BGDiffSeg: a Fast Diffusion Model for Skin Lesion Segmentation via Boundary Enhancement and Global Recognition Guidance

Yilin Guo[1] and Qingling Cai[1*]

[1] Sun Yat-sen University, Shenzhen 518107, China
guoylin8@mail2.sysu.edu.cn,caiqingl@mail.sysu.edu.cn

Abstract. In the study of skin lesion segmentation, models based on convolution neural networks (CNN) and vision transformers (ViT) have been extensively explored but face challenges in capturing fine details near boundaries. The advent of Diffusion Probabilistic Model (DPM) offers significant promise for this task which demands precise boundary segmentation. In this study, we propose BGDiffSeg, a novel skin lesion segmentation model utilizing a wavelet-transform-based diffusion approach to speed up training and denoising, along with specially designed Diffusion Boundary Enhancement Module (DBEM) and Interactive Bidirectional Attention Module (IBAM) to enhance segmentation accuracy. DBEM enhances boundary features in the diffusion process by integrating extracted boundary information into the decoder. Concurrently, IBAM facilitates dynamic interactions between conditional and generated images at the feature level, thus enhancing the global recognition of target area boundaries. Comprehensive experiments on the ISIC 2016, ISIC 2017, and ISIC 2018 datasets demonstrate BGDiffSeg's superiority in precision and clarity under limited computational resources and inference time, outperforming existing state-of-the-art methods. Our code will be available at https://github.com/erlingzz/BGDiffSeg.

Keywords: Skin lesion segmentation, Denoising Probabilistic Models

1 Introduction

Malignant melanoma is one of the fastest-growing cancers globally, with an estimated 97,610 new cases in the US in 2023 alone [1]. Accurate and rapid segmentation of these lesions is critical for early detection and treatment planning, making precision in automated skin lesion segmentation imperative. The task is challenging as lesions in skin images are often obscured by natural artifacts like hair and blood vessels, as well as artificial ones like surgical markings, which can closely resemble the texture, color, and shape of lesions. Moreover, low contrast and indistinct boundaries make it difficult to distinguish lesions from healthy skin. Accurately locating skin lesion areas and precisely predicting clear lesion boundaries are crucial. As deep learning evolves, neural

* Qingling Cai is the corresponding author.

network models, from convolution neural networks (CNN) to the recent vision transformers (ViT) [2], have advanced medical image segmentation. UNet [3], known for its stellar performance and efficient CNN-based design, is widely used in medical imaging. Its extensibility has led to enhancements like UNet++ [4], AttentionUNet [5] and 3D-UNet [6]. TransUNet [7] uses ViT for encoding and CNN for decoding, showing great global information capture ability in medical image segmentation. Similarly, MedT [8] leverages a transformer-based encoder and adds control within self-attention for impressive results. Swin-UNet [9] combines Swin Transformer with the UNet structure, introducing a fully transformer-based model using self-attention within shifting windows. CTO [10] employs a combination of CNN, ViT and a boundary detection operator to achieve high accuracy. However, these methods exhibit inherent limitations. CNN-based approaches suffer from detail loss due to downsampling and upsampling, particularly affecting fine features near boundaries, leading to boundary misalignments. Similarly, ViT-based methods, constrained by fixed windows, struggle to capture the fine contextual details necessary for precise pixel-level segmentation. These shortcomings necessitate the development of a new medical image segmentation architecture, designed to segment skin lesions with greater accuracy and clarity.

Recently, Diffusion Probabilistic Models (DPMs) [11] have attracted considerable attention for their superior performance [12][13][14][15]. Inspired by DPMs, we find their particular suitability for skin lesion segmentation tasks that demand precise and clear boundary predictions. Unlike CNNs and Transformers, DPMs model the evolution of lesion boundaries as a parameterized process, which aids in learning the distribution of the target for clearer segmentation. Moreover, DPMs perform denoising at the original image size, effectively avoiding boundary shifts caused by downsampling and upsampling in CNNs. Utilizing these unique characteristics, we aim to employ diffusion models for tackling the challenges of skin lesion segmentation. Although studies like MedSegDiff [16] have begun exploring diffusion models in medical image segmentation, their slow training, lengthy inference time, and suboptimal accuracy limit their clinical applicability in dynamic and real-time settings. Thus, achieving precise and clear lesion segmentation with limited computational resources and inference time remains a significant challenge.

In this study, we propose BGDiffSeg, a diffusion model-based skin lesion segmentation model to address the challenges of speed, inference time, and accuracy in existing diffusion models for medical image segmentation, demonstrating its potential for dynamic, real-time, and precise segmentation. To enhance training speed and reduce inference time for segmentation tasks, we utilize a wavelet transform-based diffusion approach, named WaveDiff [17]. To improve segmentation accuracy, we specifically design two key modules: the Diffusion Boundary Enhancement Module (DBEM) and the Interactive Bidirectional Attention Module (IBAM). On one hand, DBEM is proposed to enhance the boundary features of the lesion area generated by the diffusion model. Specifically, it extracts low-frequency information from the conditional encoder to reduce high-frequency noise interference and employs the Sobel operator for boundary extraction. This extracted boundary information is then integrated into the diffusion model's decoder to refine the prediction of lesion boundaries. On the other hand, IBAM

is proposed to enhance global recognition of target area boundaries. Specifically, it fosters interactions between the generative encoder and the conditional encoder, enables dynamic feature-level interaction between conditional and generated images, and makes the most of features at various levels to enhance global boundary recognition. By combining the aforementioned modules with the baseline diffusion model, we propose BGDiffSeg and conduct extensive experiments on multiple skin lesion segmentation datasets, including ISIC 2016 [18][19], ISIC 2017 [20][21], and ISIC 2018 [22][23]. The results demonstrate that BGDiffSeg can segment lesion boundaries more accurately within limited resources and inference time, achieving state-of-the-art performance in various skin lesion segmentation tasks.

In summary, our contributions are as follows: 1) We propose BGDiffSeg, a skin lesion segmentation model based on the diffusion model that utilizes a wavelet-transform-based denoising generative adversarial approach, combined with DBEM and IBAM, for more precise segmentation of skin lesions with limited resources and inference time. 2) We specifically design DBEM and IBAM, where DBEM enhances boundary features of target areas generated during the diffusion process, and IBAM improves global recognition of target area boundaries. 3) We conduct extensive experiments, which demonstrate the effectiveness of our methods in achieving state-of-the-art performance.

2 Preliminaries

Current studies have demonstrated the immense potential of DPMs in image tasks, yet their extensive training and sampling times restrict real-time applications. To speed up both training and sampling, we adopt WaveDiff [17]. Given an input image $x \in R^{1\times H\times W}$, it is decomposed into a set of low and high subbands, which are then concatenated to form a matrix $y \in R^{4\times\frac{H}{2}\times\frac{W}{2}}$, effectively reducing the network's computational load by decreasing spatial dimensions fourfold.

The forward diffusion process gradually adds noise to the data $x_0 \sim q(x_0)$ in T steps with pre-defined variance schedule β_t:

$$q(x_{1:T}|x_0) = \prod_{t\geq 1} q(x_t|x_{t-1}), \qquad q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

where $q(x_0)$ is a data-generating distribution, and I is the identity matrix.

Compared to DPM's unimodal distribution, WaveDiff models the denoising distribution as a complex multimodal one, reducing sampling steps while maintaining high generative quality. Specifically, it uses conditional generative adversarial networks (GANs) to approximate the true denoising distribution:

$$p_\theta(x_{t-1}|x_t) = \int p(z)q(x_{t-1}|x_t, x_0 = G_\theta\ (x_t, z, t))dz \tag{2}$$

where $p_\theta\ (\mathrm{x}_{t-1}|\mathrm{x}_t)$ denotes the implicit distribution imposed by the generator $G_\theta\ (\mathrm{x}_t, z, t)$ that outputs x_0, given x_t and a latent variable $z \sim\ p(z) := \mathcal{N}(z;\ 0, I)$.

3 BGDiffSeg

Using the diffusion model introduced in preliminaries as our foundational framework, we focus on overcoming artifacts and blurred lesion boundaries in skin disease images. To achieve this, we specifically design the DBEM and the IBAM, collectively establishing BGDiffSeg. The comprehensive framework is depicted in Fig. 1.

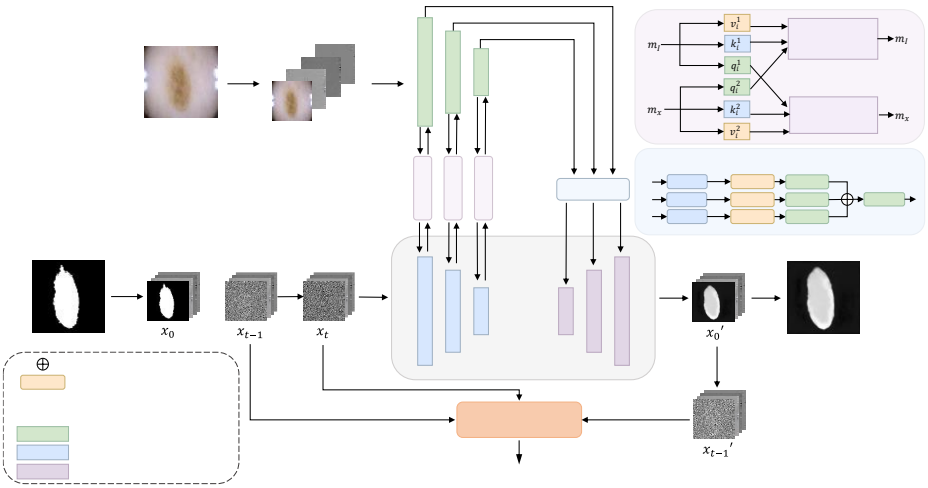

Fig. 1. The overall architecture of BGDiffSeg

3.1 Diffusion Boundary Enhancement Module (DBEM)

In our diffusion model, we meticulously design the Diffusion Boundary Enhancement Module (DBEM) inspired by [10], illustrated in Fig. 1, to enhance boundary features of skin lesions. The process begins with extracting features from the conditional encoder. To counteract the interference caused by high-frequency noise in the images, we first apply a wavelet transform to these features, isolating the low-frequency information that more closely represents the primary structure of the image. Following this, the Sobel operator is employed to extract boundary information from each low-frequency subband in both the horizontal G_x and vertical G_y directions, with the horizontal and vertical Sobel kernels defined as follows:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{3}$$

For each pixel, gradients in the horizontal and vertical directions are calculated. These gradients are combined to calculate the total gradient magnitude as follows:

$$G = \sqrt{{G_x}^2 + {G_y}^2} \tag{4}$$

After calculating the total gradient magnitude, we fuse it with the input feature map through element-wise multiplication to enhance boundary features. Subsequently, a simple convolution layer merges these enhanced features at various levels. The final enhanced feature is then multiplied with features from skip connections, improving the decoder-generated feature representation for more accurate image segmentation.

3.2 Interactive Bidirectional Attention Module (IBAM)

While traditional image generation tasks typically involve unidirectional information flow from conditional to generative encoding, we recognize that while the original image contains precise segmentation targets, they often blend with the background. Moreover, intermediate segmentation maps, although highlighting target areas, can sometimes inaccurately include non-target regions. To dynamically focus the model on lesion-relevant features and suppress potentially misleading artifacts, IBAM is developed, as shown in Fig. 1. This module facilitates feature-level interactions between conditional and generated images and fosters interactions between two encoders. It leverages multi-level features to enhance the model's ability to accurately identify and process subtle structural differences in medical images, thereby improving global recognition of target area boundaries. IBAM features a focused linear attention module [24], with a novel mapping function f_p for adjusting query and key features, and a depthwise convolution (DWC) module to increase feature diversity, achieveing high expressiveness with linear complexity. The focused linear attention module can be written as:

$$O = Sim\ (Q, K)\ V = \emptyset_p(Q)\emptyset_p(K)^T V + DWC(V) \tag{5}$$

$$where\ \emptyset_p(x) = f_p\ (ReLU(x))\ ,\ f_p(x) = \frac{\|x\|}{\|x^{**p}\|} x^{**p} \tag{6}$$

and x^{**p} represents element-wise power p of x. IBAM consists of two parallel branches, one for refined conditional encoding and the other for refined generative encoding. Specifically, we pair intermediate layer features m_I and m_x from each encoder to calculate query, key, and value vectors respectively. In the generative encoding attention branch, we propose calculating cross-attention between m_I and m_x, with m_I serving as the query matrix and m_x acting as both key and value matrices, followed by a residual connection to enhance the output. The calculation can be written as:

$$m_x = FLMHCA_x\ (m_I, m_x, m_x) + m_x \tag{7}$$

where $FLMHCA_x$ stands for the focused linear multi-head cross-attention in the generative encoding branch. A similar method is employed in the conditional encoding branch, as follows:

$$m_I = FLMHCA_I(m_x, m_I, m_I) + m_I \tag{8}$$

3.3 Loss Function

We employ the Least Squares GAN (LSGAN) loss function to train the diffusion discriminator in order to prevent gradient vanishing. The diffusion discriminator, denoted as $D_d(x_{t-1}, x_t, t)$, is trained to minimize this loss:

$$\mathcal{L}_{diff} = \sum_{t\geq 1} \mathbb{E}_{q(x_t)} \mathbb{E}_{q(x_{t-1}|x_t)} [(D_d(x_{t-1}, x_t, t) - 1)^2] + \mathbb{E}_{p_\theta(x_{t-1}|x_t)} \mathbb{E}_{q(x_{t-1}|x_t)} [D_d(x_{t-1}, x_t, t)^2] \quad (9)$$

where t represents the diffusion time step index. Additionally, we utilize the $Dice$ loss function to supervise the boundaries for more precise boundary information:

$$\mathcal{L}_{boundary} = Dice(y, \hat{y}) \quad (10)$$

Therefore, our loss function can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{diff} + \mathcal{L}_{boundary} \quad (11)$$

4 Experiments

4.1 Datasets and Implementation Details

We evaluate our model, BGDiffSeg, using three public skin lesion segmentation datasets: ISIC2016 [18][19], ISIC2017 [20][21], and ISIC2018 [22][23]. For ISIC2016, we use the default dataset partitioning method. For ISIC2017 and ISIC2018, datasets are split into training and testing sets at a 7:3 ratio. We resize all the images to a resolution of 256 × 256 and apply various data augmentation, including horizontal flipping, vertical flipping, and random rotation. Adam [25] is utilized as the optimizer, initialized with a learning rate of 0.1 and the CosineAnnealingLR [26] is employed as the scheduler with a minimum learning rate of 1e-5. A total of 200 epochs are trained with a batch size of 8. To evaluate our method, we employ Mean Intersection over Union (mIoU), Dice similarity score (DSC) as metrics. All the experiments are conducted using a NVIDIA RTX TITANX GPU with 12 GB RAM.

4.2 Comparative Results

We compare BGDiffSeg with widely-used medical image segmentation methods on several skin lesion segmentation datasets, using the same experimental protocol for each dataset to ensure fairness. For a fair comparison with MedSegDiff, we retrain it for 200,000 steps. The results, listed in Table 1, show BGDiffSeg outperforming all other methods across these datasets, demonstrating its effectiveness and generalizability. Fig. 2 offers a clearer visual representation of these results, with our model accurately predicting lesion boundaries closer to the ground truth, as opposed to others that either over-segment (e.g., MedT) or under-segment (e.g., TransUnet).

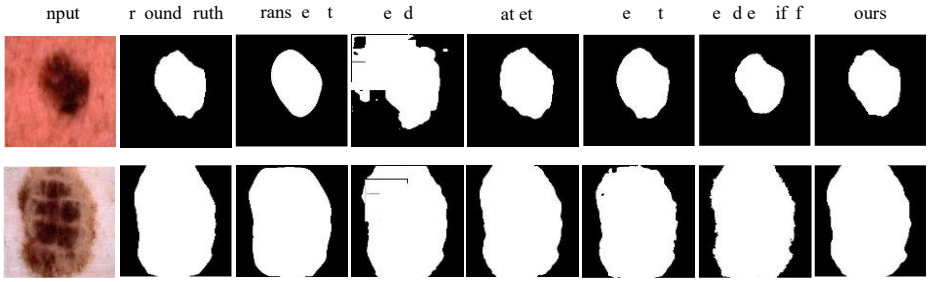

Fig. 2. Qualitative comparisons on ISIC2017 dataset (row 1) and ISIC2018 dataset (row 2)

Table 1. Comparative experimental results on the ISIC2017, ISIC2018 and ISIC2016(**Bold** indicates the best.)

Dataset	Model	mIoU(%)	DSC(%)
ISIC2017	UNet [3]	76.98	86.99
	UTNetV2 [27]	77.35	87.23
	TransUNet [7]	78.42	87.90
	MedT [8]	76.55	86.72
	Fat-Net [28]	78.36	87.86
	UNeXt [29]	76.57	86.74
	MedSegDiff [16]	73.41	84.67
	BGDiffSeg(ours)	**79.73**	**88.72**
ISIC2018	UNet [3]	77.86	87.55
	UNet++ [4]	78.31	87.83
	Att-UNet [5]	78.43	87.91
	UTNetV2 [27]	78.97	88.25
	TransUNet [7]	79.74	88.73
	MedT [8]	76.96	86.98
	Fat-Net [28]	79.74	88.73
	UNeXt [29]	78.01	87.64
	MedSegDiff [16]	74.71	85.52
	BGDiffSeg(ours)	**80.22**	**89.02**
ISIC2016	UNet [3]	80.25	87.81
	UNet++ [4]	81.84	88.93
	Att-UNet [5]	79.70	87.43
	TransUNet [7]	84.89	91.26
	MedT [8]	83.35	90.92
	UNeXt [29]	84.32	91.49
	MedSegDiff [16]	81.97	89.81
	BGDiffSeg(ours)	**85.52**	**92.19**

Additionally, we compare the training resources and sampling times required by MedSegDiff and BGDiffSeg during the training process, as shown in Table 2. The table reveals that our BGDiffSeg requires fewer GPU resources and FLOPs calculations. Moreover, while MedSegDiff takes approximately 85.5s to sample and generate a segmentation map, BGDiffSeg requires only about 0.2s, making it significantly faster than MedSegDiff by a factor of over 400 times.

Table 2. Comparison of training resources and sampling time between MedSegDiff and BGDiffSeg (**Bold** indicates the best.)

Model	Params(M)↓	FLOPs(G)↓	MEM(G)↓	Time(s)↓
MedSegDiff	129.34	2083.06	25.03	85.51
BGDiffSeg(ours)	**34.37**	**309.78**	**5.18**	**0.23**

4.3 Ablation Results

We conduct extensive ablation experiments on the ISIC2017 dataset to validate the effectiveness of our proposed modules, with results shown in Table 3. Our baseline model, based on Wavediff, integrates features from the conditional encoder and the generative encoder via direct addition, as shown in Table 3(a). The ablation on IBAM, shown in Table 3(b), indicates a significant performance boost with the addition of IBAM, enhancing interaction between the two encoders and increasing mIoU and DSC by 2.67% and 1.7%, respectively. To validate the necessity of the bidirectional information flow between encoders, we remove the branch that integrates generative encoding into conditional encoding within IBAM. This deletion significantly reduces IBAM's effectiveness, underscoring the importance of this key design. Table 3(c) outlines the ablation on DBEM. Adding DBEM, which focuses on boundary features enhancement, notably improves performance, raising mIoU and DSC by 1.6% and 1.03%. Moreover, removing low-frequency feature extraction within DBEM leads to a performance drop, confirming that high-frequency noise interferes with boundary detection and validating the effectiveness of our key designs.

Table 3. Our ablation studies on the ISIC2017 dataset cover: (a) the result of baseline, (b) the ablation on IBAM, and (c) the ablation on DBEM.

Type	Model	mIoU(%)	DSC(%)
(a)	Baseline	75.88	86.28
(b)	Baseline + IBAM	78.55	87.98
	w/o generative encoding attention branch	77.29	87.19
(c)	Baseline + DBEM	77.48	87.31
	w/o wavelet transform	77.21	87.14

5 Conclusions

In this paper, we propose BGDiffSeg, a new fast diffusion-based skin lesion segmentation model. Leveraging WaveDiff, alongside innovations like the Diffusion Boundary Enhancement Module (DBEM) and Interactive Bidirectional Attention Module (IBAM), BGDiffSeg generates precise segmentation masks with limited computational resources and time. Extensive experiments demonstrate its superiority quantitatively and qualitatively, highlighting its potential to enhance subsequent tasks in an end-to-end manner.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. Ca Cancer J Clin **73**, 17-48 (2023)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241. Springer, Munich (2015)
4. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3-11. Springer, Granada (2018)
5. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424-432. Springer, Istanbul (2016)
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
8. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 36-46. Springer, Strasbourg (2021)
9. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205-218. Springer, Tel Aviv (2022)
10. Lin, Y., Zhang, D., Fang, X., Chen, Y., Cheng, K.-T., Chen, H.: Rethinking boundary detection in deep learning models for medical image segmentation. In: International Conference on Information Processing in Medical Imaging, pp. 730-742. Springer, Switzerland (2023)

11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840-6851 (2020)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684-10695. IEEE, New Orleans (2022)
13. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479-36494 (2022)
14. Chung, H., Ryu, D., McCann, M.T., Klasky, M.L., Ye, J.C.: Solving 3d inverse problems using pre-trained 2d diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22542-22551. IEEE, Vancouver (2023)
15. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. arXiv preprint arXiv:2112.07804 (2021)
16. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)
17. Phung, H., Dao, Q., Tran, A.: Wavelet diffusion models are fast and scalable image generators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10199-10208. IEEE, Vancouver (2023)
18. ISIC2016 Homepage, https://challenge.isic-archive.com/data/#2016, last accessed 2024/7/7
19. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397 (2016)
20. ISIC2017 Homepage, https://challenge.isic-archive.com/data/#2017, last accessed 2024/7/7
21. Berseth, M.: ISIC 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523 (2017)
22. ISIC2018 Homepage, https://challenge.isic-archive.com/data/#2018, last accessed 2024/7/7
23. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
24. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: Flatten transformer: Vision transformer using focused linear attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5961-5971. IEEE, Paris (2023)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
27. Gao, Y., Zhou, M., Liu, D., Metaxas, D.: A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv preprint arXiv:2203.00131 (2022)
28. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Medical image analysis **76**, 102327 (2022)
29. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 23-33. Springer, Singapore (2022)