Template-guided Hierarchical Feature Restoration for Anomaly Detection (Supplementary Material)

Hewei Guo[1*‡] Liping Ren[2*‡] Jingjing Fu[3†] Yuwang Wang[2†] Zhizheng Zhang[3]
Cuiling Lan[3] Haoqian Wang[2] Xinwen Hou[1]
[1]Institute of Automation, Chinese Academy of Sciences
[2]Tsinghua University [3]Microsoft Research Asia
{guohewei2020, xinwen.hou}@ia.ac.cn
{rlp20@mails., wang-yuwang@mail., wanghaoqian@}tsinghua.edu.cn
{jifu, zhizzhang, culan}@microsoft.com

1. Implementation details

1.1. Feature visualization network

We propose a feature restoration paradigm to restore anomaly-free features from the anomalous features, and detect anomalies in terms of the cosine distance between the pre-trained features of an inference image and the corresponding restored anomaly-free features. The restored features tend to be close to the input image feature for the normal regions and depart on anomalies. In order to explicitly demonstrate the effectiveness of our method, we visualize the restored features using a visualization network [10], which is trained on normal and anomaly samples to reconstruct images from the pre-trained embedding features. The visualization network follows a reversed architecture of WideResNet [11], and the down-sampling in the original network is replaced by up-sampling. Targeting high-fidelity visualization, we use MSE loss to train the restoration networks.

2. MVTec LOCO AD dataset

MVTec LOCO AD dataset [1] is proposed to cover representative examples of structural and logical anomalies in industrial inspection scenarios. Structural anomalies appear as scratches, dents, or contamination in manufactured products, while logical anomalies are defined by violating underlying logical constraints, e.g. an allowed object appears in an invalid location or a required object does not exist at all. The dataset consists of five sub-datasets, including breakfast box, pushpins, splice connectors, screw bag, and juice bottle.

All anomalies in the dataset are categorized as structural or logical anomalies, which enable independent evaluation on the anomaly detection performance of each type. Previous works predominantly focus on the detection of structural anomalies, while the proposed THFR network is developed for the detection of structural and logical anomalies.

3. More ablation results

3.1. Compression level

In our experiments on LOCO, we investigated the effect of varying channel size of the bottleneck, which corresponds to the compression level of the bottleneck. As shown in Figure 1, we observe that the localization accuracy increases with the channel size during the first half, indicating that a higher preservation of information could lead to better restoration quality after passing through the bottleneck. However, beyond a certain point, increasing the channel size no longer improves performance. When the bottleneck is too loose, it is unable to effectively filtered out anomaly features by compression, leading to decrease in restoration quality and accuracy. Based on the observation, we set channel size to 2048 in our bottleneck design, to find a balance between preserving sufficient information while still accomplishing compression.

3.2. Template bank subsampling

Since the template feature is retrieved using image-level nearest neighbor search, reducing template bank size by subsampling will inevitably lower the feature similarity between the inference image and the selected template. To investigate the impact of template bank subsampling on template retrieval and anomaly detection performance, we evaluate the anomaly localization accuracy and average fea-

*Equal contribution.
†Corresponding author.
‡Work done during an internship at Microsoft Research Asia.

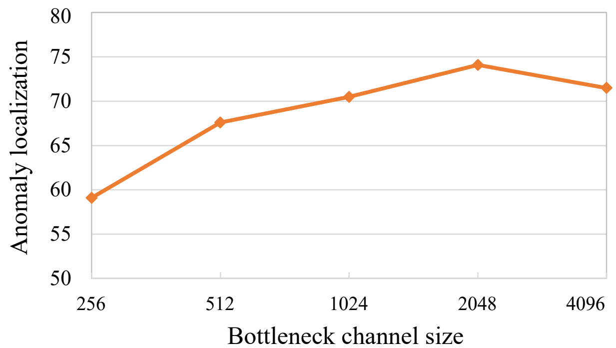

Figure 1. Anomaly localization accuracy of bottleneck compression with different channel sizes. We examined the impact of modifying the channel size of the bottleneck, which correlates to the compression level of the bottleneck.

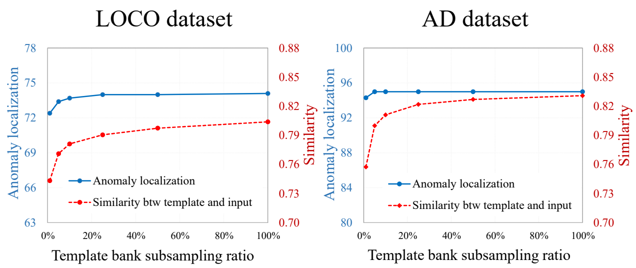

Figure 2. Anomaly localization accuracy and average feature similarity vary with different template subsampling ratios. Our method is robust to variations in feature similarity between the input feature and the template feature.

ture similarity between the inference image and the corresponding template at different subsampling ratios on two datasets. The results are shown in figure 2. The average feature similarity of LOCO dataset is lower than that of AD dataset, where samples are better aligned. As a result, previous works have demonstrated good performance on the AD dataset but unsatisfactory performance on the LOCO dataset. We leverage the relationship between the input and the template to guide the restoration process, which makes our method less sensitive to the feature similarity variation between the input and the template. When we decrease the number of templates, the average similarity also decreased. However, the performance remained consistent until subsampling ratio drops below 5%, and even so, the drop in performance is very limited.

3.3. Restored feature visualization

To explicitly demonstrate the effectiveness of our network design, we employ a visualization network to visualize the restored features using various restoration networks on LOCO and AD datasets, and more visualization results are provided in figure 3. We compare the restored features of GBN-only without compensation, LBN-only without compensation, G-L bottleneck without compensation and our template-guided hierarchical feature restoration (THFR) framework.

The GBN-only network is proved to be effective in filtering out anomalies, like missing juice bottle labels and misplacement of transistors. However, it could not restore the details properly, such as missing textures on hazelnuts and blurry wires in cables. The LBN-only network could only partially remove the anomalies, like the orange peel in the breakfast box and the crack on hazelnut, but it is more efficient in restoring the details. By combining these two networks, G-L bottleneck may introduce anomalous features, such as the color on the background of the juice bottle and deformed transistors. With template-guided compensation, THFR restores anomaly-free features. Most of anomalies are removed in the visualization images of the restored features, which further benefits the anomaly detection performance.

4. More experimental results

4.1. MVTec LOCO AD

We provide a detailed comparison of structural and logical anomalies of MVTec LOCO AD dataset [1] in Table 1 and Table 2. We can observe that most existing methods focus on the detection of structural anomalies and achieve satisfactory detection and localization results on structural anomalies. But these methods cannot handle the logical anomalies determined by various complicated logical constraints. Our method introduces a hierarchical structure to correct the semantic-level anomalies and local anomalies at the same time. As a result, THFR outperforms the other works by a remarkable margin in terms of the mean accuracy of detection and localization. We also show visualization results of anomaly detection on all categories in MVTec LOCO AD [1] (seeing Figures 4 to 8), including breakfast box, screw bag, pushpins, splicing connectors, and juice bottle.

4.2. MVTec AD

We provide a detailed comparison on all MVTec AD categories [2] in Tables 3 to 5. For overall categories, we receive comparable results with the other advanced methods. Our method could achieve high accuracy on both image-level detection and pixel-level localization at the same time. Moreover, we show the visualization results of anomaly detection on MVTec AD [2] in Figures 9 and 10.

5. Limitations and discussion

Our method has two potential limitations. Firstly, our network employs a fixed ImageNet [7] pre-trained classification model to extract features for effective feature representation. However, the feature representation that is

Table 1. Pixel-level anomaly localization accuracy on MVTec LOCO AD dataset (sPRO) [1]. Best and second-best scores are bolded and underlined.

Method	Structural anomalies	Logical anomalies	Mean
S-T [3]	<u>75.6</u>	49.7	62.6
DRAEM [12]	74.4	43.7	59.1
CFLOW [8]	70.9	56.7	63.8
RD4AD [6]	**76.0**	50.9	63.5
PatchCore [9]	74.0	55.4	64.7
GCAD [1]	69.2	<u>71.1</u>	<u>70.1</u>
THFR (ours)	75.1	**73.0**	**74.1**

Table 2. Image-level anomaly detection accuracy on MVTec LOCO AD dataset (AUROC) [1]. Best and second best scores are bolded and underlined.

Method	Structural anomalies	Logical anomalies	Mean
S-T [3]	<u>88.3</u>	66.4	77.3
DRAEM [12]	**88.4**	71.9	80.1
CFLOW [8]	87.3	74.4	80.8
RD4AD [6]	87.2	72.2	79.7
PatchCore [9]	87.3	74.7	81.0
GCAD [1]	80.6	**86.0**	<u>83.3</u>
THFR (ours)	86.7	<u>85.2</u>	**86.0**

friendly to classification may lack discriminative low-level features, and therefore make it insensitive to the subtle anomalies, like the subtle color differences between banana juice and orange juice in the juice bottle category. It is possible for the pre-trained models obtained by self-supervised learning to generate feature representation that covers both high-level and low-level features, and we plan to employ it in our future work.

Secondly, we leverage the relationship between the input feature and the template feature to guide the anomaly-free feature restoration process and it could resolve misalignment between the input feature and the template feature. However, when the normal pattern is defined based on the specific rules on object numbers, like the screw bag category in LOCO dataset. It is difficult for relation-based compensation to restore the correct number of objects. To address this issue, we plan to investigate adaptive template composition methods, which could take advantage of local template prior and global semantic relation prior for anomaly-free feature restoration in the future.

6. Evaluation Metrics

6.1. MVTec AD

For simple per-pixel measures, such as AUROC, a single large region segmented correctly can make up for many incorrectly segmented small ones. Therefore, we compute the per-region-overlap (PRO) [2] as an additional anomaly localization metric, which weights ground-truth regions of different sizes equally. While computing the PRO metric, anomaly scores are first thresholded to make a binary decision for each pixel whether an anomaly is present or not. For each connected component within the ground truth, the relative overlap with the thresholded anomaly region is computed. Following the protocol mentioned in [3], we evaluate the PRO value for a large number of increasing thresholds until an average per-pixel false-positive rate of 30% for the entire dataset is reached, and use the area under the PRO curve as a measure of anomaly localization performance.

6.2. MVTec LOCO AD

Unlike structural anomalies, logical anomalies usually have multiple correct predicted anomaly maps, and the union of all areas that could potentially be the cause for the anomaly is labeled as the final anomaly area in MVTec LOCO AD. However, a method is not necessarily required to predict the whole ground truth area. To reflect this, Bergmann *et al.* [1] propose a suitable performance metric that saturates once the overlap with the ground truth exceeds a certain saturation threshold. The metric is called saturated-per-region-overlaps (sPRO) that generalized from the PRO metric as follows:

$$sPRO(M) = \frac{1}{m}\sum_{i=1}^{m} min(\frac{|A_i \cap M|}{s_i}, 1), \qquad (1)$$

where the $\{A_i, ..., A_m\}$ are the set of all defect ground truth regions, $\{s_i, ..., s_m\}$ are a set of corresponding saturation thresholds and the M is the predicted anomaly map. An illustrative example of the sPRO metric with a single ground-truth region is provided in Figure 6. In the fourth row, an additional cable appears between two connectors. The annotated area A covers both the cables and the corresponding saturation threshold s is set to the area of one cable, i.e., half of the annotated region. Hence, all predictions M for which the overlap with A exceeds s fully solve the segmentation task, i.e., $sPRO(M) = 1$. Following [1], we evaluate the sPRO value for a large number of increasing thresholds until an average per-pixel false-positive rate of 5% for the entire dataset is reached, and use the area under the sPRO curve as a measure of anomaly localization performance.

References

[1] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022. 1, 2, 3, 5

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 4, 5, 11, 12

Table 3. Image-level anomaly localization accuracy on MVTec AD (AUROC) [2]. Best and second best scores are bolded and underlined. * To ensure fair comparison with previous studies, CFLOW [8] is evaluated using input images with resolution of 256x256 pixels.

Method \Dataset	Bottle	Cable	Capsule	Carpet	Grid	Hazeln.	Leather	Metal	Pill	Screw	Tile	Toothb.	Trans.	Wood	Zipper	Mean
SPADE [4]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	85.5
PaDiM [5]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	95.3
DRAEM [12]	99.2	91.8	**98.5**	97.0	99.9	**100**	**100**	98.7	**98.9**	93.9	99.6	**100**	93.1	99.1	**100**	98.0
CFLOW [8] *	**100**	97.6	97.7	99.2	91.5	**100**	**100**	99.1	97.1	83.2	**100**	91.9	96.1	99.0	99.4	96.8
RD4AD [6]	**100**	95.0	96.3	98.9	**100**	99.9	**100**	**100**	96.6	97.0	99.3	99.5	96.7	**99.2**	98.5	98.5
PatchCore [9]	**100**	**99.5**	98.1	98.7	98.2	**100**	**100**	**100**	96.6	**98.1**	98.7	**100**	**100**	**99.2**	99.4	99.1
THFR (ours)	**100**	99.2	97.5	**99.8**	**100**	**100**	**100**	**100**	97.8	97.1	99.3	**100**	99.7	**99.2**	97.7	**99.2**

Table 4. Pixel-level anomaly localization accuracy on MVTec AD (AUROC) [2]. Best and second best scores are bolded and underlined. * To ensure fair comparison with previous studies, CFLOW [8] is evaluated using input images with resolution of 256x256 pixels.

Method \Dataset	Bottle	Cable	Capsule	Carpet	Grid	Hazeln.	Leather	Metal	Pill	Screw	Tile	Toothb.	Trans.	Wood	Zipper	Mean
SPADE [4]	98.4	97.2	**99.0**	97.5	93.7	99.1	97.6	98.1	96.5	98.9	87.4	97.9	94.1	88.5	96.5	96.0
PaDiM [5]	98.3	96.7	98.5	99.1	97.3	98.2	99.2	97.2	95.7	98.5	94.1	98.8	**97.5**	94.9	98.5	97.5
DRAEM [12]	**99.1**	94.7	94.3	95.5	**99.7**	**99.7**	98.6	**99.5**	97.6	97.6	**99.2**	98.1	90.9	**96.4**	**98.8**	97.3
CFLOW [8] *	98.8	97.6	97.7	**99.2**	96.9	98.8	**99.6**	98.6	**98.9**	98.1	97.7	98.6	93.9	94.5	98.4	97.9
RD4AD [6]	98.7	97.4	98.7	98.9	99.3	98.9	99.4	97.3	98.2	**99.6**	95.6	99.1	92.5	95.3	98.2	97.8
PatchCore [9]	98.6	98.4	98.8	99.0	98.7	98.7	99.3	98.4	97.4	99.4	95.6	98.7	96.3	95.0	**98.8**	98.1
THFR (ours)	98.9	**98.5**	98.7	**99.2**	99.3	99.2	99.4	97.4	98.0	99.5	95.5	**99.2**	95.9	95.3	98.7	**98.2**

Table 5. Pixel-level anomaly localization accuracy on MVTec AD (AUPRO) [2]. Best and second-best scores are bolded and underlined. * To ensure fair comparison with previous studies, CFLOW [8] is evaluated using input images with resolution of 256x256 pixels.

Method \Dataset	Bottle	Cable	Capsule	Carpet	Grid	Hazeln.	Leather	Metal	Pill	Screw	Tile	Toothb.	Trans.	Wood	Zipper	Mean
SPADE [4]	95.5	90.9	93.7	94.7	86.7	95.4	97.2	**94.4**	94.6	96.0	75.9	93.5	**87.4**	87.4	92.6	91.7
PaDiM [5]	94.8	88.8	93.5	96.2	94.6	92.6	97.8	85.6	92.7	94.4	86.0	93.1	84.5	91.1	95.9	92.1
S-T [3]	93.1	81.8	**96.8**	87.9	95.2	96.5	94.5	94.2	96.1	94.2	94.6	93.3	66.6	91.1	95.1	91.4
CFLOW [8] *	93.3	93.5	93.4	96.3	90.9	**96.7**	98.9	91.7	95.4	93.1	**91.3**	87.4	84.1	91.4	93.4	92.7
RD4AD [6]	96.6	91.0	95.8	97.0	97.6	95.5	99.1	92.3	**96.4**	**98.2**	90.6	94.5	78.0	90.9	95.4	93.9
PatchCore [9]	96.2	92.5	95.5	96.6	96.0	93.8	98.9	91.4	93.2	97.9	87.3	91.5	83.7	89.4	**97.1**	93.4
THFR (ours)	**97.2**	**94.8**	95.9	**97.7**	**97.7**	96.2	**99.2**	90.5	**96.4**	**98.2**	90.8	**94.7**	85.9	**93.3**	96.6	**95.0**

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 4

[4] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 4

[5] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 475–489, Cham, 2021. Springer International Publishing. 4

[6] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, June 2022. 3, 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[8] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 98–107, January 2022. 3, 4

[9] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, June 2022. 3, 4

[10] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1

[11] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. 1

[12] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, October 2021. 3, 4

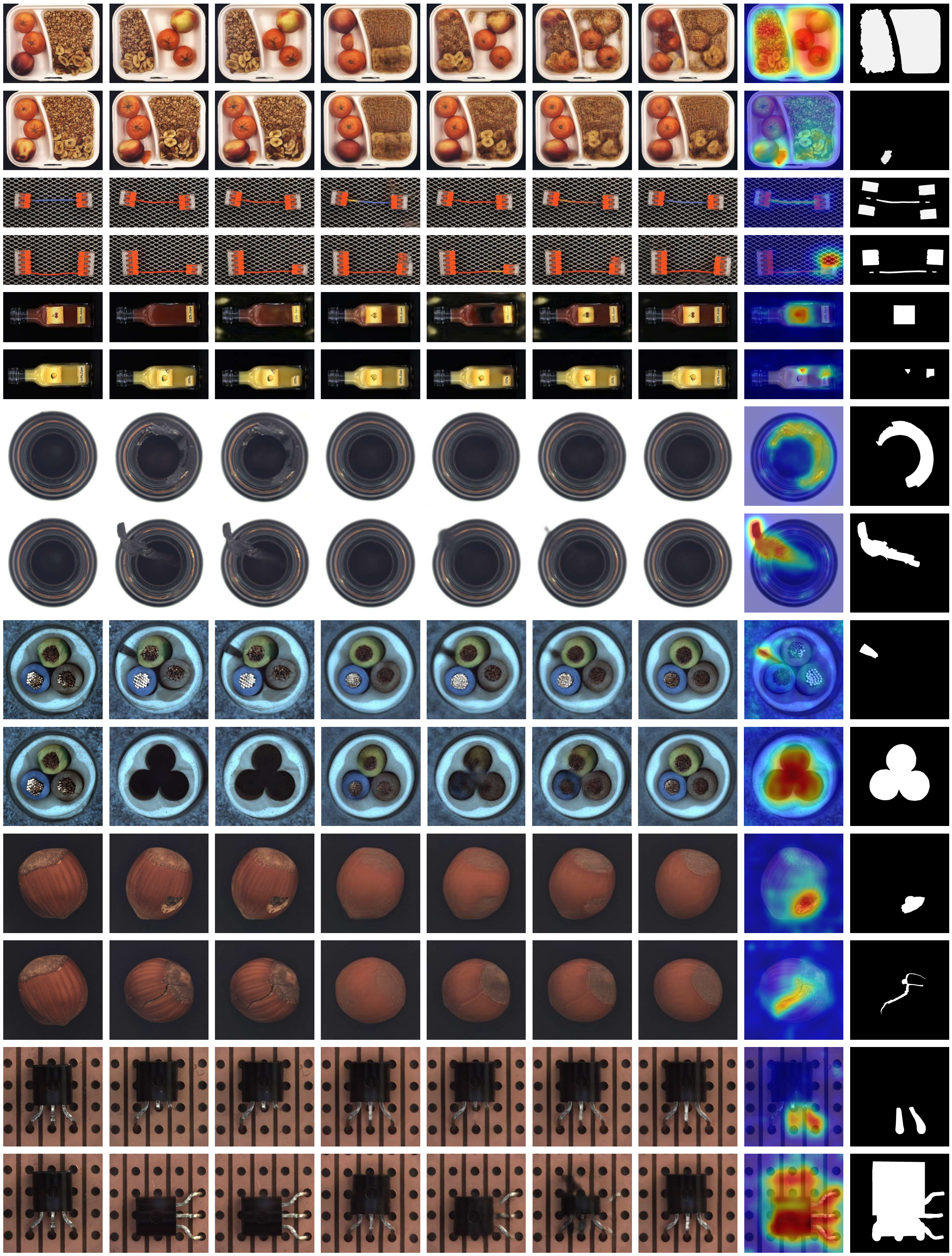

Figure 3. Visualization of restored features from different restoration networks on MVTec LOCO AD [1] and MVTec AD [2]. From left to right: normal image, anomaly image, reconstructed images from the pre-trained embedding features (Recon), restored features of GBN-only without compensation, LBN-only without compensation, G-L bottleneck without compensation and our template-guided hierarchical feature restoration (THFR) framework, anomaly map, and ground truth. The visualization results show how restoration networks progressively achieve anomaly-free restoration with G-L bottleneck and template-guided hierarchical feature compensation.

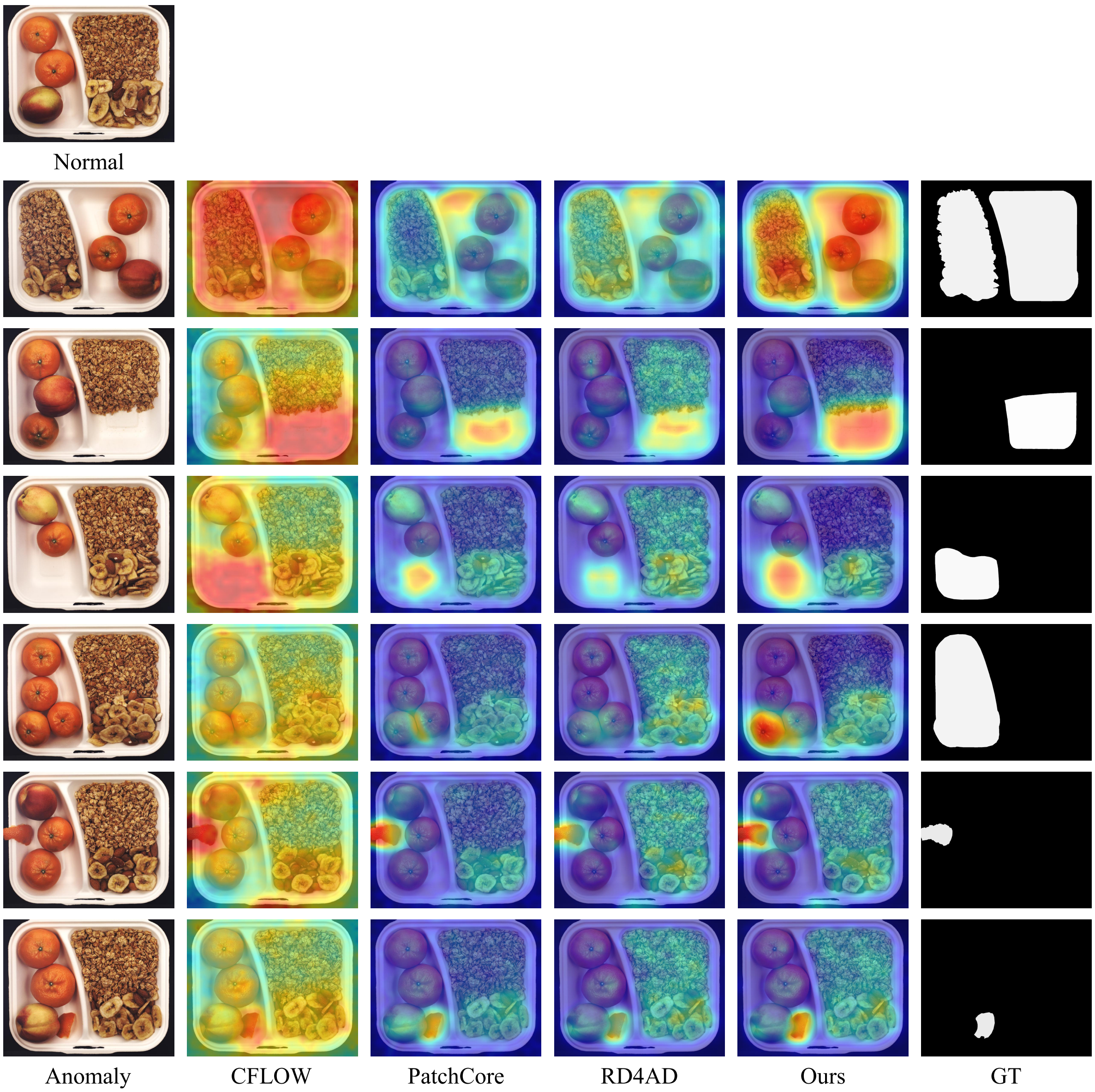

Figure 4. Qualitative results of breakfast box. Normal: breakfast box contains exactly two tangerines and one nectarine which are always located on the left-hand side of the box. Furthermore, the ratio and relative position of the cereals and the mix of banana chips and almonds on the right-hand side are fixed. We show results on typical anomalies, including arrangement changes, missing banana chips and almonds, missing orange, and additional objects.

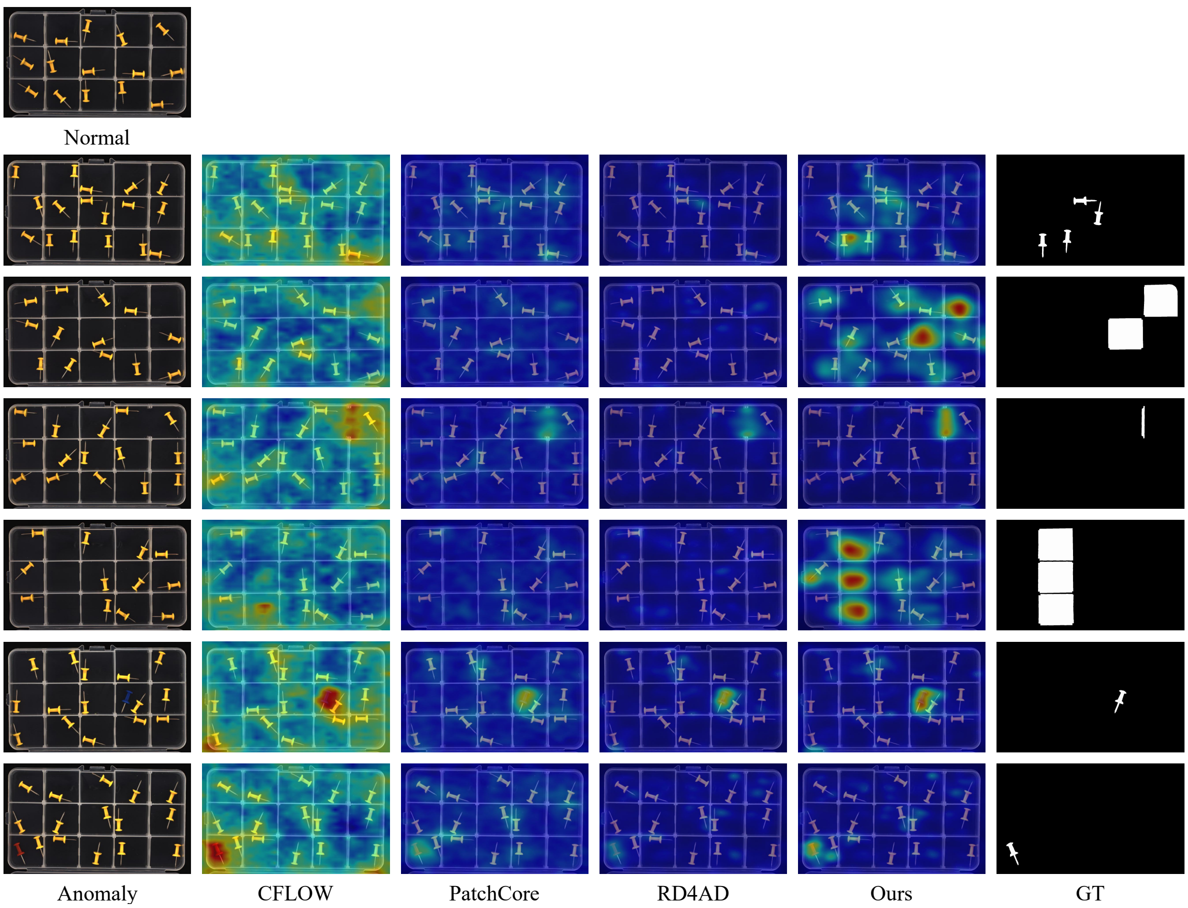

Figure 5. Qualitative results of pushpins. Normal: each compartment of the box contains exactly one pushpin. We show results on typical anomalies, including one additional pushpin, missing pushpin, missing separator, and additional object.

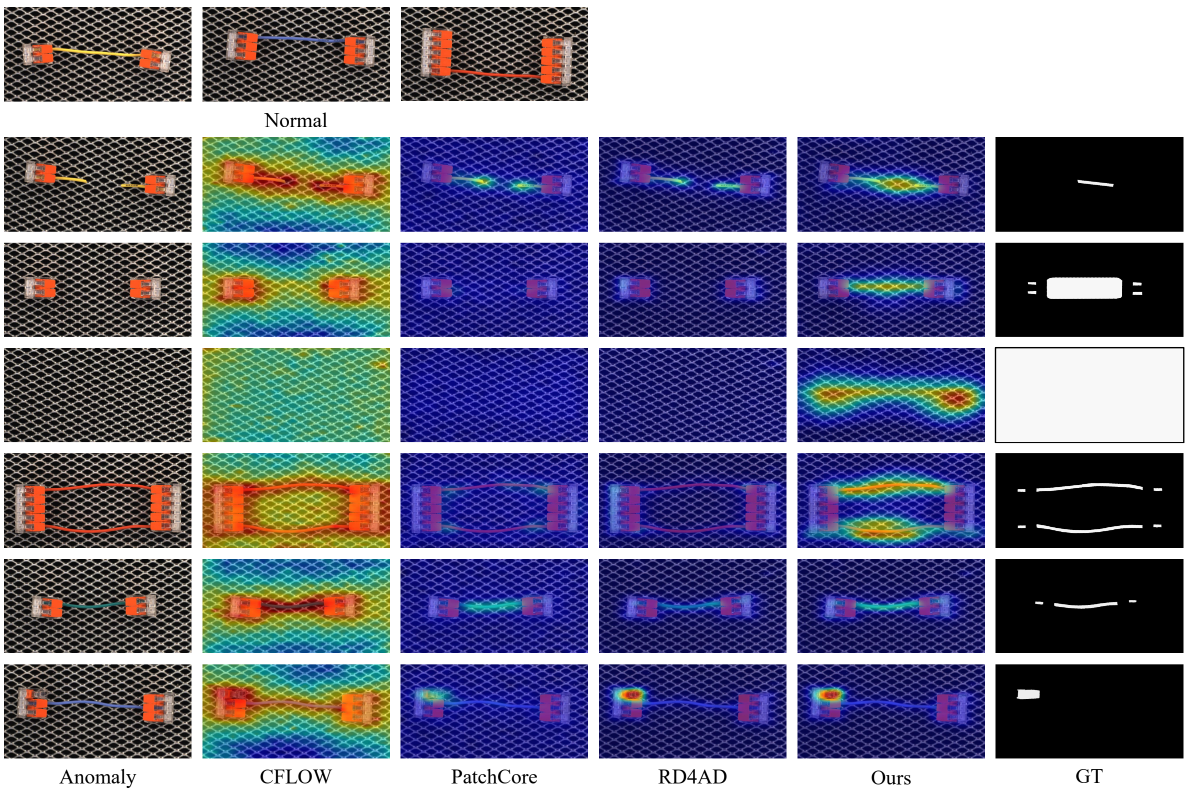

Figure 6. Qualitative results of splicing connectors. Normal: exactly two splicing connectors with the same number of cable clamps are linked by exactly one cable. In addition, the number of clamps has a one-to-one correspondence to the color of the cable, and the cable has to terminate in the same relative position on its two ends such that the whole construction exhibits mirror symmetry. We show results on typical anomalies, including cable broken, missing cable, missing connectors, additional cable, wrong cable color, and connector broken.

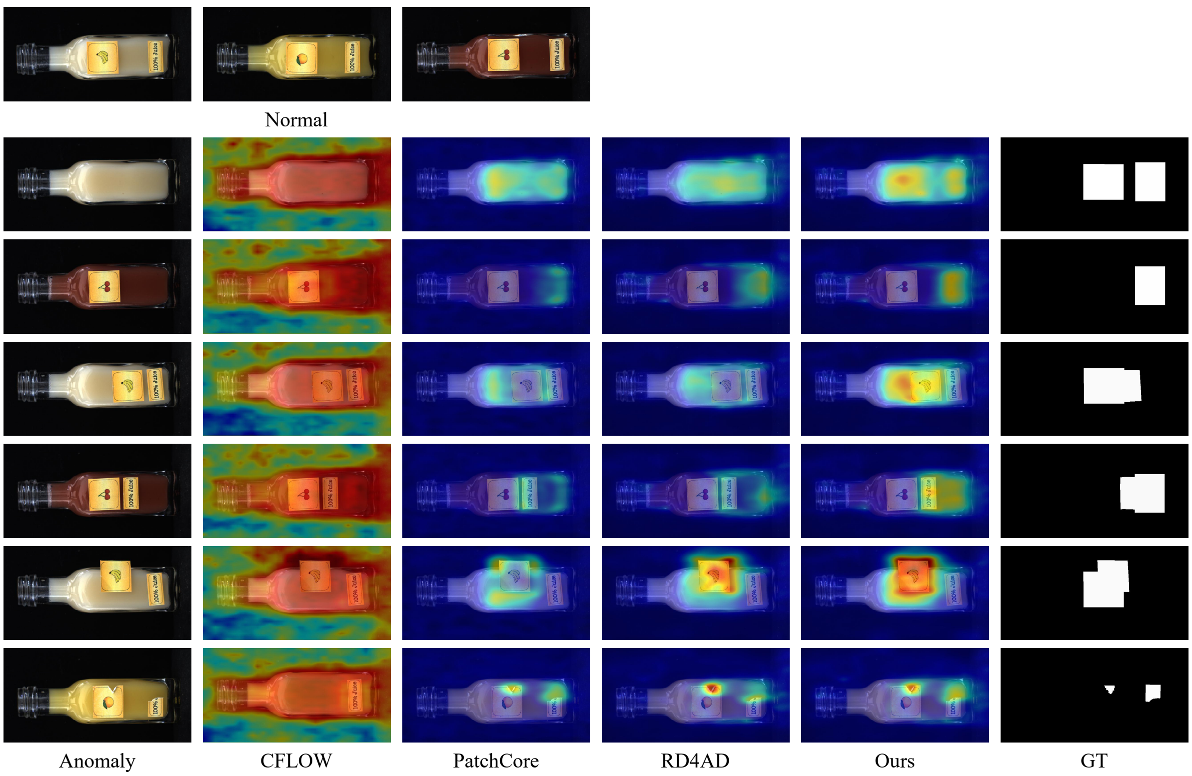

Figure 7. Qualitative results of juice bottle. Normal: each juice bottle is filled with one of three differently colored liquids and carries exactly two labels. The first label is attached to the center of the bottle and displays an icon that determines the type of liquid. The second is attached to the lower part of the bottle with the text "100% Juice" written on it. The fill level is the same for each bottle. We show results on typical anomalies, including missing all labels, missing bottom label, misplaced top label, misplaced bottom label, label location wrong and label broken. The failure cases of juice color are analyzed in Section 5 of the paper.

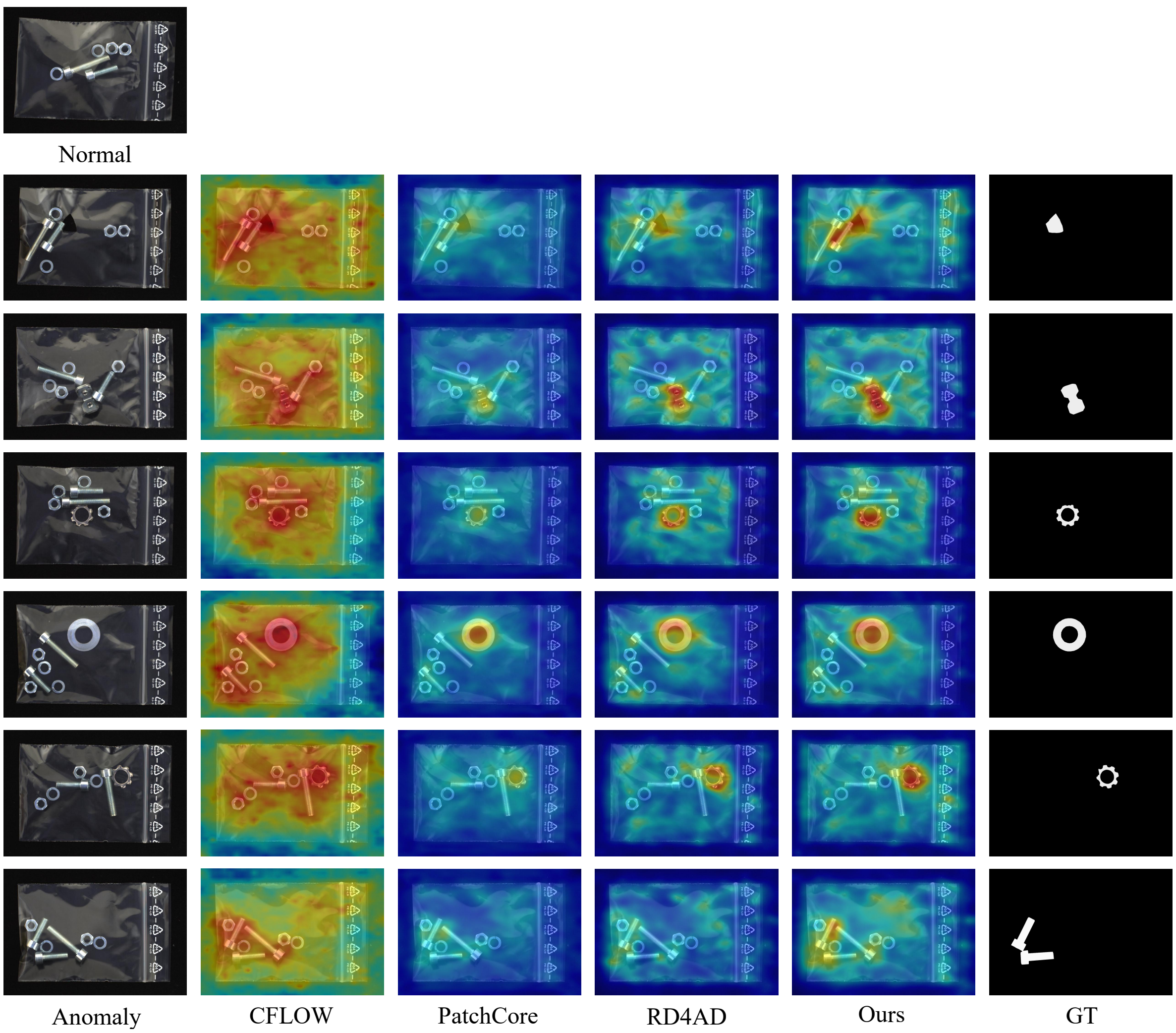

Figure 8. Qualitative results of screw bag. Normal: a screw bag contains exactly two washers, two nuts, one long screw, and one short screw. We show results on typical anomalies, including bag broken, additional object, and one additional short screw. The existing methods failed to achieve satisfactory results due to the interference caused by the reflection of screw bags.

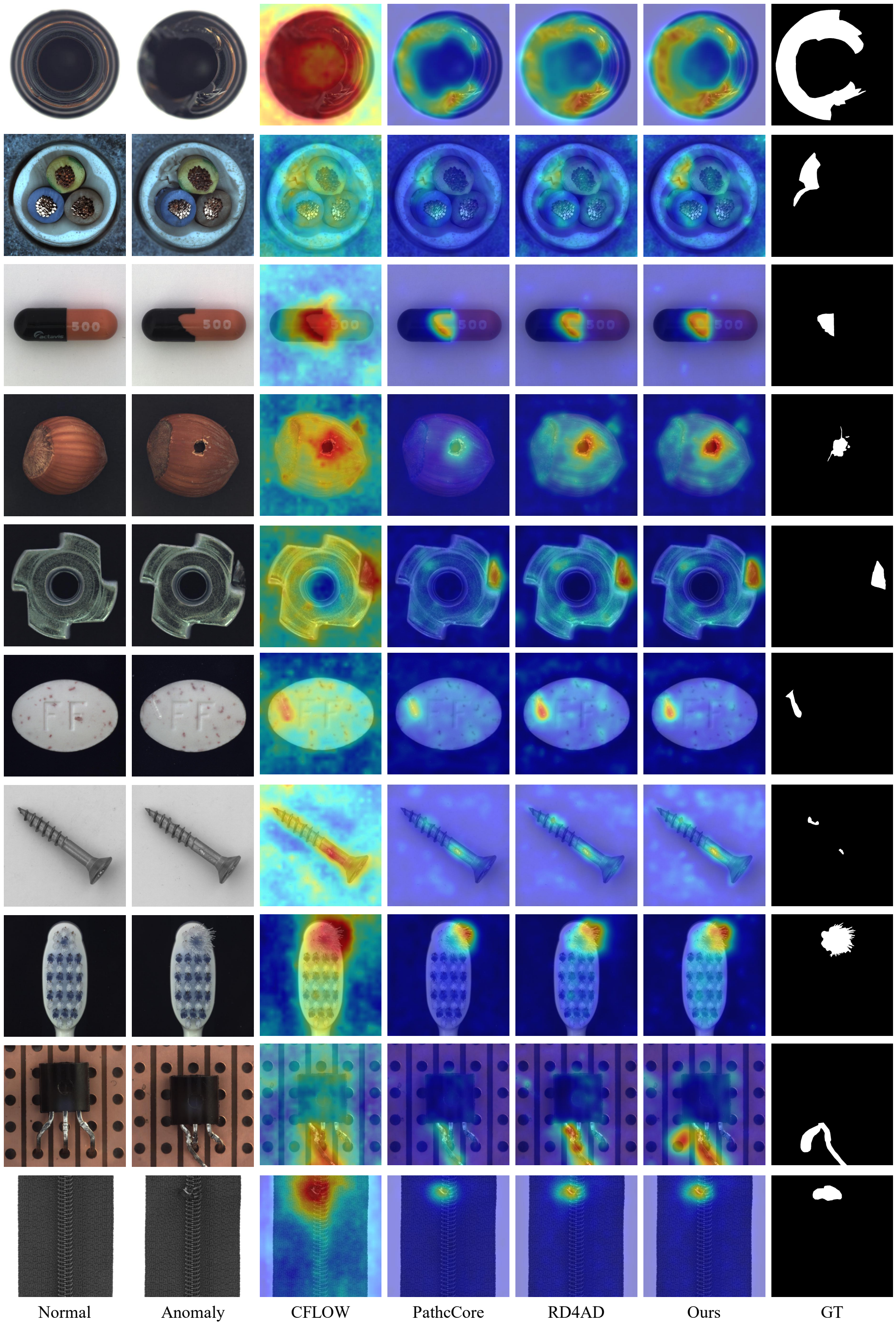

Figure 9. Qualitative results of object categories in MVTec AD [2].

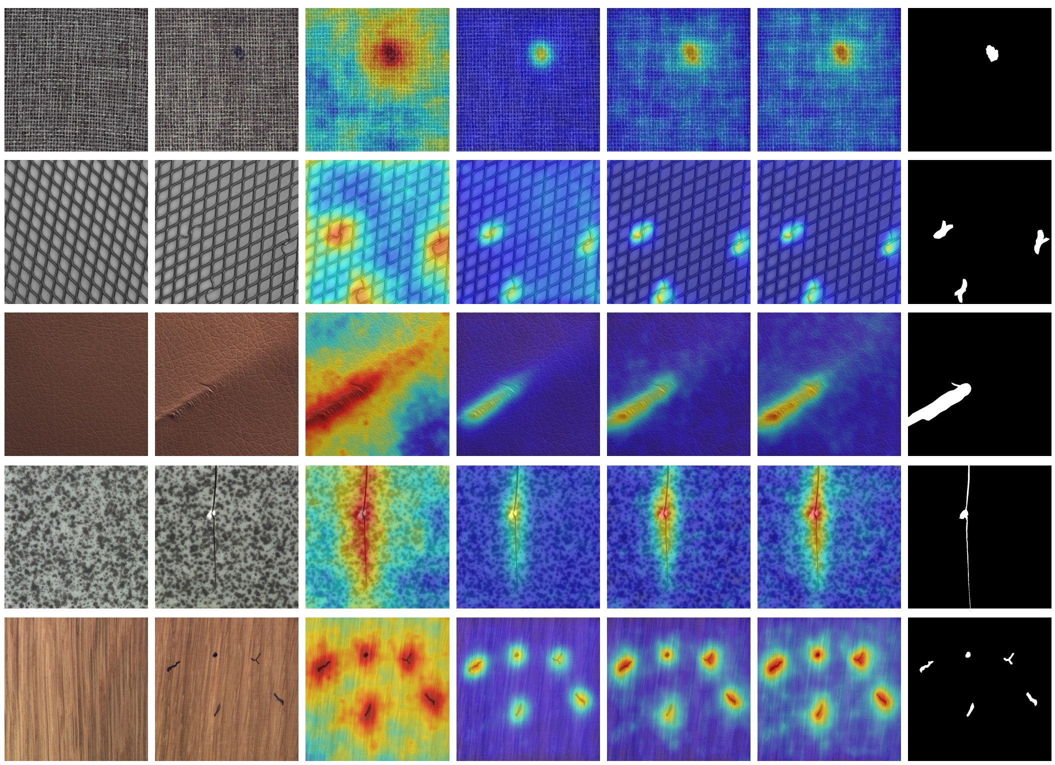

Figure 10. Qualitative results of texture categories in MVTec AD [2].