Supplementary Material for LSFA

Yuanpeng Tu[1*], Boshen Zhang[2*], and Liang Liu[2*], Yuxi Li[2], Jiangning Zhang[2], Yabiao Wang[3,2†], Chengjie Wang[2], Cairong Zhao[1†]

[1] Tongji University, Shanghai
{2030809, zhaocairong}@tongji.edu.cn
[2] YouTu Lab, Tencent, Shanghai
{boshenzhang,leoneliu,yukiyxli,vtzhang,caseywang,jasoncjwang}@tencent.com
[3] Zhejiang University

1 Datasets

MVTec-3D [10] dataset is the first comprehensive benchmark in 3D industrial anomaly detection, which consists of 10 object categories and 2,657/1,137 samples for training/testing respectively. Both the anomaly-free training set and mixed test set are available in MVTec-3D AD. Since all the point samples are collected from the same angle, only one RGB sample is recorded for each point cloud. The collected samples are generated from industrial sensors with structured light to facilitate the detection of anomalies that is inconspicuous only based on RGB modality.

Eyecandies [11] is the latest synthetic dataset for 3D unsupervised anomaly detection and localization, containing 10 classes of procedurally generated images with depth and normal maps. It has a larger intra-class variation than the MvTec-3D dataset, and the anomalies are automatically generated to reduce human bias. The dataset is split into training, validation, and test sets, with 1,000, 100, and 400 instances respectively, and half of the samples in the test sets are anomaly candies.

Real3D-AD [9] is collected using high-resolution laser scans, perfect for spotting the product's defects everywhere. It comprises a total of 1,254 samples that are distributed across 12 distinct categories. Each training set for a specific category contains only four samples, similar to the few-shot scenario in 2D anomaly detection. These categories include but are not limited to Airplane, Candybar, Chicken, Diamond, Duck, Fish, Gemstone, Seahorse, Shell, Starfish, and Toffees. All these categories are toys from manufacturing lines. Real3D-AD demonstrates a point resolution and precision of 0.04mm and 0.011mm, respectively. This is notably higher than MVTec 3D [10], with a factor of 4.28 and 9 for point resolution and precision, respectively.

* Equal contribution.
† Corresponding author.

2 Results on Eyecandies

To further verify the effectiveness of our method, we conduct experiments on the Eyecandies dataset. Besides the comparison of I-AUROC in the main text, we further provide the P-AUROC results on RGB/RGB+3D modality as shown in Table. 5 and 6 respectively. For I-AUROC results, our LSFA outperforms AE [11] by a large margin of 18.9% and 5.4% for I-AUROC/P-AUROC. For multimodal inputs, LSFA also achieves state-of-the-art performance and obtains 21.6%/5.4% accuracy improvement compared with AE. Different from MVTec-3D AD, the performance of combining RGB and 3D modalities is generally worse than the results of RGB modality. This is mainly due to that there exist defects of some classes (i.e., Chocolate C.) that are only visible in RGB modality while having little difference in the 3D modality(see Fig. 3), making it generate deteriorated results when directly fusing the unreliable scores of 3D modality and precise scores of RGB modality.

3 Results on Real3D-AD

To further verify the effectiveness of our LSFA, we conduct experiments on the most recent 3D industrial anomaly detection benchmark Real3D-AD. Specifically, we utilize the FPFH+Raw as the pre-trained features and introduce additional adaptors (i.e., single vanilla transformer encoder layer) for both RGB and 3D modalities to perform multi-scale feature adaptation. As shown in Table. 1, M3DM [8] is much inferior to its baseline Patchcore [20], which indicates its poor generality and thus may not applicable in real-world scenarios. And our LSFA acheieves 71.2% object-level AUROC and outperforms all the compared methods, demonstrating its great generality in various scenarios.

Implementation details.

For the feature extractors of the RGB modality, a ViT-B/8 [14] with DINO [12] is adopted. The 768-dim output of the final layer is used and then pooled into 56×56 for subsequent training. For the 3D modality, a point transformer [19] pre-trained on ShapeNet [13] dataset is utilized and the outputs from 3/7/11 layer are concatenated to fuse multi-scale information. Similar to patches in ViT, the point transformer clusters point clouds into multiple local groups and these groups have their corresponding center points for position and neighbor numbers for group size. For data processing, the background area of depth and RGB images is removed by estimating the background plane of depth images with RANSAC [15], where points within 5×10^{-2} are ignored to alleviate the influence of background. Finally, the RGB and point cloud tensor are both resized to 224 × 224 to be consistent with the input size. The projected features for point clouds and RGB samples in CLC are 512-dim. The AdamW optimizer is used and its learning rate is set as 2×10^{-3} with cosine warm-up. The batch size N_b for adaptation is set as 8.

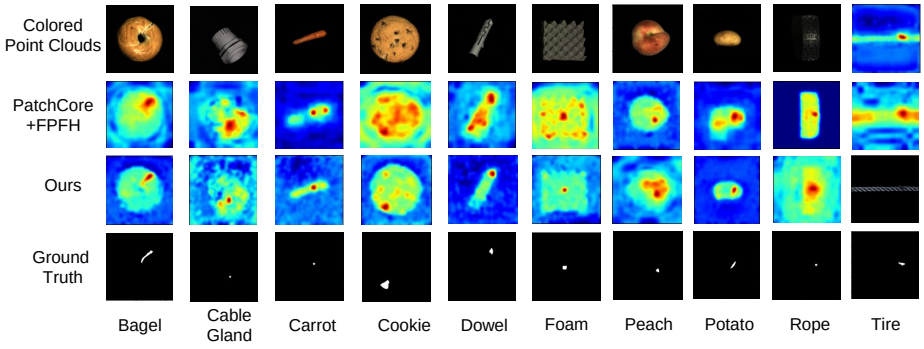

Fig. 1: Visualization of anomaly segmentation results for all the categories. Compared with PatchCore+FPFH, LSFA consistently performs more precise defect localization.

Table 1: ADBench-3D for Real3D-AD. The score indicates object-level AUROC ↑. The best results are highlighted in bold.

Category	**BTF** [17]		**M3DM** [8]		**PatchCore** [20]			**Reg3D-AD** [9]	**LSFA**
	Raw	*FPFH*	*PointMAE*	*PointBERT*	*FPFH*	*FPFH+Raw*	*PointMAE*		
Airplane	0.730	0.520	0.434	0.407	**0.882**	0.848	0.726	0.716	0.854
Car	0.647	0.560	0.541	0.506	0.590	0.777	0.498	0.697	**0.791**
Candybar	0.539	0.630	0.552	0.562	0.541	0.570	0.663	**0.685**	0.651
Chicken	0.789	0.432	0.683	0.673	0.837	0.853	0.827	0.852	**0.887**
Diamond	0.707	0.545	0.602	0.627	0.574	0.784	0.783	**0.900**	0.849
Duck	0.691	**0.784**	0.433	0.466	0.546	0.628	0.489	0.584	0.673
Fish	0.602	0.549	0.540	0.556	0.675	0.837	0.630	**0.915**	0.872
Gemstone	**0.686**	0.648	0.644	0.617	0.370	0.359	0.374	0.417	0.391
Seahorse	0.596	**0.779**	0.495	0.494	0.505	0.767	0.539	0.762	0.773
Shell	0.396	**0.754**	0.694	0.577	0.589	0.663	0.501	0.583	0.682
Starfish	0.530	**0.575**	0.551	0.528	0.441	0.471	0.519	0.506	0.490
Toffees	0.703	0.462	0.450	0.442	0.565	0.626	0.585	**0.827**	0.641
Average	0.635	0.603	0.552	0.538	0.593	0.682	0.594	0.704	**0.712**

3.1 Visualization Analysis

To verify the superiority of LSFA, we perform visualization of the anomaly detection results for all the categories of the MVTec-3D AD dataset as shown in Fig. 4. The activation map results show that our LSFA consistently generates more precise anomaly segmentation scores than PatchCore+FPFH [16] baseline.

4 Difference from Existing Works

Since our LSFA is slightly similar to 3D CLIP [1], we analyze their essential differences. Specifically, 3D CLIP only uses coarse-level global feature, while ignoring the finer-grained local feature alignment that is more essential for AD. For [5,7], they are designed for reconstruction based methods rather than memory bank based ones. Thus the purpose of IFC and [5,7] is essentially different.

Table 2: Performance of LSFA with confidence intervals over 10 runs on MVTec-3D AD.

Modality	RGB	3D	RGB+3D
AUROC	0.9212 ± 0.0038	0.8609 ± 0.0026	0.9713 ± 0.0034
AUPRO	0.9341 ± 0.0035	0.9449 ± 0.0024	0.9683 ± 0.0041

The latter aims to model pixel-level feature prototype to maximize reconstruction error. Such an update scheme is not suitable for our LSFA, since it makes the memory bank lack sufficient feature diversity of local-wise normal patterns, thus leading to inferior performance. Moreover, we have conducted experiments by adopting the same memory update scheme in [5, 7] and only achieve 94.1% I-AUROC for RGB+3D modality on MVTec-3D AD, validating the superiority of our IFC.

5 Parameter Sensitivity

Since our LSFA achieves similar performance across various values for both n_I^L and λ as shown in the main text. Here we conduct further analysis on the influence of n_I^L and λ by setting them as more extreme values.

Investigation on n_I^L. We have conducted experiments by setting n_I^L from [100, 1,000, 4,000, 7,000, 10,000], which achieve 89.6%/90.1%/93.5%/94.2%/94.9% for I-AUROC of RGB+3D modality on MVTec-3D AD respectively. Notably, LSFA suffers from accuracy degradation when n_I^L <10000. This is because small N_I^L leads to decreased feature diversity in the memory bank and performing alignment with such limited banks as in Eq.5&6 may lead to over-fittig to the training set, thus resulting in mis-classification of normal patterns. Therefore, we set $N_I^L = 20,000$ for trade-off between accuracy and memory.

Investigation on λ. Here we set λ from [0, 0.05, 0.1, 0.15] and our LSFA achieves 95.9%/96.1%96.4/%/96.5% for I-AUROC of RGB+3D modality on MVTec-3D AD respectively. And with λ becoming even larger, the accuracy is improved from 95.9% to 96.7%. Such a phenomenon indicates that cross-modal representation misalignment in both local and global views will inevitably induce a negative impact on the complementary fusion of features from two modalities, leading to deteriorated accuracy.

6 Performance Stability

Table 2 shows the performance of LSFA over 10 runs with confidence intervals on MVTec-3D AD. On Eyecandies, I-AUROC is 90.59%±0.20%. The results demonstrates the performance stability of the proposed LSFA.

7 Comparison with inputing RGBD data

Here we have evaluated the performance of feeding RGB-D image into a single model and separately feeding RGB and depth images into two models in the proposed LSFA. The mean I-AUROC/AUPRO of a single model and two models for RGB+3D are 87.63%/93.25% and 88.12%/93.74% respectively, which is much inferior to the results of adopting two networks. This may be due to the large domain discrepancy between RGB and depth images, thus the features of depth generated from the pre-trained models is of low quality.

8 Comparison with Fusion based Methods

Given the multi-modal nature of 3D industrial anomaly detection, here we compare our LSFA with multi-modal fusion based works. Specifically, we have re-implemented four fusion based methods, including [4], [3], [6], [2] and adopt a similar scheme to [8] for AD. On the MVTec3D AD, [4], [3], [6], [2] achieve 92.1% /93.0% /91.9% /92.4% I-AUROC for RGB+3D modality, which are much inferior to LSFA. We attribute such results for that they fail to take intra-modality compactness into consideration. Thus directly combining them for 3D industrial anomaly detection will inevitably result in much inferior performance.

9 Sensitivity to Pre-training weights

To verify the robust performance of LSFA when loading different pre-trained weights, we replace the DINO with SAM [18] for the model of RGB-modality. The I-AUROC and AUPRO results are: 86.31%/97.02% and 94.47%/96.29% for RGB/RGB+3D, which is similar to the performance of adopting the pre-trained DINO.

10 P-AUROC for Segmentation

Besides the I-AUROC and AUPRO shown in the manuscript, we also report the P-AUROC metric on MVTec3D for more comprehensive evaluation of our method. As shown in Tab. 7, we compare our method with FPFH [16], PatchCore [20], AST [21]. Moreover, for a fair comparison, we also use the features extracted from the same pre-trained backbone for FPFH and PatchCore. The results demonstrate that our method consistently outperforms the compared methods across all the settings. For example, LSFA surpasses AST by 2.1% with the multimodal inputs.

Table 3: Investigation on the structure of Ψ_I/Ψ_P.

Structure Ψ_I/Ψ_P	I-AUROC	AUPRO	P-AUROC
Linear projection	0.953	0.959	0.989
Single encoder layer	**0.974**	**0.968**	**0.993**
Two encoder layers	0.954	0.963	0.984
1×1 Convolution	0.951	0.962	0.986

11 Investigation on Adaptor Structure Ψ_I/Ψ_P

In this section, we investigate the influence of different structures of adaptor within LSFA, including linear projection layer, single vanilla transformer encoder layer [14], two vanilla transformer encoder layers, and single 1×1 convolution layer are evaluated. As shown in Tab. 3, the single vanilla transformer encoder layer performs better than linear projection since it is better at integrating local geometry information for local defect localization. However, the performance deteriorates when multiple encoder layers are integrated. This may explained by the over-fitting problem, i.e., over-parameterized network leads to fully memorize the training sets, thus resulting in misestimating a large number of normal samples as abnormal ones in the test set. Moreover, the performance of the 1×1 convolution layer is also inferior to the results of a single vanilla transformer encoder layer. Thus we choose a single vanilla transformer encoder layer as our adaptors for two modalities.

12 Rotation-Invariant Intra-modal Augmentation

Besides improving intra-modal feature compactness in IFC, a patch-wise rotation strategy is introduced in both RGB/3D modalities to improve the rotation invariant properties of deep model. As shown in Fig. 2, patches of I_i are randomly rotated 90°/180°/270° and then the rotated image $\widehat{I_i}$ are fed into $\phi_I(\cdot)$ and $\Psi_I(\cdot)$ for feature generation. Denote the adapted feature of $\widehat{I_i}$ as $\widehat{F_{I_i}}$, a patch-wise contrastive loss is introduced to maximize/minimize the similarity between features of the same/different j-th patch in I_i and $\widehat{I_i}$:

$$\mathcal{L}_{\text{RI}} = \frac{F_{I_i}^{(i,j)} \cdot F_{\widehat{I_i}}^{(i,j)T}}{\sum_{t=1}^{N_b}\sum_{k=1}^{N_m} F_{I_i}^{(t,k)} \cdot F_{\widehat{I_i}}^{(t,k)T}}. \tag{1}$$

However, since neighboring patches tend to have a more similar appearance, minimizing similarity for features of neighboring patches is unreasonable. Therefore, in implementations, we randomly sample patch-wise features from F_{I_i} and $F_{\widehat{I_i}}$ to alleviate the mentioned issue. A similar strategy is operated on the 3D point clouds as well, which rotates 90°/180°/270° with the depth coordinates fixed.

The ablation study result of the designed augmentation strategy is shown in Table. 4. The results show that by enhancing the robustness against angle

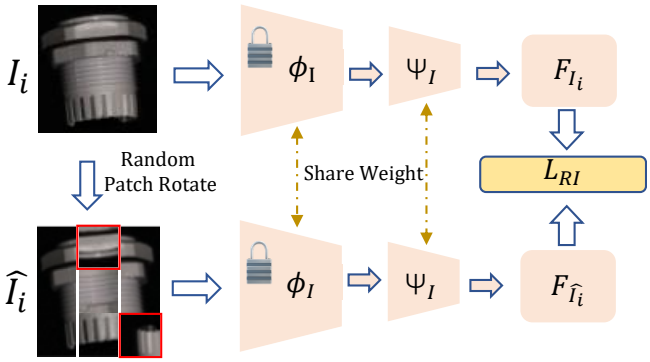

Fig. 2: The proposed Rotation-Invariant Augmentation Strategy for RGB modality.

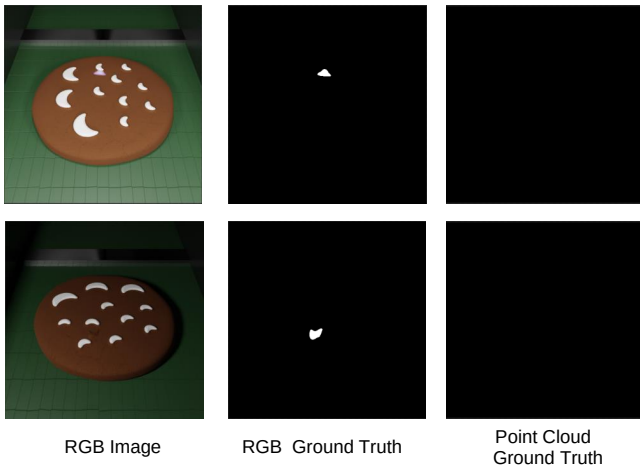

Fig. 3: Examples in Eyecandies dataset, where some defects are invisible in 3D modality.

variation, the proposed augmentation strategy can slightly improve the baseline results.

Table 4: Investigation on the augmentation strategy.

Component $\mathcal{L}_{RI}$	I-AUROC	AUPRO	P-AUROC
✗	0.929	0.953	0.987
✓	**0.935**	**0.957**	**0.988**

Table 5: P-AUROC score for anomaly segmentation of all categories of Eyecandies [11] dataset on RGB modality.

	Method	Candy Cane	Choco late C.	Choco late P.	Confetto	Gummy Bear	Hazel nut T.	Licor ice S.	Lollipop	Marsh mallow	Peppe rmint C.	Mean
RGB	AE [11]	0.972	0.933	0.960	0.945	0.929	0.815	0.855	0.977	0.931	0.928	0.924
	PatchCore*/M3DM [8]	0.962	0.981	0.963	**0.998**	0.974	0.938	**0.978**	**0.987**	**0.997**	**0.990**	0.976
	Ours	**0.964**	**0.983**	**0.964**	**0.998**	**0.976**	**0.939**	**0.978**	**0.987**	**0.997**	**0.990**	**0.978**

Table 6: P-AUROC score for anomaly segmentation of all categories of Eyecandies [11] dataset on RGB+3D modalities.

	Method	Candy Cane	Choco late C.	Choco late P.	Confetto	Gummy Bear	Hazel nut T.	Licor ice S.	Lollipop	Marsh mallow	Peppe rmint C.	Mean
RGB+3D	AE [11]	0.973	0.927	0.958	0.945	0.929	0.806	0.827	0.977	0.931	0.928	0.920
	PatchCore*/M3DM [8]	0.967	0.936	0.963	**0.996**	0.968	0.928	0.969	0.989	**0.998**	0.978	0.969
	Ours	**0.969**	**0.957**	**0.967**	**0.996**	**0.971**	**0.938**	**0.970**	**0.990**	**0.998**	**0.987**	**0.974**

13 Detailed Results of Ablation Study

In the Section 4.3 of the manuscript, we perform ablation studies on the components IFC and CLC of LSFA together with the investigation on the each module in IFC/CLC. Here we first present the detailed performance for each variant of our methods in Table 10 and 11 respectively. The results show that by optimizing the intra-modal compactness, IFC brings significant performance boost to the accuracy of Cable Gland/Foam. Similarly, CLC also prominently improves the performance on the Tire class. Moreover, the detailed results of Table.5 in the manuscript are shown in Table. 12 and 15. The Table. 14 and 15 present the detailed ablation study results of the proposed CLC. To summarize, all the results demonstrate the effectiveness of each component in our LSFA.

14 Detailed Results of Few-shot Settings

For a better evaluation of LSFA on few-shot settings, the detailed performance of Tab.7 in the manuscript on all the classes of MVTec-3D AD is shown in Table. 8 and 9. The results clearly show the superiority of our LSFA for each category when compared with some non-few-shot methods.

References

1. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In: ICCV
2. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: ICCV
3. Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)

Table 7: P-AUROC score for anomaly segmentation of all categories of MVTec-3D AD [10] dataset. The P-AUROC is a saturated metric for anomaly segmentation, and the difference between methods is smaller than the AUPRO. '*' denotes using the same pre-trained backbone as ours.

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
3D	FPFH [16]	**0.994**	0.966	**0.999**	0.946	0.966	0.927	0.996	**0.999**	**0.996**	0.990	0.978
	FPFH* [16]	0.981	0.949	0.997	0.932	0.959	0.925	0.989	0.995	0.994	0.981	0.970
	Ours	**0.994**	**0.969**	0.998	**0.953**	**0.967**	**0.938**	**0.998**	**0.999**	**0.996**	**0.998**	**0.981**
RGB	PatchCore [20]	0.983	0.984	0.980	0.974	0.972	0.849	0.976	0.983	0.987	0.977	0.967
	PatchCore* [20]	0.992	0.990	0.994	**0.977**	0.983	0.955	0.994	**0.990**	**0.995**	0.994	**0.987**
	Ours	**0.993**	**0.992**	**0.996**	**0.977**	**0.984**	**0.957**	**0.996**	**0.990**	0.994	**0.995**	**0.987**
RGB+3D	AST [21]	-	-	-	-	-	-	-	-	-	-	0.976
	PatchCore + FPFH [16]	**0.996**	0.992	0.997	**0.994**	0.981	0.974	0.996	0.998	0.994	0.995	0.992
	PatchCore* + FPFH* [16]	0.995	0.993	0.997	0.985	**0.985**	0.984	0.996	0.994	0.997	0.996	0.992
	Ours	**0.996**	**0.994**	**0.998**	0.973	0.981	**0.988**	**0.999**	**0.999**	**0.998**	**0.999**	**0.993**

Table 8: Detailed I-AUROC results of our method under few-shot settings.

Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
5-shot	0.981	0.772	0.872	0.942	0.682	0.798	0.827	0.781	0.920	0.762	0.834
10-shot	0.987	0.842	0.880	0.969	0.767	0.827	0.908	0.832	0.931	0.773	0.871
50-shot	0.997	0.860	0.957	0.966	0.910	0.915	0.937	0.910	0.946	0.863	0.926
Full dataset	**1.000**	**0.939**	**0.982**	**0.989**	**0.961**	**0.951**	**0.983**	**0.962**	**0.989**	**0.951**	**0.971**

Table 9: Detailed AUPRO results of our method under few-shot settings.

Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
5-shot	0.959	0.919	0.974	0.916	0.891	0.872	0.968	0.957	0.967	0.939	0.936
10-shot	0.963	0.923	0.975	0.923	0.906	0.899	0.969	0.957	0.969	0.943	0.943
50-shot	0.981	0.962	0.979	0.941	0.920	0.932	0.978	0.973	0.972	0.979	0.962
Full dataset	**0.986**	**0.974**	**0.981**	**0.946**	**0.925**	**0.941**	**0.983**	**0.983**	**0.974**	**0.983**	**0.968**

Table 10: Detailed ablation results of I-AUROC for the component IFC and CLC.

Component												
IFC	CLC	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
✗	✗	0.981	0.831	0.980	0.985	0.96	0.905	0.936	0.964	0.967	0.780	0.929±0.0024
✗	✓	0.987	0.892	0.981	0.986	0.96	0.937	0.959	0.959	0.972	0.932	0.957±0.0041
✓	✗	0.989	0.901	0.980	0.988	0.959	0.942	0.967	0.960	0.970	0.929	0.959±0.0018
✓	✓	**1.000**	**0.939**	**0.982**	**0.989**	**0.961**	**0.951**	**0.983**	**0.962**	**0.989**	**0.951**	**0.971±0.0031**

4. Murel: Multimodal relational reasoning for visual question answering. In: CVPR (2019)
5. Learning memory-guided normality for anomaly detection. In: CVPR (2020)
6. Reasoning with heterogeneous graph alignment for video question answering. In: AAAI (2020)

Table 11: Detailed ablation results of AUPRO for the component IFC and CLC.

Component IFC	Component CLC	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
✗	✗	0.968	0.925	0.979	0.914	0.909	0.948	0.975	0.976	0.967	0.965	0.953±0.0011
✗	✓	0.979	0.968	0.980	0.937	0.919	0.945	0.976	0.979	0.970	0.976	0.963±0.0025
✓	✗	0.985	0.967	0.979	0.940	0.923	0.942	0.971	0.980	0.972	0.979	0.964±0.0020
✓	✓	**0.986**	**0.974**	**0.981**	**0.946**	**0.925**	**0.941**	**0.983**	**0.983**	**0.974**	**0.983**	**0.968±0.0016**

Table 12: Detailed investigation of I-AUROC on the components of IFC.

Component $\mathcal{L}_{GC}$	Component $\mathcal{L}_{LC}$	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
✗	✗	0.981	0.831	0.980	0.985	**0.960**	0.905	0.936	**0.964**	0.967	0.780	0.929±0.0024
✓	✗	0.986	0.871	0.979	0.985	0.959	0.932	0.949	0.957	0.970	0.912	0.950±0.0016
✗	✓	0.982	0.886	0.978	0.984	**0.960**	0.936	0.955	0.953	0.971	0.919	0.952±0.0023
✓	✓	**0.987**	**0.892**	**0.981**	**0.986**	**0.960**	**0.937**	**0.959**	0.959	**0.972**	**0.932**	**0.957±0.0041**

Table 13: Detailed investigation of AUPRO on the components of IFC.

Component $\mathcal{L}_{GC}$	Component $\mathcal{L}_{LC}$	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
✗	✗	0.968	0.925	0.979	0.914	0.909	**0.948**	0.975	0.976	0.967	0.965	0.953±0.0011
✓	✗	0.976	0.967	0.979	0.929	0.917	0.942	0.975	0.976	0.967	0.975	0.960±0.0014
✗	✓	0.972	0.965	**0.980**	0.931	0.913	0.943	0.975	0.978	0.969	0.971	0.960±0.0021
✓	✓	**0.979**	**0.968**	**0.980**	**0.937**	**0.919**	0.945	**0.976**	**0.979**	**0.970**	**0.976**	**0.963±0.0025**

Table 14: Detailed investigation of I-AUROC on the components of CLC.

Component L_{GA}	Component L_{LA}	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
✗	✗	0.981	0.831	**0.980**	0.985	0.960	0.905	0.936	**0.964**	0.967	0.780	0.929±0.0024
✗	✓	0.985	0.862	0.979	0.987	**0.961**	0.930	0.961	0.961	0.969	0.893	0.949±0.0015
✓	✗	0.986	0.874	**0.980**	0.986	0.960	0.935	0.958	0.960	0.968	0.912	0.952±0.0022
✓	✓	**0.989**	**0.901**	**0.980**	**0.988**	0.959	**0.942**	**0.967**	0.960	**0.970**	**0.929**	**0.959±0.0018**

Table 15: Detailed investigation of AUPRO on the components of CLC.

Component L_{GA}	Component L_{LA}	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
✗	✗	0.968	0.925	**0.979**	0.914	0.909	**0.948**	**0.975**	0.976	0.967	0.965	0.953±0.0011
✗	✓	0.972	0.966	0.978	0.939	0.920	0.941	0.969	0.979	0.966	0.978	0.961±0.0012
✓	✗	0.976	0.963	0.978	0.936	0.918	0.945	0.970	**0.980**	0.967	0.976	0.961±0.0019
✓	✓	**0.985**	**0.967**	**0.979**	**0.940**	**0.923**	0.942	0.971	**0.980**	**0.972**	**0.979**	**0.964±0.0020**

7. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In: ICCV (2021)
8. Multimodal industrial anomaly detection via hybrid fusion. In: CVPR (2023)
9. Real3d-ad: A dataset of point cloud anomaly detection. arXiv (2023)

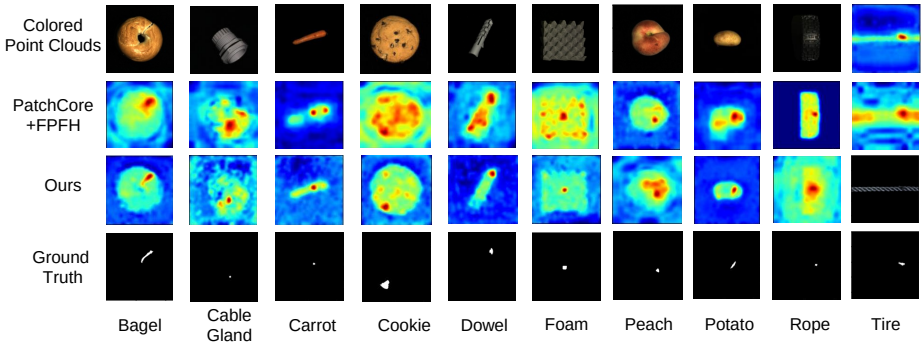

Fig. 4: Visualization of anomaly segmentation results for all the categories. Compared with PatchCore+FPFH, our LSFA consistently performs more precise defect localization.

10. Bergmann, P., Jin, X., Sattlegger, D., Steger, C.: The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In: VISIGRAPP (2022)
11. Bonfiglioli, L., Toschi, M., Silvestri, D., Fioraio, N., De Gregorio, D.: The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In: ACCV (2022)
12. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
13. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM (1981)
16. Horwitz, E., Hoshen, Y.: An empirical investigation of 3d anomaly detection and segmentation. arXiv preprint arXiv:2203.05550 (2022)
17. Horwitz, E., Hoshen, Y.: An empirical investigation of 3d anomaly detection and segmentation. arXiv preprint arXiv:2203.05550 (2022)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
19. Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: ECCV (2022)
20. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: CVPR (2022)
21. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: Asymmetric student-teacher networks for industrial anomaly detection. arXiv preprint arXiv:2210.07829 (2022)