https://doi.org/10.7840/kics.2025.50.10.1505

Few-Shot Anomaly Detection for Medical Ultrasound Images Using Metric Learning and Multimodal BiomedCLIP Embeddings

Haeyun Lee◆, Kyungsu Lee*, Jihun Kim°

ABSTRACT

Medical ultrasound imaging is extensively utilized in clinical practice due to its advantages of safety, cost-effectiveness, and real-time imaging capability. Nevertheless, inherent issues such as low signal-to-noise ratios, operator dependency, and speckle noise introduce significant challenges in automated anomaly detection. To overcome these limitations, we propose a novel few-shot anomaly detection framework specifically designed for medical ultrasound imaging. Our method employs BiomedCLIP, a multimodal model tailored for biomedical applications, to jointly encode ultrasound images and clinically relevant textual descriptions into semantically rich embeddings. Subsequently, these embeddings are refined through a projection network to create compact, discriminative representations optimized for anomaly classification. A prototype-based metric learning approach further enhances the separability of these embeddings by explicitly clustering normal and abnormal cases. Extensive evaluations conducted on representative ultrasound datasets demonstrate that our proposed method achieves superior anomaly detection performance compared to existing contrastive and multimodal learning frameworks, particularly in severely limited data scenarios. Our findings underscore the efficacy and clinical potential of combining multimodal embeddings and metric learning for robust and interpretable anomaly detection in medical ultrasound images.

Key Words : Few-shot Learning, Anomaly Detection, Ultrasound Images, Metric Learning

Ⅰ. Introduction

Medical ultrasound imaging is widely used in clinical diagnosis due to its significant advantages, such as safety, real-time capability, cost-effectiveness, and absence of ionizing radiation[1]. It plays a pivotal role across numerous medical fields, including cardiology, oncology, obstetrics, gynecology, vascular imaging, and emergency medicine. Despite these clear advantages, ultrasound imaging presents inherent difficulties for automated interpretation. Common issues include speckle noise, poor contrast resolution, significant operator variability, and image artifacts resulting from patient movement or limitations of the imaging equipment[2]. These factors complicate the accurate automated detection of anomalies.

Automated anomaly detection-identifying pathological conditions within medical images-is crucial in clinical practice, because timely and precise diagnosis directly influences patient outcomes. Traditional machine learning methods rely on handcrafted features; however, these approaches have limitations because of their inability to capture the complex patterns inherent in medical images[3,4]. In contrast, deep learning

※ This work was supported by the new professor research program of KOREATECH in 2024 and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-00556581).
◆ First Author : KOREATECH, School of Computer Science and Engineering, haeyun@koreatech.ac.kr, 정회원
° Corresponding Author : Kangnam University, Division of Electronic and Semiconductor Engineering, Major in Electronic Engineering, jihunk@kangnam.ac.kr, 정회원
* Jeonbuk National University, Department of Computer Science & Artificial Intelligence, ksl@jbnu.ac.kr
논문번호 : 202504-075-A-RN, Received April 2, 2025; Revised May 19, 2025; Accepted May 28, 2025

approaches significantly advance anomaly detection by directly learning meaningful features from raw data. However, deep learning methods typically require large, meticulously labeled datasets, the creation of which is challenged by ethical considerations, patient privacy constraints, and the limited availability of expert annotators[5].

To address the challenges associated with data scarcity, Few-Shot Learning (FSL) has emerged as an effective approach. FSL methods focus on training models to perform well even when only a very limited amount of labeled data is available. By leveraging prior knowledge obtained from related tasks or extensive datasets, FSL significantly reduces the required volume of annotated data. Recent advances in FSL, such as Prototypical Networks[6], which classify new data based on proximity to prototype representations, and Model-Agnostic Meta-Learning (MAML)[7], which rapidly adapts models with minimal gradient updates, have demonstrated notable success in medical image analysis tasks.

Concurrently, Contrastive Learning (CL), a form of self-supervised learning, has shown considerable promise for learning robust representations without relying on extensive labeled datasets. The fundamental principle of CL involves training models to pull similar samples closer in a learned embedding space, while pushing dissimilar samples apart. Prominent CL methods, including SimCLR[8], Momentum Contrast (MoCo)[9], and Bootstrap Your Own Latent (BYOL)[10], have exhibited exceptional performance in various vision tasks by utilizing augmented views of data and contextually similar pairs.

Integrating Few-Shot Learning with Contrastive Learning paradigms presents a compelling solution for medical imaging analysis, particularly when annotated data are limited. The additional integration of multimodal data, particularly textual descriptions accompanying medical images, further enhances the model interpretability and diagnostic accuracy. By aligning textual and visual information, these multimodal methods facilitate the learning of more informative context-aware embeddings critical to clinical diagnostics.

Contrastive Language-Image Pretraining (CLIP)[11] is a prominent multimodal CL approach that learns joint representations of images and textual descriptions, thereby demonstrating remarkable success in tasks such as zero-shot classification and general image understanding. BiomedCLIP extends the CLIP model specifically to biomedical applications, integrating visual and textual biomedical data into a unified representation and effectively capturing semantic nuances critical for medical image analysis[12]. Biomed-CLIP provides rich semantic representations, enhanced generalization, and robust performance across diverse biomedical tasks, making it particularly suitable for few-shot learning scenarios in the medical context.

In this study, we leveraged the advantages of BiomedCLIP and Few-Shot Learning to propose a novel anomaly detection framework tailored explicitly for medical ultrasound imaging. Our proposed method aims to maximize the semantic coherence between the ultrasound image data and clinical textual descriptions, thereby enhancing the discriminative power of the learned embeddings. Extensive evaluations across a representative medical ultrasound dataset highlight significant improvements in the detection performance, reinforcing the practical applicability of our proposed approach in clinical environments.

Ⅱ. Preliminary

In this section, we review the key concepts and recent research trends that serve as the basis for understanding this study. In particular, we will explain in depth the three core concepts utilized in this study: Few-Shot Learning (FSL), Contrastive Learning (CL), and biomedCLIP.

2.1 Few-Shot Learning (FSL)

FSL addressed the critical challenge of effectively training machine learning models when only a limited number of labeled examples per class are available. This scenario frequently arises in medical imaging, where obtaining labeled data is costly and time consuming. FSL methods broadly fall into three categories: metric-based, optimization-based, and hallucination-based methods.

Metric-based methods learn an embedding space in which samples belonging to the same class cluster closely and samples from different classes are separated. For instance, Prototypical Networks compute class prototypes by averaging the feature embeddings of labeled samples within each class[6]. Formally, the prototype c_k for class k is computed as $c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} f_\theta(x_i)$, where S_k denotes labeled samples of class k and $f_\theta(x_i)$ represents the embedding function parameterized by θ. The classification of new samples is then performed by identifying the nearest class prototype in the learned embedding space.

Optimization-based methods, such as Model-Agnostic Meta-Learning (MAML), aim for the rapid adaptation of models to new tasks using minimal gradient updates[7]. The MAML learns a set of initial model parameters that enable effective adaptation to new tasks with only a few gradient steps. This approach effectively reduces the number of labeled samples required by reusing the knowledge learned across multiple related tasks.

Hallucination-based methods generate synthetic or augmented data to alleviate data-scarcity issues[13]. These methods often involve data transformations or generative models to simulate realistic training samples, significantly enhancing model robustness and generalization despite the limited labeled data.

However, these existing few-shot learning methods often face difficulties when applied to medical imaging scenarios, particularly due to image variability, speckle noise, and operator dependency inherent in ultrasound imaging, limiting their robustness and clinical applicability[14].

In our approach, we utilized a metric-based FSL to explicitly construct class prototypes from ultrasound images, facilitating robust anomaly detection through distance-based classification.

2.2 Contrastive Learning (CL)

Contrastive Learning is a type of unsupervised learning or self-supervised learning that operates by increasing the similarity between similar samples and decreasing the similarity between different samples to learn the inherent representation of data[15]. At this time, similarity is usually defined as the relationship between identical samples generated by different augmentation methods. The representative algorithms of Contrastive Learning include the following.

SimCLR is one of the most influential CL approaches and employs a simplified, yet effective methodology[8]. It generates augmented versions of each input image using various transformations, thereby creating pairs of positive samples. These pairs were then contrasted with a large set of negative examples (different images). The SimCLR framework maximizes the similarity between augmented views (positive pairs) while minimizing the similarity with other distinct samples (negative pairs), utilizing a contrastive loss function. This approach results in the learning of highly discriminative and generalized feature embeddings that are suitable for downstream tasks.

Momentum Contrast is an influential contrastive learning method that maintains a dynamic queue of previously processed embeddings as negative samples by utilizing a momentum encoder to ensure consistency and enhance feature representations[9]. The momentum encoder is updated as a weighted average of the query encoder weights (θ_q) and the previous momentum encoder weights (θ_k), as shown in the following equation: $\theta_k \leftarrow m\theta_k + (1 - m)\theta_q$, where m is the momentum coefficient. Nevertheless, Momentum Contrast's dependence on an extensive memory bank of negative samples may present difficulties in few-shot learning contexts, in which data availability is constrained. Additionally, typical contrastive learning models, such as SimCLR and MoCo, are primarily tailored for natural image datasets, which means they often fail to address the unique challenges of medical imaging. This is particularly true for ultrasound images, which are marked by noise, artifacts, and subtle anomalies[16].

2.3 BiomedCLIP

Recently, multimodal approaches that learn embeddings by combining text and images have attracted considerable attention in the field of medical AI. A representative example is Contrastive Language-Image Pretraining (CLIP)[11]. CLIP is designed to learn rich multimodal representations through seman-

tic relationships between images and texts and generally shows excellent performance in image classification and zeroshot classification. BiomedCLIP is an adaptation of the CLIP approach tailored explicitly for biomedical applications[12]. Unlike traditional CLIP, which handles general image-text pairs, BiomedCLIP integrates biomedical images with medically specialized textual data, such as clinical annotations and medical literature excerpts. By aligning these multimodal biomedical embeddings within a shared semantic space, Biomed-CLIP effectively captured detailed biomedical semantics and context. Consequently, it provides rich semantic features, enhanced model generalization, and improved robustness. This makes BiomedCLIP particularly advantageous for few-shot learning scenarios prevalent in medical image analysis, enabling accurate and interpretable model predictions, even with limited annotated data.

In this study, the BiomedCLIP was used as a base model to maximize the anomaly detection performance by utilizing the rich expressive power of multimodal information in ultrasound medical images.

Ⅲ. Proposed Method

In this section, we describe in detail our proposed approach for anomaly detection specifically tailored for medical ultrasound imaging. Our framework effectively integrates few-shot learning and metric learning methodologies, leveraging the capabilities of Biomed-CLIP to achieve robust and accurate anomaly detection despite the inherent data scarcity commonly encountered in medical imaging tasks. The core of our approach relies on three essential components: First, the BiomedCLIP encoder is utilized to process multimodal data, encoding ultrasound images along with corresponding clinical textual descriptions into shared multimodal embeddings. These embeddings encapsulate critical semantic information pertinent to both visual anomalies and their associated medical contexts. Second, we introduce a carefully designed projection network that refines these multimodal embeddings by mapping them into a discriminative, lower-dimensional representation space explicitly optimized for anomaly detection. Lastly, a prototype-based metric learning loss mechanism, based on an adapted prototypical network framework, is employed to enhance the discriminability of the resulting embeddings[6]. This loss function explicitly promotes tighter clustering of embeddings belonging to the same class (normal or abnormal) while maximizing their distances to embeddings from the other class.

3.1 Model Architecture

Figure 1 shows the proposed architecture based on the BiomedCLIP model. We employ the pretrained BiomedCLIP model, which consists of a vision transformer-based vision encoder[17] and a PubMedBERT-based textual encoder[18], to generate embeddings from ultrasound images and clinically relevant textual prompts. For each ultrasound image, embeddings are extracted by passing it through the visual encoder. Concurrently, textual embeddings are computed from clinically relevant descriptions (e.g., "normal ultrasound image", "abnormal ultrasound image with lesion"), producing informative semantic representations.

We propose a carefully designed projection network to effectively map multimodal embeddings obtained from the BiomedCLIP model into a discriminative representation space suitable for anomaly detection. This projection network consists of three fully-connected layers, each coupled with batch normalization and nonlinear activation functions (ReLU). Specifically, given an initial multimodal embedding $x \in \mathbb{R}^{512}$, the projection network progressively refines and compresses this embedding into a lower-dimensional representation as follows:

First, the original 512-dimensional embedding vector x obtained from BiomedCLIP is transformed through a linear layer that reduces its dimensionality to a hidden dimension of 256. Subsequently, batch normalization is applied to stabilize learning, followed by a ReLU activation to introduce nonlinearity and enhance representational capacity. This first transformation can be mathematically represented as $x^{(1)} = ReLU\,(Bat\,chNorm(W_1 x + b))$, where $W_1 \in \mathbb{R}^{256\times512}$ and $b_1 \in \mathbb{R}^{256}$. Next, the resulting intermediate representation $x(1)$ undergoes a second linear transformation, further reducing the dimensionality from

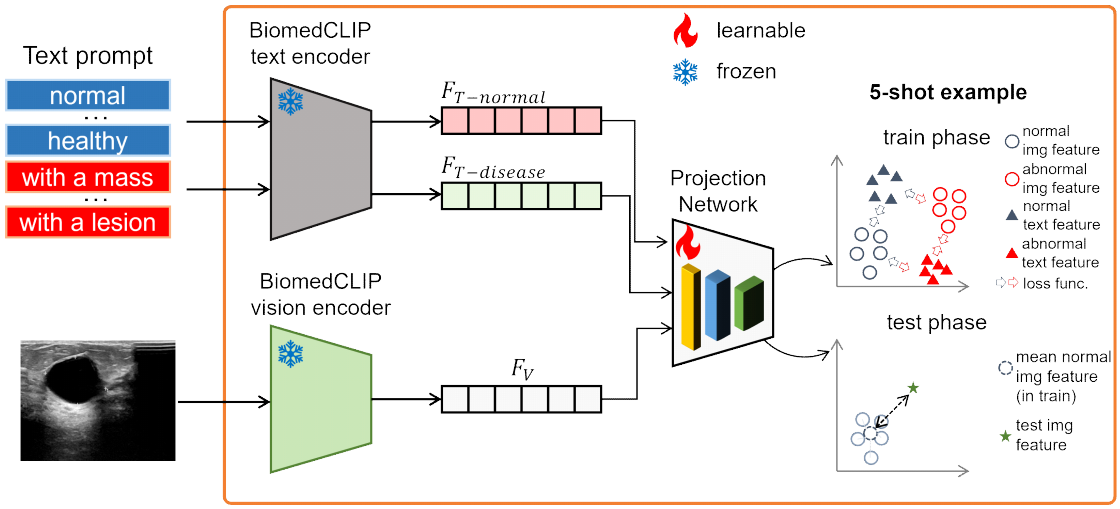


Fig. 1. Illustration of the proposed metric-learning-based few-shot anomaly detection method for medical ultrasound images.

256 to 128. Again, this step employs batch normalization and ReLU activation, ensuring the embeddings maintain discriminative power while remaining compact. Formally, this step is expressed as $x^{(2)} = ReLU(BatchNorm(W_2x^{(1)} + b_2))$, with parameters $W_2 \in \mathbb{R}^{128\times256}$ and $b_2 \in \mathbb{R}^{128}$ Finally, a third linear transformation is applied to map the embedding into a final, discriminative space of 128 dimensions, specifically optimized for anomaly detection. Unlike previous layers, this final layer does not use an activation function, preserving the embedding' s linear properties for subsequent similarity computations. This final embedding is mathematically given as $z = W_3x^{(2)} + b_3$, where $W_3 \in \mathbb{R}^{128\times128}$ and $b_3 \in \mathbb{R}^{128}$. The resulting embeddings z facilitate effective discrimination between normal and abnormal ultrasound images by clearly clustering semantically similar data points and separating dissimilar ones.

The projection network proposed here projects the feature maps of medical ultrasound images that pass through the vision encoder in BiomedCLIP and the feature maps of text prompts that pass through the text encoder in medCLIP into one space. In this space, during training, the features of normal text and normal images are mapped to similar spaces, the features of abnormal text and abnormal images are mapped to similar spaces, and the features of normal and abnormal features are mapped to distant spaces. During the evaluation, the average features of the data used in the training are mapped to values between 0 and 1, depending on how far they are from each other.

3.2 Loss Function

In our proposed framework, we leverage an adapted prototype-based metric learning loss function inspired by Prototypical Networks[6] to enhance the anomaly detection performance, particularly in scenarios with limited annotated ultrasound images. The core intuition behind this approach is to explicitly generate representative embeddings, known as prototypes, for each class and to optimize the embedding space by minimizing the distances between each embedding and its corresponding class prototype.

Formally, given a batch of training data consisting of N samples with corresponding labels $\{(x_i, y_i)\}_{i=1}^{N}$, the embeddings of these samples, denoted as $f_\theta(x_i) \in \mathbb{R}^D$, are computed through our backbone encoder followed by the proposed projection head. Here, D represents the embedding dimension of 128 in our specific implementation. We first compute class-specific prototypes ck by taking the mean embedding of all samples belonging to the same class within each batch. Mathematically, for each unique class label $k \in \{0, 1\}$, the prototype c_k is defined as follows:

$$c_k = \frac{1}{\|S_k\|}\Sigma_{x_i \in S_k} f_\theta(x_i) \quad (1)$$

where S_k denotes the set of samples in the batch belonging to class k. Once these prototypes are computed, we measure the Euclidean distance between each sample embedding $f_\theta(x_i)$ and each of the prototypes c_k. Specifically, for each embedding in the batch, the distance to all class prototypes is computed as:

$$d(f_\theta(x_i), c_k) = \|f_\theta(x_i) - c_k\|_2 \quad (2)$$

To optimize the embeddings, we employ a cross-entropy loss function based on these distances, ensuring that embeddings cluster tightly around their respective class prototypes and remain distinct from prototypes of the opposite class. The cross-entropy loss function is formulated as:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(-d(f_\theta(x_i), c_{y_i}))}{\sum_k \exp(-d(f_\theta(x_i), c_k))} \quad (3)$$

Here, y_i is the ground-truth class label for the sample x_i. Minimizing this loss ensures embeddings from the same class are closely grouped and clearly differentiated from embeddings of the opposite class, significantly enhancing anomaly detection performance.

3.3 Training Details

To ensure reproducibility and clearly illustrate the training methodology of our proposed framework, we provide comprehensive implementation details here in a descriptive form. The BiomedCLIP model used as the foundation of our approach comprises a ViT-Base visual encoder and a PubMedBERT textual encoder, with an original embedding dimension of 512. The proposed projection network detailed above maps these 512-dimensional embeddings down to a final embedding dimension of 128, leveraging a hidden layer dimension of 256 with intermediate batch normalization and ReLU activations to enhance feature quality.

For training optimization, we adopt the AdamW optimizer with an initial learning rate set at 1×10^{-4} and a weight decay parameter also at 1×10^{-4}, which effectively regularizes the model parameters. The training procedure employs a batch size of 32 image-text pairs, and the entire training spans a total of 200 epochs, sufficient for convergence and robust representation learning.

Data augmentation strategies play a critical role during training in improving the robustness and generalization of the model. Specifically, each ultrasound image is randomly subjected to horizontal flipping with a probability of 0.5, random rotations within a range of 10 degrees, and color jittering that includes slight variations in brightness, contrast, saturation, and hue. Then, all the images are uniformly resized to a resolution of 224×224 pixels.

Ⅳ. Experiments

In this section, we describe the experimental setting and dataset used to evaluate the performance of our approach. We evaluate our proposed method using a publicly available medical ultrasound imaging dataset to demonstrate its applicability and robustness. Specifically, we selected the breast ultrasound imaging (BUSI) dataset[19], one of the most widely-used datasets in medical ultrasound research, to validate the effectiveness of our method in anomaly detection. This dataset contains labeled normal and abnormal cases, providing an appropriate basis for evaluating the performance and clinical utility of our proposed approach in ultrasound anomaly detection.

To evaluate our approach, we selected a simple framework for contrastive learning (SimCLR)[8], and a version of our approach in which BiomedCLIP is replaced by CLIP[11]. The SimCLR is widely recognized as a leading and highly effective method for few-shot learning, and numerous studies have selected it for comparative analyses. We also demonstrated the effectiveness of BiomedCLIP by comparing it with a version in which we replaced CLIP with our proposed method.

We adopt a few-shot evaluation protocol, randomly selecting a small set of normal samples (k-shots) for training. In contrast, evaluation involves detecting anomalies in a balanced set of normal and abnormal images. To ensure reliability, experiments are repeated across five random seeds, and the mean AUC (area under the curve) of ROC (receiver operating charac-

teristic) is reported as the primary evaluation metric. We conducted comparative experiments using 1-shot, 2-shot, and 4-shot settings. We trained and evaluated each shot five times and compared the average performances.

Table 1 shows the few-shot classification results of our approach and comparison methods. As shown in Table 1, our proposed method significantly outperforms the baseline SimCLR-based methods across all tested scenarios (1-shot, 2-shot, and 4-shot). Specifically, in the 1-shot scenario, our method achieves an AUROC of 0.8004±0.0299, marking a noticeable improvement over the SimCLR-CLIP (0.5499±0.1907) and SimCLR-BiomedCLIP (0.6627±0.0669) methods. This substantial margin demonstrates that our model can efficiently leverage limited labeled data, a crucial advantage in medical imaging applications where labeled samples are often scarce.

When comparing the results of our proposed method, the incorporation of BiomedCLIP showed a clear performance advantage over standard CLIP implementation. In the 1-shot scenario, our BiomedCLIP variant (0.8004±0.0299) outperformed our standard CLIP variant (0.7087±0.0262), illustrating the benefit of integrating domain-specific biomedical knowledge into the feature-encoding process. Similarly, this pattern consistently persisted in the 2-shot and 4-shot settings, with Ours-BiomedCLIP obtaining AUROC values of 0.8289 ± 0.0370 and 0.8402 ± 0.0325, respectively, compared to 0.7167 ± 0.0144 and 0.7657 ± 0.0396, respectively, achieved by Ours-CLIP.

Table 1. A quantitative comparison of different methods in terms of the AUROC on the BUSI dataset. The best performance is in **bold**.

Methods		AUROC
3*SimCLR-CLIP	1-shot	0.5499 ± 0.1907
	2-shot	0.6187 ± 0.0825
	4-shot	0.5445 ± 0.0438
3*SimCLR-BiomedCLIP	1-shot	0.6627 ± 0.0669
	2-shot	0.7207 ± 0.0608
	4-shot	0.7477 ± 0.0325
3*Ours-CLIP	1-shot	0.7087 ± 0.0266
	2-shot	0.7167 ± 0.0124
	4-shot	0.7657 ± 0.0396
3*Ours-BiomedCLIP	1-shot	**0.8004 ± 0.0299**
	2-shot	**0.8289 ± 0.0370**
	4-shot	**0.8402 ± 0.0325**

Table 2. A comparison of training and inference times based on 4-shot setting.

Methods	Training time(s)	Inference Time(s)
SimCLR-CLIP	383.802	0.206
SimCLR-BiomedCLIP	338.559	0.201
Ours-CLIP	390.474	0.214
Ours-BiomedCLIP	354.879	0.211

We also compared the computational costs of our method with those of other comparison methods. In terms of computational complexity, our proposed BiomedCLIP-based framework required a training time of approximately 354.879 seconds for 200 epochs using an NVIDIA RTX 4090 GPU. This training time represents a 4.82% increase compared with the baseline SimCLR-BiomedCLIP model (338.559 seconds). However, inference speeds remained competitive at approximately 211 ms per image, comparable to baseline models (201 ms), thus ensuring clinical practicality and feasibility.

Overall, these experimental results confirm the efficacy and superiority of our proposed approach in anomaly detection tasks, particularly emphasizing the critical role of BiomedCLIP in enhancing the feature representations for medical imaging analysis. The consistent and significant improvement over the SimCLR-based methods further underscores the robustness of our few-shot learning methodology.

Ⅴ. Conclusions

In this study, we introduced a novel few-shot anomaly detection framework for medical ultrasound imaging, leveraging multimodal contrastive learning with BiomedCLIP. By combining the strengths of few-shot learning paradigms and contrastive multimodal embedding approaches, our method effectively addresses the critical challenge of limited labeled data in medical image analysis. The proposed architecture,

built upon the BiomedCLIP encoder, efficiently integrates ultrasound image data with clinically relevant textual prompts, thereby enhancing the semantic coherence and discriminative power of the learned embeddings.

Extensive experimental evaluations conducted on representative medical ultrasound datasets demonstrated that our method achieves superior performance compared with state-of-the-art contrastive learning baselines, particularly under severely limited labeled data scenarios (1-shot, 2-shot, and 4-shot settings). The inclusion of BiomedCLIP notably contributed to the robustness and interpretability of the model, resulting in significant improvements in anomaly detection accuracy, as quantified by AUROC metrics.

Our results underscore the effectiveness and practical utility of leveraging domain-specific multimodal pretrained models in medical anomaly detection. Furthermore, this methodology shows promise for future extensions to broader medical applications, including the detection and diagnosis of diverse diseases such as pneumonia, Kawasaki disease, and hepatocellular carcinoma, thereby contributing meaningfully to improving clinical decision-making and patient outcomes. In addition, a comparative study involving gradient-based meta-learning methods, particularly Model-Agnostic

Meta-Learning (MAML), will be conducted to provide deeper insights into the relative advantages and limitations of prototype-based metric learning versus gradient-based adaptation methods. Such comparative analyses will inform decisions regarding the selection of optimal few-shot learning strategies under various clinical conditions and further refine the clinical applicability of our proposed approach.

References

[1] B. B. Goldberg, "International arena of ultrasound education," *J. Ultrasound in Medicine*, vol. 22, no. 6, pp. 549-551, 2003.

[2] H. Lee, M. H. Lee, S. Youn, K. Lee, H. M. Lew, and J. Y. Hwang, "Speckle reduction via deep content-aware image prior for precise breast tumor segmentation in an ultrasound image," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 9, pp. 2638-2650, 2022.

[3] H. Lee, J. Park, and J. Y. Hwang, "Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 7, pp. 1344-1353, 2020.

[4] D. Kwak, J. Choi, and S. Lee, "Method of the breast cancer image diagnosis using artificial intelligence medical images recognition technology network," *J. KICS*, vol. 48, no. 2, pp. 216-226, 2023.

[5] K. H. Le, T. V. Tran, H. H. Pham, H. T. Nguyen, T. T. Le, and H. Q. Nguyen, "Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis," *IEEE Access*, vol. 11, pp. 14105-14114, 2023.

[6] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in NIPS*, vol. 30, 2017.

[7] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *Int. Conf. Machine Learn., PMLR*, pp. 1126-1135, 2017.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *Int. Conf. Machine Learn., PMLR*, pp. 1597-1607, 2020.

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. CVPR*, pp. 9729-9738, 2020.

[10] J.-B. Grill, F. Strub, F. Altché, et al., "Bootstrap your own latent: A new approach to self-supervised learning," *Advances in NIPS*, vol. 33, pp. 21271-21284, 2020.

[11] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," *Int. Conf. Machine Learn., PMLR*, pp. 8748-8763, 2021.

[12] S. Zhang, Y. Xu, N. Usuyama, et al., "Biomedclip: A multimodal biomedical foun-

dation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[13] K. Li, Y. Zhang, K. Li, and Y. Fu, "Adversarial feature hallucination networks for few-shot learning," in *Proc. IEEE/CVF Conf. CVPR*, pp. 13470-13479, 2020.

[14] E. Pachetti and S. Colantonio, "A systematic review of few-shot learning in medical imaging," *Artificial Intell. Medicine*, p. 102949, 2024.

[15] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193 907-193 934, 2020.

[16] Z. Fu, J. Jiao, R. Yasrab, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Anatomy-aware contrastive representation learning for fetal ultrasound," *ECCV*, pp. 422-436, Springer, 2022.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Y. Gu, R. Tinn, H. Cheng, et al., "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Computing for Healthcare* (*HEALTH*), vol. 3, no. 1, pp. 1-23, 2021.

[19] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.

Haeyun Lee

Feb. 2016 : B.S. Mathematics, Jeonbuk National University, Jeonju, South Korea

Feb. 2018 : M.S. Information & Communication Engineering, DGIST, Daegu, South Korea

Feb. 2022 : Ph.D. Information & Communication Engineering, DGIST, Daegu, South Korea

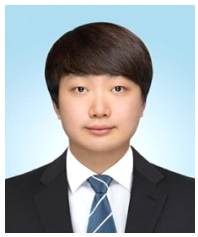
Mar. 2022-Feb. 2024 : Staff Engineer, Samsung SDI, Yongin, South Korea

Mar. 2024-Current : Assistant Professor, School of Computer Science and Engineering, KOREATECH, Cheonan, South Korea

<Research Interest> Computational Photography, Medical Image Analysis, Artificial Intelligence

[ORCID 0000-0002-7572-1705]

Kyungsu Lee



Feb. 2018 : B.S., Dept. of Computer Science and Electrical Engineering, Handong Global Uni- versity, Pohang, South Korea

Feb. 2023 : Ph.D. Information & Communication Engineering, DGIST, Daegu, South Korea

Mar. 2024-Current : Assistant Professor, Department of Computer Science & Artificial Intelligence, Jeonbuk National University, Jeonju, South Korea

<Research Interests> Medical Imaging, Artificial Intelligence.

[ORCID:0009-0000-7223-0903]

Jihun Kim

Feb. 2015 : B.S., Dept. of Electronic Engineering, Hannam University, Daejeon, South Korea

Aug. 2019: Ph.D, Dept. of Information & Communication Engneering, DGIST, Daegu, South Korea

Aug. 2019-Dec. 2020 : Postdoc, University of Notre Dame, IN, USA

Jan. 2021-Mar. 2022 : Postdoc, University of Illinois at Urbana-Champaign, IL, USA

Apr. 2022-Current : Assistant Professor, Kangnam University, Yongin, South Korea

<Research Interests> Ultrasound Imaging, Artificial Intenlligence

[ORCID:0000-0003-3833-4201]