DENOISING DIFFUSION MEDICAL MODELS

Pham Ngoc Huy[†], and Tran Minh Quan^{†‡} Email : {hpham, qtran}@talosix.com

[†]Talosix, Ho Chi Minh City, Vietnam

[‡]VinUniversity, Ha Noi, Vietnam

ABSTRACT

In this study, we introduce a generative model that can synthesize a large number of radiographical image/label pairs, and thus is asymptotically favorable to downstream activities such as segmentation in bio-medical image analysis. Denoising Diffusion Medical Model (DDMM), the proposed technique, can create realistic X-ray images and associated segmentations on a small number of annotated datasets as well as other massive unlabeled datasets with no supervision. Radiograph/segmentation pairs are generated jointly by the DDMM sampling process in probabilistic mode. As a result, a vanilla UNet that uses this data augmentation for segmentation task outperforms other similarly data-centric approaches.

Index Terms— Image Synthesis, Generative Models, Denoising Diffusion, NeRP, ChestXR

1. INTRODUCTION

X-Ray (XR) images play a critical role in medical diagnosis, especially in the early stage of detecting many diseases [1, 2]. Deep learning applications for XR image processing have evolved in the era of artificial intelligence and achieved some remarkable feats, such as automated segmentation, lesion categorization, quantification, diagnosis and treatment. [3]. To gather such high-quality labels in the biomedical area, however, calls for a significant amount of annotated data, which can be expensive and difficult to obtain. When training a deep learning model on limited data, a variety of data augmentation approaches are used to avoid overfitting [4, 5]. For instance, simple image alterations like morphological geometry adjustments can successfully regularize the training process and aid deep learning techniques in yielding meaningful results. Recently, diffusion-based models emerged as a promising deep learning method for image generation and have demonstrated their advantages in terms of image quality and training stability.

In this paper, we propose a multi-branch model based on the Denoising Diffusion Probabilistic Model (DDPM)[6] and its improved method, the Denoising Diffusion Implicit Model (DDIM)[7], to synthesize more image/label tuples for use in training models for other biomedical imaging tasks such as segmentation. The generated samples not only improve diversity and photo-realism in several common metrics, but also generalize dataset distributions with minimal supervisions. In addition, the pairs of image/label sampled by DDMM also improve the results of downstream tasks compared to other state-of-the-art generative models.

Our contributions for the proposed method, namely Denoising Diffusion Medical Models (DDMM), are twofold: First, (1) the DDMM model is built from one or more branches of DDPM (radiographs and segmentation branches) that share the same noise scheduler and latent code, which enforce semantic consistency. Second, (2) each branch in DDMM can use other unlabeled large-scale datasets to increase diversity and generalization. These settings make DDMM useful for (a) generalizing the dataset distribution by simultaneously generating high-quality XR-like images and their annotations, and (b) leveraging the synthesized data to improve other image analysis tasks (such as segmentation) or broaden out-domain image-to-image translation.

2. RELATED WORK

2.1. Generative models for XR-like image generation

XR-like image generation methods were developed using multiple approaches, both physics-based and otherwise. The physics-based models, such as XRaySyn [8] or Deep-DRR [9], produce high-quality images. However, their training procedures require knowledge from other modalities, and they can not generate the corresponding segmentation. In terms of non-physics-based methods, the GAN-based models showed improvements in image synthesis [10], data augmentation [11], and style augmentation [12]. However, the GAN approaches often experience unstable training and mode collapse, particularly when generating images from random noise.

Recently, diffusion-based solutions (DDPM [6], DDIM [7], etc.) have emerged as a new method for image synthesis by gradually denoising a random noisy image in large timesteps, during which a temporal encoding is attached to guide a reconstruction UNet [13] model. These aforementioned problems and innovations motivated us to develop a novel approach of generating paired XR-like images from random noises without facilitating other modality knowledge.



Fig. 1: The upper branch is a DDPM model that attempts to denoise the random Gaussian noisy input and produce the XR-like image, while the lower branch tries to generate the corresponding segmentation. To ensure semantical consistency, both use the same initialized noises and noise scheduler.

2.2. Denoising diffusion probabilistic model

With an input data sample $x_0 \sim q(x_0)$, the forward process q adds the Gaussian noise with variance $\beta_t \in (0, 1)$ at each time-step to the given input x_{t-1} and produces T latents x_t where the subscription t ranges from 1 to T:

$$q(x_1, ..., x_T | x_0) := \prod_{t=1}^T q(x_t | x_{t-1})$$
(1)

$$q(x_t|x_{t-1}) := \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \quad (2)$$

Supposing that the time-steps T is long enough, and a good beta scheduler is properly designed, the latent x_t approximates a Gaussian distribution. Therefore, if the distribution $q(x_{t-1}|x_t)$ is known, we can sample $x_T \sim N(0, \mathbf{I})$ and feed it to the forward process to get $q(x_0)$. On the other hand, a reverse process can be defined as a routing that gradually removes the noise in the inputs, begins at the point $p(x_T) = \mathcal{N}(x_t, \mathbf{0}, \mathbf{I})$. The join distribution $p_{\theta}(x_{0:T})$ is calculated from the starting point by the following Markov chain:

$$p_{\theta}(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_{\theta}(x_{t-1}|x_t)$$
 (3)

$$p_{\theta}(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, t), \Sigma_{\theta}(x_t, t))$$
(4)

In this case, p could be considered an approximation of q in each time step t. Therefore, q and p are components of a variational auto-encoder. The loss can be defined as the variational lower bound on negative log-likelihood, and is formally rewritten as the sum of loss at each step:

$$L_{\text{vlb}} := L_0 + \dots + L_{t-1} + \dots + L_T \tag{5}$$

where

$$L_0 \quad := -\log p_\theta(x_0|x_1) \tag{6}$$

$$L_{t-1} := D_{KL}(q(x_{t-1}|x_t, x_0)||p_{\theta}(x_{t-1}|x_t))$$
(7)

$$L_T := D_{KL}(q(x_T|x_0)||p(x_T))$$
(8)

3. METHOD

3.1. Model overview and Training Procedure

As shown in Fig. 1, our DDMM method consisted of **two** separate DDPM models that shared the noise latent code. The model p_{θ} takes responsibility for generating the XR images while the model p_{ϕ} produces the corresponding segmentation. The supervised datasets, which include both XR images and labels, are introduced in each training step to compute the reconstruction loss. The p_{θ} branch can be further trained with other unlabeled datasets at the same stage, producing an unsupervised loss. The total loss of the training step is the combination of the supervised loss and unsupervised loss. While the former loss component drives the model to generate the corresponding pairs of images and labels since both share the same latent noise, the latter loss component supports in expanding data distribution coverage with different sampling points.

3.2. Probabilistic Sampling Procedures

Our multi-branch DDMM method can be extended to multimodal image-to-image translation with a minimal number of paired samples and extensive unpaired images. Equivalently, the Gaussian-noise in the latent space plays as a semantic-sharing code across the multiple domains that it can span. Therefore, we can perform the sampling procedures in a probabilistic way: we start at one side of both radiograph/segmentation branches, initialize and fetch the noise into the processes, then gradually denoise them using p_{θ} and p_{ϕ} networks. This approach results in both image/label tuples generated simultaneously, with one of them satisfying the XR image distribution, while the other one is tightly coupled with its semantic segmentation. With this approach, we can synthetically generate a massive amount of semanticallyconsistent pairs of images, which is helpful to finetune other downstream tasks.

Datasets	Images	Annotation
ChinaSet [14]	566	Lung
Montgomery [14]	138	Lung
JSRT [15]	247	Lung
VinDr-CXR [16]	18,000	N/A

Table 1: Dataset summary. Those which do not have segmentation are used to train the unsupervised branches.

Table 2:	Image	quality	metrics
----------	-------	---------	---------

Method	FID ↓	KID↓	SSIM ↑	UQI↑	SCC↑
GAN	279.869	0.3765	0.3724	0.0661	0.1291
XRaySyn	181.390	0.2256	0.3317	0.0318	0.0701
NeRP	174.294	0.1875	0.3268	0.0415	0.1077
DDMM (1)	155.772	0.1913	0.3592	0.0686	0.1325
DDMM (2)	93.998	0.0976	0.4258	0.1012	0.1750

DDMM (1) is trained with the supervised branch only, while DDMM (2) is trained with both supervised and unsupervised branches.

3.3. Implementation Details

Our DDMM framework is implemented based on the available open source of DDPM [6] with the learning rate of Adam optimizer set to 1e-4. The models p_{θ} and p_{ϕ} share the same cosine-based noise scheduler β_t but the supervised training noise is initialized with a different seed than the unsupervised scheme. We set the total time-steps T = 100 for all diffusion branches and trained the DDMM for 100 epochs on a workstation equipped with 64 GB of RAM and an NVIDIA GTX 3090 GPU. The training took approximately three days to complete.

4. DATA

We demonstrated the experiments on the ChestXR images. The number of images available in each sub-dataset is shown in Table 1. There are 951 accompanying images with their lung region segmentations [14]. These labeled pairs are split into the training and test sets at a ratio 80:20. In the ChestXR experiment, we also leverage 15,000 out of 18,000 images from a large-scale public VinDr-CXR dataset [16], which does not have pixel-level segmentation for the lung regions, to adapt on to our unsupervised training step.

5. RESULTS

5.1. Quality of generated XR images

Fig. 2 illustrates the generated samples from DDMM and other methods compared to the actual images. The other two approaches, GAN and XRaySyn, produce low-quality images with much distortion and blur. The images from NeRP are higher quality, as compared to GAN and XRaySyn, but still



Fig. 2: Samples of XR images generated by our method DDMM and others, compared to the real samples in VinDr-CXR test sets (last row).

miss the bone details in ChestXR. These problems are not observed in the DDMM-generated images. The image quality is further evaluated quantitatively by collecting the numbers of Frechet Inception Distance (FID) [17], Kernel Inception Distance (KID) [18], the Structural Similarity Index Measurement (SSIM) [19], the Universal Image Quality Index (UQI) [20], the Spatial Correlation Coefficient (SCC) [21]. While the better methods have lower FID and KID scores, the SSIM, UQI, and SCC should be as high as possible. These metrics are calculated on 1,000 XR images drawn randomly from 10,000 pre-generated images for each method. On the ChestXR dataset, the results shown in Table 3 indicate that DDPM produces the highest quality of synthesized images. In general, without the unsupervised branch, we can not enhance the images.

5.2. Image Segmentation

 Table 4: Segmentation results

Anatomy	Method	Dice Score ↑	Rand Score ↑
Chest	NeRP	0.6619	0.4057
Chest	DDMM	0.7649	0.5763

In-domain segmentation: For each experiment on the above Chest datasets, we pre-generate 10,000 pairs of image/label using NeRP [22] and our method DDMM for training vanilla UNets [13] without other augmentation techniques. We do not reuse the previously split training sets for



Fig. 3: Qualitative evaluations on ChestXR test set of NeRP and DDMM methods against the ground truth.

this downstream task. The segmentation performances are assessed directly on the test sets by Dice-score and Rand-score metrics, which measure the semantic-aware and instanceaware significance. As can be seen in Fig. 3 and Table 4, our DDMM method extracts the lung areas qualitatively better than the physics-based method NeRP, and also achieves better segmentation metrics.

Out-of-domain segmentation: The segmentation models are tested on the out-of-domain samples. The model trained on the DDMM images performed well on the chest XR images even though the input sources are extreme outliers. For example, the pleural space with partially visible dark textured lung regions, or even the severe cases of the skeleton and t-shirt images (see Fig. 4), can be inferred more appropriately with our method. Interestingly, the model based on DDMM data still detects the lung area on the non-XR images, while the NeRP model cannot produce the proper segmentation. These extreme out-of-domain samples showcase the remarkable generalization capability.

6. CONCLUSION

We present DDMM, a diffusion-based multi-branch model that can jointly produce realistic XR medical images and their associated segmentation masks. The generated images outperform other similar work qualitatively and quantitatively. Our method is also beneficial for downstream tasks, such as improving the segmentation results in biomedical image analysis. In addition, DDMM is scalable and can be extended to other cross- and intra-modality such as CT, MRI, or multi spatial omics analysis. By leveraging the unlabeled datasets, DDMM better generalizes the data distribution and helps to capture more useful information, which in turn supports better diagnosis, treatment and precision medicine.





7. COMPLIANCE WITH ETHICAL STANDARDS

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of retrospective study, formal consent is not required.

8. CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

9. ACKNOWLEDGMENT

The authors would like to extend their thanks to Truc Thai, Trista Pham, Dennis Chien, Adam Rhine, and Arsenio Valdez (at Talosix LLC.) for their invaluable support and contribution towards the methodology and discussion of this paper. The authors are truly grateful for the support and assistance provided during the course of this project.

10. REFERENCES

- S. H. Bradley et al., "Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: a systematic review," *British Journal of General Practice*, vol. 69, no. 689, pp. e827–e835, Dec 2019.
- [2] F. Malgo et al., "Value and potential limitations of vertebral fracture assessment (VFA) compared to conventional spine radiography: experience from a fracture liaison service (FLS) and a meta-analysis," *Osteoporosis international*, vol. 28, no. 10, pp. 2955–2965, 10 2017.
- [3] Morgan P. McBee, Omer A. Awan, Andrew T. Colucci, Comeron W. Ghobadi, Nadja Kadom, Akash P. Kansagra, Srini Tridandapani, and William F. Auffermann, "Deep learning in radiology," *Academic Radiol*ogy, vol. 25, no. 11, pp. 1472–1480, 2018.
- [4] Luis Perez and Jason Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621, 2017.
- [5] Connor Shorten and Taghi M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal* of Big Data, vol. 6, no. 1, July 2019.
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020, vol. 33, pp. 6840–6851.
- [7] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.
- [8] Cheng Peng, Haofu Liao, Gina Wong, Jiebo Luo, S Kevin Zhou, and Rama Chellappa, "XraySyn: Realistic View Synthesis From a Single Radiograph Through CT Priors," in AAAI, 2021, vol. 35, pp. 436–444.
- [9] Mathias Unberath, Jan-Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab, "DeepDRR–a catalyst for machine learning in fluoroscopy-guided procedures," in *MICCAI*, 2018, pp. 98–106.
- [10] Bing Cao, Han Zhang, Nannan Wang, Xinbo Gao, and Dinggang Shen, "Auto-GAN: self-supervised collaborative learning for medical image synthesis," in *MIC-CAI*, 2020, vol. 34, pp. 10486–10493.
- [11] Hoo-Chang Shin et al., "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *IWSSMI*, 2018, pp. 1–11.
- [12] Sophia J. Wagner, N. Khalili, R. Sharma, M. Boxberg, C. Marr, Walter de Back, and T. Peng, "Structurepreserving multi-domain stain color augmentation using style-transfer with disentangled representations," in *MICCAI*, 2021, pp. 257–266.

- [13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.
- [14] S. Jaeger, S. Candemir, S. Antani, Y. X. Wang, P. X. Lu, and G. Thoma, "Two public CXR datasets for computeraided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475– 477, Dec 2014.
- [15] J. Shiraishi et al., "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, Jan 2000.
- [16] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Nguyen, et al., "VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations," *Scientific Data*, vol. 9, no. 1, pp. 1–7, 2022.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2017, vol. 30.
- [18] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton, "Demystifying MMD GANs," in *ICLR*, 2018.
- [19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [20] Zhou Wang and Alan C Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.
- [21] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge landsat tm and spot panchromatic data," *International Journal of Remote Sensing*, vol. 19, no. 4, pp. 743–757, 1998.
- [22] Pham Ngoc Huy and Tran Minh Quan, "Neural radiance projection," in *IEEE ISBI*, 2022, pp. 1–5.