Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images - Supplementary Material

Chaoqin Huang[1,2,3]*, Aofan Jiang[1,3]*, Jinghao Feng[1,3], Ya Zhang[1,3], Xinchao Wang[2,†], Yanfeng Wang[1,3,†]
[1] Shanghai Jiao Tong University, [2] National University of Singapore
[3] Shanghai Artificial Intelligence Laboratory
{huangchaoqin,stillunnamed,fjh1345528968,ya_zhang,wangyanfeng622}@sjtu.edu.cn; {xinchao}@nus.edu.sg

Datasets	Sources	Train (all-normal)	Train (with labels)	Test	Sample size	Annotation Level
BrainMRI	BraTS2021 [1, 2, 15]	7,500	83	3,715	240*240	Segmentation mask
LiverCT	BTCV[14] + LiTs [5]	1,452	166	1,493	512*512	Segmentation mask
RESC	RESC [10]	4,297	115	1,805	512*1,024	Segmentation mask
OCT17	OCT2017 [13]	26,315	32	968	512*496	Image label
ChestXray	RSNA [18]	8,000	1,490	17,194	1,024*1,024	Image label
HIS	Camelyon16 [3]	5,088	236	2,000	256*256	Image label

Table 1. Summary of datasets from different medical modalities.

A. Medical Anomaly Detection Benchmark

The details of the medical anomaly detection (AD) benchmark are concisely summarized in Table 1. For the few-shot AD scenario, we select a random subset of labeled training samples, with $K \in \{2, 4, 8, 16\}$, from the labeled training set (designated as "Train (with labels)" in Table 1). These samples are employed in various competing baselines, including CLIP [11], WinCLIP [12], DRA [8], BGAD [19], and April-GAN [6]. Furthermore, consistent with the original methodologies that require training on a substantial amount of normal data, such as CFlowAD [9], RD4AD [7], PatchCore [16], and MKD [17], we employ a dataset exclusively comprising normal images. This dataset is referred to as "Train (all-normal)" for training purposes. It is important to highlight that this "all-normal" training set encompasses considerably more data compared to the limited data used in the few-shot scenario. Below are the detailed descriptions of the datasets used in the medical AD benchmark:

BrainMRI: This dataset is built upon the BraTS2021 dataset [1, 2, 15], utilizing 3D FLAIR volumes. To account for variations in brain images at different depths, slices within the depth range of 60 to 100 of the 3D FLAIR volumes are selected. Each extracted 2D slice was saved in PNG format and has an image size of 240×240 pixels. The training set encompasses 7,500 normal samples, while the test set comprises 3,715 samples with a balanced ratio of normal to anomaly instances.

LiverCT: Derived from two distinct datasets, BTCV [14] and LiTS [5], this dataset is structured to facilitate anomaly detection. The anomaly-free BTCV set, consisting of 50 abdominal 3D CT scans, constitutes the training set, while the test data comprises 131 abdominal 3D CT scans from LiTS. For both datasets, Hounsfield-Unit (HU) of the 3D scans are transformed into grayscale with an abdominal window. The scans are then cropped into 2D axial slices, containing 1,452 2D slices for training and 1,493 2D slices for testing.

Retinal OCT: The benchmark includes two different OCT AD datasets. The **RESC** dataset [10] offers pixel-level segmentation labels, delineating regions affected by retinal edema. In contrast, the **OCT17** dataset [13] primarily serves for classification tasks, featuring retinal OCT images categorized into three types of anomalies.

ChestXray: This dataset comprises lung images, utilizing RSNA [18] which was originally provided for a lung pneumonia detection task. Abnormal data encompasses cases of "Lung Opacity" and cases of "No Lung Opacity/Not Normal". The dataset is partitioned into 8,000 normal training images and 17,194 images for testing.

HIS: Based on Camelyon16 [3], this dataset encompasses 400 whole-slide images (WSIs) of lymph node sections stained with hematoxylin and eosin (H&E) from breast cancer patients. The training set incorporates 5,088 randomly extracted normal patches from the original training set. For testing, 1,003 normal and 997 abnormal patches from the

(a) *State*-level (-:normal, +:abnormal)

```
- c := "[o]"
- c := "flawless [o]"
- c := "perfect [o]"
- c := "unblemished [o]"
- c := "[o] without flaw"
- c := "[o] without defect"
- c := "[o] without damage"
+ c := "damaged [o]"
+ c := "[o] with flaw"
+ c := "[o] with defect"
+ c := "[o] with damage"
```

(b) *Template*-level

- `"a photo of a/the/one [c]."`
- `"a photo of a/the cool [c]."`
- `"a photo of a/the small [c]."`
- `"a photo of a/the large [c]."`
- `"a bright photo of a/the [c]."`
- `"a dark photo of a/the [c]."`
- `"a blurry photo of a/the [c]."`
- `"a bad photo of a/the [c]."`
- `"a good photo of a/the [c]."`
- `"a cropped photo of a/the [c]."`
- `"a close-up photo of a/the [c]."`
- (cont'd) `"a photo of my [c]."`
- `"a low resolution photo of a/the [c]."`
- `"a black and white photo of a/the [c]."`
- `"a jpeg corrupted photo of a/the [c]."`
- `"there is a/the [c] in the scene."`
- `"this is a/the/one [c] in the scene."`

Figure 1. Lists of state and template level prompts employed in this paper to construct text features.

Table 2. Comparisons with state-of-the-art **few-shot** anomaly detection methods with $K = 2, 4, 8, 16$. The AUCs (in %) for anomaly classification (AC) and anomaly segmentation (AS) are reported. The best result is in bold, and the second-best result is underlined.

Shot Number	Method	Source	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
			AC	AC	AC	AC	AS	AC	AS	AC	AS
2-shot	DRA [8]	CVPR 2022	72.91	72.22	98.08	71.78	72.09	57.17	63.13	85.69	65.59
	BGAD [19]	CVPR 2023	-	-	-	78.70	92.42	72.27	**98.71**	83.58	92.10
	APRIL-GAN [6]	arXiv 2023	69.57	69.84	**99.21**	78.45	94.02	57.80	95.87	89.44	96.39
	MFA	ours	**82.61**	**81.32**	97.98	**92.72**	**96.55**	**81.08**	96.57	**91.36**	**98.11**
4-shot	DRA [8]	CVPR 2022	68.73	75.81	99.06	80.62	74.77	59.64	71.79	90.90	77.28
	BGAD [19]	CVPR 2023	-	-	-	83.56	92.68	72.48	98.88	86.22	93.84
	APRIL-GAN [6]	arXiv 2023	76.11	77.43	**99.41**	89.18	94.67	53.05	96.24	94.70	97.98
	MFA	ours	**82.71**	**81.95**	99.38	**92.44**	**97.30**	**81.18**	**99.73**	**96.18**	**98.97**
8-shot	DRA [8]	CVPR 2022	74.33	82.70	99.13	85.94	75.32	72.53	81.78	93.06	83.07
	BGAD [19]	CVPR 2023	-	-	-	88.01	94.32	74.60	99.00	89.96	96.06
	APRIL-GAN [6]	arXiv 2023	81.70	73.69	**99.75**	88.41	95.50	62.38	97.56	91.36	97.36
	MFA	ours	**85.10**	**83.89**	99.64	**92.61**	**97.21**	**85.90**	**99.79**	**96.57**	**99.00**
16-shot	DRA [8]	CVPR 2022	79.16	85.01	99.87	82.99	80.45	80.89	93.00	94.88	84.01
	BGAD [19]	CVPR 2023	-	-	-	88.05	95.29	78.79	99.25	91.29	97.07
	APRIL-GAN [6]	arXiv 2023	81.16	78.62	**99.93**	94.03	96.17	82.94	99.64	95.96	98.47
	MFA	ours	**82.62**	**85.72**	99.66	**94.40**	**97.70**	**83.85**	**99.73**	**97.25**	**99.07**

115 testing WSIs are utilized.

B. Text Prompt Formatting

In this study, we adopt a combination of state-level and template-level prompts for generating textual input for the text encoder, as detailed in Figure 1 and following the approach in [12]. The state-level prompts are ingeniously designed by substituting the token [o] with names of human organs such as "brain", "liver", etc. This substitution strategy allows us to create a varied range of prompts that can categorize images as "normal" or "abnormal" based on the organ context. We then incorporate these state-level prompts into broader template-level constructs. By replacing the placeholder [c] in a template-level prompt with a corresponding state-level prompt, we formulate prompts that are both comprehensive and contextually rich. This systematic approach enables the creation of detailed, context-specific prompts that accurately distinguish between the normal and abnormal states.

C. Additional Quantitative Results

Results Varied Shot Numbers: Table 2 provides a detailed quantitative analysis on the performance of our medical AD approach, benchmarking it against leading few-shot AD methodologies. This analysis is meticulously tabulated, showing our approach's performance specificity across different shot numbers ($K \in \{2, 4, 8, 16\}$). These results lay the groundwork for a line chart featured in the main paper, which visually captures the subtle differences in performance under various conditions.

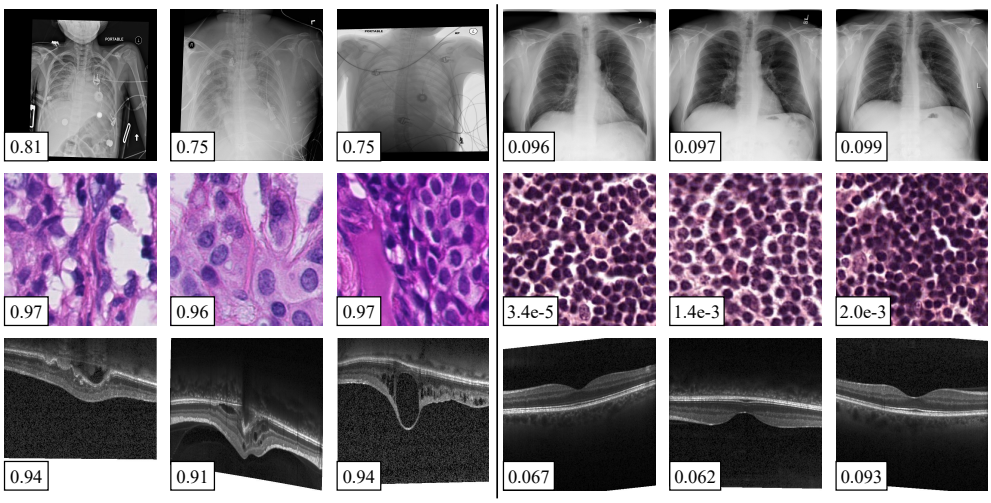
Ablation Study on Multi-level Features: We carried out an extensive ablation study to evaluate the effectiveness of

Table 3. Ablation study of multi-level features **without multi-level training**. The AUCs (in %) for classification (AC) and segmentation (AS) under the few-shot setting (k=4) are reported. The best result is in bold, and the second-best result is underlined.

Layers	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
	AC	AC	AC	AC	AS	AC	AS	AC	AS
Layer 1	74.54	78.69	97.75	87.84	97.05	58.15	98.47	88.76	97.58
Layer 2	<u>81.36</u>	<u>81.09</u>	**99.84**	<u>90.81</u>	**97.34**	**85.36**	<u>99.58</u>	<u>94.54</u>	<u>98.81</u>
Layer 3	69.00	79.75	98.68	83.01	94.34	63.78	95.35	93.29	98.40
Layer 4	71.02	72.84	99.37	86.92	95.42	76.09	97.92	93.72	98.23
Ensemble	**82.71**	**81.95**	<u>99.38</u>	**92.44**	<u>97.30</u>	<u>81.18</u>	**99.73**	**96.18**	**98.97**

Table 4. Ablation study of multi-level features **with multi-level training**. The AUCs (in %) for classification (AC) and segmentation (AS) under the few-shot setting (k=4) are reported. The best result is in bold, and the second-best result is underlined.

Layers	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
	AC	AC	AC	AC	AS	AC	AS	AC	AS
Layer 1	71.19	78.80	94.82	87.47	97.13	80.04	<u>99.67</u>	88.03	97.15
Layer 2	80.88	**83.56**	**99.49**	92.41	**97.35**	80.98	99.61	<u>95.73</u>	<u>98.90</u>
Layer 3	<u>82.35</u>	58.32	96.96	**93.01**	97.14	81.10	99.59	94.18	<u>98.90</u>
Layer 4	81.43	64.58	95.36	<u>92.86</u>	94.43	**81.32**	99.59	92.17	98.31
Ensemble	**82.71**	<u>81.95</u>	<u>99.38</u>	92.44	<u>97.30</u>	<u>81.18</u>	**99.73**	**96.18**	**98.97**



(a) Abnormal Samples (b) Normal Samples

Figure 2. Examples of (a) abnormal samples and (b) normal samples on chest X-ray, histopathology, and retinal OCT. The predicted scores by our method are shown with each sample. The higher the score, the more likely to be an anomaly.

utilizing single-layer features for each dataset, in line with the average performances across all datasets discussed in the main paper. The outcomes, elucidated in Table 3, provide a comprehensive understanding of the performance of single-layer features obtained without the implementation of multi-level training. In contrast, Table 4 presents the results attained through the strategic implementation of multi-level training techniques.

D. Additional Qualitative Results

Anomaly Classification Instances: In Figure 2, we display the results of anomaly classification from datasets that only provide anomaly classification labels. These results were obtained using our method in a few-shot setting ($K = 4$). Each instance is accompanied by a predicted score, ranging from zero to one, where higher scores indicate a higher likelihood of an anomaly.

Table 5. Comparisons with state-of-the-art methods on in-domain dataset MVTec AD. The AUCs (in %) for classification (AC) and segmentation (AS) under the few-shot setting (k=4) are reported.

Category (k=4)	RegAD		WinCLIP		April-GAN		MVFA	
	AC	AS	AC	AS	AC	AS	AC	AS
bottle	99.3	98.5	99.3	97.8	94.2	97.2	99.8	98.7
cable	82.9	95.5	90.9	94.9	76.7	91.8	88.0	87.3
capsule	77.3	98.3	82.3	96.2	93.5	97.5	93.9	96.0
carpet	97.9	98.9	100	99.3	99.9	98.7	100	99.4
grid	87.0	85.7	99.6	98.0	99.2	97.6	100	96.9
hazelnut	95.9	98.4	98.4	98.8	98.8	97.7	99.7	98.1
leather	99.9	99.0	100	99.9	100	99.5	99.9	99.4
metal nut	94.3	96.5	99.5	92.9	91.0	93.1	99.4	99.3
pill	74.0	97.4	92.8	97.1	84.1	95.5	95.1	96.8
screw	59.3	96.0	87.9	96.0	83.7	98.5	88.3	98.5
tile	98.2	92.6	99.9	96.6	99.1	96.0	99.7	98.7
toothbrush	91.1	98.5	96.7	98.4	93.2	98.8	95.8	98.8
transistor	85.5	93.5	85.7	88.5	84.1	83.7	84.3	80.9
wood	98.9	96.3	99.8	95.4	98.7	96.2	99.7	97.2
zipper	95.8	98.6	94.5	94.2	95.4	96.6	99.3	98.9
average	89.2	96.2	95.2	**96.3**	92.8	95.9	**96.2**	**96.3**

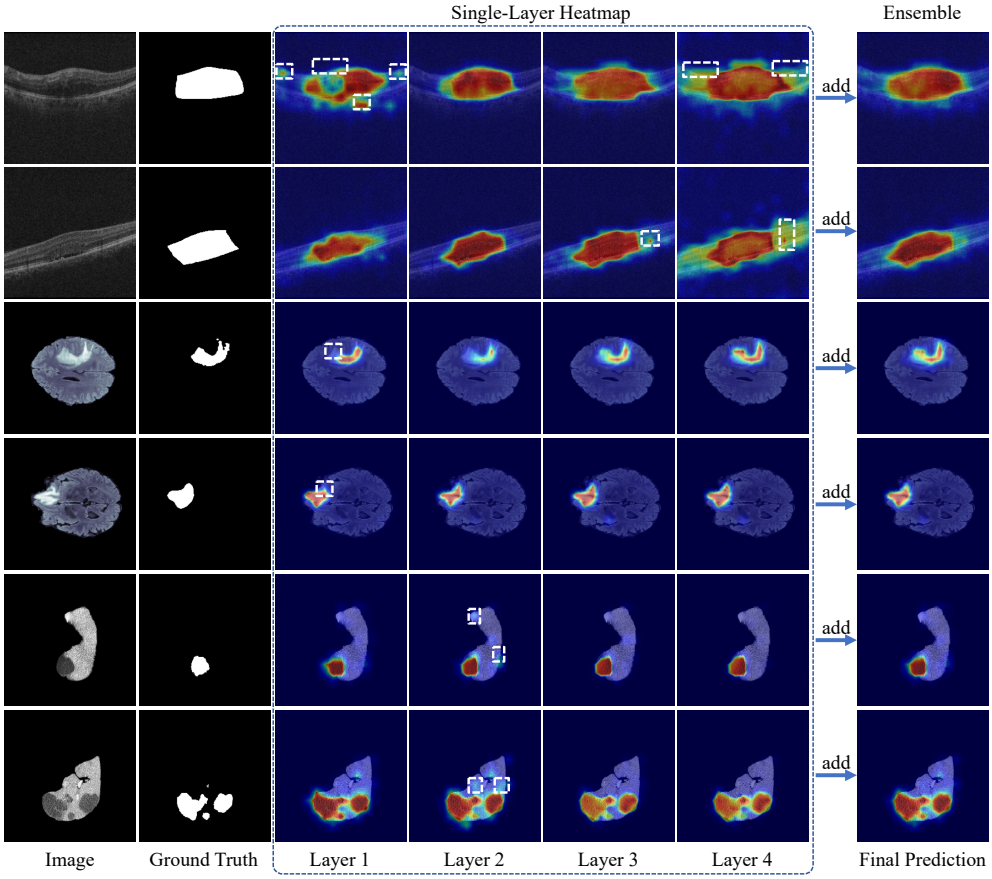


Figure 3. Visualization of anomaly segmentation heatmaps from the four single layers and the multi-layer ensemble results. The white dashed boxes demarcate regions that have been missed or erroneously segmented.

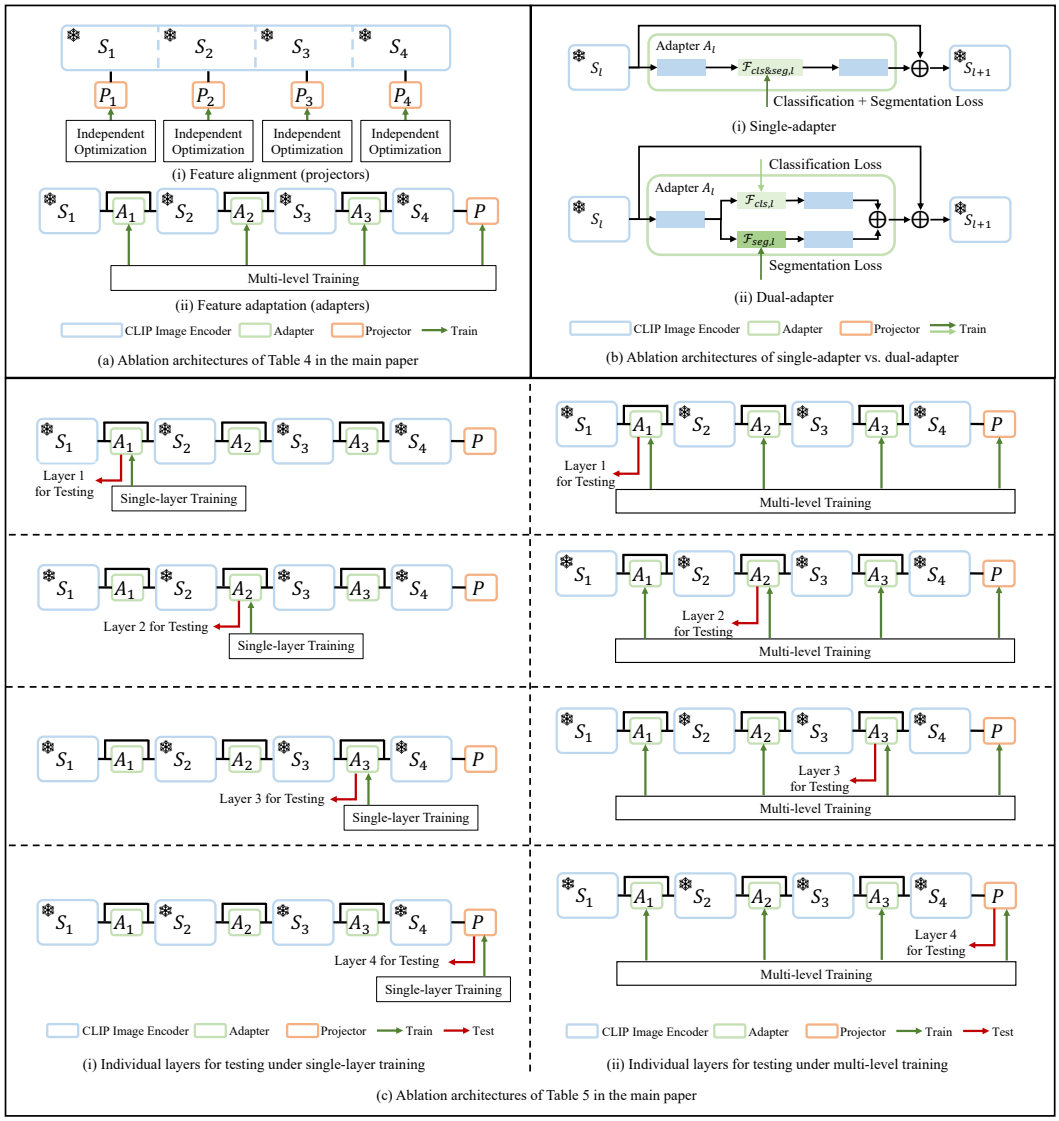


Figure 4. Model structures corresponding to the ablation experimental settings.

Ensemble of Multi-level Features: Figure 3 showcases visualizations from different layers used in the anomaly segmentation task. These visualizations include results from datasets with segmentation labels, such as BrainMRI, LiverCT, and RESC.

Evaluation on Industrial Anomaly Detection: For in-domain evaluation, the MVTec AD benchmark [4], consisting of 15 industrial defect detection sub-datasets, is considered. MVFA significantly outperforms competing methods, highlighting its superior generalization capabilities. Detailed results for each sub-dataset are included in Table 5.

E. Ablation Model Structure

To effectively convey the nuances of our ablation study in the main paper, we utilized Figure 4 to graphically demonstrate the configurations of the models used in our experiments. Specifically, Figure 4 (a) visually details the designs of both the adapter and projector as outlined in Table 4 of the main paper, where part (i) illustrates the projector and part (ii) depicts the adapter. In Figure 4 (b), we present the configurations for both the single-adapter and dual-adapter models, shown in subfigures (i) and (ii) respectively. Furthermore, Figure 4 (c) illustrates the testing pipeline for assessing the impact of training at different levels. Subfigure (i) represents the scenario of single-layer training, while subfigure (ii) demonstrates the approach for multi-level training, corresponding to the discussions and findings presented in Table 5 of the main paper.

Dual-Adapter vs. Single-Adapter. We compare the performance of the dual-adapter architecture against the single-

Table 6. Ablation studies of the architecture of dual-adapter against single-adapter in MVFA. The AUCs (in %) for classification (AC) and segmentation (AS) under the few-shot setting (K=4) are reported, with the best result marked in bold.

Datasets	AC		AS	
	single-adapter	dual-adapter	single-adapter	dual-adapter
HIS	80.80	**82.71**	-	-
ChestXray	78.02	**81.95**	-	-
OCT17	**99.87**	99.38	-	-
BrainMRI	92.28	**92.44**	96.98	**97.30**
LiverCT	81.07	**81.18**	99.42	**99.73**
RESC	94.06	**96.18**	98.53	**98.97**
average	87.68	**88.97**	98.31	**98.67**

adapter setup within the few-shot setting. The dual-adapter design, as implemented in our MVFA model, generates two parallel sets of features at each level, catering to both global (classification) and local (segmentation) aspects. The corresponding architectures are shown in Figure 4 (b). According to the results in Table 6, the dual-adapter approach outperforms the single-adapter model on almost all the datasets. We observed an enhancement in the average AUC for AC, improving from 87.68% to 88.97%, and for AS, rising from 98.31% to 98.67%. This improvement indicates that the dual-adapter architecture is more effective in managing the demands of both AC and AS in medical images.

References

[1] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 1

[2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1): 1–13, 2017. 1

[3] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 1

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019. 5

[5] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 1

[6] 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*, 2023. 1, 2

[7] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 2022. 1

[8] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, 2022. 1, 2

[9] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *WACV*, 2022. 1

[10] 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. 1

[11] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 1

[12] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023. 1, 2

[13] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 1

[14] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 1

[15] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE TMI*, 34(10):1993–2024, 2014. 1

[16] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022. 1

[17] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad Hossein Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, 2021. 1

[18] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017. 1

[19] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *CVPR*, 2023. 1, 2