DA-CLIP: Mitigating Granularity Mismatch in Zero-Shot Anomaly Detection via Decoupled Text-Visual Alignment

Supplementary Material

1. Supplementary

This Supplementary Material contains the following three parts: 1) Additional Information about text prompt, SOTA models, datasets, implementation details and details of Mamba; 2) Additional analysis about decoupled alignment and token adapter; 3) Additional Experiments about hyper-parameters, SOTA methods and qualitative results.

2. Additional Information

2.1. Text Prompt

Our text prompt templates follow the same format as prior work [10]. However, during both training and inference, we do not replace "[CLS]" with specific class names. Instead, we employ an object-agnostic prompt, using the single term "object" uniformly to describe all categories.

Table 1. Templates of Text Prompts

State	Prompt
Prompt Template	{} a photo of a {}
Normal Descriptors	[CLS] a [CLS] the [CLS]
Anomaly Descriptors	damaged [CLS] broken [CLS] [CLS] with flaw [CLS] with defect [CLS] with damage

2.2. SOTA models

To demonstrate the superiority of DA-CLIP, we compare our methods with baseline CLIP [13]. The introductions are given as follows:

- **CLIP** [13], is a multimodal model introduced by OpenAI in 2021, renowned for its powerful zero-shot classification capabilities. It can classify images directly without any dataset-specific training. To extend CLIP to the task of anomaly segmentation, we We use the prompts in APRIL-GAN [2] to derive the segmentation.
- **WinCLIP** [7], published in CVPR 2023, enhances CLIP with a compositional ensemble of prompt templates and an efficient feature extraction method across image patches. It efficiently extracts and aggregates window/patch/image-level features to align with text, making progress in anomaly detection and segmentation tasks. The results for WinCLIP in our paper are referenced from [16].
- **APRIL-GAN** [2] is a winning approach for the CVPR 2023 Zero/Few-shot Track of the Visual Anomaly and Novelty Detection (VAND) 2023 Challenge. APRIL-GAN aligns local visual features and text embeddings using a simple linear projection, resulting in a significant improvement in segmentation performance. The results for APRIL-GAN in our paper are obtained using the official weights
- **AnomalyCLIP** [16] is published in ICLR 2024, which is the first work to introduce learnable prompts for ZSAD. It proposes an object-agnostic text prompt that directs CLIP's attention toward anomalies rather than the objects themselves. The results in our paper are obtained using the official weights.
- **TPS** [9] is published in AAAI 2025, which introduces fine-grained shunt text prompts to enable pixel-level precise alignment between the textual cues and the segmentation task. The results in our paper are obtained using the official weights.
- **Bayes-PFL** [12] is published in CVPR 2025, models the text prompt space as a probabilistic distribution and applies regularization to this space in order to improve generalization to novel categories. The results in our paper are obtained using the official weights.

2.3. Datasets

We conduct extensive experiments on 11 public datasets from industrial and medical domains, covering five modalities. For each dataset, we use only the test set. Relevant details of the datasets are presented in Tab.2.

2.4. Implementation Details

During training, we follow the same data preprocessing pipeline as AnomalyCLIP [16] and keep all CLIP parameters frozen. We integrate our Token Adapters into every layer of the text encoder. We allocate 10 learnable tokens at each layer of the text encoder. At inference time, we apply a top-k selection strategy by retaining the top-80 scoring outputs.

2.5. Details of Mamba

State-space models (SSMs) [4] has recently emerged as a compelling paradigm for sequence modeling, offering a

Table 2. Key Statistics of The Datasets Used

Dataset	Domain	Modalities	$\|C\|$	Number of Samples	
				Normal	Anomaly
MVTec-AD	Industrial	Photography	15	467	1258
VisA	Industrial	Photography	12	962	1200
MPDD	Industrial	Photography	6	176	282
BTAD	Industrial	Photography	3	451	290
DTD	Industrial	Photography	12	357	947
Brain MRI	Medical	MRI	1	640	1013
HeadCT	Medical	CT	1	833	660
Br35H	Medical	OCT	1	1041	764
ClinicDB	Medical	Endoscopy	1	0	612
ISIC	Medical	Endoscopy	1	0	1000
MSD	Medical	CT	1	0	577

Table 3. AnomalyCLIP with Decoupled Alignment (DA).

Method	Pixel-Level		Image-Level	
	AUROC	PRO	AUROC	AP
AnomalyCLIP	95.4(-0.1)	**86.0**(+0.5)	82.1(-1.5)	85.4(-0.5)
AnomalyCLIP + DA	**95.5**	85.5	**83.6**	**85.9**

Table 4. The Role of Residual Connections in the Token Adapter.

Method	Pixel-Level		Image-Level	
	AUROC	PRO	AUROC	AP
w/o residual	95.3(-0.2)	91.4(-0.3)	87.8(-0.7)	90.0(-0.4)
w/ residual	**95.5**	**91.7**	**88.5**	**90.4**

promising alternative to established architectures such as transformers and CNNs in visual understanding tasks. With the introduction of the S4 model [14], SSMs have been widely adopted across computer vision [5], medical imaging [8], and anomaly detection [6] applications. Building on S4, Mamba [3] integrates a selective scanning mechanism and hardware-aware design to achieve linear scalability for long-sequence processing. [15] have adopted Mamba as a bridging module to jointly fine-tune the vision foundation models (VFMs) and the vision-language models (VLMs).

Different from previous modal alignment methods [11, 12], we adopt a novel hybrid attention mechanism. We modify the original Mamba block to encourage the two branches to capture the sequential dynamics and spatial relationships respectively in parallel. The specific details of the Mamba block are as follows:

$$\begin{aligned} x_l^{(ssm)} &= SSM(conv(Linear(x_l))) \\ x_l^{(conv)} &= conv(proj(x_l)) \\ [\Delta x_l^s, \Delta x_l^a] &= proj(x_l^{(ssm)} \odot x_l^{(conv)}) \end{aligned} \tag{1}$$

x_l denotes the input feature to the Mamba block. Δx_l^s and Δx_l^a are the learned feature offsets, $\odot$ denotes the element-wise multiplication.

3. Additional Analysis

3.1. More Visualizations for Decoupled Alignment

To provide a more intuitive visualization of the alignment process, we offer a 3D t-SNE analysis. In the original CLIP, the normal and abnormal text features are entangled, exhibiting an inherent lack of anomaly awareness. Simultaneously, there is a large modality gap between text and visual features, and the heatmaps show instances of contradictory predictions, as shown in Fig. 2(a). While joint alignment solves the text feature entanglement problem, it causes the text features to preferentially align with local visual features, thereby neglecting the alignment with the global visual feature. The modality gap between them is further widened, which limits its classification performance (see Fig. 2(b)). Our DA-CLIP solves this problem well. By decoupling the alignment process into two subspaces to prevent mutual interference, it thoroughly unleashes the capability of ZSAD.

To further validate the effectiveness of our decoupled alignment strategy, we also apply it to AnomalyCLIP under otherwise identical conditions. We fine-tune the model on the MVTec dataset and test it on the VisA dataset, with the results shown in Tab. 3. We find that the decoupled alignment strategy is also compatible with other methods, yielding improvements in both classification and segmentation performance.

3.2. More about Token Adapter

Most prior works [1, 11, 16] have attempted to refine the text space by adding extra learnable tokens into the text encoder, as shown in Fig. 1(a). However, this method faces the following issues: insufficient fine-tuning may result in the learnable tokens not fully adapting to the knowledge of the anomaly domain, while excessive fine-tuning can cause the model to quickly overfit on the auxiliary dataset, which harms its performance on unseen classes.

To solve these problems, we attempt to directly fine-tune the embeddings after they are encoded by the tokenizer, as depicted in Fig. 1(b). We attribute the success of our token adapter to its direct fine-tuning of CLIP's original semantic knowledge. This ensures the model can adapt to the knowledge of the anomaly domain, while the residual connection simultaneously guarantees that the model preserves its original semantic information, as shown in Tab. 4.

4. Additional Experiments

4.1. Additional Ablations for Hyper-parameters

We conduct hyperparameter ablation studies on the VisA dataset. Our main hyperparameters include: 1) the number

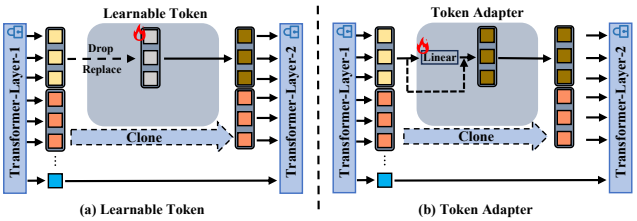

Figure 1. Schematic of Learnable Tokens and Our Token Adapter.

of embedding layers, 2) different values for Top-k strategy, and 3) fusion weight of different anomaly maps for α.

Embedding Layers. We investigate the effect of the layers of inserted Token Adapters on model performance, as shown in Fig. 3(a). Our Token Adapter achieves controllable generalization through intra-token interactions, eliminating the need to worry about overfitting issues caused by an excessive number of layers.

Top-k Strategy. Fig. 3(b) illustrates the effect of different k values on inference performance. We swept k from 1 to 90, achieving the optimal result when k was 80. Note that $k = 1$ corresponds to taking only the maximum value. We observed that this strategy degrades performance due to noise in the anomaly maps. Consequently, we set k to 80 for all subsequent experiments.

Fusion Weight of Different Anomaly Maps. To investigate the impact of the hyperparameter α on our model's performance, we conduct a comprehensive ablation study on two widely-used benchmark datasets: MVTec-AD and VisA. We evaluate its effect on both image level classification performance, measured by Image AUROC, and pixel-level segmentation performance, measured by Pixel PRO. The value of α is varied from 0.3 to 0.8, and the results are presented in Fig. 4. The analysis reveals that setting $\alpha = 0.7$ provides the best trade-off across both metrics and datasets. This value not only maximizes the image-level classification accuracy on MVTec-AD but also ensures high and robust segmentation performance on both benchmarks without the degradation observed at higher values. Therefore, based on this empirical evidence, we selected $\alpha = 0.7$ as the definitive value for all subsequent experiments.

4.2. Additional Comparisons with SOTA Methods

We provide additional visualizations comparing our method with AnomalyCLIP and Bayes-PFL to demonstrate its superiority. The first row displays the ground truth for reference. As shown in Fig. 6, we can discern clear and consistent patterns in the behavior of each method. AnomalyCLIP, as shown in the second row, frequently suffers from false negatives. It tends to produce sparse or weak activations, often failing to detect the full extent of the anomalous region or missing it entirely. Conversely, Bayes-PFL (third row) exhibits the opposite problem, showing a high propensity for false positives. Its activation maps often over-activate, incorrectly highlighting large, non-anomalous areas of the object or background. In contrast, our method achieves more precise localization.

To demonstrate the superiority of our approach, we benchmark DA-CLIP against the latest SOTA methods, AA-CLIP and AdaCLIP. Fig. 5 summarizes the comparison in terms of image-level AUROC and pixel-level PRO, highlighting the advantages of our method.

4.3. Additional Qualitative Results

To demonstrate the powerful performance of DA-CLIP, we provide more visualization results, as shown in Fig.7 - Fig.24. For each figure, the first row shows the input, with anomaly regions highlighted in red. The final row displays the segmentation results generated by DA-CLIP. Our model shows precise localization results in general.

References

[1] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 2

[2] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2(4), 2023. 1

[3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[4] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 1

[5] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025. 2

[6] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024. 2

[7] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 1

[8] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pages 615–625. Springer, 2024. 2

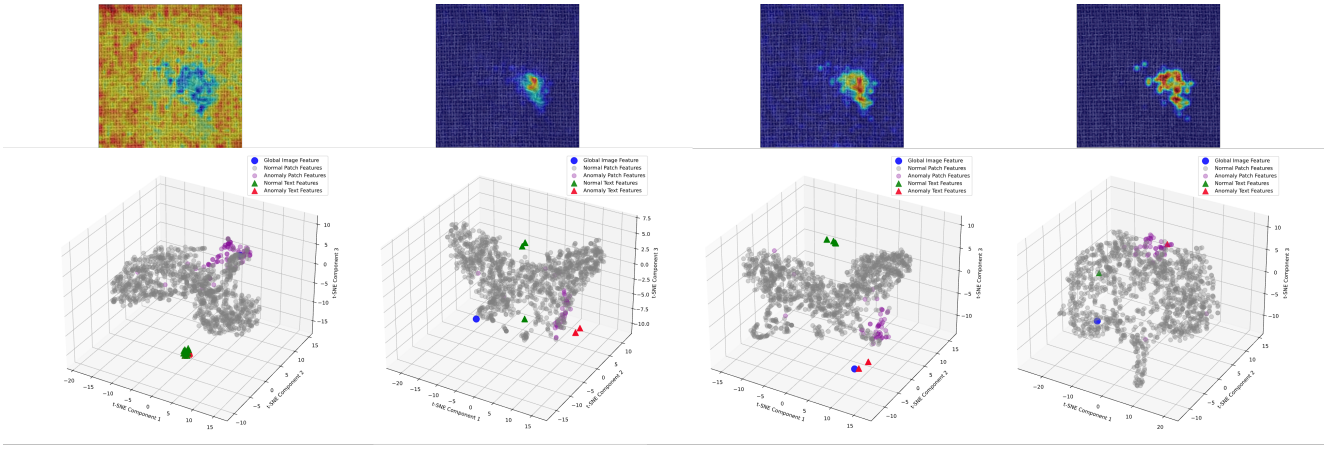

Figure 2. t-SNE Visualization of Text-Visual Alignment.

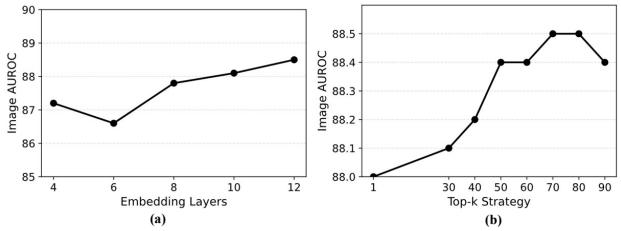

Figure 3. (a) Ablation on layers of inserted token adapters. (b) Ablation on top-k-strategy.

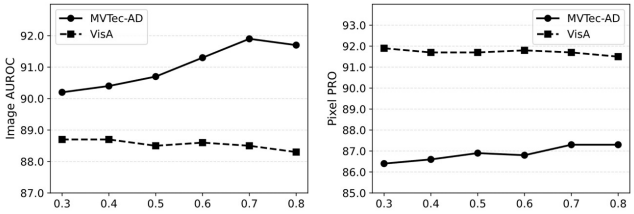

Figure 4. Ablation on the fusion weight of different anomaly maps.

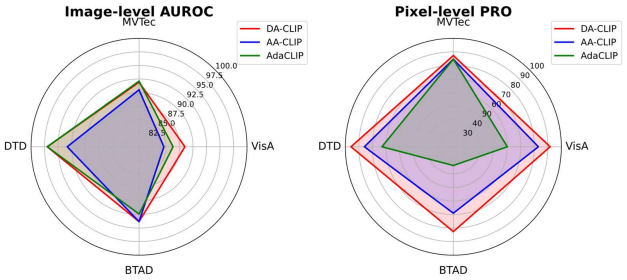

Figure 5. Performance comparison between DA-CLIP and recent SOTA methods.

[9] Jitao Ma, Weiying Xie, Hangyu Ye, Daixun Li, and Leyuan Fang. Aligning and prompting anything for zero-shot generalized anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5964–5972, 2025. 1

[10] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4744–4754, 2025. 1

[11] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, pages 301–317. Springer, 2024. 2

[12] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30398–30408, 2025. 1, 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[14] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 2

[15] Xin Zhang and Robby T Tan. Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14527–14537, 2025. 2

[16] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learn-

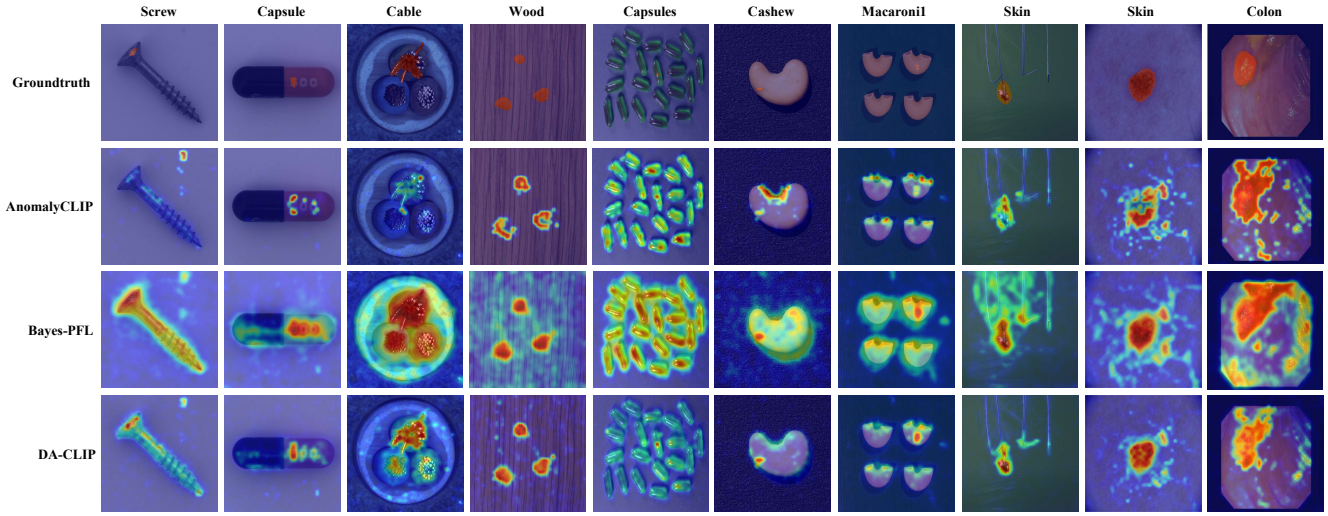

Figure 6. Qualitative comparison of anomaly localization results.

Figure 7. Visualization Examples of Class capsule in MVTec-AD dataset.

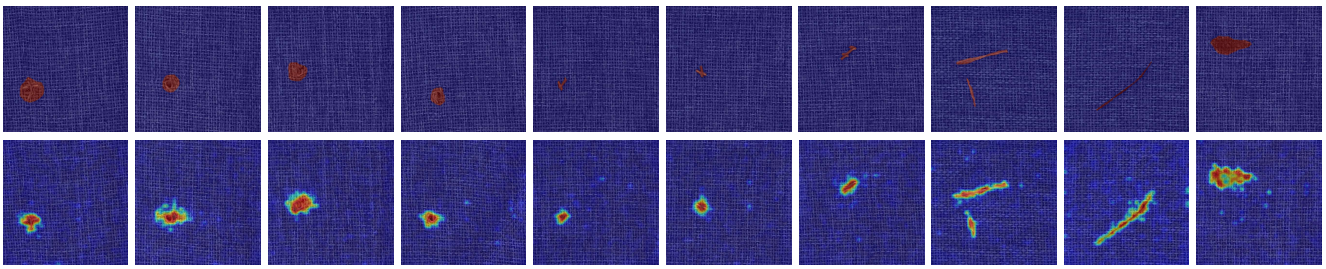

Figure 8. Visualization Examples of Class carpet in MVTec-AD dataset.

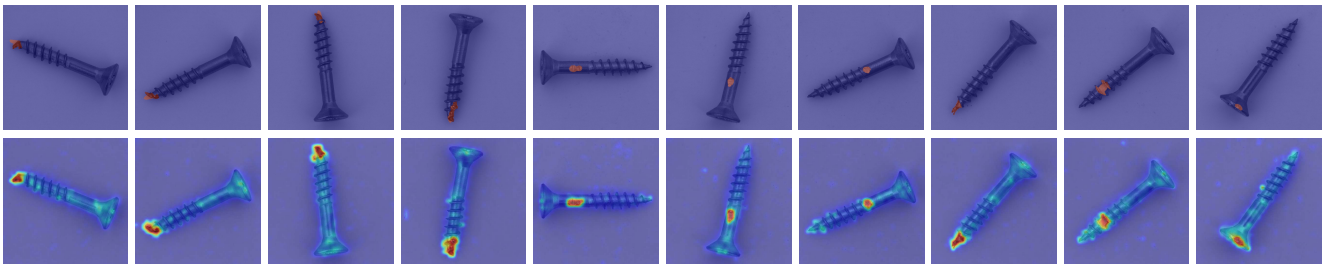

Figure 9. Visualization Examples of Class screw in MVTec-AD dataset.

ing for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 1, 2

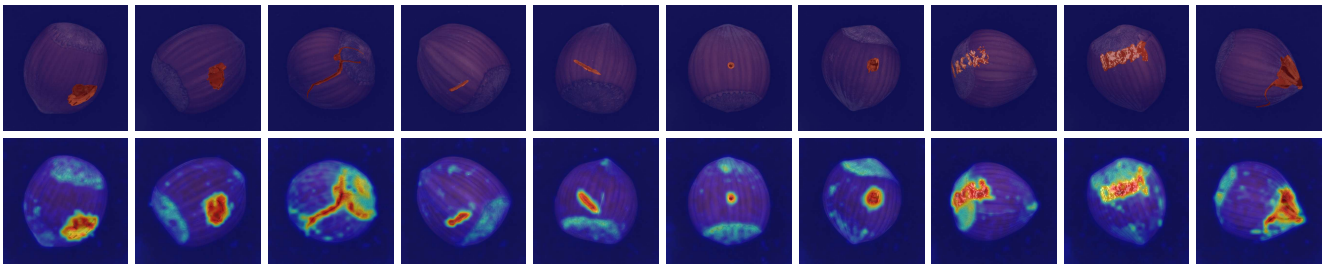

Figure 10. Visualization Examples of Class hazelnut in MVTec-AD dataset.

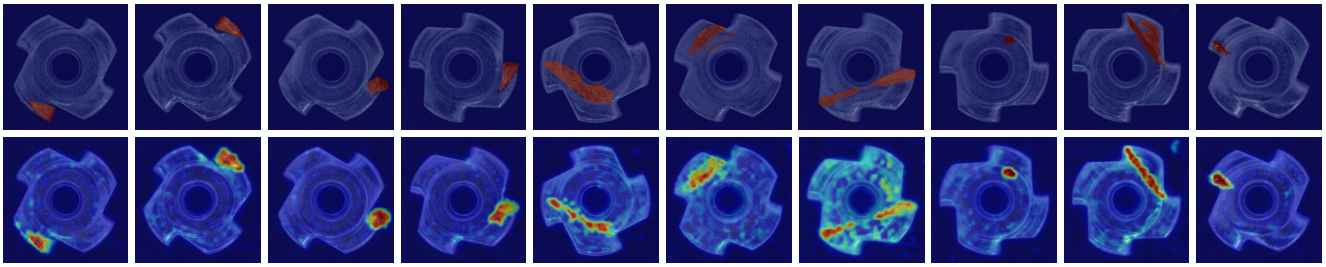

Figure 11. Visualization Examples of Class meta-nut in MVTec-AD dataset.

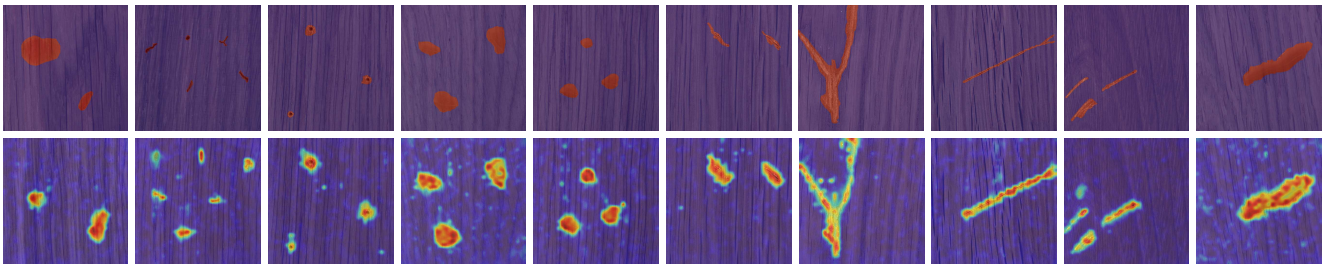

Figure 12. Visualization Examples of Class wood in MVTec-AD dataset.

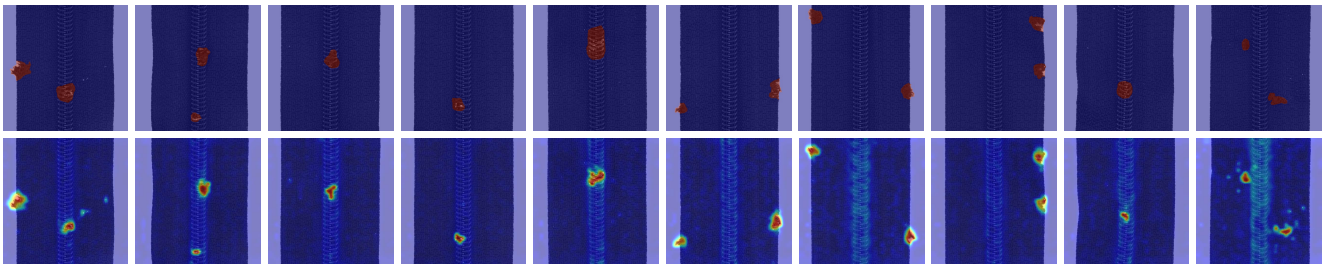

Figure 13. Visualization Examples of Class zipper in MVTec-AD dataset.

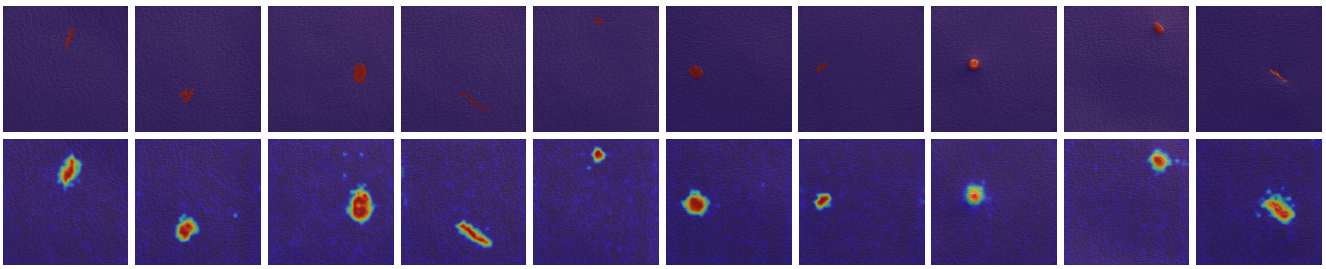

Figure 14. Visualization Examples of Class leather in MVTec-AD dataset.

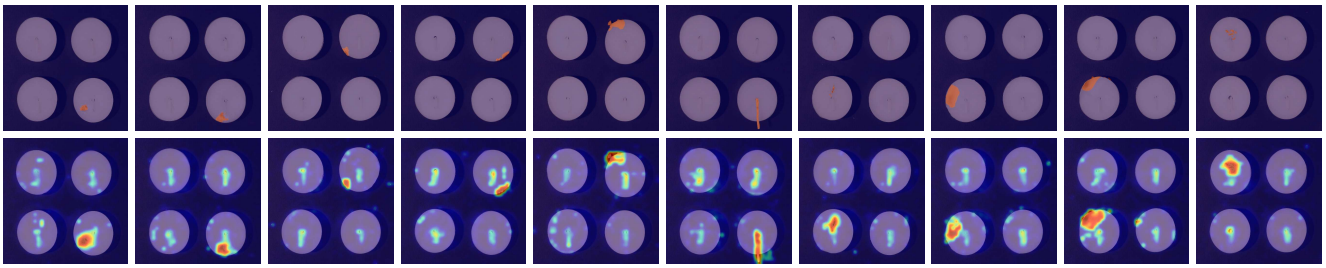

Figure 15. Visualization Examples of Class candle in VisA dataset.

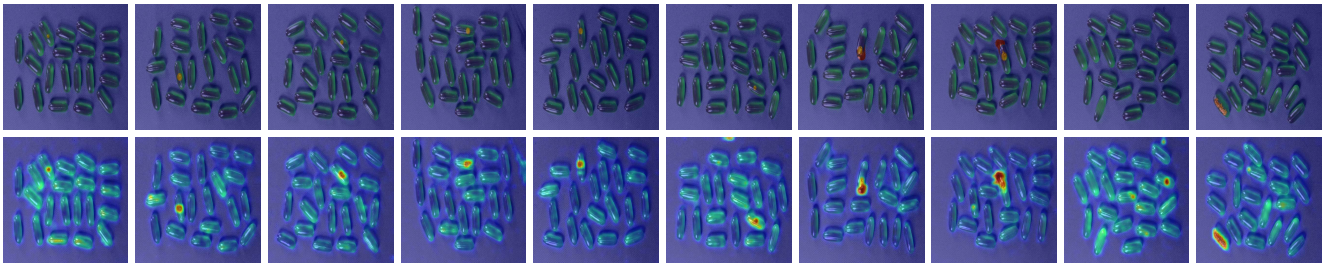

Figure 16. Visualization Examples of Class capsules in VisA dataset.

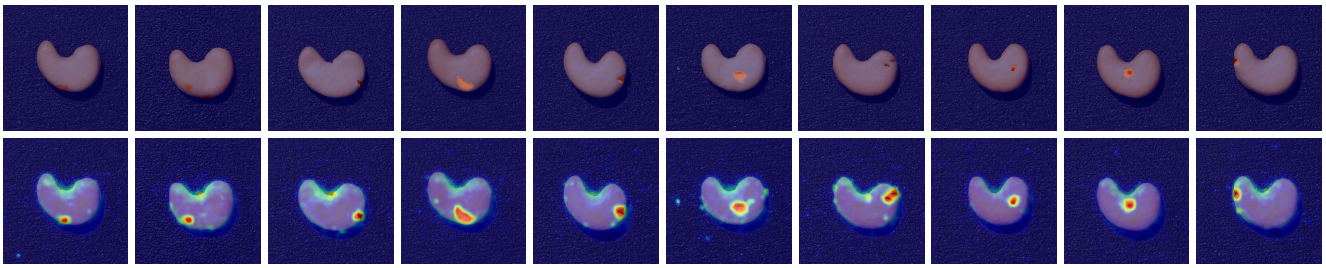

Figure 17. Visualization Examples of Class cashew in VisA dataset.

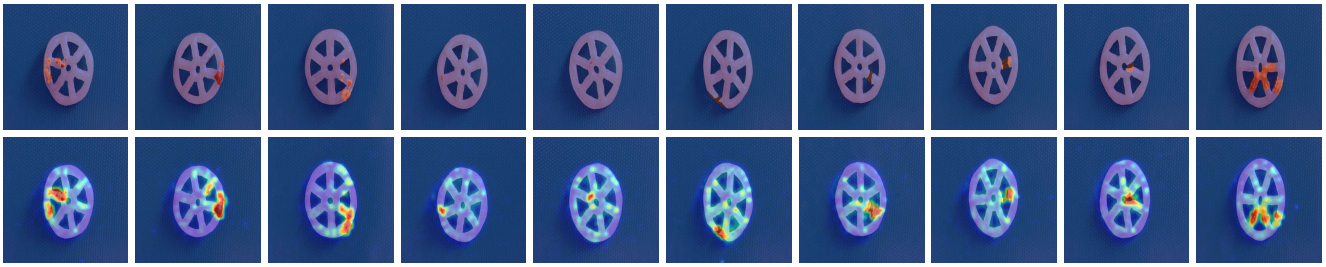

Figure 18. Visualization Examples of Class fryum in VisA dataset.

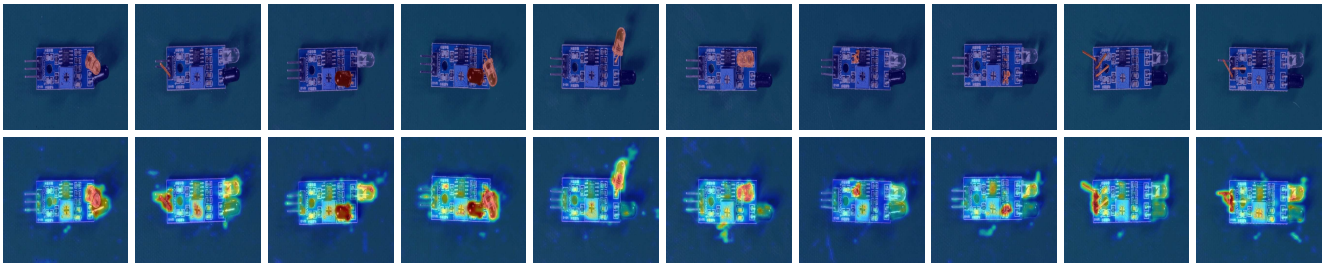

Figure 19. Visualization Examples of Class pcb3 in VisA dataset.

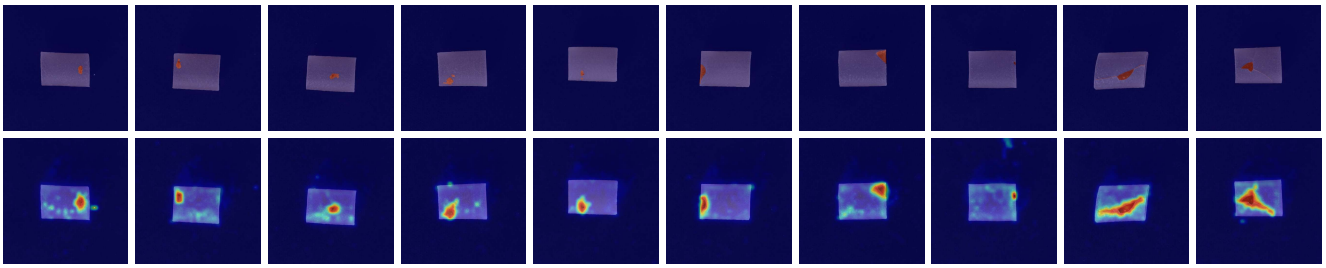

Figure 20. Visualization Examples of Class pipe fryum in VisA dataset.

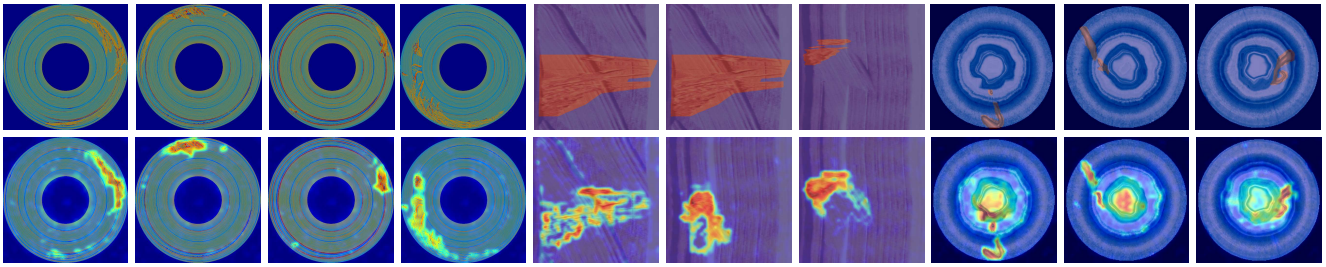

Figure 21. Visualization Examples of Class 01, 02, 03 in BTAD dataset.

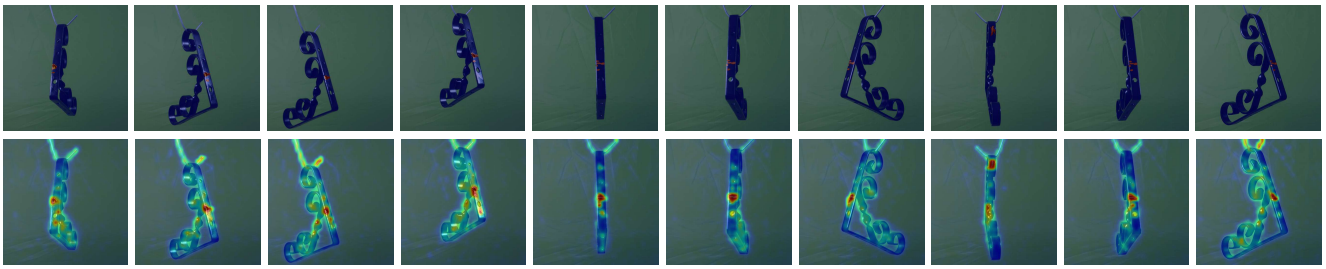

Figure 22. Visualization Examples of Class bracket black in MPDD dataset.

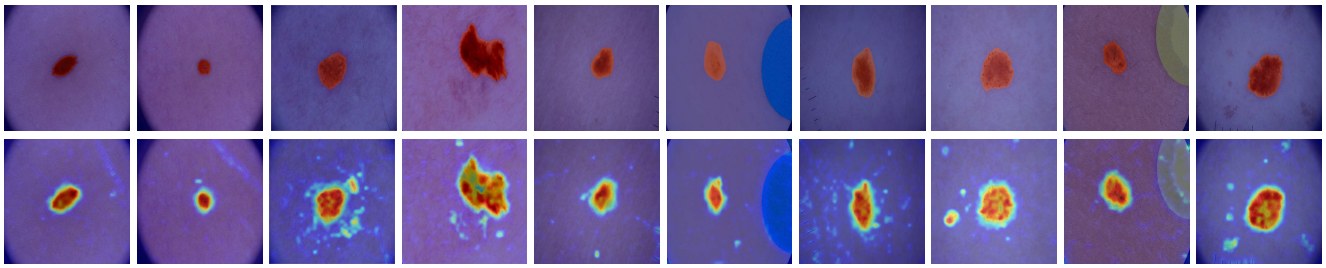

Figure 23. Visualization Examples of Class skin in ISIC dataset.

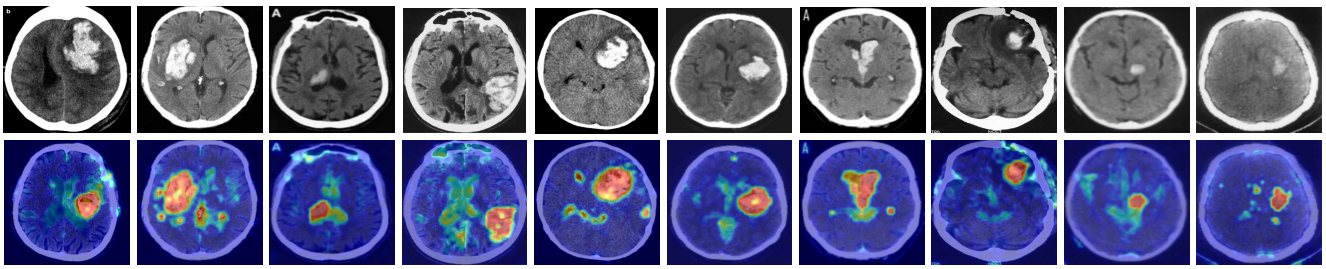

Figure 24. Visualization Examples of Class head in HeadCT dataset.