TSMCR: Two-stage Supervised Multi-modality Contrastive Representation for Ultrasound-based Breast Cancer Diagnosis

Bangming Gong GBM351@SHU.EDU.CN
School of Communication and Information Engineering, Shanghai University, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Contrastive learning has demonstrated great performance in breast cancer diagnosis. However, few existing works inspect label information in contrastive representation learning, especially for multi-modality ultrasound scenes. In this work, a two-stage supervised multi-modality contrastive representation classification network (TSMCR) is proposed for assisting breast cancer diagnosis on the multimodality ultrasound. TSMCR consists of two-stage supervised multimodality contrastive learning (SMCL) and deep support vector machine (DSVM). By a novel contrastive loss, SMCL handles the consistency between modalities and the sample separability. Further, two-stage SMCL learns expressive representation by gradually pulling the similar samples of positive pairs closer and pushing the dissimilar samples of negative pairs apart in the projection space. Besides, on the fusion of the multi-level contrastive representation, DSVM is to jointly learn the representation network and classifier again in a unified framework to improve the generation performance. The experimental results on the multimodality ultrasound dataset show the proposed TSMCR achieves superior performance with an accuracy of 87.51%, sensitivity of 86.67%, and specificity of 88.36%.

Keywords: Multimodality breast ultrasound; Supervised contrastive learning; Representation learning; Deep support vector machine

1. Introduction

Breast cancer is the most commonly diagnosed cancer and the second leading cause of cancer death among women worldwide Giaquinto et al. (2022). Multimodality ultrasound, including B-mode ultrasound (BUS) and ultrasound elastography (USE) has become a main tools in the clinic for breast cancer diagnosis Goswami et al. (2020). Representation learning is a critical factor in designing a breast cancer diagnostic system. Some efforts of breast cancer classification based on computer-aided diagnosis have been carried out to provide effective decision supports Han et al. (2017); Yap et al. (2018); Wu et al. (2020); Lahoura et al. (2021); Ding et al. (2022). However, learning the expressive representation remains an challenge question.

In recent years, contrastive learning methods have attracted considerable attention in representation learning for various tasks Chen et al. (2020); Radford et al. (2021); Zeng et al. (2023); Tian et al. (2020), also including medical images analysis Azizi et al. (2021); Chen et al. (2021); Muller et al. (2022); Hager et al. (2023). However, they all belong to the self-supervised algorithms, so the label information cannot be used to guide the representation learning and enhance the classification performance simultaneously. In addition, the typical contrastive losses, e.g. InfoNCE Oord et al. (2018); Hjelm et al. (2019), are originally developed for data with a single modality. If a task requires a multimodality dataset, the contrastive methods mentioned above feed the multimodality samples into two encoders directly and then calculate the loss with the typical contrastive losses, e.g. InfoNCE used in Hager et al. (2023). Therefore, the unexplored challenge in these methods is the correlation between modalities and the separability of different categories.

In this work, we propose a two-stage supervised multi-modality contrastive representation classification method (TSMCR) to assist in the diagnosis of breast cancer on multimodality ultrasound (BUS and USE). TSMCR firstly trains a two-stage supervised multimodality contrastive learning (SMCL) for multi-level representation learning from BUS and USE images, and then a DSVM for breast cancer diagnosis. Firstly, SMCL designs a novel multimodality supervised contrastive loss intended to learn powerful and effective representation. Especially, according to the guidance of label information, 3 pairs (1 positive pair and 2 negative pairs) are selected from different modalities, and then the contrastive loss is calculated by optimizing the similarity measurement of pairs. Secondly, to fuse the rich information of multimodality data, multi-level representation is learned by a two-stage SMCL with pseudo-siamese encoders (without weight-sharing) and siamese encoders (with weight-sharing). Finally, DSVM integrates DNN (consist of two-stage contrastive encode networks) and SVM classifier into the unified framework, which achieves better model training on the limited dataset and enhances classification performance. Meanwhile, the objective function fine-tunes the whole networks in a supervised manner, including encoder networks in the two-stage SMCL.

The main contributions of our proposal are shown as follows:

1) We propose a novel SMCL algorithm for representation learning on the multimodality breast ultrasound (BUS and USE). SMCL effectively learns powerful and effective representation by increasing the consistency between modalities and handling the sample separability well.

2) We develop the TSMCR framework by combining two-stage SMCL with DSVM to improve the performance of representation learning and classification simultaneously. In particular, the two-stage contrastive manner gradually improves the expressiveness of representation by pulling the similar samples closer and pushing the dissimilar samples apart, and then on the integrated multi-level representation, DSVM improves the classification performance of the whole framework.

3) The experiments verify the feasibility of the proposed TSMCR framework. Additionally, the evaluation of a real-world breast cancer dataset confirms that the proposed method outperforms several existing algorithms.

2. Related Work

Representation learning in breast images. To design an effective breast cancer diagnostic system, representation learning is a critical factor. Han et al. (2017) exploited the deep learning framework to classify breast lesions with ultrasound imaging. Yap et al. (2018) conducted breast lesion detection by deep learning that is investigated from three aspects

(Patch-based LeNet, U-Net, and Transfer Learning FCN-AlexNet). Wu et al. (2020) leveraged deep neural network (DNN) to learn the representation of breast images and classified them according to breast imaging-reporting and data system (BI-RADS). Lahoura et al. (2021) built a cloud-based system to conduct remote diagnosis, in which an extreme learning machine is applied for representation learning and classification. Ding et al. (2022) used USE to improve the performance of the learned representation from BUS by their method ResNet-GAP, which conducts localization and classification of breast lesions simultaneously.

Contrastive learning. Contrastive-based algorithms aim that the representation of similar samples are mapped close together, while that of dissimilar samples are apart in projection space. To improve the performance of representation learning, multiple effective contrastive functions have been designed, such as vanilla contrastive loss by optimizing the Euclidean distance of a pair to predict whether two inputs are similar or not Chopra et al. (2005); Hadsell et al. (2006), triple loss by enforcing the difference between one positive pair and one negative pair to be greater than a given margin Chechik et al. (2010), InfoNCE by maximizing the mutual information estimation on one positive pair and multi-negative pairs with cosine similarity Oord et al. (2018); Hjelm et al. (2019). Besides, various research works in contrastive learning have been developed and applied to different tasks, e.g. language prediction of a given image Radford et al. (2021), representation of language-image-point cloud Zeng et al. (2023), and multiview contrastive method on image and video datasets Tian et al. (2020).

Contrastive learning has also been applied for medical image analysis successfully Shurrab and Duwairi (2022). Azizi et al. (2021) conducted dermatology and chest X-ray classification by a SimCLR-based algorithm that introduced a multi-instance contrastive learning to satisfy the scene of multiple images. Chen et al. (2021) classified a novel coronavirus on the limited training samples of chest computed tomography (CT) imaging by momentum contrastive algorithm that captures expressive feature representation with a pre-text task. Muller et al. (2022) studied the relationship between local and global contrastive losses in image-text contrastive learning so that the medical downstream tasks work well. Hager et al. (2023) predicted risks of myocardial infarction and coronary artery disease on cardiac MR images and corresponding tabular data by combining two contrastive learning strategies SimCLR and SCARF.

DSVM. A unified framework combining DNN with SVM effectively improves the model performance on a small-sample. Li and Zhang (2017) proposed the deep neural mapping support vector machine (DNMSVM) for representation and classification successfully by DNN explicit mapping instead of a traditional kernel mapping. Later, some variants have also been successfully proposed to prove the effectiveness of the unified framework in different applications Okwuashi and Ndehedehe (2020); Xie et al. (2023). These works indicate that the unified framework is feasible to improve the representation learning and classification performance even training on a limited dataset.

3. Preliminaries

Vanilla Contrastive Learning. Given a dataset $T = \{(\boldsymbol{x}_i, y_i)\}_{i=1,\ldots,N}$ with N sample-label pairs. $\boldsymbol{x}_i \in R^D$ is the i-th sample corresponding label $y_i \in \{-1, 1\}$. For convenience, we further organize the input and the corresponding output by $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ and

$\boldsymbol{y} = [y_1, \ldots, y_N]$, respectively. Arbitrary $\boldsymbol{x}_i$ selected from the dataset T is taken as an anchor, and another sample $\boldsymbol{x}_j$ $(i \neq j)$ is termed as positive sample of $\boldsymbol{x}_i$ if they are from the same category (namely $y_i = y_j$), otherwise negative sample. The encoder function $f(\cdot)$ generates an embedding vector $f(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in T$. According to the vanilla contrastive functions Chopra et al. (2005); Hadsell et al. (2006), they train two encoder networks of shared parameters to decrease the distance of the positive pair, namely tuple (anchor, positive), and generate a large distance of the negative pair, namely tuple (anchor, negative). Specifically, the loss is formularized as follows:

$$L_{CL}(\boldsymbol{x}_i, \boldsymbol{x}_j; f) = \delta(y_i \neq y_j) \cdot |m - \|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\||_+^2 + \delta(y_i = y_j) \cdot \|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\| \tag{1}$$

where m is a non-negative constant, which enforces the distance between samples of the different categories to be greater than m as much as possible.

Supervised contrastive learning (SCL). Using the label information, SCL is capable of handling the case where many samples have the same category rather than InfoNCE with only one similar sample. By two random augmentations, the dataset T is transferred to $T' = \{(\bar{\boldsymbol{x}}_i, \bar{y}_i)\}_{i=1,\ldots,2N}$, where $\bar{\boldsymbol{x}}_{2i}$, $\bar{\boldsymbol{x}}_{2i-1}$ corresponds to $\boldsymbol{x}_i$, the label $\bar{y}_{2i} = \bar{y}_{2i-1} = y_i$. Let $P(\boldsymbol{x}_i) = \{j : y_j = y_i\}$ is a set of indices with the same label as y_i in the dataset T, and $|P(\boldsymbol{x}_i)|$ is its cardinality. For a given anchor $\boldsymbol{x}_i$, it is combined with multiple positive samples $\boldsymbol{x}_j$ to construct positive pairs. According to the supervised contrastive learning Khosla et al. (2020), the loss can be written as:

$$L_{SCL}(\boldsymbol{x}_i) = \frac{-1}{|P(\boldsymbol{x}_i)|} \sum_{j \in P(\boldsymbol{x}_i)} \log \frac{exp(f(\boldsymbol{x}_i) \cdot f(\boldsymbol{x}_j)/\tau)}{G} \tag{2}$$

where the denominator $G = \sum_{k \neq i} exp(f(\boldsymbol{x}_i) \cdot f(\boldsymbol{x}_k)/\tau)$. The temperature parameter τ is a non-negative constant.

4. Method

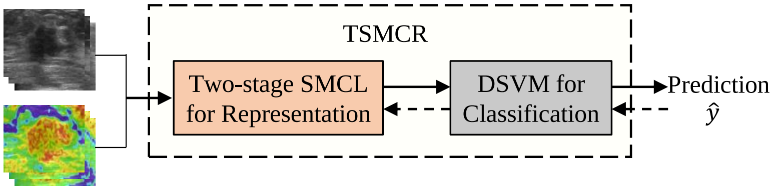


Figure 1: The overall architecture of multi-level supervised contrastive representation classification TSMCR. The supervised training manner are employed to learn TSMCR with iterated forward (solid arrows) and backward (dashed arrows).

In this section, we describe the proposed algorithm TSMCR, as shown in Fig. 1. A novel contrastive network, two-stage SMCL, is designed to extract multi-level representation with rich and diverse information. Further, DSVM is used to fine-tune the whole network and improve the generalization performance. The learning procedure of the TSMCR in detail is given below.

4.1. Supervised Multimodal Contrastive Learning

To learn powerful representation from multimodality samples, we design a supervised multimodal contrastive learning SMCL, which includes two branches of encoder and projection head (as shown in Fig. 2). In SMCL, a novel supervised contrastive loss is developed to extract expressive representation, in which the loss is calculated on three pairs, including 1 postive pair and 2 negative pairs. In particular, given an anchor, positive pair is built by the combination of the anchor and another sample (called as positive sample), which is from another modality corresponding to the anchor. The sample of positive pair naturally have the same category. Negative pair consists of anchor and another sample (called as negative sample) with different categories, but the selection of the negative sample is not constrained by the data modal. SMCL aims at pulling the similar samples (positive pair) closer and pushing the dissimilar samples (negative pair) apart. SMCL uses a pseudo-siamese architecture with two different encoders to meet the requirement of multimodality data. After training, SMCL enhances the modal consistency effectively and handles the sample separability well.

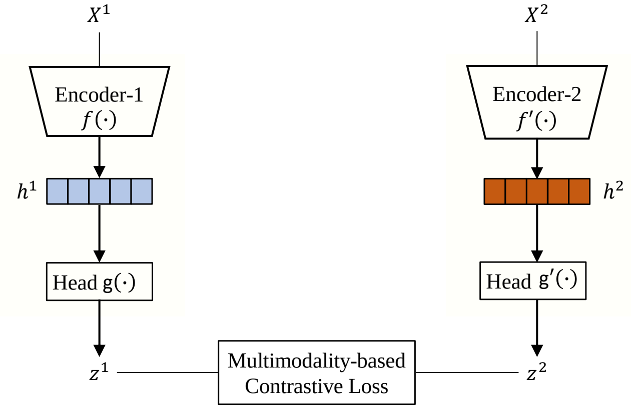


Figure 2: The architecture of supervised multimodal contrastive learning. The multimodal data are fed into the different branches of encoder and head to obtain two correlated representation.

Given a multimodal dataset $T = \{(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i)_{i=1,\ldots,N}\}$ with N sample-label tuples, the right superscript $v(=1,2)$ is used to denote the v-th modality. The $\boldsymbol{x}_i^v \in R^{D_v}$ is the feature of the i-th element corresponding the label $y_i \in \{-1, 1\}$. The sample $\boldsymbol{x}_i^1$, $\boldsymbol{x}_i^2$ are fed into encoders $f(\cdot)$ and $f'(\cdot)$, resulting in a pair of representation vectors $\boldsymbol{h}_i^1 = f(\boldsymbol{x}_i^1)$ and $\boldsymbol{h}_i^2 = f'(\boldsymbol{x}_i^2)$. Further, the projection head with a single hidden layer maps the learned representation $\boldsymbol{h}_i^1$, $\boldsymbol{h}_i^2$ to a pair normalized vector $\boldsymbol{z}_i^1 = g(\boldsymbol{h}_i^1)$, $\boldsymbol{z}_i^2 = g'(\boldsymbol{h}_i^2)$, in which they are used for calculating our contrastive loss.

Under the guidance of label information, SMCL selects 4-tuple samples from multimodality samples to construct 3 pairs. We set a sample $\boldsymbol{x}_i^1$ with label y_i as an anchor from the dataset T, while the positive sample $\boldsymbol{x}_i^2$ is selected from another modal , which has the same label as the anchor $\boldsymbol{x}_i^1$. Let the set $Y^+ = \{\tau_q | y_{\tau_q} = +1, \tau_q \in 1, \ldots, N, q \in 1, \ldots, Q\}$ and $Y^- = \{\tau_p | y_{\tau_p} = -1, \tau_p \in 1, \ldots, N, p \in 1, \ldots, P\}$ (P+Q=N) are the index set. We select a $j \in Y^+ or Y^-$ randomly to satisfy $y_j \neq y_i$, and then the corresponding samples $\boldsymbol{x}_j^1$ and $\boldsymbol{x}_j^2$ are called negative. Finally, the tuple of the learned representation and head vectors are denoted as $\{\boldsymbol{h}_i^1, \boldsymbol{h}_i^2, \boldsymbol{h}_j^1, \boldsymbol{h}_j^2\}$ and $\{\boldsymbol{z}_i^1, \boldsymbol{z}_i^2, \boldsymbol{z}_j^1, \boldsymbol{z}_j^2\}$. To measure the similarity between positive/negative pairs, SMCL employs a Euclidean distance in the projection space following as:

$$\begin{cases} D(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2) = \|\boldsymbol{z}_i^1 - \boldsymbol{z}_i^2\| \\ D(\boldsymbol{z}_i^1, \boldsymbol{z}_j^1) = \|\boldsymbol{z}_i^1 - \boldsymbol{z}_j^1\| \\ D(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2) = \|\boldsymbol{z}_i^1 - \boldsymbol{z}_j^2\| \end{cases} \tag{3}$$

The contrastive loss requires decreasing the distance of the positive pair and increasing the distance of the negative pairs, so it is formulated as:

$$L^1_{SMCL}(\boldsymbol{z}_i^1) = D(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2) + |\lambda_1 - D(\boldsymbol{z}_i^1, \boldsymbol{z}_j^1)|_+^2 + |\lambda_2 - D(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2)|_+^2 \tag{4}$$

where $[\cdot]_+$ denotes the hinge function, which keeps the inputs unchanged if the input is non-negative, returns zero otherwise. λ_1, λ_2 are non-negative constants, defining a constrictive radius so that the negative pairs contribute to the loss only if their measures are less than the radius. For constraint consistency between different modalities, we use the same margin value, i.e. $\lambda_1 = \lambda_2$. In the Eq. (4), loss L^1_{SMCL} treats the 1st modality as anchor. Symmetrically, we can get L^2_{SMCL} by anchoring at 2nd modality. Finally, we add them up as our multimodality contrastive loss:

$$L_{SMCL}(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2) = L^1_{SMCL}(\boldsymbol{z}_i^1) + L^2_{SMCL}(\boldsymbol{z}_i^2) \tag{5}$$

Algorithm 1 Supervised multimodal contrastive learning (SMCL)

Input: Dataset $T = \{(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i)\}$; Hyper-parameters set $\lambda_1, \lambda_2, Max\_iter$.
Output: Weights of encoder f, f' and head g, g', denoted by $\boldsymbol{\theta} = [\boldsymbol{\theta_1}, \boldsymbol{\theta_2}]$ and $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2]$.

```
Initialize the parameters $\boldsymbol{\theta}$ and $\boldsymbol{P}$ randomly;
Generate an index set $I^v = \{(anchor = i, postive = i, negative1 = j, negative2 = j)\}$,
  $i \in \{1, \ldots, N\}$, $v \in \{1, 2\}$, and $j \in Y^+$ or $Y^-$ $(y_j \neq y_i)$;
for iter $\in \{1, \ldots, Max\_iter\}$ do
    for $i \in \{1, \ldots, N\}$ do
        $\boldsymbol{h}_i^1 = f(\boldsymbol{x}_i^1; \boldsymbol{\theta}_1)$        # representation
        $\boldsymbol{h}_i^2 = f'(\boldsymbol{x}_i^2; \boldsymbol{\theta}_2)$
        $\boldsymbol{z}_i^1 = g(\boldsymbol{h}_i^1; \boldsymbol{p}_1)$        # projection
        $\boldsymbol{z}_i^2 = g(\boldsymbol{h}_i^2, \boldsymbol{p}_2)$
    end for $i$
    Calculate the contrastive loss $L_{SMCL}$ by Eq. (5) on the index set $I^v$;
    Update encoders $f$, $f'$ and heads $g$, $g'$ to minimize $L_{SMCL}\big(\boldsymbol{X}^1, \boldsymbol{X^2}\big)$;
end for $Max\_iter$
Return parameters $\boldsymbol{\theta}$ and $\boldsymbol{P}$.
```

where it is worth noting that $L^1_{SMCL}(\boldsymbol{z}^1_i) \neq L^2_{SMCL}(\boldsymbol{z}^2_i)$ because of the different negative samples by generating j twice randomly. For the whole dataset, the contrastive loss can be rewritten as:

$$L_{SMCL}\left(\boldsymbol{X}^1, \boldsymbol{X^2}\right) = \frac{1}{2N}\sum_{i=1}^{N} L_{SMCL}(\boldsymbol{z}^1_i, \boldsymbol{z}^2_i) \tag{6}$$

where $\boldsymbol{X}^v = [\boldsymbol{x}^v_1, \ldots, \boldsymbol{x}^v_N]$, $v = 1, 2$. By minimizing the above loss function, the increase/decrease of the similarity measurement of the positive/negative pairs in the projective space is done simultaneously. Algorithm 1 summarizes the process of the SMCL training.

4.2. Two-stage Supervised Multi-modality Contrastive Representation Classification

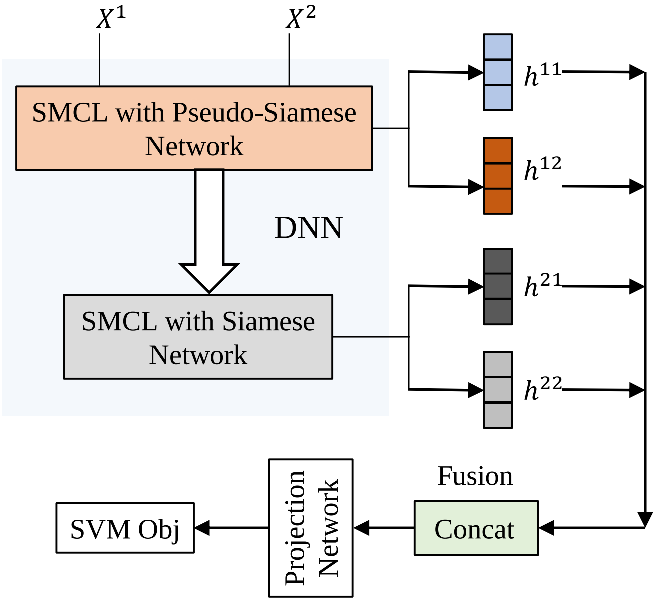


Figure 3: The architecture of our TSMCR network. Multimodality samples are fed into a two-stage supervised contrastive network to generate the multi-level representation, which is then fused together and classify by DSVM.

The overall architecture of TSMCR is shown in Fig. 3. By the pseudo-siamese and siamese encoder networks, two-stage SMCL learns the multi-level contrastive representation, in which DSVM is then used for classification based on the fused representation. First, 1st stage encoder is capable of extracting representation from the original inputs, in which using a pseudo-siamese network provides a feasible manner so that the inputs can be fed into SMCL, even if inputs have a different dimensional structure. Second, 2nd stage encoder is capable of further extracting effective representation based on the learned representation from the 1st stage SMCL. Finally, DSVM classification refines the whole network again and makes good use of the fused multi-level representation to improve the generalization performance, rather than the direct representation from contrastive learning only once. By getting together them, the TSMCR builds a framework of original features (input layer) → primary contrastive representation → high-level contrastive representation → classification, in which the supervised manner is employed in the TSMCR training. The detailed procedure of TSMCR construction is followed as:

Given dataset $T = \{(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i)_{i=1,\ldots,N}\}$, the SMCL with pseudo-siamese network is first used to learn the primary contrastive representation $\boldsymbol{h}_i^{11} = f^1(\boldsymbol{x}_i^1)$ and $\boldsymbol{h}_i^{12} = f^{1\prime}(\boldsymbol{x}_i^2)$. In the 2nd stage, the $\boldsymbol{h}_i^{11}$ and $\boldsymbol{h}_i^{12}$ are fed into the SMCL with siamese network to generate high-level representation $\boldsymbol{h}_i^{21} = f^2(\boldsymbol{h}_i^{11})$ and $\boldsymbol{h}_i^{22} = f^{2\prime}(\boldsymbol{h}_i^{12})$, in which $f^{2\prime}(\cdot) = f^2(\cdot)$. Next, the all learned representation is fused with concatenation, denoted as $\boldsymbol{h}_i^f = \left[\boldsymbol{h}_i^{11}; \boldsymbol{h}_i^{12}; \boldsymbol{h}_i^{21}; \boldsymbol{h}_i^{22}\right]$ Further, the fused representation is fed into a small neural network with a single hidden layer that maps it to the space where margin-based L2-SVM is applied. The mapping and predictive output are given by

$$\boldsymbol{h}_i = \sigma\left(\left(\boldsymbol{W}^f\right)^T \boldsymbol{h}_i^f + \boldsymbol{b}^f\right) \tag{7}$$

$$\hat{y}_i = (\boldsymbol{w})^T \boldsymbol{h}_i + b \tag{8}$$

where σ is the activation function $1/(1 + exp(-x))$. Meanwhile, we use a symbol $f(\cdot)$ to denote the subpart network with weights $\boldsymbol{W}^f, \boldsymbol{b}^f, \boldsymbol{w}, b$. Therefore, the objective function is formalized as

$$L_{obj}(\boldsymbol{X}^1, \boldsymbol{X}^2; \boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}[1 - y_i\hat{y}_i]_+^2 \tag{9}$$

where C is a non-negative trade-off parameter, $\boldsymbol{\theta}$ denote the weights of two-stage encode networks and fused network, including $f^1, f^{1\prime}, f^2, f^{2\prime}, f$. Therefore, DSVM consists of the combination of two-stage contrastive networks and SVM classification.

The whole procedure of the TSMCR is summarized in Algorithm 2. After the TSMCR has been trained, the predictive label could be achieved for a testing sample by calculating the sign on Eq. (8).

4.3. Relationship with Previous Contrastive Loss

Our supervised multimodality contrastive loss L_{SMCL} is an extend of the vanilla contrastive loss on multimodality samples. On Eq. (1), the indicative function $\delta(\cdot)$ could be removed when the inputs are from multimodality samples, since the similar and dissimilar pairs exist simultaneously. Meanwhile, our loss L_{SMCL} has a 3rd term, which constrains the loss of a new negative pair from the new modality and is similar to the 2nd term on Eq. (1).

L_{SMCL} is also closely related to the supervised contrastive loss L_{SCL} on Eq. (4). If L_{SCL} is used to multimodality samples with one positive and two negatives, the connection of between L_{SMCL} and L_{SCL} could be given by $L_{SMCL} \propto L_{SCL}$.

Algorithm 2 Training with Two-stage Supervised Multi-modality Contrastive Representation (TSMCR)

```
Input: Dataset $T$; Hyper-parameters set $\{C, \lambda_1, \lambda_2, Iter\}$.
Output: Weights of 1st, 2nd-stage contrastive encoders and classification network, denoted
        by $\boldsymbol{\theta}^{PS} = [\boldsymbol{\theta}_1^{PS}, \boldsymbol{\theta}_2^{PS}]$, $\boldsymbol{\theta}^{S} = [\boldsymbol{\theta}_1^{S}, \boldsymbol{\theta}_2^{S}]$, $\boldsymbol{\theta}^{C}$.
Initialization: randomly initialize $\boldsymbol{\theta}^{PS}$, $\boldsymbol{\theta}^{S}$, and $\boldsymbol{\theta}^{C}$.
for iter $\in \{1, \ldots, Iter\}$ do
    Stage-1 SMCL: Update network $f^1, f^{1\prime}$ to minimize $L_{SMCL}(\boldsymbol{X}^1, \boldsymbol{X}^2)$ according to
    Algorithm 1;
    Generate the stage-1 representation $\boldsymbol{H}^{11}, \boldsymbol{H}^{12}$;
    Stage-2 SMCL: Update network $f^2, f^{2\prime}$ to minimize $L_{SMCL}(\boldsymbol{H}^{11}, \boldsymbol{H}^{12})$ according to
    Algorithm 1;
    Generate the stage-2 representation $\boldsymbol{H}^{21}, \boldsymbol{H}^{22}$;
    Calculate the final objection $L_{Obj}$ according to Eq. (9);
    Use the optimizer to minimize $L_{Obj}$ to update all parameters of the TSMCR;
end for iter
Return parameters $\hat{\boldsymbol{\theta}}^{PS}$, $\hat{\boldsymbol{\theta}}^{S}$, and $\hat{\boldsymbol{\theta}}^{C}$.
```

5. Experiments

To validate the performance of the proposed TSMCR algorithm, a real-world multimodality breast ultrasound dataset with the diagnostic label is used to conduct the experiments. After the description of the dataset, we report and analyze the experimental results in detail.

Dataset. The multimodality ultrasound (BUS and USE) dataset of breast cancer was collected from the Nanjing Drum Tower Hospital, China. By the Mindray Resona 7 ultrasound scanner with the L11-3 probe, the BUS and USE images were acquired by experienced sonologists. The dataset of breast cancer consists of 264 pairs of BUS and USE images from 129 patients with benign tumors and 135 patients with malignant cancers. All the malignant cancers have been proved by the pathological diagnosis. Therefore, it could ensure the label validity for all the BUS and USE images. Moreover, this study was approved by the Research Ethics Board, and all patients had signed informed consent.

Experiment Design. In order to demonstrate the effectiveness and efficiency of our TSMCR, extensive experiments are conducted from the following aspects.

Firstly, our proposed TSMCR is compared with several classical and state-of-the-art algorithms by the three group experiments: a) SVM and DNMSVM: the primary SVM and a method Li and Zhang (2017) by combining DNN with SVM are used for breast cancer diagnosis on concatenated feature of BUS and USE. b) Kernel-based multimodal methods, including SM-MKL Xu et al. (2013), Simple-MKL Rakotomamonjy et al. (2008), where they conduct classification directly on the multimodality ultrasound. c) Contrastive learning methods: Vanilla CL Hadsell et al. (2006), Triplet Chechik et al. (2010), SimCLR Chen et al. (2020), CLIP Radford et al. (2021), and SupCon Khosla et al. (2020), where the different ultrasound modalities are taken as the different augmented views to feed into the encoders of the contrastive methods. In addition, Vanilla CL, Triplet, SimCLR, and CLIP belong to self-supervised learning, while SupCon trains a model in a supervised manner.

Secondly, we conducted an ablation experiment to verify the effectiveness of TSMCR by comparing it with the following algorithm: 1) SMCL: our TSMCR uses only 1st contrastive learning. Specially, our supervised contrastive loss directly is used to train encode network to extracts representation on multimodality samples and then classify with DSVM; 2) TSMCR-CE: the classifier based cross-entropy is used to replace SVM classifier.

Thirdly, we investigate the convergence of TSMCR and whether our supervised contrastive loss handle the similarity of positive/negative pairs effectively.

Finally, in order to investigate the multi-level contrastive representation learning, we provide a more intuitive evaluation by representation visualization in low-dimensional embedding space. All feature representation was mapped to 2-dimensional space by t-SNE.

Implementation. The code of the proposed TSMCR algorithm was implemented in Pytorch with Python 3.6. The compared algorithms, including SVM, SM-MKL, and Simple-MKL, were conducted with Matlab R2018b (MathWorks company). The rest of the comparison methods were conducted similarly to the TSMCR. All the experiments were performed using an NVIDIA Geforce 1080Ti GPU and an Intel Xeon Silver 4116 CPU on Win10. The code is available in: https://github.com/351gbm/TSMCR-ACML.

Evaluation. The variety of indices are calculated to evaluate the comprehensive performance, including classification accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), and negative predictive value (NPV). Moreover, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are also used for evaluation. Results were given by the format of the mean ± SD (standard deviation) and boldface means significantly better than other algorithms.

6. Experimental Results

In this section, we demonstrate the results of experiments designed above to investigate the effectiveness of TSMCR in the task of breast cancer diagnosis.

6.1. Performance Comparison

Table 1 shows the classification results of different algorithms on the BUS and USE datasets. The proposed TSMCR with the best results of breast diagnosis outperforms the others compared algorithms, as shown in the last row of Table 1. The best values of classification ACC, SEN, SPE, PPV, and NPV are $87.51 \pm 2.51\%$, $86.67 \pm 6.01\%$, $88.36 \pm 2.43\%$, $88.72 \pm 1.83\%$, $86.76 \pm 4.90\%$, respectively. In addition, it can be observed that 1) our TSMCR outperforms better compared with the single-modality classification methods SVM and DNMSVM, indicating that TSMCR benefits from rich information extracted by multi-level contrastive learning on multimodality ultrasound. 2) compared to the typical multimodality methods (SM-MKL, Simple-MKL), our TSMCR exhibit better results, which suggests that the effectiveness of representation learning in the unified framework by combining DNN and SVM. 3) SupCon and our TSMCR achieve better performance compared to other self-supervised-based contrastive methods, indicating that the supervised contrastive manner is conducive to learning expressive representation. Meanwhile, due to multimodality-specific contrastive and multi-level representation manner, our TSMCR further outperforms SupCon. Summarily, these facts demonstrate that the fused multi-level contrastive representation learned by TSMCR has a better power to enhance the classification performance.

Table 1: Classification results of the proposed method and existing algorithms. The best classification results are highlight in boldface. (Unit: %)

Method	ACC	SEN	SPE	PPV	NPV
SVM	83.33 ± 3.07	81.48 ± 4.54	85.29 ± 9.92	86.14 ± 7.91	81.69 ± 2.67
DNMSVM	84.85 ± 5.14	85.92 ± 6.63	83.75 ± 10.27	85.33 ± 7.67	85.38 ± 5.65
SM-MKL	84.09 ± 4.12	86.66 ± 6.19	81.41 ± 8.72	83.43 ± 6.23	85.82 ± 6.02
Simple-MKL	84.48 ± 4.03	85.92 ± 6.08	82.98 ± 6.34	84.25 ± 4.81	85.23 ± 5.36
Vanilla CL	84.48 ± 4.33	81.48 ± 6.19	87.66 ± 7.42	87.84 ± 6.75	82.18 ± 5.24
Triplet	84.48 ± 2.98	84.44 ± 6.37	84.52 ± 7.64	85.61 ± 4.97	84.37 ± 4.75
SimCLR	84.85 ± 3.35	82.96 ± 5.02	86.83 ± 3.89	86.89 ± 3.59	83.14 ± 4.11
CLIP	84.48 ± 2.73	86.67 ± 3.77	82.21 ± 4.49	83.72 ± 3.59	85.63 ± 3.67
SupCon	85.99 ± 2.54	86.67 ± 4.44	85.32 ± 4.39	86.21 ± 3.68	86.13 ± 4.01
SMCL	86.75 ± 2.62	88.14 ± 4.32	85.35 ± 7.80	86.82 ± 5.91	87.67 ± 3.95
TSMCR-CE	86.37 ± 2.98	85.18 ± 5.23	87.63 ± 2.78	87.82 ± 2.62	85.19 ± 4.52
TSMCR	**87.51** ± **2.51**	**86.67** ± **6.01**	**88.36** ± **2.43**	**88.72** ± **1.83**	**86.76** ± **4.90**

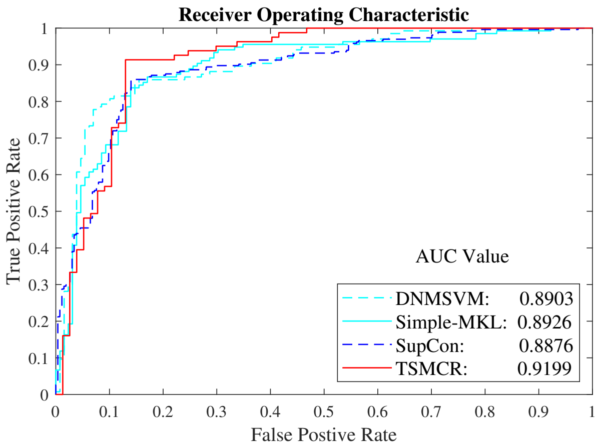


Figure 4: ROC curves and AUC values of different algorithms with the best results in each group experiments.

Fig. 4 and Table 2 show the ROC curves and the corresponding values of AUC, in which the ROC curve of TSMCR is drawn by the solid red line. For clarity, Fig. 4 only presents four ROC curves with the best performance from three group experiments, namely DNMSVM, Simple-MKL and SupCon, and the corresponding AUC is exhibited on the bottom-right. It can be found that our proposed TSMCR achieves the AUC value of 0.9199, which is superior to all the examined algorithms.

6.2. Ablation Studies

The last second and third rows of Table 1 also show the classification results of ablation studies, namely SMCL and TSMCR-CE. Compared to the contrastive methods, including

Table 2: AUC value of all experimental algorithms.

Method	AUC	Method	AUC
SVM	0.8721	SM-MKL	0.8935
DNMSVM	0.8903	Simple-MKL	0.8926
Vanilla	0.8748	SimCLR	0.8670
Triplet	0.8698	SupCon	0.8876
CLIP	0.8670	TSMCR	**0.9199**

Vanilla CL, Triplet, SimCLR, and SupCon, the results of ablation experiments demonstrate that the supervised contrastive strategy is successful with better classification performance on multimodality dataset.

The experiment results also illustrate the effectiveness of the proposed TSMCR from two aspects: multi-level contrastive representation and classifier. Firstly, benefiting from multi-level contrastive representation with rich information, the classification performance by TSMCR has been improved compared with SMCL with only 1st contrastive representation. Secondly, compared with TSMCR-CE calculated the loss by cross-entropy, our TSMCR combining DNN with SVM obtains superior performance, because the manner of the margin measurement in SVM is more related to the Euclidean distance-based loss in our supervised contrastive loss rather than probability-based measurement of cross-entropy. Therefore, by combining two-stage SMCL and DSVM, our TSMCR can be trained better in a supervised manner and generate representation with more expressiveness to improve the classification performance.

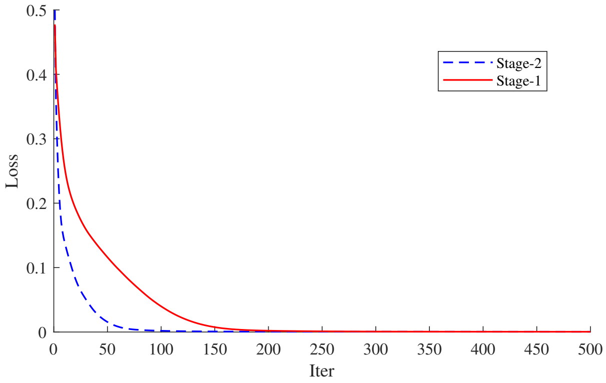
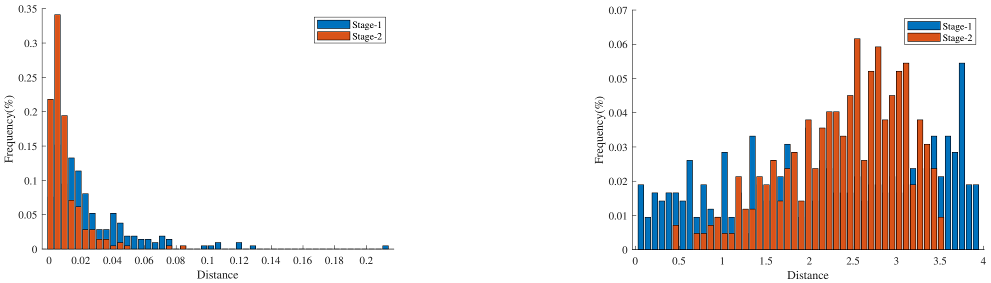


Figure 5: Training curve of TSMCR.

6.3. Convergence and Distance of Pairs

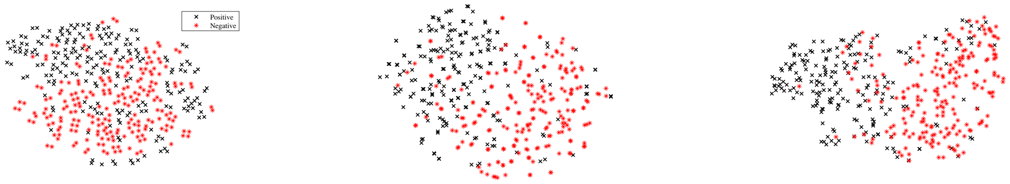
We provide the training curves in Fig. 5 to observe the effects of two-stage manner on convergence. From the figure, we can see that our TSMCR is trained well with small iterations and the contrastive loss on 2nd stage converges much faster than 1st stage learning.

The distance distribution of positive/negative pairs is explored in Fig. 6. Both of two subfigures show that the distance distribution of 2st stage contrastive learning (with brown) is more compact than 1nd stage (with blue) , demonstrating that the two-stage contrastive manner effectively handles the similarity of positive/negative pairs. Beside, in the two-stage manner, Fig. 6(a) shows that the distance of positive pairs decreases gradually while Fig. 6(b) shows that that of negative pairs rises. Therefore, our TSMCR effectively increases the modal consistency by selecting the positive pairs from the same category of different modalities. Meanwhile, our TSMCR handles sample separability well, since the negative pairs are chosen from different categories.



(a) Positive pairs (b) Negative pairs

Figure 6: The distance distribution of pairs. (a) positive pairs. (b) negative pairs.



(a) Original features (b) Stage-1 (c) Stage-2

Figure 7: Visualization of representation learning is obtained by t-SNE. Subfigure (a) is from the original features of the multimodality ultrasound. Subfigure (b) and (c) present the t-SNE projection of the contrastive representation from the 1st and 2nd stages in our TSMCR.

6.4. Visualization of Feature Representation

Fig. 7 shows the visualization results of representation learning, in which t-SNE maps the representation to 2D space. Fig. 7 (a) is the original features mapping, which indicates that the positive and negative instances, corresponding to malignant and benign breast lesions, are hardly separated on a 2D projection plane. As shown in Fig. 7 (b), in the 1st stage of contrastive learning, the mapping outputs with different categories are pushed in the opposite direction, namely black ‘ x ’to the left and red ‘ $*$ ’to the right. Further, through the

2nd stage of contrastive learning (Fig. 7 (c)), the significant separability of representation is achieved. These results demonstrate that the multi-level supervised contrastive manner in TSMCR effectively enhances the expressiveness of representation in the feature mapping space.

7. Conclusion

A two-stage contrastive classification algorithm TSMCR is proposed to assist breast cancer diagnosis on the multimodality ultrasound. TSMCR makes full use of the two-stage contrastive learning to gradually enhance the expressiveness of representation, in which the multimodality supervised contrastive loss is designed to meet the multimodal scene. On the limited dataset, TSMCR uses the unified framework by combining DNN and SVM to conduct classification well on the fused multi-level contrastive representation. Experimental results show that the proposed TSMCR outperforms all the comparative algorithms on a real-world dataset of breast cancer. It indicates that TSMCR has the potential to aid breast diagnosis on the scene of multimodality ultrasound.

Acknowledgments

This research was supported by Natural Science Foundation of Guizhou Provincial Department of Education (KY2022072). The author would like to thank Weijun Zhou and Pengli Yu from the Nanjing Drum Tower Hospital for contributing to data acquisition and annotation, as well as Hui Li for helpful discussions.

References

S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi. Big self-supervised models advance medical image classifications. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 3478–3488, 2021.

G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010.

T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

X. C. Chen, L. N. Yao, T. Zhou, J. M. Dong, and Y. Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern Recognition*, 113:107826, 2021.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively with application to face verification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005.

W. C. Ding, J. Wang, W. J. Zhou, S. C. Zhou, C. Chang, and J. Shi. Joint localization and classification of breast cancer in b-mode ultrasound imaging via collaborative learning with elastography. *IEEE Journal of Biomedical and Health Informatics*, 26(9):3374–4485, 2022.

A. N. Giaquinto, H. Sung, K. D. Miller, J. L. Kramer, L. A. Newman, A. Minihan, A. Jemal, and R. L. Siegel. Breast cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(6):524–541, 2022.

S. Goswami, R. Ahmed, S. M. Khan, M. Doyley, and S. A. McAleavey. Shear induced non-linear elasticity imaging: elastography for compound deformations. *IEEE Transactions on Medical imaging*, 39(11):3559–3570, 2020.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742, 2006.

P. Hager, M. J. Menten, and D. Rueckert. Best of both worlds: multimodal contrastive learning with tabular and imaging data. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23924–23935, 2023.

S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, and Y.-K. Seong. A deep learning framework for supporting the classification of breast lesions in ultrasound image. *Physics in Medicine and Biology*, 62(19):7714–7728, 2017.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019.

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. L. Tian, P. Isola A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS))*, volume 33, pages 18661–18673, 2020.

V. Lahoura, H. Singh, A. Aggarwal, B. Sharma, M. A. Mohammed R. Damaševičius, S. Kadry, and K. Cengiz. Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. *Diagnostics*, 11(2):241, 2021.

Y. J. Li and T. Zhang. Deep neural mapping support vector machines. *Neural Networks*, 93:185–194, 2017.

P. Muller, G. Kaissis, and D. Rueckert. The role of local alignment and uniformity in image-text contrastive learning on medical images. *arXiv preprint arXiv:2211.07254*, 2022.

O. Okwuashi and C. E. Ndehedehe. Deep support vector machine for hyperspectral image classification. *Pattern Recognition*, 103:107298, 2020.

A. V. D. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021.

A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

S. Shurrab and R. Duwairi. Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Computer Science*, 8:e1045, 2022.

Y. L. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, volume 12356, pages 776–794, 2020.

N. Wu, J. Phang, J. Park, Y. Q. Shen, Z. Huang, M. Zorin, and et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2020.

X. J. Xie, Y. F. Li, and S. L. Sun. Deep multi-view multiclass twin support vector machines. *Information Fusion*, 91:80–92, 2023.

X. X. Xu, I. W. Tsang, and D. Xu. Soft margin multiple kernel learning. *IEEE Transactions on Neural Networks and Learning Systems*, 24(5):749–761, 2013.

M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 22(4):1218–1226, 2018.

Y. H. Zeng, C. H. Jiang, J. G. Mao, J. H. Han, and C. Q. Ye. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15244–15253, 2023.