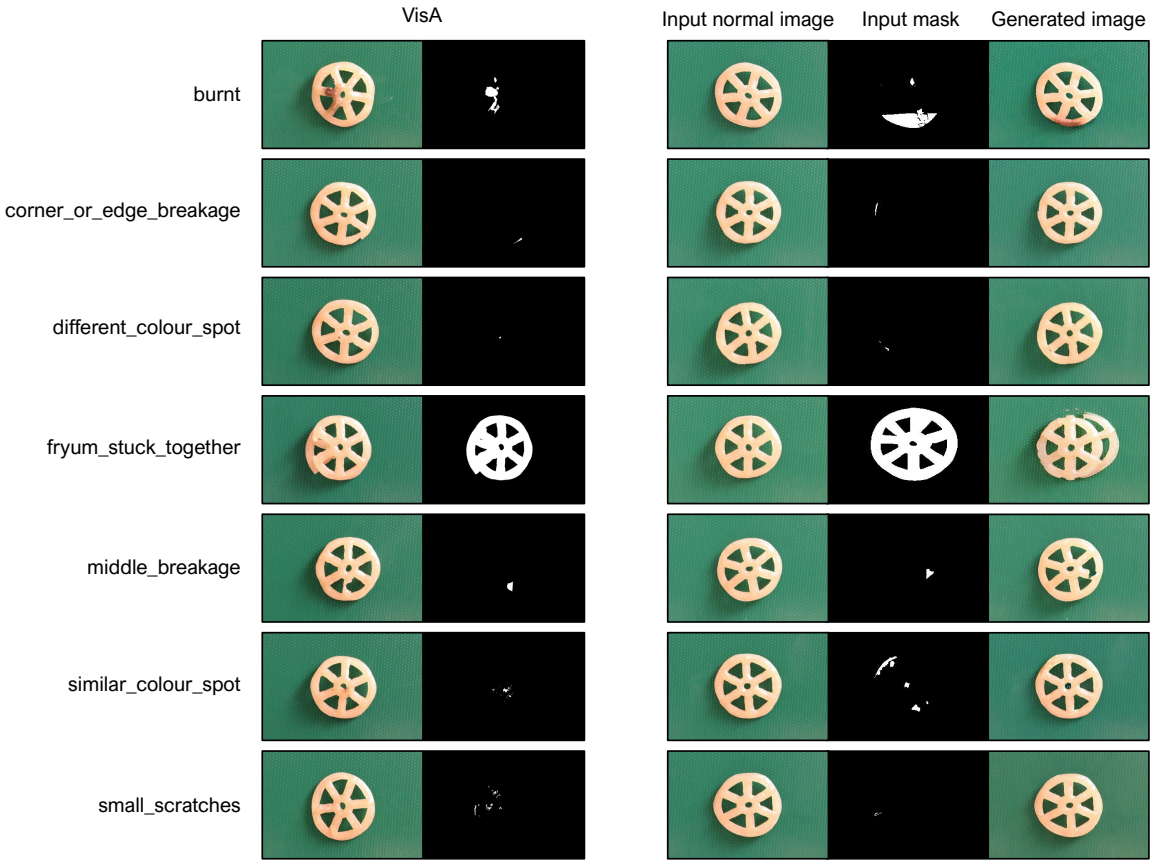

Figure 32. Generated images on *fryum*

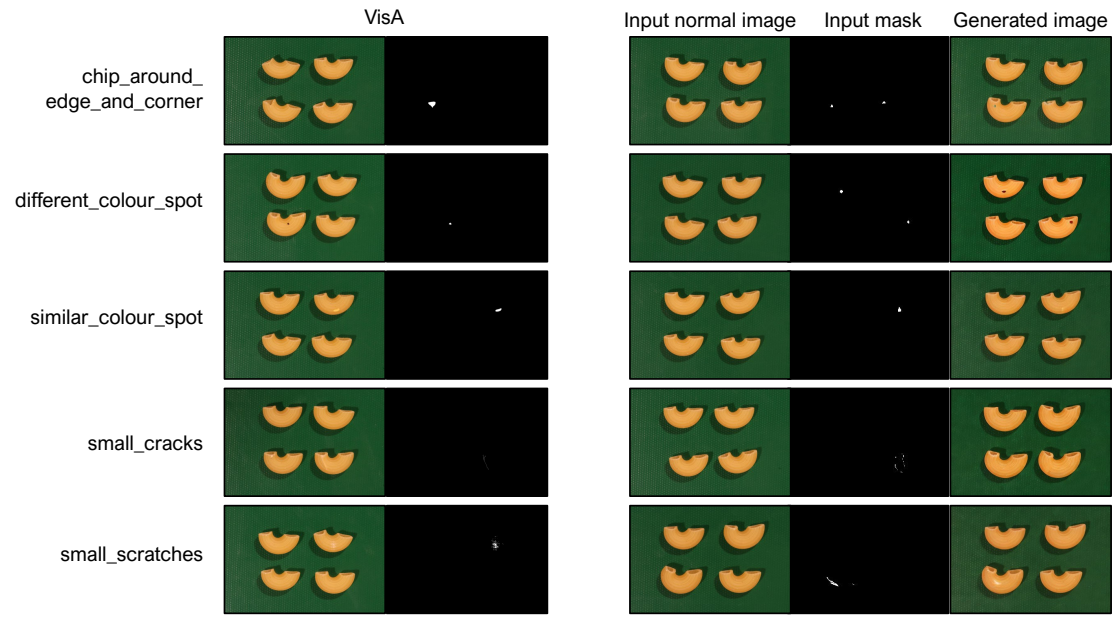

Figure 33. Generated images on *macaroni1*

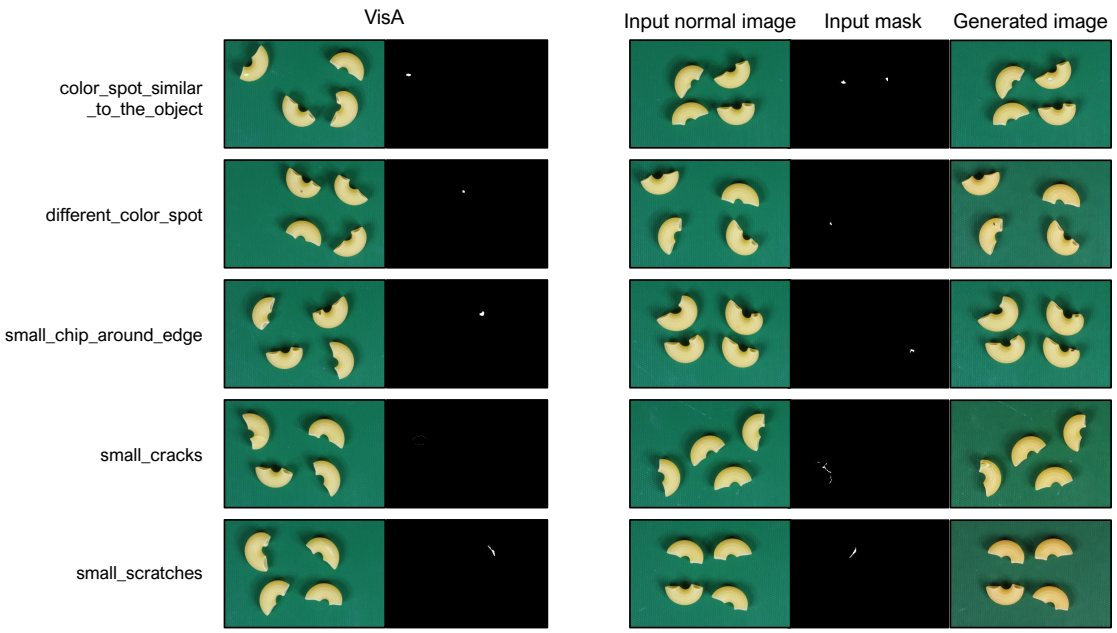

Figure 34. Generated images on *macaroni2*

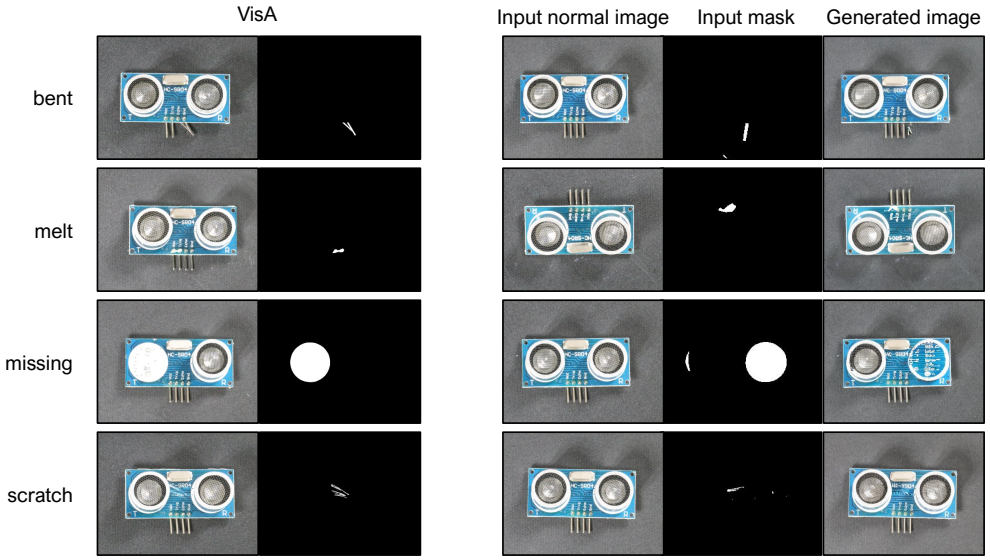

Figure 35. Generated images on *pcb1*

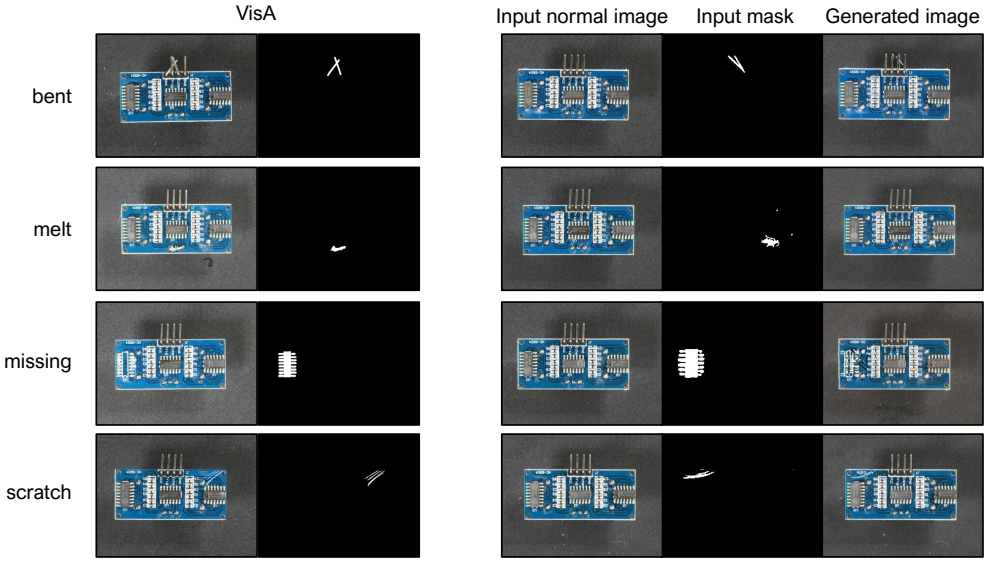

Figure 36. Generated images on *pcb2*

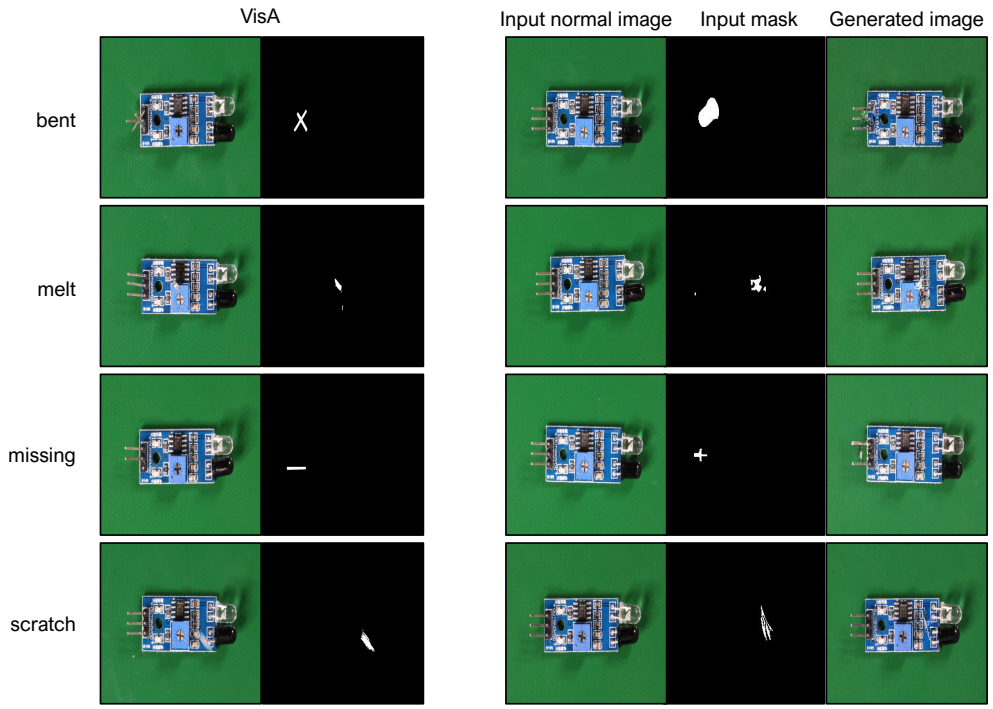

Figure 37. Generated images on *pcb3*

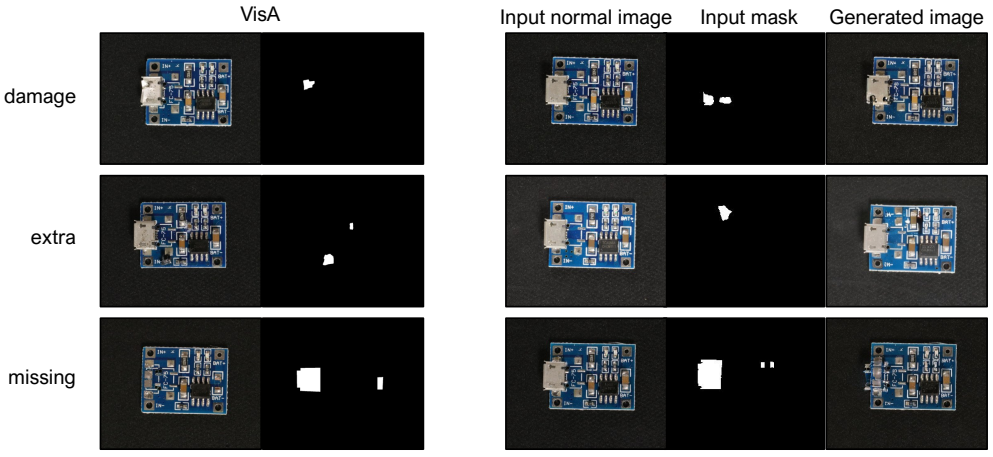

Figure 38. Generated images on *pcb4*

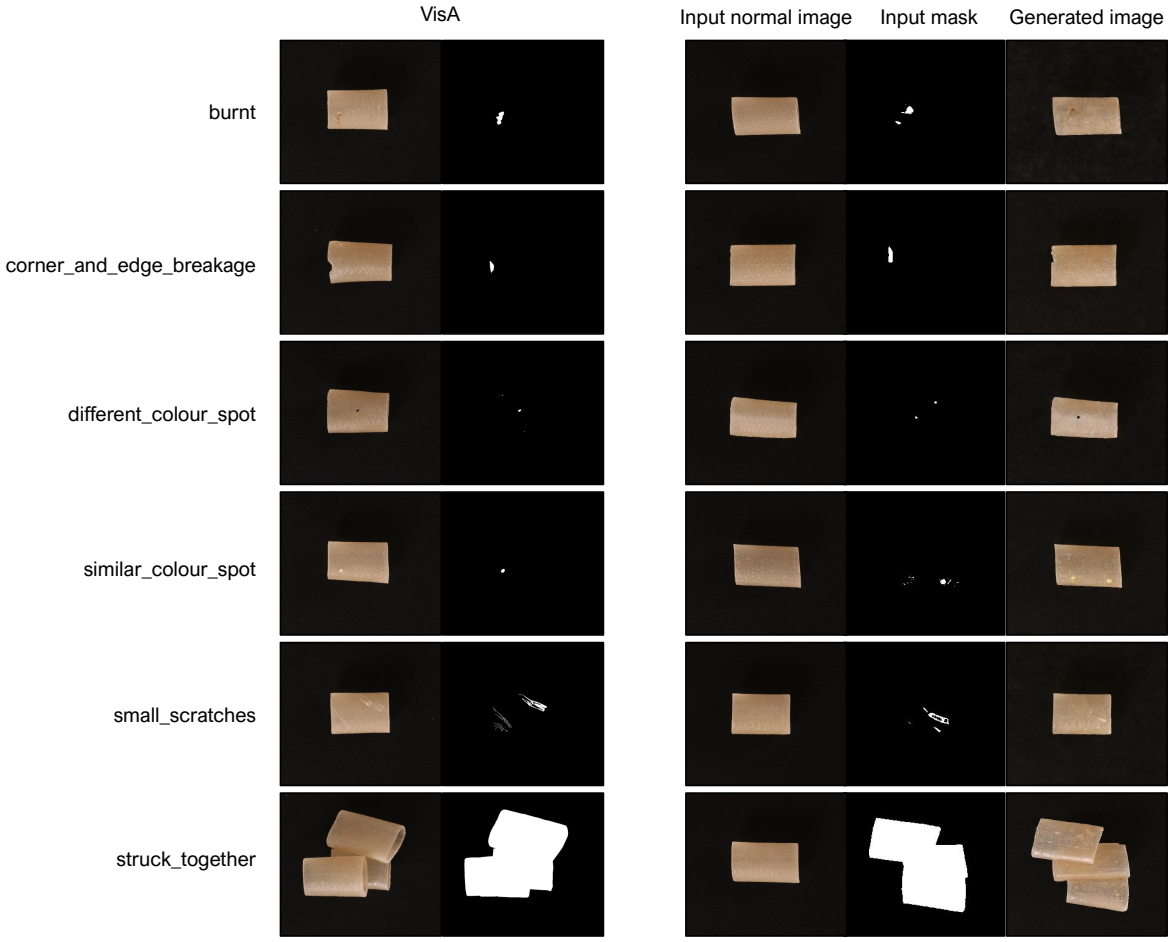

Figure 39. Generated images on *pipe_fryum*

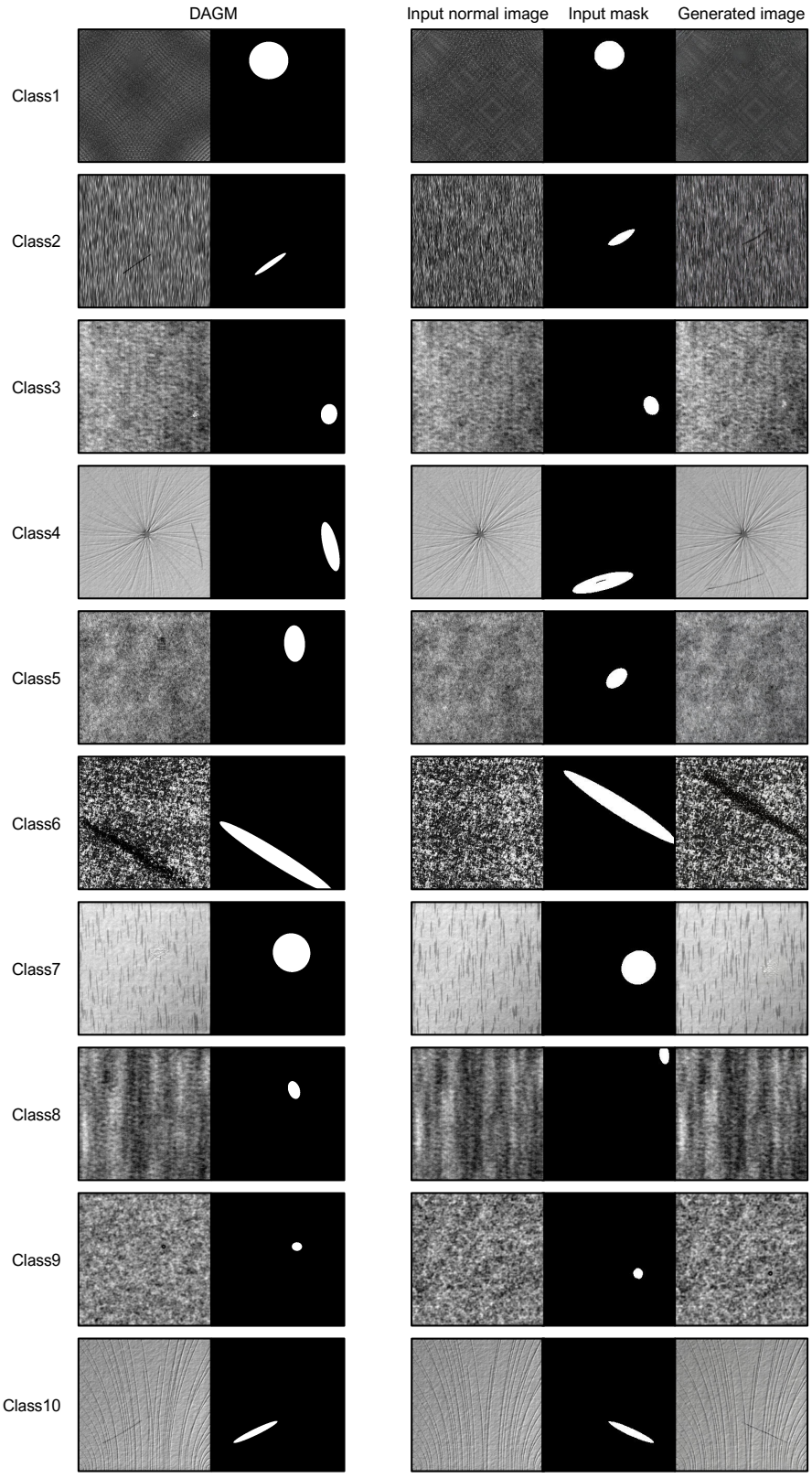

Figure 40. Generated images on every classes of *DAGM*

References

[1] JaeHyuck Choi, Minjun Kim, and Je Hyeong Hong. MAGIC: Few-shot mask-guided anomaly inpainting with prompt perturbation, spatially adaptive guidance, and context awareness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026. 1, 3, 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 4

[3] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 1, 2, 3

[4] Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, and Yabiao Wang. Dual-interrelated diffusion model for few-shot anomaly image generation. *arXiv preprint arXiv:2408.13509*, 2024. 2

[5] Tobias Lingenberg, Markus Reuter, Gopika Sudhakaran, Dominik Gojny, Stefan Roth, and Simone Schaub-Meyer. Diagen: Diverse image augmentation with generative models. *arXiv preprint arXiv:2408.14584*, 2024. 5

[6] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *ArXiv*, abs/2310.17347, 2023. 5

[7] Jaewoo Song, Daemin Park, Kanghyun Baek, Sangyub Lee, Jooyoung Choi, Eunji Kim, and Sungroh Yoon. DefectFill: Realistic defect generation with inpainting diffusion model for visual inspection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18718–18727, 2025. 4

[8] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1