DLVP-CLIP: Enhancing Fine-Grained Zero-Shot Anomaly Detection via Dynamic Local Visual Prompting

Supplementary Material

This appendix contains the following five sections: 1) Supplementary notes on SOTA models and implementation details in Section A; 2) Supplementary ablation experiments and text t-SNE visualization in Section B; 3) Supplementary notes on the dataset in Section C; 4) Visualization results for various datasets in Section D.

A. Additional Information

A.1. SOTA Models

To validate the superiority of DLVP-CLIP, we compare this method with recent state-of-the-art (SOTA) models. Details are as follows:

- **WinCLIP** [9], was presented at CVPR 2023, enhancing CLIP by combining and integrating prompt templates and adopting an efficient image block feature extraction method. This method efficiently extracts and aggregates window/image-level features to match text content, achieving significant progress in anomaly detection and segmentation tasks. Since its code is not open-source, we use the comparison results provided by AnomalyCLIP. For datasets not compared by AnomalyCLIP, we use the AnomalyCLIP reproduction code for inference.
- **AnomalyCLIP** [17], published at ICLR 2024, is the first study to improve CLIP using prompt learning techniques. The model learns object-independent text prompts to capture general indicators of normal and abnormal states, enabling it to detect anomalies without being constrained by specific object semantics. The experimental results in our paper are based on the original data. For datasets not compared in the original paper, we used the official weights for inference.
- **AdaCLIP** [4], presented at ECCV 2024, combines static and dynamic prompt mechanisms to support shared real-time adaptation, demonstrating strong zero-shot performance and cross-dataset generalization capabilities. Its original implementation includes training on both industrial-grade and medical-grade datasets. We retrained the model using the MVTec dataset with the same parameter settings.
- **TPS** [12], was presented at the AAAI 2025 conference. This method proposes a novel fine-grained text prompt generation strategy to achieve precise image-text alignment. Additionally, it introduces the Text Prompt Splitting (TPS) model, which reconstructs the complementary dependencies between two tasks to enable joint learning, thereby enhancing anomaly detection performance. TPS provides the model weights trained on the MVTec dataset. We use this weight file for inference and validation of the model's performance on other datasets. The original paper does not provide the weight file for inference on MVTec, so we retrained VisA to validate the model's performance on the MVTec dataset.
- **Bayes-PFL** [14] was published in CVPR 2025. It designed a prompt stream module that can simultaneously learn image-specific distributions and image-independent distributions, which work together to normalize the text prompt space and improve the model's generalization ability for unknown categories. All experimental results in this paper are based on official weight parameters.

A.2. Additional Implementation Details

During training, all images were resized to 518×518 pixels. We used the AdaCLIP [4] method to combine four images into one for data augmentation. Additionally, during both training and inference, all datasets were processed using the default normalization settings from OpenCLIP. The depth of the dynamic prompt was set to 9, and the length was set to 4. The hyperparameters for the Adam optimizer are set to $\beta_1 = 0.5$, $\beta_2 = 0.999$, following previous work.

B. Additional Results and Analysis

B.1. Prompt depth ablation.

We conducted ablation experiments on prompt depth, with the results shown in Figure A.6. When the prompt depth is 9, the model achieves optimal performance. Therefore, this work sets the prompt depth to 9.

B.2. Ablation on backbone and input image size.

In Table A.7, we analyze the impact of different pre-trained CLIP backbone networks and input image resolutions on model performance. The results show that larger backbone networks and appropriately increased input resolutions can achieve more accurate pixel-level segmentation. Notably, when the input image resolution is 518^2, the ViT-L-14-336 backbone network achieves the best zero-shot anomaly detection (ZSAD) performance.

B.3. Text t-SNE Visualization.

To validate the model's ability to detect abnormal samples, we used the t-SNE algorithm to perform a visualization analysis of the feature embeddings of the prompt text. Figure A.7 shows the distribution results of various categories

Backbone	Size	MVTec	
		I-AUROC	P-AUROC
ViT-14-224	336^2	89.0	89.1
ViT-14-224	518^2	90.2	89.6
ViT-14-336	336^2	92.4	89.8
ViT-14-336	518^2	94.2	90.4

Table A.7. Ablation on different backbone and input image size.

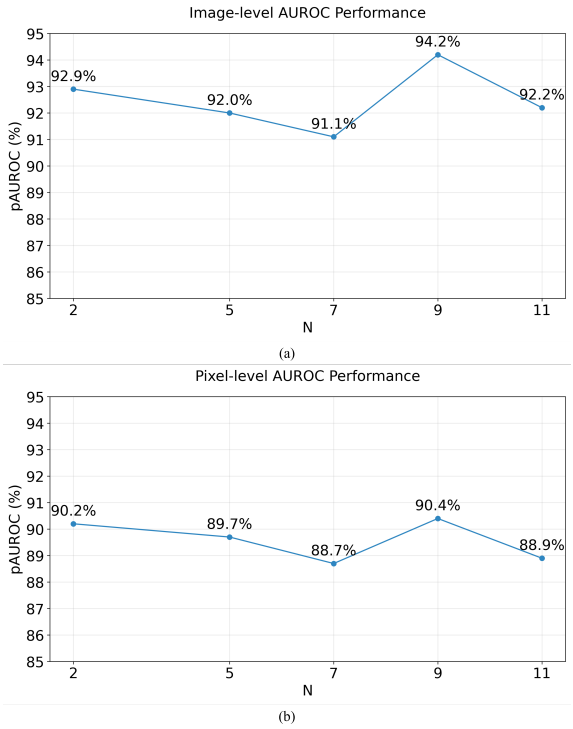

Figure A.6. Prompt depth ablation results comparison.

of VisA, while Figure A.8 presents the situation of Br35H. As shown in FiguresA.7 and A.8, normal samples (blue) and abnormal samples (red) form distinguishable cluster structures in the two-dimensional projection space, indicating that the model can effectively capture the semantic differences between the two types of samples.

C. Datasets

In this section, we provide a brief introduction to the 13 datasets used in this work.

C.1. Industrial Domain

- MVTec-AD [2] is a model specifically designed for industrial anomaly detection, comprising 15 different categories (e.g., bottles, wood). This study only uses its annotated test set, which includes 467 normal images and 1,258 abnormal images. This dataset covers both texture types and object types, making it more comprehensive, and is therefore widely used in our ablation experiments.
- VisA [18] is a highly challenging industrial dataset containing 12 object types (e.g., candles, capsules). Its test set includes 962 normal images and 1,200 abnormal images, which were primarily used for auxiliary training in this study. Since the abnormal regions are relatively small compared to the background, this poses significant challenges for both pixel-level classification and image-level segmentation.
- SDD2 [3] is a dataset specifically designed for industrial

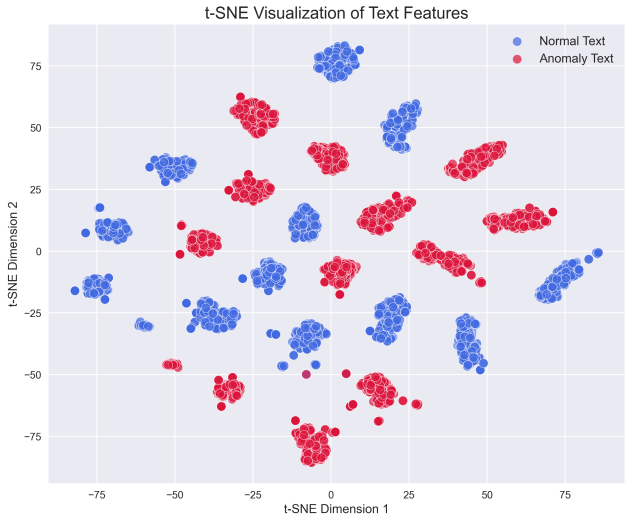

Figure A.7. t-SNE visualization results of text features on the VisA dataset.

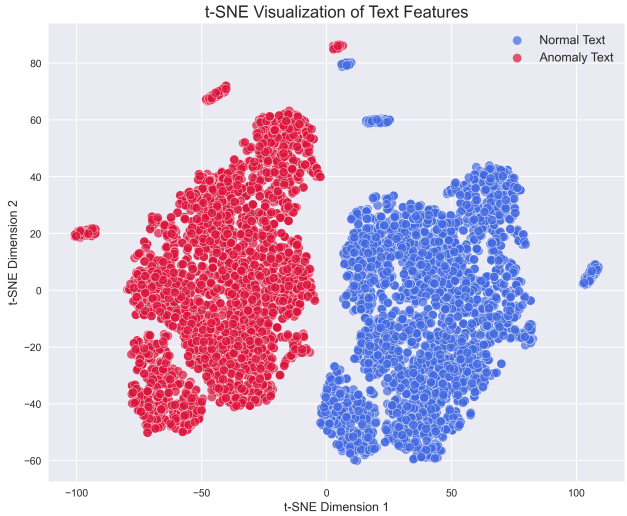

Figure A.8. t-SNE visualization results of text features on the Br35H dataset.

defect detection. The original training set contains 2,085 normal images and 246 abnormal images, while the original test set contains 894 normal images and 110 abnormal images. All images are of similar size, with a width of approximately 230 pixels and a height of approximately 630 pixels.

- DAGM [16] is a texture dataset specifically designed for weakly supervised anomaly detection, comprising 10 categories. The dataset includes 6,996 normal images and 1,054 abnormal images. The original pixel-level annotations are presented as elliptical shapes as weak labels.
- DTD-Synthetic [1] is a synthetic dataset specifically designed for texture anomaly detection, comprising 12 categories. The dataset includes 357 normal images and 947 abnormal images.
- BTAD [13] contains three categories, all of which are object types, with a resolution range of 600 to 1600. This dataset includes 451 normal images and 290 abnormal images, used to evaluate the performance of ZSAD.

C.2. Medical Domain

- HeadCT [15] is a dataset for analyzing head CT scans, which contains image data of various brain lesions (such as hemorrhage and tumors), and is specifically used for anomaly detection. This dataset includes 100 normal images and 100 abnormal images. This study directly adopted the dataset organized by AdaCLIP [4].
- BrainMRI [11] is a dataset for brain magnetic resonance imaging analysis, containing healthy and abnormal brain scan images, including conditions such as tumors and lesions. The dataset includes 98 abnormal images and 155 normal images, with only image-level labels provided. In this study, we directly adopted the dataset compiled by AdaCLIP [4].
- Br35H [6] is a dataset for brain tumor detection in MRI images, containing 1,500 normal images and 1,500 abnormal images. Since this dataset only provides image-level labels, it is only used for abnormal classification tasks. In this study, we directly adopted the dataset compiled by AdaCLIP [4].
- ISIC [5] is a dataset for analyzing skin lesions in endoscopic images, containing a large number of dermatoscopic images, each annotated as either melanoma or non-melanoma lesions. The dataset includes 379 abnormal images with pixel-level annotations, suitable only for abnormal segmentation tasks. In this study, we directly adopted the dataset compiled by AdaCLIP [4].
- Endo [7] is a dataset containing 200 abnormal images with pixel-level annotations. It is used for colon polyp detection. In this study, we directly adopted the dataset compiled by AdaCLIP [4].
- Kvasir [10] is a dataset specifically designed for detecting colon polyps in endoscopic images. The dataset contains 1,000 abnormal images with pixel-level annotations, and this study applies it to abnormal segmentation tasks in the medical field. In this study, we directly adopted the dataset compiled by AdaCLIP [4].
- The RESC dataset [8] provides pixel-level segmentation labels to define the areas affected by retinal edema. We used abnormal images for validation.

D. Anomaly Maps Visualization

We have provided more examples of DLVP-CLIP predictions, as shown in Figures A.9 to A.23. In each figure, the first row shows the input image, the second row highlights the abnormal areas in red, and the last row displays the segmentation result generated by DLVP-CLIP. Overall, our model can provide accurate positioning results.

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

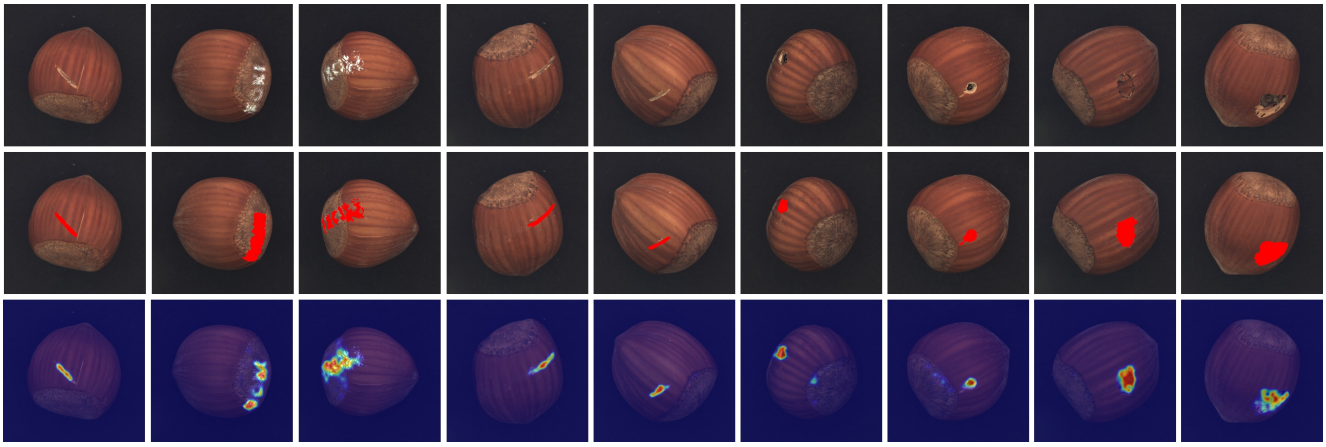

Figure A.10. Visualization Examples of Class Hazelnut in MVTec-AD dataset.

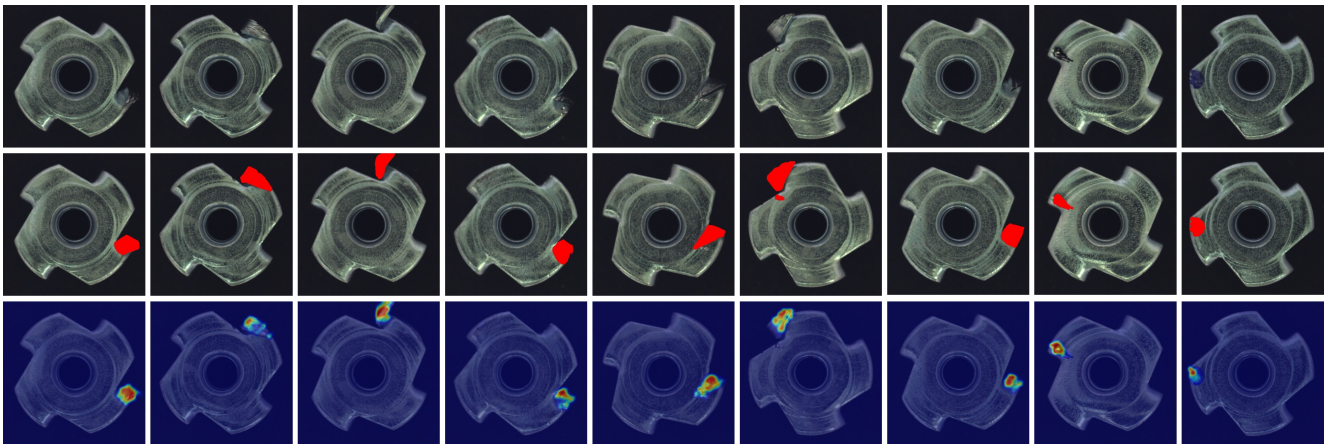

Figure A.11. Visualization Examples of Class Metalnut in MVTec-AD dataset.

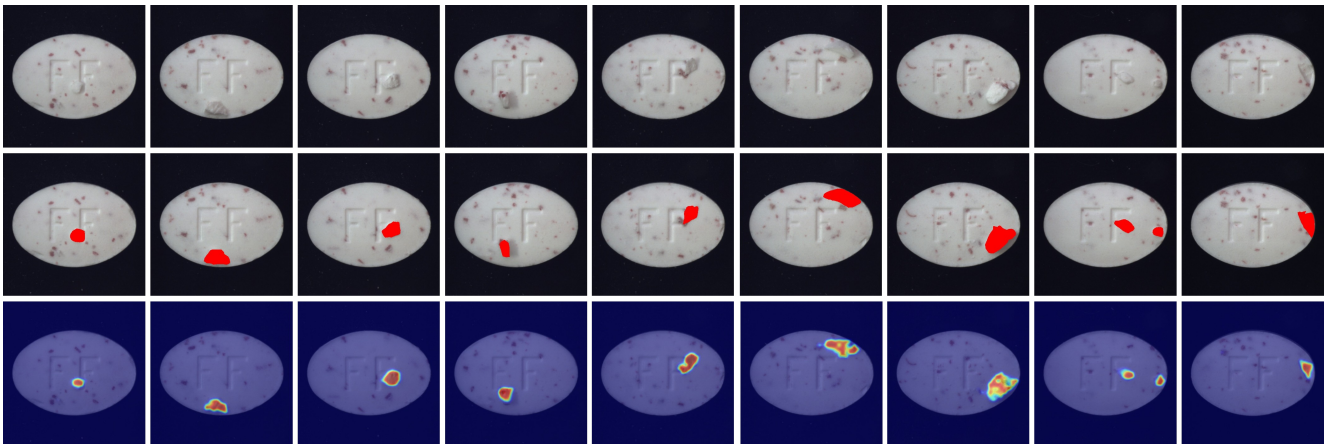

Figure A.12. Visualization Examples of Class Pill in MVTec-AD dataset.

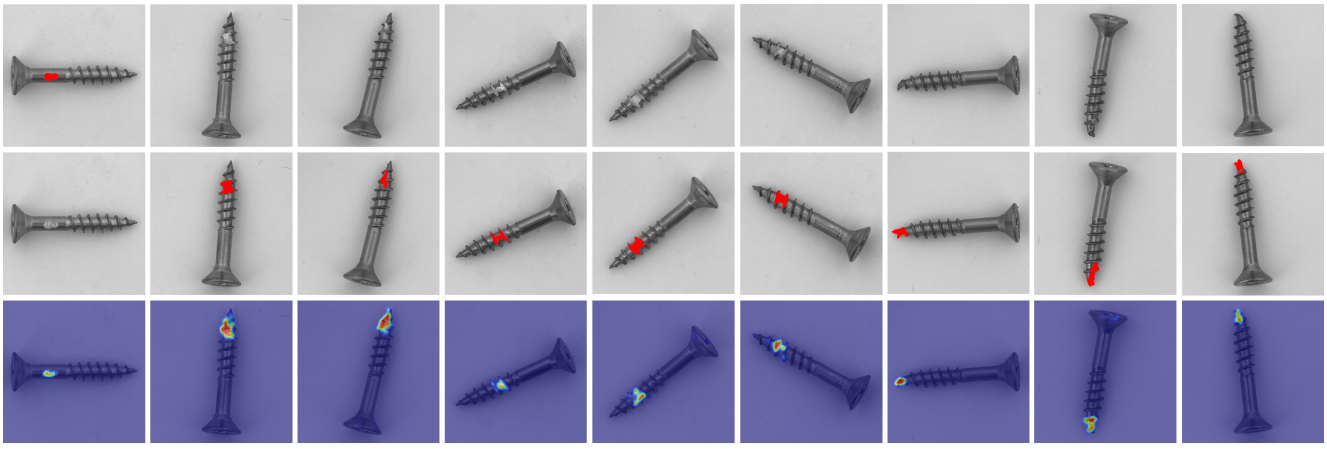

Figure A.13. Visualization Examples of Class Screw in MVTec-AD dataset.

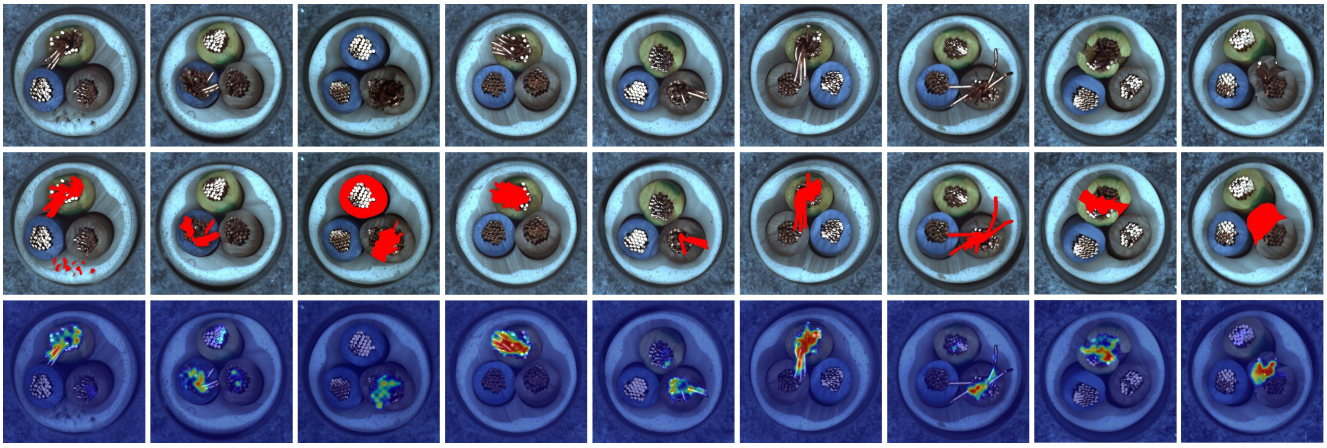

Figure A.14. Visualization Examples of Class Cable in MVTec-AD dataset.

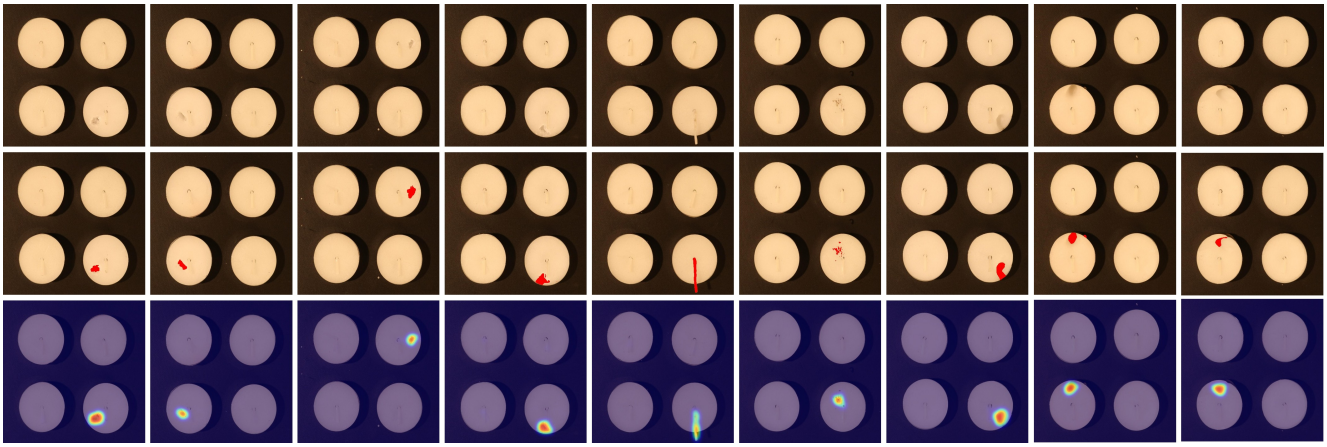

Figure A.15. Visualization Examples of Class Candle in VisA dataset.

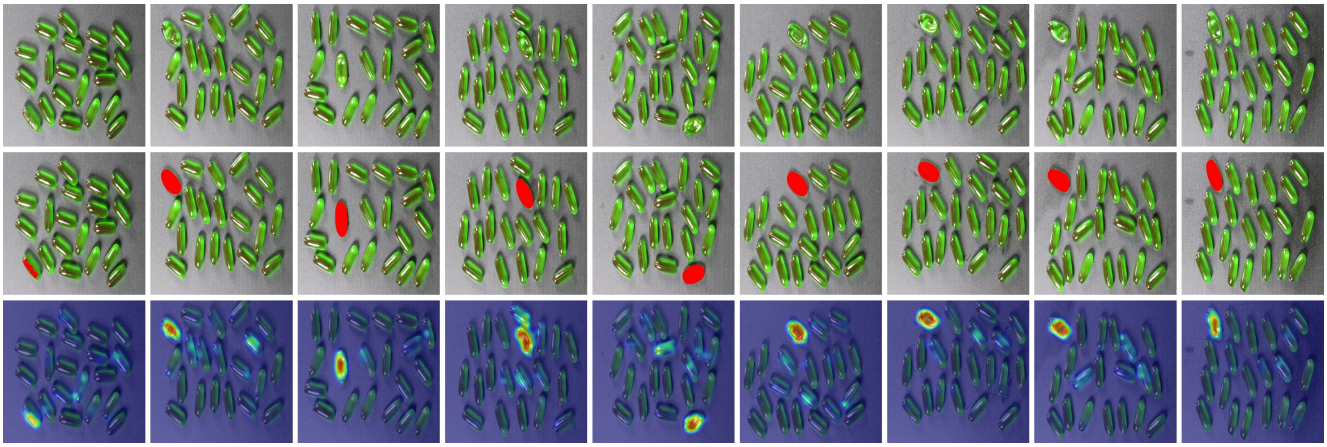

Figure A.16. Visualization Examples of Class Capsules in VisA dataset.

Figure A.17. Visualization Examples of Class Cashew in VisA dataset.

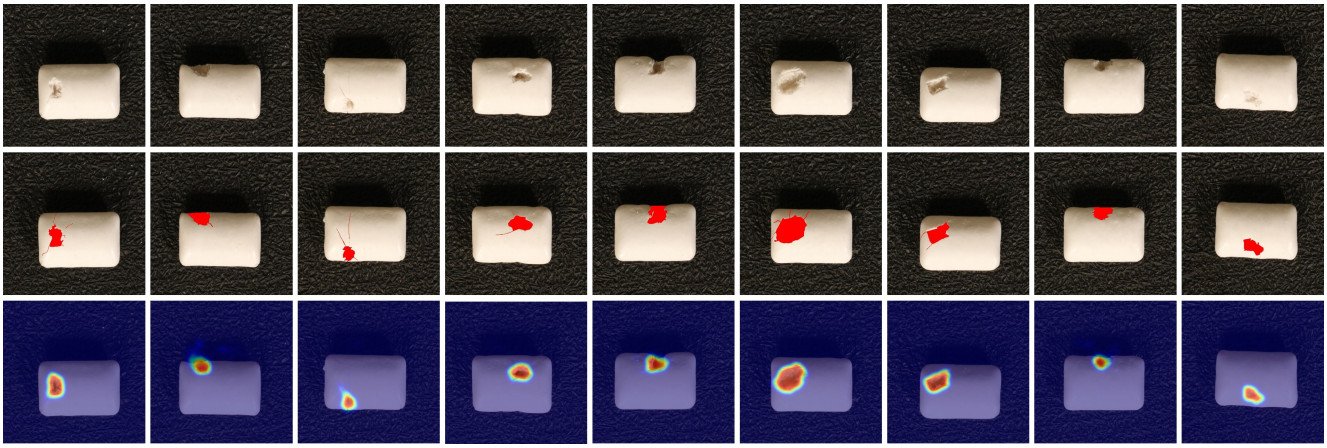

Figure A.18. Visualization Examples of Class Chewinggum in VisA dataset.

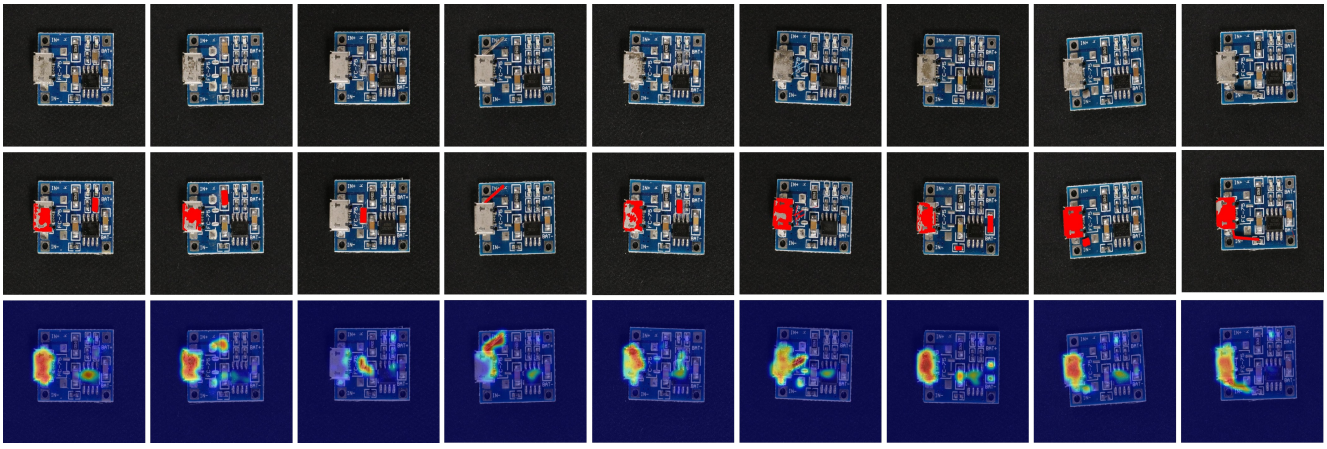

Figure A.19. Visualization Examples of Class PCB in VisA dataset.

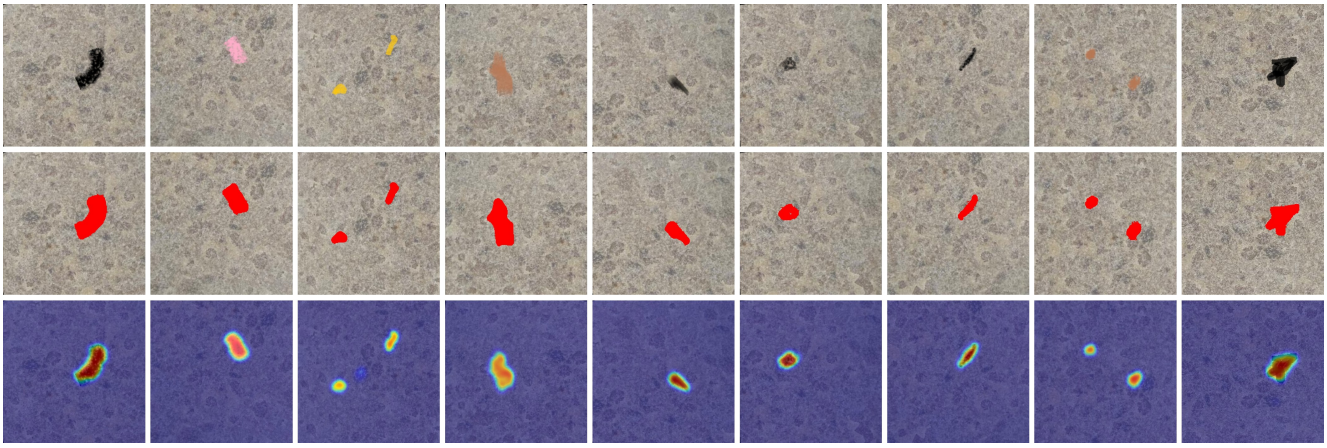

Figure A.20. Visualization Examples of Class Blotchy in DTD dataset.

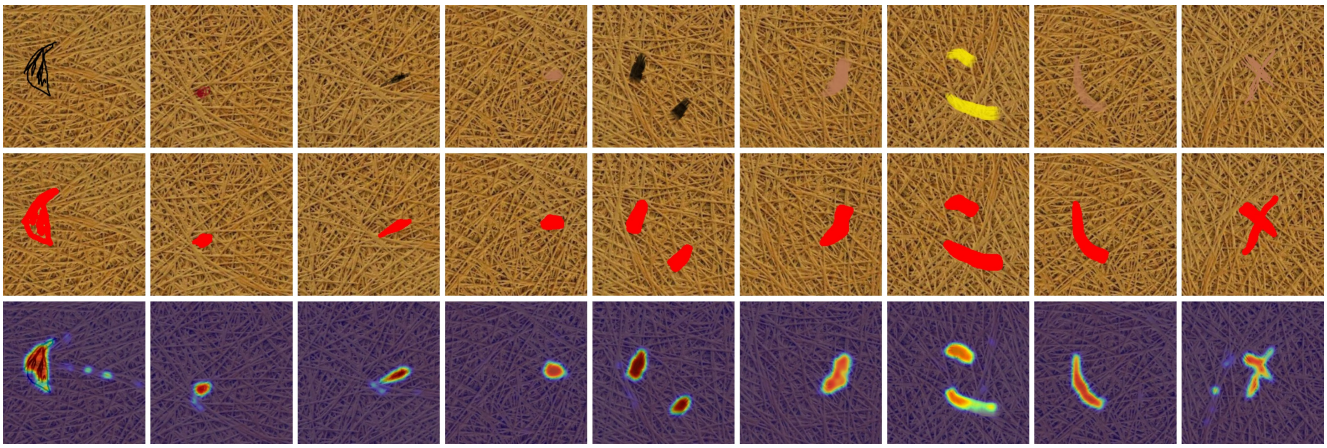

Figure A.21. Visualization Examples of Class Fibrous in DTD dataset.

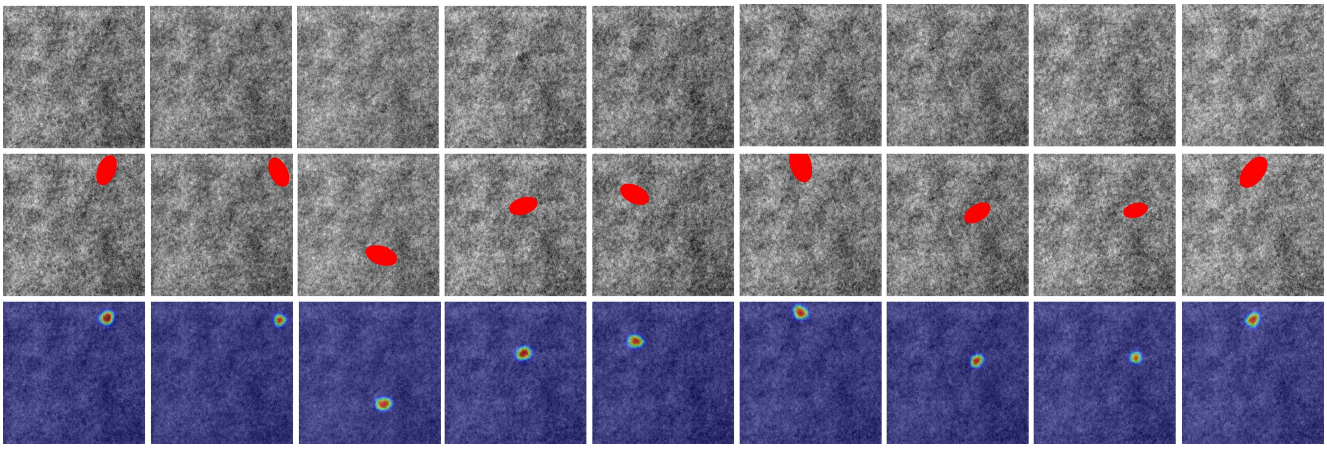

Figure A.22. Visualization Examples in DAGM dataset.

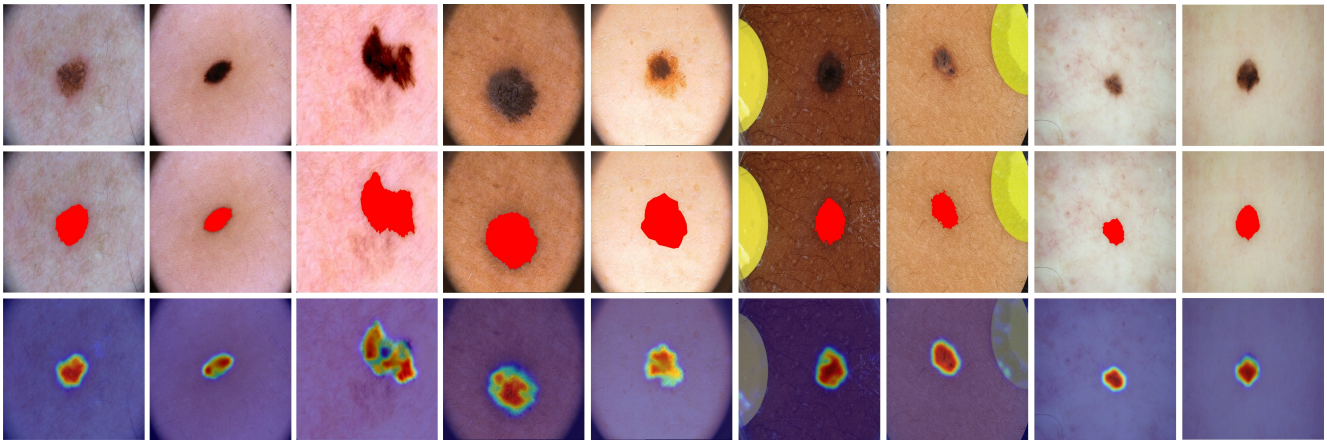

Figure A.23. Visualization Examples in ISIC dataset.

References

[1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023.

[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*, pages 9592–9600, 2019.

[3] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129: 103459, 2021.

[4] 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.

[5] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[6] A. Hamada. Br35h: Brain tumor detection 2020. *Online. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection*, 2020.

[7] Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *International Conference on Pattern Recognition*, pages 263–274. Springer, 2021.

[8] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical image analysis*, 55:216–227, 2019.

[9] 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.

[10] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019.

[11] Pranita Balaji Kanade and PP Gumaste. Brain tumor detection using mri images. *Brain*, 3(2):146–150, 2015.

[12] 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.

[13] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021.

[14] 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.

[15] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021.

[16] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, page 11, 2007.

[17] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022.