2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635486

BOOSTING DERMATOSCOPIC LESION SEGMENTATION VIA DIFFUSION MODELS WITH VISUAL AND TEXTUAL PROMPTS

Shiyi Du[1] *Xiaosong Wang*[1] *Yongyi Lu*[2] *Yuyin Zhou*[2,4]
Shaoting Zhang[1] Alan Yuille[2] Kang Li[1,3] Zongwei Zhou[2]

[1]Shanghai Artificial Intelligence Lab, Shanghai, China
[2]Johns Hopkins University, Baltimore, USA
[3]West China Biomedical Big Data Center, Sichuan University West China Hospital, Chengdu, China
[4]University of California Santa Cruz, Santa Cruz, USA

ABSTRACT

Image synthesis approaches, e.g., generative adversarial networks, have been popular as a form of data augmentation in medical image analysis tasks. It is primarily beneficial to overcome the shortage of publicly accessible data and associated quality annotations. However, the current techniques often lack control over the detailed contents in generated images, e.g., the type of disease patterns, the location of lesions, and attributes of the diagnosis. In this work, we adapt the latest advance in the generative model, i.e., the diffusion model, with the added control flow using lesion-specific visual and textual prompts for generating dermatoscopic images. We further demonstrate the advantage of our diffusion model-based framework over the classical generation models in both the image quality and boosting the segmentation performance on skin lesions. It can achieve a 9% increase in the SSIM image quality measure and an over 5% increase in Dice coefficients over the prior arts.

Index Terms— Diffusion Model, Controllable Image Generation, Skin Lesion Segmentation.

1. INTRODUCTION

Image synthesis methods have played an important role in the development of machine vision-based applications as a data augmentation tool to enrich and expand the limited distribution of training data. It is especially helpful for those domains in which sample data and quality annotation are scarce and not cost-effective to obtain, e.g., anonymous driving and medical imaging. Massive research has been conducted in controlling the generated contents for the actual need in model training, from first manipulating the noise parameters Z [1], Conditional GAN (cGAN) [2], to supervised [3] and unsupervised [4] image style transfer, to decoupling the style and content parts in images [5], to latent diffusion model via text2image [6] and most recently diffusion model based ControlNet [7]. However, the application of such algorithms in medical image analysis remains limited due to the special characteristics of medical images.

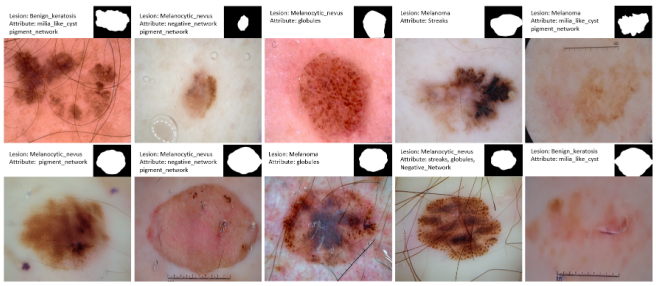

Fig. 1. Quiz: which ones are real samples or synthesized by the proposed work? Answers are in the experiments section. Lesion types, attributes, and masks are used as prompts in the training and inference.

In most medical diagnosis scenarios, anomaly studies (e.g., image scans with lesions or other abnormalities) are in the minority. Although increasing the number of anomalous samples could potentially help improve the performance of subsequent tasks like segmentation tasks, the generation of such corner cases is not well-controlled and mostly customized at the image level, e.g., the pioneering work in adopting GAN for data augmentation in medical image segmentation [8], adopting cGAN for colon polyp generation [9], diversify the generated image using radiogenomic features [10] and pseudo labels [11].

Fundamentally, the data sample of such abnormalities remains low in terms of a normal clinical distribution, and it is even hard to obtain enough data for training purposes. Then, the scarce corner cases are often flooded with normal cases or other cases with more common diseases in the set of generated images. There is a critical demand for generating data with desired categories, i.e., specific disease types and disease attributes (shape, location, appearance, severity).

Additionally, the quality and efficiency of the generation of lesion images, as well as the transferability of the gener-

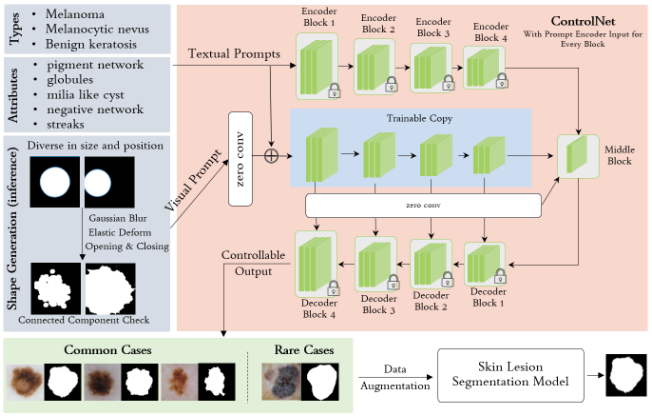

Fig. 2. Overview of our visual and textural prompted image generation framework.

ation method, are essential to improve the performance of subsequent tasks and to apply the method to anomaly images of different organs in different modalities. The handcrafted methods [12, 13, 14] are effective when applied to a specific organ and modality, but they are not automated enough and lack universality among various organs and modalities.

In this paper, we propose to use the diffusion models as the backbone to generate skin lesion images. Examples are shown in Fig. 1. The proposed framework largely leverages the recent work of ControlNet [7] while we attempt to integrate the controllable lesion function (with desired lesion type, attributes in text, and shapes with locations in masks images) into the framework for both the training and inference stage. The correlation could be first learned by linking the visual and textual prompts with the detailed image contents and then prompted during the inference by focusing on rare cases. We also proposed an automatic module to generate lesion shapes and masks. We conducted the experiments and the comparison study (mainly with a classic GAN method, Pix2PixHD [15]) on a publicly accessible skin lesion dataset, ISIC [16, 17]. The result demonstrates the superiority of the diffusion model-based framework over the classical generation models in both the image quality and boosting the segmentation performance on skin lesions. To our knowledge, we are arguably the first to utilize the diffusion model for skin lesion generation. A PyTorch implementation of our method can be found later on our GitHub repository.

2. METHODS

The overall workflow of our method is illustrated in Fig. 2. In the training stage, we utilize the skin images with lesions from the public dataset and their corresponding visual and textual prompts, i.e., the masks of these skin lesions, indicating the shape and location information for the generation and the lesion types and associated attributes. They are the input to the ControlNet [7] for training the lesion image generator, which is imposed on top of the Stable Diffusion model [6]. In order to generate the necessary amount of synthetic data to augment the downstream lesion segmentation task, we build an automatic shape generation module, which can produce lesion segmentation masks with a diversity of lesion sizes, shapes, and locations. Theoretically, there is no limit to the number of these generated masks, together with randomly picked textual tags of lesion categories and attributes, that can maximally facilitate the training of downstream lesion segmentation tasks.

In the following, we first introduce the principle of Denoising Diffusion Implicit Models [18] as the common diffusion model backbone in §2.1. Then, we explain the diffusion models with visual and textual prompts in §2.2. Specifically, we introduce the process of automatically generating lesion shapes in §2.3.

2.1. Backbone Diffusion Model

Denoising Diffusion Implicit Model (DDIM) is a recent technique for generative modeling that builds on the common framework of Diffusion Models. DDIM leverages a denoising process to generate high-quality samples from the underlying probability distribution of a dataset.

The key idea behind DDIM is to use a diffusion process to smooth out the noise in an image or data sample, gradually moving it toward the true data distribution. This process is performed by iteratively applying a diffusion equation, where the noise in the data is diffused according to a predetermined schedule of diffusion steps. At each step, the data is corrupted by a certain amount of noise, and the resulting noisy data is used as input for the next step.

The diffusion equation used in DDIM can be written as:

$$\frac{\partial x}{\partial t} = \frac{\beta(t)}{2}\nabla^2 x + \sqrt{\beta(t)}\epsilon,$$

where x represents the data sample, t is the diffusion time, ∇^2 is the Laplacian operator, $\beta(t)$ is the diffusion coefficient, and ϵ is Gaussian noise with zero mean and unit variance.

In a DDIM model, we use a denoising objective function that encourages the model to minimize the distance between the diffused noisy data and the original data. Specifically, we use a maximum likelihood approach to learn the parameters of the diffusion process that minimize the negative log-likelihood of the training data. The denoising objective function can be written as:

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\frac{1}{2\sigma_t^2}|x_{i,t} - \widetilde{x}i, t|^2 - \frac{1}{2}\sum t = 0^T \log\sigma_t^2,$$

where θ represents the parameters of the DDIM model, n is the number of data samples, T is the number of diffusion steps, $x_{i,t}$ is the data sample for the i-th training example at

time t, $\widetilde{x}_{i,t}$ is the diffused noisy data, and σ_t is the diffusion scale parameter.

2.2. Diffusion Models with Multi-modality Prompts

We aim to use diffusion models with multimodal prompts to generate skin images with lesions and then further benefit the downstream segmentation task. First, we train a diffusion model with multimodal prompts for medical images based on Stable Diffusion Models [6] with the multimodal condition, also known as ControlNet [7]. ControlNet is designed to control the diffusion model by adding additional conditions to facilitate what we call medical image generation through Stable Diffusion Models with multi-modality prompts. As shown in Fig. 2, the network structure is divided into trainable and locked sections in ControlNet. The trainable part is the controllable part that is initialized with the same encoder of the stable diffusion model and then connects the prompts and the detailed generation output. The locked part retains the original parameters of the trainable-diffusion model, so a small amount of data is used to bootstrap. Therefore, we can ensure the adapted model learns the desired controlling constraints while retaining the generation capability of the original diffusion model itself. Specifically, the parameters in the upper and lower Encoder and Decoder blocks are locked, and the parameters in the middle blocks are the "trainable" ones.

Then, we use the trained model to generate a large number of skin lesion images. At this time, we incorporate the automatically generated skin lesion mask module we mentioned earlier and use the lesion masks generated by this module as input to the trained model to generate the complete skin image with a lesion. According to our subsequent experiments, the segmentation of the model is better to a certain extent as we add more synthetic images to the training.

2.3. Automatic Lesion Mask Generation

We experiment with two ways to generate the shape of skin lesions, one of which is the direct transformation according to the existing segmentation masks, and the other is automatically constructing the segmentation masks. However, according to our experimental results, we found that the latter is more flexible and productive in terms of shape diversity and subsequent contribution to the model effect in the segmentation process. Especially, the automated process offers an almost unlimited number of data samples. Therefore, we will mainly focus on and discuss more of the automatic lesion shape generation below.

Overall, we generate synthetic images of circles of different sizes and positions and applied some post-processing methods to make the images more realistic. First, a blank canvas is created, then a random point is chosen at the center of the canvas, and a random radius is chosen between a defined threshold and a minimum distance from that point to the edge of the canvas. Next, Gaussian blurring is applied to the image with the inserted circle. Alternatively, the image could be elastically deformed using the elastic deformation library. Finally, morphological on and off operations are performed with elliptical structure elements. The resulting image is labeled using the classic region-growing function to figure out if it has only one connected component (i.e., circle). On the other hand, transformation-based shape generation constructs the shape directly from the original masks, where we operate resizing, rotating, and elastic transforming of the original segmentation masks.

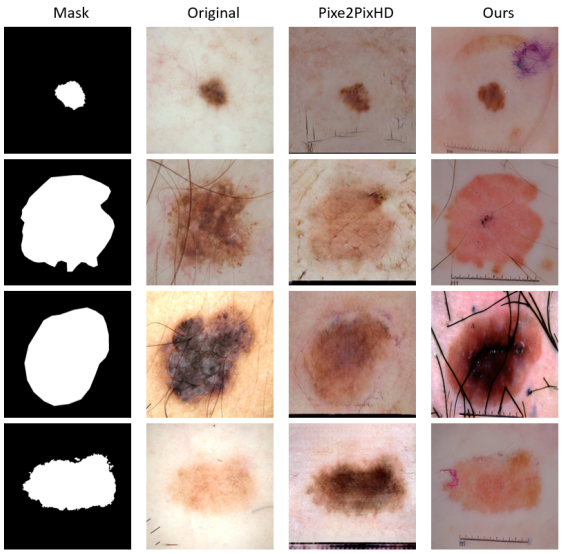

Fig. 3. An example of image generation based on masks, from left to right: a skin lesion mask, a Pix2PixHD model generated image, an original image, a diffusion model generated image

3. EXPERIMENTS

We conduct the experiments in this study to mainly two aspects of the proposed generation framework, i.e., the quality of generated images and how the controlled generation of sample data (images with associated masks) could benefit the downstream skin lesion segmentation task.

3.1. Dataset and Experiment Setup

The International Skin Imaging Collaboration (ISIC) skin segmentation dataset [16, 17] is from the world's largest skin image analysis challenge, hosted by ISIC, a global partnership that has organized the world's largest public repository of dermoscopic images of skin. Though there are many application tasks, e.g., lesion type and attribute classification, the goal of this work is to create a model for segmenting lesion boundaries (same task as ISIC 2018 task 1). There were 2,594

S +	S+1K	S+3K	S+5K	SOTA	S only
P2PHD	0.871	0.903	0.912	U-Net	0.861
Ours	**0.912**	**0.913**	**0.914**	DCSAU-Net	0.903

Table 1. Performance of model lesion segmentation with different generative models adding different amounts of generative images marked by DSC ('S+' and 'P2PHD' indicates 'S+generated images' and 'Pix2PixHD').

Model	MSE	PSNR	SSIM
Pix2pixHD	0.09	58.80	0.71
Ours	0.06	61.64	0.80

Table 2. Comparison of image generation between Pix2PixHD and ours.

dermoscopic images provided with ground truth segmentation masks. We further cross-correlate the lesion images with the diagnosis and attribute information from ISIC 2017 and 2019 to form the prompts we use for the training of ControlNet. Sample images and associated information are shown in the top row of Fig. 1, and images on the bottom row of Fig. 1 are generated via our framework.

We divided the dataset into two parts. Roughly half of the data (1,594 data samples in total) is reserved for the training of the generation model, defined as the G set. The rest (1,000 samples) is employed for the experiments of the lesion segmentation task, defined as the S set. We split the 1,000 images in S set randomly into training, validation, and testing with a ratio of 7:1:2 for evaluating the segmentation performance. We compute the Sørensen–Dice coefficient (DSC) for all the segmentation results.

3.2. Implementation Details

The diffusion models with visual and text prompts are implemented using Pytorch Lightning based on the codebase provided by [7]. The Segmentation and Pix2PixHD model experiments are carried out on NVIDIA A100 Tensor Core GPUs. To train all models, we use a common segmentation loss function, i.e., Dice loss, since we make it clean and target evaluating the performance gain from additional data samples generated from our framework. The Adam optimizer with a learning rate of 1e-3 is set with a batch size of 256 and a total epoch number of 350 during the training.

3.3. Results on Image Generation

Here, we utilize the testing images of the S set to evaluate the image generation quality. After both the image generation methods, e.g., ours and Pix2PixHD, we use the mask of the original images in the testing of S for image generation. No prompt is entered, i.e., remaining empty in the prompts for our framework for a fair comparison. In this way, we can compare the generated image with the original image to see how the methods work in a quantitative manner.

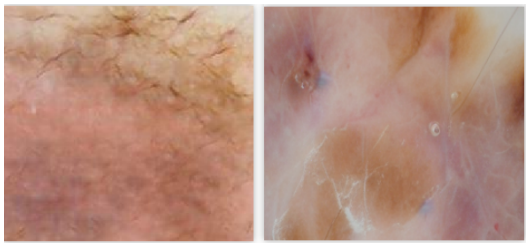

Fig. 4. On the left are the texture details generated by the Pix2PixHD model, and on the right are the texture details generated by the Stable Diffusion model with linguistic and visual prompts.

As shown in Table 2, ours outperforms the Pix2PixHD method by a significant margin. In general, the lesion content and the detailed texture of our generated image are better controlled and produced in comparison to the GAN counterpart. We also illustrate four examples of generation results for both methods in Fig. 3. and enlarged texture regions from both methods are shown in Fig. 4. The diffusion model-based image generation can achieve much better generation results with all the details preserved, a clear difference in the degree of texture detail generation as shown in Fig. 4.

3.4. Lesion Segmentation Results

As shown in Table 1, the diffusion model outperforms the Pix2PixHD by a large margin (over 5%). We also experiment with how different amounts (e.g., 1K, 3K, and 5K) of synthetic data will affect performance. Indeed, more data can benefit the segmentation training, with the potential to increase the data amount.

4. CONCLUSION

In this work, we present a diffusion model-based image generation framework with detailed lesion characteristics as the prompts for the lesion image generation task. We demonstrate both quantitatively and qualitatively that the quality of the resulting images is significantly better than the counterpart with the GAN framework, i.e., the popular Pix2PixHD approach. The proposed framework opens the gate of precise data sample generation for multi-tasks, e.g., the segmentation task in this work and possible lesion diagnosis and attributes for image classification.

Acknowledgments

This work is funded by the National Key R&D Program of China (2022ZD0160700).

5. REFERENCES

[1] Ian J. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *CoRR*, vol. abs/1701.00160, 2017.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2016.

[3] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2018.

[4] Harshad Rai and Naman Shukla, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2018.

[5] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision*, 2018.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[7] Lvmin Zhang and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," *ArXiv*, vol. abs/2302.05543, 2023.

[8] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine P. Andriole, and Mark H. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *SASHIMI@MICCAI*, 2018.

[9] Younghak Shin, Hemin Ali Qadir, and Ilangko Balasingham, "Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance," *IEEE Access*, vol. 6, pp. 56007–56017, 2018.

[10] Ziyue Xu, Xiaosong Wang, Hoo-Chang Shin, Dong Yang, Holger R. Roth, Fausto Milletarì, Ling Zhang, and Daguang Xu, "Correlation via synthesis: End-to-end image generation and radiogenomic learning based on generative adversarial network," in *International Conference on Medical Imaging with Deep Learning*, 2020.

[11] Fei Lyu, Mang Ye, Jonathan Frederik Carlsen, Kenny Erleben, Sune Darkner, and Pong C Yuen, "Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[12] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jie-Neng Chen, Alan Yuille, and Zongwei Zhou, "Synthetic tumors make ai segment tumors better," *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022.

[13] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou, "Label-free liver tumor segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7422–7432.

[14] Bowen Li, Yu-Cheng Chou, Shuwen Sun, Hualin Qiao, Alan Yuille, and Zongwei Zhou, "Early detection and localization of pancreatic cancer by label-free tumor synthesis," *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023.

[15] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2017.

[16] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[17] Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1902.03368, 2019.

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *CoRR*, vol. abs/2010.02502, 2020.