IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 June 2025, accepted 18 July 2025, date of publication 22 July 2025, date of current version 28 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3591544

RESEARCH ARTICLE

Dual-Stream Contrastive Learning for Medical Visual Representations Using Synthetic Images Generated by Latent Diffusion Model

WEITAO YE[1], **LONGFU ZHANG**[2], **XIAOBEN JIANG**[1], **DAWEI YANG**[3,4,5], **AND YU ZHU**[1]

[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[2]Department of Pulmonary and Critical Care Medicine, Shanghai Xuhui Central Hospital, Zhongshan-Xuhui Hospital, Fudan University, Shanghai 200237, China
[3]Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital (Xiamen), Fudan University, Huli District, Xiamen, Fujian 361015, China
[4]Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[5]Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai 200032, China

Corresponding authors: Yu Zhu (zhuyu@ecust.edu.cn) and Dawei Yang (yang_dw@hotmail.com)

This work was supported in part by the National Scientific Foundation of China under Grant 62476088 and Grant 82170110, in part by Xuhui District Medical Research Project under Grant SHXH202113, in part by the Science and Technology Commission of Shanghai Municipality under Grant 20DZ2254400 and Grant 20DZ2261200, and in part by Fujian Provincial Department of Science and Technology under Grant 2022D014.

ABSTRACT Deep learning-based medical image processing methods can enhance diagnostic accuracy while significantly accelerating clinical decision workflows. However, in order to learn better visual representations, such approaches usually need substantial amount of expert-annotated data, which are highly costly. To address this issue, we propose a novel approach called Dual-Stream Contrastive Learning with Cross-Scale Token Projection (DCL-CsTP), which aims to enhance visual representations and transferable initializations. Specifically, a latent diffusion model (LDM) is leveraged to generate high-quality synthetic medical images in order to expand the dataset. Then we utilize the proposed dual-stream architecture that consists of a global semantic relations stream and a local detail relations stream to learn discriminative medical image representations from the dataset. Furthermore, a cross-scale token projection is designed to enable the model to capture various scales of focus in medical images. Comprehensive experiments are performed on two downstream tasks: medical image classification and segmentation. For multi-classification of pneumonia, our DCL-CsTP method achieves 95.90% accuracy. For lesions segmentation, our DCL-CsTP method attains 89.73% dice coefficient on the International Skin Imaging Collaboration 2018 (ISIC 2018) dataset and 82.50% dice coefficient on the Kvasir-SEG dataset. The performance superiority of the model pre-trained by DCL-CsTP is conclusively demonstrated through the above experiments on various dataset, which shows that DCL-CsTP can enhance diagnostic precision and alleviate radiologists' image screening burdens.

INDEX TERMS Contrastive learning, cross-scale token projection, dual-stream, latent diffusion model, medical visual representations.

I. INTRODUCTION

In recent years, due to the exponential growth of medical imaging and the continuous enhancement of computational power, an increasing number of researchers have been utilizing deep learning to enhance the performance of medical image processing by constructing deeper and more complex networks [1], [2]. To achieve outstanding performance, networks typically leverage supervised learning with large amounts of labeled data [3], [4]. Nevertheless, within the realm of medical image processing, the collection and creation of extensive datasets present significant challenges, largely due to privacy protection considerations [5]. Moreover, even professional radiologists encounter considerable hurdles in ensuring the reliability and consistency of annotations, particularly when dealing with large-scale medical

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

data annotation [6]. Medical image segmentation annotations acutely demonstrate this challenge, where variations in annotations made by different radiologists for the same lesion can be significant [7]. Therefore, performing medical image processing method under constrained availability of high-quality annotated data poses a significant technical hurdle.

To expand the amount of data, generative adversarial networks (GANs) [8], [9] have been widely used for generating a large number of synthetic images from existing datasets. However, the pattern coverage of these GANs is constrained, rendering them unable to adequately capture the true breadth of features [10], [11]. Moreover, challenges including training instability and mode collapse impose inherent constraints on GANs' practical applicability [12]. Recently, diffusion models have established a novel pathway to image generation tasks. In this study, we utilize the latent diffusion model (LDM) [13], [14] to generate substantial amount of high-quality synthetic medical images in order to expand the dataset.

Moreover, to leverage abundant unlabeled synthetic medical images, we employ self-supervised learning (SSL) [15], [16] to obtain appropriate medical visual representations. Unlike traditional supervised learning, SSL does not require manual annotations. Instead, it automatically generates labels from the input data itself and guides the model through pre-training with proxy tasks, allowing for the exploration of intrinsic feature structures and latent relationships within the dataset [17]. Following this, we fine-tune the network in downstream tasks with minimal labeled data. SSL can substantially diminish model reliance on labels and enhance generalization performance. Recently, contrastive learning [18], [19], [20], [21], [22], [23], [24], a successful variant of SSL, has garnered significant attention due to its exceptional visual feature learning capabilities. Generally, contrastive learning can be categorized into two approaches: one utilizes both positive samples and negative samples during the training process, while the other solely relies on positive samples. The first kind of methods [18], [19], [20], [21] train model by pushing augmented views of distinct images (negative pair) apart while pulling different augmented views of the same image (positive pair) closer. These methods employ InfoNCE [22] as their contrastive loss function. On the contrary, the second kind of methods [23], [24] solely rely on utilizing the positive pair to maximize the similarity between two augmented views of an image, using cosine similarity as their loss function. In detail, the former places greater emphasis on global semantic relations, while the latter can pay closer attention to local detail relations.

Inspired by contrastive learning, recent studies [25], [26], [27], [28], [29] have adopted it as a pre-training strategy to improve the performance of downstream tasks. While these approaches demonstrate promising results, we identify two critical limitations in current contrastive learning frameworks when applied to medical imaging:

(1) **Incomplete Global-Local Feature Modeling**: Most existing contrastive learning methods focus either on global semantic relations or local detail relations, but neglect to fuse both representations effectively.

(2) **Limited Capacity of MLP Projection Head**: The multi-layer perceptron(MLP) projection head, commonly used to map features before contrastive loss computation, lack the capacity to effectively model multi-scale features.

To fill these gaps, we propose Dual-stream Contrastive Learning with Cross-scale Token Projection (DCL-CsTP) for medical visual representations. Our main contributions are summarized as follows:

- We leverage a LDM to generate high-quality synthetic medical images in order to expand the dataset.
- To address the incomplete global-local feature modeling issue, we propose a novel dual-stream contrastive learning architecture that consists of a global semantic relations stream and a local detail relations stream. The global semantic relations stream utilizes both positive and negative samples during training to learn global semantic relations, meanwhile the local detail relations stream solely leverages positive pairs to capture local detail relations.
- To overcome the limitation of the MLP projection head, we design cross-scale token projection to enable the model to capture various scales of focus in medical images.
- Combining dual-stream contrastive learning and cross-scale token projecton, we propose the DCL-CsTP. Its effectiveness is evaluated on downstream tasks: medical image classification and segmentation. The results demonstrate that our DCL-CsTP can surpass other state-of-the-art contrastive learning methods.

II. RELATED WORK

A. SYNTHETIC IMAGE GENERATION

GANs [8], [9] have been widely used for synthetic image generation. However, GANs have a problem with mode collapse, because of this, they fail to capture the entire diversity of the data distribution and instead produce limited variations or even repetitions of a few modes. In addition, GAN training can be notoriously unstable, with the generator and discriminator struggling to find a Nash equilibrium, leading to missing or unrealistic samples in the generated data. Recently, the diffusion model [13] has garnered significant interest in tackling various problems in deep learning generation. Compared to traditional GANs, the advantages of the diffusion model are its lower sensitivity to hyperparameters and its stability during both training and generation. Moreover, LDM [14] can reduce computational complexity by combining the variational autoencoder (VAE) [30] and the diffusion model while generating high-resolution synthetic images. In this study, we leverage LDM to generate high-quality synthetic medical images in order to expand the dataset.

B. SELF-SUPERVISED LEARNING FOR MEDICAL IMAGES

Given the unique characteristics of medical image data, manual annotation of massive medical image data is highly costly. To tackle this issue, SSL has emerged as a pre-training approach that leverages unlabeled data to achieve a suitable initialization for subsequent tasks that have limited annotations available. Chen et al. [31] proposed a novel self-supervised learning strategy based on context restoration, aiming to effectively leverage unlabelled images for improved learning. Cheng et al. [32] introduced an effective adaptive local prototype pooling module that eliminates the need for annotations during training. In addition, CAiD [33] aimed to enhance instance discrimination learning by leveraging diverse local context information from unlabeled medical images, thereby providing more precise and discriminative information. DiRA [34] utilized restorative learning, discriminative learning, and adversarial learning techniques to improve the effectiveness of self-supervised learning approaches in the domain of medical image analysis.

Contrastive learning, a highly effective form of SSL, has proven to be an effective approach to extract visual representations from unlabeled images [35]. Motivated by the success of contrastive learning, numerous researchers have incorporated contrastive learning techniques into the field of medical image processing. Sowrirajan et al. [25] employed MoCo (Momentum Contrast) pretraining to enhance the representation learning and transferability of Chest X-ray models. Zhang et al. [26] derived medical visual representations by leveraging contrastive learning with paired image-text data. Furthermore, researchers [36], [37], [38] have also utilized contrastive learning techniques to advance medical image segmentation tasks. Despite these efforts, we believe the existing contrastive mechanisms retains optimization potential in medical image processing.

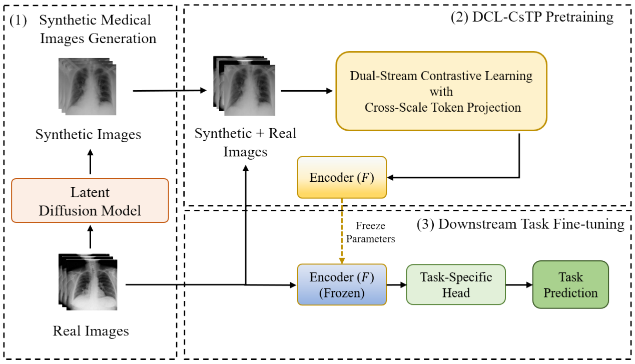


FIGURE 1. Overview of the proposed three-stage approach for medical image model training with limited data.

III. THE PROPOSED METHOD

A. OVERVIEW

As illustrated in Fig. 1, we propose a three-stage approach to enhance medical image processing with limited labeled data:

(1) **Synthetic Medical Images Generation**: Employ LDM to generate substantial amount of high-quality synthetic medical images.

(2) **DCL-CsTP Pretraining**: Leverage DCL-CsTP to pretrain the encoder using both synthetic and real medical images;

(3) **Downstream Task Fine-tuning**: Build network for downstream task with the frozen pretrained encoder (trained via DCL-CsTP) and a task-specific head. Fine-tuning is performed on limited labeled real medical images using supervised learning.

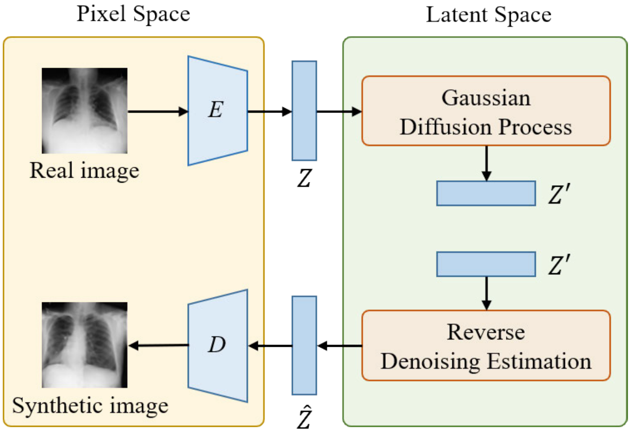


FIGURE 2. The pipeline for medical image generation using latent diffusion model.

B. MEDICAL IMAGE GENERATION BASED ON LATENT DIFFUSION MODEL

Compared to traditional diffusion model, we use the LDM [14] which can reduce computational complexity by combining the VAE [30]. In detail, LDM consists of two parts: pixel space and latent space, as shown in Fig. 2.

1) PIXEL SPACE

An encoder E is utilized to compress the pixel-level medical image into latent feature Z. In addition, the decoder D is employed to restore the denoised latent feature $\hat{Z}$ to pixel-level image. As shown in (1):

$$Z = E(x); \hat{x} = D(\hat{Z}) \tag{1}$$

where x denotes the pixel-level real medical image, and $\hat{x}$ denotes the generated pixel-level synthetic medical image.

2) LATENT SPACE

The latent diffusion model diffuses the latent feature Z by introducing Gaussian noise ε, and then estimating the data distribution by gradual denoising the Gaussian noise. The complete denoising learning process corresponds to the inverse operation of fixing a Markov chain. The denoising autoencoder ε_θ is used to predict the denoising distribution variable of Z_t which represents the result of diffusing latent feature Z with t times. (2) is the loss function of LDM:

$$\mathcal{L} = \|\varepsilon - \varepsilon_\theta(Z_t, t)\|_2^2 \tag{2}$$

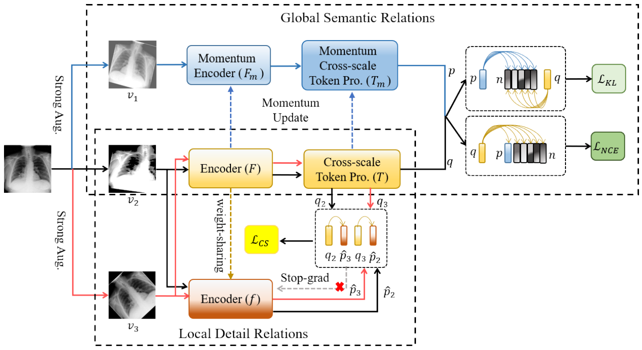


FIGURE 3. **The overall architecture of the proposed DCL-CsTP. The upper stream utilizes both positive and negative samples during the training process to focus on global semantic relations. The other lower stream solely leverages positive pairs to capture local detail relations.**

C. DUAL-STREAM CONTRASTIVE LEARNING ARCHITECTURE

As shown in Fig. 3, we design the DCL-CsTP to learn both global semantic relations and local detail relations. First, we utilize strong augmentations to create two views v_1 and v_2 as positive pair and take out negative samples n which are pre-stored in the memory bank. To capture global semantic relations, we push (v_1, n) and (v_2, n) apart and simultaneously pull (v_1, v_2) closer. Then, we take another randomly strong augmented view v_3 from the same image. To extract local detail relations, we solely rely on maximize the similarity between v_2 and v_3. Moreover, a cross-scale token projection is proposed to fuse the multi-scale features. Additionally, to strengthen model robustness against noise interference, random Poisson Noise Perturbation (PNP) augmentation is applied.

1) CAPTURE GLOBAL SEMANTIC RELATIONS

First, we apply strong data augmentations to medical image X to generate two distinct views, v_1 and v_2. Strong augmentation is a random combination of 8 types of augmentations: Rotate, Resized crop, Color jitter, Auto contrast, Equalize, Sharpness, Brightness, and PNP. Then, we employ ResNet50 [39] backbone as the feature encoder. The positive pairs are mapped via encoder (F) and momentum encoder (F_m) to extract feature. Then, cross-scale token projection (T) and momentum cross-scale token projection (T_m) are leveraged to enable the model to capture various scales of focus in medical images. Most contrastive learning methods [18], [19], [20], [21] employed InfoNCE [22] as their contrastive loss function which focuses on pushing negative pair apart while pulling positive pair closer. These methods foucs on learning the relation between query views and other views without considering the relation between positive views and other view. Inspired by [40] and [41], we use the modified version of InfoNCE loss shows in (3), where q is a query, p stands for the positive sample, n represents the negative samples, τ is a temperature hyper-parameter [42] and N is the number of negative samples that are stored in the memory bank [19] in advance. The modified version of InfoNCE enhance the role of p beyond merely being a target to be pulled closer for q. To further exploit p, we calculate the similarity between the positive and negative samples and the similarity between the query and the negative sample as $P(p, n)$ and $Q(q, n)$. Then, the symmetric Kullback-Leibler (KL) Divergence is utilized to calculate the difference between $P(p, n)$ and $Q(q, n)$, we employ $\mathcal{L}_{\mathrm{KL}}$ to keep the agreement between $P(p, n)$ and $Q(q, n)$ by minimize their difference, as demonstrated in (4).

$$\mathcal{L}_{\mathrm{NCE}} = -\log\left(\frac{\exp(q \cdot p/\tau)}{\exp(q \cdot p/\tau) + \sum_{i=0}^{N} \exp(q \cdot n_i/\tau)}\right) \tag{3}$$

$$P(p, n) = [P_1, P_2, \ldots, P_N], P_j = \frac{\exp(p \cdot n_j/\tau)}{\sum\limits_{i=0}^{N} \exp(p \cdot n_i/\tau)}$$

$$Q(q, n) = [Q_1, Q_2, \ldots, Q_N], Q_j = \frac{\exp(q \cdot n_j/\tau)}{\sum\limits_{i=0}^{N} \exp(q \cdot n_i/\tau)}$$

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2} D_{\mathrm{KL}}(P \parallel Q) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel P) \tag{4}$$

We update the weights ω_q of the encoder (F) and cross-scale token projection (T) by back-propagation, while the weights ω_k of the encoder (F_m) and token projection (T_m) are updated by momentum update [19], as shown in (5), where m is 0.999 to update the weights slowly.

$$\omega_k = m\omega_k + (1 - m)\omega_q \tag{5}$$

2) EXTRACT LOCAL DETAIL RELATIONS

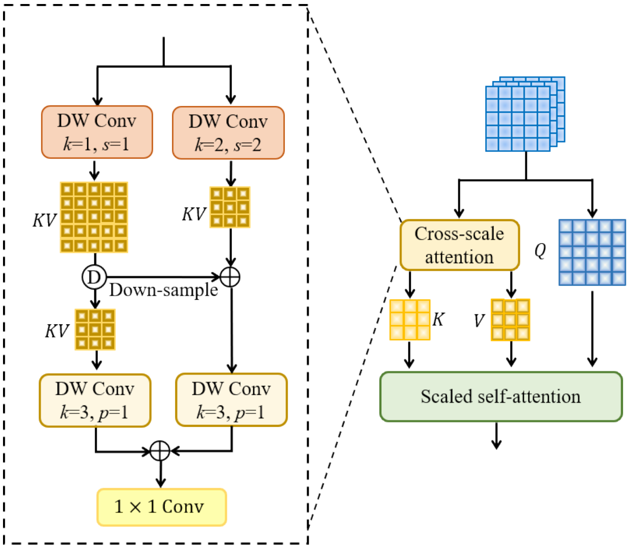
To extract local detail relations, we take two randomly generated strongly augmented views v_2 and v_3 from the same image. They are encoded by encoder (F) and then processed by cross-scale token projection (T), resulting in q_2 and q_3. Simultaneously, the encoder (f) shares weights with encoder (F) takes the same input and generates $\hat{p}_2$ and $\hat{p}_3$. To maintain a clean gradient flow, we apply stop-gradient operation to encoder (f). The loss function shown in (6) is utilized to minimize the negative cosine similarity between $q_2, \hat{p}_3$ and $\hat{q}_3, p_2$. By performing this operation, we can achieve a more refined expression of the local features.

$$\mathcal{L}_{\mathrm{CS}} = -\frac{1}{2}\left(\frac{q_2}{\|q_2\|_2} \cdot \frac{\hat{p}_3}{\|\hat{p}_3\|_2}\right) - \frac{1}{2}\left(\frac{q_3}{\|q_3\|_2} \cdot \frac{\hat{p}_2}{\|\hat{p}_2\|_2}\right) \tag{6}$$

D. CROSS-SCALE TOKEN PROJECTION

To enhance multi-scale representation learning, we propose a cross-scale token projection module as illustrated in Fig. 4(b). The core component is the cross-scale attention (CSA) block, which aims to decrease the spatial size of key/value pair while integrating multi-scale representations. The structure of the CSA is shown in Fig. 4(a). We utilize two depth-wise convolutions with different kernel sizes and strides to generate multi-scale keys and values. Following this, the feature with higher resolution undergoes down-sampling through a depth-wise convolution (kernel size 3 × 3, stride 2, padding 1), and the down-sampled feature will be added to the feature with lower resolution. Next, after applying a depth-wise convolution (kernel size 3 × 3, padding 1), we combine the two features by adding them directly. Then, a 1 × 1 convolution is employed to fuse the features extracted from different scale. Finally, we tokenize the features and feed them into scaled self-attention (SSA), which can be described by (7), where d is the dimension of the features.

$$\mathrm{SSA}(Q, K, V) = \mathrm{softmax}(\frac{Q \cdot K^T}{\sqrt{d}}) \cdot V \tag{7}$$



(a) Cross-scale attention (b) Cross-scale token projection

FIGURE 4. The pipeline of the proposed cross-scale attention (a) and cross-scale token projection (b).

E. ARCHITECTURES FOR DOWNSTREAM TASKS

After performing DCL-CsTP pre-training, we added task-specific heads after the encoder (ResNet50 [39]) to fine-tune downstream tasks:

Medical image classification: We append a simple linear classifier after a pre-trained encoder. Since a chest X-ray may contain more than one label, binary cross-entropy (BCE) loss is employed to fine-tune the entire network using, as shown in (8):

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i)] \tag{8}$$

Here, N represents the sample size for one training instance, y_i denotes the label of the sample, and p_i is the prediction of the model.

Medical image Segmentation: Following the architecture of U-Net [43], skip connections are used to merge down-sampled features with up-sampled features between the encoder and decoder. During the fine-tuning phase, the network integrates a joint application of Intersection over Union (IoU) loss and BCE loss, as shown in (9):

$$\mathcal{L}_{\mathrm{IoU}} = 1 - \frac{|A \cap B|}{|A \cup B|}$$
$$\mathcal{L} = \mathcal{L}_{\mathrm{IoU}} + \mathcal{L}_{\mathrm{BCE}} \tag{9}$$

IV. EXPERIMENTS AND ANALYSIS

A. DATASETS

1) MULTI-CLASSIFICATION OF PNEUMONIA

The COVID-19 Radiography Database [44] is collaboratively created by research teams from Qatar University and Bangladesh's Dhaka University, along with collaborating physicians. The dataset comprises 3,616 X-ray images of COVID-19 positive cases, 10,192 normal images, 6,012 images showing pulmonary opacities (non-COVID lung infections), and 1,345 images of viral pneumonia. In this experiment, 80% of the dataset (2,892 COVID, 8,153 normal, 4,809 lung opacity, 1,076 viral pneumonia) is allocated for training purposes, while the remaining 20% (724 COVID, 2,039 normal, 1,203 lung opacity, 269 viral pneumonia) is reserved for testing. All images are resized to 256 × 256.

2) SKIN LESIONS SEGMENTATION

ISIC 2018 dataset [45] is created to facilitate research and development in the field of computer-aided diagnosis (CAD) of skin lesions. The dataset is partitioned into 2,594 training samples and 100 test samples offically. All images are rescaled to standardized dimensions of 256 × 256 pixels.

3) POLYP SEGMENTATION

Kvasir-SEG dataset [46] is a publicly available dataset designed for research in semantic segmentation of gastrointestinal (GI) endoscopy images. It focuses specifically on the task of segmenting anatomical structures and lesions within the GI tract, such as the esophagus, stomach, duodenum, and colon. It randomly assigned 880 images for training and 120 for testing. All images are resized to 256 × 256.

All datasets used in this study are publicly available, de-identified, and comply with ethical/legal requirements.

B. IMPLEMENTATION AND EVALUATION

We use Python 3.7 with PyTorch 1.7.0 as our coding environment, and our experiments run on a PC with an Intel(R) i9-10940X CPU and 2 NVIDIA RTX 3090 GPUs (24 GB memory each).

1) NETWORK TRAINING AND HYPERPARAMETER SETTING

Stage 1: Latent diffusion model training. Following the Stable Diffusion model [14], the autoencoder using VAE [30] is trained to compress the pixel-level medical image with 256 × 256 into the latent feature with 64 × 64. The VAE training system is configured as follow:

(1) optimizer: Adam.
(2) learning rate: 1×10^{-6}.
(3) batch size: 4.
The training is conducted for 1000 epochs.

Then, the pre-trained VAE maintains frozen weights while encoding the pixel image into the latent space.

Subsequently, we train the latent space denoising autoencoder. We utilize the AdamW optimizer for training over 1000 epochs, employing a learning rate of 1×10^{-4}.

When generating medical images, 1000 steps of gaussian noise diffusion is applied.

Stage 2: DCL-CsTP pre-training. At the DCL-CsTP pre-training stage, the system is configured as follow:

(1) optimizer: stochastic gradient descent (SGD) with 1×10^{-4} weight decay and 0.9 momentum.

(2) mini-batch size: 128.

(3) initial learning rate: 3×10^{-2}.

Following former research [19], total training epoch number is 200 and the learning rate is multiplied by 0.1 at 120 and 160 epochs. The overall loss function employed for pre-training is shown in (10). Here, we set $\alpha = 0.3$ and $\beta = 0.7$ as hyperparameters.

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \alpha \cdot \mathcal{L}_{\text{KL}} + \beta \cdot \mathcal{L}_{\text{CS}} \tag{10}$$

During the subsequent fine-tuning stage, the system is configured as follow:

(1) optimizer: AdamW with 1×10^{-3} weight decay.

(2) mini-batch size: 32.

(3) initial learning rate: 1×10^{-4}.

The learning rate decreases according to the cosine schedule [47] over a period of 30 epochs. The number of negative samples stored in the memory bank, denoted as N, is set to 4,096 for pneumonia classification and 512 for segmentation.

Stage 3: Downstream tasks fine-tuning. After performing DCL-CsTP pre-training, we add task-specific heads after the encoder (ResNet50 [39]) to fine-tune downstream tasks. For the classification task, the system is configured as follow:

(1) optimizer: AdamW with 1×10^{-3} weight decay.

(2) mini-batch size: 32.

(3) initial learning rate: 1×10^{-4}.

We trained the network for 30 epochs.

For the segmentation task, the system is configured as follow:

(1) optimizer: AdamW with 1×10^{-5} weight decay.

(2) mini-batch size: 8.

(3) initial learning rate: 1×10^{-4}.

We trained the network for 100 epochs.

2) EVALUATION

The classification performances are assessed using accuracy (ACC), Recall (REC), and precision (PRE). The average dice coefficient (Dice), average Intersection over Union (IoU), and average Hausdorff distance (HD) are utilized for segmentation performance evaluation.

C. RESULTS OF SYNTHETIC MEDICAL IMAGES GENERATION

We employ LDM to generate synthetic medical images on the datasets. After noticing some low-quality synthetic samples in the generated results, we seek expertise from professionals in the field and opt for high-quality synthetic samples. In detail, we select 1,984 high-quality synthetic X-ray samples for the COVID-19 Radiography Database, 1,297 samples for the ISIC 2018 dataset, and 440 samples for Kvasir-SEG dataset. As shown in Fig. 5, the real images are displayed in the first and second rows, while the synthetic images are shown in the subsequent two rows. We find that LDM can generate authentic and diverse synthetic medical images.

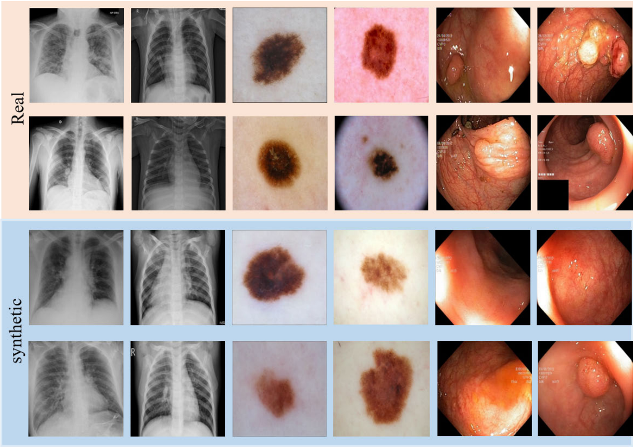
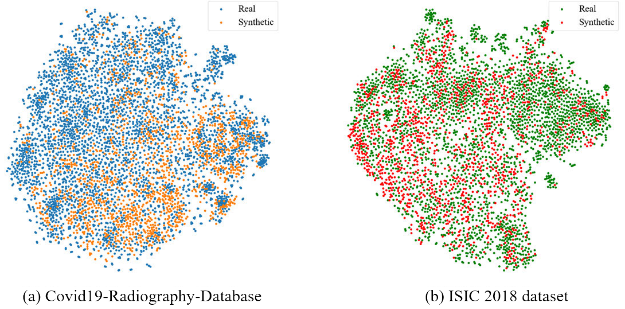


FIGURE 5. The real images and synthetic images.



(a) Covid19-Radiography-Database

(b) ISIC 2018 dataset

FIGURE 6. Visualization of the distribution of real and synthetic images.

Additionally, we conduct t-Distributed Stochastic Neighbor Embedding (t-SNE) [48] to visualize the distribution of real and synthetic images. From Fig. 6, it can be seen that the distribution of real data and synthetic data is very close. It demonstrates that LDM is a valuable tool for medical image generation.

D. RESULTS OF DOWNSTREAM TASKS

1) EXPERIMENTS ON MULTI-CLASSIFICATION OF PNEUMONIA

To evaluate the effectiveness of DCL-CsTP on multi-classification of pneumonia, we employ 6 existing supervised methods and 3 contrastive learning methods (i.e., SimCLR [18], MoCo-v2 [21] and SimSiam [24]) for comparison. Note that the supervised pre-training methods utilize the ImageNet-1k [53] dataset for its pre-training

TABLE 1. Performance comparison of pneumonia multi-classification on the COVID-19 Radiography Database. For each column, the best results are emphasized in bold.

Architectures	Pre-training		ACC(%)	REC(%)	PRE(%)
	Method	Dataset			
VGG16 [49]	Supervised	ImageNet-1K	93.65	94.54	94.24
ResNet50 [39]			94.89	94.81	95.74
ResNet50 + SE-Net [50]			95.32	95.39	95.92
ResNet50 + CBAM [51]			95.40	95.42	95.98
ResNet101 [39]			95.39	95.48	95.89
DenseNet121 [52]			94.86	95.11	95.46
ResNet50 [39]	SimCLR [18]	Self	95.21	95.89	95.80
	MoCo-v2 [21]		95.27	95.92	95.83
	SimSiam [24]		95.33	96.01	96.12
	DCL-CsTP(Ours)		95.65	96.33	96.50
	DCL-CsTP(Ours)	Self + Synthetic	**95.90**	**96.69**	**96.74**

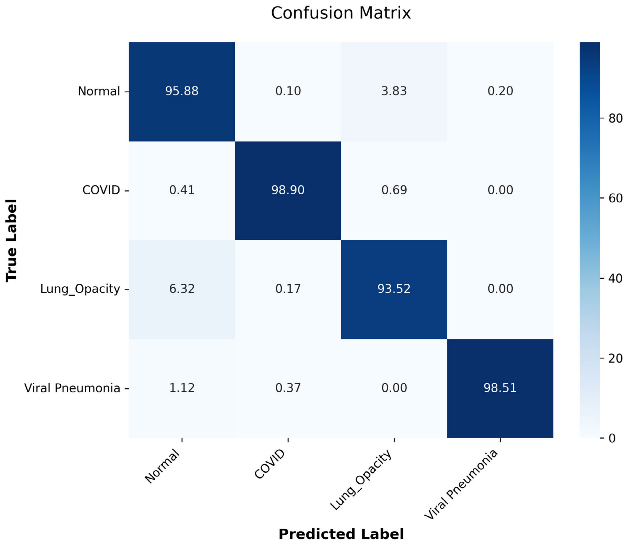


FIGURE 7. Confusion matrix of DCL-CsTP on multi-classification of pneumonia.

phase, while the contrastive learning methods leverage the self-training dataset for pre-training. For our proposed DCL-CsTP method, we perform pre-training using two different data configurations: (1) the self-training dataset alone, and (2) the self-training dataset combined with the synthetic dataset. From Table 1, we can gain the following observations:

(1) Contrastive learning methods can significantly enhance the classification performance of backbone. SimCLR [18] pre-training outperforms the ImageNet pre-trained method by 0.32% accuracy when employing ResNet50 [39] as architecture.

(2) Our DCL-CsTP achieves higher by 0.41% in recall and 0.32% in recall than MoCo-v2 [21] and SimSiam [24], respectively.

(3) With the help of synthetic images, the performance of ResNet50 [39] is improved further achieving 95.90% accuracy and 96.69% recall.

The confusion matrix of DCL-CsTP on pneumonia multi-classification task in Fig. 7 quantitatively validates the effectiveness of DCL-CsTP. This confirms its ability to learn transferable feature patterns that are applicable to downstream medical image analysis tasks.

TABLE 2. Performance comparison of skin lesion segmentation on the ISIC 2018 dataset. For each column, the best results are emphasized in bold.

Architectures	Pre-training		Dice(%)	IoU(%)	HD(%)
	Method	Dataset			
U-Net [43]	Supervised	ImageNet-1K	87.55	77.86	41.28
U-Net++ [54]			87.83	78.31	42.75
Attention U-Net [55]			87.91	78.43	41.90
ResUNet [56]			86.20	78.60	–
DeepLabV3+ [57]			88.49	80.62	34.74
U-Net [43] + ResNet50 [39]	SimCLR [18]	Self	88.29	79.87	35.68
	MoCo-v2 [21]		88.32	79.96	35.75
	SimSiam [24]		88.45	80.08	35.06
	DCL-CsTP		88.72	80.90	34.89
	DCL-CsTP	Self + Synthetic	**89.73**	**81.24**	**34.06**

2) EXPERIMENTS ON SKIN LESION SEGMENTATION

To evaluate the effectiveness of DCL-CsTP on skin lesion segmentation, we employ 5 existing supervised methods for comparison. In addition, we incorporat ResNet50 [39] as the encoding module of the U-Net network [43] and apply different contrastive learning methods to pre-train the ResNet50 network. Table 2 lists the results of comparative methods on the ISIC 2018 dataset. It can be observed that contrastive learning methods can significantly enhance the segmentation performance of the backbone. SimSiam [24] pre-training outperforms the ImageNet pre-trained ResUNet [56] by 2.25% Dice when "U-Net [43] + ResNet50 [39]" is employed as architecture. Furthermore, we utilize synthetic images to expand the dataset, thereby enhancing the quality of medical visual representations. The proposed DCL-CsTP can achieve the highest 89.73% Dice and 81.24% IoU with the help of synthetic images.

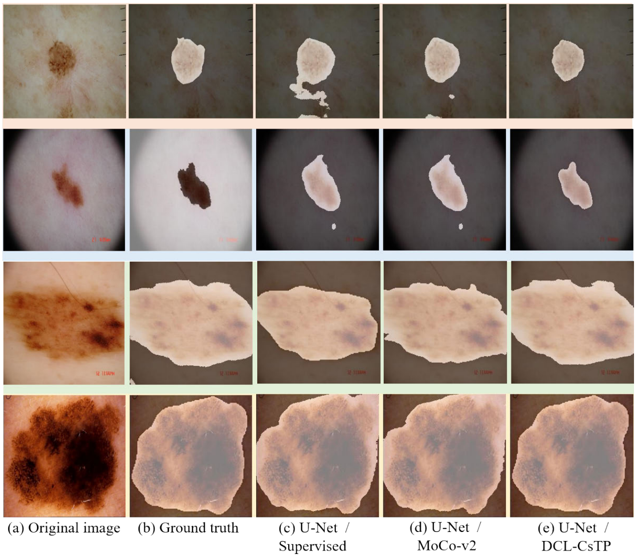

(a) Original image (b) Ground truth (c) U-Net / Supervised (d) U-Net / MoCo-v2 (e) U-Net / DCL-CsTP

FIGURE 8. Visual skin lesions segmentation results in comparison of different pre-trained methods. The initial two columns are original image data and ground truth annotations. (c)-(e) are results from supervised, MoCo-v2 and the proposed DCL-CsTP pre-trained methods.

We substantiate the efficacy of our proposed method by visualizing several segmentation results in Fig. 8. Segmentation outcomes are rendered as translucent masks and composited with source images to enhanced visualization on lesion regions' edge clarity. On the ISIC 2018 dataset, it can be observed that supervised U-Net (Fig. 8 (c)) has produced poor segmentation. Compared to the ground truth, the segmentation results of supervised U-Net have enlarged the lesion area, leading to misdiagnosis. Although the MoCo-v2 method has improved the results, it still performs poorly on the boundary of the lesion. In summary, the proposed DCL-CsTP network outperforms other methods in handling complex scenarios characterized by diverse scales and indistinct boundaries.

3) EXPERIMENTS ON POLYP SEGMENTATION

To evaluate the effectiveness of DCL-CsTP on polyp segmentation, we illustrate the results between Dice and IoU of our DCL-CsTP on the Kvasir-SEG dataset. According to Fig. 9, our DCL-CsTP attains the highest Dice score and IoU score, demonstrating precise alignment of its predictions with ground truth annotations across both lesion interiors and boundary regions. In addition, we also presented the visualization of segmentation results in Fig. 10. The qualitative results demonstrate that the proposed DCL-CsTP accurately predicts the location and boundaries of polyp lesions, irrespective of their size, whether they are large-scale or small-scale lesions.

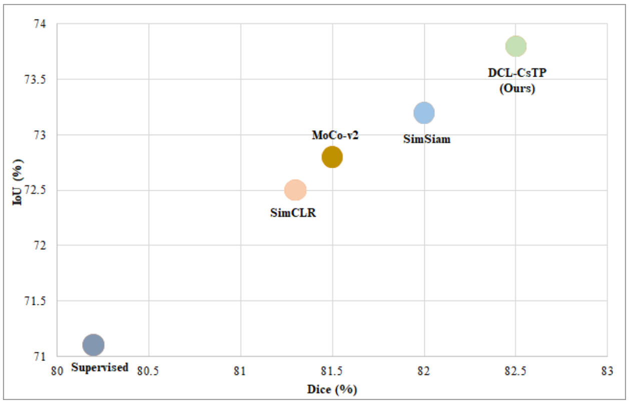


FIGURE 9. Generalizability results on Kvasir-SEG dataset. The horizontal axis corresponds to the intervals of Dice, and The vertical axis stands for the intervals of IoU. The green circle is our proposed DCL-CsTP method.

E. ABLATION STUDIES

In this section, various ablation studies are designed to validate the effectiveness of each component in our DCL-CsTP network on COVID-19 Radiography Database. First, we independently verify both local detail relations and global semantic relations, as illustrated in the first two rows of Table 3. The accuracy decreased by 0.30% and 0.48% compared to our DCL-CsTP, respectively. Then, we utilize the traditional MLP instead of cross-scale token projection and observe a decrease in the mean AUC by 1.11%. Finally, when using incorporate traditional self-attention instead of the cross-scale attention, model accuracy decreases by 0.28%. Overall, the components we have designed effectively enhance the effectiveness of pre-training.

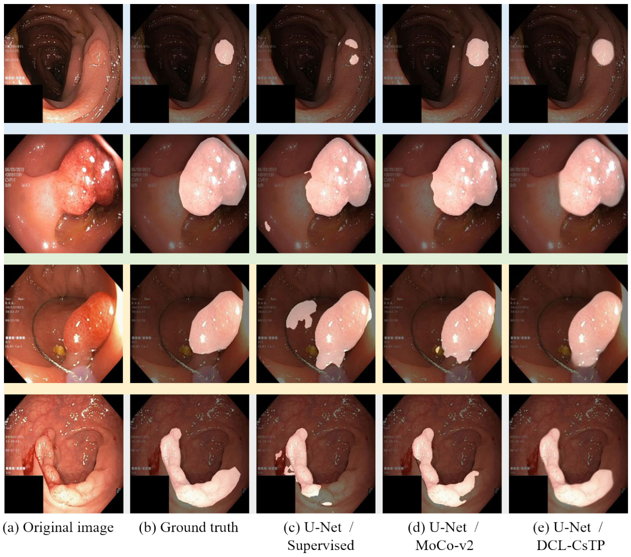


FIGURE 10. Visual polyp segmentation result comparison of different pre-trained methods. The initial two columns are original image data and ground truth annotations. (c)-(e) are results from supervised, MoCo-v2 and the proposed DCL-CsTP pre-trained methods.

TABLE 3. Ablation studies to evaluate components of DCL-CsTP on COVID-19 Radiography Database.

Methods	ACC(%)
w/o global stream	95.60
w/o local stream	95.42
w/o cross-scale token projection	94.79
w/o cross-scale attention	95.62
DCL-CsTP (Ours)	**95.90**

TABLE 4. Impact of the ratio of synthetic data on the performance of DCL-CsTP on the ISIC 2018 dataset. For each column, the best results are emphasized in bold.

Architectures	Pre-training		Dice(%)	IoU(%)	HD(%)
	Method	Dataset			
U-Net [43] + ResNet50 [39]	Supervised	ImageNet-1K	86.20	78.60	–
	DCL-CsTP(Ours)	Self	88.72	80.90	34.89
		Synthetic	88.49	79.80	36.72
		Self + Synthetic(10%)	88.91	80.98	34.67
		Self + Synthetic(50%)	89.40	81.12	34.29
		Self + Synthetic(100%)	**89.73**	**81.24**	**34.06**

Furthermore, we conducted experiments on ISIC 2018 to investigate the impact of the ratio of synthetic data on the performance of DCL-CsTP. Table 4 presents the comparison results. Note that there are 2,594 samples of ISIC 2018 and 1,297 high-quality synthetic samples. The proposed DCL-CsTP outperformed the supervised U-Net in terms of Dice scores, even when utilizing only synthetic samples. Furthermore, it can be inferred that the anticipated trend of enhancing the Dice score corresponds to an increase in synthetic samples during pre-training. These findings suggest that high-quality synthetic samples generated by LDM can effectively enhance the model's medical visual representation.

V. DISCUSSION

Recently, contrastive learning has garnered significant attention due to its exceptional visual feature learning capabilities. However, we find that the existing literature on contrastive learning has largely overlooked three crucial aspects that have the potential to advance the current state-of-the-art in medical image classification significantly. First, the collection and creation of extensive datasets present significant challenges, largely due to privacy protection considerations. Secondly, most existing works primarily focus on extracting either global semantic relations or local detail relations, but they lack effective fusion methods. Thirdly, traditional MLP projection cannot effectively model multi-scale features. To address these issues, we first leverage the latent diffusion model to generate high-quality synthetic medical images and in order to expand the dataset. Moreover, we propose the DCL-CsTP for learning better medical visual representations. According to the above experiments, we have detailed discussions as follows:

(1) Based on Table 3, it is evident that the components of DCL-CsTP we have designed effectively enhance the effectiveness of pre-training.

(2) As shown in Fig. 5 and Fig. 6, we can find that LDM can generate authentic and diverse synthetic medical images and has the potential to be a valuable tool for medical image generation. Table 4 presents the positive impact of the ratio of synthetic data on the performance of DCL-CsTP.

(3) Drawing from the data in Table 1, Table 2, Fig. 8, Fig. 9, and Fig. 10, it is evident that our DCL-CsTP method effectively boosts the performance of medical visual representations, leading to improved outcomes in medical image classification and segmentation.

The above discussion demonstrates that the proposed DCL-CsTP framework serves as an effective unsupervised learning approach for small-scale medical image datasets, The innovative dual-stream architecture significantly enhances encoder pretraining performance by simultaneously capturing both local and global image features. Moreover, our LDM-based data expansion strategy further boosts pretraining efficacy by providing diverse synthetic samples during pretraining phase. These two improvements collectively optimize the performance of the pretrained encoder, consequently improving downstream task performance.

Despite the promising advantages of our method, there are still some remaining shortcomings that need to be addressed. First, we introduce LDM-based synthetic image generation to expand the dataset. However, since LDM-generated image quality cannot be guaranteed, manual data curation remains necessary. Second, while the proposed DCL-CsTP demonstrates promising performance, its robustness against diverse data distribution shifts and adversarial perturbations remains an area for further enhancement. We will continue further investigate more effective contrastive learning approaches for medical image classification in the future, particularly in scenarios with limited annotations.

VI. CONCLUSION

In conclusion, we propose a novel method, DCL-CsTP, for medical visual representations. Specifically, the synthetic medical images generated by the latent diffusion model are utilized to expand the dataset for pre-training. Furthermore, we design a dual-stream contrastive learning architecture, where one stream utilizes both positive and negative samples during the training process to focus on global semantic relations. while the other solely leverages positive pairs to capture local detail relations. In addition, we propose a cross-scale token projection to enable the model to capture various scales of focus in medical images. Comprehensive experiments on various datasets demonstrate that DCL-CsTP can benifit medical image classification and segmentation. Overall, our proposed method holds the promise of serving as a valuable tool to aid radiologists in diagnosing and treating diseases.

ACKNOWLEDGMENT

This manuscript utilizes Deepseek-R1 [58] for grammar enhancement.

(Weitao Ye and Longfu Zhang are co-first authors.)

REFERENCES

[1] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[2] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: A review," *J. Med. Syst.*, vol. 42, no. 11, pp. 1–13, Nov. 2018.

[3] K. Zhou, E. Diehl, and J. Tang, "Deep convolutional generative adversarial network with semi-supervised learning enabled physics elucidation for extended gear fault diagnosis under data limitations," *Mech. Syst. Signal Process.*, vol. 185, Feb. 2023, Art. no. 109772.

[4] R. Jiao, Y. Zhang, L. Ding, B. Xue, J. Zhang, R. Cai, and C. Jin, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," *Comput. Biol. Med.*, vol. 169, Feb. 2024, Art. no. 107840.

[5] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger, "Deep learning in medical imaging and radiation therapy," *Med. Phys.*, vol. 46, no. 1, pp. e1–e36, Jan. 2019.

[6] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020.

[7] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101693.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[10] A. Ferreira, J. Li, K. L. Pomykala, J. Kleesiek, V. Alves, and J. Egger, "GAN-based generation of realistic 3D volumetric data: A systematic review and taxonomy," *Med. Image Anal.*, vol. 93, Apr. 2024, Art. no. 103100.

[11] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li, "Photorealistic facial texture inference using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2326–2335.

[12] D. Saxena and J. Cao, "Generative adversarial networks (GANs): Challenges, solutions, and future directions," *ACM Comput. Surveys*, vol. 54, no. 3, pp. 1–42, Apr. 2022.

[13] J. Ho, A. N. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[15] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomed. Eng.*, vol. 6, no. 12, pp. 1346–1352, Aug. 2022.

[16] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[17] K. Ohri and M. Kumar, "Review on self-supervised image recognition using deep neural networks," *Knowledge-Based Syst.*, vol. 224, Jul. 2021, Art. no. 107090.

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

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

[20] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 22243–22255.

[21] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[22] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[23] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent–A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[24] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[25] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest X-ray models," in *Proc. Medical Imaging Deep Learning*, 2021, pp. 728–744.

[26] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proc. Mach. Learn. Healthcare Conf.*, 2020, pp. 2–25.

[27] M. Fischer, T. Hepp, S. Gatidis, and B. Yang, "Self-supervised contrastive learning with random walks for medical image segmentation with limited annotations," *Computerized Med. Imag. Graph.*, vol. 104, Mar. 2023, Art. no. 102174.

[28] Y. Zhang, L. Luo, Q. Dou, and P.-A. Heng, "Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification," *Med. Image Anal.*, vol. 86, May 2023, Art. no. 102772.

[29] K. Guo, S. Zheng, R. Huang, and R. Gao, "Multi-task learning for lung disease classification and report generation via prior graph structure and contrastive learning," *IEEE Access*, vol. 11, pp. 110888–110898, 2023.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[31] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539.

[32] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervised learning for few-shot medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1837–1848, Jul. 2022.

[33] M. R. H. Taher, F. Haghighi, M. B. Gotway, and J. Liang, "CAiD: Context-aware instance discrimination for self-supervised learning in medical imaging," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 535–551.

[34] F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang, "DiRA: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20792–20802.

[35] P. Yang, X. Yin, H. Lu, Z. Hu, X. Zhang, R. Jiang, and H. Lv, "CS-CO: A hybrid self-supervised visual representation learning method for H&E-stained histopathological images," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102539.

[36] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12546–12558.

[37] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu, "Distributed contrastive learning for medical image segmentation," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102564.

[38] D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, and Y. Shi, "Positional contrastive learning for volumetric medical image segmentation," in *Proc. MICCAI*, Strasbourg, France, Sep. 2021, pp. 221–230.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] C. Wei, H. Wang, W. Shen, and A. Yuille, "CO_2: Consistent contrast for unsupervised visual representation learning," 2020, *arXiv:2010.02217*.

[41] J. Zhang, T. Lin, Y. Xu, K. Chen, and R. Zhang, "Relational contrastive learning for scene text recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5764–5775.

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

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234–241.

[44] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. H. Chowdhury, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104319.

[45] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[46] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. D. Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proc. 26th Int. Conf. MultiMedia Model.*, Daejeon, South Korea, 2019, pp. 451–462.

[47] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[48] G. Hinton and L. Van Der Maaten, "Visualizing data using t-sne journal of machine learning research," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[51] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[54] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, Sep. 2018, pp. 3–11.

[55] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[56] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[57] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder–decoder with Atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.

[58] DeepSeek-AI et al., “DeepSeek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning,” 2025, *arXiv:2501.12948*.

WEITAO YE received the B.S. degree from the East China University of Science and Technology, in 2022, where he is currently pursuing the M.S. degree.

His research interests include medical image processing, computer vision, and computer-aided diagnosis systems (CAD), with a particular focus on developing intelligent solutions for clinical applications. His work aims to bridge the gap between theoretical algorithms and clinical implementation, specifically targeting improvements in diagnostic accuracy and workflow efficiency.

LONGFU ZHANG received the M.D. degree from the Respiratory Department, Zhongshan Hospital, Fudan University, China, in 2015.

He is currently the Associate Chief Physician of the Department of Respiratory, Shanghai Xuhui Central Hospital, Zhongshan-Xuhui Hospital. He has published more than 20 papers in journals and conferences. His research interests include the personal treatment of lung cancer, respiratory intervention, CRISPR, and artificial intelligence.

XIAOBEN JIANG received the Ph.D. degree from the East China University of Science and Technology.

His experience includes the denoising method on chest X-ray images and CT images, and the detection of COVID-19 cases from denoised CXR images. He has published in journals in the cross-field of medical science and computer vision. He has been involved in publicly and privately funded projects. His current research interests include digital image processing and computer vision.

DAWEI YANG received the M.D. degree in internal medicine from Fudan University, in 2012.

He is currently the Associate Chief Physician of Zhongshan Hospital (Xiamen), Fudan University, and the Vice-Director of Shanghai Engineering and Technology Research Center of Internet of Things for Respiratory Medicine. He is dedicated to the early diagnosis of lung cancer and relevant studies. Since 2011, he has published 41 SCI research articles and 16 as the first author or corresponding author, including those on *American Journal of Respiratory and Critical Care Medicine*, in 2013, *Cancer Letters*, in 2015 and 2020, and *Cancers*, in 2015, 2018, and 2020.

YU ZHU received the Ph.D. degree in optical engineering from Nanjing University of Science and Technology, China, in 1999.

She is currently a Professor with the Department of Electronics and Communication Engineering, East China University of Science and Technology. She has published more than 150 papers in journals and conferences. Her research interests include artificial intelligence, image processing, computer vision, multimedia communications, and deep learning.