Masked Modeling-Based Ultrasound Image Classification via Self-Supervised Learning

Kele Xu, *Member, IEEE*, Kang You, Boqing Zhu, Ming Feng, Dawei Feng, and Cheng Yang

***Abstract*—Recently, deep learning-based methods have emerged as the preferred approach for ultrasound data analysis. However, these methods often require large-scale annotated datasets for training deep models, which are not readily available in practical scenarios. Additionally, the presence of speckle noise and other imaging artifacts can introduce numerous hard examples for ultrasound data classification. In this paper, drawing inspiration from self-supervised learning techniques, we present a pre-training method based on mask modeling specifically designed for ultrasound data. Our study investigates three different mask modeling strategies: random masking, vertical masking, and horizontal masking. By employing these strategies, our pre-training approach aims to predict the masked portion of the ultrasound images. Notably, our method does not rely on externally labeled data, allowing us to extract representative features without the need for human annotation. Consequently, we can leverage unlabeled datasets for pre-training. Furthermore, to address the challenges posed by hard samples in ultrasound data, we propose a novel hard sample mining strategy. To evaluate the effectiveness of our proposed method, we conduct experiments on two datasets. The experimental results demonstrate that our approach outperforms other state-of-the-art methods in ultrasound image classification. This indicates the superiority of our pre-training method and its ability to extract discriminative features from ultrasound data, even in the presence of hard examples.**

***Index Terms*—Pre-training, self-supervised, ultrasound image, masked modeling.**

***Impact Statement*—An ultrasound images classification approach via mask modeling without human annotation.**

I. Introduction

OVER the past few decades, multiple medical imaging modalities have been used to create images of the human body, such as Computed Tomography (CT), Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), and ultrasound, all of which are widely used in various clinical settings [1]. Compared with other imaging modalities, medical ultrasound imaging has the following advantages: non-invasiveness, real-time imaging, and cost-effectiveness [2]. Therefore, ultrasound is one of the widespread methods for visualizing human soft tissue. However, despite sustainable efforts, the interpretation of ultrasound data remains challenging in the real-world settings due to contamination from speckle noise, hidden fields of view, signal dropout, and other imaging artifacts [3]. The potential applications for efficient and accurate ultrasound data analysis seem to be evident in many fields [4], [5], [6].

Classification task is essential for ultrasound interpretation, which aims to extract representative features and distinguish different ultrasound signals (either radio-frequency (RF) signals or ultrasound images), such as tissue classification [7]. Leveraging the data-driven classifiers, traditional shallow-architecture classification algorithms can distinguish different kinds of ultrasound data. Ultrasound signal classification has important implications for many practical applications, from cancer diagnosis to cardiovascular disease diagnosis [8], [9].

Due to the practical clinical need for ultrasound data interpretation, automatic and accurate classification has drawn increasing interest in the past few years. For example, during natural speech production, ultrasound tongue imaging (UTI) is one of the appealing ways for vocal tract modeling [10], as it can capture the tongue movement at a high frame rate (60 Hz or higher). Moreover, UTI does not expose the speakers to radiation and the machines are of lower costs presently [11], [12]. Automatic classification of tongue gesture shapes from raw ultrasound can facilitate the understanding of speech production, which has attracted increasing attention during the last years [11], [13]. Previous attempts employed the Principal Component Analysis (PCA) [14], Discrete Cosine Transform (DCT) [15], AutoEncoder [16] for the feature extraction in the UTI, leveraging the unsupervised learning manner. Since the revolution of deep learning, convolutional neural network (CNN)-based supervised learning has been successfully applied in UTI processing [17], [18], [19], [20]. Generally speaking, supervised deep learning often requires a large number of labeled examples [21], which is difficult to obtain in practical settings. The effective utilization of large-scale unlabeled ultrasound data for representation learning necessitates comprehensive investigation. In recent years, self-supervised learning (SSL) has shown significant advancements [22], enabling the learning of informative representations

Manuscript received 11 January 2024; revised 21 February 2024 and 4 March 2024; accepted 5 March 2024. Date of publication 12 March 2024; date of current version 1 April 2024. This work was supported by the major Science and Technology Innovation 2030 “New Generation Artificial Intelligence” Project under Grant 2021ZD0112904. The review of this article was arranged by Editor Hassan Rivaz. *(Kele Xu and Kang You contributed equally to this work.) (Corresponding authors: Boqing Zhu; Ming Feng.)*

Kele Xu, Boqing Zhu, Dawei Feng, and Cheng Yang are with the National University of Defense Technology, Changsha 410073, China (e-mail: zhuboqing@outlook.com).

Kang You is with the Shanghai Jiao Tong University, Shanghai 200240, China.

Ming Feng is with the TongJi University, Shanghai 200070, China (e-mail: andrew_tal@163.com).

Digital Object Identifier 10.1109/OJEMB.2024.3374966

from unlabeled data without the need for human annotation. The SSL paradigm for ultrasound is under-explored in previous studies.

In this paper, we explore the SSL paradigm for ultrasound image classification. In the pre-training stage, we proposed three different mask modeling strategies, with the goal to learn discriminative features which can be further deployed to ultrasound image classification through a fine-tuning approach. In addition, due to the low signal-to-noise ratio (SNR) and high speckle noise of ultrasound imaging [10], [23], [24], imaging artifacts, sensor noise, acquisition errors, or even mislabeling in the dataset, there are multiple hard examples in the datasets, and even domain experts can be easily confused and difficult to draw conclusions. Here, we aim to design learning strategies to mimic the learning paths of domain experts. Specifically, our algorithm first picks out hard examples and then puts them back into the network for training, gradually improving the network's ability to distinguish hard examples. Combining a novel mask modeling-based SSL with hard example mining, we can improve the classification performance on ultrasound data.

The subsequent section provides an overview of relevant literature, discussing related works in detail. Our methodology is elucidated in Section III. Furthermore, experimental results are presented in Section IV. Section VI summarizes the methods employed in this paper and presents an exploratory discussion.

II. Related Work

A. Ultrasound Classification

Traditional ultrasound classification methods start by extracting discriminative features from radio frequency (RF) signals or ultrasound images. In the feature extraction phase, the researchers propose to use statistical modeling to extract discriminative features from ultrasound data and differentiate between different ultrasound data. For example, Grey Level Co-occurrence Matrix [25] measures how often different combinations of pixel intensities occur in an ultrasound image. Local Binary Patterns (LBP) [26] is a non-parametric method for extracting the local structural features of an image. Morphological features [27] focus on local characteristics of the certain ultrasound image region, such as the shape and margin. Statistical distribution modeling is also used as a feature in [28]. In order to exploit the complementary information of different ultrasound data, feature combination and selection are used in many attempts [29]. It is desirable to build a suitable classifier automatically. In the classification module, data-driven machine learning algorithms separate different ultrasound data by defining optimal decision boundaries in the feature space. Commonly-used classifiers include Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Adaboost, Random Forest, k-Nearest Neighbors (KNN), Neural Networks (NN) and Bayesian classifiers [30], [31], [32], [33].

Deep learning methods for ultrasound classification have gained increasing attention in recent years, where the features are automatically learned by neural networks rather than a-priori-defined features features [34], [35]. These methods are rapidly becoming the de-facto solution for ultrasound classification tasks, outperforming other traditional shallow architectures-based methods. Among them, CNN and their variants [36], [37], [38], [39] are widely-adopted. Long short-term memory (LSTM) [40] has also been applied which alleviates the vanishing gradient problem of standard deep neural networks. Transformer models [41], which are based on self-attention between image patches, have shown their great potential in ultrasound image classification. Due to the learnability of deep neural networks, the performance usually improves with the increase of labeled samples. Transfer Learning [42], weak-supervised learning [43] and unsupervised learning [41] have been proposed to tackle the problem of limited availability of the labeled data. These methods mainly focus on the ultrasound imaging data of human breast [41], liver [38], [42], brain [44]. To the best of our knowledge, how to learn from large-scale unlabeled ultrasound images is under-explored in the previous studies [45], [46].

B. Self-Supervised Learning

Self-supervised learning methods mainly uses an auxiliary task (pretext task) to predict its own supervised information from large-scale unsupervised data. By training the network with this constructed supervisory information, valuable representations for downstream tasks can be learned. For example, early self-supervised learning attempts are based on devising classification tasks that try to predict the properties of a transformation (e.g. rotations [47], colorization [48], [49], orderings or relative positions [50], [51]) applied on the input data.

Contrastive SSL methods, which use the *Instance Discrimination* as the auxiliary task, currently achieve state-of-the-art performance in SSL. These methods include CPC [52], MoCo [53], SimCLR [54], SvAV [55], CMPC [56] and BYOL [57]. The core idea of contrastive approaches is bringing the representation of different views of the same image closer ("positive pairs"), and spreading representations of views from different images ("negative pairs") apart [58]. In practice, contrastive SSL learning methods benefit from a large number of negative samples [52], [53], [54], [59]. These negative samples can be kept in a memory bank [59] or queue [53]. SimCLR [54] directly uses negative samples coexisting in the current batch, and it requires a large batch size for training purposes. Moreover, contrastive SSL relies on the construction of positive and negative sample pairs, and how to construct sample pairs for contaminated ultrasound image data has not been fully explored [46].

As revealed by recent studies [54], [60], [61], data augmentations are crucial in the representation learning. Data augmentations are also useful for the hard example mining. For example, MoCHi [58] employ a hard negative mixing for contrastive learning. More recently, masked modeling has evolved as a new kind of SSL method. These predict or reconstruct the partially masked parts of the data. For example, MAE [62] masks random patches of the input image and reconstruct the missing pixels. MaskFeat [63] masks out a portion of the input and then predicts the feature of the masked regions. These SSL approaches play an important role in the pre-training method and leading performance in downstream tasks. Most closely related

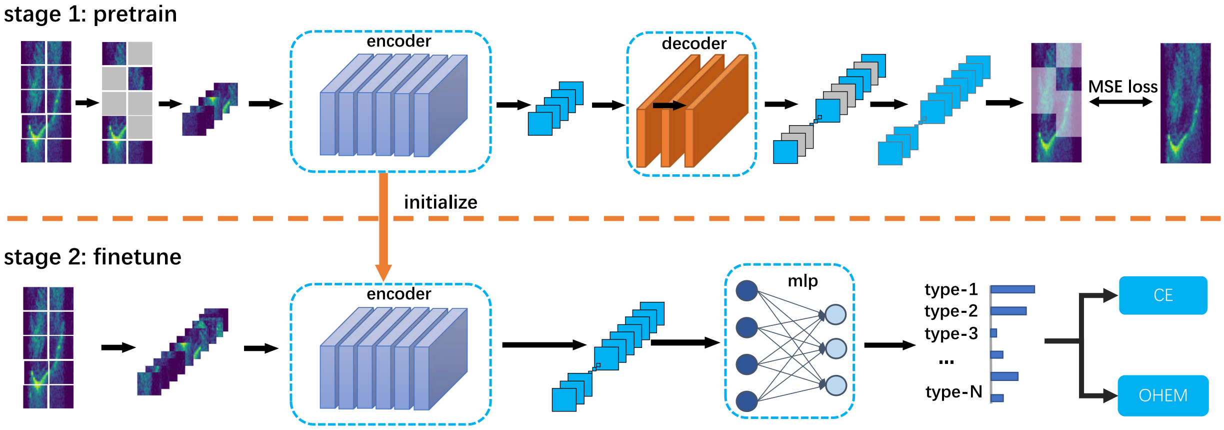

Fig. 1. Provided diagram illustrates the comprehensive flowchart of the proposed framework, encompassing two primary stages: pre-training utilizing unlabeled datasets and fine-tuning based on labeled datasets. Notably, both stages employ the same encoder but distinct decoders. During the pre-training stage, we employ a mask modeling-based self-supervised learning (SSL) method to train the neural network, obviating the necessity for expert labeling. This approach enables the network to learn an optimal representation of ultrasound images. The pre-trained encoder parameters are subsequently inherited for initialization in the fine-tuning stage. In the fine-tuning stage, the initialized encoder and classification head are fine-tuned employing a limited number of labeled datasets specific to the ultrasound classification task. To enhance the model's ability to differentiate challenging examples, we introduce a novel strategy for hard example mining.

to our work are the semi-supervised contrastive learning method (USCL) proposed in [45] and the hierarchical contrastive (HiCo) learning methods for ultrasound data. Despite the efforts that have been made in self-supervised learning, the applications of SSL for ultrasound data are under-explored in previous studies, which is the main goal of this paper.

III. Methodology

Fig. 1 shows the overall framework of our proposed method, which consists of two main stages: the pre-training stage of SSL using the Transformer architecture, and the fine-tuning stage using supervised learning with limited labeled datasets. The two stages adopt the same encoder, but different decoders. For the pre-training stage, the widely-used SSL approach is explored to train the neural network without human annotation. After the pre-training phase, the encoder has learned discriminative representation of ultrasound images, and its trained parameters will be inherited to the fine-tuning stage for the initialization of the network's parameters.

For the fine-tuning stage, we use the labeled datasets to fine-tune the initialized encoder and a classification head for the ultrasound classification tasks. Moreover, to enable the model to distinguish hard examples, we propose a novel hard example mining strategy. Compared to the method without pre-training, our method can improve classification performance, due to the superior representation abilities.

For SSL from ultrasound data, we aim to address the following challenges in this paper:

- 1) How to perform representation learning for ultrasound data based on unlabeled data?
- 2) For hard samples in ultrasound data, how to design learning strategies to improve the performance of classification models?
- 3) How to quantitatively and qualitatively evaluate the effectiveness of the ultrasound data representation?

In the following subsections, we will explain the components in more detail subsequently.

A. Mask Modeling

Currently, many natural language processing (NLP) tasks employ the mask modeling for SSL, where a random token of the sentence is masked and the pre-training goal of the model is to predict the masked part from the unmasked information, so that pre-training has the ability to model the contextual knowledge. Deep neural networks (DNN) like BERT fall into this category [64], and this approach has achieved promising performance on tasks for NLP. Mask modeling method possesses commendable representation and generalization capabilities, the intricacies of which will be expounded upon subsequently.

Compared with NLP tasks, there are certain differences in the analysis tasks of ultrasound images. First, nearby pixels are highly correlated in ultrasound data, so even if a pixel is masked, its value can be inferred relatively easily by analyzing its neighbors. Moreover, ultrasound data are continuous unlike tags in NLP which are discrete. However, ultrasound images are affected by speckle noise and many other imaging artifacts, how to overcome the influence is still under-explored in previous studies. Designing a suitable SSL paradigm for ultrasound data is the main goal of this paper.

In this paper, inspired by mask modeling in NLP, we explore a pre-training approach for ultrasound images based on mask modeling. The pre-training strategy for mask modeling aims to enable deep neural networks to reason about mask parts using neighborhood information in ultrasound images. Specifically, we tested the effect of different masking strategies on the performance of the pre-trained model. The mask strategy is mainly determined by three factors: mask shape, mask ratio and mask size. Specifically, we mainly tested three different mask shapes: random patch mask, vertical mask, and horizontal-mask. Fig. 2 shows an example of masking strategy for an ultrasound

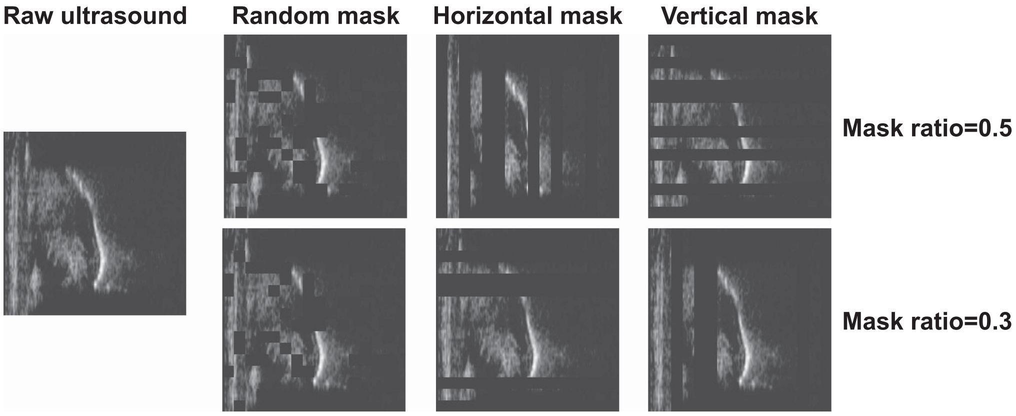

Fig. 2. Running example of masking strategies for ultrasound data, including random patch level, horizontal masking, and vertical masking. The first column gives the raw ultrasound data, and the next three columns give different masking strategies. It is worth noting that the first row has a mask ratio of 0.5 (50% of the original pixels are set to 0), while the second row has a mask ratio of 0.3.

image. It is noteworthy that the selection of hyper-parameters for the mask involves the mask scale and mask size, which play a crucial role in defining the level of difficulty during pre-training. Optimal prediction difficulty is vital for the model to acquire superior ultrasound data representations. Low difficulty may induce the model to underfit, whereas excessively high difficulty may impede convergence. Through our pre-training experiments, we discovered that the pre-trained model exhibits remarkable accuracy in recovering the masked portions, even at high masking ratios (e.g., masking 60% of ultrasound image pixels). Further verification of this observation can be found in the subsequent experimental section.

B. Network Architecture

Traditional approaches for ultrasound data interpretation tasks predominantly rely on domain knowledge, aiming to engineer task-specific representations (e.g., LBP). However, as previously mentioned, the manual feature design by domain experts necessitates extensive prior knowledge. Moreover, the performance of ultrasound classification models is significantly influenced by the quality of ultrasound data representation. Notably, feature extraction algorithms often exhibit limited generalization capabilities [65], [66]. On the other hand, affected by ultrasound imaging settings, these factors will further restrict the performance of the model, that is, the artificially designed features are not robust. CNNs are considered as one of the typical architectures for ultrasound image analysis [67], [68]. They employ a hierarchical data representation method, where high-level feature representation relies on low-level feature representation. This allows for the extraction of features with higher-level semantic information, progressing from shallow to deep abstraction features.

Given that self-supervised training necessitates the acquisition of high-quality, general-purpose representations from vast amounts of unlabeled data, Transformer-based networks are considered more adaptable to handling large-scale dataset. Consequently, in our SSL framework, we adopt ViT-large [69] as the encoder, leveraging its suitability for this purpose. For pixel-level prediction as the generator, we utilize a lightweight Transformer structure as the decoder. It is important to note that the decoder is exclusively employed during the pre-training stage and subsequently substituted with a multi-layer perceptron (MLP) header for classification in the fine-tuning stage. Specifically, the encoder consists of 24-layer Transformer blocks, and the decoder consists of 8-layer Transformer blocks. Each block includes a Multi-Head Attention layer, an MLP layer, and two Normalization layers, in which the Multi-Head Attention layer is a signature component playing an important role. The self-attention model establishes the relationship between different elements in a vector through triples (query, key, value), thus enhancing valuable information and weakening irrelevant information. The self-attention can be represented as the (1):

$$y_i = \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{1}$$

where x_i refers to the query, x_j refers to the key, and $g(x_j)$ is an embedding of x_j and refers to the value.

Single attention creates only one query and key dependency. We aim to learn about different dependencies and then combine these dependencies to capture various ranges of dependencies within the sequence, thus employing the multi-headed attention in our practical implementation. To fit the input size of the Transformer, we first convert the ultrasound image into sequence data. Suppose we have an ultrasound image $x \in R^{H \times W \times 1}$, we can divide it into N patches, and each patch with the size of $p \times p$. The patch can be expressed as $x_p \in R^{N \times (p^2)}$, and N can be calculated as $N = \frac{HW}{p^2}$, which is the length of the sequence. Then we flatten the patch into a one-dimensional vector and project it to a smaller vector through a linear layer, which is called as token. The tokens are fed into the encoder for subsequent processing.

C. Fine-Tuning With Limited Labeled Datasets

Once the pre-training of the model is performed using unlabeled ultrasound data, it can be fine-tuned on downstream tasks. As shown in Fig. 1, the fine-tuning stage aims to further optimize the model parameters with the labeled data. In this phase, we directly utilize the pre-trained encoder network and learned weights as the network backbone for ultrasound feature extraction, and then use a randomly initialized linear layer as the classification head to perform the ultrasound classification task. Thus, the fine-tuning problem can be regarded as a transfer learning problem, so the weight adjustment for pre-training is small, aiming to maintain the representation ability of the pre-trained model. We evaluate the effectiveness of this pre-trained algorithm by calculating the accuracy of ultrasound image classification, which will be further discussed in the experimental sections.

D. Hard Example Mining

Given the low signal-to-noise ratio (SNR) characteristic of ultrasound data, the presence of challenging examples in classification is inevitable. As the pre-trained model undergoes successive iterations, a substantial number of examples can be accurately classified by the model. Consequently, our primary objective is centered around enhancing the classification accuracy of hard examples. By enabling the model to allocate more attention to challenging instances, it holds the potential to further elevate the overall performance of the ultrasound classification task.

Standard cross-entropy is commonly used for classification tasks, as shown in the Equation 2:

$$L_{CE} = -\sum_{i=1}^{c} p_h(i) \log p_s(i) \tag{2}$$

where c is the number of categories, p_h refers to the hard label, p_s is the output of softmax function. Standard cross-entropy treats all samples equally, which may not be suitable when the categories are not balanced or the examples are not equally hard.

1) Focal Loss: To handle with the issue of standard cross-entropy, the focal loss is firstly proposed in the object detection field [70]. The object detection task can be regarded as a special dichotomous task, where candidate boxes with objects are regarded as a positive category, while the candidate boxes without objects are regarded as a negative category. In the object detection task, there are a large number of negative examples but only a few positive examples.

To solve the problem of unbalanced positive and negative examples, [70] proposed the focal loss. This function uses prediction confidence as a measure of the example's identification difficulty. As shown in (3), the neural network model can not only give the category judgment but also output the confidence degree of the deep models.

$$(y', conf) = f_\theta(x) \tag{3}$$

where θ present the parameters of the model, $conf$ presents the confidence, and y' is the category judgment of the model. Specially, both $conf$ and y' are calculated from p, the probability distribution of the model output, e.g., $conf = max(p)$ and $y' = argmax(p)$. Hence, the greater the confidence, the more definitive the model's categorization judgment, indicating that the example is relatively easier to identify. Building upon this understanding, the Focal loss employs confidence as a criterion for assessing example difficulty and adjusts the weight of examples within the loss function, as demonstrated in (4):

$$FL(p) = -(1-p)^\mu \log p \tag{4}$$

where p is the probability of prediction as a positive example, μ is a hyper-parameter, and a larger μ gives greater weight to hard examples.

Expanding focal loss to multi-classification, the formula is transformed into the form shown in Equation 5:

$$FL(p_{gt}, p_{pd}) = -\sum_{i=1}^{c} (1 - p_{gt}(i) * p_{pd}(i))^\mu \log p_{pd}(i) \tag{5}$$

where p_{gt} is the ground truth, p_{pd} is the prediction, c is the number of category.

2) GHM-C Loss: Due to the contamination of speckle noise, there also may be outliers in the dataset, and these outliers will still be wrongly judged when the model has converged. If the model pays attention to the excessively hard examples, the convergent model will deviate from the optimal state. In [71], a new gradient coordination mechanism (GHM) is proposed to hedge the discordance between examples, which can overcome the disadvantages of outliers. It utilizes the gradient norm to represent difficult levels of examples, which is defined as:

$$g = |p_h - p_{pred}| \tag{6}$$

which measures the difference between the prediction and the label, and takes values in the range $0 \sim 1$. On this basis, for representing the distribution of gradient norm, a so-called gradient density function is defined as formula 7:

$$GD(g) = \frac{1}{l_\epsilon(g)} \sum_{k=1}^{N} \delta_\epsilon(g_k, g) \tag{7}$$

where ϵ refers to the neighborhood of gradient norm g, g_k is the gradient norm of the k-th example, N is the total number of examples, and $\delta_\epsilon(g_k, g)$ indicates whether g_k is distributed in the neighborhood ϵ (defined as (8)), and $l_\epsilon(g)$ represents the interval length of the neighborhood (defined as (9)).

$$\delta_\epsilon(x, y) = \begin{Bmatrix} 1, & \text{if} \quad y - \frac{\epsilon}{2} \le x < y + \frac{\epsilon}{2} \\ 0, & \text{otherwise} \end{Bmatrix} \tag{8}$$

$$l_\epsilon(g) = \min\left(g + \frac{\epsilon}{2}, 1\right) - \max\left(g - \frac{\epsilon}{2}, 0\right) \tag{9}$$

With the iterations of deep model, the gradient norm (g) gradually approaches to the value 0. When the model gradually converges, a large number of gradient norms are clustered near 0, a small number of gradient norms are near 1, and only a few gradient norms are between 0 and 1. The gradient norm near 0 represents easy examples, the gradient norm near 1 refers to excessively hard outliers, and the others represent moderately hard examples.

As mentioned above, we can utilize the reciprocal of gradient density as a weight factor of the corresponding example in the loss function, so we can increase the weight of the general hard examples in the loss function, and suppress the weight of the easy examples and outliers. Thus the weight factor and loss function can be formulated as follows:

$$\beta_j = \frac{N}{GD(g(j))} \tag{10}$$

$$\begin{aligned} L_{GHM-C} &= \frac{1}{N}\sum_{j=1}^{N} \beta_j L_{CE}(p_h(j), p_{pred}(j)) \\ &= \sum_{j=1}^{N} \frac{L_{CE}(p_h(j), p_{pred}(j))}{GD(g_j)} \end{aligned} \tag{11}$$

where N is the total number of examples, L_{CE} refers to Cross Entropy loss, $p_h(j)$ and $p_{pred}(j)$ refer to the label and prediction of the j_{th} example. We can see that the reciprocal of gradient density is treated as a weight factor for L_{CE}.

Finally, the model is trained to minimize a joint loss function, including the cross-entropy loss and GHM-C loss:

$$L = L_{CE} + \alpha L_{GHM\_C} \tag{12}$$

where α is a hyper-parameter used to adjust the weight of cross entropy and GHM-C.

Based on the pre-training using mask-modeling strategy, we can utilize the unlabeled ultrasound data. Then we rely on the labeled data for fine-tuning purposes, combined with new hard sample mining methods. Combining the above improvements, our framework will substantially improve the performance of ultrasound image classification, which will be further verified in the following experiments.

IV. Experiments

In this section, we first conduct experiments on different ultrasound datasets using the proposed method, and then present quantitative comparison and analysis. Furthermore, to verify the generalizability of representation learning for ultrasound images under different settings, we validate the transferability of representation learning between different datasets.

A. Datasets

We validate the effectiveness of the algorithm based on two different ultrasound datasets: ultrasound tongue images and breast tumor ultrasound images.

1) Ultrasound Tongue Images: As mentioned earlier, ultrasound images can characterize tongue movement with real-time visualization of natural vocalizations [23]. Quantitative analysis of ultrasound tongue data (UTI) has a wide range of practical applications, such as silent speech recognition [72] and speech disorder classification [73], [74]. The performance of ultrasound tongue gesture classification based on deep learning has surpassed traditional feature extraction methods, while however relying on the labeled data.

In this paper, we evaluate the classification performance of the proposed framework on UTI data. Specifically, we aim to classify the data into four categories based on different articulation positions: (1) bilabial and labial-dental (e.g. /p/, /b/, /v/...); (2) dental, alveolar and postalveolar sounds (e.g. /th/, /d/, /t/, /z/...); (3) velar sounds (e.g. /k/, /g/...); (4) the alveoli approximate /r/. The used dataset comes from the UXTD of [75] in the publicly available UltraSuite dataset. The UXTD dataset contains 58 speakers (31 female and 27 male), aged 5–12 years. Based on phoneme boundaries using force alignment, we extracted the middle part of each available utterance, resulting in a dataset of approximately 10,700 examples. Then, we split these data into disjoint training, validation, and test sets. For the datasets used for self-supervised learning, we utilize all frames of each available utterance as unlabeled examples and obtain approximately 460,819 samples. Similar to [11], we evaluate our method in four different settings: subject-dependent, multi-subject, subject-independent, and subject-adaptive. In subject-independent and subject-adapted, for N patients, we create N datasets. The training subset of each dataset contains sample data from (N-1) patients, and the testing data of each dataset contains sample data from another patient. The patients in the training and test subsets are different. In subject-dependent, each patient constitutes a separate dataset, and in multi-subject, all patient data constitute a single dataset. Both these two settings have a training/validation/test set ratio of 6:2:2. For example, assuming that we sampled 100 ultrasound tongue images from three patients. The training/validation/test dataset division with subject-dependent, multi-subject, subject-independent, and subject-adaptive settings are shown in Table I It is noted that in subject-dependent, subject-independent, and subject-adaptive settings, we create 3 different datasets. Meanwhile, in multi-subject, we only create 1 dataset.

2) Breast Tumor Ultrasound Images: Ultrasound imaging is one of the effective means for early screening of breast cancer [76], [77]. In 2018, the American Cancer Society evaluated the incidence and mortality of 36 cancers in 185 countries around the world and performed statistical analysis showing that breast cancer is one of the three major malignant tumors in women, of which breast cancer accounts for 8.6 million female cancers 24.2% of new cases accounted for 15% of the 4.2 million female cancer deaths, ranking first in the incidence and mortality of female cancer [78].

In our experiments, we employ the BUSI-BUI joint dataset [79], [80] for the classification task. The BUSI datasets collects breast ultrasound images including women aged 25 to 75 years [80]. The dataset consists of 780 images, including normal, benign and malignant, with an average image size of 500 × 500 pixels. BUI dataset consists of 250 breast cancer ultrasound images, with 100 images classified as benign and 150 images classified as malignant. For dataset creation, we randomly mixed the two datasets into one dataset and use 5-fold cross validation method to test models performance, which means 824 images for the training set, 206 images for the validation set in one fold.

B. Experimental Setting

We train the SSL model using 8 GPUs (NVIDIA A40). For the pre-training of SSL models, we employ the UTI for the training purpose, with the input size of 256 × 256. We use Adamw optimizer, where $\beta_1 = 0.9$, $\beta_2 = 0.95$. We set the batch size to 512, the weight_decay to 0.05, and the learning rate to $6 \times e^{-4}$ [62]. We take a 16 × 16 pixel block as a unit, and the

TABLE I
Dataset Division With Different Settings

Dataset	subject-dependent			multi-subject	subject-independent			subject-adaptive		

Train	A(80)	B(80)	C(80)	A+B+C=180	A(100)
					B(100)
						A(100)
						C(100)
							B(100)
							C(100)
								A(100),
								B(100)
									A(100),
									C(100)
										B(100),
										C(100)

Validation	/			A+B+C=60	/	/	/	/	/	/
Test	A(20)	B(20)	C(20)	A+B+C=60	C(100)	B(100)	A(100)	C(20)	B(20)	A(20)
Fine-tune	/			/	/	/	/	C(80)	B(80)	A(80)

A, B, and C respectively represent the different patients. The number in parentheses indicates the number of data samples sampled for this patient. For each patient, we sample 100 ultrasound tongue images.

mask pixel ratio is 0.75 for masking (the selection of a mask pixel ratio of 0.75 is anchored in empirical evidence derived from the study conducted by [62]). The model undergoes a comprehensive training regimen spanning 100 epochs, with the inclusion of a warmup strategy for the initial 40 epochs. The training process lasted approximately 10 hours. During the pre-training phase, we exclusively employ a limited array of online data augmentation techniques, which is performed with a certain probability during the model training, specifically random crop and horizontal Flip operations, to cultivate resilient representations. For the fine-tuning phase, the input is consistent with the settings of pre-training, using AdamW optimizer, where $\beta_1 = 0.9$, $\beta_2 = 0.999$. We set the batch size to 128, the learning rate to $5e^{-4}$, the weight decay to 0.05. The model is trained with the early stopping strategy until no improvement on the validation set, and the maximum epoch is set as 50. Online data augmentation is also used, including the random crop and horizontal flip. It is worth noting that many techniques mentioned above are employed to effectively adjust the fine-tuning process. Firstly, conservative learning rates allow for gradual adjustments to pre-trained weights while preventing significant biases. Additionally, weight decay techniques can further stabilize the fine-tuning process and prevent overfitting.

C. Competitive Baselines

For quantitative comparisons, we tested the following competitive approaches:

1) Raw UTI+DNN;
2) PCA+DNN;
3) DCT+DNN;
4) Raw UTI+CNN;
5) SimCLR;
6) SimSiam;
7) USCL [45];
8) HiCo [46].

Specifically, the raw UTI input denotes the mean-variance normalized raw frame. PCA is a linear transformation that transforms high-dimensional data into low-dimensional data, simplifies data and retains the most important features of the data. We retain the first 1000-dimensional features in our experiments. In our experiments, the hyper-parameter (1000) is empirically selected to trade off between dimensionality reduction and information preservation. DCT denotes Discrete Cosine Transform, which is mainly used to compress data or images and can convert spatial domain signals to the frequency domain. In our experiments, we retain the 40×40 sub-matrix in the upper left corner of the transformed DCT coefficient matrix, which is the most important 1,600 feature dimension. When applying the DCT to an image, it transforms the image from the spatial domain to the frequency domain, representing the image in terms of its frequency components. The DCT coefficients represent the magnitudes of these frequency components. The energy compaction property of the DCT means that a majority of the signal's energy tends to be concentrated in a small number of DCT coefficients. In the case of images, the energy tends to be concentrated in the lower-frequency components, which correspond to large-scale variations in the image. In a typical DCT coefficient matrix, the top-left corner contains the low-frequency coefficients, while the bottom-right corner contains the high-frequency coefficients. By convention, the top-left corner of the DCT coefficient matrix corresponds to the lower-frequency components that capture the most important and visually significant information. In your scenario, selecting the 40×40 sub-matrix in the upper left corner of the transformed DCT coefficient matrix implies retaining the coefficients associated with the most significant frequency components. These components capture the bulk of the energy and represent the essential visual features of the image.

In our practical implementation, the DNN architecture consists of three hidden layers with inputs being either raw, PCA or DCT. Each hidden layer is composed of 512 rectified linear units (ReLU) and softmax activation function. For the CNN architectures, we use two convolutional layers and max-pooling layers followed by two fully connected layers. The convolutional layer consists of 16 filters, with the kernel sizes of 8×8 and 4×4.

In our experiments, we also test contrasting self-supervised learning algorithms and we chose SimCLR and SimSiam as the representative algorithms, due to its versatile representation abilities. Specifically, SimCLR introduces a learnable nonlinear transformation between representation and contrastive losses to improve the model to learn high-quality representations. This transformation employs a straightforward single-layer architecture, consisting of a Multi-Layer Perceptron (MLP) with Rectified Linear Unit (ReLU) activation. The advantage of this approach lies in its ability to circumvent the loss of critical features during the calculation of the similarity loss function. This, in turn, contributes to an enhancement in the quality of the representation within the preceding layer. The SimSiam algorithm can obtain a discriminative and meaningful representation by directly maximizing the similarity of two views of an input based on a simple Siamese network without using

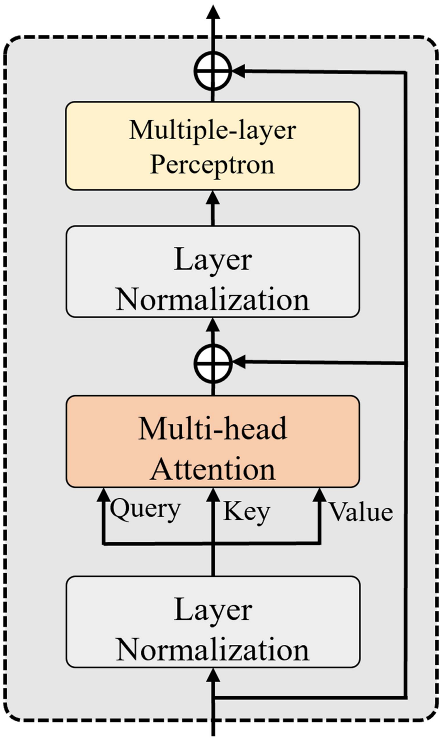

Fig. 3. Internal structure of the Transformer block.

negative samples or momentum encoders. The two branches of the Siamese network use a shared backbone network (such as ResNet), one of which is connected with a simple MLP head, and the other branch has no MLP head. We also compare our methods with two state-of-the-art ultrasound image classification methods: USCL and HiCo. Additionally, the table includes the performance results of the Vision Transformer model trained from scratch [82].

D. The Visualization for Pre-Training

In the context of pre-training utilizing unlabeled ultrasound data, the acquisition of a representation endowed with robust generalization capabilities assumes paramount significance. In our experiments, we first validate the effectiveness of the pre-trained model and then provide a quantitative evaluation of the proposed framework. Specifically, we randomly select several UTI data that are not used for training and apply random masking to the input. Subsequently, using the pre-trained model, we infer the masked regions of the UTI data. The prediction visualizations are presented in Fig. 4. The first column displays the raw UTI data. The second and third columns illustrate the patch-level random masking and the corresponding predicted UTI data, respectively. The fourth and fifth columns show the results obtained from the horizontal mask, while the sixth and seventh columns correspond to the results obtained from the vertical mask modeling.

From a visual analysis of the image, the pre-trained model demonstrated satisfactory reconstruction performance when predicting occluded areas, irrespective of the particular occlusion strategy used during the pre-training phase of the ultrasound image representation. Notably, even with a considerably high masking rate applied to ultrasound images, the pre-trained model consistently infers the contents of the masked portions with precision. Experimental findings highlight a distinctive characteristic of our pre-trained model, whereby it exhibits inherent capabilities for neighborhood modeling and inference within ultrasound images, distinguishing it from many existing methods. To a certain extent, the visualization outcomes suggest that our pre-trained model effectively learns and captures features associated with masked regions and contextual information.

E. Quantitative Comparison

Table II provides a quantitative comparison between different methods, including the proposed methods and competitive baselines, in terms of their performance on two datasets: UTI dataset and BUSI-BUI dataset.

As can be seen from the table, the pre-training methods combined with fine-tuning (including SimCLR [53], SimSiam [81], HiCo [46], and our methods with different masking strategies) provide better performance than supervised learning. This is mainly because that the labeled data for real ultrasound images is usually limited, so that supervised learning usually could not achieve a satisfactory performance. In our approach, we learn a representation with good generalization by using a large amount of unlabeled data through the SSL approach. It is worth noting that our method outperforms other competitive methods. Unlabeled data are not employed in the supervised methods (row 2 to row 5), while self-supervised approaches learns patterns and structures in unlabeled data. We present the performance based on three different masking-reconstruction strategies, among which the method based on vertical masking has the best performance. For the UTI data, our framework leveraging the random patch masking strategy can provide higher accuracy than other competitive methods in the "subject-dependent" scenario, which is 2.19% higher than HiCo (which is the previous state-of-the-art method). In the "multiple-subject" setting, all of our three methods outperform other methods, with at least 6.69% higher accuracy. In the "subject-independent" scenario, the vertical masking slightly performs better than the horizontal masking by 0.24%. Our patch masking method outperforms previous HiCo [46] by 8.35%. In the "subject-adaptive" scenario, the method based on contrastive SSL also achieved satisfactory performance, and the accuracy of SimCLR, SimSiam, USCL and HiCo reaching 87.46%, 85.50%, 82.59%, and 84.69%, respectively. Compared to the second-best method (HiCo), our method improves the average accuracy by 5.2%. For the BUSI-BUI dataset, our framework surpasses SimCLR, SimSiam, HiCo, and USCL by significant margins of 13.04%, 1.08%, 1.64%, and 2.14%, respectively.

The proposed framework based on mask modeling achieves higher classification accuracy on both UTI and BUSI-BUI datasets compared to recently proposed contrastive SSL methods. We attribute this performance boost primarily to

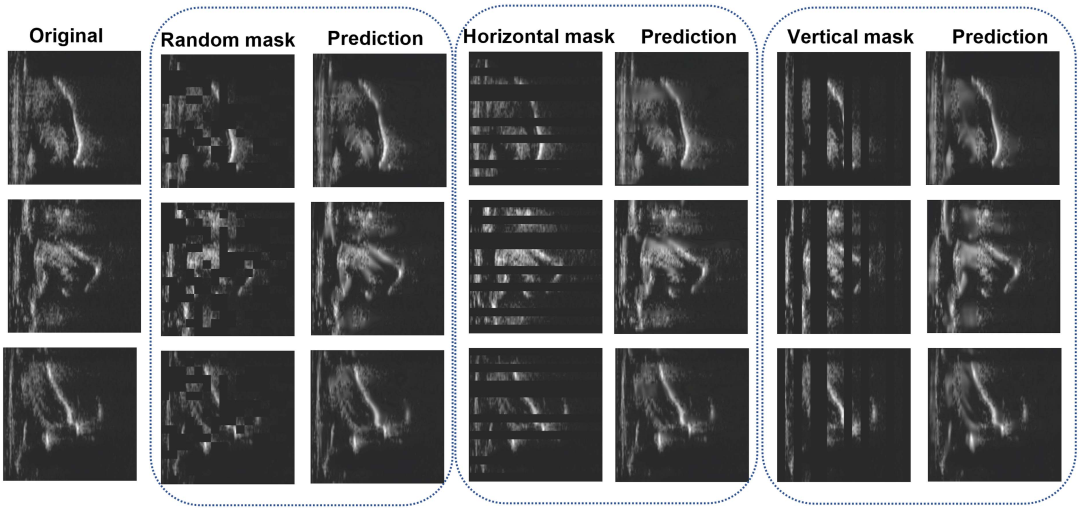

Fig. 4. Prediction visualizations using the pre-trained model. The first column gives the original ultrasound image. The second and third columns give the patch-level random masking and the predicted ultrasound image respectively; the fourth and fifth columns give the results based on the horizontal mask, and the sixth and seventh columns correspond to the results based on the vertical mask modeling.

TABLE II
Quantitative Comparison Between Our Proposed Methods and Competitive Baselines

	UTI dataset					BUSI-BUI dataset
Method	Subject-dependent	Multiple-subject	Subject-independent	Subject-adapted	Mean accuracy	Mean accuracy
Raw UTI + DNN	62.15%	69.62%	54.15%	69.26%	63.80%	67.47%
PCA + DNN	57.78%	66.30%	55.14%	68.37%	61.90%	67.25%
DCT + DNN	68.38%	71.91%	55.36%	67.76%	65.85%	69.41%
Raw UTI + CNN	66.56%	74.70%	59.42%	72.67%	68.34%	72.35%
SimCLR [53]	45.29%	77.02%	70.14%	87.46%	69.98%	74.60%
ViT (from scratch) [69]	40.71%	60.71%	48.79%	45.45%	48.90%	71.3%
SimSiam [81]	82.79%	78.31%	71.41%	85.50%	79.50%	86.56%
USCL [45]	83.45%	76.26%	74.57%	82.59%	79.22%	85.50%
HiCo [46]	84.41%	77.38%	75.68%	84.69%	80.54%	86.00%
Our method (Patch masking)	**86.60%**	85.46%	84.03%	89.13%	86.31%	**87.64%**
Our method (Horizontal-masking)	83.11%	85.00%	84.94%	**90.00%**	85.74%	85.10%
Our method (Vertical-masking)	85.00%	**85.85%**	**85.18%**	89.57%	**86.40%**	85.51%

TABLE III
Results of F1-Score, Recall, and Precision on BUSI-BUI Dataset

Method	F1-Score	Recall	Precision
SimCLR	52.60%	54.93%	52.99%
SimSiam	79.01%	78.51%	80.24%
Our method (Vertical-masking)	82.86%	81.19%	85.39%
Our method (Horizontal-masking)	83.13%	82.08%	84.74%
Our method (Patch-masking)	82.48%	81.04%	84.61%

mask-based reconstruction methods that enable the model to reason about neighborhood information, which is important for ultrasound image interpretation. As our approach incorporates a hard example mining strategy, we will conduct a series of ablation experiments to analyze the impact of each strategy on performance in the subsequent sections.

Apart from accuracy, we also validated the model's performance on other metrics, as shown in Table III. From the perspectives of F1 score, recall, and precision, our method performed well across these three metrics, demonstrating its effectiveness. Firstly, for vertical and horizontal masking, we observed their performances to be very close, both outperforming the SimSiam method by approximately 3 percentage points. This indicates that whether masking in the vertical or horizontal direction, our method can better capture key features in images, thus improving classification accuracy. This also reflects the effectiveness of our method in capturing different directional features in images. Additionally, for patch masking, we noted its performance to be slightly lower than vertical and horizontal masking, but still significantly better than the SimCLR method. This suggests that although patch masking may lose some local information, it still provides sufficient contextual information to effectively learn image features. Overall, our experimental results demonstrate that our method performs well under different types of masking, confirming its robustness and versatility. These results provide strong support for the feasibility of our method in practical applications.

TABLE IV
ABLATION STUDIES FOR HARD EXAMPLE MINING METHODS, USING THE PROPOSED FRAMEWORK

Cross Entropy Loss	Focal loss	GHM-C Loss	Accuracy
✓			82.62%
	✓		83.90%
		✓	84.94%
	✓	✓	**85.46%**

It is worth noting that in this ablation experiment, we evaluate the classification performance based on the multiple-subject scene and use the random patch masking measurement.

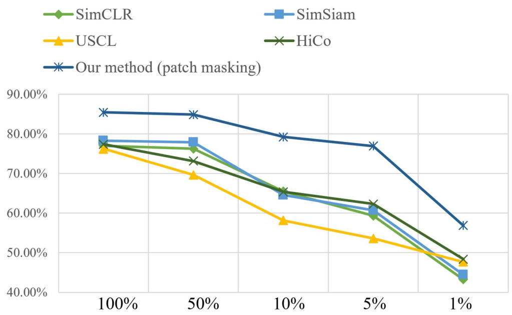

Fig. 5. Classification performance under few-shot settings. The horizontal axis denotes the percentage of the labeled datasets which are utilized for the fine-tuning purpose. The horizontal axis denotes the classification accuracy.

In the multi-subject scenario of the UTI classification task, we compare various hard example mining strategies, and the corresponding results are presented in Table IV. From the table, it is evident that both the focal loss (83.90%) and GHM-C (84.94%) outperform the standard cross-entropy loss (82.62%) in terms of performance. Notably, the combined utilization of focal loss and GHM-C achieves the highest accuracy of 85.46%, further validating the effectiveness of our proposed hard example mining method. The integration of mining and analysis of challenging samples resulted in a 2.84% increase in accuracy compared to the standard cross-entropy loss. This improvement bears significant practical significance for the interpretation of ultrasound data.

F. Few-Shot Settings

For many ultrasound image classification tasks, we often need to fine-tune the deep models under few-shot settings. In the few-shot scenarios, the deep models are confronted with more challenges. To further verify the superior performance of our pre-trained model, we fine-tune the model by randomly sampling the training set of ultrasound classification and reduced the scale of the labeled data set for fine-tuning to 50% and 10%, 5% and 1% of the original data set, respectively. We then use the sampled dataset to train and evaluate performance with different algorithms. The experimental results of quantitative analysis are shown in Fig. 5.

It can be seen that the classification results of our proposed framework are less adversely affected by few-shot scenarios, while other algorithms are more sensitive to the reduction of samples. Analyzing the underlying reasons, we posit that our pre-trained model can acquire more robust representations and deliver superior performance on designated tasks, even when provided with limited training samples. On the other hand, the unlabeled ultrasound data can be fully utilized based on pre-training and fine-tuning paradigms.

V. DISCUSSION

In this work, we propose a masked modeling-based pretraining method specifically designed for ultrasound images. We examine the effectiveness of the proposed model by leveraging two different types of ultrasound datasets: ultrasound tongue and breast tumor ultrasound datasets. Additionally, we investigate three different masking strategies: random masking, vertical masking, and horizontal masking. The proposed framework consists of two stages: pretraining on unlabeled datasets and fine-tuning on labeled datasets. In the pretraining stage, we employ a masked modeling-based SSL approach to train the neural network. During fine-tuning stage, we fine-tune the encoder and classification head using a limited number of labeled datasets specific to ultrasound classification tasks, while introducing a novel hard example mining strategy. The model achieves an accuracy and F1 Score of 87.64% and 82.48%, respectively, on the BUSI-BUI breast cancer classification dataset.

The absence of annotated data presents a formidable challenge for medical practitioners, particularly in the field of ultrasound imaging, where annotating detailed information for low signal-to-noise ratio ultrasound images proves to be arduous. To tackle this hurdle, we employ self-supervised learning to extract features from ultrasound images without relying on annotations. Moreover, given the inherent low signal-to-noise ratio in ultrasound images, the presence of challenging samples in classification tasks is inevitable. Hence, we introduce a novel hard example mining strategy to mitigate this challenge. Leveraging self-supervised learning for feature extraction from ultrasound images leads to a significant enhancement in accuracy scores. Our study demonstrates that ViT, initialized with parameters acquired through self-supervised learning, outperforms ViT pretrained with ImageNet parameters. This underscores the notion that manual augmentation of annotated data volume is dispensable when employing self-supervised learning. Our framework is well-suited for ultrasound image learning and possesses sufficient versatility to address the issue of inadequate annotation in various other medical image types.

VI. CONCLUSION

Pre-training using large-scale unlabeled data has become a dominant paradigm in the field of artificial intelligence. However, previous studies have not fully explored the potential of utilizing large-scale unlabeled ultrasound data. In this paper, we investigate the application of the masking-reconstructing strategy-based semi-supervised learning (SSL) paradigm for ultrasound image classification tasks. Our aim in the pre-training stage is to develop three different mask modeling strategies that enable the deep model to learn representations with strong generalization ability. Ultrasound images pose unique challenges due to their low signal-to-noise ratio (SNR) and high speckle noise, resulting in the presence of numerous hard examples in the dataset. These hard examples can be perplexing even for domain experts, making it difficult to draw conclusive classifications. To address this issue, we propose a hard example mining strategy that mimics the learning process of domain experts. By combining the mask modeling-based SSL approach with hard example mining, our method significantly enhances the classification performance of ultrasound data. Furthermore, we assess the transferability of our learned representations on different datasets and evaluate their generalization ability in few-shot learning scenarios. It is important to note that our proposed method can also be extended to ultrasound radiofrequency (RF) signals. In future work, we plan to explore a multi-representation pre-training approach that combines ultrasound RF signals and ultrasound images, thus further enhancing the robustness and effectiveness of our method.

Conflict of Interest

No potential confilct of interest was reported by all authors.

Author Contribution

Kele Xu: Writing, Project administration. Kang You: Writing, Validation. Boqing Zhu: Methodology. Ming Feng: Validation, Visualization. Dawei Feng: Investigation. Cheng Yang: Resources.

References

[1] A. Feher and A. J. Sinusas, “Quantitative assessment of coronary microvascular function: Dynamic single-photon emission computed tomography, positron emission tomography, ultrasound, computed tomography, and magnetic resonance imaging,” *Circulation: Cardiovasc. Imag.*, vol. 10, no. 8, 2017, Art. no. e006427.

[2] B. C. Marincola et al., “High-intensity focused ultrasound in breast pathology: Non-invasive treatment of benign and malignant lesions,” *Expert Rev. Med. Devices*, vol. 12, no. 2, pp. 191–199, 2015.

[3] H. Li et al., “Cr-unet: A composite network for ovary and follicle segmentation in ultrasound images,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 974–983, Apr. 2020.

[4] M. S. Firouz, A. Farahmandi, and S. Hosseinpour, “Recent advances in ultrasound application as a novel technique in analysis, processing and quality control of fruits, juices and dairy products industries: A review,” *Ultrason. Sonochemistry*, vol. 57, pp. 73–88, 2019.

[5] M. Mischi, M. A. L. Bell, R. J. Van Sloun, and Y. C. Eldar, “Deep learning in medical ultrasound–from image formation to image analysis,” *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2477–2480, Dec. 2020.

[6] G. R. Harris et al., “Hydrophone measurements for biomedical ultrasound applications: A review,” *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 70, no. 2, pp. 85–100, Feb. 2023.

[7] J. Mamou and M. L. Oelze, *Quantitative Ultrasound in Soft Tissues*. New York, NY, USA: Springer, 2013.

[8] M. Alsharqi, W. Woodward, J. Mumith, D. Markham, R. Upton, and P. Leeson, “Artificial intelligence and echocardiography,” *Echo Res. Pract.*, vol. 5, no. 4, pp. R115–R125, 2018.

[9] R. A. Dar et al., “Breast cancer detection using deep learning: Datasets, methods, and challenges ahead,” *Comput. Biol. Med.*, vol. 149, 2022, Art. no. 106073.

[10] M. Li, C. Kambhamettu, and M. Stone, “Automatic contour tracking in ultrasound images,” *Clin. Linguistics Phonetics*, vol. 19, no. 6/7, pp. 545–554, 2005.

[11] M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals, “Speaker-independent classification of phonetic segments from raw ultrasound in child speech,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 1328–1332.

[12] J. Cleland, J. M. Scobbie, Z. Roxburgh, C. Heyde, and A. Wrench, “Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback,” *J. Speech, Lang., Hear. Res.*, vol. 62, no. 2, pp. 229–246, 2019.

[13] K. Xu, P. Roussel, T. G. Csapó, and B. Denby, “Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images,” *J. Acoustical Soc. Amer.*, vol. 141, no. 6, pp. EL531–EL537, 2017.

[14] T. Hueber et al., “Eigentongue feature extraction for an ultrasound-based silent speech interface,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. I-1245–I-1248.

[15] J. Cai, B. Denby, P. Roussel-Ragot, G. Dreyfus, and L. Crevier-Buchman, “Recognition and real time performances of a lightweight ultrasound based silent speech interface employing a language model,” in *Proc. InterSpeech*, 2011, pp. 1005–1008.

[16] Y. Ji, L. Liu, H. Wang, Z. Liu, Z. Niu, and B. Denby, “Updating the silent speech challenge benchmark with deep learning,” *Speech Commun.*, vol. 98, pp. 42–50, 2018.

[17] M. H. Mozaffari, M. Ratul, A. Rab, and W.-S. Lee, “IrisNet: Deep learning for automatic and real-time tongue contour tracking in ultrasound video data using peripheral vision,” 2019, *arXiv:1911.03972*.

[18] K. Xu, T. G. Csapó, and M. Feng, “Deep learning-based age estimation using b-mode ultrasound tongue imaging,” *J. Acoustical Soc. Amer.*, vol. 150, no. 4, pp. A190–A190, 2021.

[19] M. Feng, Y. Wang, K. Xu, H. Wang, and B. Ding, “Improving ultrasound tongue contour extraction using U-Net and shape consistency-based regularizer,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6443–6447.

[20] Y. Xiong, K. Xu, M. Jiang, L. Cheng, Y. Dou, and J. Wang, “Improving the classification of phonetic segments from raw ultrasound using self-supervised learning and hard example mining,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8262–8266.

[21] J. Zhu, W. Styler, and I. C. Calloway, “Automatic tongue contour extraction in ultrasound images with convolutional neural networks,” *J. Acoustical Soc. Amer.*, vol. 143, no. 3, pp. 1966–1966, 2018.

[22] K. You et al., “Raw ultrasound-based phonetic segments classification via mask modeling,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[23] M. Stone, “A guide to analysing tongue motion from ultrasound images,” *Clin. Linguistics Phonetics*, vol. 19, no. 6/7, pp. 455–501, 2005.

[24] J. Berry and I. Fasel, “Dynamics of tongue gestures extracted automatically from ultrasound,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 557–560.

[25] U. R. Acharya et al., “Data mining framework for fatty liver disease classification in ultrasound: A hybrid feature extraction paradigm,” *Med. Phys.*, vol. 39, no. 7Part1, pp. 4255–4264, 2012.

[26] D. V. Pazinato et al., “Pixel-level tissue classification for ultrasound images,” *IEEE J. Biomed. Health Inform.*, vol. 20, no. 1, pp. 256–267, Jan. 2016.

[27] T. Tan, B. Platel, H. Huisman, C. I. Sánchez, R. Mus, and N. Karssemeijer, “Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation,” *IEEE Trans. Med. Imag.*, vol. 31, no. 5, pp. 1034–1042, May 2012.

[28] F. Destrempes and G. Cloutier, “A critical review and uniformized representation of statistical distributions modeling the ultrasound echo envelope,” *Ultrasound Med. Biol.*, vol. 36, no. 7, pp. 1037–1051, 2010.

[29] J. A. Noble, “Ultrasound image segmentation and tissue characterization,” *Proc. Inst. Mech. Engineers, Part H: J. Eng. Med.*, vol. 224, no. 2, pp. 307–316, 2010.

[30] M.-H. Horng, “Multi-class support vector machine for classification of the ultrasonic images of supraspinatus,” *Expert Syst. with Appl.*, vol. 36, no. 4, pp. 8124–8133, 2009.

[31] W.-L. Lee and K.-S. Hsieh, “A robust algorithm for the fractal dimension of images and its applications to the classification of natural images and ultrasonic liver images,” *Signal Process.*, vol. 90, no. 6, pp. 1894–1904, 2010.

[32] F. Ciompi, “Multi-class learning for vessel characterisation in intravascular ultrasound,” *ELCVIA: Electron. Lett. Comput. Vis. Image Anal.*, vol. 13, no. 2, pp. 47–48, 2014.

[33] M. Moradi, P. Mousavi, and P. Abolmaesumi, “Tissue characterization using fractal dimension of high frequency ultrasound RF time series,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2007, pp. 900–908.

[34] G.-Q. Zhou et al., "A single-shot region-adaptive network for myotendinous junction segmentation in muscular ultrasound images," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2531–2542, Dec. 2020.

[35] A. Sahoo, H. He, D. Darrow, C. C. Chen, and E. S. Ebbini, "Image-guided measurement of radiation force induced by focused ultrasound beams," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 70, no. 2, pp. 138–146, Feb. 2023.

[36] J. Jang, Y. Park, B. Kim, S. M. Lee, J.-Y. Kwon, and J. K. Seo, "Automatic estimation of fetal abdominal circumference from ultrasound images," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1512–1520, Sep. 2018.

[37] M. H. Yap et al., "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, Jul. 2018.

[38] D. Mishra, S. Chaudhury, M. Sarkar, S. Manohar, and A. S. Soin, "Segmentation of vascular regions in ultrasound images: A deep learning approach," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2018, pp. 1–5.

[39] M. Amiri, R. Brooks, and H. Rivaz, "Fine-tuning U-Net for ultrasound image segmentation: Different layers, different outcomes," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2510–2518, Dec. 2020.

[40] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, pp. 115–143, 2002.

[41] B. Gheflati and H. Rivaz, "Vision transformers for classification of breast ultrasound images," in *Proc. IEEE 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2022, pp. 480–483.

[42] D. Meng, L. Zhang, G. Cao, W. Cao, G. Zhang, and B. Hu, "Liver fibrosis classification based on transfer learning and FCNet for ultrasound images," *IEEE Access*, vol. 5, pp. 5804–5810, 2017.

[43] R. J. Van Sloun and L. Demi, "Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 957–964, Apr. 2020.

[44] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 919–923.

[45] Y. Chen et al., "USCL: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 627–637.

[46] C. Zhang, Y. Chen, L. Liu, Q. Liu, and X. Zhou, "HICO: Hierarchical contrastive learning for ultrasound video model pretraining," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 229–246.

[47] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018.

[48] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[49] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 577–593.

[50] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[51] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "VideoMoCo: Contrastive video representation learning with temporally adversarial examples," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit. Conf.*, 2021, pp. 11205–11214.

[52] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[53] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit. Conf.*, 2020, pp. 9729–9738.

[54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[55] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in*Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

[56] B. Zhu et al., "Unsupervised voice-face representation learning by cross-modal prototype contrast," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 3787–3794.

[57] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[58] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21798–21809.

[59] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[60] A. Sinha, K. Ayush, J. Song, B. Uzkent, H. Jin, and S. Ermon, "Negative data augmentation," in *Proc. Int. Conf. Learn. Representations*, 2021.

[61] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1920–1929.

[62] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[63] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14668–14678.

[64] Z. A. Yilmaz, S. Wang, W. Yang, H. Zhang, and J. Lin, "Applying bert to document retrieval with birch," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.: Syst. Demonstrations*, 2019, pp. 19–24.

[65] Q. Meng et al., "Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 722–734, Feb. 2021.

[66] X. Xu, T. Sanford, B. Turkbey, S. Xu, B. J. Wood, and P. Yan, "Shadow-consistent semi-supervised learning for prostate ultrasound segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1331–1345, Jun. 2022.

[67] J. Ma, F. Wu, T. Jiang, J. Zhu, and D. Kong, "Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images," *Med. Phys.*, vol. 44, no. 5, pp. 1678–1691, 2017.

[68] W. Gómez-Flores, W. C. de, and A. Pereira, "A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumors in ultrasound," *Comput. Biol. Med.*, vol. 126, 2020, Art. no. 104036.

[69] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[70] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[71] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8577–8584.

[72] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.

[73] E. Sugden, S. Lloyd, J. Lam, and J. Cleland, "Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders," *Int. J. Lang. Commun. Disord.*, vol. 54, no. 5, pp. 705–728, 2019.

[74] J. Cleland, A. Wrench, S. Lloyd, and E. Sugden, "ULTRAX2020: Ultrasound technology for optimising the treatment of speech disorders: Clinicians' resource manual," Univ. Strathclyde, 2018.

[75] A. Eshky et al., "UltraSuite: A repository of ultrasound and acoustic data from child speech therapy sessions," in *Proc. InterSpeech*, 2018, pp. 1888–1892.

[76] H.-D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, no. 1, pp. 299–317, 2010.

[77] E. Kozegar, M. Soryani, H. Behnam, M. Salamati, and T. Tan, "Breast cancer detection in automated 3D breast ultrasound using iso-contours and cascaded rusboosts," *Ultrasonics*, vol. 79, pp. 68–80, 2017.

[78] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[79] P. S. Rodrigues, "Breast ultrasound image," *Mendeley Data*, vol. 1, 2017.

[80] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, 2020, Art. no. 104863.

[81] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15750–15758.

[82] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.