Pattern Recognition 156 (2024) 110745

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Global contrast-masked autoencoders are powerful pathological representation learners

Hao Quan [a,b,c,1], Xingyu Li [a,1], Weixing Chen [d], Qun Bai [a], Mingchen Zou [a], Ruijie Yang [e], Tingting Zheng [a], Ruiqun Qi [f], Xinghua Gao [f], Xiaoyu Cui [a,c,*]

[a] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
[b] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[c] The Key Laboratory of Biomedical Imaging Science and System, Chinese Academy of Sciences, Shenzhen, China
[d] The School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[e] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[f] Department of Dermatology, The First Hospital of China Medical University, Shenyang, China

ARTICLE INFO

Keywords:
Self-supervised learning
Representation learning
Pathological image

ABSTRACT

Using digital pathology slide scanning technology, artificial intelligence algorithms, particularly deep learning, have achieved significant results in the field of computational pathology. Compared to other medical images, pathology images are more difficult to annotate, and thus, there is an extreme lack of available datasets for conducting supervised learning to train robust deep learning models. In this paper, we introduce a self-supervised learning (SSL) model, the Global Contrast-masked Autoencoder (GCMAE), designed to train encoders to capture both local and global features of pathological images and significantly enhance the performance of transfer learning across datasets. Our study demonstrates the capability of the GCMAE to learn transferable representations through extensive experiments on three distinct disease-specific hematoxylin and eosin (H&E)-stained pathology datasets: Camelyon16, NCT-CRC, and BreakHis. Moreover, we propose an effective automated pathology diagnosis process based on the GCMAE for clinical applications. The source code of this paper is publicly available at https://github.com/StarUniversus/gcmae.

1. Introduction

Pathology is considered the gold standard of diagnosis. Traditionally, pathology relies heavily on macroscopic observation through microscopes. Pathologists assess the nature of lesions and classify tissues based on their subjective experience, making the diagnostic outcomes susceptible to variability due to factors such as the observer's experience and fatigue [1-3]. Computational pathology, by converting glass slides into digital images and applying image processing technologies, facilitates a shift from qualitative analysis to quantitative evaluation in pathological diagnosis. In recent years, empowered by whole-slide image (WSI) scanning technology [4], artificial intelligence algorithms, especially those involving deep learning, have made significant strides in the field of computational pathology [1,5,6]. This has not only marked a current research trend but also outlined a pivotal direction for the future development of pathological methodologies.

Deep learning (DL) is one of the common methods used to extract computational pathology features, as it can directly learn subvisual image features that are difficult for humans to find with their eyes [7]. However, most DL methods require a large amount of high-quality labeled data, making them difficult to transfer to other datasets with different feature spaces or probability distribution functions [8]. Different staining methods, scanning equipment variations, different diseases and intraclass differences across organs and tissues lead to data feature differences and long-tailed problems, especially in the field of computational pathology [7]. Maximizing the use of source domain datasets for representation learning has become an important method for alleviating the poor model performance caused by data scarcity in the target domain [8].

Recent advances in self-supervised visual representation learning have led to significant progress in the field of natural images [9-11]. Self-supervised learning (SSL) employs pretext tasks to extract valuable

* Corresponding author.
E-mail addresses: h.quan@siat.ac.cn (H. Quan), cuixy@bmie.neu.edu.cn (X. Cui).
[1] Co-first author

https://doi.org/10.1016/j.patcog.2024.110745
Received 13 August 2022; Received in revised form 12 April 2024; Accepted 28 June 2024
Available online 4 July 2024

representation information from a large volume of unsupervised data. In the last few years, contrastive learning-based SSL methods such as SimCLR [10] and MoCo [9] have been transferred to computational pathology for use in downstream tasks. This transfer has significantly narrowed the performance gap between unsupervised and supervised learning methods [12]. Furthermore, specific data augmentation strategies and complex pretext tasks have been developed to better capture the unique representation spaces of pathological images. For example, Yang et al. [13] developed a cross-stain prediction and a new data augmentation technique, stain vector perturbation, tailored to the characteristics of pathological images, and proposed the CS-CO method using contrastive learning, which was proven effective on the NCTCRC datasets. Similarly, Li et al. [14] created the SSLP method, which explores pathological features from three perspectives—self-invariance, intra-invariance, and inter-invariance—by designing complex pretext tasks; this method exceeded the performance of supervised methods on the Camelyon16 dataset. However, the abovementioned self-supervised method based on contrastive learning has the problems of high hardware resource consumption, high training difficulty in multitask learning scenarios, and lower cross-dataset transfer learning performance than supervised learning [11,15]. Therefore, simplifying pretext tasks and improving the overall representational capability of models are essential challenges in pathological representation learning.

In 2021, as an extensible SSL method, a masked autoencoder (MAE) achieved state-of-the-art (SOTA) results on the ImageNet dataset [11]. This method randomly masks part of the input image and employs a lightweight decoder to rebuild the obscured pixels, which can not only yield improved accuracy but also speed up the training process. Additionally, it demonstrates better learning efficiency than contrastive learning [11]. Pathological diagnosis often requires the consideration of both the global and local features of WSIs [1]. Due to the morphological similarity between cells and tissues of the same type, MAEs may have the potential to discover correlations within pathological image tiles, that is, to extract local features. Correspondingly, if we use the memory bank structure [16] of contrast learning to store the features between each pair of tiles, MAEs may also be capable of capturing global features.

Based on the above analysis, we propose a global contrast-masked autoencoder (GCMAE)-based SSL model that can extract both global and local features from pathological images. On the one hand, based on the MAE network structure, the model can obtain the internal hidden space feature representation of each patch in pathological images. On the other hand, the model integrates the memory bank structure to store the global features of pathological images, and contrastive learning is used to mine the feature associations between tiles. Second, we also design an automatic pathological image diagnosis process based on the GCMAE for clinical application, which can make full use of unlabeled pathological data to further improve the performance of the model. Finally, we also attempt to utilize a lightweight modeling method to increase the confidence of GCMAE in clinical application. The main contributions of this study are as follows.

1. We have proposed GCMAE, which integrates two self-supervised pretext tasks, masking image reconstruction and contrast learning, to produce effective supervision. These tasks also train the encoder to represent local-global features of pathological images.
2. We analyzed the mask ratio suitable for pathological images, and provided guidance for pathology-specific training methods related to the masked image modeling (MIM) paradigm.
3. We selected three pathological image benchmark data sets, and proved that GCMAE has a tangible improvement over other state-of-the-art self-supervised and transfer learning methods through extensive experiments.
4. In this paper, an automatic diagnosis process and a lightweight modeling method for pathological images based on the GCMAE are designed for clinical application purposes.

2. Related works

Pathological image analysis constitutes a critical branch within the field of medical image processing, playing an indispensable role in disease diagnosis, treatment strategy formulation, and prognostic evaluation. In recent years, the rapid development of deep learning technologies has introduced new methodologies for automatic disease feature extraction, lesion detection, and classification from pathological images, leading to significant advancements in the field of pathological image analysis [17,18]. These advanced deep learning methods are capable of autonomously identifying patterns and features of clinical significance within complex pathological images, such as whole slide image classification [19] and prognostic indicator evaluation [20], and can further enhance the decision-making performance of models by constructing graph representations of spatial correlations between tissue components [21]. However, the heterogeneity of data resulting from multi-center sources presents a major challenge in pathological image analysis, compelling researchers to explore strategies like federated learning [22] and representation learning [23] to improve model generalization capabilities, with preliminary successes achieved in these fields.

In addressing the challenge of data heterogeneity, self-supervised learning has demonstrated tremendous potential. This paradigm, by leveraging the input data itself as a supervisory signal, has been proven beneficial across a variety of downstream tasks in the realm of representation learning [24]. Consequently, through self-supervised learning, a vast array of unlabeled pathological image samples can be utilized to learn a universal data representation, thereby assisting downstream tasks with limited labeled samples. Currently, contrastive learning and masked image modeling, as the predominant methods of self-supervised learning, have both shown effectiveness in the representation task of pathological images, despite each having its limitations [25].

Contrastive learning, a self-supervised learning paradigm extensively utilized in the domain of pathological image representation, has been validated as effective by multiple studies [26]. Despite some success achieved by directly applying self-supervised algorithms designed for natural image representation to the field of medical image analysis [12,27], the lack of targeted design has limited performance improvements. Effective data augmentation methods are key to realizing the superior representational capabilities of contrastive learning [28]. Accordingly, researchers have devised data augmentation strategies tailored to the characteristics of pathological images, such as elastic deformation [29] and stain vector perturbation [13]. Additionally, specially designed pretext tasks have been shown to enhance the performance of contrastive learning in pathological image representation tasks [14]. Methods like SDSCL [30] employ uniquely crafted self-distillation strategies to improve the feature representation encoded by contrastive learning. SSLP [14] and RSP [31], respectively, explore from the perspectives of intrinsic characteristics of pathological images and the multi-resolution contextual information inherent in their pyramidal nature, achieving commendable results. However, contrastive learning still faces shortcomings in the realm of pathological image representation, primarily due to the homogeneity in the visual appearance of pathological images which constrains its representational capabilities, and the common practice of random cropping as a data augmentation method, which only inputs the main part of the sample into the encoder, thus limiting the universality of the learned representations.

On the other hand, MAE and their derivatives within the masked image modeling (MIM) paradigm offer innovative solutions to the limitations faced by contrastive learning. To date, methods within the MIM paradigm have been applied to representation learning tasks for natural images and videos [32,33], but only a handful of studies have ventured into the medical imaging domain [34]. Among these, a notable approach for pathological images is the SDMAE [35], which, in addition to preserving the original task of image reconstruction, incorporates a

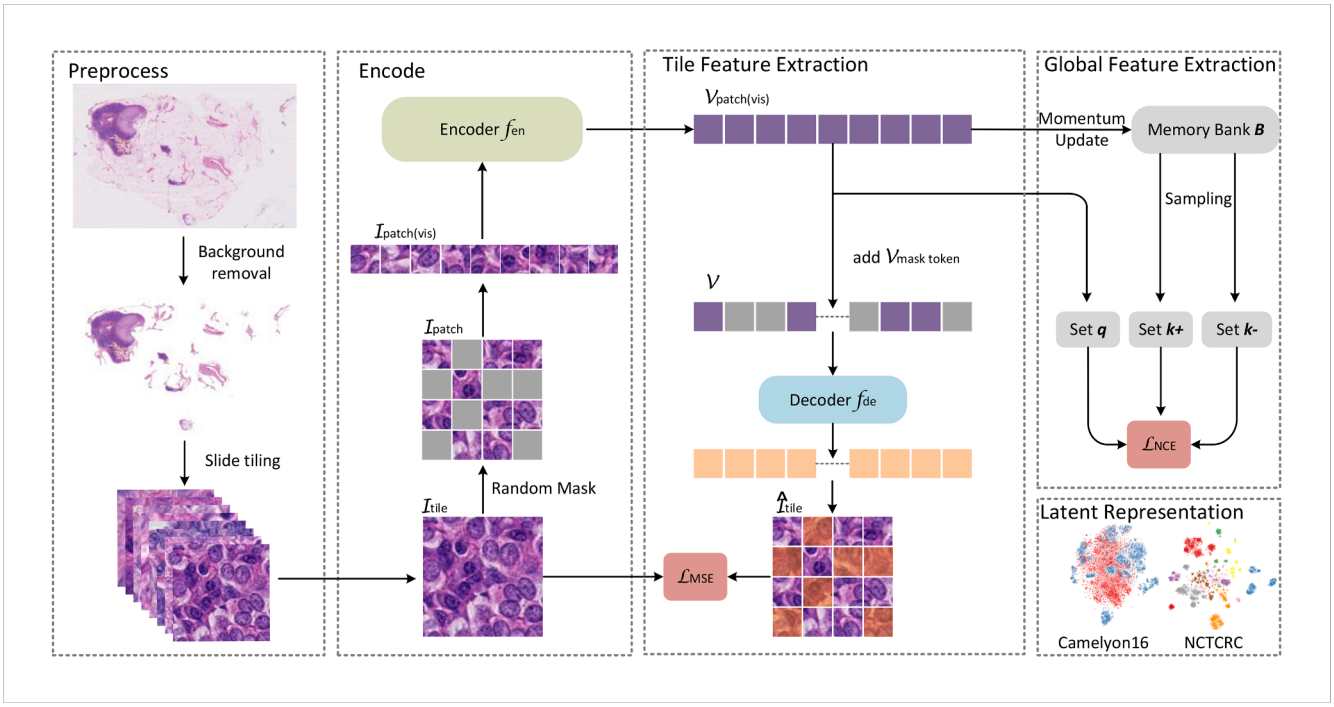

Fig. 1. Framework of the GCMAE. Tile feature extraction is the pretext task of image reconstruction, while global feature extraction is the pretext task of contrastive learning. The latent representation is the t-distributed stochastic neighbor embedding (t-SNE) result of the encoder output.

self-distillation module for the visible parts of the image to enhance the learning of high-level semantic information. However, the focus of MAE models on learning representations within individual samples, without fully exploiting the relationships between samples, has limited their applicability in pathological image representation tasks. Therefore, a self-supervised model that integrates the strengths of both self-supervised learning paradigms, tailored to the unique characteristics of pathological image representation tasks, will offer a new avenue to enhance model generalization capabilities and address data heterogeneity issues.

3. Methodology

In this section, we describe the GCMAE algorithm in detail. Fig. 1 shows the framework diagram of our proposed GCMAE-based SSL algorithm. In summary, the GCMAE consists of four parts, a preprocessor, an encoder, a tile feature extractor and a global feature extractor, as well as two pretext tasks: image reconstruction and contrast learning. The GCMAE inherits and optimizes these methods for pathological images. As shown in Eq. (1), the weighted sum of the mean squared error (MSE) loss of tile feature extraction and the noise contrastive estimation (NCE) loss of global feature extraction is used as the cost function to reduce the distance between similar features while learning high-level image features, thus improving the generalization of the model and the accuracy achieved in the cross-dataset transfer learning task.

$$L = \lambda_1 L_{MSE} + \lambda_2 L_{NCE} \tag{1}$$

3.1. Preprocessing

The large amount of redundant information contained in WSIs, such as non-tissue background regions, can reduce the training performance of the model. Therefore, it is necessary to perform preprocessing operations on WSIs. First, the optimal segmentation threshold of the given WSIs is calculated based on the Otsu threshold segmentation algorithm, and the tissue regions are extracted. Finally, the mean and standard deviation of the tiles are calculated to achieve the normalization operation, and a normalized image with a mean of 0 and a standard deviation of 1 is output to accelerate the convergence of the model.

3.2. Encoder

The vision transformer (ViT) [36] is regarded as the encoder backbone f_{en}. Compared with the classic convolutional neural network (CNN), the ViT model without an inductive bias has a very high capacity and a good generalization ability; it can also learn more abundant pathological representations and transfer them to downstream tasks. Because the ViT is a large model, we need to consider an efficient pretraining method to train the visual representation ability of the ViT. An MAE randomly masks some details of the input image with a high mask ratio (MR%) and reconstructs missing pixels only from the visible part of the feature space, enabling it to achieve excellent performance in natural image representation tasks. This study attempts to extend a simple and efficient MAE to pathological image representation tasks. Specifically, first, 224 × 224 tiles $I_{tile} \in R^{H\times W\times C}$ are divided into regular nonoverlapping patches (16 × 16) $I \in R^{N\times(p^2\times C)}$; a 2D patch sequence is output, where $H \times W$ is the size of the original image, C is the number of channels in the image, p^2 is the resolution of each patch, and $N = \frac{HW}{p^2}$ is the number of patches cut from the original image. This setting has two main purposes. 1. It is convenient for randomly masking some images. 2. When a 2D image sequence is processed into a 1D image sequence, the sequence length can be reduced. Then, the patches are randomly sampled by a uniform distribution to mask some patches, and the visible parts form a new subset of patches $I_{patch(vis)} \in R^{V\times(p^2\times C)}$. $V = N \times$ (1-MR %) is the number of visible partial patches and the length of the sequence of valid inputs for the transformer block.

The image features of the visible parts are embedded by linear projection, and the position information is encoded by positional embeddings. Specifically, the 2D image of the visible part $I_{patch(vis)} \in R^{V\times(p^2\times C)}$ is flattened and mapped to the Dth dimension through a trainable linear projection, and the output vector is a patch embedding. Then, standard learnable 1D vectors are added to the patch embedding to preserve the position information. The embedded features and position information

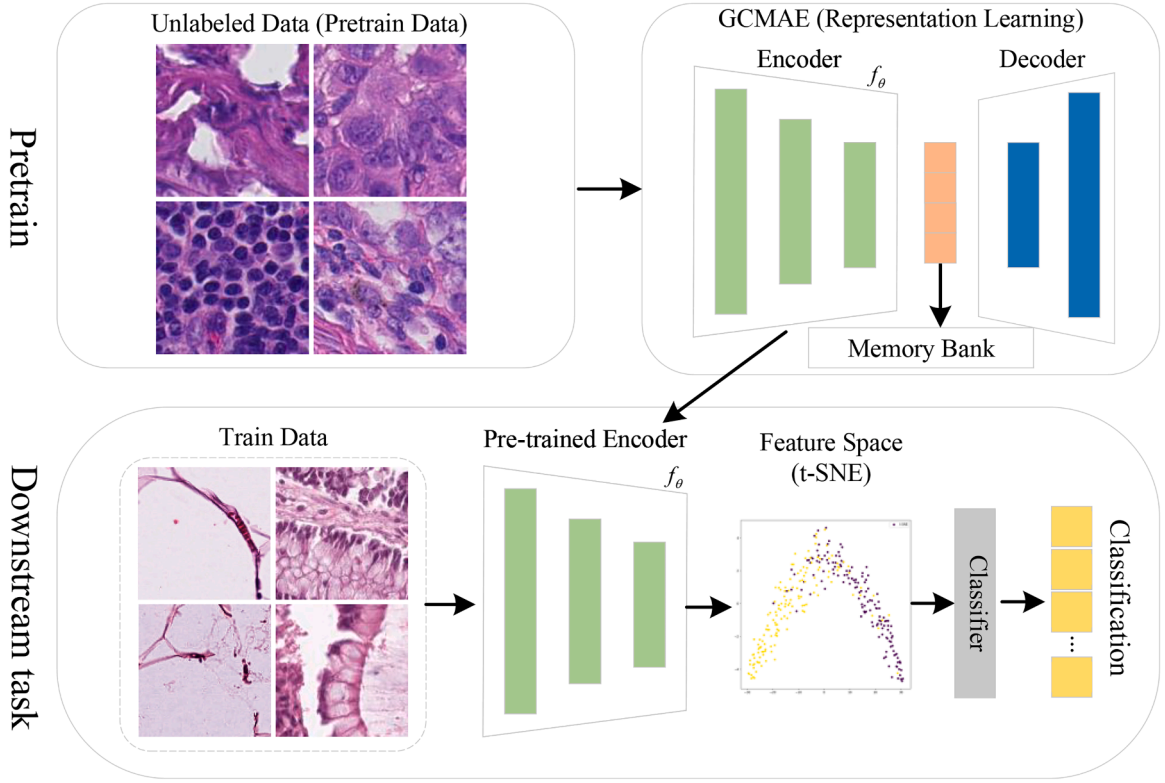

Fig. 2. Automatic pathological diagnosis process based on the GCMAE.

are fed into the transformer block to extract a latent representation of the visible parts of the tile.

3.3. Tile feature extraction

As a decoder f_{de}, the tile feature extraction module consists of eight transformer blocks, which form an asymmetric structure with an encoder possessing at least 12 transformer blocks (ViT-base). The asymmetric encoder-decoder structure shows that the encoder and decoder are decoupled, which is beneficial for the encoder to learn more generalized representations. In this study, the decoder mainly assists the encoder in learning general representations. However, while training the encoder, the reconstruction ability of the decoder is also optimized. If a symmetrical structure design is adopted and the encoder and decoder are coupled, even if the encoder's representation ability is insufficient, a powerful decoder can minimize the loss by optimizing the reconstruction ability, thus limiting the feature expression of the encoder. The key to self-supervision lies in pretraining the encoder to attain a strong representation ability and transferring it to downstream tasks. Therefore, it is necessary to adopt an asymmetric encoder-decoder structure, which has also been proven in the field of natural images [11]. At the same time, the lightweight decoder design reduces the memory consumption and further expands the application range of the algorithm in clinical environments. The experimental MAE results verify that a decoder with 8 transformer blocks can effectively assist the encoder in learning general representations, and this study follows this setting.

In addition to the latent representation $\mathscr{V}_{patch(vis)}$, the input also includes a shared, learned vector mask token $\mathscr{V}_{mask-token}$, which is used to indicate the missing patches. The mask token also contains the position embeddings of all patches, which are used to reconstruct the missing pixels. The normalized tiles are used as the target to calculate the MSE loss, as shown in Eq. (4).

$$V = f_{en}\left(I_{patch(vis)}\right) + V_{mask-token} \tag{2}$$

$$I'_{tile} = f_{de}(V) \tag{3}$$

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(I_{tile} - I'_{tile}\right)^2 \tag{4}$$

The random tile sampling strategy can remove redundant information and realize the difficult pretext process of directly reconstructing the original missing pixels from their adjacent patches. However, the strategy can only represent the internal features of each tile and cannot represent the feature relationships between different tiles.

3.4. Global feature extraction

Global feature extraction is implemented through contrastive learning [16]. The latent representation $V_{patch(vis)}$ is not only inputted to the decoder for image reconstruction but also updated to a memory bank B with a momentum coefficient of t for storing global features. The memory bank is a fixed-length, dynamically updated queue for storing feature embeddings of data samples. In the contrastive learning task, some features can be randomly selected from the memory bank and embedded as negative samples, which can alleviate the limitation of batch size on contrastive learning performance. We design the momentum update feature differently from the momentum update model parameters in MoCo, using only a separate encoder and a memory bank to enable contrastive learning. Specifically, it is known that the latent representation at the output of the encoder in the current epoch is $V_{patch(vis)}$, and the latent representation deposited in B from the previous epoch is $V'_{patch(vis)B}$. The latent representation $V_{patch(vis)B}$ deposited in B during the current epoch is

$$V_{patch(vis)B} = \frac{0.5V_{patch(vis)} + 0.5V'_{patch(vis)B}}{\| 0.5V_{patch(vis)} + 0.5V'_{patch(vis)B}\|^2} \tag{5}$$

The main reason for this design of $V_{patch(vis)B}$ is twofold: 1. It can

alleviate the differences among the features of the same sample in different epochs caused by the use of different model parameters and random masks during the training process of the network. 2. Combining features from adjacent epochs enables the construction of pathological feature representations with higher information density, thus improving the generic feature representation capability of the GCMAE.

We construct $V_{patch(vis)B}$ as a dictionary to be stored as a data sample queue in B. The $V_{patch(vis)B}$ corresponding to the input minibatch of the current epoch is extracted from B as the key value $k+$. The current minibatch of $V_{patch(vis)}$ (as query q) forms a positive pair with k. The potential features of the n samples drawn randomly from the memory bank form a negative pair with q. Cosine similarity is used as a means of evaluating the distances between features to calculate the similarity between the $V_{patch(vis)}$ values. We consider an efficient form of the contrast learning loss function, called infoNCE, to minimize the distances between positive pairs and maximize the distances between negative pairs; the loss function is defined as follows.

$$L_{q,k^+,B} = -log\frac{exp(q^T k^+/\tau)}{exp(q^T k^+/\tau) + \sum_{k^- \in B} exp(q^T k^-/\tau)} \quad (6)$$

τ is a temperature parameter that controls the concentration level of the distribution [37]. Because the memory bank dynamically stores a large number of data sample features, the constant comparison and discrimination between q and the memory bank features can help the encoder to effectively mine rich semantic information between tiles and then help the encoder learn global features.

3.5. Workflow of the proposed GCMAE

Based on the GCMAE, we propose a reasonable pathological diagnosis process for images with hematoxylin and eosin (HE) staining. The overall flow chart is shown in Fig. 2, which is mainly divided into three parts: pretraining data collection, the GCMAE and downstream tasks. The pretraining data part mainly prepares a large number of unlabeled pathological image datasets, which are easier to obtain than labeled data. At the same time, the pretraining dataset and the target dataset can be different datasets, which further expands the data sources and reduces the difficulty of data collection. The GCMAE is the key to the whole process, and the encoder is pretrained in a self-supervised way to perform representation learning for pathological images. The good cross-dataset representation ability of the GCMAE-based self-supervised algorithm improves the classification performance and expansion ability attained in downstream tasks. The downstream task is mainly to fine-tune the pretrained encoder to adapt to the target task by using the target domain dataset. According to different downstream task objectives, the pretraining encoder can be fine-tuned in three ways: 1. taking the pretraining parameters as the initialization parameters of the downstream task model, the best performance can be achieved in the target task by training the model from scratch; 2. the feature extraction part is frozen, only the linear probing method of the classifier is fine-tuned, and the model size and training time are greatly reduced; 3. the model storage weight is reduced from 32 bits to 16 bits by quantization, and the model size is reduced as much as possible while ensuring maximal performance.

4. Experimental results and analysis

4.1. Dataset and data settings

In this paper, three pathological datasets, Camelyon16 [38], NCTCRC [39] and BreakHis [40], were collected to fully evaluate the general visual pathological image representation ability of the GCMAE-based self-supervised algorithm. The Camelyon16 data set contains two types of breast cancer WSIs. We randomly cut 270k non-overlapping images with the size of 224 × 224 at 40 magnification.

Table 1
Data settings for conducting pretraining, training and testing on the three pathological datasets.

Dataset	Pretrain	Downstream task		Overall
		Train	Test	
Camelyon16	220,000	40,000	10,000	270,000
NCTCRC	100,000*	100,000*	7180	107,180
BreakHis	N/A	1274	546	1820

Note: annotation * is the same data.

The NCTCRC data set consisted of nine categories of 100k pathological image patches with a size of 224 × 224, which were scanned at a spatial resolution of 0.5 μm/pixel. BreakHis data set contains 7909 breast tumor pathological images with a size of 700 × 460 in eight categories. This study mainly uses 400 magnification images in this data set, with a total of 1820 images. Among them, the Camelyon16 and NCTCRC datasets were used for the pretraining and downstream tasks of SSL, and BreakHis was only used as an extended experimental dataset for the downstream tasks. The data setting are shown in Table 1, and the specific data details are shown in the supplementary file part A.

In this study, all pathological images used for model development were resized to 224 × 224 pixels via bilinear interpolation and normalized using the calculated means and standard deviations. We refrained from employing conventional data augmentation and stain normalization techniques in the preprocessing of pathological images, with the intent to evaluate the representation learning capabilities and robustness of the SSL algorithms under more stringent training conditions. Additionally, the specific impact of stain normalization on model performance was assessed independently.

4.2. Hyperparameter settings and evaluation criteria

The hyperparameters of the GCMAE method were set as follows: τ = 0.07, t = 0.5, k^-= 8192, batch_size=128, epochs = 80, and the other hyperparameters were consistent with those of the MAE. With regard to the loss function, we confirmed that $\lambda_1 = 1$ and $\lambda_2 = 0.1$ through pre-experiments to ensure that the losses of MSE and NCE were on the same order of magnitude and that the performance of the GCMAE could be optimized. In order to avoid over-fitting of ViT on relatively small datasets, a reasonable epoch=80 was determined by pre-experiment. The details of the experiment are shown in the supplementary material part B. The optimizer was AdamW with betas= (0.9, 0.95). We set up two classic models as the backbones: ResNet 50 [41] and ViT-base (ViT-B/16). ResNet 50 is a classic CNN and the backbone of contrastive learning paradigm. ViT-B/16 is a high-capacity model with a stronger generalization ability. It is beneficial to build a general representation model in the pathological field when the available pathological data are sufficient, and this model is also the backbone network of the MAE and GCMAE. See part C in the supplementary file for details regarding the hyperparameter settings utilized for the downstream tasks and comparison methods. All self-supervised pretraining experiments did not involve any labels, and all experiments were conducted on an RTX A6000 GPU.

Linear probing and end-to-end fine-tuning are common evaluation methods for self-supervised models. Specifically, linear probing freezes the backbone parameters and trains classifiers in a supervised way. This task focuses on the feature extraction ability of the tested pretraining model and is widely used to evaluate the representation performance of self-supervised models. The end-to-end fine-tuning task involves training a model from scratch on the target task, and the pretraining model is equivalent to the parameter initialization method of the model fine-tuning task. This task is a classic downstream task for a self-supervised model. In practical self-supervised applications, end-to-end fine-tuning can better optimize the target task. This study mainly reports accuracy and Area Under Curve (AUC) to evaluate the model

Table 2
Influences of different mask ratios on pathological representation with SSL and the MIM paradigm (mean±std%).

Mask ratio	Linear probing		Fine-tunning	
	Accuracy	AUC	Accuracy	AUC
10 %	80.87±0.64	89.92±0.84	83.41±0.45	92.53±0.61
20 %	81.79±0.91	90.14±0.86	83.11±0.79	92.07±0.63
30 %	82.11±0.88	90.75±0.74	83.01±0.44	91.96±0.31
40 %	82.41±0.76	90.78±0.69	83.24±0.38	92.39±0.64
50 %	**85.21±0.58**	**93.18±0.47**	82.43±0.22	92.02±0.33
60 %	81.88±0.69	90.02±0.74	83.22±0.59	92.18±0.46
70 %	82.25±0.79	90.85±0.81	83.02±0.37	92.33±0.29
75 %	81.79±0.84	89.82±0.76	84.01±0.53	92.82±0.49
80 %	80.52±0.76	88.07±0.59	**84.97±0.31**	**93.52±0.29**
90 %	80.94±0.73	87.14±0.53	83.79±0.49	92.44±0.36

performance. The mean and standard deviation are obtained by running Monte Carlo cross-validation ten times.

4.3. Mask ratio

For SSL with the MIM paradigm, images with different information densities are suitable for different mask ratios. Therefore, this study discusses the suitable mask ratio for pathological image representation. Table 2 shows the influence of the mask ratio on the pathological representation of MIM-based SSL, represented by the MAE. In the pathological representation application, a 50 % mask ratio is suitable for linear probing, and an 80 % mask ratio is suitable for fine-tuning, which contrasts with the optimal mask ratio of 75 % in natural image applications. Pathological images contain abundant tissue features, and their information density is higher than that of natural images. Therefore, when using the MAE model to reconstruct pathological images, more information needs to be known. However, the optimal result of model fine-tuning is achieved at a higher mask ratio of 80 %, which shows that the suitable mask ratio for linear probing in the pathological image field is not suitable for model fine-tuning, further proving the result in [11]. Consistent with the natural image results, the influences of pretraining models with different mask ratios on model fine-tuning are lower than that of linear probing, and the accuracy of model fine-tuning is better than that of the ViT-B/16 trained from scratch (81.9 %).

4.4. Cross-dataset and cross-disease transfer task

The scarcity of abundant high-quality labeled data is one of the key factors limiting the performance of deep learning methods. A common strategy employed is the use of readily available source domain data to pre-train models, enhancing performance on target domain data through feature transfer. Hence, cross-dataset transfer tasks serve as an effective benchmark for assessing the practical performance of SSL algorithms.

This experiment was designed as a challenging cross-dataset and cross-disease transfer task. Specifically, the source and target domain datasets were entirely distinct, with no overlap (i.e., cross-dataset transfer), and involved different diseases (i.e., cross-disease transfer). Within this context, we evaluated four categories of methods. As a baseline for transfer learning, we utilized the ResNet50 and ViT-B/16 models, employing parameters from random initialization, ImageNet pre-training, and pathology-specific pre-training derived from supervised classification tasks for both linear probing and fine-tuning tasks. Additionally, we compared three classical methods from the contrastive learning paradigm: SimCLR [10], MoCo v1 [9], and MoCo v2. The third category comprised two SSL algorithms specifically designed for pathology images: TransPath [42] and CS-CO [13]. Lastly, we included the representative method MAE from the MIM paradigm. The backbone for the contrastive learning and pathology-specific SSL algorithms was ResNet 50, while for MAE and GCMAE, it was ViT-B/16.

Table 3
Performance comparison among different SSL models when transferring from Camelyon16 to the NCTCRC task (mean±std%).

Methods	Linear probing		Fine-tuning	
	Accuracy	AUC	Accuracy	AUC
Randomly initailized ResNet 50	50.12 ±1.52	63.45 ±1.56	86.75 ±1.34	94.56 ±1.23
ImageNet pre-trained ResNet50	62.32 ±0.85	75.23 ±0.98	88.12 ±0.89*	98.39 ±0.85
Camelyon16 pre-trained ResNet50	78.69 ±0.79	88.65 ±0.95	89.12 ±0.96	98.42 ±0.85
Randomly initailized ViT-B/16	43.59 ±2.32	68.89 ±2.14	76.58 ±2.15	84.56 ±2.06
ImageNet pre-trained ViT-B/16	52.68 ±1.39	88.95 ±1.62	81.46 ±0.92*	97.35 ±0.87
Camelyon16 pre-trained ViT-B/16	73.46 ±0.85	82.23 ±0.75	82.01 ±0.78	96.85 ±0.84
SimCLR	80.95 ±0.87	97.73 ±0.79	90.67 ±0.52	98.99 ±0.68
MoCo v1	78.40 ±0.95	92.52 ±0.98	89.54 ±0.45	98.57 ±0.42
MoCo v2	81.75 ±0.74	93.26 ±0.92	91.29 ±0.85	99.02 ±0.72
TransPath	82.41 ±0.79	94.62 ±0.65	91.89 ±0.46	99.05 ±0.41
CS-CO	85.58 ±0.54	98.33 ±0.41	92.01 ±0.33	98.58 ±0.31
MAE	85.25 ±0.43	98.28 ±0.74	93.43 ±0.47	99.12 ±0.41
GCMAE	**89.22 ±0.32**	**98.74 ±0.15**	**93.89 ±0.25**	**99.46 ±0.19**

Note: The result of marking * is also the baseline result of fully-supervised learning.

4.4.1. Transferring from Camelyon16 to NCTCRC

In this study, we utilized the Camelyon16 dataset for model pretraining before transferring to the NCTCRC to perform a nine-class classification task, with detailed results presented in Table 3. The experimental outcomes demonstrate that GCMAE achieved the highest

Table 4
Performance comparison among different SSL methods when transferring from NCTCRC to Camelyon16 (mean±std%).

Methods	Linear probing		Fine-tuning	
	Accuracy	AUC	Accuracy	AUC
Randomly initailized ResNet 50	68.72 ±2.32	79.42 ±2.15	80.41 ±1.75	87.62 ±1.31
ImageNet pre-trained ResNet50	72.96 ±0.79	83.41 ±0.76	82.15 ±0.49*	89.68 ±0.62
NCTCRC pre-trained ResNet50	77.28 ±0.91	85.51 ±0.76	81.12 ±0.45	89.53 ±0.79
Randomly initailized ViT-B/16	63.51 ±2.42	69.75 ±2.74	78.47 ±2.21	86.22 ±2.04
ImageNet pre-trained ViT-B/16	69.41 ±1.25	78.27 ±1.43	81.13 ±0.49*	88.06 ±0.52
NCTCRC pre-trained ViT-B/16	75.26 ±0.85	84.02 ±0.49	80.25 ±0.35	87.91 ±0.51
SimCLR	79.29 ±0.83	89.89 ±0.78	80.25 ±0.85	90.69 ±0.89
MoCo v1	77.82 ±0.56	85.43 ±0.64	80.93 ±0.45	88.75 ±0.37
MoCo v2	80.21 ±0.45	89.73 ±0.75	81.76 ±0.43	90.87 ±0.35
TransPath	79.85 ±0.47	88.75 ±0.67	82.23 ±0.51	91.45 ±0.81
CS-CO	79.74 ±0.28	88.43 ±0.32	82.47 ±0.21	91.77 ±0.29
MAE	80.72 ±0.37	89.04 ±0.2	83.32 ±0.15	**92.85 ±0.18**
GCMAE	**81.56 ±0.23**	**90.52 ±0.32**	**83.92 ±0.24**	92.69 ±0.16

Note: The result of marking * is also the baseline result of fully-supervised learning.

accuracy and AUC values in both linear probing and fine-tuning tasks. Specifically, in the linear probing task, GCMAE's performance was significantly superior to the comparative methods. To illustrate, compared to Camelyon16 pre-trained ResNet50, MoCo v2, CS-CO, and MAE, GCMAE's accuracy improved by 10.53 %, 7.47 %, 3.64 %, and 3.97 %, respectively. The ViT-B/16 model pre-trained with GCMAE achieved an accuracy of 89.22 % through mere fine-tuning of the classifier (linear probing), which is a 7.76 % increase over the supervised training baseline of ViT-B/16 (accuracy of 81.46 %). This indicates that GCMAE can effectively train models to learn pathology representations with significant generalizability. Additionally, the strategy of merely fine-tuning the classifier is advantageous due to its high training efficiency and low computational resource consumption, enabling the model to rapidly adapt to various downstream tasks, thus confirming the excellent scalability of GCMAE. Despite maintaining the best performance in fine-tuning tasks, GCMAE did not show a marked improvement in performance compared to other SSL algorithms, largely because fine-tuning performance is greatly influenced by the training strategy on the target data, which diminishes the distinctive performance of SSL algorithms.

4.4.2. Transferring from NCTCRC to camelyon16

In this experiment, the NCTCRC dataset was used to pre-train models which were then transferred to the Camelyon16 dataset to assess performance through a binary classification task; the specific results are presented in Table 4. In the linear probing task, GCMAE achieved an accuracy improvement of 4.28 %, 1.35 %, 1.71 %, and 0.84 % compared to Camelyon16 pre-trained ResNet50, MoCo v2, CS-CO, and MAE, respectively. In the fine-tuning task, while GCMAE's AUC was slightly lower than that of MAE by a margin of 0.16 %, GCMAE's accuracy consistently outperformed the other methods.

4.4.3. Performance analysis in cross-dataset and cross-disease transfer tasks

Upon further analysis of the results presented in Tables 3 and 4, several insights emerge. Initially, ResNet50 outperformed ViT-B/16 in the baseline tests for transfer learning, potentially due to the limited scale of the dataset used in our experiment, which may have restricted the full potential of ViT-B/16. ViTs have a stronger representation capability due to their reduced inductive bias, yet are more dependent on larger quantities of data [15]. Fortuitously, with only 200,000 pathology images for pre-training via GCMAE, the ViT models were able to surpass the performance of ResNet on smaller-scale pathology image datasets, demonstrating the broad applicability and cost-effectiveness of GCMAE for pathology image representation.

Moreover, as a prototypical algorithm within the MIM paradigm, MAE surpassed the contrastive learning approaches in pathology image representation tasks. MAE excels by randomly masking image patches and reconstructing the missing parts, encouraging the model to capture a global representation of the image. In contrast, random cropping, a common data augmentation method used in contrastive learning, can only input the main part of the cropped picture into the network training process, and no other means are taken to urge the encoder to learn the representation for the missing part. As a result, contrastive learning usually pays more attention to the features of the subject part, which inevitably reduces the universality and generalization of the representational transfer process to downstream tasks.

Lastly, other SSL methods specifically designed for pathology image representation, such as CS-CO, TransPath, and SSLP, also exhibit certain limitations. CS-CO's model training is divided into two stages: a generative task for cross-stain prediction and a contrastive learning task for fine-tuning the encoder, whereas GCMAE utilizes a more streamlined and efficient approach with two pretext tasks conducted simultaneously. TransPath implements contrastive learning via data augmentation without considering inter-patch relationships. GCMAE, on the other hand, not only delves deep into the features of each image patch but also explores the rich semantics between them through contrastive learning. SSLP designs three pretext tasks utilizing the spatial adjacency of patches in WSIs, but its effectiveness may be diminished for datasets containing only patch-level pathology images due to the loss of spatial information. As SSLP is not open-sourced, we did not directly include it in our comparison. Furthermore, GCMAE, benefiting from the design of the masked image generation task, also outperforms the aforementioned methods in terms of hardware consumption.

Table 5
Data settings for different dataset sizes.

Dataset	Train				Test
	100 %	80 %	50 %	10 %	
Camelyon16	40,000	32,000	20,000	4000	10,000
NCTCRC	100,000	80,000	50,000	10,000	7180

Table 6
Experimental results obtained by the model on tasks with different dataset sizes (mean±std%).

Task	Method	Dataset size			
		10 %	50 %	80 %	100 %
Transferring from Camelyon16 to NCTCRC	ImageNet pre-trained ViT-B/16	49.92 ±1.41	54.83 ±0.98	76.53 ±0.93	81.46 ±0.92
	GCMAE	**86.73 ±0.45**	**88.52 ±0.39**	**89.13 ±0.31**	**89.22 ±0.32**
Transferring from NCTCRC to Camelyon16	ImageNet pre-trained ViT-B/16	54.59 ±0.89	78.93 ±0.59	80.58 ±0.41	81.13 ±0.49
	GCMAE	**79.91 ±0.43**	**80.94 ±0.28**	**81.39 ±0.17**	**81.56 ±0.23**

4.5. Robustness test

Robustness tests are designed to evaluate a model's resistance to anomalous inputs, such as sudden changes in dataset size and various noise attacks. Specifically, we randomly selected 10 %, 50 %, and 80 % of the data from the training set used in Section 4.4′s downstream tasks to investigate the impact of dataset size on model performance. Regarding noise attacks, we introduced five different types of noise into the dataset to verify the model's noise immunity. In this experiment, we solely assessed the performance of GCMAE using linear probing as the downstream task.

4.5.1. Influence of the dataset size on the representation ability of the GCMAE

This experiment primarily investigates the performance of models pre-trained with GCMAE across different dataset sizes. The experiment utilized the full training dataset (i.e., 100 %) and its randomly extracted subsets of 80 %, 50 %, and 10 %, while the test set remained unchanged, all within the context of cross-dataset and cross-disease transfer tasks. For details of the data, see Table 5. As shown in Table 6, when the dataset size was reduced from 100 % to 10 %, for both tasks, the accuracy of ViT-B/16 models without GCMAE pre-training decreased by 31.54 % and 26.54 % respectively, whereas the accuracy of the GCMAE pre-trained ViT-B/16 models only decreased by 2.49 % and 1.65 %, improving stability by 12.7 and 16 times respectively. Furthermore, the impact of different dataset sizes on the model's classification performance is visualized in Fig. 3, which shows that the performance of the ViT-B/16 models pre-trained with GCMAE is more stable, with almost no noticeable decline, whereas the accuracy of ViT-B/16 models without GCMAE pre-training drops more significantly in complex tasks (as shown in Fig. 3(a)). This further validates the effectiveness of GCMAE in reducing the model's dependence on large amounts of high-quality

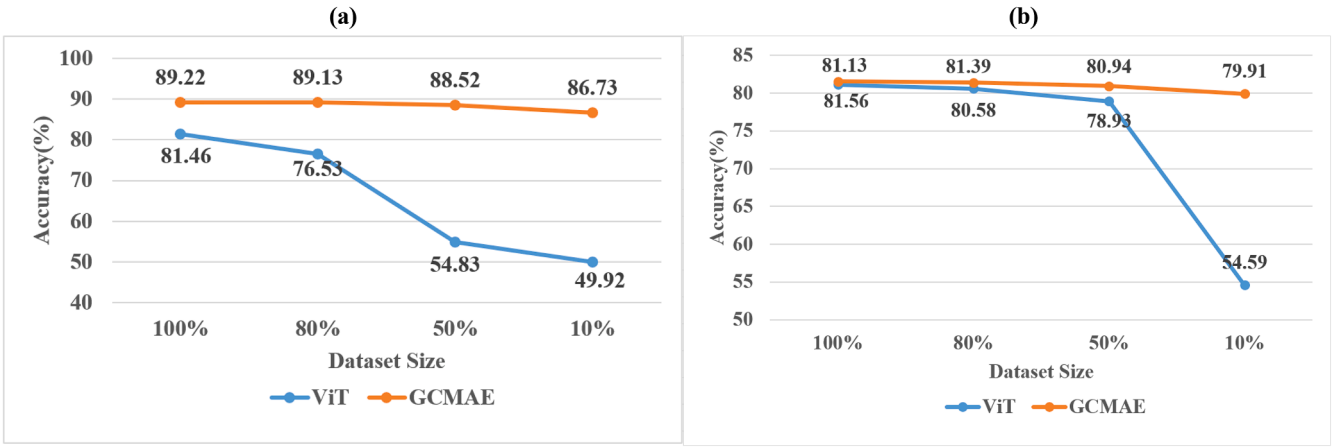

Fig. 3. Effects of different dataset sizes on model performance. (A) is the experimental result obtained by pretraining on Camelyon16 and transferring to NCTCRC for downstream tasks, and (b) is the experimental result obtained by pretraining on NCTCRC and transferring to Camelyon16 for downstream tasks.

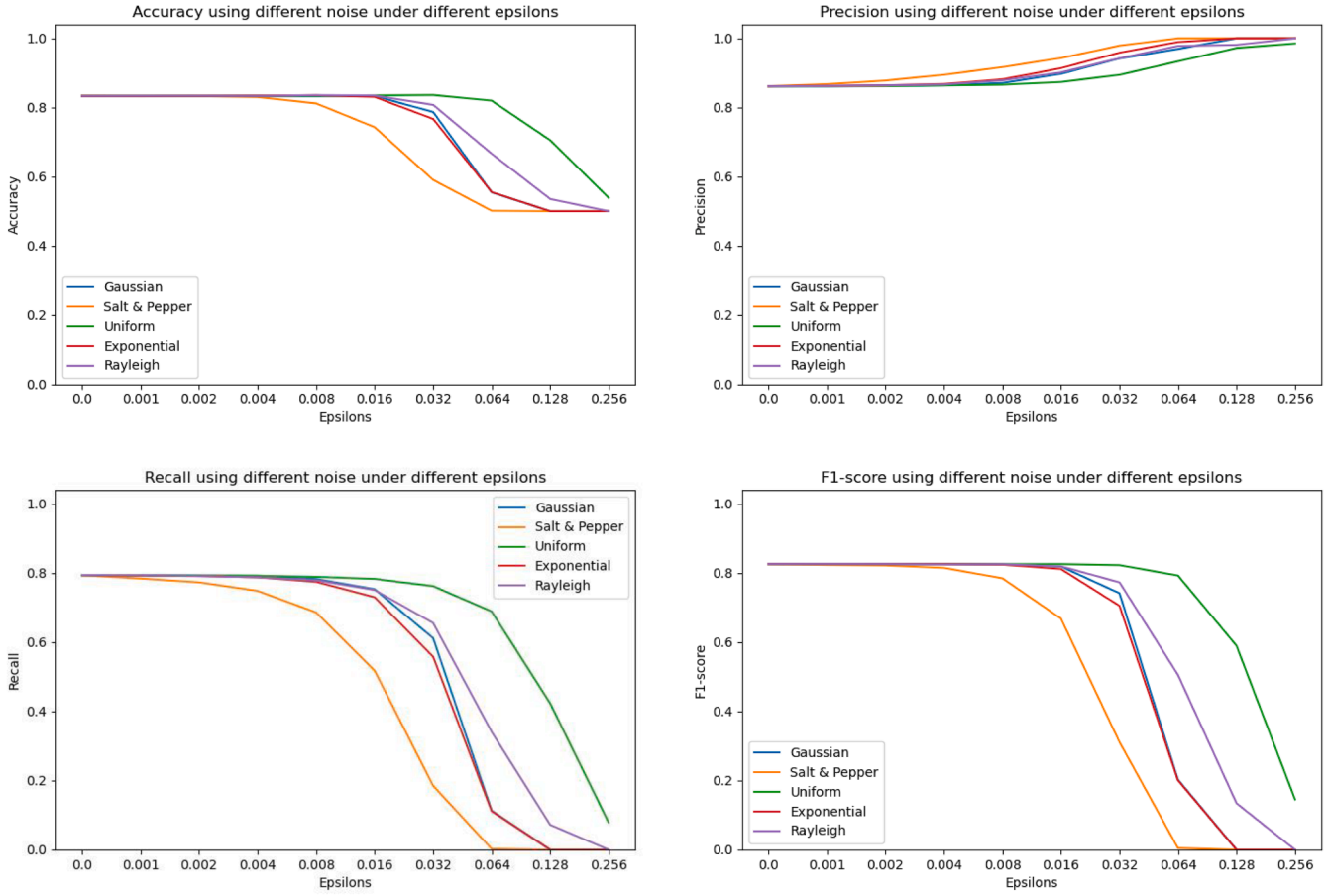

Fig. 4. Effects of five noises on GCMAE performance.

labeled data and in handling complex tasks, which is of significant importance for expanding the application scope of deep learning in the field of medical image analysis.

4.5.2. Influence of noise attacks on the representation ability of the GCMAE

This experiment primarily examines the impact of noise attacks on GCMAE's performance. We introduced Gaussian noise, salt-and-pepper noise, uniform noise, exponential noise, and Rayleigh noise into the test dataset to interfere with the model's predictions. The initial epsilon value for noise intensity was set at 0.001, incrementing in powers of 2, ranging from 0.001 to 0.256. We evaluated the impact of noise on GCMAE's performance using four metrics: accuracy, precision, recall, and F1 score. The results are displayed in Fig. 4. GCMAE exhibited the strongest robustness against uniform noise. The model's performance began to significantly decline when the epsilon value exceeded 0.064. For epsilon values below 0.016, GCMAE showed good robustness against all types of noise except for salt-and-pepper noise; when the epsilon value was below 0.004, it maintained good resistance to all noise types. Additionally, we observed that as the epsilon value approached 0.256, GCMAE tended to predict high-noise tumor images as normal, thereby

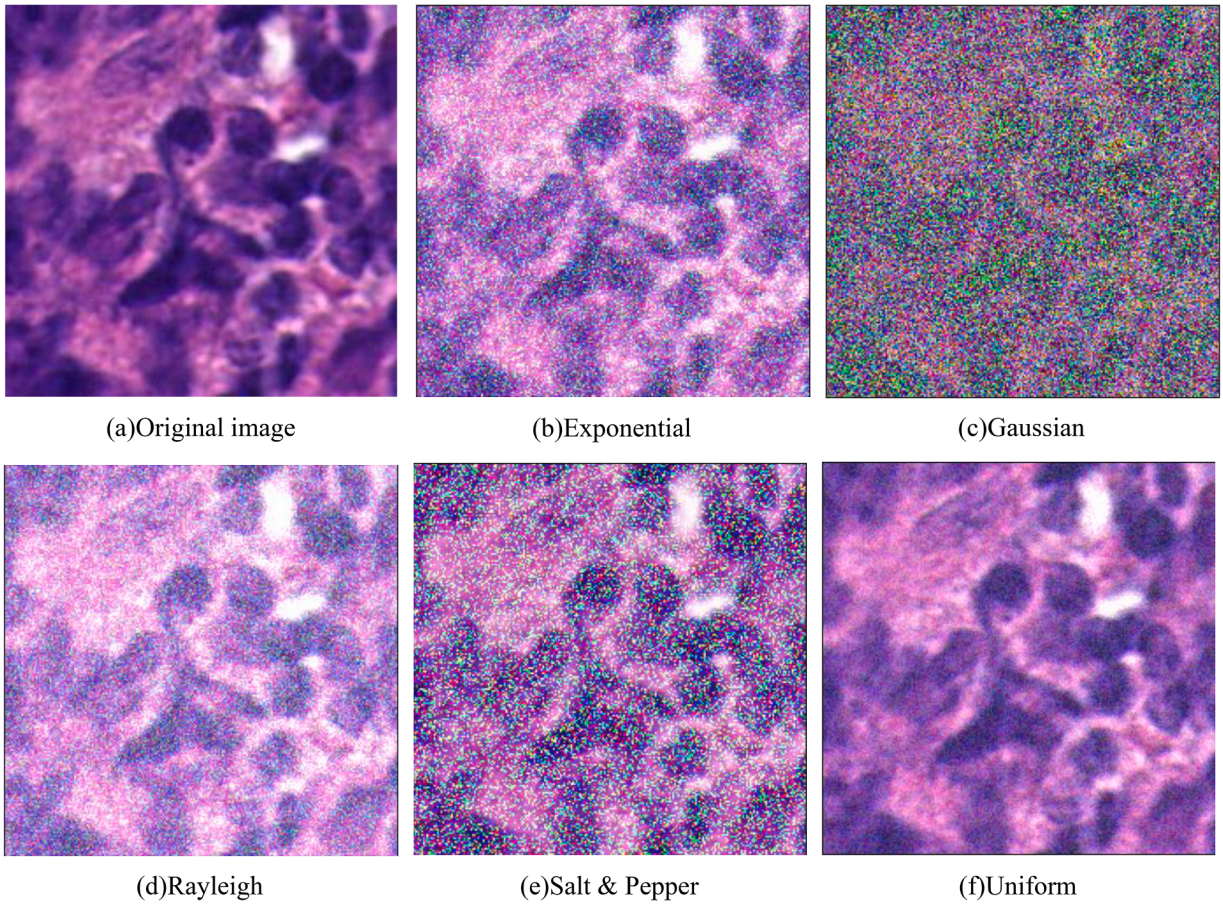

Fig. 5. Effects of adding five kinds of noise to pathological images.

Table 7
Pathological image classification results obtained on the BreakHis dataset (mean ±std%).

Method	Linear probing		Fine-tuning	
	Accuracy	AUC	Accuracy	AUC
ImageNet pre-trained ViT-B/16	51.24 ±0.74	72.21 ±0.59	52±0.65	74.18 ±0.87
Camelyon16 pre-trained ViT-B/16	52.42 ±0.41	75.84 ±0.58	51.85 ±0.37	73.22 ±0.37
MAE	53.14 ±0.34	**83.92 ±0.41**	89.52 ±0.27	98.51 ±0.32
GCMAE	**53.74 ±0.26**	83.79 ±0.29	**92.14 ±0.17**	**99.23 ±0.16**

reducing the number of false positives and leading to an increase in precision. Finally, Fig. 5 demonstrates the effect of adding noise to pathology images, with a visual comparison revealing that the uniform noise images most closely resemble the original images, explaining why GCMAE exhibits the best robustness against uniform noise. In summary, GCMAE demonstrated positive effects in resisting noise interference, proving its excellent robustness.

4.6. Extended experiment

4.6.1. Extended experiment on BreakHis

In this experiment, we further assessed the performance of GCMAE in pathology image representation using the BreakHis dataset. On the Camelyon16 dataset, we pre-trained the ViT-B/16 model using supervised learning, MAE, and GCMAE, and set the ViT-B/16 model pre-trained on ImageNet with supervised learning as the performance baseline. The specific experimental results are shown in Table 7. In the linear probing task, GCMAE achieved the best accuracy at 53.74 %, but it did not have a significant advantage over MAE and methods based on supervised pre-training. For the fine-tuning task, GCMAE significantly outperformed other methods, reaching an accuracy of 92.14 % and an AUC of 99.23 %. In this task, the accuracy of models pre-trained with self-supervision was significantly higher than those pre-trained with

Table 8
Influence of Stain Normalization on Transfer Learning Performance (mean±std%).

Method	Linear probing		Fine-tuning	
	w/o stain normalization	w/ stain normalization	w/o stain normalization	w/ stain normalization
Camelyon16 pre-trained ResNet50	78.69±0.79	79.58±0.86	89.12±0.96	89.33±0.41
Camelyon16 pre-trained ViT-B/16	73.46±0.85	74.91±0.74	82.01±0.78	83.92±0.57
MoCo v2	81.75±0.74	82.41±0.62	91.29±0.85	90.74±0.45
CS-CO	85.58±0.54	87.56±0.85	92.01±0.33	93.74±0.51
MAE	85.25±0.43	87.91±0.47	93.43±0.47	93.91±0.47
GCMAE	**89.22±0.32**	**90.58±0.49**	**93.89±0.25**	**94.21±0.27**

Table 9
Results of Whole Slide Image Classification on the Camelyon16 Dataset (mean ±std%).

Method	TransMIL	
	Accuracy	AUC
ImageNet pre-trained ViT-B/16	86.05±0.59	85.15±0.72
Camelyon16 pre-trained ViT-B/16	88.49±0.61	88.58±0.93
MAE	90.08±0.43	91.92±0.46
GCMAE	**91.86±0.52**	**92.79±0.76**

supervision, especially GCMAE, which saw an accuracy increase of 40.14 % compared to the ViT-B/16 pre-trained on ImageNet with supervised learning. These results clearly demonstrate that GCMAE outperforms other comparative algorithms on the BreakHis dataset in both linear probing and fine-tuning tasks, further validating GCMAE's universal representational capability and superior robustness.

4.6.2. *Influence of stain normalization on transfer learning performance*

In transfer learning tasks involving pathology images, stain normalization plays a crucial role in reducing data heterogeneity. To this end, we compared the performance of various methods on linear probing and fine-tuning tasks, both with and without stain normalization. We selected the best-performing methods from transfer learning, contrastive learning, pathology-specific approaches, and the MIM paradigm from Table 3 for comparison against GCMAE. The experimental results in Table 8 indicate that the accuracy of all methods improved on both linear probing and fine-tuning tasks after applying stain normalization. Notably, GCMAE achieved the highest accuracy following stain normalization, with rates of 90.58 % and 94.21 %, respectively. Particularly in the linear probing task, the performance increase of MAE was the most significant, with an accuracy improvement of 2.66 %, reaching 87.91 %, although this still fell short of GCMAE's 89.22 % accuracy without stain normalization. These findings suggest that stain normalization has a positive impact on enhancing model performance in pathology image tasks, especially evident in the linear probing task.

4.6.3. *Whole slide image classification experiment*

WSI classification is a crucial task in clinical diagnosis. Within this task, the Multi-Instance Learning (MIL) algorithm is extensively applied, where the quality of training for feature extractors directly impacts its classification performance. To further validate the effectiveness of GCMAE, we conducted experimental research on WSI classification within the MIL framework. Specifically, we utilized a classic MIL algorithm, TransMIL, and performed validation experiments on the Camelyon16 dataset. The experimental results (see Table 9) demonstrated that the ViT-B/16 model pre-trained with GCMAE achieved the highest accuracy rate of 91.86 % in the WSI classification task, marking an accuracy improvement of 5.81 % compared to the ViT-B/16 baseline model pre-trained with supervised learning on the ImageNet dataset. It is worth emphasizing that models based on self-supervised pre-training outperformed those pre-trained with supervision, mainly due to the self-supervised learning's ability to effectively avoid the decrease in feature generalization performance caused by overfitting labels, thereby possessing better feature generalization capability. These results prove the efficacy of GCMAE in enhancing WSI classification performance, offering robust technical support for the future development of clinical diagnostic tools.

5. Conclusion and future work

In this study, we introduce the GCMAE self-supervised learning framework specifically designed for pathology image representation, aimed at encoding a universal representation of pathology images. By integrating contrastive learning with masked image reconstruction tasks, GCMAE effectively pre-trains models to learn general pathology image representations, significantly enhancing model performance in cross-dataset and cross-disease transfer learning tasks. This approach promises to mitigate the challenges of modeling rare diseases caused by the long-tail problem in medical imaging through cross-dataset pre-training. Moreover, we explored the optimal mask ratio suitable for pathology image representation, providing a reference for future self-supervised algorithms in the MIM paradigm within the pathology image domain. Based on GCMAE, we also proposed a rational automatic diagnostic process for HE-stained pathology images, exploring its potential feasibility in clinical applications. However, this work has its limitations. Currently, GCMAE has been pre-trained only on a limited pathology image dataset from a single organ, and due to data scale constraints, only the vit-base model has been pre-trained, limiting its ability to construct a broadly applicable representation model. Future work will encompass integrating large-scale pathology image datasets from multiple organs and centers, and utilizing GCMAE to pre-train larger models such as ViT-Large/-Huge, aiming to build a truly universal large-scale pathology image representation model. At the same time, we will also consider improving the masking strategy to extend the applicability of GCMAE to the Pyramid-based ViTs model.

CRediT authorship contribution statement

Hao Quan: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Xingyu Li:** Conceptualization, Data curation, Methodology, Validation, Visualization, Writing – review & editing. **Weixing Chen:** Formal analysis, Methodology, Software, Writing – review & editing. **Qun Bai:** Data curation, Investigation. **Mingchen Zou:** Data curation, Investigation. **Ruijie Yang:** Data curation, Validation, Visualization. **Tingting Zheng:** Data curation, Investigation. **Ruiqun Qi:** Project administration, Supervision. **Xinghua Gao:** Project administration, Supervision. **Xiaoyu Cui:** Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code of this paper is publicly available at https://github.com/StarUniversus/gcmae.

Acknowledgment

This work is supported by the National High-tech Research and Development Program (Grant No. 2023YFC2508200), Liaoning Provincial Natural Science Foundation (Grant No. 2022-MS-105) and Department of Science and Technology of Liaoning Province (Grant No. 2022-YGJC-76). There is no conflict of interest in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2024.110745.

References

[1] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (2019) 1301–1309.

[2] J.G. Elmore, G.M. Longton, P.A. Carney, B.M. Geller, T. Onega, A.N.A. Tosteson, H. D. Nelson, M.S. Pepe, K.H. Allison, S.J. Schnitt, Diagnostic concordance among pathologists interpreting breast biopsy specimens, JAMa 313 (2015) 1122–1132.
[3] D. Wang, A. Khosla, R. Gargeya, H. Irshad, A.H. Beck, Deep learning for identifying metastatic breast cancer, arXiv preprint arXiv:1606.05718, (2016) .
[4] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Snchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[5] M.Y. Lu, D.F.K. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (2021) 555–570.
[6] H. Quan, X. Xu, T. Zheng, Z. Li, M. Zhao, X. Cui, DenseCapsNet: detection of COVID-19 from X-ray images using a capsule neural network, Comput. Biol. Med. 133 (2021) 104399.
[7] P. Pati, A. Foncubierta-Rodrguez, O. Goksel, M. Gabrani, Reducing annotation effort in digital pathology: a Co-Representation learning framework for classification tasks, Med. Image Anal. 67 (2021) 101859.
[8] N. Agarwal, A. Sondhi, K. Chopra, G. Singh, Transfer learning: survey and classification, in: Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS 2020, 2021, pp. 145–155.
[9] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, 2020, pp. 9729–9738 .
[10] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, 2020, pp. 1597–1607 .
[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollr, R. Girshick, Masked autoencoders are scalable vision learners, 2022, pp. 16000–16009 .
[12] O. Dehaene, A. Camara, O. Moindrot, A. de Lavergne, P. Courtiol, Self-supervision closes the gap between weak and strong supervision in histology, arXiv preprint arXiv:2012.03583, (2020) .
[13] P. Yang, Z. Hong, X. Yin, C. Zhu, R. Jiang, Self-supervised visual representation learning for histopathological images, 2021, pp. 47–57 .
[14] J. Li, T. Lin, Y. Xu, Sslp: spatial guided self-supervised learning on pathological images, 2021, pp. 3–12 .
[15] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098, (2017) .
[16] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, 2018, pp. 3733–3742 .
[17] X. Li, C. Li, M.M. Rahaman, H. Sun, X. Li, J. Wu, Y. Yao, M. Grzegorzek, A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches, Artif. Intell. Rev. 55 (2022) 4809–4878.
[18] H. Quan, X. Li, D. Hu, T. Nan, X. Cui, Dual-Channel Prototype Network for Few-Shot Pathology Image Classification, IEEE J. Biomed. Health Inform. 28 (7) (2024) 4132–4144.
[19] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, Transmil others, Transformer based correlated multiple instance learning for whole slide image classification, Adv. Neur. Inf. Process. Syst. 34 (2021) 2136–2147.
[20] J. Liang, W. Zhang, J. Yang, M. Wu, Q. Dai, H. Yin, Y. Xiao, L. Kong, Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer, Nat. Mach. Intell. 5 (2023) 408–420.
[21] Z. Gao, Z. Lu, J. Wang, S. Ying, J. Shi, A convolutional neural network and graph convolutional network based framework for classification of breast histopathological images, IEEE J. Biomed. Heal. Inform. 26 (2022) 3163–3173.
[22] Y. Zhang, Z. Li, X. Han, S. Ding, J. Li, J. Wang, S. Ying, J. Shi, Pseudo-Data based Self-Supervised Federated Learning for Classification of Histopathological Images, IEEE Trans. Med. ImAging (2023).
[23] J. Shi, X. Zheng, J. Wu, B. Gong, Q. Zhang, S. Ying, Quaternion Grassmann average network for learning representation of histopathological image, Pattern. Recognit. 89 (2019) 67–76.
[24] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: generative or contrastive, IEEe Trans. Knowl. Data Eng. 35 (2021) 857–876.
[25] A. Chowdhury, J. Rosenthal, J. Waring, R. Umeton, Applying self-supervised learning to medicine: review of the state of the art and medical implementations, 2021, pp. 59 .
[26] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, 2022, pp. 2–25 .
[27] X. Chen, L. Yao, T. Zhou, J. Dong, Y. Zhang, Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images, Pattern. Recognit. 113 (2021) 107826.
[28] C.J. Reed, S. Metzger, A. Srinivas, T. Darrell, K. Keutzer, Selfaugment: automatic augmentation policies for self-supervised learning, 2021, pp. 2674–2683 .
[29] J. Xu, J. Hou, Y. Zhang, R. Feng, C. Ruan, T. Zhang, W. Fan, Data-efficient histopathology image analysis with deformation representation learning, 2020, pp. 857–864 .
[30] R. Gong, L. Wang, J. Wang, B. Ge, H. Yu, J. Shi, Self-Distilled Supervised Contrastive Learning for diagnosis of breast cancers with histopathological images, Comput. Biol. Med. 146 (2022) 105641.
[31] C.L. Srinidhi, S.W. Kim, F.-D. Chen, A.L. Martel, Self-supervised driven consistency training for annotation efficient histopathology image analysis, Med. Image Anal. 75 (2022) 102256.
[32] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, C. Feichtenhofer, Masked feature prediction for self-supervised visual pre-training, 2022, pp. 14668–14678 .
[33] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, L. Yuan, Bevt: bert pretraining of video transformers, 2022, pp. 14733–14743 .
[34] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, P. Prasanna, Self pre-training with masked autoencoders for medical image analysis, arXiv preprint arXiv: 2203.05573, 1 (2022) .
[35] Y. Luo, Z. Chen, S. Zhou, X. Gao, Self-distillation augmented masked autoencoders for histopathological image classification, arXiv preprint arXiv:2203.16983, (2022) .
[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, others, An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv: 2010.11929, (2020) .
[37] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531, (2015) .
[38] B.E. Bejnordi, M. Veta, P.J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J.A.W.M. Van Der Laak, M. Hermsen, Q.F. Manson, M. Balkenhol, Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, JAMA 318 (2017) 2199–2210.
[39] J.N. Kather, N. Halama, A. Marx, 100,000 histological images of human colorectal cancer and healthy tissue, Zenodo10 5281 (2018).
[40] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, IEEE Trans. Biomed. Eng 63 (2015) 1455–1462.
[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016, pp. 770–778 .
[42] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, X. Han, Transpath: transformer-based self-supervised learning for histopathological image classification, 2021, pp. 186–195 .

Hao Quan received the B.S. degree from the Department of Electrical and Information Engineering, Beihua University, China in 2018. He received the M.S. degree from College of Medicine and Biological Information Engineering, Northeastern University, China in 2022. Since 2023, he has been a joint Ph.D. student in Biomedical Engineering at Northeastern University and Shenzhen Institutes of Advanced Technology. His research interests include computer vision, self-supervised learning, and medical image analysis.

Xingyu Li is currently working toward a B.S. degree at the College of Medicine and Biological Information Engineering, Northeastern University, China. His research is focused on computer vision and machine learning.

Weixing Chen got the B.S. degree from Northeastern University, and now is working toward the M.S. degree in Shenzhen Institute of advanced technology, Chinese Academy of Sciences. His research interest mainly includes medical image analysis and multimodal understanding and generation.

Qun Bai is an undergraduate student at the college of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China. His research interests are computer vision and image processing.

Mingchen Zou is a M.S. student at the College of Medicine and Biological Information Engineering, Northeastern University, China. And he received B.S. degree in Mathematics from Southwest Minzu University, China in 2020. His research interests are reinforcement learning and self supervised learning.

Ruijie Yang is an undergraduate majoring in artificial intelligence in the school of computer science and engineering of Northeast University of China. His research interests are deep learning.

Tingting Zheng was born in Henan, China. She received her Bachelor's degree in Software Engineering from Henan University of Economics and Law. She is pursuing a Master's degree in Electronic Information at Northeastern University in Shenyang, China. Her research interests include weakly supervised learning, reinforcement learning medical image disease detection, and artificial intelligence-based learning.

Prof. Ruiqun Qi, male, born in 1978, researcher/professor, doctoral supervisor. Now he works in the Department of Dermatology, No.1 Hospital of China Medical University, and is a member of the Experimental Group of Dermatology and Venereology Branch of Chinese Medical Association. Henry Ford Health System returned from an exchange visit to the United States. He has published more than 30 SCI papers by the first work or correspondent author in Hepatology, J Invest Dermatol, JAAD and other journals. A total of 36 patents were obtained, including 5 international patents, and patents in 3 directions were successfully transformed. He presided over 4 projects of National Natural Science Foundation of China. The first finisher won one first prize of scientific and technological progress in Liaoning Province, and the second finisher won one second prize of scientific and technological progress in Liaoning Province. He was awarded the 2016 Outstanding Young and Middle-aged Doctor of Dermatology in China, the top-notch young talent of China Medical University, and won the 21st Century Important Medical Achievement Award as the third finisher.

Prof. Xing-Hua Gao, a Changjiang Scholarship Professor, is Ph.D. supervisor. He was studied in Oxford University, UK. And he currently serves as vice president of No.1 Hospital of China Medical University, director of department of dermatology, vice director of Key Lab of Immunodermatology, Ministry of Health, director of Institute of Dermatology and Cosmetic Dermatology of China Medical University, leader of the Innovation Team of

Ministry of Education on "Experimental and clinical studies of immune related skin diseases" (2008–2010), member of Scientific Committee of Ministry of Education, grantee of program for New Century Excellent Talents Support Program, expert who enjoying the State Council Subsidy, President designate of Chinese Society of Dermatology, committee member of Chinese Society of Medical Aesthetics and Cosmetology, President designate of Chinese Medical Association Dermatology Branch. He is the vice Editor-in-Chief of Journal of Applied Cosmetology (Europe), Journal of Dermatology and Venereology and Journal of Clinical Dermatology. He also is member of editorial board of International Journal of Dermatology (US), Chinese Journal of Medical Aesthetics and Cosmetology, Journal of Practical Dermatology, China Journal of Dermatology, China Journal of Leprosy and Skin Diseases, and Chinese Journal of Dermatology. He had published over 200 academic papers, including over100 papers in SCI indexed journals. And he is the chief editor, author or referee of more than 20 books (four books in English). He had won 7 patents, 1 First Class Award for Chinese Medical Science and Technology Prize, 5 items of Science and Technology Progress Award on Provincial and ministerial-level, and 1 Muto Journal paper Award.

Prof. Xiaoyu Cui received his Bachelor degrees in Electronics and Information Engineering in 2007 from Shenyang University of Technology and received his Master and Doctor degrees in Biomedical Engineering in 2009 and 2013, respectively, from Northeastern University. He is currently an associate professor in College of Medicine and Biological Information Engineering at Northeastern University in China. His research interests include optical imaging and machine learning.