Guided Diffusion: Balancing Fidelity and Diversity in Dataset Augmentation with EfficientNet and Gaussian Noise

Tianyi Ouyang*

Glasgow College, University of Electronic Science and Technology of China, 611731, Chengdu, Sichuan, China

Abstract. The denoising diffusion probabilistic model (DDPM) has recently attracted massive attention due to its better capability of synthesizing high-quality and diverse synthetic data than generative adversarial network (GAN), paving the way for its application in data augmentation scenarios. However, balancing fidelity and diversity remains a challenge. To address the problem, a novel architecture is proposed, incorporating EfficientNet to extract features from the original dataset and fuse them with those of noise samples, guiding the denoising process and ensuring fidelity between synthetic samples and the original data. Additionally, random Gaussian noise is introduced to the UNet bottleneck output at each timestep to enhance diversity. A pre-trained CNN classification network follows to ensure label consistency between the reference and the synthetic images. The approach is evaluated through experiments on lung cancer prediction using a chest CT-scan dataset, achieving a 13.6% improvement in classification accuracy over baseline methods, 9.8% over the traditional cropping and rotation approach, and 4.1% over the GAN-based approach. These results validate the effectiveness of the proposed method for dataset augmentation.

1 Introduction

Deep learning (DL) is increasingly pivotal in diverse vision comprehension applications such as autonomous driving and object detection in modern computer vision. However, its performance can be compromised due to dataset scarcity, especially in the medical area, where privacy regulations and the rarity of specific syndromes become barriers to deep learning training and deployment [1].

A dataset augmentation technique is proposed to mitigate the dataset limitation. Along with the development of dataset augmentation, from random cropping and rotation to more recent popular approaches such as CutMix, they have exhibited an effect on alleviating dataset constraints [2]. However, these approaches failed to bring new semantic information and introduce diversity to the dataset to enhance the generality of DL further. A generative adversarial network (GAN), which synthesizes samples based on the original content in the

* Corresponding author: 2021190501014@std.uestc.edu.cn

dataset, has presented a preliminary promise in diversifying the dataset by synthesizing new samples [3]. Recently, the denoising diffusion probabilistic model (DDPM), which produces new samples via forward noise diffusion and backward denoising processes, has shown an impressive performance in image synthesis and has become a new latent solution to dataset augmentation tasks [4]. According to [5], DDPM outperforms its GAN-based counterparts in high-quality image synthesis tasks, determining its value for further research. Despite the advantages of DDPM, the trade-off between diversity and fidelity becomes a critical challenge. DDPM might present undue diversity and synthesize samples that far deviate from the original dataset, while if the fidelity is paid more attention to, the diversity of DDPM becomes unsatisfactory [6].

Grounded on the limitation, a refined DDPM is proposed to balance the model's fidelity and diversity in this paper. Specifically, an EfficientNet is introduced into DDPM to extract the critical semantic information of the reference image from the original dataset, which is fused with the down-sampling block's output to provide a reference and constraint for the UNet's noise prediction. Random Gaussian noise is added to the UNet's bottleneck output to enhance the synthetic diversity. The refined model experiments on a lung cancer prediction scenario, a typical dataset scarcity situation due to privacy regulations, which, in essence, is a classification task based on the chest CT-scan images. Considering the possibility of fidelity distortion, though a reference is provided by EfficientNet, due to the introduced Gaussian noise, a pseudo-label is proffered for each synthetic sample using a classification network pre-trained on the original training dataset. The pseudo-label is kept if it is consistent with the reference image. The labeled synthetic samples are eventually incorporated into the training dataset, while the unlabeled ones are discarded to complete the dataset augmentation tasks.

The refined model's performance is evaluated on the dataset augmentation task on a chest CT-scan dataset with traditional dataset augmentation approaches (GAN-based, random cropping, and rotation). The result shows that after dataset augmentation via the refined DDPM, the classification accuracy increases by 9.8% compared with the random cropping and rotation counterpart and 4.1% compared with the GAN-based counterpart. The contribution is concluded as follows.

(1) Enhance DDPM's fidelity of synthetic images by integrating EfficientNet with DDPM to provide a preliminary reference and constraint for the noise prediction process at each timestep.

(2) Introduce a random disturbance to the fused feature to balance DDPM's diversity of synthetic images by adding random Gaussian noise to UNet's bottleneck output.

2 Methodology

2.1 Framework overview

Fig. 1 is an overview of the refined DDPM framework. This model fuses the features extracted from the training dataset with those from the noise sample during the denoising process of DDPM to enhance the fidelity between the synthetic samples and the original training dataset. The model also introduces diversity to the synthetic samples by adding random Gaussian noise to the UNet's bottleneck's output at each timestep.

The refined DDPM framework comprises an EfficientNet, a UNet, and a pre-trained CNN classification network. At each timestep t during the denoising process, the semantic features F_R, and classification results C_R of the reference image I_R from the original training dataset are obtained by EfficientNet. The extracted features are then fused with the features of the noise sample X_t from the UNet down-sampling block's output F_N by a feature fusion block

(FF). This fusion process offers the denoising process a reference and constraint to ensure better fidelity. The fused features pass through the UNet's bottleneck and are added with random Gaussian noise ϵ_G to improve diversity during the subsequent denoising process. After image synthesis, the synthesized image X_0 is forwarded to a CNN classification network pre-trained on the original training dataset to generate a pseudo-label L_S for the synthesized image. This pseudo-label is kept if it is consistent with the label of the reference image. Eventually, the labeled synthesized images are incorporated into the original training dataset, and the remaining unlabeled synthesized images are discarded to accomplish the dataset augmentation task.

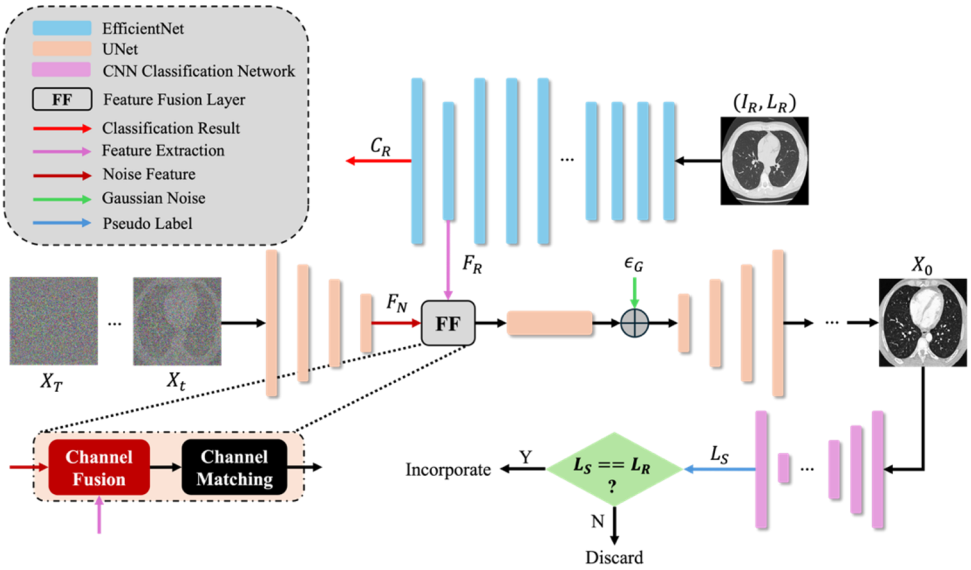

Fig. 1. The framework of the refined DDPM (Photo/Picture credit: Original).

2.2 EfficientNet

EfficientNet plays a vital role in feature extraction in the refined DDPM. In this framework, an EfficientNet for classification tasks is utilized. EfficientNet, compared with traditional convolutional neural networks (CNN), makes a better trade-off between computational complexity and feature comprehension capability [7]. In this framework, the model EfficientNetB0 is adopted as a feature extractor, which has the fewest parameters compared with its counterparts, with an acceptable sacrifice of accuracy. At each timestep during the denoising stage, the feature of the image from the original training dataset extracted by the last MBConv layer before the full-connection layer is fused with the feature of the noise sample to provide a preliminary reference for the UNet on noise prediction to mitigate the randomness.

2.3 Feature fusion block

Feature Fusion Block (FF) aims to combine the semantic features from two sources and adjust the dimension of the combined features to satisfy the UNet bottleneck input's dimension requirement while minimizing the information loss. Fig. 2 shows the details of FF. FF contains a channel fusion module and a channel adjustment module. The semantic features of the reference image and the noise sample have identical resolutions and different channel

numbers. The channel fusion module first receives features from two sources and combines them on the channel dimension. The combined features are then forwarded to the channel adjustment module. The channel adjustment module adjusts the channel number to match the dimension requirement of the UNet bottleneck's input through a learnable 1x1 convolution layer. The adjustment can be dynamically optimized through a learnable 1x1 convolutional layer to minimize the information loss.

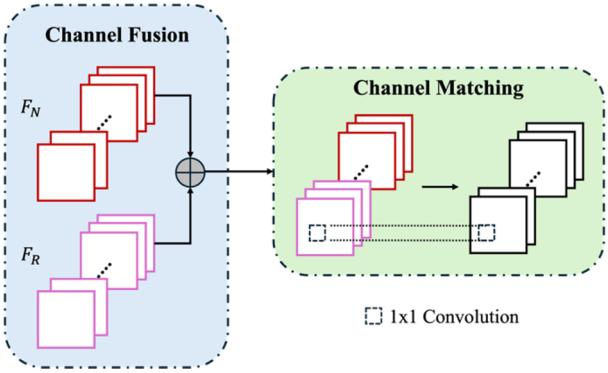

Fig. 2. Structure of feature fusion (FF) block (Photo/Picture credit: Original).

2.4 Random gaussian noise

At each timestep, random Gaussian noise is added to the UNet's bottleneck output to enhance the diversity of the synthetic sample. According to [8], features from the bottleneck block of the UNet contain the most abstract and comprehensive semantic information on the noise sample. Adding a random Gaussian noise to the bottleneck output achieves higher diversity than adding it to UNet's other modules.

2.5 Pseudo labeling

Though the preliminary reference discussed in 2.2 is provided, the possibility of the deviation of the synthetic image from the original label still exists due to the addition of the random Gaussian noise mentioned in 2.4. To mitigate its adverse impact, a classification network, a CNN, pre-trained on the original training dataset is introduced to provide a pseudo label for each synthetic sample. The CNN receives the synthetic sample and provides a label with the prediction confidence. The synthetic sample with the pseudo label whose prediction confidence satisfies the threshold is incorporated into the training dataset, while those who fail to meet the threshold are dropped out.

2.6 Loss function

A typical loss function (shown in Equation (1)) designed in [6] is adopted to optimize the noise prediction capability of the UNet during the denoising stage.

$$L_{noise} = \left\| \epsilon_t - U\left(x_t, t, E(I_R)\right) \right\|_2^2 \tag{1}$$

Where L_{noise} denotes the loss function of UNet, ϵ_t denotes the actual noise added at timestep t, $U(\cdot)$ denotes the noise prediction of the UNet, which receives the noise sample at

timestep t, the timestep t, and the features of the original image I_R from the training dataset extracted by the EfficientNet $E(\cdot)$.

Equation 2 introduces the cross-entropy loss function typical for classification tasks to synchronously optimize the EfficientNet's feature extraction capability.

$$L_{en} = H(C_R, y_o) \tag{2}$$

Where L_{en} denotes the loss function of EfficientNet, $H(\cdot)$ represents the cross-entropy, C_{R} denotes the classification results, and y_o denotes the ground-truth label of the image I_o.

The total loss, L_{total}, is designed in Equation 3.

$$L_{total} = L_{noise} + \lambda \cdot L_{en} \tag{3}$$

Where λ denotes the weighting parameter.

3 Experiment

3.1 Dataset

The refined DDPM framework is trained on an open-source chest CT-scan image dataset on Kaggle designed to train the neural network for lung cancer prediction [9]. This dataset consists of four statuses: normal, adenocarcinoma, large cell carcinoma, and squamous cell carcinoma. The dataset contains clear and high-quality chest CT-scan images. At the same time, more training samples are expected to train a large classification neural network better to optimize its diagnosis accuracy, providing a suitable scenario for dataset augmentation performance examination using the refined DDPM framework. The dataset is proportionated as follows: 70% for training, 20% for testing, and 10% for validation. The effect of the dataset augmentation is evaluated based on the classification accuracy obtained on the validation dataset.

3.2 Implementation configurations

Table 1 lists details of the training parameter configuration. The model is employed on the PyTorch platform and trained using one NVIDIA 4090 GPU. The basic modules of the diffusion model are broadly cited from the diffusers library developed by Hugging Face. The construction of the EfficientNet is broadly cited from the open-source code released in [7]. The image resolution is 256 × 256 pixels, and the batch size is 8 to reduce computational complexity. The learning rate is set to 1×10^{-4} with learning rate warm-up enabled. The diffusion timestep is set to 1000. The training epoch is set to 200. The weighting parameter is set to 0.1. A typical CNN framework proposed in [10] is pre-trained to generate pseudo-labels and re-trained using the combination of the synthetic and original training samples to measure the dataset augmentation performance. During validation, the original-to-synthetic image ratio of the augmented training dataset, both for DDPM-based and GAN-based approaches, is set to 0.4. The DAGAN proposed in [11] is selected as the GAN-based approach's model to compare with the DDPM-based counterpart.

Table 1. Training parameter configuration.

Parameter	Configuration
Resolution	256 × 256
Learning Rate	1×10^{-4}
Timestep	1000
Training Epoch	200
Weighing Parameter	0.1
Original-to-Synthetic Image Ratio	0.4

4 Results

Table 2 presents the accuracy of lung cancer prediction implemented by the CNN framework trained on datasets with different pre-processing.

Table 2. Dataset augmentation performance comparison: random cropping and rotation V.S. DAGAN V.S. DDPM.

Dataset Augmentation Technique	Accuracy (%)
No Pre-Processing (Baseline)	65.6
Random Cropping and Rotation	69.4 (+3.8%)
DAGAN-Based	75.1 (+9.5%)
DDPM-Based (this paper proposed method)	79.2 (+13.6%)
Dataset Augmentation Technique	Accuracy (%)

Fig. 3 shows examples of synthesized images and their corresponding reference images. The refined DDPM framework successfully synthesizes images that bring diversity to the original dataset while keeping the labels unchanged. The synthetic results verify the framework's ability to comprehend the semantic features of the original dataset and bring features based on the known distribution, guaranteeing a qualified balance of the refined DDPM framework between fidelity and diversity.

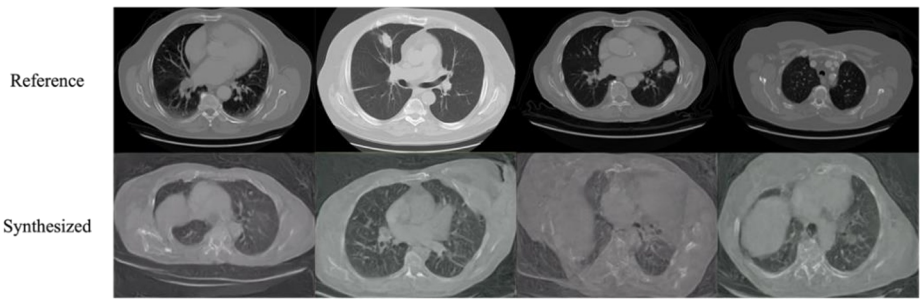

Fig. 3. Synthesized image examples (Photo/Picture credit: Original).

Table 2 shows that accuracy increases by 3.8% compared with the baseline after random cropping and rotation. These operations bring spatial diversity into the training dataset, making the model robust against size, location, and orientation variations. The GAN-based approach presents 9.5% higher accuracy compared with the baseline and 5.7% compared with the random cropping and rotation approach. The GAN-based approach can capture the underlying feature distribution of the original training dataset and synthesize corresponding samples to enrich the dataset. The enriched dataset exposes the model to a wider range of scenarios, enabling it to learn more robust and generalizable patterns. Compared with the previous trio, the DDPM-based approach exhibits superior performance, 13.6% higher than the baseline, 9.8% higher than the random cropping and rotation approach, and 4.1% higher than the GAN-based approach. This can be explained by the DDPM's stronger performance in producing high-quality and diverse synthetic data. DDPM achieves better sampling quality and a more appropriate balance between the fidelity and diversity of the synthetic samples than GAN, which also provides the theoretical foundation for the feasibility of the refined DDPM framework. The refined DDPM framework introduces EfficientNet to determine the feature distribution of the original training dataset to enhance the synthetic data's fidelity further. Introducing the random Gaussian noise at the output of the UNet's bottleneck helps improve the diversity of the generated data. The assistance of the pre-trained classifier avoids the contamination of the dataset, which becomes an additional rationale for the better performance of the refined DDPM framework compared with GAN.

5 Conclusions

This paper presents a refined DDPM-based framework that better balances fidelity and diversity in dataset augmentation tasks. By integrating EfficientNet with diffusion models, the semantic features of the reference image from the original dataset are extracted to guide the denoising process. Random Gaussian noise is added to the UNet bottleneck output to enhance the diversity of synthetic samples. The refined approach outperforms traditional techniques, random cropping and rotation, and GAN-based methods in the context of lung cancer detection. This work verifies the potential of the refined DDPM framework in generating high-quality, diverse datasets, which is crucial for improving model robustness in scenarios with limited datasets. In future work, more attention could be paid to reducing the computational complexity during inference, such as simplifying the diffusion model or adopting a more efficient feature-extracting network.

References

1. H.-T. Gayap, M.-A. Akhloufi, Deep machine learning for medical diagnosis, application to lung cancer detection: A review. BioMedInformatics. **4**(1), 236–284 (2024)
2. D. Walawalkar, Z. Shen, Z. Liu, M. Savvides, Attentive CutMix: An enhanced data augmentation approach for deep learning based image classification. arXiv preprint arXiv:2003.13048 (2020)
3. C. Bowles et al., GAN augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018).
4. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239 (2020)
5. B. Trabucco, K. Doherty, M. Gurinas, R. Salakhutdinov, Effective data augmentation with diffusion models. arXiv preprint arXiv:2302.07944 (2023)

6. Y. Fu, C. Chen, Y. Qiao, Y. Yu, DreamDA: Generative data augmentation with diffusion models. arXiv preprint arXiv:2403.12803 (2024)
7. M. Tan, Q.V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2020)
8. J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, Y. Xu, MedSegDiff-V2: Diffusion-based medical image segmentation with transformer. AAAI, **38**(6), 6030–6038 (2024)
9. M. Hany, Chest CTScan images. Kaggle (2019). Available: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images
10. M.A. Hossain, M.S.A. Sajib, Classification of image using convolutional neural network (CNN). Glob. J. Comput. Sci. Technol. **19**(2), 13-14 (2019)
11. A. Antoniou, A. Storkey, H. Edwards, Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340 (2018)