UltraAD: Fine-Grained Ultrasound Anomaly Classification via Few-Shot CLIP Adaptation

Yue Zhou[1,2], Yuan Bi[1,2], Wenjuan Tong[3], Wei Wang[3], Nassir Navab[1], and Zhongliang Jiang[1]

[1] Computer Aided Medical Procedures (CAMP), TU Munich, Germany
zl.jiang@tum.com
[2] Munich Center for Machine Learning (MCML), Munich, Germany
[3] The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China

Abstract. Precise anomaly detection in medical images is critical for clinical decision-making. While recent unsupervised or semi-supervised anomaly detection methods trained on large-scale normal data show promising results, they lack fine-grained differentiation, such as benign vs. malignant tumors. Additionally, ultrasound (US) imaging is highly sensitive to devices and acquisition parameter variations, creating significant domain gaps in the resulting US images. To address these challenges, we propose UltraAD, a vision-language model (VLM)-based approach that leverages few-shot US examples for generalized anomaly localization and fine-grained classification. To enhance localization performance, the image-level token of query visual prototypes is first fused with learnable text embeddings. This image-informed prompt feature is then further integrated with patch-level tokens, refining local representations for improved accuracy. For fine-grained classification, a memory bank is constructed from few-shot image samples and corresponding text descriptions that capture anatomical and abnormality-specific features. During training, the stored text embeddings remain frozen, while image features are adapted to better align with medical data. UltraAD has been extensively evaluated on three breast US datasets, outperforming state-of-the-art methods in both lesion localization and fine-grained medical classification. Project page: https://karolinezhy.github.io/UltraAD/

Keywords: Ultrasound image analysis · Anomaly detection · Few-shot adaptation

1 Introduction

Medical ultrasound (US) is a widely used imaging modality for examining internal organs, such as the breast and thyroid, due to its real-time capability, non-radiative nature, and accessibility. In remote or low-income regions, it is often the only available diagnostic tool. However, ultrasound images frequently suffer from low quality and significant domain variations due to differences in imaging devices and acquisition parameters [4, 16], which pose challenges for ultrasound

image understanding [18]. To address these challenges, a generalized anomaly detection (AD) algorithm that adapts across diverse anatomical structures and imaging domains is highly demanded for supporting clinical decision-making, particularly for junior clinicians with limited clinical experience.

While supervised deep learning has achieved phenomenal success in medical image segmentation [10, 11, 14, 17, 23, 33], these methods require large labeled training data, which is costly in the US imaging field. To overcome this limitation, unsupervised and self-supervised approaches leveraging autoencoders and generative models have been developed using only unlabeled data [2, 3]. Despite success in MRI anomaly detection, these methods cannot be directly adopted to US images because of the noisy nature and limited imaging field of view.

To improve generalization across unseen domains or objects, recent advancements in vision-language models (VLMs) have gained increasing attention. A pioneering study, WinCLIP [13], employed pre-trained CLIP [22] with binary textual prompts for anomaly detection. Subsequent methods, including VAND [7], AnomalyCLIP [32], AdaClip [5], and VCP-CLIP [21], have further integrated prompt learning and adapter mechanisms to better align natural image representations with specific application domains. MediCLIP [29] adapts the pretrained CLIP model for anomaly detection in medical images by leveraging few-shot normal images with synthetic anomalies. The improved generalization of these approaches is largely attributed to the universality of text prompts across diverse datasets. However, these methods focus on binary anomaly detection, neglecting the need for fine-grained classification, which is crucial for distinguishing lesion phases or benign and malignant tumors in medical applications.

For few-shot classification tasks using VLMs, Radford *et al.* introduced linear probing, which optimizes a classifier on top of frozen vision encoders [22]. Building on this, LP++ [12] has demonstrated promising results in the natural image domain. To better align language and image features while preserving pretrained representations, prompt learning methods incorporate learnable tokens [25, 30, 31]. To enhance few-shot classification, CLIP-Adapter [8] introduced lightweight non-linear projections via MLP layers. Alternatively, Tip-Adapter [28] offers a training-free approach that achieves comparable classification performance through direct feature optimization. In the medical domain, few-shot adaptation has also demonstrated strong performance on surgical data, as evidenced by recent work [6, 27]. Although promising results have been demonstrated on natural images, their extension to US imaging remains unexplored. Furthermore, due to the specific demands of clinical diagnosis, few-shot anomaly detection with fine-grained anomaly classification is crucial yet remains an unmet need in the research community.

In this study, we introduce UltraAD, a CLIP-based framework for lesion localization and fine-grained anomaly classification through few-shot adaptation. Inspired by multi-task learning in medical image segmentation [15, 26], UltraAD unifies pixel-wise anomaly localization and image-level classification to boost performance across tasks. To address the domain gap between natural and US images, a small set of US samples (4/8/16 shots in this study) is needed to

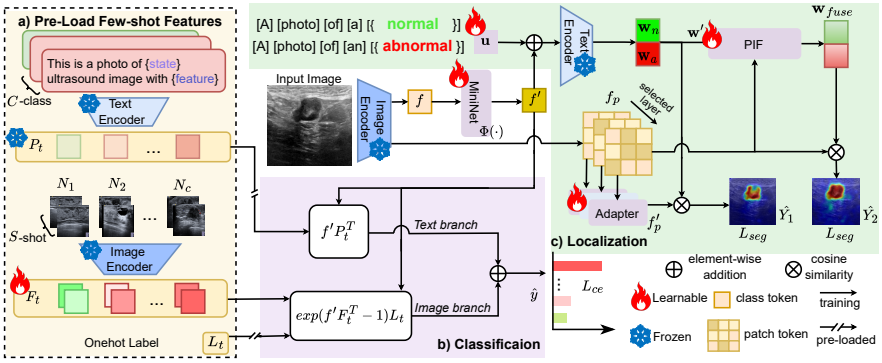

Fig. 1. Proposed pipeline. Via pretrained image and text encoders, text token embeddings P_t, image token embeddings F_t and onehot label L_t are preloaded from the S-shot C-class dataset. A mini-net projects image token embeddings f to obtain f'. For anomaly classification, F_t, P_t and L_t are utilized to generate a classification score with f' through feature similarity. For anomaly localization, two initial text prompts indicating normal and abnormal are added with the projected image token embedding f'. Two localization masks are produced by calculating the cosine similarity between an updated image token embedding w' and the projected patch token f'_p, and a text-image fused text-embedding w_{fuse} with the original patch token f_p. The final anomaly map prediction is computed as $\hat{Y} = \frac{1}{2}(\hat{Y}_1 + \hat{Y}_1)$.

adapt the CLIP model. To the best of our knowledge, this is the first CLIP-based model tailored for meeting US diagnosis requirements, leveraging domain-specific pathological information to improve generalization across imaging variations. Notably, the few-shot adaptation is performed on a single public breast dataset, while validation is conducted on two unseen breast datasets acquired using different machines, patients, and probe types. This highlights the robustness and effectiveness of the proposed method in real scenarios. The key contributions are: (1) A simple yet effective few-shot adaptation approach that generalizes across unseen US images of the same anatomy, despite significant domain variations in machine types and acquisition parameters; (2) a vision-language fusion strategy, combining learnable prompt tokens and image embeddings at both global (image-level) and local (patch-wise) scales to enhance anomaly localization; (3) a feature memory bank built from a small set of paired US images and text descriptions, enabling enhanced anomaly classification.

2 Methodology

Problem Formulation US domain gaps can arise from variations in machines and imaging parameters. We aim to develop a generalized model for pixel-wise anomaly localization and image-level classification indicating varying severity

through few-shot CLIP adaptation. The S-shot training dataset from a specific domain is denoted as $\mathcal{D}_{train} = \{\mathbf{X}_i, \mathbf{Y}_i, y_i\}_{i=1}^{C\times S}$, where C represents the number of image classes. Here, $\mathbf{X}_i \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{Y}_i \in \mathbb{R}^{H\times W\times 1}$ are input image and its annotation mask, respectively. $y_i \in \{0, 1, \ldots, C-1\}$ is the class label. If $y_i = 0$, then $\mathbf{Y}_i = \mathbf{0}$, indicating no anomaly; otherwise, y_i denotes a specific anomaly type, and $\mathbf{Y}_i$ has pixels with a value of one, marking the anomaly region. The unseen test data is denoted as $\mathcal{D}_{test}^k = \{\mathbf{X}_i, \mathbf{Y}_i, y_i\}_{i=1}^{N_k}$, which consists of images from the k-th domain (where $k \in \{1, \ldots, K\}$ representing an unseen dataset).

2.1 CLIP for Anomaly Detection

To adapt CLIP [22] for anomaly detection, two contrasting text prompts are used: A photo of a normal *[CLS]* and A photo of an abnormal *[CLS]*, where *[CLS]* represents the specific image class category, e.g., breast US image. The text encoder $E_T(\cdot)$ processes text prompts to generate the text embedding $\mathbf{w}_n \in \mathbb{R}^D$ for the normal class and $\mathbf{w}_a \in \mathbb{R}^D$ for the abnormal class. Given an input image X, the image encoder $E_I(\cdot)$ outputs a global class token $f \in \mathbb{R}^D$ and local patch tokens $f_p \in \mathbb{R}^{h\times w\times D}$, where h and w denote the spatial dimensions in feature space. The anomaly scores are obtained based on cosine similarity $\langle\cdot\rangle$ between text and visual embedding for both global image-level and local pixel-level, i.e.,

$$\hat{y} = \frac{\exp(\langle \mathbf{w}_a, f\rangle)}{\exp(\langle \mathbf{w}_n, f\rangle) + \exp(\langle \mathbf{w}_a, f\rangle)}, \hat{Y} = \sigma\left(\frac{\exp(\langle \mathbf{w}_a, f_p\rangle)}{\exp(\langle \mathbf{w}_n, f_p\rangle) + \exp(\langle \mathbf{w}_a, f_p\rangle)}\right) \quad (1)$$

where $\sigma(\cdot)$ is an interpolation function to upsample patch-level anomaly scores into the final predicted anomaly map $\hat{Y} \in \mathbb{R}^{H\times W}$.

Image-Aware Prompting Module: Inspired by [30], to incorporate few-shot adaptation that improves generalization ability to unseen images, we proposed a simple unified template to generate both normal and abnormal text prompts based on input US images instead of using a pre-defined *[CLS]* token. Specifically, the template replace the pre-defined *[CLS]* token by learnable tokens, i.e.

$$P = [\mathrm{A}][\mathrm{photo}][\mathrm{of}][\mathrm{a/an}][state][u + f']$$

where $u \in \mathbb{R}^D$ is global learnable token and $f' \in \mathbb{R}^D$ is instance-specific token. To obtain the instance-specific token f', a MiniNet $\Phi(\cdot)$ takes the global image class token f as input, thus adding global image information into the text token.

Image Feature Adapter: To further improve anomaly localization performance, we utilize the lightweight linear adapters [7] to refine image features obtained from pretrained CLIP image encoder. The linear adapters are optimized during few-shot adaptation. The adapters project patch token features f_p from each encoder layer l to f_p' for anomaly map computation. Finally, the final anomaly map $\hat{Y}$ is obtained by averaging anomaly maps from all layers.

2.2 Memory-Boosted Few-Shot Adaptation

Previous modifications are aimed at enhancing CLIP's performance in anomaly localization in US imaging. However, it still lacks the capacity for fine-grained

anomaly classification, which is crucial for accurate disease grading and severity assessment. To address this limitation, our method utilizes memory-based few-shot adaptation and thus can predict different anomaly types.

A support dataset with S-shot C-class training samples for few-shot adaptation denoted as $\mathcal{I}$ with corresponding labels $\mathcal{L}$, as illustrated in the left yellow part of Fig.2. The total number of training data is $N = S \times C$. To enhance adaptation in the medical domain, beyond class labels, a lesion-aware prompt template $\mathcal{P}$ is incorporated to provide domain-specific contextual information for each class. The template is structured as: [This] [is] [a] [*type*] [ultrasound] [image] [with] [*pathological features*]. For instance, in the case of a benign lesion, the prompt is instantiated as: *"This is a benign ultrasound image with round shape and sharply demarcated margins."* We employ the pretrained CLIP text encoder $E_T(\cdot)$ and image encoder $E_I(\cdot)$ to obtain both text tokens and image tokens that are pre-computed and stored in the memory, i.e.,

$$\mathbf{F}_{\mathrm{t}} = E_I(\mathcal{I}) \in \mathbb{R}^{N \times D},\ \mathbf{P}_{\mathrm{t}} = E_T(\mathcal{P}) \in \mathbb{R}^{C \times D},\ \mathbf{L}_{\mathrm{t}} = \mathrm{OneHot}(\mathcal{L}) \in \mathbb{R}^{N \times C} \quad (2)$$

The classification result is determined by the similarity between the text prompt and the projected image class token f' using a lightweight network. Additionally, it considers the interaction between the few-shot image memory and the image class token, as proposed in [28]. The classification score for a sample is computed as follows, also shown in pink in Fig.2:

$$\hat{y} = f'\mathbf{P}_{\mathrm{t}}^T + \exp(f'\mathbf{F}_{\mathrm{t}}^T - 1)L_{\mathrm{t}} \quad (3)$$

2.3 Patch-Wise Image-Language Fusion

To enhance the alignment between US-informed text embeddings [30] and image patch features, a further feature fusion of patch-wise image embedding and prompts representation is carried out (see PIF block in green in Fig. 1). To effectively leverage the few-shot samples, we utilize class-specific prompts from few-shot prompt embedding $\mathbf{P}_t$ by ensembling them with normal and abnormal prototypes for anomaly detection. The normal prototype is defined as $w'_n = w_n + \mathbf{P}_t^0$, while the abnormal prototype is given by $w'_a = w_a + \frac{1}{C-1}\sum_{i=1}^{C} \mathbf{P}_t^i$, ensuring diversity in normal and abnormal prompts by incorporating US-aware information. Hence, we get $w' = [wn', wa']$. To further refine feature fusion, we adopt an M-head cross-attention module that learns three projection matrices, W_Q, W_K, W_V, to compute image-language fused text embedding:

$$Q = \mathbf{w}'W_Q,\ K = f_pW_K,\ V = f_pW_V,\ \mathbf{w}^{fuse} = \mathrm{softmax}\left(QK^T\right)V \quad (4)$$

2.4 Multitask Learning for Anomaly Classification and Localization

The class token, which acts as an embedding containing global image information, can be used for image classification and integrated into the prompt for conditioned prompt learning. To achieve this, we combine anomaly classification and anomaly localization by using a shared, dynamically updated class

Table 1. Summary of breast ultrasound datasets used in experiments

Dataset	Total	Normal	Benign	Malignant	Ultrasound System	Year
BUS-UCLM	683	419	174	90	Siemens ACUSON S2000TM	2022-23
BUSI	780	133	437	210	LOGIQ E9, LOGIQ E9 Agile	2018
BUSZS	300	100	100	100	Mindray, Toshiba, GE, Canon, PHILIPS, Esaote	2023

token within a unified framework. This approach seamlessly integrates anomaly classification with anomaly localization within a cohesive learning pipeline. The initially extracted CLIP image feature, denoted as f, is projected via a MiniNet $\Phi(\cdot)$, resulting in f'. During training, UltraAD refines the pixel-level anomaly maps $\hat{Y}_1$ and $\hat{Y}_2$ using a combination of Dice loss [20] and Focal loss [19], equally weighted, on auxiliary data (few-shot samples) to ensure accurate segmentation. Meanwhile, it refines classification performance by employing cross-entropy loss. The image feature memory $\mathbf{F}_t$ will be refined during the training while text feature $\mathbf{P}_t$ remains frozen. The network is trained end-to-end with a joint optimization of segmentation and classification losses. The shared feature f' is jointly refined, which will be used to compute the classification result as well as the segmentation result.

3 Experiments

3.1 Experimental Setup

Training and Testing Tab.1 details the datasets used for training and testing. We employ a few-shot subset of BUS-UCLM [24] for model adaptation and assess performance on a public breast dataset BUSI [1] and an in-house dataset collected using six different US machines with both convex and linear probes, denoted as BUSZS without additional adaption. For the few-shot setting, we select 4, 8, and 16 samples per class from the BUS-UCLM [24]. For fair comparisons, all baseline methods use identical training shots, except WinCLIP [13], which only uses a pre-trained CLIP model without any further training. Other anomaly detection methods [5,9,21,32] and few-shot CLIP adaptation approaches [8, 12, 30] require training on the few-shot data. For consistency, we use the "ViT-L/14@336px" CLIP backbone across all methods.

Following the evaluation methods used in anomaly detection and localization tasks [5, 13, 21, 32], we utilize the AUROC score for classification and AUROC/AUPRC for localization. All the values in Tab. 2 and Tab. 3 shows the average metrics of three experiments using 3 seeds for few-shot sample selection to ensure fair evaluation and reduce random few-shot sample variations.

3.2 Performance Analysis

Anomaly Detection and Classification We conduct comprehensive evaluations on two distinct few-shot classification tasks: anomaly detection (binary)

Table 2. Comparison of anomaly detection (left) and classification (right) performance across benchmarks, evaluated using image-level AUROC score. The best values are **bold**, and the second best is underlined.

	Anomaly Detection (binary)							Anomaly Detection (multi-class)						
	BUSI			BUSZS				BUSI			BUSZS			
	4	8	16	4	8	16		4	8	16	4	8	16	
WinCLIP [13]	79.7			59.3				-	-	-	-	-	-	
AdaCLIP [5]	54.7	78.7	86.5	66.9	83.5	92.8	LP++ [12]	56.8	62.3	70.1	54.4	56.5	62.2	
AnomalyCLIP [32]	62.9	81.9	85.6	76.6	87.4	90.8	ClipAdapter [8]	51.8	51.8	52.0	50.7	50.5	51.0	
VCP-CLIP [21]	84.8	85.7	90.0	77.3	86.8	89.6	TipAdapter [28]	58.9	61.4	68.8	59.0	57.2	58.7	
MVFA [29]	78.4	84.0	90.6	83.3	88.9	94.8	COOP [31]	58.8	61.8	67.7	57.9	58.9	62.1	
Ours	**87.0**	**90.7**	**91.4**	**95.9**	**98.2**	**98.2**	Ours	**68.6**	**72.0**	**73.8**	**78.9**	**82.8**	**84.1**	

and anomaly classification (multi-class). For anomaly detection, we formulate the task as a binary classification problem and evaluate against leading CLIP-based AD approaches [5, 9, 13, 21, 32]. Tab.2 (left) demonstrate UltraAD's superior performance across all settings, particularly in low-shot scenarios where it achieves a substantial 10% gain over baselines in the challenging 4-shot setting. In addition to anomaly detection, we address the more complex task of anomaly classification, which involves fine-grained classification of anomaly types under limited data constraints. Comparing against state-of-the-art CLIP-based few-shot learning techniques including advanced prompt tuning and adapter-based methods [8, 12, 28, 31], our results in Tab.2 (right) shows UltraAD's significant advantages, especially on the complex BUSZS dataset, achieving 80% score on most settings. We attribute this superior performance to our novel integration of segmentation-guided learning. Additionally, we employ a mask-guided post-processing technique, commonly used in recent anomaly detection methods [13, 32], which utilizes segmentation masks to refine the separation between normal and abnormal classes. This approach further enhances overall performance, specifically using the formulation $y_{pred} = \frac{1}{2}(\max(\hat{Y}) + \hat{y})$.

Table 3. Comparison of anomaly localization performance across benchmarks, evaluated using pixel-level AUROC and AUPRC, shown as (AUROC, AUPRC). Best values are **bold**, and seconds are underlined.

	BUSI			BUSZS		
Method	4	8	16	4	8	16
WinCLIP [13]	(79.7, 5.4)			(62.1, 8.4)		
AdaCLIP [5]	(79.4, 32.7)	(87.5, 51.5)	(86.8, 52.4)	(85.9, 42.1)	(94.3, 60.8)	(95.4, 70.5)
AnomalyCLIP [32]	(87.8, **50.2**)	(90.0, 54.8)	(90.6, 56.0)	(**94.1**, **64.4**)	(96.1, 68.9)	(96.6, 70.9)
VCP-CLIP [21]	(75.0, 27.6)	(79.0, 34.4)	(80.4, 38.4)	(75.3, 21.3)	(81.7, 32.0)	(83.5, 36.2)
MVFA [9]	(69.9, 16.1)	(66.2, 15.1)	(69.5, 16.2)	(87.9, 42.4)	(92.4, 65.0)	(93.5, 68.9)
Ours	(**88.3**, 48.7)	(**91.5**, **58.9**)	(**91.8**, **58.7**)	(93.9, 61.1)	(**96.5**, **70.5**)	(**96.7**, **73.3**)

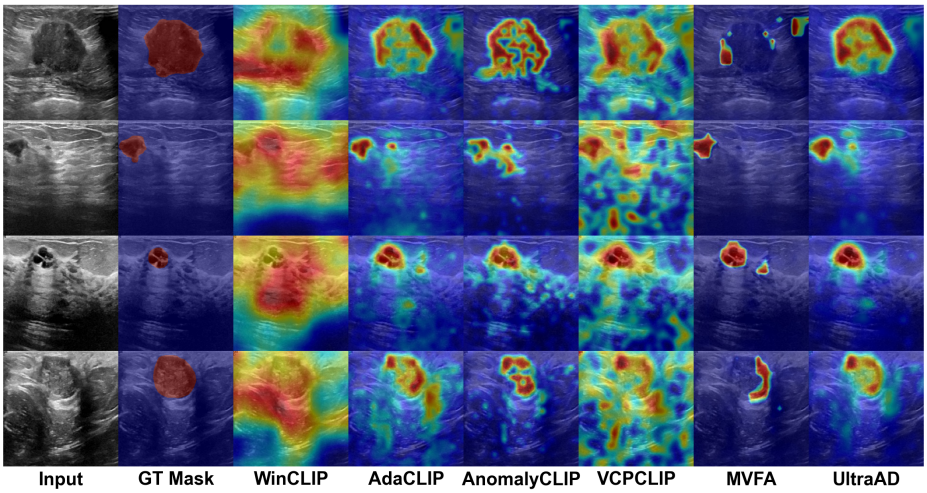

Fig. 2. Visualization of anomaly detection maps from multiple baseline methods. UltraAD achieves the most precise localization on unseen ultrasound images, effectively identifying anomalies as cohesive regions rather than fragments.

Anomaly Localization Tab. 3 presents a comprehensive analysis of our method against state-of-the-art baseline models for anomaly localization. Our proposed UltraAD demonstrates superior performance compared to most existing methods. AnomalyCLIP [32] demonstrates competitive performance attributed to its V-V attention mechanism, which significantly enhances local feature perception. Nevertheless, AnomalyCLIP exhibits limitations in image level anomaly detection show in Tab.2. Fig.2 presents the quantitative results of localization. WinCLIP [13] and VCPCLIP [21] failed to produce satisfactory maps, while MVFA [9] and AnomalyCLIP [32] struggled to generate continuous regions as anomaly areas. AdaCLIP [5] included some normal regions in its predictions. Our method detects anomaly locations and predicts them as continuous regions, which can be further used for precise segmentation.

3.3 Ablation Study

We evaluate UltraAD modules via an ablation study in a 4-shot (one-seed) setting on BUSZS Dataset. The results are reported using the notation (I-AUROC/P-AUROC). Our final method achieves scores of (82.3/93.4). Specifically, we examine the impact of employing US-specific terminologies by using US-unaware prompts to generate $\mathbf{P}_t$, resulting in scores of (72.3/88.6). To verify the functionality of Memory-Boosted Few-Shot Adaptation, we remove classification and focus on segmentation loss, reducing localization performance and yielding (-/84.3) without joint learning. To verify the Patch-Wise Image-Language Fusion, we omit the Patch-Wise Image-Language Fusion module, instead directly

utilizing the linearly projected patch tokens f'_p for anomaly localization, resulting in (71.2/90.3).

4 Conclusion

We introduce UltraAD, an innovative framework that enables simultaneous abnormal localization and fine-grained classification in US imaging through few-shot adaptation of the CLIP model. Few-shot adaptation was applied to a breast dataset and validated on two unseen datasets with different machines, patients, and probes, demonstrating robustness in real-world scenarios. An ablation study further validates the effectiveness of the memory-boosted few-shot adaptation and PIF module. The promising results on breast US images highlight the potential for developing a generalized model for US imaging, with future extensions to diverse anatomies using corresponding datasets.

Acknowledgments. This work was supported in part by the Multi-Scale Medical Robotics Center, AIR@InnoHK, Hong Kong; and in part by the SINO-German Mobility Project under Grant M0221.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Evaluating normative representation learning in generative ai for robust anomaly detection in brain imaging. Nature Communications **16**(1), 1624 (2025)
3. Bi, Y., Huang, L., Clarenbach, R., Ghotbi, R., Karlas, A., Navab, N., Jiang, Z.: Synomaly noise and multi-stage diffusion: A novel approach for unsupervised anomaly detection in ultrasound imaging (2024), `https://arxiv.org/abs/2411.04004`
4. Bi, Y., Jiang, Z., Duelmer, F., Huang, D., Navab, N.: Machine learning in robotic ultrasound imaging: Challenges and perspectives. Annual Review of Control, Robotics, and Autonomous Systems **7**
5. Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In: European Conference on Computer Vision. pp. 55–72. Springer (2024)
6. Chen, T., Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Text-driven adaptation of foundation models for few-shot surgical workflow analysis (2025), `https://arxiv.org/abs/2501.09555`
7. Chen, X., Han, Y., Zhang, J.: April-gan: 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)
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)

9. Huang, C., Jiang, A., Feng, J., Zhang, Y., Wang, X., Wang, Y.: Adapting visual-language models for generalizable anomaly detection in medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11375–11385 (2024)
10. Huang, D., Bi, Y., Navab, N., Jiang, Z.: Motion magnification in robotic sonography: Enabling pulsation-aware artery segmentation. In: IROS. pp. 6565–6570. IEEE (2023)
11. Huang, D., Li, C., Karlas, A., Chu, X., Au, K.S., Navab, N., Jiang, Z.: Vibnet: Vibration-boosted needle detection in ultrasound images. IEEE Transactions on Medical Imaging (2025)
12. Huang, Y., Shakeri, F., Dolz, J., Boudiaf, M., Bahig, H., Ben Ayed, I.: Lp++: A surprisingly strong linear probe for few-shot clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23773–23782 (2024)
13. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: CVPR. pp. 19606–19616 (2023)
14. Jiang, Z., Bi, Y., Zhou, M., Hu, Y., Burke, M., Navab, N.: Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations. The International Journal of Robotics Research **43**(7), 981–1002 (2024)
15. Jiang, Z., Kang, Y., Bi, Y., Li, X., Li, C., Navab, N.: Class-aware cartilage segmentation for autonomous us-ct registration in robotic intercostal ultrasound imaging. IEEE Transactions on Automation Science and Engineering (2024)
16. Jiang, Z., Salcudean, S.E., Navab, N.: Robotic ultrasound imaging: State-of-the-art and future perspectives. Medical image analysis **89**, 102878 (2023)
17. Lee, H., Park, J., Hwang, J.Y.: Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image. IEEE transactions on ultrasonics, ferroelectrics, and frequency control **67**(7), 1344–1353 (2020)
18. Li, X., Huang, D., Zhang, Y., Navab, N., Jiang, Z.: Semantic scene graph for ultrasound image explanation and scanning guidance (2025), https://arxiv.org/abs/2506.19683
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
20. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
21. Qu, Z., Tao, X., Prasad, M., Shen, F., Zhang, Z., Gong, X., Ding, G.: Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In: European Conference on Computer Vision. pp. 301–317. Springer (2024)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
23. Tang, F., Wang, L., Ning, C., Xian, M., Ding, J.: Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In: 2023 IEEE 20th international symposium on biomedical imaging (ISBI). pp. 1–5. IEEE (2023)
24. Vallez, N., Bueno, G., Deniz, O., Rienda, M.A., Pastor, C.: Bus-uclm: Breast ultrasound lesion segmentation dataset. Scientific Data **12**(1), 242 (2025)
25. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6757–6767 (2023)

26. You, C., Zhao, R., Liu, F., Dong, S., Chinchali, S., Topcu, U., Staib, L., Duncan, J.: Class-aware adversarial transformers for medical image segmentation. Advances in neural information processing systems **35**, 29582–29596 (2022)
27. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. Medical Image Analysis p. 103644 (2025)
28. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
29. Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N., Zhou, X.: Mediclip: Adapting clip for few-shot medical image anomaly detection (2024), `https://arxiv.org/abs/2405.11315`
30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
32. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. arXiv preprint arXiv:2310.18961 (2023)
33. Zhuang, Z., Li, N., Joseph Raj, A.N., Mahesh, V.G., Qiu, S.: An rdau-net model for lesion segmentation in breast ultrasound images. PloS one **14**(8), e0221535 (2019)