DiffMIC-v2: Medical Image Classification via Improved Diffusion Network

Yijun Yang, Huazhu Fu, *Senior Member, IEEE*, Angelica I. Aviles-Rivero, Zhaohu Xing, and Lei Zhu, *Member, IEEE*

***Abstract*— Recently, Denoising Diffusion Models have achieved outstanding success in generative image modeling and attracted significant attention in the computer vision community. Although a substantial amount of diffusion-based research has focused on generative tasks, few studies apply diffusion models to medical diagnosis. In this paper, we propose a diffusion-based network (named DiffMIC-v2) to address general medical image classification by eliminating unexpected noise and perturbations in image representations. To achieve this goal, we first devise an improved dual-conditional guidance strategy that conditions each diffusion step with multiple granularities to enhance step-wise regional attention. Furthermore, we design a novel Heterologous diffusion process that achieves efficient visual representation learning in the latent space. We evaluate the effectiveness of our DiffMIC-v2 on four medical classification tasks with different image modalities, including thoracic diseases classification on chest X-ray, placental maturity grading on ultrasound images, skin lesion classification using dermatoscopic images, and diabetic retinopathy grading using fundus images. Experimental results demonstrate that our DiffMIC-v2 outperforms state-of-the-art methods by a significant margin, which indicates the universality and effectiveness of the proposed model on multi-class and multi-label classification tasks. DiffMIC-v2 can use fewer iterations than our previous DiffMIC to obtain accurate estimations, and also achieves greater runtime efficiency with superior results. The code will be publicly available at https://github.com/scott-yjyang/DiffMICv2.**

***Index Terms*— Medical image classification, diffusion models, chest X-ray, ultrasound, skin lesion, diabetic retinopathy.**

Received 3 December 2024; accepted 6 January 2025. Date of publication 15 January 2025; date of current version 2 May 2025. This work was supported in part by The Hong Kong University of Science and Technology (Guangzhou) [HKUST(GZ)] Joint Funding Program under Grant 2023A03J0671; in part by Guangzhou Municipal Science and Technology Project under Grant 2024312139; in part by Guangdong Provincial Key Laboratory of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007; and in part by the Huazhu Fu's Agency for Science, Technology and Research (A*STAR) Central Research Fund ("Robust and Trustworthy AI System for Multi-Modality Healthcare"). *(Corresponding author: Lei Zhu.)*

Yijun Yang and Zhaohu Xing are with the Robotics and Autonomous Systems, The Hong Kong University of Science and Technology (Guangzhou), Nansha, Guangzhou, Guangdong 511400, China (e-mail: yyang018@connect.hkust-gz.edu.cn; zxing565@connect.hkust-gz.edu.cn).

Huazhu Fu is with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: hzfu@ieee.org).

Angelica I. Aviles-Rivero is with the Yau Mathematical Sciences Center, Tsinghua University, Beijing 100190, China (e-mail: aviles-rivero@tsinghua.edu.cn).

Lei Zhu is with the Robotics and Autonomous Systems, The Hong Kong University of Science and Technology (Guangzhou), Nansha, Guangzhou, Guangdong 511400, China, and also with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: leizhu@ust.hk).

Digital Object Identifier 10.1109/TMI.2025.3530399

I. Introduction

MEDICAL image analysis plays an indispensable role in clinical therapy because of the implications of digital medical imaging in modern healthcare [1]. A key component of this analysis is medical image classification, which aims to differentiate medical images based on specific criteria for various modalities. Developing an automatic and reliable classification system can greatly assist radiologists in interpreting medical images quickly and accurately. Consequently, a large number of solutions for automatic medical image classification have been developed over the past decades in the literature, ranging from handcrafted features to deep learning. While handcrafted features are not flexible and have poor generalization on unseen data [2], [3], [4], deep features are data-driven and are becoming the approach of choice in learning medical image representation for classification, most of which are based on the convolutional neural networks (CNNs) and vision transformer architectures [5], [6], [7], [8], [9], [10], [11], [12], [13]. These methods have the potential to reduce the time and effort required for manual classification and improve the accuracy and consistency of results. However, medical images with diverse modalities still frequently challenge existing methods due to the presence of various ambiguous lesions and fine-grained tissues, such as ultrasound (US), dermatoscopic, and fundus images. Moreover, generating medical images under hardware limitations can cause noisy and blurry effects, which can degrade image quality and thus demand a more robust feature representation modeling for effective classifications.

Recently, Denoising Diffusion Probabilistic Models (DDPM) [14] have achieved excellent results in image generation and synthesis tasks [15], [16], [17], [18] by iteratively improving the quality of a given image. Specifically, DDPM is a generative model based on a Markov chain, which models the data distribution by simulating a diffusion process that evolves the input data towards a target distribution. Although a few pioneer works tried to adopt the diffusion model for image segmentation and object detection

tasks [19], [20], [21], [22], [23], [24], their potential for high-level vision has yet to be fully explored.

Motivated by the achievements of diffusion probabilistic models in generative image modeling, in this work, we present a novel denoising diffusion-based model named DiffMIC-v2 for accurate diagnosis of diverse medical image modalities. As far as we know, we are the first to propose a diffusion-based model for general medical image classification. DiffMIC-v2 empowers an improved diffusion network to robustly eliminate undesirable noise in deep features from medical images. Specifically, we introduce a Dual-granularity Conditional Guidance (DCG) strategy to condition each step of the denoising procedure with global and local priors in the diffusion process. We develop a dense guidance map and image feature prior to aggregate the global and local information from the DCG model, achieving more precise guidance for denoising. By conducting the diffusion process on smaller patches, our method can distinguish critical tissues with fine-grained capability. Also, we design a novel Heterologous Diffusion Process, allowing various noise points to interact in the 2D latent feature. This strategy explores the common feature representation under different perturbations by convolution operations. We evaluate the effectiveness of DiffMIC-v2 on four 2D medical image classification tasks including chest x-ray classification, placental maturity grading, skin lesion classification, and diabetic retinopathy grading.

In summary, the contributions of our work are four-fold:

- We develop an improved diffusion-based model for generic medical image classification. Our DiffMIC-v2 provides a promising solution for the accurate and robust classification of diverse medical image modalities (*e.g.*, X-Ray, ultrasound, dermatoscopy and fundus).
- We introduce a Dual-granularity Conditional Guidance (DCG) strategy by leveraging both the global and local information to construct the 2D dense guidance map and image feature prior and guide the conditional diffusion process.
- A novel Heterologous Diffusion Process is developed to enforce various noise perturbations to interact in one feature instance for efficient visual representation learning.
- Extensive experimental results demonstrate that our diffusion-based classification method consistently and significantly surpasses state-of-the-art methods for all four tasks.

Different from our previous work DiffMIC [25], DiffMIC-v2 adopts an improved diffusion process to construct the general feature representation under simultaneous noise perturbations. This accelerates the convergence efficiency of the diffusion models by the aggregation of diverse noise points, which originally independently vary in multiple iterations. We also develop a 2D dense guidance map to appropriately guide the denoising of each noise point. In our DiffMIC-v2, we only preserve the shared stream and thus spare the computations of the global and local stream in DiffMIC to further improve the model's efficiency. As observed, DiffMIC-v2 is about 3× more efficient than DiffMIC with better performance.

II. Related Work

A. Deep Learning for Medical Image Classification

Medical image classification has been thoroughly studied over the past decades with massive solutions in the literature [3], [4], [26], most of which are based on handcrafted features. Despite the success of these methods, it is usually laborious and time-consuming to design the optimal handcrafted features for a specific classification task.

Recently, the development of deep learning techniques has significantly promoted the advancement of medical image classification, particularly deep convolutional neural networks (DCNN) and vision transformers [5], [6], [7], [8], [9], [10], [11], [27], [28], [29]. These deep neural network (DNN) models offer a unified feature extraction-classification framework, relieving human users from the tedious task of manually crafting features for medical image classification. For instance, Zhang et al. [30] jointly used deep and handcrafted visual features for medical image classification and found that handcrafted features could complement the image representation learned by DCNNs on small training datasets. Lei et al. [31] proposed integrating deep descriptors extracted from convolutional neural networks and hand-crafted features on ultrasound images to produce hybrid descriptors for boosting placental maturity grading performance. Esteva et al. [32] utilized 129,450 clinical images to train a DCNN for diagnosing the most common and deadliest skin cancers and achieved comparable performance against 21 board-certified dermatologists. Zhang et al. [33] developed the synergic deep learning method that takes multiple images as input, allowing multiple DCNN components to enhance each other's learning of more discriminative representation.

Since Chest X-ray14 [34] dataset was released, its high interclass similarity, dirty atypical data, complex symbiotic relationships between diseases, and long-tailed or imbalanced data distribution frequently challenged traditional methods. To tackle these issues, deep neural networks have been applied to assist radiologists in the analysis of chest x-ray [35], [36], [37], [38], [39], [40]. For instance, DNetLoc [37] proposes a location-aware dense network by incorporating the spatial information of chest X-ray pathologies. Differently, ImageGCN [40] involves the natural relations between CXR images and view information into their framework to learn the image representation. However, both of them introduce external data or information to improve the multi-label classification performance. Imbalanced data distribution also impedes the accurate diagnosis of common pigmented skin lesions [41] and diabetic retinopathy [42]. Marrakchi et al. [43] emphasize the issue of class imbalance and design a contrastive learning mechanism to arrange the feature space for minority and majority classes. ProCo [44] presents a prototype-based framework to recalibrate the category distribution and update the frequency of tailed classes.

B. Conditional Generative Models

Deep generative models, such as Variational autoencoder (VAE) [45] and Generative Adversarial Network (GAN) [46],

have gained popularity as effective techniques for generating synthetic samples to address imbalanced data distributions in image classification tasks. These models learn to mimic the data generation process, enabling the creation of new instances that help compensate for skewed distributions. The VAE excels at capturing the underlying data distribution and can be directly applied to imbalanced datasets to model dependencies within the data through its latent variables. For instance, Guo et al. [47] proposed a method to handle binary imbalance classification by modeling latent embeddings with two Gaussian distributions that have opposite means. Additionally, deep latent variable models, such as DGCMM [48], use a Gaussian Mixture Model to represent latent variables and enforce uncertainty in the model.

GANs offer an alternative generation strategy by learning a mapping from the latent space to the original data space. To generate class-specific minority samples, various conditional GANs (cGAN) have been developed, which incorporate label information into the generation process to make latent variables class-specific [49], [50]. Ghorbani et al. [51] demonstrated the effectiveness of cGANs in generating high-quality synthetic images to augment small datasets in skin lesion classification, which helped improve classification accuracy and generalization. NAGAN [52] reformulated noise adaptation in medical imaging using a generator and two discriminators. It adjusted noise patterns from test data to match those from training data while preserving content, thereby enhancing image classification performance across different devices and settings. BAGAN [53] combined VAE and cGAN in a two-step framework to oversample minority classes. However, oversampling with cGAN can sometimes result in boundary distortion [54]. Additionally, the separation between learning the data distribution and generating new samples can lead to inconsistencies between these two processes. To address this issue, the Generative Adversarial Minority Oversampling (GAMO) method [55] was introduced. It integrates data distribution learning and sample generation into a unified framework through a three-player adversarial game involving a convex generator, a multi-class classifier, and a real/fake discriminator, helping to alleviate boundary distortion and improve the consistency between learning and generation.

By introducing heterologous noise and multi-granular guidance, our latent diffusion model can generate high-quality latent representations in sparse data while preserving class boundaries. This enables more accurate sampling of minority classes within latent space, improving performance on imbalanced datasets without distorting boundaries.

C. Diffusion Models

Recently, Diffusion model is well-known as a novel generative modeling approach for its superior achievements in image synthesis and generation tasks [14], [17], [18], [56], [57], [58]. In essence, Denoising Diffusion Probabilistic Models (DDPM) [14] adopted parameterized Markov chain to optimize the lower variational bound on the likelihood function, which can simulate a diffusion process to iteratively improve the quality of target distribution than other generative models. DDIM [59] develops iterative implicit probabilistic models based on DDPM and introduces deterministic sampling process for the trade-off between computational costs and sample quality. Very recently, a few pioneer works tried to adopt diffusion models for high-level perceptual tasks, such as image classification, segmentation and object detection [19], [20], [21], [22], [23], [24], [25], [60]. Han et al. [22] combine a denoising diffusion-based conditional generative model and a pre-trained conditional mean estimator for natural image classification and regression. Chen et al. [21] propose a diffusion-based framework that formulates object detection as a denoising diffusion process from noisy boxes to object boxes. Diff-UNet [23] utilizes the diffusion models to solve 3D medical image segmentation problems and designs a Step-Uncertainty based Fusion (SUF) module during inference to bolster the robustness of the diffusion model's predictions. Their potential for high-level vision in medical images has yet to be fully explored.

Following our previous method [25], in this work, we employ an improved diffusion network to enhance the diagnosis of medical images and prove its universality across diverse universal modalities, *e.g.*, X-ray, ultrasound, dermatoscopic and fundus.

III. Method

A. Preliminaries: Diffusion Models

The diffusion model is composed of a forward and a reverse process. The forward process is defined as a discrete Markov chain of length T: $q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$. For each step $t \in [1, T]$ in the forward process, a diffusion model adds noise ϵ_t sampled from the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ to data x_{t-1} and obtains disturbed data x_t from $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$. β_t decides the ratio of noise at timestep t. Noticeably, instead of sampling sequentially along the Markov chain, we can sample x_t at any time step t in the closed form via $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_{t-1}, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. To parameterize the Gaussian distribution, the neural network ϵ_θ is introduced, which is optimized by the objective of DDPM [14]:

$$\mathcal{L}_\epsilon = ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||_2^2. \quad (1)$$

In the reverse process, the diffusion model gradually denoises the randomly sampled Gaussian noise to the high-quality output x_0 through the predicted noise by the well-trained ϵ_θ. This process is also defined as a Markov chain: $p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$, and $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I})$. The mean and variance are $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t))$, $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, respectively.

B. Overview

Figure 1 shows the schematic illustration of our network for medical image classification. Given an input medical image x, we pass it to a dual-granularity conditional guidance (DCG) model to produce the global prior $\hat{y}_g$ and local prior $\hat{y}_l$, which are further aggregated into the dense guidance map $\mathcal{M}$. Then, we extract the image feature prior by summing the

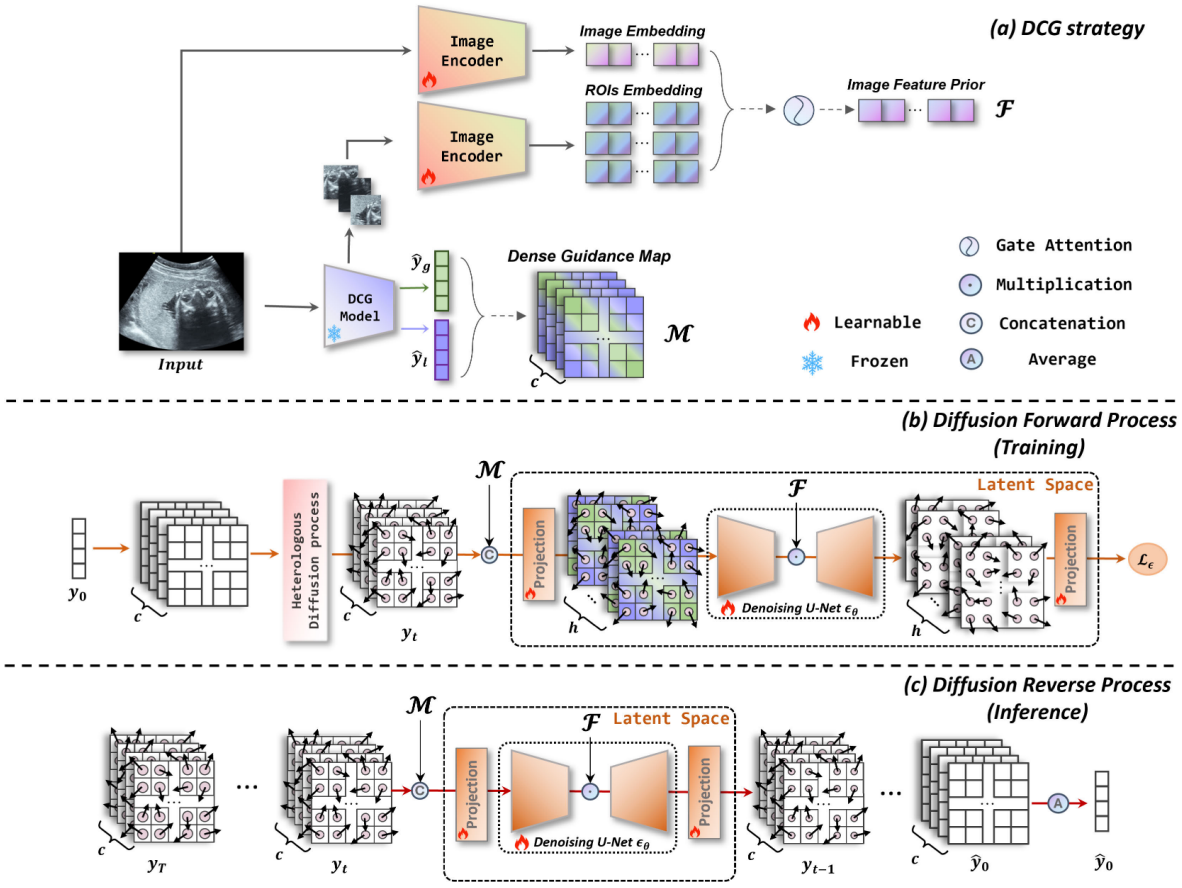

Fig. 1. **Overview of our DiffMIC-v2 framework.** (a) The DCG strategy generates two conditions, which are the dense guidance map $\mathcal{M}$ and image feature prior $\mathcal{F}$, to guide the diffusion process by aggregating the global and local information from the raw image and ROIs. The DCG model is pre-trained and frozen during the diffusion training while two image encoders are trained collaboratively with diffusion models. (b) The training phase (forward process) and (c) The inference phase (reverse process) of diffusion models are constructed, respectively. (The pink circle's size represents the noise's magnitude while different arrow directions mean different timesteps.)

image embedding and ROIs embedding using the attention mechanism. At the training stage, we apply the heterologous diffusion process on ground truth y_0 to generate the noisy variable $y_\mathbf{t}$ with the randomly sampled timestep $\mathbf{t}$. Then, we combine the noisy variable $y_\mathbf{t}$ and the dense guidance map and project them into the latent space. We further integrate the projected embedding with the image feature prior $\mathcal{F}$ in the denoising U-Net and predict the noise distribution sampled for $y_\mathbf{t}$. We employ the noise estimation loss by mean squared error (MSE) on the predicted noise of $y_\mathbf{t}$ to train our DiffMIC-v2 network. At the inference stage, we apply the diffusion reverse process under the guidance of $\mathcal{M}$ and $\mathcal{F}$, and iteratively recover the final prediction from the randomly sampled variable y_T based on DDIM sampling paradigm [59].

C. Dual-Granularity Conditional Guidance Strategy

1) DCG Model: In most conditional DDPM, the conditional prior will be a unique given information. However, medical image classification is particularly challenging due to the ambiguity of objects. It is difficult to differentiate lesions and tissues from the background, especially in low-contrast image modalities, such as ultrasound images. Moreover, unexpected noise or blurry effects may exist in regions of interest (ROIs), thereby hindering the understanding of high-level semantics. Taking only a raw image x as the condition in each diffusion step will be insufficient to robustly learn the fine-grained information, resulting in classification performance degradation.

To alleviate this issue, we design a Dual-granularity Conditional Guidance (DCG) for encoding each diffusion step as illustrated in Fig. 1(a). Specifically, we introduce a DCG model $\tau_\mathcal{D}$ to compute the global and local conditional priors for the diffusion process. Similar to the diagnostic process of a radiologist, we can obtain a holistic understanding from the global prior and also concentrate on areas corresponding to lesions from the local prior when removing the negative noise effects. As shown in Figure 2, for the global stream, the raw image data x is fed into the global encoder τ_g and then a 1×1 convolutional layer to generate a saliency map of the whole image. The global prior $\hat{y}_g$ is then predicted from the whole saliency map by averaging the responses. For the local stream, we further crop the top K ROIs whose responses are most significant in the saliency map of the image following [61]. Each ROI is fed into the local encoder τ_l to obtain a feature vector. We then leverage the gated attention mechanism [62] to fuse all feature vectors $\mathbf{r}_k$ to obtain a

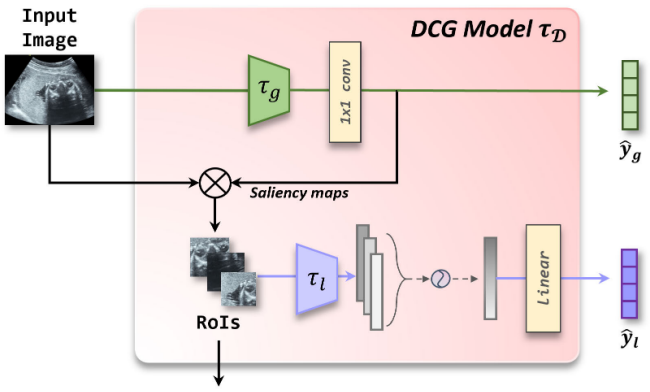

Fig. 2. **Overview of our DCG model.** The DCG Model τ_D guides the diffusion process by the dual priors $\hat{y}_g$, $\hat{y}_l$ from the raw image and ROIs. τ_g is the global encoder while τ_l is the local encoder.

weighted vector $\mathbf{z}$:

$$\alpha_k = \frac{\exp\{\mathbf{w}^T(tanh(\mathbf{V}\mathbf{r}_k^T) \odot sigm(\mathbf{U}\mathbf{r}_k^T))\}}{\sum_{j=1}^{K}\exp\{\mathbf{w}^T(tanh(\mathbf{V}\mathbf{r}_j^T) \odot sigm(\mathbf{U}\mathbf{r}_j^T))\}}, \quad (2)$$

$$\mathbf{z} = \sum_{k=1}^{K}\alpha_k\mathbf{r}_k, \quad (3)$$

where $\odot$ denotes element-wise multiplication and $\mathbf{w} \in \mathbb{R}^L$, $\mathbf{V} \in \mathbb{R}^{L\times M}$, $\mathbf{U} \in \mathbb{R}^{L\times M}$ are learnable parameters achieved by the linear layer. The attention-weighted representation is then utilized for computing the local prior $\hat{y}_l$ by one linear layer. Note that we pre-train the DCG model and adopt the standard cross-entropy loss as its training objective before the diffusion training.

2) Dense Guidance Map: Once obtaining the global prior $\hat{y}_g \in \mathbb{R}^C$ and local prior $\hat{y}_l \in \mathbb{R}^C$, we construct a 2D dense guidance map to provide each noise point with a specific prior. More specifically, we first calculate the distance matrix $\mathcal{D} \in \mathbb{R}^{1\times N_p\times N_p}$ as follows:

$$d_{ij} = \frac{|i-j|}{N_p - 1}, i, j \in [0, N_p - 1], \quad (4)$$

where d_{ij} is the element of the distance matrix in the i-th row and j-th column. The distance matrix is designed to capture all the possible distribution means between the global and local prior for more sophisticated guidance of diffusion models. Then, we expand the probability vectors $\hat{y}_g$ and $\hat{y}_l$ to 2D probability maps by duplicating points, respectively. The distance matrix is utilized as weights for the interpolation of the global and local priors, and the dense guidance map $\mathcal{M}$ is computed as:

$$m_{ij} = (1 - d_{ij}) \cdot \hat{y_g} + d_{ij} \cdot \hat{y}_l, i, j \in [0, N_p - 1], \quad (5)$$

Thus, $\mathcal{M} \in \mathbb{R}^{C\times N_p\times N_p}$ is a symmetric matrix, whose values are the point-wise linear interpolation between $\hat{y_g}$ and $\hat{y}_l$. Such a dense guidance map considers the complementary information of the global and local view (*i.e.*, $d_{ij} \in (0, 1)$) and also preserves the specificity of each view (*i.e.*, $d_{ij} = 0/1$), so that DiffMIC-v2 obsoletes the global and local stream in the original DiffMIC to improve its efficiency. Our approach improves upon a linear interpolation with fixed weights by performing the diffusion process using different probability distribution means, thereby eliminating the need for weight optimization.

Algorithm 1 Training Scheme

```
def train(images, gt, np, T):
"""
np: number of noise point, T: training steps
images: [b, 3, h, w], gt: [b, c]
"""
# generate conditional priors
# dense_map: [b, c, np, np], img_enc: [b, h, 1, 1]
dense_map, img_enc = dcg(images)
# corrupt gt
gt = gt.expand(b*np*np,c)
t = uniform(0, T-1) # t: [b*np*np, 1]
eps = normal(mean=0, std=1) # eps: [b*np*np, c]
gt_crpt = sqrt(alpha_cumprod(t)) * gt +
sqrt(1 - alpha_cumprod(t)) * eps
gt_crpt = gt_crpt.reshape(b,c,np,np)
x = cat(dense_map,gt_crpt)
t = t.reshape(b, 1, np, np)
# predict and backward
eps_pred = denoising_unet(x, img_enc, t)
loss = objective_func(eps_pred, eps)
return loss
```

3) Image Feature Prior: Deep semantic feature from the raw image is commonly used in diffusion probabilistic models as the condition to mitigate the convergence difficulties. Nevertheless, the subtle changes in local tissues also make a great difference in diagnosing medical images. To this end, we effectively leverage the cropped ROIs from the DCG model to offer fine-grained cues. Specifically, we introduce a Transformer-based image encoder to abstract the deep semantic feature from the raw image by its dynamic and global nature, and a CNN-based image encoder to extract the statistics feature from the ROIs by its static and local nature. Once obtaining the image embedding and ROIs embedding, we fuse them for the image feature prior $\mathcal{F}$ via the learnable parameter $\mathbf{Q}$:

$$\mathcal{F} = \sum \mathbf{Q} \odot [\mathcal{F}_{raw}, \mathcal{F}_{roi}^1, \ldots, \mathcal{F}_{roi}^K], \quad (6)$$

where $\mathcal{F}_{raw}, \mathcal{F}_{roi}^1, \ldots, \mathcal{F}_{roi}^K \in \mathbb{R}^{H\times 1}$, $\mathbf{Q} \in \mathbb{R}^{H\times(K+1)}$, H is the dimension of the latent space, $[\cdot]$ denotes the concatenation operation.

The DCG strategy explores the common priors by constructing the dense guidance map and image feature prior based on the global and local view. These conditional priors effectively help the convergence of the denoising diffusion training.

D. Heterologous Diffusion Process

The traditional diffusion model is trained to denoise unique noise from a perturbed feature instance. The noise ϵ in Eq. (1) is sampled from an i.i.d. Gaussian distribution $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. In order to explore robust representation under the guidance of 2-D dense guidance map and improve the efficiency of diffusion training, we extend the original diffusion process to the heterologous counterpart. This allows various noise points to interact within one feature instance. The details of training

Algorithm 2 Sampling Strategy

```
def sample(images, np, steps, T):
"""
np: number of noise point
steps: sample steps, T: training steps
"""
# generate conditional priors
# dense_map: [b, c, np, np], img_enc: [b, h, 1, 1]
dense_map, img_enc = dcg(images)
noisy_y = normal(0, 1) # [b, c, np, np]
# time intervals
td = T // steps
timesteps = (np.arange(0,steps) * td).round()[::-1]
for t in timesteps:
# predict y_0 from y_T
x = cat(dense_map,noisy_y)
t = t.expand(-1, -1, np, np)
eps_pred = denoising_unet(x, img_enc, t)
# estimate y_t at t
noisy_y = ddim(noisy_y, eps_pred, t)
y_pred = noisy_y.mean(-1).mean(-1)
return y_pred
```

and sampling process are illustrated in Alg. 1 and Alg. 2, respectively.

1) Training Scheme: Specifically, in the training stage (Fig 1(b)), the ground truth $y_0 \in \mathbb{R}^C$ is broadcast to 2-D space, *i.e.*, $y_0 \in \mathbb{R}^{C\times N_p\times N_p}$. Then, the noisy variable $y_{\mathbf{t}} \in \mathbb{R}^{C\times N_p\times N_p}$ is sampled in the diffusion process following:

$$y_{\mathbf{t}} = \sqrt{\bar{\alpha}_{\mathbf{t}}}y_0 + \sqrt{1-\bar{\alpha}_{\mathbf{t}}}\epsilon, \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\epsilon \in \mathbb{R}^{C\times N_p\times N_p}$, $\mathbf{t} \in \mathbb{R}^{1\times N_p\times N_p}$. Each point of $\mathbf{t}$ is independently sampled from the uniform distribution between 0 and the total training timesteps T.

After that, we feed the concatenated map of the noisy variable $y_{\mathbf{t}}$ and dense guidance map $\mathcal{M}$ into our denoising model UNet ϵ_θ to estimate the noise distribution, which is formulated as:

$$\epsilon_\theta(y_{\mathbf{t}}, \mathcal{M}, \mathcal{F}, \mathbf{t}) = D(E(f([y_{\mathbf{t}}, \mathcal{M}]), \mathcal{F}, \mathbf{t}), \mathbf{t}), \tag{8}$$

where $f(\cdot)$ denotes the projection layer to the latent space, specifically a 1×1 convolutional layer. $E(\cdot)$ and $D(\cdot)$ are the encoder and decoder of UNet. Note that the timestep embedding is parallelly extracted in each point of $\mathbf{t}$. In the diffusion forward process, we seek to minimize the noise estimation loss $\mathcal{L}_\epsilon$:

$$\mathcal{L}_\epsilon = ||\epsilon - \epsilon_\theta(y_{\mathbf{t}}, \mathcal{M}, \mathcal{F}, \mathbf{t})||_2^2. \tag{9}$$

Our method improves the vanilla diffusion model by conditioning each step estimation function on priors that combine information derived from the raw image and ROIs.

2) Inference Scheme: As displayed in Fig. 1(c), given an input image x, we first feed it into the DCG model for the dense guidance map and image encoders for the image feature prior. Then, following the pipeline of DDIM [59], in the diffusion reverse process, the trained UNet ϵ_θ generates the final prediction $\hat{y}_0$ by iteratively transforming the noisy variable distribution $p_\theta(y_T)$ to the ground truth distribution $p_\theta(y_0)$ with time intervals:

$$p_\theta(y_{0:T-td}|y_T, \mathcal{M}, \mathcal{F}) = \prod_{t=1}^{T} p_\theta(y_{t-td}|y_t, \mathcal{M}, \mathcal{F}),$$
$$p_\theta(y_T) = \mathcal{N}(\mathcal{M}, \mathbb{I}), \tag{10}$$

where the timestep t is constant across all noise points when inference, td is the time interval between sample steps, $\mathbb{I}$ is the identity matrix.

IV. Experiments

A. Experimental Setup

We evaluate the effectiveness of our network on an in-home dataset and three public datasets, which are PMG2000 [25], HAM10000 [41], APTOS2019 [42] and ChestX-ray14 [34] datasets. We only utilize image data and the disease labels for training across all datasets.

1) PMG2000: We collect and annotate a benchmark dataset for placental maturity grading (PMG) with four categories. PMG2000 is a class-balanced dataset composed of 2,098 B-mode ultrasound images captured by GE Voluson E8 Expert/Phillip EPIQ7 vendors. All images are taken from the anterior wall placenta, and the subjects involved in PMG2000 are pregnant women aged from 18 to 40 weeks. They are taken by ultrasound doctors with more than 5 years of clinical experience to ensure the image quality. we randomly divide the entire dataset into a training part and a testing part at an 8:2 ratio.

2) HAM10000: HAM10000 [41] is a class-imbalanced dataset from the Skin Lesion Analysis Toward Melanoma Detection 2018 challenge, and is publicly available through the ISIC archive. The dataset is a large collection of multi-source dermatoscopic images of common pigmented skin lesions containing 10,015 skin lesion images with predefined 7 categories labeled by human experts.

3) APTOS2019: APTOS2019 [42] is also a class-imbalance dataset with a total of 3,662 fundus images. These images have been labeled to classify diabetic retinopathy into five different grades. Following the same protocol in [70], we split HAM10000 and APTOS2019 into a train part and a test part at a 7:3 ratio.

4) ChestX-ray14: NIH ChestX-ray14 [34] is a significant publicly available chest x-ray dataset consisting of 112,120 frontal-view CXR images from 32,717 patients, many of whom have advanced lung diseases. Each image is labeled with one or multiple pathology keywords, such as atelectasis, cardiomegaly or pneumonia. The dataset includes complicated diseases with potential interrelations, making the classification task challenging. With 14 different labels, the image classification problem involves associating each instance with a subset of those labels, making it a multi-instance, multi-label classification problem. We follow the official data split for a fair comparison.

For PMG2000, HAM10000 and APTOS2019, we introduce two widely-used metrics Accuracy and F1-score to quantitatively compare our framework against existing SOTA methods. For ChestX-ray14, we adopt the area under the receiver operating curve (AUROC) as the metric.

TABLE I

QUANTITATIVE COMPARISON TO STATE-OF-THE-ART METHODS ON PMG2000. THE BEST RESULTS ARE MARKED IN BOLD FONT. ACCURACY AND F1-SCORE ARE ADOPTED AS THE METRICS. *, ** AND *** DENOTE THE SIGNIFICANCE LEVELS OF 0.1, 0.05 AND 0.01, RESPECTIVELY

Methods		ResNet18 [6]	ViT [63]	Swin [64]	PVT [65]	GMIC [61]	DGCMM [48]	UniFormer [66]	DiffMIC [25]	**DiffMIC-v2**
PMG2000	Accuracy	0.879	0.886	0.893	0.907	0.900	0.910	0.915	0.931	**0.935**
	F1-score	0.881	0.890	0.892	0.902	0.901	0.910	0.919	0.926	**0.933**
	p-value	***	***	***	***	***	**	**	*	–

TABLE II

QUANTITATIVE COMPARISON TO STATE-OF-THE-ART METHODS ON HAM10000 AND APTOS2019. THE BEST RESULTS ARE MARKED IN BOLD FONT. ACCURACY AND F1-SCORE ARE ADOPTED AS THE METRICS. *, ** AND *** DENOTE THE SIGNIFICANCE LEVELS OF 0.1, 0.05 AND 0.01, RESPECTIVELY

Methods		LDAM [67]	OHEM [68]	MTL [69]	DANIL [70]	CL [43]	ProCo [44]	DGCMM [48]	UniFormer [66]	DiffMIC [25]	**DiffMIC-v2**
HAM10000	Accuracy	0.857	0.818	0.811	0.825	0.865	0.887	0.886	0.889	0.906	**0.909**
	F1-score	0.734	0.660	0.667	0.674	0.739	0.763	0.794	0.802	0.816	**0.839**
	p-value	***	***	***	***	***	**	**	*	**	–
APTOS2019	Accuracy	0.813	0.813	0.813	0.825	0.825	0.837	0.845	0.847	0.858	**0.871**
	F1-score	0.620	0.631	0.632	0.660	0.652	0.674	0.685	0.690	0.716	**0.721**
	p-value	***	***	***	***	***	***	***	**	**	–

B. Implementation Details

Our framework is implemented with the PyTorch on one NVIDIA RTX 4090 GPU. For pre-processing, we center-crop the image and then resize the spatial resolution of the cropped image to 224×224. Random flipping and rotation for data augmentation are implemented during the training processing. For the first three datasets, we extract six 32×32 ROI patches from each image. For ChestX-ray14, we extract four 112×112 ROI patches from each image. We first pre-trained the DCG model using the batch size of 64 and the Adam optimizer with a learning rate of $2e-4$. The cross-entropy loss is employed on both the global and local predictions for classification tasks. The number of pre-training epochs is set as 100 for the first three datasets and 20 for ChestX-ray14. We utilized the same training sets during pre-training as the training of the diffusion model and did not incorporate additional datasets. Then, we trained our diffusion models together with two image encoders end-to-end using the batch size of 64 and the Adam optimizer. The initial learning rate is set as 1×10^{-3} and decayed to 1×10^{-5} with the cosine scheduler. The number of training epochs is set as 1,000 for the first three datasets and 100 for ChestX-ray14. During the training, we used the diffusion process linearly increasing from $\beta_1 = 0.0001$ to $\beta_T = 0.02$ with $T = 1000$ timesteps. Each point of the timestep $\mathbf{t}$ is randomly selected from a uniform distribution of $[0, T-1]$. During the testing, the inference step is reduced to 10 by deterministic implicit sampling following a standard DDIM training process [59]. To construct image feature prior, we adopt EfficientSAM [71] as the backbone of the Transformer-based encoder to embed the raw image, and adopt ResNet18/50, DenseNet169 as the backbone of the CNN-based encoder to embed the ROIs for the first three datasets, ChestX-ray14, respectively.

C. Comparison With State-of-the-Art Methods

1) Results on PMG2000: In Table I, we compare our DiffMIC-v2 and DiffMIC against many state-of-the-art CNNs and transformer-based networks, including ResNet18 [6], Vision Transformer (ViT) [63], Swin Transformer (Swin) [64], Pyramid Transformer (PVT) [65], UniFormer [66] and one medical image classification method GMIC [61], one conditional generative method DGCMM [48] on PMG2000. Apparently, UniFormer has the largest Accuracy of 0.915, and the largest F1-score of 0.919 among these competing methods. More importantly, our method DiffMIC-v2 further outperforms the compared method UniFormer by improving the Accuracy from 0.915 to 0.935, and the F1-score from 0.919 to 0.933. DiffMIC-v2 also advances our previous work DiffMIC by a margin of 0.004, 0.007 in Accuracy and F1-score.

2) Results on HAM10000 and APTOS2019: Considering the class imbalance issue of both datasets, we compare our DiffMIC-v2 and DiffMIC against state-of-the-art long-tailed medical image classification methods in Table II. These methods include LDAM [67], OHEM [68], MTL [69], DANIL [70], CL [43], ProCo [44]. We also reproduce one transformer-based method UniFormer [66] and one conditional generative method DGCMM [48]. For HAM10000, our method produces a promising improvement over the second-best method UniFormer of 0.020 and 0.037 in terms of Accuracy and F1-score, respectively. DiffMIC-v2, as the extension of DiffMIC, improves its performance over 0.003, 0.023 in Accuracy and F1-score. For APTOS2019, our method obtains a considerable increase over UniFormer of 0.024 and 0.031 in Accuracy and F1-score, respectively. DiffMIC-v2 outperforms DiffMIC by a significant margin of 0.013, 0.005 in Accuracy and F1-score, respectively.

3) Results on ChestX-ray14: Table III illustrates the overall performance of the NIH Chest-Xray14 dataset of our proposed method compared with previous works, the best performance of each pathology is shown in bold. These compared methods only involve image data and labels in training and testing, which keep the same setting as ours. Our method achieves the best average performance of 14 pathologies and obtains a significant improvement of 0.007 in AUROC compared to the second-best method. DiffMIC-v2 improves the diffusion process of DiffMIC and brings a considerable increase of 0.006. The experiments show that our method can not only detect fine-grained lesions such as nodule, but also capture some pathologies that require holistic understanding such as pneumothorax.

TABLE III

AUROC COMPARISONS ON CHESTX-RAY14 DATASET. THE BEST RESULTS ARE MARKED IN BOLD GREEN FONT WHILE THE SECOND-BEST RESULTS ARE UNDERLINED. THE LAST ROW IS THE SIGNIFICANCE TEST RESULTS (P-VALUE). *, ** AND *** DENOTE THE SIGNIFICANCE LEVELS OF 0.1, 0.05 AND 0.01, RESPECTIVELY

Pathology	Wang *et al.* [34]	Yao *et al.* [72]	Li *et al.* [73]	Kumar *et al.* [36]	Tang *et al.* [74]	Shen *et al.* [75]	Mao *et al.* [76]	Guan *et al.* [38]	Liu *et al.* [77]	Zhang *et al.* [78]	Xiao *et al.* [79]	DiffMIC [25]	**DiffMIC-v2** (Ours)
Atelectasis	0.716	0.772	**0.800**	0.762	0.756	0.766	0.750	0.781	0.773	0.778	0.731	0.779	0.788
Cardiomegaly	0.807	0.904	0.870	**0.913**	0.887	0.801	0.869	0.883	0.889	0.890	0.885	0.871	0.880
Effusion	0.784	0.859	**0.870**	0.864	0.819	0.797	0.810	0.831	0.821	0.833	0.810	0.829	0.831
Infiltration	0.609	0.695	0.700	0.692	0.689	**0.751**	0.687	0.697	0.710	0.695	0.687	0.691	0.685
Mass	0.706	0.792	0.830	0.750	0.814	0.760	0.782	0.830	0.829	0.784	0.788	0.835	**0.837**
Nodule	0.671	0.717	0.750	0.666	0.755	0.741	0.726	0.764	0.770	0.776	0.750	0.779	**0.783**
Pneumonia	0.633	0.713	0.670	0.715	0.729	**0.778**	0.695	0.725	0.713	0.765	0.742	0.728	0.742
Pneumothorax	0.806	0.841	0.870	0.859	0.850	0.800	0.845	0.866	0.869	0.841	0.844	0.860	**0.871**
Consolidation	0.708	0.788	**0.800**	0.784	0.728	0.787	0.728	0.758	0.749	0.781	0.775	0.760	0.762
Edema	0.835	0.882	0.880	**0.888**	0.848	0.820	0.834	0.853	0.847	0.831	0.829	0.851	0.857
Emphysema	0.815	0.829	0.910	0.898	0.906	0.773	0.870	0.911	**0.934**	0.925	0.901	0.923	0.926
Fibrosis	0.769	0.767	0.780	0.756	0.818	0.765	0.798	0.826	0.845	0.835	0.811	0.824	**0.849**
Pleural_Thickening	0.708	0.765	0.760	0.774	0.765	0.759	0.758	0.780	0.773	0.768	0.750	**0.790**	0.785
Hernia	0.767	0.914	0.770	0.802	0.875	0.748	0.877	0.918	0.925	0.924	0.912	0.946	**0.956**
Average	0.738	0.803	0.804	0.794	0.803	0.775	0.788	0.816	0.818	0.816	0.801	0.819	**0.825**
p-value	***	**	*	**	***	***	***	***	**	*	***	***	–

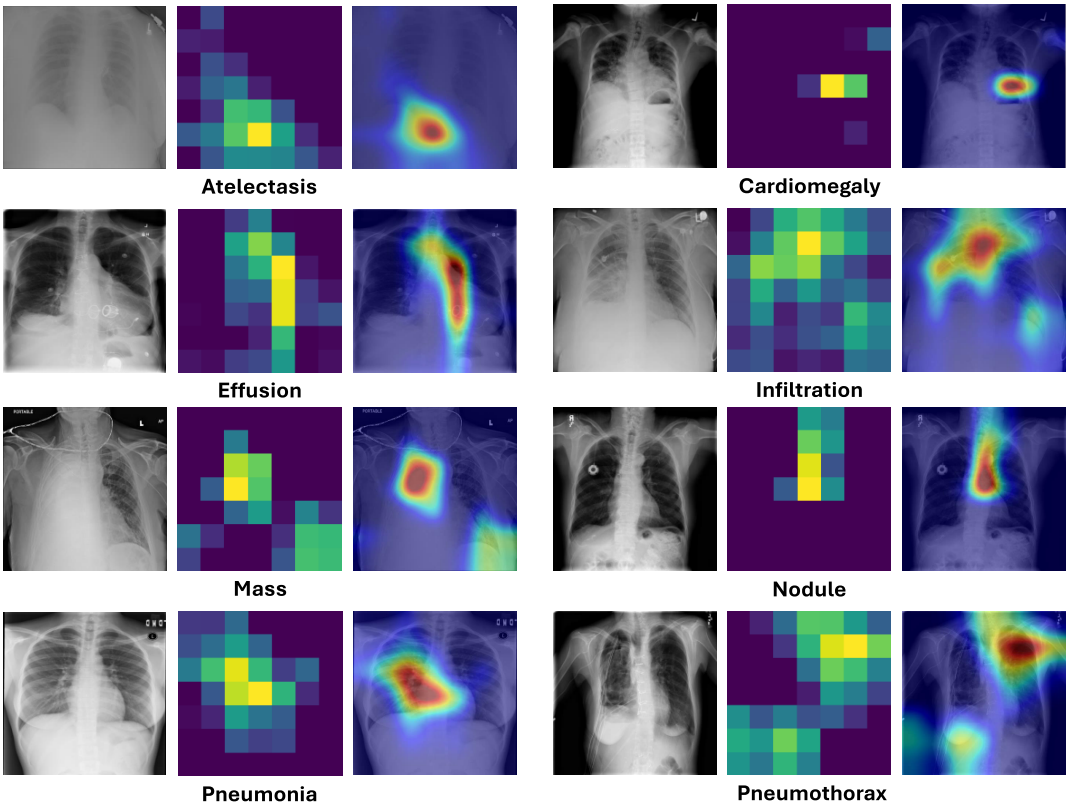

Fig. 3. Eight chest radiographs with different pathologies and the corresponding attention heat maps are visualized by Grad-CAM. Red: Most indicative pathology regions with high confidence, Yellow: pathology regions with low confidence, Blue: Regions without abnormalities. As observed, our method is significantly aware of both large masses and tiny nodules.

Fig. 3 visualizes the weighted heat maps of our proposed method for some tough cases to validate its effectiveness. For instance, (a) displays one irregular case with low contrast, and our DiffMIC-v2 can capture the right pathology region. (e),(f) and (g) demonstrate our method is also robust to letters and medical devices on images.

4) Statistical Analysis: To statistically evaluate the performance improvements of our method, we conducted paired t-tests between our approach and each compared method, as shown in the last rows of Tables I, II, and III. The significance levels in our t-test analysis are categorized into three tiers following [40]. The highest significance level (***) represents a p-value of less than 0.01, corresponding to a confidence level of 99%. The second level (**) indicates a p-value of less than 0.05, with a confidence level of 95%, while the third level (*) signifies a p-value of less than 0.1, with a confidence level of 90%. For placental maturity grading and diabetic retinopathy grading, the significance level between DiffMIC-v2 and the best-compared method, UniFormer [66], is 0.05 for both accuracy and F1-score, while for skin lesion classification, it is 0.1. In the ChestX-ray14 classification task, the significance level between DiffMIC-v2

TABLE IV
QUANLITATIVE EVALUATION ON EACH COMPONENT OF OUR DIFFMIC-V2 ON THE PMG2000 DATASET. "DIFFUSION" DENOTES VANILLA DIFFUSION PROCESS, "GP" DENOTES THE GLOBAL PRIOR FROM THE RAW IMAGE, "LP" DENOTES THE LOCAL PRIOR FROM THE ROIS, "DCG" DENOTES DUAL-GRANULARITY CONDITIONAL GUIDANCE STRATEGY, "$\mathcal{F}$" DENOTES THE IMAGE FEATURE PRIOR, "$\mathcal{M}$" DENOTES DENSE GUIDANCE MAP AND "HETER. DIFF." DENOTES HETEROLOGOUS DIFFUSION PROCESS

Combinations	Diffusion	GP	LP	DCG		Heter. Diff.	Metrics	
				$\mathcal{F}$	$\mathcal{M}$		Accuracy	F1-score
C1	–	–	–	–	–	–	0.879	0.881
C2	✓	–	–	–	–	–	0.906	0.899
C3	✓	✓	–	–	–	–	0.915	0.910
C4	✓	–	✓	–	–	–	0.917	0.912
C5	✓	–	–	✓	–	–	0.913	0.905
C6	✓	–	–	✓	✓	–	0.927	0.925
Our method	✓	–	–	✓	✓	✓	**0.935**	**0.933**

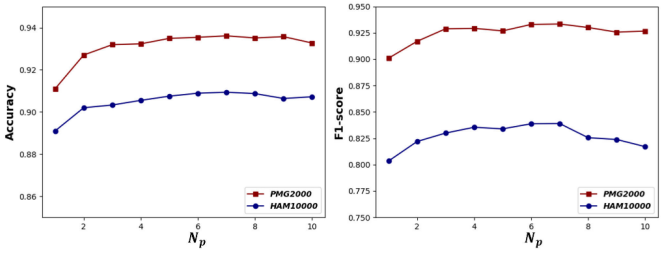

Fig. 4. The Analysis of the number of noise points N_p in one feature instance related to our Heterologous diffusion process.

and the top-performing method by Liu et al. [77] is 0.05 in AUROC. These results indicate that our method demonstrates statistically significant improvements over other methods, highlighting its superior performance. Additionally, the t-tests between DiffMIC-v2 and its predecessor, DiffMIC, further confirm the effectiveness of our proposed enhancements.

D. Ablation Study

1) Major Module: Extensive experiments are conducted to evaluate the effectiveness of major modules of our network from Accuracy and F1-score. To do so, we build six baseline networks from our method. The first baseline (denoted as "C1") is to remove all diffusion operations from our network, which means that "C1" is equal to the traditional ResNet18 network. We apply the vanilla diffusion process onto "C1" to construct another baseline network (denoted as "C2") with only traditional raw image feature embedding as guidance. Then, we consider the global prior from the raw image and the local prior from the ROIs as the extra condition to guide the training of diffusion models in "C3" and "C4", respectively. On the other hand, we add our dual-granularity conditional guidance into the diffusion process to build two baseline networks, which are denoted as "C5" and "C6". More specifically, "C5" introduces our image feature prior $\mathcal{F}$ by replacing the raw image embedding in "C2" as the single condition while "C6" further constructs the dense guidance map $\mathcal{M}$ from the global and local prior to condition the vanilla diffusion process together with $\mathcal{F}$. Hence, "C6" is equal to degenerating the heterologous diffusion process into the vanilla counterpart in our full method for image classification.

Table IV reports our method's Accuracy and F1-score results and six baseline networks on our PMG2000 dataset. Apparently, compared to "C1", "C2" has an Accuracy improvement of 0.027 and an F1-score improvement of 0.018, which indicates that the diffusion mechanism under the guidance of vanilla image feature embedding can learn more discriminate semantics for medical image classification, thereby improving the PMG performance. The improvements in Accuracy and F1-score of "C3" and "C4" indicate that both the global and local prior contribute greatly to the training of diffusion-based classification. Simultaneously, the better Accuracy and F1-score results of "C5" over "C2" demonstrate that leveraging our image feature prior $\mathcal{F}$ as the condition to guide the vanilla diffusion process can benefit the PMG performance. "C6" introduces the 2D dense guidance map $\mathcal{M}$ and significantly advances the performance of "C5" by the improvement of 0.014,0.020 in Accuracy, F1-score, respectively. Finally, our full method outperforms "C6" in terms of both Accuracy and F1-score, which indicates that our heterologous diffusion process can effectively help to enhance the PMG results by exploring the common information of various noise points in one feature instance.

2) Hyperparameter Optimization: We evaluate our model's sensitivity to the number of noise points N_p, which is a critical hyper-parameter in our designed heterologous diffusion process. Fig 4 displays the results of Accuracy and F1-score on PMG2000 and HAM10000 datasets when N_p varied from 1 to 10. As observed, the performance is usually stable while the number of noise points is increasing, especially for [4, 7]. Considering both performance and computational costs, we set N_p as 6.

We also evaluate our model's sensitivity to the number of regions of interest (ROIs), denoted as K, a crucial hyper-parameter in the DCG model. Table V presents the results for Accuracy and F1-score on the PMG2000 dataset and AUROC on the ChestX-ray14 dataset as K varies from 2 to 10. As observed, the model's performance remains generally stable as the number of ROIs increases, particularly in the range of [4, 8]. Based on these results, we set K to 6 for PMG2000 and 4 for ChestX-ray14.

Additionally, we optimize the dimension of the latent space (H). As shown in Table V, increasing H enhances the denoising ability of latent diffusion models but also leads to a higher parameter count. Balancing performance and computational efficiency, we select 6144 as the optimal value for H.

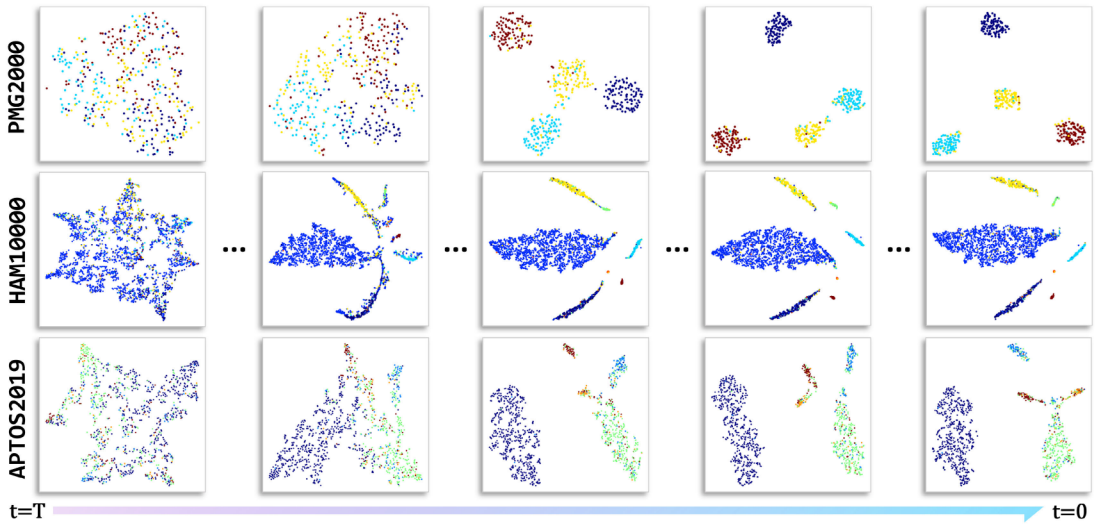

Fig. 5. t-SNE obtained from the denoised feature embedding by the diffusion reverse process during inference on three datasets. As the time step encoding progresses, the noise is gradually removed, thereby obtaining a clear distribution of classes.

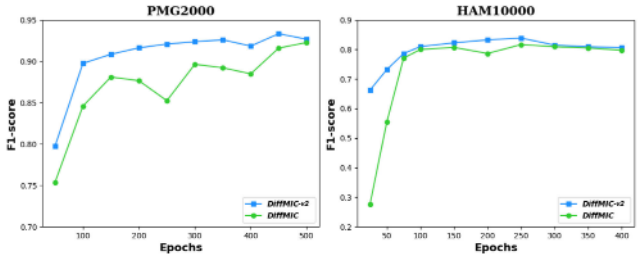

Fig. 6. The convergence efficiency of DiffMIC-v2 *vs.* DiffMIC during training.

TABLE V
THE ANALYSIS OF THE NUMBER OF ROIS (K) AND THE DIMENSION OF THE LATENT SPACE (H) ON PMG2000 AND CHESTX-RAY14 DATASETS. THE BEST RESULTS ARE MARKED IN BOLD FONT

K		2	4	6	8	10
PMG2000	Accuracy	0.928	0.934	**0.935**	0.934	0.930
	F1-score	0.925	0.932	**0.933**	0.933	0.930
ChestX-ray14	AUROC	0.824	**0.825**	0.824	0.824	0.822

H		2048	4096	6144	8192
PMG2000	Accuracy	0.920	0.924	**0.935**	0.934
	F1-score	0.919	0.925	0.933	**0.935**
ChestX-ray14	AUROC	0.820	0.822	**0.825**	0.825

3) Efficiency of DiffMIC-v2 Vs. DiffMIC: In this part, we compare the efficiency of DiffMIC-v2 against our previous work DiffMIC. Fig. 6 visualizes the performance of F1-score on PMG2000 and HAM10000 datasets with the number of epochs increasing during training. The results prove that DiffMIC-v2 has greater convergence efficiency by taking fewer iterations to achieve better performance than DiffMIC across two datasets. Table VI further compares their computational costs in view of FLOPs and the average run-time during inference. By adopting fewer sampling steps, DiffMIC-v2 is about 3× cheaper in FLOPs and faster in run-time than DiffMIC to predict an image with the patch size of 224 × 224.

4) Visualization of Our Diffusion Procedure: To qualitatively illustrate the diffusion reverse process, we used the t-SNE tool to visualize the denoised feature embeddings at consecutive time steps. Figure 5 presents the results of this process on all three datasets. As the time step encoding progresses, the denoise diffusion model gradually removes noise from the feature representation, resulting in a clearer distribution of classes from the pure Gaussian distribution.

TABLE VI
COMPARISON OF FLOPS AND RUN-TIME ON AN IMAGE WITH THE SIZE OF 224 × 224 DURING INFERENCE

Method	DiffMIC	DiffMIC-v2
FLOPs/G	2.27×100+ steps	7.65×10 steps
Run-time/s	0.056	0.021

V. DISCUSSION

Latent Diffusion Models (LDMs) present an innovative approach to developing robust representations for medical image classification. By modeling data distribution in a latent space rather than directly on pixel space, LDMs effectively capture complex patterns, which is particularly beneficial in medical imaging where noise and variability (such as inter-patient differences) are common. This approach enables more accurate classification of diseases or conditions that exhibit subtle variations across images. Medical imaging often faces challenges from artifacts, noise, and variations introduced by acquisition devices, which can significantly impact classification performance. LDMs excel at learning to remove noise in the latent space, thereby enhancing the model's robustness to these imperfections. This capability is especially valuable in scenarios involving low-quality or incomplete scans, where traditional models may struggle. DiffMIC-v2 introduces an improved diffusion process that allows multiple noise points to interact within a single feature instance. The key distinction between our heterologous diffusion process and conventional methods is that each noise point in the 2D map is assigned different sampled timesteps during training. Originally varying independently across multiple iterations, these noise points are aggregated into one 2D map. We use consecutive 2D

convolutional layers in the Denoising U-Net to explore correlations among diverse noise points, which enhances the convergence efficiency of the diffusion models.

Our proposed dual-granularity conditional guidance in LDMs involves leveraging both global and local views of medical images to boost classification accuracy. This approach mirrors the radiologists' workflow: first, they assess the overall context of the image, and then focus on specific areas of interest. By incorporating this dual-granularity method, our model aligns closely with the radiologists' analytical process, providing a natural extension of their methods into the deep network. This enhances both the interpretability and reliability of the model.

Currently, in DiffMIC-v2, we formulate the dense guidance map and image feature prior as conditions for LDMs, both derived from image data. Future work could explore the introduction of clinical textual data as additional conditions. Integrating textual data with imaging data could enrich the model's contextual understanding, thereby improving classification accuracy. Conditional LDMs could adjust their outputs based on specific textual inputs during both generation and inference processes, enhancing model controllability. In medical scenarios, this would allow models to generate diagnostic results that align with specific clinical symptoms or disease descriptions. Providing text-based conditions would enable the model to produce image features that meet clinical needs, assisting doctors in understanding the likelihood of certain pathologies and improving result interpretability.

VI. Conclusion

In this work, we present an improved diffusion network (DiffMIC-v2) to boost generic medical image classification. The main idea of our DiffMIC-v2 is to introduce Dual-granularity Conditional Guidance over our proposed Heterologous Diffusion Process. The dense guidance map and image feature prior from the global and local views offer a significant help to the fast convergence of diffusion models. Our Heterologous Diffusion Process further advances the vanilla counterpart by providing the occurrence of multiple noise points. Experimental results on four medical image classification datasets of diverse image modalities exhibit the superior performance of our network over state-of-the-art methods. These demonstrate that DiffMIC-v2 can robustly tackle the common issues in general 2D medical image classification, such as class imbalance, interclass similarity and complex symbiotic relationships between diseases or severities. DiffMIC-v2 also develops our previous work DiffMIC in view of both performance and computational costs, making it an essential baseline for future research in this area.

References

[1] M. De Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," *Med. Image Anal.*, vol. 33, pp. 94–97, Oct. 2016.

[2] C. Varela, S. Timp, and N. Karssemeijer, "Use of border information in the classification of mammographic masses," *Phys. Med. Biol.*, vol. 51, no. 2, pp. 425–441, Jan. 2006.

[3] Y. Song, W. Cai, Y. Zhou, and D. D. Feng, "Feature-based image patch approximation for lung tissue classification," *IEEE Trans. Med. Imag.*, vol. 32, no. 4, pp. 797–808, Apr. 2013.

[4] S. Koitka and C. M. Friedrich, "Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016," in *Proc. CLEF (Working Notes)*, Jan. 2016, pp. 304–317.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, May 2012, pp. 1–9.

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

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

[8] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[9] A. Esteva et al., "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[10] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Med.*, vol. 28, no. 1, pp. 31–38, Jan. 2022.

[11] F. Shamshad et al., "Transformers in medical imaging: A survey," 2022, *arXiv:2201.09873*.

[12] Y. Yang, S. Wang, L. Zhu, and L. Yu, "HCDG: A hierarchical consistency framework for domain generalization on medical image segmentation," 2021, *arXiv:2109.05742*.

[13] Y. Yang, A. I. Aviles-Rivero, H. Fu, Y. Liu, W. Wang, and L. Zhu, "Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13200–13210.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[15] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 139, M. Meila and T. Zhang, Eds., Jul. 2021, pp. 8162–8171.

[16] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," 2021, *arXiv:2111.13606*.

[17] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8780–8794.

[18] J. Singh, S. Gould, and L. Zheng, "High-fidelity guided image synthesis with latent diffusion models," 2022, *arXiv:2211.17084*.

[19] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, "SegDiff: Image segmentation with diffusion probabilistic models," 2021, *arXiv:2112.00390*.

[20] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *Proc. Int. Conf. Med. Imag. With Deep Learn.*, 2022, pp. 1336–1348.

[21] S. Chen, P. Sun, Y. Song, and P. Luo, "DiffusionDet: Diffusion model for object detection," 2022, *arXiv:2211.09788*.

[22] X. Han, H. Zheng, and M. Zhou, "CARD: Classification and regression diffusion models," 2022, *arXiv:2206.07275*.

[23] Z. Xing, L. Wan, H. Fu, G. Yang, and L. Zhu, "Diff-UNet: A diffusion embedded network for volumetric segmentation," 2023, *arXiv:2303.10326*.

[24] H. Chen et al., "Robust classification via a single diffusion model," 2023, *arXiv:2305.15241*.

[25] Y. Yang, H. Fu, A. I. Aviles-Rivero, C.-B. Schönlieb, and L. Zhu, "DiffMIC: Dual-guidance diffusion network for medical image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2023, pp. 95–105.

[26] Y. Song, W. Cai, Y. Wang, and D. D. Feng, "Location classification of lung nodules with optimized graph construction," in *Proc. 9th IEEE Int. Symp. Biomed. Imag. (ISBI)*, May 2012, pp. 1439–1442.

[27] S. Ali et al., "An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion," 2024, *arXiv:2401.15753*.

[28] Y. Yang et al., "MammoDG: Generalisable deep learning breaks the limits of cross-domain multi-center breast cancer screening," 2023, *arXiv:2308.01057*.

[29] Y. Yang, Z. Xing, L. Yu, C. Huang, H. Fu, and L. Zhu, "Vivim: A video vision mamba for medical video segmentation," 2024, *arXiv:2401.14168*.

[30] J. Zhang, Y. Xia, Y. Xie, M. Fulham, and D. D. Feng, "Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1521–1530, Sep. 2018.

[31] B. Lei et al., "Hybrid descriptor for placental maturity grading," *Multimedia Tools Appl.*, vol. 79, nos. 29–30, pp. 21223–21239, Aug. 2020.

[32] A. Esteva et al., "Correction: Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 546, no. 7660, p. 686, 2017.

[33] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Med. Image Anal.*, vol. 54, pp. 10–19, May 2019.

[34] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[35] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[36] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded ConvNets for multilabel classification of thoracic diseases in chest radiographs," in *Proc. 15th Int. Conf. Image Anal. Recognit. (ICIAR)*, Póvoa de Varzim, Portugal. Cham, Switzerland: Springer, Jun. 2018, pp. 546–552.

[37] S. Guendel et al., "Learning to recognize abnormalities in chest X-rays with location-aware dense networks," in *Proc. 23rd Iberoamerican Congr. Prog. Pattern Recognit., Image Anal., Comput. Vis., Appl.*, Madrid, Spain. Cham, Switzerland: Springer, Mar. 2018, pp. 757–765.

[38] Q. Guan and Y. Huang, "Multi-label chest X-ray image classification via category-wise residual attention learning," *Pattern Recognit. Lett.*, vol. 130, pp. 259–266, Feb. 2020.

[39] X. Gong, X. Xia, W. Zhu, B. Zhang, D. Doermann, and L. Zhuo, "Deformable Gabor feature networks for biomedical image classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 4004–4012.

[40] C. Mao, L. Yao, and Y. Luo, "ImageGCN: Multi-relational image graph convolutional networks for disease identification with chest X-rays," *IEEE Trans. Med. Imag.*, vol. 41, no. 8, pp. 1990–2003, Aug. 2022.

[41] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Aug. 2018.

[42] S. D. Karthik. (2019). *Aptos 2019 Blindness Detection*. Maggie. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection

[43] Y. Marrakchi, O. Makansi, and T. Brox, "Fighting class imbalance with contrastive learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 466–476.

[44] Z. Yang et al., "ProCo: Prototype-aware contrastive learning for long-tailed medical image classification," in *Proc. MICCAI*. Cham, Switzerland: Springer, Jan. 2022, pp. 173–182.

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

[46] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–10.

[47] T. Guo, X. Zhu, Y. Wang, and F. Chen, "Discriminative sample generation for deep imbalanced learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2406–2412.

[48] X. Wang et al., "Deep generative mixture model for robust imbalance classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2897–2912, Mar. 2023.

[49] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project Stanford CS231N, Convolutional Neural Netw. Vis. Recognit., Winter Semester*, vol. 2014, no. 5, pp. 1–9, 2014.

[50] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Exp. Syst. Appl.*, vol. 91, pp. 464–471, Jan. 2018.

[51] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu, "DermGAN: Synthetic generation of clinical skin images with pathology," in *Proc. Mach. Learn. Health Workshop*, Jan. 2019, pp. 155–170.

[52] T. Zhang et al., "Noise adaptation generative adversarial network for medical image analysis," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1149–1159, Apr. 2019.

[53] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "BAGAN: Data augmentation with balancing GAN," 2018, *arXiv:1803.09655*.

[54] S. Santurkar, L. Schmidt, and A. Madry, "A classification-based study of covariate shift in GAN distributions," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4480–4489.

[55] S. S. Mullick, S. Datta, and S. Das, "Generative adversarial minority oversampling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1695–1704.

[56] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.

[57] X. Liu et al., "More control for free! Image synthesis with semantic diffusion guidance," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 289–299.

[58] Y. Yang, H. Wu, A. I. Aviles-Rivero, Y. Zhang, J. Qin, and L. Zhu, "Genuine knowledge from practice: Diffusion test-time adaptation for video adverse weather removal," 2024, *arXiv:2403.07684*.

[59] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.

[60] Y. Lu, Y. Yang, Z. Xing, Q. Wang, and L. Zhu, "Diff-VPS: Video polyp segmentation via a multi-task diffusion network with adversarial temporal reasoning," 2024, *arXiv:2409.07238*.

[61] Y. Shen et al., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101908.

[62] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

[63] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[64] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[65] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[66] K. Li et al., "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023.

[67] K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019, pp. 1–16.

[68] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[69] H. Liao and J. Luo, "A deep multi-task learning approach to skin lesion classification," 2018, *arXiv:1812.03527*.

[70] L. Gong, K. Ma, and Y. Zheng, "Distractor-aware neuron intrinsic learning for generic 2D medical image classifications," in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent.* Cham, Switzerland: Springer, Jan. 2020, pp. 591–601.

[71] Y. Xiong et al., "EfficientSAM: Leveraged masked image pretraining for efficient segment anything," 2023, *arXiv:2312.00863*.

[72] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*.

[73] Z. Li et al., "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8290–8299.

[74] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," in *Proc. 9th Int. Workshop Mach. Learn. Med. Imag.*, Granada, Spain. Cham, Switzerland: Springer, Sep. 2018, pp. 249–258.

[75] Y. Shen and M. Gao, "Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization," in *Proc. 9th Int. Workshop Mach. Learn. Med. Imag.*, Granada, Spain. Cham, Switzerland: Springer, Sep. 2018, pp. 389–397.

[76] C. Mao, L. Yao, Y. Pan, Y. Luo, and Z. Zeng, "Deep generative classifiers for thoracic disease diagnosis with chest X-ray images," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 1209–1214.

[77] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3429–3440, Nov. 2020.

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

[79] J. Xiao et al., "Multi-label chest X-ray image classification with single positive labels," *IEEE Trans. Med. Imag.*, vol. 43, no. 12, pp. 4404–4418, Dec. 2024.