[Supplementary Material] Learning Diffusion Models for Multi-View Anomaly Detection

Chieh Liu[1], Yu-Min Chu[1], Ting-I Hsieh[1], Hwann-Tzong Chen[1,2], and Tyng-Luh Liu[3]

[1] National Tsing Hua University, Taiwan
[2] Aeolus Robotics, Taiwan
[3] Institute of Information Science, Academia Sinica, Taiwan

A Experimental Results on MVTec 3D-AD

A.1 Settings of MVTec 3D-AD Dataset

The MVTec 3D-AD dataset [1] is different from the Eyecandies dataset [2]. Each instance in the MVTec 3D-AD dataset contains a single view RGB image and the corresponding 3D point cloud data. We simply convert the point cloud data into a normal map using the open3d software and then fine-tune the UNet and ControlNet in the same configurations. Since there is only a single image view in this case, we omit the feature loss during training. We update the weight of the pre-trained UNet, which is the Stable Diffusion Model version 1-5 utilized in the same architecture. In the inference phase, we set the noise intensity to 40 and the step size to 20, and we use the concatenated features from the decoder block-1 and block-2 of the UNet. In other methods such as BTF [6], M3DM [7], Shape-Guided [3], etc., the anomaly score for image level typically implemented a reweight mechanism between RGB score and 3D score. Therefore, our approach only performs a weighted sum of RGB and 3D image level scores. An important difference is that we do not adjust the score map.

A.2 Comparisons with Other Methods

We compare our approach with the state-of-the-art methods and evaluate them on the MVTec 3D-AD dataset [1] using the Img-AUROC and the Pix-AUROC metrics. Because current reconstruction-based methods CFM [4] and MMRD [5] typically require training a model for each category rather than employ a single model for all classes. In addition, CFM requires different backbones to handle 2D RGB and 3D data. Our approach is competitive to reconstruction-based methods and shows strong capabilities across various datasets even without using the key technique of feature loss. In Table 1, our method achieves above-average Img-AUROC scores in each category, which allows us to attain an overall competitive average score. In Table 2, the Pix-AUROC score approaches near-perfect performance. Even though the MVTec 3D-AD provides only a single image view, our method performs especially well.

B Qualitative Comparisons

Fig. 1 illustrates the heatmaps for comparing our proposed method with state-of-the-art. In the first row, M3DM [7] and BTF [6], tend to classify the entire `CandyCane` as an abnormal area. On the contrary, our method identifies abnormal regions accurately without false positives. In the second and third rows, BTF [6] struggles to detect 3D geometric anomalies on the surfaces of `GummyBear` and `Confetto`. Furthermore, in the fourth row (`ChocolateCookie`), our method shows a lower false alarm rate than the other two approaches. In the last row, our method effectively detects anomalies in multiple regions of `PeppermintCandy`. By these samples in Fig. 1, our approach can detect anomalies accurately compared with other embedding-based methods.

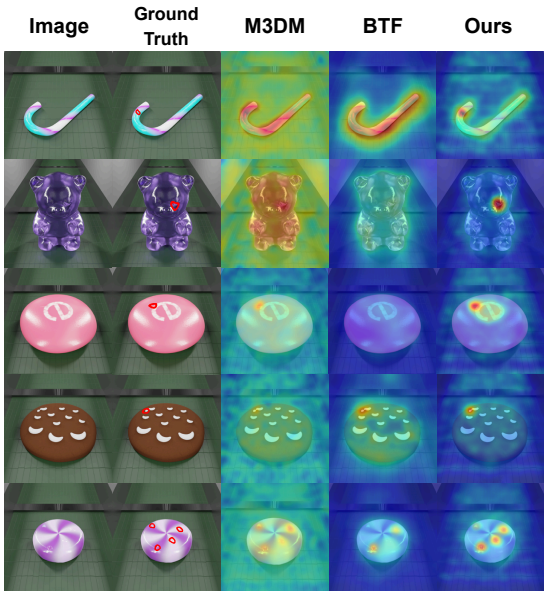

Fig. 1: Visualization of ours and other methods.

C More Qualitative Results

Fig. 2 displays additional visualizations of our anomaly localization results for all categories in the Eyecandies [2]. Our prediction in the score map closely matches the ground truth, indicating our proposed method's effectiveness and precision. Additionally, our approach excels in identifying 3D geometric anomalies that cannot be distinguished solely based on appearance colors. This capability is attributed to our method successfully integrating information from the normal

map to pinpoint these abnormal regions. Regarding the MVTec 3D-AD [1], Fig. 3 illustrates that our method also possesses capabilities to detect anomalies within this dataset, regardless of whether these anomalies are RGB-based or 3D-based.

Table 1: Assessing anomaly detection performance on the MVTec 3D-AD dataset [1] using the Img-AUROC metric. "SG" means Shape-Guided [3]. The top-performing outcomes are highlighted in red, while the second-best results are indicated in blue.

	Method	Bagel	Cable -Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	**Mean**
RGB	BTF [6]	0.854	0.840	0.824	0.687	0.974	0.716	0.713	0.593	0.920	0.724	0.785
	SG [3]	0.911	0.936	0.883	0.662	0.974	0.772	0.785	0.641	0.884	0.706	0.815
	M3DM [7]	0.944	0.918	0.896	0.749	0.959	0.767	0.919	0.648	0.938	0.767	0.850
	Ours	0.910	0.871	0.854	0.687	0.908	0.859	0.885	0.556	0.914	0.720	0.816
3D	BTF [6]	0.820	0.533	0.877	0.769	0.718	0.574	0.774	0.895	0.990	0.582	0.753
	M3DM [7]	0.941	0.651	0.965	0.969	0.905	0.760	0.880	0.974	0.926	0.765	0.874
	SG [3]	0.983	0.682	0.978	0.998	0.960	0.737	0.993	0.979	0.966	0.871	0.916
	Ours	0.965	0.852	0.962	0.988	0.963	0.900	0.990	0.984	0.965	0.823	0.939
3D+RGB	BTF [6]	0.938	0.765	0.972	0.888	0.960	0.664	0.904	0.929	0.982	0.726	0.873
	M3DM [7]	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	0.945
	SG [3]	0.986	0.894	0.983	0.991	0.976	0.857	0.990	0.965	0.960	0.869	0.945
	MMRD [5]	0.999	0.943	0.964	0.943	0.992	0.912	0.949	0.901	0.994	0.901	0.950
	CFM [4]	0.988	0.875	0.984	0.992	0.997	0.924	0.964	0.949	0.979	0.950	0.960
	Ours	0.967	0.880	0.964	0.984	0.963	0.978	0.990	0.973	0.960	0.829	0.949

Table 2: Assessing anomaly detection performance on the MVTEC 3D-AD dataset [1] using the Pix-AUROC metric. "SG" means Shape-Guided [3]. The top-performing outcomes are highlighted in red, while the second-best results are indicated in blue.

	Method	Bagel	Cable -Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	**Mean**
RGB	BTF [6]	0.983	0.984	0.980	0.974	0.973	0.851	0.976	0.983	0.987	0.977	0.967
	SG [3]	0.990	0.993	0.990	0.981	0.992	0.927	0.984	0.987	0.995	0.991	0.983
	M3DM [7]	0.992	0.990	0.994	0.977	0.983	0.955	0.994	0.990	0.995	0.994	0.987
	Ours	0.990	0.985	0.988	0.976	0.990	0.957	0.992	0.984	0.997	0.977	0.984
3D	M3DM [7]	0.981	0.949	0.997	0.932	0.959	0.925	0.989	0.995	0.994	0.981	0.970
	BTF [6]	0.995	0.965	0.999	0.947	0.966	0.928	0.996	0.999	0.996	0.991	0.978
	SG [3]	0.992	0.961	0.998	0.948	0.960	0.931	0.996	0.999	0.995	0.996	0.978
	Ours	0.996	0.981	0.998	0.976	0.990	0.897	0.998	0.997	0.997	0.980	0.981
3D+RGB	M3DM [7]	0.995	0.993	0.997	0.985	0.985	0.984	0.996	0.994	0.997	0.996	0.992
	BTF [6]	0.996	0.992	0.997	0.994	0.981	0.973	0.996	0.998	0.994	0.995	0.992
	MMRD [5]	-	-	-	-	-	-	-	-	-	-	0.992
	CFM [4]	0.997	0.992	0.999	0.972	0.987	0.993	0.998	0.999	0.998	0.998	0.993
	SG [3]	0.996	0.993	0.998	0.992	0.992	0.997	0.997	0.999	0.998	0.997	**0.996**
	Ours	0.996	0.988	0.998	0.988	0.992	0.994	0.998	0.998	0.997	0.996	**0.994**

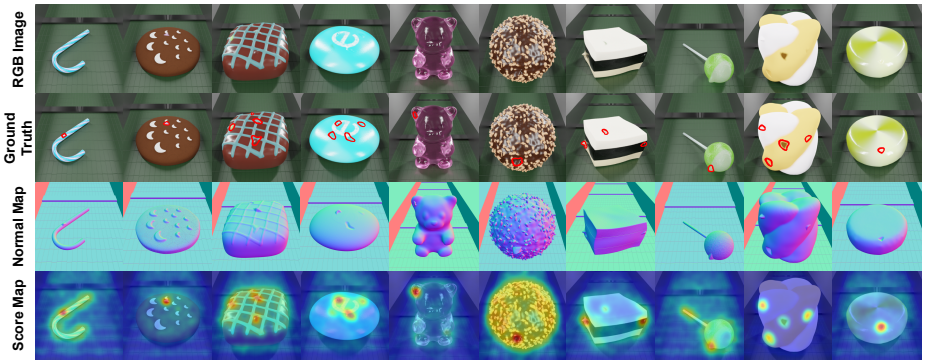

Fig. 2: Our anomaly localization results on Eyecandies dataset [2].

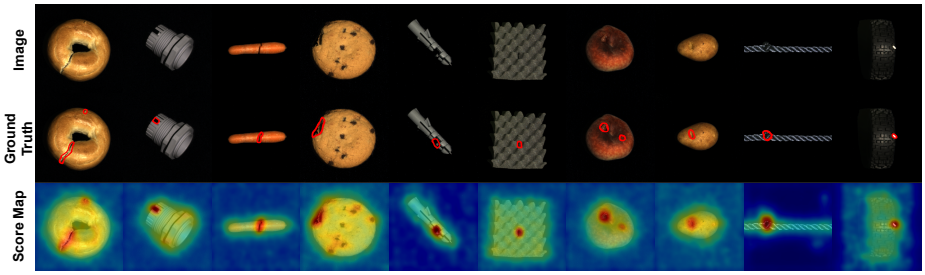

Fig. 3: Our anomaly localization results on MVTec 3D-AD dataset [1].

D Multi-view Anomaly Detection

Table 3 shows that our approach, in conjunction with the 3D normal map, performs more effectively when six views are used simultaneously compared to a single view. This improvement is due to the multi-view configuration's ability to more effectively detect anomalies that might only be visible under certain lighting conditions.

E Computational Cost of ControlNet

In Table 4, we examine the parameter count (Params), inference speed (FPS), and the anomaly detection scores(Img-ROC) for UNet and UNet+ControlNet on the Nvidia RTX 3090Ti. ControlNet utilizes fewer than half the Params of UNet, without notably reducing FPS. Furthermore, ControlNet facilitates the generation of more expressive features.

Table 3: Comparison between all-view and single-view.

View	All	1	2	3	4	5	6
Img-ROC	**0.948**	0.936	0.929	0.933	0.923	0.936	0.938

F Applying Feature Loss at Different Layers

Table 5 analyzes feature consistency loss impacts among different decoder blocks. This analysis indicates that applying the loss to all blocks improves performance compared to using it on individual blocks or not using it at all.

Table 4: Comparison: UNet(U) vs. UNet+ControlNet ($U+C$).

Model	Params	FPS	Img-ROC
U	860M	**0.225**	0.934
$U+C$	1222M	0.196	**0.948**

Table 5: Feature consistency loss for different decoder blocks.

Block	wo Loss	1	4	All
Img ROC	0.934	0.937	0.932	**0.948**

G Detailed Hyperparameter Settings

Table 6 shows that different top-k values only slightly impact on AD scores. Following the settings in [19, 28], we experiment with different step sizes and report the Img-ROC results and their FPS in Table 7.

Table 6: Different Top-k scores with Image-ROC.

Top-k	$k=1$	$k=3$	$k=5$	$k=7$
Img ROC	0.945	**0.948**	0.944	0.942

Table 7: Diffusion step sizes with Image-ROC and frames per second (FPS).

Step Size	10	20	40
Img-ROC	0.946	**0.948**	0.944
FPS	0.169	0.196	**0.205**

References

1. Bergmann, P., Jin, X., Sattlegger, D., Steger, C.: The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In: Farinella, G.M., Radeva, P., Bouatouch, K. (eds.) VISIGRAPP (2022) 1, 3, 4
2. Bonfiglioli, L., Toschi, M., Silvestri, D., Fioraio, N., Gregorio, D.D.: The eyecandies dataset for unsupervised multimodal anomaly detection and localization (2022) 1, 2, 4

3. Chu, Y., Liu, C., Hsieh, T., Chen, H., Liu, T.: Shape-guided dual-memory learning for 3d anomaly detection (2023) 1, 3
4. Costanzino, A., Zama Ramirez, P., Lisanti, G., Di Stefano, L.: Multimodal industrial anomaly detection by crossmodal feature mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024), cVPR 1, 3
5. Gu, Z., Zhang, J., Liu, L., Chen, X., Peng, J., Gan, Z., Jiang, G., Shu, A., Wang, Y., Ma, L.: Rethinking reverse distillation for multi-modal anomaly detection. Proceedings of the AAAI Conference on Artificial Intelligence (2024) 1, 3
6. Horwitz, E., Hoshen, Y.: Back to the feature: Classical 3d features are (almost) all you need for 3d anomaly detection (2023) 1, 2, 3
7. Wang, Y., Peng, J., Zhang, J., Yi, R., Wang, Y., Wang, C.: Multimodal industrial anomaly detection via hybrid fusion (2023) 1, 2, 3