Generative Self-Supervised Learning for Medical Image Classification

Inhyuk Park[1,3][0009−0001−2883−8994], Sungeun Kim[1][0009−0006−2819−8006], and Jongbin Ryu*[1,2][0000−0001−5574−5358]

[1] Deparment of Artificial Intelligence, Ajou University, Suwon, Korea
[2] Department of Computer Engineering, Ajou University, Suwon, Korea
{kimsungeun,jongbinryu}@ajou.ac.kr
[3] VUNO Inc., Korea
{inhyuk.park}@vuno.co

Abstract. This paper introduces the generative self-supervised learning method in medical image recognition. We use the generative models in two main ways: 1) creating diversified training data and 2) learning domain-aligned pretext knowledge for self-supervised learning. In general, gathering real-world medical data can be quite difficult, so we generate synthetic training data using the diffusion model with elaborated prompts. We also propose a domain-aligned generative approach for our self-supervised learning algorithm. Our approach learns the robust visual representation from the masked autoencoder model with adaptive instance normalization. It minimizes the domain gap between our synthetic training data and real-world data when training the masked autoencoder model. In this self-supervised learning process, we rely solely on generative data, allowing our approach to achieve state-of-the-art performance without utilizing any real-world medical data. We demonstrate that our approach surpasses the previous best results by significant margins of CheXpert, COVIDx, and ChestX-ray14 datasets. These results highlight the potential of generated data in medical image recognition, a field that has historically faced data scarcity. We open-source our implementation of the generative self-supervised learning method at: https://github.com/inhyukpark2/gen-ssl.

Keywords: Medical Image Classification · Self-Supervised Learning · Generative Model

1 Introduction

Accurate labeling of medical data is quite costly and challenging because it requires a deep understanding of the disease. Hence, label-free Self-Supervised Learning (SSL) methods [56,62,19] have garnered significant interest in the area of medical image recognition. Recently, SSL methods have mainly utilized Vision Transformer (ViT) [16] as a backbone model. Therefore, because of ViT's less

* Corresponding author

emphasis on inductive bias properties [46], a tremendous amount of training data is required to learn sufficient self-supervision [68,28]. However, constructing large-scale medical training data has essential limitations. It has 1) private data that could cause privacy issues [1] and 2) class-imbalanced distribution data that may result in unfair classification. [45] Regarding this matter, we raise a critical issue of self-supervised learning from medical data: the scarcity of real-world medical data, regardless of its labels. In other words, while previous studies have utilized SSL on real-world data due to a lack of labels, we tackle the problem that their approach lacks adequate training data for learning effective self-supervision. Insufficient training data may negatively affect the model's ability to generalize its usefulness in real-world clinical settings. Therefore, it is crucial to tackle the limited training data issue to enhance the efficacy of self-supervised approaches.

Therefore, we introduce a new approach to SSL for medical image recognition through generative models. In the proposed approach, we generate diversified medical images from the diffusion models with Large Language Model (LLM)-guided prompts. Although we can generate different images from the diffusion model [24] using random vectors, there is a concern about bias toward the data distribution the generative model was trained on. These concerns suggest that the generated images might not accurately contain the information of real-world medical data. Therefore, we guarantee the diversification of images generated by the diffusion model through the use of prompts that fairly reflect the data distribution of medical images. Despite our approach being able to generate unlimited images, we still face the challenge of ensuring that its distribution is aligned with real-world data. Therefore, we employ the Adaptive Instance Normalization(AdaIN) [29] method to align our generated data with real-world ones, followed by a pretraining stage of SSL. For this SSL on the domain-aligned generated data, we adopt the Masked Autoencoder(MAE) [21]. Currently, most SSL methods for medical image recognition employ a contrastive learning approach [9] to learn the semantic information between positive and negative pairs. However, we argue that this approach restricts applicability to our generated medical images due to the uncertainty in determining positive pairs. This is because, in medical images, the lesion information can be found in small regions, so when the lesion region of an anchor image is removed or transformed, it might not be the positive sample of the anchor image. Put differently, the contrastive learning method, which uses the augmentation method to produce positive samples, is hard to use on medical images due to the risk of distortion in small lesion areas. Thus, we use the MAE model in our generative SSL approach.

In all these processes of our approach, we do not use any real-world data; instead, we exploit the generative models to learn the pretext knowledge from the SSL method. We refer to this GENerative SSL method as **GEN-SSL**. By leveraging our generative model's capacity to create unlimited training data, our GEN-SSL method can enhance classification performance. This performance improvement is noteworthy, where the lack of training data is an unavoidable problem in medical image recognition tasks. Further, we show that our GEN-SSL performs favorably against existing SSL methods as well as other medical image

recognition approaches. This result indicates that our generated approach is carefully designed to represent real-world data precisely, enabling the model to learn the generalized feature distribution from a large amount of the generated medical images. In our experiments, we generate up to 3M images to pre-train the models with the SSL method and fine-tune them on three widely used medical image datasets such as CheXpert, COVIDx, and ChestX-ray14.

Our paper's contribution is outlined below:

- We introduce a self-supervised learning framework using only generated images. The proposed framework efficiently integrates image generation methods with the diffusion model and AdaIN, a domain-aligning normalization, and elaborated prompt generation methods.
- We specifically present an SSL method adopting adaptive instance normalization to align the generated medical images with real ones. In addition, we propose a practical method to generate high-quality medical images with prompts generated by LLM's guidance, which results in higher SSL performance.
- We demonstrate that our generative approach improves medical image recognition performance as more images are used. This result highlights the significance and potential of the generative models in the medical image recognition field, where collecting real data is quite challenging.

2 Related Work

2.1 Generative Models in Medical Image Recognition

Generative models have been effectively utilized in various ways to improve the performance of medical image recognition models. Generative Adversarial Networks(GANs) [18] have been exploited to train models from generated images with real ones in several tasks such as cardiac segmentation [11], brain vessel segmentation [35] and class imbalance medical image recognition [45]. In recent years, diffusion models [24] have gained much popularity in the realm of image generation due to their outstanding performance [15,42]. It is remarkable that the diffusion model produces high-resolution images by progressively removing noise during its generation process. Therefore, research is currently underway on generating medical images using a diffusion model for nuclei segmentation [55] and skin lesion segmentation [7] tasks. These approaches aimed to generate medical images to supplement the limited train data of real images, thereby enhancing the model's performance.

2.2 Prompt Method in Medical Image Recognition

The Large Language Model (LLM) has demonstrated its proficiency in understanding text information within the medical domain [43]. There is a lot of ongoing research using LLM, such as GPT-4 [2], to obtain medical knowledge [65,49]. In

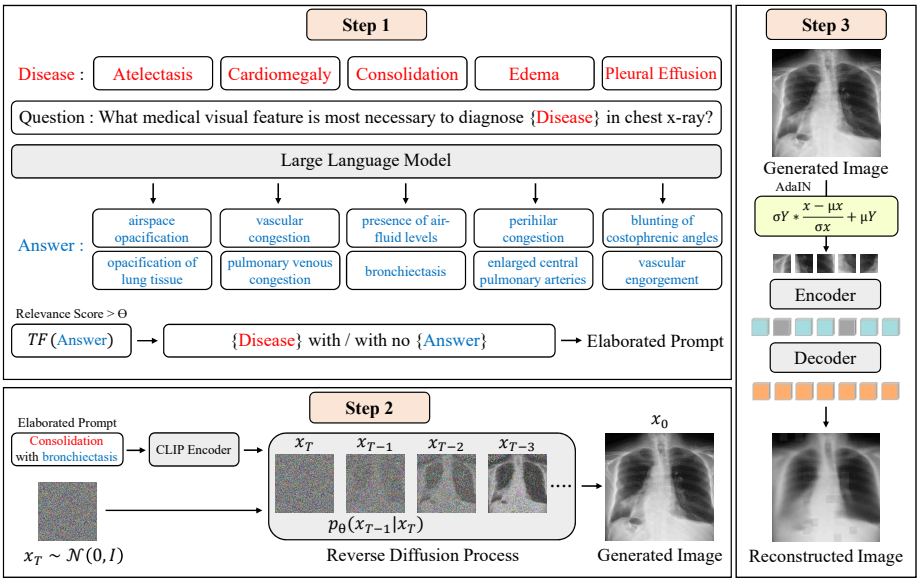


Fig. 1. Workflow of the proposed GEN-SSL. **Step1.** We generate elaborated prompts using relevance scores between LLM-guided prompts and real medical reports. **Step2.** We generate medical images with the reverse diffusion process from elaborated prompts and Gaussian noises. **Step3.** We train the MAE model using the statistical normalization method of AdaIN.

addition, research has been carried out to utilize CLIP [52] zero-shot medical image recognition by incorporating medical knowledge obtained from LLM as a text prompt [48,39]. This research explores new possibilities in medical image recognition by harnessing the capabilities of LLM. Based on recent research, we utilize LLM to create medical images that precisely capture the visual information of the disease from a diffusion model.

2.3 SSL in Medical Image Recognition

Traditionally, SSL has utilized Siamese architecture for training models, where the weight parameters of neural networks are shared [10]. Specifically, contrastive learning [8] employs data augmentations on an anchor image to train models by producing positive and negative pairs without labels. However, recently, the MAE [21] method based on masked image modeling, which erases and restores image patches, has also received attention. The MAE method enables the model to learn important visual representations of the image through the repeated process of removing and restoring image patches. This approach can be particularly useful in fields where structural information is crucial, such as medical images. In the medical image recognition field, SSL is particularly widely used because labeling

disease classes is time-consuming and expensive for constructing training data. SSL has been applied to various studies such as disease recognition [40,58,4], 3D image classification [12], and brain tumor classification and staging tasks [37]. These studies contribute to effectively classifying lesions in medical images by reducing the burden of labeling tasks. However, all of these SSL-enabled medical image recognition studies use only real images, thereby restricting their capacity to learn sufficient self-supervision. This limitation arises from the inevitably limited number of real medical images, which may not be sufficient for SSL methods that require large amounts of training data. Therefore, in this paper, we introduce the generative SSL approach that leverages the generation models for easily obtaining large-scale medical images. The proposed approach pre-trains the model using the SSL method with generated medical images and fine-tunes it using real ones.

3 Method

Here, we initially introduce our approach to medical image generation, leveraging the diffusion model with elaborated prompts. Then, we present our MAE method with statistical normalization to train SSL model medical images effectively. The overall workflow of the proposed GEN-SSL is illustrated in Fig. 1.

3.1 Elaborated Prompts for Medical Image Generation

When working in the medical domain, generative models must generate realistic images while maintaining the disease information [45]. Thus, it is crucial that the disease information should be accurately depicted in an image, and at the same time, a background image should also be created to complement the disease's information. To consider this matter in generating medical images, it is essential to utilize generative models trained on large datasets containing diverse cases. Therefore, for this reason, we adopt the diffusion model [51] trained on MIMIC-CXR [31] that contains relatively large-scale medical training images. We used the implementation of the diffusion model available in the GitHub repository[4]. The diffusion model is published in MONAI [5], the open-source framework for medical image processing. The MIMIC-CXR dataset includes five main disease labels, which are used in the prompts to generate images. The simplest strategy for creating the prompts for the diffusion model is to use the disease name itself, such as 'atelectasis' or 'pleural effusion'. While this approach is accurate in terms of disease information itself, its capability to create a diverse range of high-quality medical images is restricted. This is because the diffusion model can produce better images with semantically detailed prompts [36,66]. For example, a simple prompt of 'cat' might be a great way to draw a cat, but there is no guarantee that it will generate an image and background containing detailed context. On the other hand, prompts with detailed semantic information, such as 'A black

[4] `https://github.com/Warvito/generative_chestxray`

cat with long hair sitting on a chair' tend to generate higher quality images with detailed context information.

Heuristic Prompt. Therefore, we elaborated on the prompts by including details about the location and size of the disease in the prompts, which we define as heuristic prompts. We produce a total of 30 prompts as Big left-sided atelectasis' or 'Small bilateral pleural effusion'. This approach can potentially enhance the quality of the generated images beyond the limitations of solely relying on the disease name. All the heuristic prompts used in this paper are introduced in Tab. S1 of supplementary material.

Elaborated Prompt. We utilize LLM to represent more accurate visual features of disease information in a diverse set of prompts. We ask the question, 'What medical visual feature is most necessary to diagnose {disease label} in chest x-ray?' to get the prompt set relevant for each disease label through GPT-4 [2](`https://chat.openai.com/`). However, they include prompts that are unrelated to the visual aspects of the real disease information, thus requiring refinement to remove the irrelevant prompts. To achieve this goal, we perform the prompt refinement by computing the relevance score ψ between the elaborated prompts and real text data of medical reports in the MIMIC-CXR dataset as

$$\psi_p = \frac{1}{N} \sum_{i=1}^{N} TF(p, r_i), \tag{1}$$

where p denotes the elaborated prompt, r represents a medical report, N is the number of reports in the MIMIC-CXR dataset used for training, and TF stands for the Term Frequency score [57]. We remove elaborated prompts where their relevance score ψ is below the threshold value of 0.015. We further augment these prompts by adding a prepend of 'with' or 'with no'. Each of the prepend defines the positive or negative class information for the diseases. We take inspiration for this prepend strategy from a previous study [48] where they found that the existence of a disease class can be effectively described by prompts of 'with' or 'with no'. Therefore, our final prompt can be 'Atelectasis with lung collapse', including the disease class ('Atelectasis'), its existence ('with'), and a visual feature ('lung collapse') guided by LLM. There are a total of 66 prompts, all of which are introduced in the Tab. S2 of supplementary material. In the subsequent paper, we will refer to this as **elaborated prompts**. Using these elaborated prompts, we generate the medical images from the diffusion model.

3.2 Learning Self-supervision from Generated Images

Image Generation from Elaborated Prompt. We introduce the image generation method from the proposed elaborated prompt. We generate medical images using a generative model with the reverse diffusion process conditioned by the elaborated prompt. For the reverse diffusion process, we employ the previous model [53,51] as:

$$x_{t-1} = D_\theta \left(x_t, t, E(p) \right) \quad \text{for } t = T, T-1, \ldots, 1, \tag{2}$$

$$x_0 = D_\theta \left(x_T, E(p) \right), \tag{3}$$

where x_0 is the final generated image, $x_T \sim \mathcal{N}(0, I)$ denotes the initial Gaussian noise, and p represents the elaborated prompt. E is the CLIP encoder [52] to get the prompt embedding $E(p)$, and t stands for the index of reverse diffusion process D_θ. While generated images from the reverse diffusion process contain disease information well through our elaborated prompts, its domain is not aligned with the real medical image dataset.

Self-supervised Learning with Domain Alignment. We present the self-supervised learning method on a generated dataset aligned with the real medical domain using the Masked Autoencoder (MAE) [21] and Adaptive Instance Normalization(AdaIN) [29]. MAE learns visual representation without labels by performing a pretext task involving randomly masking and restoring it in their encoder-decoder architecture. It compares the reconstructed images from MAE with their unmasked original images so that the common visual representation can be learned. We use the generated images from LLM guidance for MAE, but there remains an important issue of domain misalignment. Although we generate images from diffusion models trained on real medical data, there is still an unavoidable domain gap. Recent studies [32,41] have also indicated that reducing the potential domain gap is effective for training medical image recognition models. Therefore, to achieve domain alignment, a statistical normalization method of AdaIN is introduced to our SSL method as

$$\mathcal{S}_Y(x) = \sigma(Y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(Y), \tag{4}$$

where x denotes a generated image, Y is a set of real images (*i.e.* MIMIC-CXR dataset), and $\mu(\cdot)$ and $\sigma(\cdot)$ are the operators to compute the mean and standard deviation. This statistically normalized data aligned in the real image space can train MAE to be more robust. Therefore, the domain gap between real and our generated medical images can be reduced by the AdaIN. We will show the performance improvement from our domain-aligned learning approach in the subsequent experimental section.

4 Experiments

4.1 Experimental Settings

Dataset and Metric. Medical images are generated using elaborated prompts in order to train models with our GEN-SSL. For all experiments, 0.3, 0.5, 1.3, and 3 million images are generated, with an equal number of images for each class label. We perform extensive experiments on three labeled downstream datasets to evaluate the performance of our GEN-SSL method. As the evaluation metric, we use Mean Area Under Curve (AUC) for the CheXpert and ChextX-ray14 datasets and Accuracy and COVID-19 sensitivity for the COVIDx dataset following the previous studies [64]. The CheXpert [30] dataset comprises 191,028 frontal-view

Table 1. Experimental results on the CheXpert dataset.

Method	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	mAUC (%)
Allaouzi *et al.* [3]	72.0	88.0	77.0	87.0	90.0	82.8
Hosseinzadeh *et al.* [25]	-	-	-	-	-	87.1
Chexclusion [54]	81.2	83.0	90.0	88.3	93.8	87.3
Dira [20]	-	-	-	-	-	87.6
Chexpert [30]	81.8	82.8	93.8	93.4	92.8	88.9
Label-assemble [33]	82.1	85.9	**94.4**	89.2	93.6	89.0
MED-MAE [64]	82.7	83.5	92.5	**93.8**	94.1	89.3
Pham *et al.* [50]	82.5	85.5	93.7	93.0	92.3	89.4
GEN-SSL (ours)	**86.2**	**90.9**	92.7	88.3	**94.4**	**90.5**

Table 2. Experimental results on the COVIDx dataset. †denotes results from [64] using 31 evaluation images due to different dataset versions. The other results are reported on 400 evaluation images. We report single model-based results.

Method	Accuracy	COVID-19 Sensitivity
COVIDNet-CXR Small [61]†	92.6	87.1
COVIDNet-CXR Large [61]†	94.4	96.8
MED-MAE [64]	95.3	95.5
GEN-SSL (ours)	**98.0**	**97.8**

images with 5 disease class labels where we follow the official training and test data split protocols. The COVIDx (version 9A) [61] includes 30,130 and 400 test images with 3 disease class labels. The ChestX-ray14 [63] dataset contains 112,210 frontal-view images, divided into a training set of 75,312 and a test set of 25,596 following the official data split. It includes a total of 14 disease class labels.

Implementation Details. We train our MAE model with ViT-S and ViT-B [16] using the hyperparameters following the previous study [64]. Specifically, we use 90% masking ratio (0.5∼1.0) for random resized crop ratio for MAE training. We utilize 800 epochs, including 40 warm-up epochs with a learning rate of 1.5^{-4}. We fine-tune all three downstream datasets with a batch size of 1024 and a learning rate of 2.5^{-4}. They are trained for 75 epochs with 5 warm-up epochs. We apply augmentation methods such as random resize crop, horizontal flip, and rotation to an input image size of 224×224. We set the RandAug [13] magnitude to 8, the DropPath [27] to 0.3, and layer-wise LR decay to 0.65 for all datasets.

4.2 Experimental Comparison with SOTA methods

We compare the proposed GEN-SSL with previous State-Of-The-Art (SOTA) methods in three downstream datasets as shown in Tab. 1, 2, and 3. The results are obtained by pertaining ViT-B only with the generated 3M images and then fine-tuning it with three different downstream datasets. We compare ours with

Table 3. Experimental results on the ChestX-ray14 dataset. We use mAUC metric for the evaluation.

AGCL[59]	Swinchex[60]	Chexclusion[54]	Acpl[38]	Xprotnet[34]	MED-MAE[64]	GEN-SSL(ours)
80.3	81.0	81.2	81.8	82.2	83.0	**83.2**

previous results obtained mostly from CNN (*i.e.* ResNet [22] and DenseNet [26]) or ViT [16] backbones, which use pre-training datasets such as ImageNet [14].

Comparison with SOTA on CheXpert dataset. We compare the proposed GEN-SSL with existing SOTA methods on the CheXpert dataset. Recently, Dira [20] presented an integrated approach that combines discriminative, restorative, and adversarial learning methods to enhance the semantic representations of networks. There have been smoothing regularization approaches [33,50,67] for multi-label classification of medical images. In this experimental comparison, our GEN-SSL works well in most cases, and in particular, ours achieve a very high mAUC score for two disease labels: Atelectasis and Cardiomegaly.

Comparison with SOTA on COVIDx dataset. We evaluate the performance of the proposed GEN-SSL on the COVIDx dataset in comparison with the COVIDNet-CXR [61] and MED-MAE [64]. COVIDNet-CXR [61] employed a human-machine collaborative approach to create a lightweight model that maintains high expressiveness of the medical image features. MED-MAE introduced the MAE-based SSL approach, which specializes in medical image recognition. In this experiment, our GEN-SSL achieved considerable performance improvement over these two SOTA methods.

Comparison with SOTA on ChestX-ray14 dataset. In Tab. 3, the classification performance on the ChestX-ray14 dataset is compared with the existing SOTA methods. Except for Swinchex [60] and MED-MAE, all results are obtained using CNN architecture such as DenseNet and ResNet. Compared with these SOTA methods, our GEN-SSL achieve better performance on the ChestX-ray14 dataset. In all three of these datasets, the proposed GEN-SSL consistently performs favorably compared to SOTA methods, notably in comparison to MED-MAE [64], which is the recent SSL approach for medical image recognition. These results demonstrate the effectiveness of our generative approach for SSL method in medical image recognition, as we focus on the SSL method using only generated images instead of real ones. This demonstrates its ability to tackle the problem of limited medical images successfully.

4.3 Ablation Study

In Tab. 4, we perform comprehensive ablation studies regarding the number of generated images, elaborated prompts, and domain alignment with the statistical normalization method. We also evaluate the performance changes due to the number of elaborated prompts in Tab. 5. In this experiment, Tab. 4 shows that the model pre-trained with our generated images performs better than

Table 4. Experimental results of the ablation study. The proposed generated dataset with elaborate prompts performs well for our SSL approach along with the statistical normalization method of AdaIN. #N denotes the number of training samples used in our SSL.

Dataset	Prompt	#N (M)	AdaIN	CheXpert		COVIDx		ChestX-ray14	
				ViT-S	ViT-B	ViT-S	ViT-B	ViT-S	ViT-B
ImageNet-1K	✗	1.3	✗	86.9	88.3	93.0	96.5	78.0	80.7
MIMIC-CXR		0.3		88.4	89.2	96.3	97.0	80.9	81.4
Generated	Disease	0.3	✗	88.0	88.9	95.8	97.0	80.8	81.1
	Disease	0.5		88.3	89.4	96.3	97.3	81.1	81.6
	Heuristic	0.5		88.5	89.6	96.3	97.5	81.3	81.7
	Elaborate	0.5		88.8	89.7	96.5	97.8	81.5	82.0
Generated	Elaborate	0.5	✓	89.1	89.9	96.8	97.8	81.6	82.4
		1.3		89.3	90.2	97.0	98.0	81.9	83.0
		3		**89.5**	**90.5**	**97.3**	**98.0**	**82.0**	**83.2**

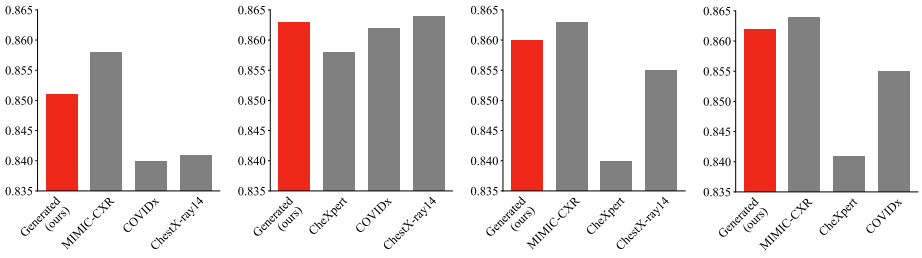
Table 5. Experimental results regarding the number of prompts (#N) used for generating images. We use 510k generated images with the elaborated prompt for all results.

#N	CheXpert	COVIDx	ChestX-ray14
5	89.1	96.8	81.7
30	89.3	97.3	81.9
66	89.7	97.8	82.0

the model pre-trained with ImageNet-1k. It is also confirmed that the more prompts we use to generate images, the more diverse visual representations the model learns, which ultimately improves the recognition performance. Based on the results of all these ablation studies, using more generated images with elaborated prompts and domain alignment methods results in notably improved performance in comparison to the respective baselines. In particular, it is evident that performance improves consistently as more generated images are used for training. This finding highlights the potential of generated images in training models for medical data, especially considering the challenge of collecting real medical images.

4.4 Quality Evaluation of Generated Images

We evaluate the quality of our generated medical images using their similarity to the real images. We assume that the data quality increases as the generated images become more similar to the real images. The reason is that when the similarity to the real images is high, it indicates the generated images have more



(a) CheXpert **(b)** MIMIC-CXR **(c)** COVIDx **(d)** ChestX-ray14

Fig. 2. The similarity between datasets with FAISS. We measure similarities between five datasets, including our generated images and four real medical image datasets. For example, (a) shows the similarity between CheXpert and the other four datasets, respectively. In this comparison, our generated dataset shows competitive similarity scores compared to the real datasets.

Table 6. Average similarity scores of five datasets. We report the average similarity scores of each dataset with the other four datasets. Our generated images show a competitive average score compared to the other real datasets. This result indicates that our generated image dataset closely resembles real ones, providing evidence that our generation can produce high-quality real images.

Dataset	Generated (ours)	MIMIC-CXR	CheXpert	COVIDx	ChestX-ray14
Similarity	0.859	0.862	0.847	0.854	0.855

Table 7. Fréchet Inception Distance (FID) Score of the generated dataset regarding different prompt strategies.

Prompt	FID↓
Disease	35.8
Heuristic	35.4
Elaborated	34.7
Elaborated+AdaIN	33.8

real clinical information. Therefore, as shown in Fig. 2, we evaluate the similarity between our generated dataset and other real datasets using the FAISS [17] method. FAISS provides accurate and efficient similarity measurements through indexing and search algorithm optimization between multiple large-scale and high-dimensional data. In this evaluation, it is shown that our generated medical image dataset is highly similar to the real datasets such as the MIMIC-CXR, CheXpert, COIVDx, and ChestX-ray14. As shown in Fig. 2.(a), the similarities between the CheXpert and the other four datasets, including our generated and three real datasets. It demonstrates that although the similarity between the

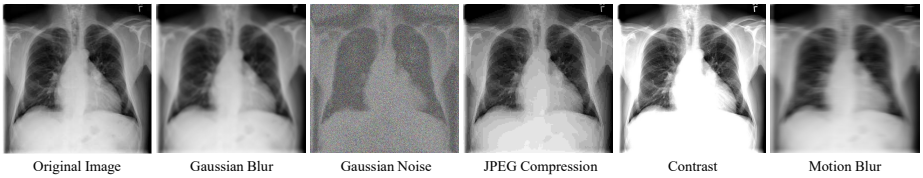


Fig. 3. Examples of noisy images. We add the five types of noises to the original images of the ChestX-ray14.

CheXpert and our generated dataset is lower than that between CheXpert and MIMIC-CXR, it is higher than in other cases, such as the COVIDx and ChestX-ray14. This trend is consistent across the other three comparisons of Fig. 2.(b), (c), and (d). The average similarity score of our generated images compared to the other four datasets is 0.859, as shown in Tab. 6, which is the second highest value, including the four real datasets. Thus, this similarity comparison confirms that our generated medical image dataset closely resembles real medical datasets. To further evaluate the quality of our generated images, we investigate the Fréchet Inception Distance(FID) [23]. FID is another representative indicator that measures the similarity between the generated and real images. A lower value implies more realistic images are generated. As shown in Tab. 7, our method achieves a better FID score compared to other approaches. Disease prompts do not have detailed disease information, and Heuristic prompts may contain information that does not exist in real medical data. On the other hand, our elaborated prompts contain disease information that exists in real medical data, so they can generate realistic medical images compared to other methods. In particular, it is evident that the generated images using AdaIN, a statistical normalization method, have considerable improvement in the FID score. These two quality evaluations using FAISS and FID support the effectiveness of the proposed generative approach on our SSL method for the medical image recognition tasks.

4.5 Analysis

Noisy images. We analyze the robustness of the proposed GEN-SSL against a noisy environment. To evaluate how our GEN-SSL robustly learns generalized features during the pre-training process, we conduct experiments with various noises such as Gaussian Blur, Gaussian Noise, JPEG Compression, Contrast, and Motion blur. Following the previous studies [47,44], we have added such noises as shown in Fig. 3. In these noisy environments, we compare ours with the previous SOTA SSL method of medical image recognition, MED-MAE. As shown in Tab. 8, our GEN-SSL consistently performs favorably in most noisy environments. Specifically, the average performance degradation due to the noise of our GEN-SSL is lower compared to MED-MAE across all three datasets.
Qualitative analysis. In Fig. 4, we conduct a qualitative analysis of the proposed GEN-SSL through the attention heatmap. We generate the attention heatmap

Table 8. Experimental results on the noisy environments. 'Average △↓' in the bottom row indicates the average performance degradation for each dataset. In this experiment, our GEN-SSL is more robust to noises than MED-MAE in all datasets. We use the model parameters of MED-MAE from its public repository: `https://github.com/lambert-x/medical_mae`.

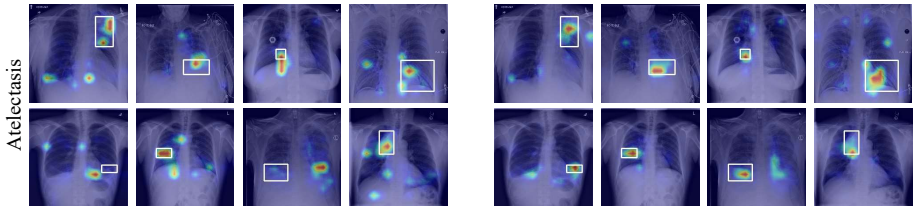
Method	CheXpert		COVIDx		ChestX-ray14	
	MED	GEN.(ours)	MED	GEN.(ours)	MED	GEN.(ours)
Original Image	89.3	90.5	95.3	98.0	83.0	83.2
Gaussian Blur	86.0	87.1	88.5	91.3	78.1	79.2
Gaussian Noise	57.4	59.3	46.7	50.0	57.9	61.4
JPEG Compression	85.7	86.3	87.7	90.5	74.3	74.8
Contrast	85.6	87.2	80.0	83.6	77.4	78.7
Motion blur	85.2	87.0	92.3	94.7	76.3	77.0
Average △↓	9.3	9.1	16.3	16.0	10.2	8.9

Table 9. Experimental comparison between our GEN-SSL and other SOTA SSL approaches.

Method	Architecture	Params (M)	CheXpert	COVIDx	ChestX-ray14
MoCo v2 [9]	ResNet-50 [22]	25.6	88.8	95.5	79.5
SimSiam [10]			88.7	95.8	80.9
GEN-SSL (ours)	ViT-S	22.2	89.1	96.8	81.6

with 14×14 size by computing the class token of the 12th layer of ViT-B following DINO [6]. It is shown that our GEN-SSL attends to the ground-truth bounding box more accurately than MED-MAE. This indicates that our approach learns thorax disease information better than the MAE-SSL method. While the MED-MAE uses real medical images for the SSL method, our generated image-based approach shows a more accurate attention heatmap by comparing it to the ground-truth bounding box. Therefore, these results emphasize the high quality of our generated images compared to the real ones.

Comparison with more SOTA SSL methods. We also compare the proposed MAE-based SSL method with other SSL approaches. While we apply the MAE method to SSL, contrastive learning and Siamese networks are also significant approaches for SSL. Thus, in Tab. 9, we compare our GEN-SSL with MoCo v2 [9] and SimSiam [10], which are the most effective methods of the two SOTA SSL learning approaches. It demonstrates that our generative method with MAE performs better than the two SOTA contrastive learning and Siamese network approaches. This result supports the efficacy of our generative approach in the proposed SSL method.



(a) MED-MAE **(b)** GEN-SSL (ours)

Fig. 4. Attention heatmap visualization. We showcase the attention heatmap of the ChestX-ray14 dataset using MED-MAE and our GEN-SSL. The white boxes denote the ground-truth labels for thorax disease. Compared to MED-MAE, the proposed GEN-SSL highlights a more accurate attention heatmap around the ground-truth bounding boxes. More visualization of this attention heatmap can be found in Fig. S1 of the supplementary material.

5 Conclusion

This paper introduces the GEN-SSL method to address the problem of limited medical data in the real world. The proposed GEN-SSL consists of elaborated prompts for image generation, domain alignment method with statistical normalization, and learning self-supervision with MAE. In this self-supervised learning process, we only use generated images instead of real data, and the results on the downstream datasets show significant performance improvements over baseline and SOTA methods. We also reduce the domain gap between the generated and real medical images using the statistical normalization method. We demonstrate that the generated images contain disease information similar to the real medical images in our analysis. Generated images from our approach achieve considerable similarity scores compared to real medical image datasets. Thus, our generative self-supervised learning approach works favorably in various medical image recognition tasks. We believe that our method will pave the way for future studies in medical image recognition.

Acknowledgments. This paper was supported in part by 'Korea Government Grant Program for Education and Research in Medical AI' through the Korea Health Industry Development Institute(KHIDI), funded by the Korea government(MOE, MOHW), under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2024-RS-2023-00255968), the Electronics and Telecommunications Research Institute (ETRI) Grant funded by Korean Government (Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems) under Grant 24ZB1200, and the National Research Foundation of Korea (NRF) from the Korea Government (MSIT) under Grant RS-2024-00356486.

References

1. Abouelmehdi, K., Beni-Hessane, A., Khaloufi, H.: Big healthcare data: preserving security and privacy. Journal of big data **5**(1), 1–18 (2018)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Allaouzi, I., Ahmed, M.B.: A novel approach for multi-label chest x-ray classification of common thorax diseases. IEEE Access **7**, 64279–64288 (2019)
4. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: IEEE International Conference on Computer Vision. pp. 3478–3488 (2021)
5. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE International Conference on Computer Vision. pp. 9650–9660 (2021)
7. Carrión, H., Norouzi, N.: Fedd-fair, efficient, and diverse diffusion-based lesion segmentation and malignancy classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 270–279. Springer (2023)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
9. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
11. Chen, X., Lian, C., Wang, L., Deng, H., Kuang, T., Fung, S.H., Gateno, J., Shen, D., Xia, J.J., Yap, P.T.: Diverse data augmentation for learning image segmentation with cross-modality annotations. Medical image analysis **71**, 102060 (2021)
12. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M.M., Brown, K.: Masked image modeling advances 3d medical image analysis. In: IEEE Winter Conference on Applications of Computer Vision. pp. 1970–1980 (2023)
13. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical data augmentation with no separate search. arXiv preprint arXiv:1909.13719 **2**(4), 7 (2019)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis **34**, 8780–8794 (2021)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024)

18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets **27** (2014)
19. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D.: A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
20. Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J.: Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 20824–20834 (2022)
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium **30** (2017)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models **33**, 6840–6851 (2020)
25. Hosseinzadeh Taher, M.R., Haghighi, F., Feng, R., Gotway, M.B., Liang, J.: A systematic benchmarking analysis of transfer learning for medical image analysis. In: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3. pp. 3–13. Springer (2021)
26. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
27. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European Conference on Computer Vision. Springer (2016)
28. Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Medicine **6**(1), 74 (2023)
29. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision (2017)
30. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019)
31. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
32. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
33. Kang, M., Li, B., Zhu, Z., Lu, Y., Fishman, E.K., Yuille, A., Zhou, Z.: Label-assemble: Leveraging multiple datasets with partial labels. In: IEEE International Symposium on Biomedical Imaging. pp. 1–5. IEEE (2023)
34. Kim, E., Kim, S., Seo, M., Yoon, S.: Xprotonet: diagnosis in chest radiography with global and local explanations. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 15719–15728 (2021)

35. Kossen, T., Subramaniam, P., Madai, V.I., Hennemuth, A., Hildebrand, K., Hilbert, A., Sobesky, J., Livne, M., Galinovic, I., Khalil, A.A., et al.: Synthesizing anonymized and labeled tof-mra patches for brain vessel segmentation using generative adversarial networks. Computers in biology and medicine **131**, 104254 (2021)
36. Lee, S.H., Li, Y., Ke, J., Yoo, I., Zhang, H., Yu, J., Wang, Q., Deng, F., Entis, G., He, J., et al.: Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. arXiv preprint arXiv:2401.05675 (2024)
37. Li, H., Xue, F.F., Chaitanya, K., Luo, S., Ezhov, I., Wiestler, B., Zhang, J., Menze, B.: Imbalance-aware self-supervised learning for 3d radiomic representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46. Springer (2021)
38. Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., Carneiro, G.: Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2022)
39. Liu, J., Hu, T., Zhang, Y., Gai, X., Feng, Y., Liu, Z.: A chatgpt aided explainable framework for zero-shot medical image diagnosis. arXiv preprint arXiv:2307.01981 (2023)
40. Marrakchi, Y., Makansi, O., Brox, T.: Fighting class imbalance with contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 466–476. Springer (2021)
41. Matsoukas, C., Haslum, J.F., Sorkhei, M., Söderberg, M., Smith, K.: What makes transfer learning work for medical images: Feature reuse & other factors. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9225–9234 (2022)
42. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
43. Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023)
44. Oh, S., Kim, N., Ryu, J.: Analyzing to discover origins of cnns and vit architectures in medical images. Scientific Reports **14**(1), 8755 (2024)
45. Park, I., Kim, W.H., Ryu, J.: Style-kd: Class-imbalanced medical image classification via style knowledge distillation. Biomedical Signal Processing and Control **91**, 105928 (2024)
46. Park, N., Kim, S.: How do vision transformers work? arXiv preprint arXiv:2202.06709 (2022)
47. Park, W., Park, I., Kim, S., Ryu, J.: Robust asymmetric loss for multi-label long-tailed learning. In: IEEE International Conference on Computer Vision (2023)
48. Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N.: Xplainer: From x-ray observations to explainable zero-shot diagnosis. arXiv preprint arXiv:2303.13391 (2023)
49. Peng, L., Cai, S., Wu, Z., Shang, H., Zhu, X., Li, X.: Mmgpl: Multimodal medical data analysis with graph prompt learning. Medical Image Analysis p. 103225 (2024)
50. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
51. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 117–126. Springer (2022)
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
54. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest x-ray classifiers. In: BIOCOMPUTING 2021: proceedings of the Pacific symposium. pp. 232–243. World Scientific (2020)
55. Shrivastava, A., Fletcher, P.T.: Nasdm: Nuclei-aware semantic histopathology image generation using diffusion models. arXiv preprint arXiv:2303.11477 (2023)
56. Shurrab, S., Duwairi, R.: Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Computer Science **8**, e1045 (2022)
57. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation **28**(1), 11–21 (1972)
58. Sriram, A., Muckley, M., Sinha, K., Shamout, F., Pineau, J., Geras, K.J., Azour, L., Aphinyanaphongs, Y., Yakubova, N., Moore, W.: Covid-19 prognosis via self-supervised representation learning and multi-image prediction. arXiv preprint arXiv:2101.04909 (2021)
59. Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M.: Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In: Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9. pp. 249–258. Springer (2018)
60. Taslimi, S., Taslimi, S., Fathi, N., Salehi, M., Rohban, M.H.: Swinchex: Multi-label classification on chest x-ray images with transformers. arXiv preprint arXiv:2206.04246 (2022)
61. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific reports **10**(1), 19549 (2020)
62. Wang, W.C., Ahn, E., Feng, D., Kim, J.: A review of predictive and contrastive self-supervised learning for medical images. arXiv preprint arXiv:2302.05043 (2023)
63. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017)
64. Xiao, J., Bai, Y., Yuille, A., Zhou, Z.: Delving into masked autoencoders for multi-label thorax disease classification. In: IEEE Winter Conference on Applications of Computer Vision. pp. 3588–3600 (2023)
65. Yan, A., Wang, Y., Zhong, Y., He, Z., Karypis, P., Wang, Z., Dong, C., Gentili, A., Hsu, C.N., Shang, J., et al.: Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 (2023)
66. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. arXiv preprint arXiv:2401.11708 (2024)
67. Yuan, Z., Yan, Y., Sonka, M., Yang, T.: Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In: IEEE International Conference on Computer Vision. pp. 3040–3049 (2021)
68. Zhu, H., Chen, B., Yang, C.: Understanding why vit trains badly on small datasets: An intuitive perspective. arXiv preprint arXiv:2302.03751 (2023)