

MedDiff-FT: Data-Efficient Diffusion Model Fine-tuning with Structural Guidance for Controllable Medical Image Synthesis

Jianhao Xie¹, Ziang Zhang¹, Zhenyu Weng², Yuesheng Zhu¹, and Guibo Luo^{*1}

¹ Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media
Technology, Peking University Shenzhen Graduate School

² South China University of Technology, China

Abstract. Recent advancements in deep learning for medical image segmentation are often limited by the scarcity of high-quality training data. While diffusion models provide a potential solution by generating synthetic images, their effectiveness in medical imaging remains constrained due to their reliance on large-scale medical datasets and the need for higher image quality. To address these challenges, we present *MedDiff-FT*, a controllable medical image generation method that fine-tunes a diffusion foundation model to produce medical images with structural dependency and domain specificity in a data-efficient manner. During inference, a dynamic adaptive guiding mask enforces spatial constraints to ensure anatomically coherent synthesis, while a lightweight stochastic mask generator enhances diversity through hierarchical randomness injection. Additionally, an automated quality assessment protocol filters suboptimal outputs using feature-space metrics, followed by mask corrosion to refine fidelity. Evaluated on five medical segmentation datasets, *MedDiff-FT*'s synthetic image-mask pairs improve SOTA method's segmentation performance by an average of 1% in Dice score. The framework effectively balances generation quality, diversity, and computational efficiency, offering a practical solution for medical data augmentation. The code is available at <https://github.com/JianhaoXie1/MedDiff-FT>.

Keywords: Medical Image Segmentation · Diffusion Model · Controlled Generation.

1 Introduction

Medical image segmentation is a key step in medical image processing and analysis, which involves separating and extracting specific structures or regions (e.g., organs, tissues, lesions, etc.) from the background or other structures in a medical image. In recent years, the development of deep learning, especially convolutional neural networks [3,6], has greatly advanced medical image segmentation. Early approaches like U-Net [1] introduced an encoder-decoder architecture with cross-layer connections, achieving remarkable success in medical imaging. Subsequent

* Corresponding author: luogb@pku.edu.cn

advancements such as nnU-Net [2] further optimized pre-processing and network adaptation based on medical data characteristics. With the rise of Transformer-

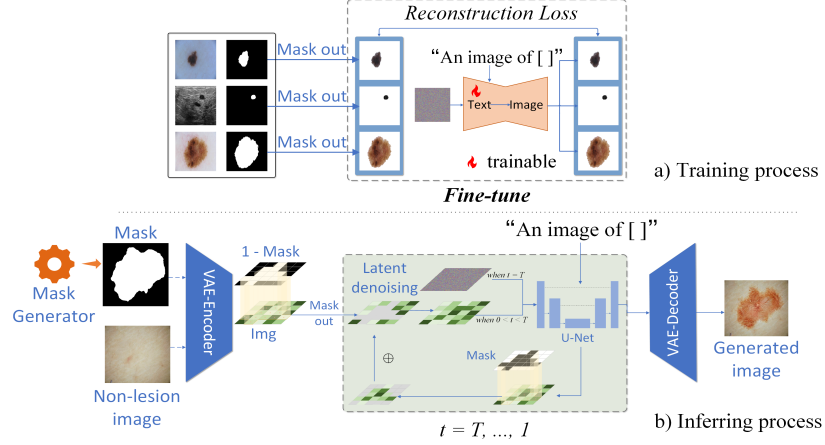


Fig. 1. Model structure. The figure is divided into parts a and b. Part a of the figure represents the training process, and part b represents the inference process. Both processes are effective, fast, and do not take much memory .

based architectures, Vision Transformer (ViT) [4] and its variants have been adapted for segmentation, yet their heavy data requirements and susceptibility to overfitting on small datasets remain challenges. Deep learning methods automatically learn feature representations without manual design, enabling robust performance in complex scenarios. However, medical images differ from natural images in data specificity, privacy constraints, and limited public availability, resulting in smaller datasets. This scarcity conflicts with the data-hungry nature of deep learning models, particularly Transformer-based architectures, making it difficult to achieve satisfactory segmentation with few labeled images.

To address data scarcity, diffusion models have emerged as powerful tools for synthetic data generation. Milestone works like DDPM [7] and DDIM [8] established frameworks for iterative denoising, while Stable Diffusion improved efficiency via latent space mapping. Further, fine-tuning methods such as Dream-Booth [12] and LoRA [13] adapt pre-trained models to downstream tasks, and controllable generation techniques like ControlNet [11] enable structural guidance. However, these methods predominantly target natural images or classification tasks, with limited exploration of medical image-mask pair generation for segmentation. This limitation stems from fundamental differences between natural and medical imaging domains.

Current medical image generation methods [9,10,25,27] focus on classification or single-image synthesis, lacking mechanisms to produce paired images and masks. This gap hinders their utility for segmentation tasks. Additionally,

fine-tuning diffusion models on limited medical data remains resource-intensive and underexplored. Some existing work [26] can also generate the corresponding masks, though they rely on DDPM trained from scratch, which can be resource-intensive.

Generating synthetic medical image-mask pairs for segmentation introduces unique challenges: (1) Structural dependency: Synthetic images must preserve precise spatial relationships between organs or lesions and their corresponding masks, requiring pixel-level alignment beyond semantic plausibility. (2) Domain specificity: Medical imaging modalities exhibit distinct noise patterns and intensity distributions that synthetic data must replicate to avoid domain shifts. (3) Data efficiency: Fine-tuning large generative models on limited medical data risks overfitting or mode collapse, especially when training pairs number in the hundreds.

To bridge these gaps, we propose a lightweight, data-efficient method for controllable medical image-mask pair generation. Our method fine-tunes Stable Diffusion with limited data (under 30 minutes and 24GB memory) and uses automated quality assessment protocol filters to enhance reliability and diversity. In the inference phase, we use guide mask for controllable generation to achieve controllable shape and location of the lesion area. We also use a lightweight diffusion model as a mask generator to improve the versatility of the generated images. Experiments on five segmentation tasks demonstrate that models trained with our synthetic data achieve an average 3% accuracy improvement.

Our contributions are: 1: A resource-efficient fine-tuning framework for medical image-mask pair generation. 2: A lightweight diffusion model with quality screening tailored for segmentation. 3: Empirical validation of synthetic data efficacy in downstream tasks.

2 Method

We present our work from two aspects: the paradigm of controlling diffusion models to generate reliable medical training data and the strategy of data selection. Our medical image generation framework with structural dependency and domain specificity requires only a few image-mask pairs. During training, a dynamic adaptive guiding mask highlights lesion regions, enabling fine-tuning of the Stable Diffusion model to focus on domain-specific feature learning. For inference, this adaptive mask guidance mechanism spatially constrains generation within predefined anatomical regions, ensuring precise controlled synthesis. We further develop a lightweight stochastic mask generator that produces both lesion-repairing patterns and anatomically plausible non-lesion maps, effectively expanding the data diversity. The post-processing phase incorporates an automated quality assessment protocol to ensure image fidelity, complemented by morphological corrosion operations to refine annotation boundaries.

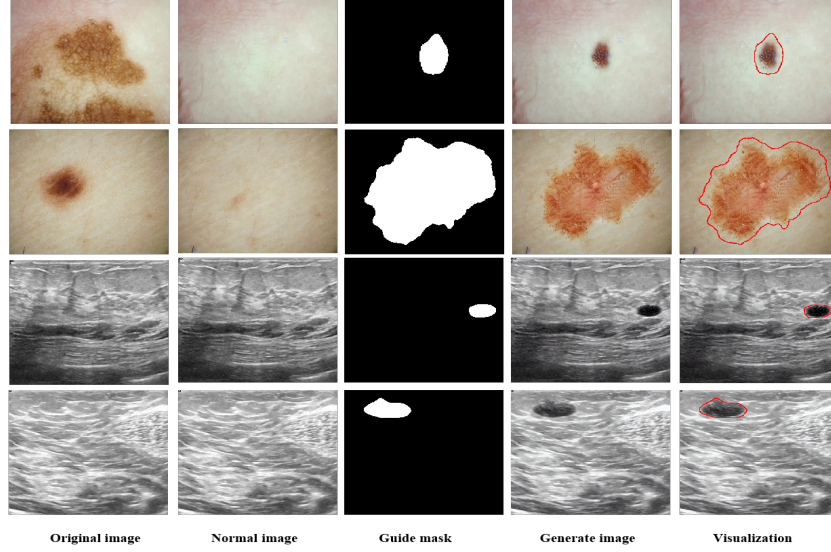


Fig. 2. The figure presents generated results from our method. The first two rows display skin images, where background images are restored from originals using our approach. The last two rows show breast ultrasound images; since background images are included in the dataset, these remain identical to the originals. Generated images are produced by applying guiding masks to background images, with the final column demonstrating controllable generation results.

2.1 Image Generation Framework

Controlled generation based on limited number of data Training a diffusion model from scratch with limited data (tens to thousands of samples) is challenging. Instead, we fine-tune the Stable Diffusion 1.5 [16]. During training, we use specialized trigger words to reference new images, allowing the fine-tuned model to generate images corresponding to these concepts during inference, similar to the Textual Inversion [15] technique. However, unlike Textual Inversion, we unfreeze the parameters of U-Net for training during the training phase.

The controlled generation paradigm consists of training and inference processes. **In training**, the model focuses on lesion regions rather than the entire image. The original image and its corresponding mask are used to extract the lesion region, which is then fine-tuned. And the input to the model is noise, which generates images guided by specific prompt, and then the generated images are used to do a reconstructed loss calculation with the original images provided by us. After such training, it makes the fine-tuned model to generate a specific type of image under a specific prompt. A text prompt, such as "An image of hta," is designed to map to the lesion area, avoiding conflicts with Stable Diffusion's original vocabulary. This process allows the model to concentrate on the lesion region, requiring only around 30 samples and completing in 20 minutes with less

than 24GB memory usage.

$$\mathcal{L} = \mathbb{E}_{t,X,c,\epsilon} [w_t \| M \odot (\hat{X}_\theta(\alpha_t X + \beta_t \epsilon, c) - X) \|_2^2] \quad (1)$$

The equation 1 is the loss function for the training phase. $\hat{X}_\theta$ is a pre-trained text-to image diffusion model like Stable Diffusion 1.5, c is a conditioning vector made by text prompt, X is the ground-truth image, α_t, β_t, w_t are terms that control the noise schedule and sample quality, M is the region mask matrix and ϵ is an initial noise map.

For inference, three components are needed: the fine-tuned model, a background (non-lesion) image, and a mask. The background image serves as the backdrop, while the mask directs the lesion region’s location and shape, forming an image-mask pair for subsequent experiments. During inference, the model generates images in specified regions while maintaining the integrity of the rest of the image. The denoising process uses two latent vectors: one from the original image and one from the preceding denoising step. These vectors are combined to produce the intermediate denoised image, ensuring controlled generation. The full structure can be seen in Fig.1.

$$x_{t-1} = M \odot \left[\frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, t) \right) \right] + (1 - M) \odot \mathcal{F}(x_t^{prev}) \quad (2)$$

The denoising process in the inference phase is formulated as equation 2. t is timesteps, M is region mask matrix, α_t, β_t are diffusion scheduling coefficients, ϵ_θ is noise predicted by U-Net, x_t^{prev} is latent variable state from the previous step and $\mathcal{F}$ is feature preservation function.

Explanation of Model Components Mask Generator: It is designed to produce highly diverse masks for guidance purposes. Implemented as a DDPM based on UNet architecture, trained on mask images. The network architecture consists of four hierarchical layers with progressively increasing channel dimensions [64,128,256,512]. Non-lesion Image Generator: Built upon Stable Diffusion 1.5 architecture, in the training phase, we use the data pairs: lesion Image, and invert the mask (swapping 0, 1 values). Then perform the original fine-tune, the inverted mask allows the model to learn how to generate healthy regions. In the inference phase, the input is the lesion image with original mask, like Fig.1 inference phase, the model can repair the lesion region to a healthy region in the mask region.

Diversity of generation In order to further enhance the diversity of generations, we adopt a dual-pronged approach to address the diversity problem.

The first strategy enhances lesion-free background diversity through conditional image restoration. When datasets contain background images, these are directly utilized; otherwise, a diffusion model trained with inverted masks (focusing attention on healthy regions) reconstructs anatomically plausible tissue (Fig.2). This enables defect repair by generating non-lesion regions guided by input masks, ultimately producing pathology-free backgrounds.

The second strategy addresses mask diversity limitations of conventional augmentations (flipping/erosion) by training a compact diffusion model specifically for mask synthesis. These generated masks condition the main diffusion

Table 1. Segmentation performance results. There are three experiments for each of these methods, original images, original images plus 1500 generated image pairs , and original images plus 2750 generated image pairs. The values in the table are Dice.

Method		PH ²	ISIC-2017	ISIC-2018	BUSI	DDTI
UPerNet	original	88.97	83.05	85.01	66.72	79.89
	original+1500	92.04	83.37	86.39	66.76	77.96
	original+2750	91.47	83.45	87.32	71.54	82.34
DeepLabV3 Plus	original	88.1	83.84	86.67	70.55	77.18
	original+1500	89.04	84.11	87.19	73.08	78.84
	original+2750	89.53	84.7	87.87	74.05	83.01
Swin Transformer	original	89.46	81.19	86.03	59.55	74.03
	original+1500	91.38	81.57	86.68	62.38	74.78
	original+2750	92.16	84.28	86.55	63.63	75.22
nnU-Net	original	94.26	83.77	87.43	77.01	79.45
	original +1500	94.75	84.17	88.06	78.1	80.51
	original+2750	94.72	84.92	88.15	78.69	81.67

model, creating medically varied images that surpass traditional augmentation constraints, significantly expanding dataset diversity.

2.2 Automated Quality Assessment Protocol

Due to variability in generated data quality, a filtering process is implemented. Generated images should exhibit a degree of similarity with real images but not be too similar or dissimilar. The DINOv2 [14] model is used as a feature extractor to calculate cosine similarity between generated and original images. Images with excessively high or low similarity scores are excluded.

Additionally, a corrosion operation is applied to the mask edges to improve the fit between the mask and the generated lesion region, enhancing annotation quality. Ablation experiments in the experimental chapter demonstrate the validity of this filtering process.

This paradigm enables controlled generation with limited samples, applicable across diverse backgrounds and lesion types, locations, and shapes.

In the experimental chapter, some relevant ablation experiments are conducted to demonstrate the validity of data filtering.

3 Experiments

The experiment section is divided into three parts: datasets and baselines, results, and ablation experiments. These sections explore the impact of data filtering strategies, hyperparameters, and the use of generated data on the final segmentation results.

3.1 Datasets and baselines

Five publicly available datasets were used: ISIC-2017 [20], ISIC-2018 [21], PH2 [19], BUSI [18], and DDTI [17]. These datasets cover various organs (skin, thyroid, breast) and modalities (picture, ultrasound). Each dataset was divided into

Table 2. Comparison of controllable generation methods. This includes the ControlNet and T2i-Adapter methods, and the results on the ISIC-2017 and BUSI datasets. Bolded numbers are the best values. The values in the table are Dice

Dataset	Method		UPerNet	DeepLabV3 Plus	Swin Transformer	nnU-Net
ISIC-2017	ControlNet	Original	83.05	83.84	81.19	83.77
		original+1500	83.29	83.31	81.06	83.67
	T2I-adapter	original+2750	83.4	83.77	81.74	84.48
		original+1500	83.54	83.72	81.22	84.47
	Ours	original+2750	82.84	82.71	79.01	84.26
		original+1500	83.37	84.11	81.57	84.17
		original+2750	83.45	84.7	84.28	84.92
		Original	66.72	70.55	59.55	77.01
BUSI	ControlNet	original+1500	69.3	71.76	59.69	78.22
		original+2750	69.69	70.33	57.9	78.04
	T2I-adapter	original+1500	69.42	72.41	59.09	76.78
		original+2750	69.81	72.54	60.93	78.30
	Ours	original+1500	66.76	73.08	62.38	78.1
		original+2750	71.54	74.05	63.63	78.69
		Original	66.72	70.55	59.55	77.01
		original+1500	69.3	71.76	59.69	78.22

training, validation, and test sets. For ISIC-2017 and ISIC-2018, the provided test sets were used directly. For PH2, BUSI, and DDTI, 20% of the data was allocated to the test and validation sets, with the remainder used for training.

Four segmentation models were chosen as baselines: UPerNet [22], DeepLabV3 Plus [23], Swin Transformer [5], and nnU-Net. These models represent both traditional and state-of-the-art approaches in medical image segmentation. Each model was trained on the training sets of the five datasets, with the best-performing weights on the validation set selected for testing.

The experimental setup was divided into two phases: segmentation and generation. The segmentation phase used Python 3.8.19, Pytorch 1.13.1+cu117, a batch size of 8, an input image size of 512x512, and 800 epochs. The loss function combined focal loss and dice loss, with dice vs. iou as the evaluation criterion. The generation phase used Python 3.10.14 and Torch 2.2.2, with a batch size of 2, 2000 iterations, and an input image size of 512x512.

3.2 Results

The original line experiments used 30 image-mask pairs for fine-tuning. Subsequently, 50 images were generated based on 50 background images, with two sets of generation quantities: 1,500 and 2,750 images. These generated image-mask pairs were added to the training sets, while the validation and test sets remained unchanged.

As shown in Table 1, incorporating 1,500 and 2,750 additional images into the training sets resulted in higher DICE scores compared to the original line results. The improvement was more pronounced with more generated data pairs, as the increased diversity and quality of the training data enhanced the model’s generalization and stability. Two controllable generation methods, ControlNet and T2I-Adapter [24], were also tested using the same number of generated image-mask pairs. These methods provided some assistance for downstream segmentation tasks but were less stable and produced suboptimal results compared

Table 3. The result of post-processing. Each method includes two comparisons: data filtering and corrosion mask. The baseline is the original image and 2750 generated images are added for the training. The evaluation metric utilized is the DICE score.

Method		PH2	ISIC-2017	ISIC-2018	BUSI	DDTI
UPerNet	Baseline	91.47	83.45	87.32	71.54	82.34
	Baseline+filter	93.76	84.11	87.88	71.75	83.26
	Baseline+corrode	93.23	84.57	87.32	71.76	83.46
DeepLabV3 Plus	Baseline	89.53	84.7	87.87	74.05	83.01
	Baseline+filter	90.91	84.74	87.9	74.54	83.25
	Baseline+corrode	91.09	84.73	87.55	73.73	83.27
Swin Transformer	Baseline	92.16	84.28	86.55	63.63	75.22
	Baseline+filter	92.1	84.33	86.33	64.03	75.26
	Baseline+corrode	91.19	84.3	86.74	63.51	75.37
nnU-Net	Baseline	94.72	84.92	88.15	78.69	81.67
	Baseline+filter	94.69	84.96	88.24	78.88	81.32
	Baseline+corrode	95.02	84.70	88.87	78.34	82.25

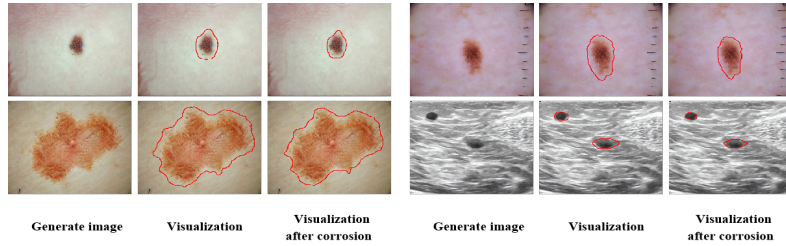


Fig. 3. Erosion effect. From left to right: generate image, visualization of image and corresponding mask, and visualization after erosion mask.

to our approach. This discrepancy may be due to the limited training data (30 pieces), which was insufficient for ControlNet and T2I-Adapter to achieve optimal results.

3.3 Ablation experiments

The ablation experiments focused on two aspects: the number of background images used for generation and the role of data filtering.

For the number of background images, experiments were conducted with 20, 30, and 50 background images. As shown in Table 4, the segmentation results improved with more background images, as increased diversity in the generated data led to better model performance.

Data filtering was applied to the generated images, removing those that were too similar or dissimilar. As shown in Table 3, filtering reduced the amount of generated data but improved segmentation performance by enhancing the quality of the training dataset. Additionally, corroding the masks of the generated images further improved the results, as it made the masks more accurate and better aligned with the lesions in the images.

In summary, the experiments demonstrated that increasing the diversity and quality of training data through controlled generation and filtering significantly

Table 4. Results of ablation experiment. Among them, 20, 30, and 50 are the number of background images used during generation. Each experiment is adding 2750 generatio images to the original images. The values in the table are Dice.

Dataset	Method	20	30	50	Method	20	30	50
ISIC-2017	UPerNet	83.12	83.42	83.45	Swin Transformer	81.6	81.78	84.28
	DeepLabV3 Plus	84.09	84.58	84.7	nnU-Net	85.31	85.37	84.92
ISIC-2018	UPerNet	86.25	86.7	88.45	Swin Transformer	86.18	86.58	86.55
	DeepLabV3 Plus	86.55	86.71	87.87	nnU-Net	87.63	87.89	88.15

improved segmentation performance. Our approach outperformed other controllable generation methods, highlighting its potential for addressing the challenge of limited medical imaging data.

4 Conclusion

This paper proposes a resource-efficient controllable generation paradigm for diffusion models using limited data. By fine-tuning Stable Diffusion with minimal training data, our method produces high-quality image-mask pairs validated through downstream segmentation tasks. The framework demonstrates hardware-friendly implementation with low-resource training/inference requirements. Extensive validation across five real-world datasets using four segmentation architectures confirms the effectiveness of the synthetic data.

Acknowledgments This work is supported by 2023 Shenzhen sustainable supporting funds for colleges and universities (20231121165240001), Shenzhen Science and Technology Program (JCYJ20230807120800001), Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (2024B1212010006).

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

References

1. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, pages 234–241(2015)
2. Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211(2021)
3. O’Shea Keiron, and Ryan Nash. An Introduction to Convolutional Neural Networks. *arxiv:1511.08458*(2015)
4. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*(2021)
5. Liu Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012-10022(2021)

6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770-778(2016)
7. Ho Jonathan, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *arXiv:2006.11239*(2020)
8. Song Jiaming, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. *arXiv:2010.02502*(2022)
9. Mohamed Akrouf, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, and others. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 99-109(2023)
10. Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Polacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, and others. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv:2211.12737*(2022)
11. Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836-3847(2023)
12. Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *ArXiv Preprint Arxiv:2208.12242*(2022)
13. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*(2021)
14. Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and others. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*(2023)
15. Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv:2208.01618*(2022)
16. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684-10695(2022)
17. Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. 2015. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, SPIE, 92870W(2015)
18. Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. BUSIS: a benchmark for breast ultrasound image segmentation. In *Healthcare*, MDPI, 729(2022)
19. Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. PH 2-A dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, IEEE, 5437-5440(2013)
20. Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and others. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by

- the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE, 168–172(2018)
21. Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, and others. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv:1902.03368*(2019)
 22. Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434(2018)
 23. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818(2018)
 24. Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4296–4304(2024)
 25. Wang Janet, Yunsung Chung, Zhengming Ding, and Jihun Hamm. From Majority to Minority: A Diffusion-Based Augmentation for Underrepresented Groups in Skin Lesion Analysis. *arxiv:2406.18375*(2024)
 26. Hanwen Zhang, Mingzhi Chen, Yuxi Liu, Guibo Luo, Yuesheng Zhu, "Non-IID Medical Image Segmentation Based on Cascaded Diffusion Model for Diverse Multi-Center Scenarios", *IEEE Journal of Biomedical and Health Informatics*(2025)
 27. Jiaying Zhang, Guibo Luo, Ziang Zhang, Yuesheng Zhu, "Data Augmentation in Class-Conditional Diffusion Model for Semi-Supervised Medical Image Segmentation", *IJCNN*(2024)