Prototype-Driven Class-Conditional Synthesis for High-Quality Chest X-ray Image Generation

Bowen Guo[1], Peng Huang[1], Yuanyuan Wang[12*], Yi Guo[12*]
[1]Department of Biomedical Engineering, School of Information Science and Technology, Fudan University, Shanghai, China
[2]Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, Shanghai, China

***Abstract*—As an advanced image argumentation approach, image generation technology offers a novel solution to the challenges of data scarcity and distribution imbalance in the medical field. However, the severe imbalance in the class distribution causes the networks to overfit to the head classes, while failing to adequately model the distribution of the tail class data during image generation, ultimately compromising the quality of the generated images. To solve this problem, we propose a Class Prototype-Driven Diffusion Model (CPDM) to improve class-conditional image synthesis on long-tailed chest X-ray images datasets. To fully extract the features of limited tail classes while avoiding overfitting to head classes, we introduce a Class Prototype Bank, which stores representative feature vectors of each class. Furthermore, by integrating cross-attention mechanisms between image features and class-specific prototypes, CPDM effectively captures fine-grained class features, enhancing both the realism and diversity of the generated images. Experiments show that our CPDM achieves the lowest FID=31.600 and highest IS=2.842, highlighting the effectiveness of CPDM in mitigating class imbalance and data scarcity in chest X-ray imaging. In downstream experiments, the classifier achieves a 17.22% improvement on the mAUC for 14 thoracic diseases when trained on a mixed dataset containing only 1% real images.**

***Keywords*-Medical Image Synthesis, Prototype-Driven, Class-Conditional Generation, Diffusion Model**

I. INTRODUCTION

Recently, deep learning has made significant strides in various auxiliary diagnostic tasks in medical imaging [1][2], yet the development of robust models remains challenged by limited access to large, high-quality datasets. By generating synthetic samples, image generation models as an advanced data augmentation technique, offer a promising solution to address data scarcity in medical imaging, particularly in chest X-ray images.

Specifically, Generative Adversarial Networks (GANs [3]) and Diffusion Models have shown great potential in images generation and reconstructing data distribution. However, GANs face inherent architectural challenges, including mode collapse and training instability [4], which hinder their ability to accurately model the complex nature of real data distributions, particularly for long-tailed class.

Diffusion Models like the Denoising Diffusion Probabilistic Models [5] (DDPM) and Stable Diffusion (SD) [6] have garnered significant attention in natural images generation. These likelihood-based models have been proven to outperform GANs with regard to both the stability of the training process and the ability to capture a wide range of data distributions [7]. The emergence of the Classifier-Free Guidance [8] has endowed diffusion models with the ability to generate images conditioned on specific classes, greatly enhancing their potential for data augmentation and improving class distribution balance in certain datasets. Due to the prevalent long-tailed distribution characteristics of existing medical datasets, currently the performance of class-conditional diffusion models is significantly constrained[9]. Overfitting to the head class data distribution and insufficient learning of the tail classes severely degrade the fidelity and diversity of generated images, resulting in a substantial reduction in the controllability of the generation process compared to natural image generation.

In this paper, we propose a Class Prototype-Driven Diffusion Model (CPDM) to enhance the overall quality of generated images including those from tail classes, with the aim of addressing the challenges of dataset scarcity and class imbalance. To maximize the utilization of existing tail-class data, we construct a class prototype bank, where each prototype is linked to the image features of its corresponding class. This approach strengthens the coupling between class information and generated images during the synthesis process. By incorporating cross-attention mechanisms between image features and class-specific prototypes, CPDM effectively enhances class-conditional image synthesis, capturing fine-grained class-specific characteristics. This approach not only improves the fidelity and diversity of generated images but also ensures alignment with the true class distributions through a class-reconstruction loss. Our proposed model demonstrates promising results in both image generation quality and downstream classification tasks, offering a practical solution for data augmentation in the medical imaging domain.

* Corresponding authors (guoyi@fudan.edu.cn, yywang@fudan.edu.cn)

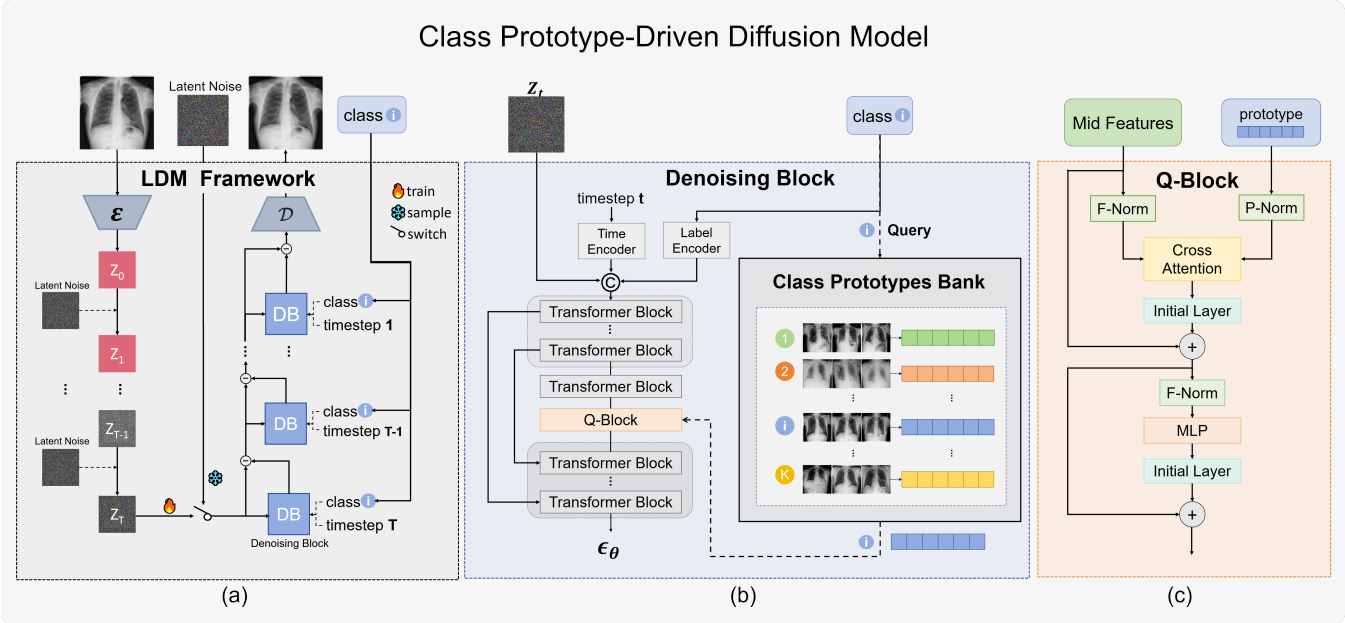


Figure 1. The overview of CPDM: Part (a) shows the framework of LDM, part (b) illustrates the Denoising Block utilized in the LDM and part (c) shows the structure of Q-Block

II. Methodology

Fig. 1 is the overview of our proposed CPDM. (a) presents the fundamental framework of The Latent Diffusion Model (LDM) training and sampling. (b) elaborates on the noise prediction module designed in our CPDM model, with the key highlights being the construction of the Class Prototype Bank (CPB) and the incorporation of the Q-Block, which integrates image-category information. (c) provides a detailed depiction of the structure of the Q-Block.

A. Latent Diffusion Models

LDM transferring the diffusion process to a lower-dimensional latent space by utilizing a pretrained autoencoder to capture latent features z_t, as demonstrated in Stable Diffusion(SD) [7]. In the latent space Z, this approach operates in two phases: forward diffusion and reverse denoising demonstrated in the Fig. 1 (a).

The forward process gradually introduces noise to a latent representation z_t over T timesteps, transforming the latent feature of a real image into pure Gaussian noise. This process is mathematically formulated as a Markov chain:

$$q(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}\left(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t \boldsymbol{I}\right) \quad (1)$$

where z_t denotes the latent representation of noisy image at step t, and β_t governs the variance of the added at each step.

In the reverse diffusion process, the model initiates from pure Gaussian noise and progressively refines the signal through iterative noise prediction and subtraction at each timestep, leveraging the shared denoising block architecture, denoted as (2). This process ultimately yields a clean latent feature representation, which is subsequently decoded into high-quality output images.

$$p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t-1}; \mu_\theta(\boldsymbol{z}_t, t), \sigma_t^2 \boldsymbol{I}) \quad (2)$$

Classifier-Free Guidance(CFG)[8] is employed to enhance the controllability of the generation process. The objective of the training process is to minimize the discrepancy between the actual data distribution and the samples generated by the model through a denoising score matching loss:

$$\mathcal{L}_{\text{denoise}} = E_{x_0,t,\epsilon}[||\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})||^2] \quad (3)$$

where ϵ is the actual noise added during the forward process and $\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$ is the denoise block's prediction.

Nonetheless, medical imaging such as chest X-ray images datasets presents unique challenges for conditional generation, including complex foreground-background relationships, heterogeneous manifestations of the same disease, and similar imaging characteristics across different pathologies. These complexities cannot be adequately addressed by label vectors alone and require more sophisticated guidance mechanisms.

B. Class Prototype Bank

To fully leverage the guidance of input class information for image generation, we aim to strengthen the alignment between class information and its corresponding typical image features over all categories. To achieve this, we introduce a Class Prototype Bank (CPB), a specialized module designed to enhance feature-class coupling. As demonstrated in the part (b) of Fig. 1, The core idea of CPB is to constructs a class-specific prototype bank in a high-dimensional feature space, denotes as $P_c \in R^{C\times(2L)\times D}$, here C represents the number of classes while L denotes the number of latent image patches and D corresponds to the embedding dimension of the latent features. Each class prototype serves as a representative feature vector, encapsulating the distinctive characteristics of images within that class. This

design enables the retrieval and alignment of image features from the CPB:

$$P_y = P_c[y] \tag{4}$$

ensuring that generated images are closely aligned with their corresponding class information. For the training of the CPB, we compute the reconstruction loss by measuring the difference between the prototype-based reconstructed features and the original image features as follows:

$$L_{\text{recon}} = \frac{1}{2}||\hat{x} - x_{\text{clean}}||_2^2 \tag{5}$$

where the $\hat{x}$ represent the reconstruction feature based on the given class and x_{clean} denotes the original latent representation of the image.

C. Q-Block and Prototype-Image Cross-Attention

Once the prototypes are retrieved, they guide the generation of class-specific images by modulating the image synthesis process using the extracted representative features. Additionally, specifically focusing on optimizing tail-class performance, we integrate a Q-block module in the mid-level layers of the U-ViT network. This integration facilitates more effective interaction between tail class information and corresponding image features, thereby improving the model's ability to generate class-consistent images particularly for categories with limited training samples. The Q-Block incorporates a Class Prototype-Image Cross-Attention Mechanism, as illustrated in Fig. 1(c), which establishes dynamic interactions between category-specific prototypes and visual features through the cross-attention paradigm defined in (6) and (7):

$$\textbf{Query}: Q_X = \hat{X}W_Q, \textbf{Key}: K_Y = P_yW_K, \textbf{Value}: V_y = P_yW_V \tag{6}$$

where the $W_Q, W_K, W_V \in R^{D\times D}$ represent the learnable query, key and value matrix respectively. $X \in R^{B\times L\times D}$ represents the latent features of the input image and $\hat{X} = LayerNorm(X)$, B is the batch size.

$$CrossAttention(Q_X, K_Y, V_y) = softmax\left(\frac{Q_X {K_Y}^T}{\sqrt{D}}\right)V_y \tag{7}$$

Building upon the "question-answer" relationship between class prototypes and image features, the Q-block employs cross-attention to dynamically focus on features most relevant to the target class. This mechanism facilitates class-conditioned image generation by emphasizing local tail-class characteristics, thereby producing high-fidelity and diverse images with enhanced representation of underrepresented categories. Following the cross-attention, the MLP layer processes the features further, enhancing non-linear representations. At the same time, we introduce zero-initialized initial layers to avoid interference from the uncertain initial state of the CPB.

III. Experimental Setup

A. Datasets and Preprocessing

ChestX-ray14 [10] comprises 112,120 frontal-view X-ray images from 30,805 patients, annotated with 14 disease labels. Following the official data split, we use 75,312 images for training and 25,596 images for testing. To better align with our experimental objectives in the long-tailed distribution of single-label dataset, we apply label filtering to retain only single-label samples, resulting in a refined dataset of 64,352 training images and 9,119 test images. The distribution of images across each category in the training set is detailed in Fig. 2, which clearly illustrates the feature of long-tailed distribution. For computational efficiency, all images are resized from their original resolution of 1024×1024 to 256×256.

B. Implement Detail

For the hyperparameter of our model, the latent dimension of images is (4, 32, 32), with a batch size of 512 and an embedding dimension of 1024 and the classifier-free guidance parameter p_{cond} is set to 0.15. We use the AdamW optimizer with the Learning rate of 1e-5 and the weight decay of 0.03. All components of the proposed model are implemented with PyTorch and trained on 2 Nvidia A100 80GB GPUs.

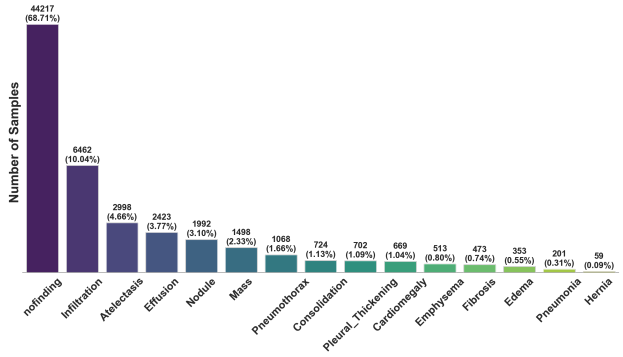


Figure 2. Category Distribution in ChestX-ray14 (Preprocessed)

C. Experiment Detail

In our comparative experiments, we evaluate our model against several state-of-the-art (SOTA) models in the class-conditional image generation domain, including both GANs and Diffusion Models. To assess image fidelity and diversity, we employ four widely-used evaluation metrics: Fréchet Inception Distance (FID), Inception Score (IS), Precision, and Recall. To rigorously assess our model's capability in generating underrepresented (tail-class) samples, we conduct additional evaluations on tail class data, focusing on four tail categories including Fibrosis(0.74%), Edema(0.55%), Pneumonia(0.31%) and Hernia(0.05%).

To further demonstrate the practical relevance and value of the generated images, we conduct downstream task experiments by mixing the generated data with varying proportions of the original dataset (1%, 10%, and 100%) to create hybrid training sets. These hybrid datasets are utilized to train deep learning classifiers for multi-label thoracic disease classification tasks, and their performance is evaluated on the original test set. To ensure the robustness of our experiments, we employ two widely-used image classifiers: DenseNet-121 pretrained on ImageNet following the chestXclusion [11] and ConvNext [12]. We quantify the classification performance using the mean Area Under the Curve (mAUC) metric, which provides a comprehensive evaluation of the classifiers' effectiveness.

IV. RESULT

A. Comparison with the State-of-the-Art Method

In the comparative experiments, our model achieves significantly superior performance in the image synthesis task compared to state-of-the-art (SOTA) models. The quantitative superiority of our model in image generation quality is substantiated by its benchmark performance: the lowest FID score (31.600) confirms enhanced fidelity to real image distributions, while the highest IS (2.84) and Precision (0.523) metrics validate improved diversity and semantic alignment of synthesized outputs, as documented in Tab. I. For tail classes, our model demonstrates remarkable performance gains achieving 10.45 reduction in FID with 0.8 improvement in IS, 0.178 increase in Precision and 0.643 enhancement in Recall. These comprehensive advancements quantitatively validate the model's dual capability in maintaining global distribution fidelity through the construction of CPB and capturing tail-class-specific features via adaptive Prototype-Image Cross-Attention mechanisms.

TABLE I. THE PERFORMANCE COMPARISON ON CHESTX-RAY14 DATASET.

Approach	*FID*↓	*IS*↑	*Precision*↑	*Recall*↑
StyleGAN3[13]	36.221	2.296	0.343	0.418
CBDM[14]	52.139	2.419	0.276	0.416
LDM[6]	47.730	2.592	0.343	0.450
DiT[15]	50.016	2.560	0.282	**0.459**
U-ViT[16]	56.204	2.723	0.295	0.455
CPDM	**27.333**	**2.842**	**0.523**	0.433
CPDM (tail)	44.027	2.989	0.717	0.701
StyleGAN3(tail)	54.442	2.189	0.539	0.058

B. Downstream task and Visualization

To further validate the effectiveness of our approach, we conducted comprehensive downstream application experiments coupled with interpretability visualization analyses on the generated data, as demonstrated in Tab. II, Fig. 3 and Fig. 4.

Tab. II presents the performance of the classifiers on mixed datasets with varying ratio of the original dataset in the downstream experiments. Horizontally, for both classifiers, the mAUC of the mixed datasets shows improvements compared to datasets without generated data. Notably, when the original dataset consists of 1% and 10% real data, the mAUC increases by 15.9% and 10.2%, respectively, for DenseNet-121. Vertically, when comparing the generated data from our model with that of the current SOTA model in class-conditional generation, we observe that mixed datasets using data generated by our CPDM model consistently yield better classification performance for both classifiers. Furthermore, as shown in the Fig. 3, when training classifiers on datasets combining real data (ratio=100%) with synthetic samples, we observe a notable performance disparity that while the overall mAUC shows a modest 2% enhancement, tail classes exhibit a more pronounced 5% gain. Specifically, classifiers trained on the ChestX-ray14+CPDM hybrid dataset outperform those using data generated by the SOTA model for tail class recognition, achieving a 0.03 higher mAUC. This differential improvement underscores our method's capability to leverage the Q-Block and cross-attention mechanisms in effectively utilizing category-prototypical features preserved in the CPB, enabling more accurate modeling of tail class distributions and generating images that better approximate the original data distribution.

TABLE II. ENHANCING CLASSIFIER PERFORMANCE WITH GENERATED DATA.

Classifier	*Data*	*Ratio of Real Images*		
		1%	*10%*	*100%*
DensNet-121	ChestX-ray14	0.5343	0.6737	0.7969
	ChestX-ray14 + styleGAN	0.6625	0.6841	0.8069
	ChestX-ray14 + CPDM	**0.6933**	**0.7757**	**0.8250**
ConvNext	ChestX-ray14	0.5321	0.7157	0.8176
	ChestX-ray14 + styleGAN	0.6534	0.7841	0.8237
	ChestXray14 + CPDM	**0.7043**	**0.8176**	**0.8252**

As evidenced by the ablation studies in Table III: The proposed CPB module achieves a 26.33 reduction in FID while maintaining the original IS. Subsequent integration of the Q-block to enhance feature interaction between prototypes and latent representations yielded further gains of -28.87 in FID and +0.12 in IS. These results demonstrate the effectiveness of our proposed CPB and Q-block modules in enhancing category-specific image generation quality.

TABLE III. THE ABLATION EXPERIMENTS RESULTS OF CPDM

Approach	*FID*↓	*IS*↑
CPDM (**w/o**) CPB, Q-Block	56.20	2.72
(**w/**) CPB	29.87**(-26.33)**	2.70
(**w/**) CPB, Q-Block	27.33 **(-28.87)**	2.84**(+0.12)**

In Fig. 4, we present visual examples from three categories including the head, middle and tail class comparing generated images from the original data, the current SOTA model and our CPDM. The experimental results demonstrate that our method achieves superior image quality with enhanced texture resolution and more accurate lesion representation compared to SOTA approaches. Specifically, for head and medium-frequency classes, our method maintains comparable lesion feature representation while significantly improving fidelity and texture clarity. More importantly, for tail classes such as Hernia (0.05% occurrence) as shown in the lower-right panel of Fig. 4, our approach not only preserves excellent fidelity and diversity but also outperforms existing methods in capturing rare disease characteristics. This breakthrough is primarily attributed to our CPB, which effectively retains the distinctive features of tail-class data and provides clearer guidance for tail-class image generation. This confirms that our model not only achieves high fidelity and diversity but also provides a more refined and precise representation of rare and challenging categories in the dataset.

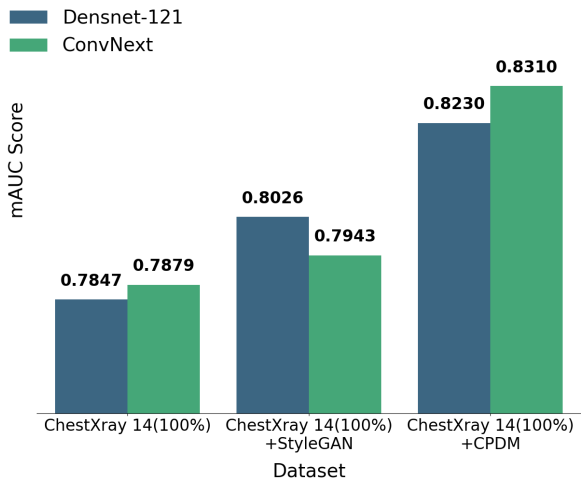
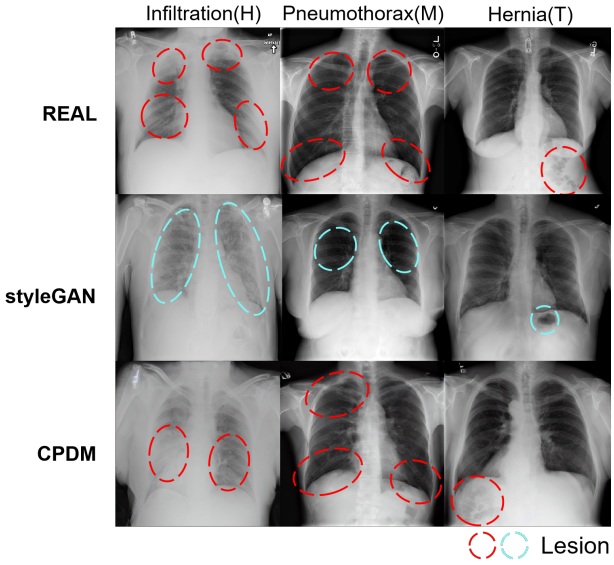


Figure 3. Performance Comparison on Thoracic Disease Classification: mAUC for Four Tail Classes Using 100% Real Image Data



(H), (M) and (T) refer to the head, middle and the tail classes respectively

Figure 4. Visualization Comparison of Synthesized Images

V. Conclusion

In this paper, we present a medical image generation framework CPDM that explores potential solutions to data scarcity and class imbalance challenges through two novel components: a Class Prototype Bank for feature preservation and a Q-block module enabling cross-modal interaction between class semantics and visual patterns. Experimental validation indicates improved performance generation especially the tail class compared to existing approaches, with downstream task analyses suggesting that synthesized images may provide meaningful augmentation for classifier training under data-limited scenarios. These results contribute to ongoing efforts in addressing long-tailed distribution problems in Chest X-ray imaging, while highlighting directions for future work in medical conditional generative augmentation.

Statement

This study did not involve human participants or animal experiments and thus no ethical approval was required.

Acknowledgment

This work was supported by Shanghai Municipality Science and Technology Commission under Grant 22ZR1404800 and Shanghai Municipal Education Commission 24KXZNA09. The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] Rahman, Md Mostafijur, Mustafa Munir, and Radu Marculescu. "Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[2] Lei, Wenhui, et al. "Medlsam: Localize and segment anything model for 3d medical images." arXiv preprint arXiv:2306.14752 (2023).

[3] Goodfellow, Ian, et al. "Generative adversarial networks." *Communications of the ACM* 63.11 (2020): 139-144

[4] Müller-Franzes, Gustav, et al. "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis." *Scientific Reports* 13.1 (2023): 12098.

[5] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

[6] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[7] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *Advances in neural information processing systems* 34 (2021): 8780-8794.

[8] Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." *arXiv preprint arXiv:2207.12598* (2022).

[9] Qin, Yiming, et al. "Class-balancing diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[10] Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[11] Seyyed-Kalantari, Laleh, et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers." *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. 2020.

[12] Liu, Zhuang, et al. "A convnet for the 2020s." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[13] Karras, Tero, et al. "Alias-free generative adversarial networks." *Advances in neural information processing systems* 34 (2021): 852-863.

[14] Qin, Yiming, et al. "Class-balancing diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[15] Peebles, William, and Saining Xie. "Scalable diffusion models with transformers." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[16] Bao, Fan, et al. "All are worth words: A vit backbone for diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.