

Tomato Disease Image Synthesis Model Base on Diffusion and Attention Mechanism

Haoyang Wei

School of Intelligent Systems Science and Engineering, JINAN UNIVERSITY, Zhuhai, Guangdong, China.

Abstract. It is difficult to collect disease pictures of the plants due to the seasonal and regional features of diseases which can lead to poor effect of recognition models. Moreover, conventional enhancement techniques like image rotation and scaling often fail to produce images that accurately reflect real disease characteristics. Taking tomato late blight disease as an example, its typical pathological features demonstrate a complex textural structure and spatial distribution pattern. This study proposes a diffusion model with an attention mechanism. By building two models, basic and enhanced, this study explores the effect of class embedding, attention mechanism and pathology feature simulation on the quality of generation. The results show that the pictures generated by the enhanced model are significantly better than the basic model in terms of rationalization of pathological features. The FID and KID values of the healthy and late blight categories are 74, 68, 0.21, 0.18 respectively, showing that the synthetic data were able to simulate the visual characteristics of real diseases to some degree. This study provides a low-cost data augmentation scheme for disease recognition model training under small sample conditions, which helps to improve the robustness of classification models.

1 Introduction

Tomatoes are one of the most productive vegetables in the world. However, the late blight disease seriously influences tomato yields, especially in high humidity and low temperatures. So it is crucial to discover whether the trees are suffering from disease. The traditional method to recognize the disease is based on experts experience which is a low efficient way.

Although Convolutional Neural Network (CNN) has good performance in general scenes, in rural areas the shortage of specialists which makes it difficult to collect enough samples for model training. What's more, the diverse light conditions in the field and limited equipment make the model significantly less robust in the agriculture scene.

Existing data enhancement methods (e.g., rotation, cropping) expand datasets through geometric transformations, but they are often constrained by the limited ability to express the features [1]. On the other hand, deep learning generative models exhibit stronger feature

hywei@stu2022.jnu.edu.cn

representation with Generative Adversarial Networks (GANs) generating realistic samples through adversarial training [2][3]. However, in the agriculture field, pictures generated by GANs based models suffer from excessive randomness in the distribution of disease and spots and blurred edges. To solve the problems, Yan et al. proposed an improved CycleGAN method by introducing a dual-path network structure and morphological constraints, which significantly enhanced the pathological feature of rice disease lesions [4]. Transfer learning was conducted on the generated samples using a YOLO model, demonstrating noticeable improvements in metrics such as precision and recall on a test set composed of real samples. This further validated the effectiveness and authenticity of the images generated by the proposed generative adversarial network. What's more, Dai et al. employed an improved FastGAN2 and EfficientNet-B5 model, using citrus disease images as the training set. The results showed that the FID and KID of the generated images improved by 31.8% and 59.86% respectively, compared to the original model [5]. However, due to the principles of adversarial learning, the models may suffer from pattern collapse and gradient vanishing. Variational Auto -Encoder (VAE) achieves data generation through latent space modeling, but the resolution may be low [6].

In recent years, Diffusion Models have achieved high-quality image generation by predicting the noise [7], and they achieved SOTA performance on general datasets such as ImageNet [8], which provides a new way to improve agriculture image generation ability. But there are some limits in the agriculture field: firstly, most studies focus on general objects and lack of agriculture data in the training process. Secondly, the traditional diffusion model does not fully use the domain knowledge in the generation process which leads to too much randomness and insufficient resemblance to the real pathology feature.

To solve the problems above, this study proposes an improved conditional diffusion model for tomato disease. By integrating the feature extraction through the attention mechanism and the generation ability of the diffusion model, the model can achieve innovation breakthroughs in the following two aspects: First, the model has a framework that can encode the disease category labels into 128-dimensional vectors and fusion with features and time step embedding to achieve precise direction in picture generated. What's more, the attention mechanism is introduced into U-net which strengthens the ability of the feature extraction of the model. This can effectively improve the generation of the texture.

2 Method

2.1 Diffusion Model

The conditional diffusion model in this paper contains two core steps, forward diffusion and backward denoising. Given the original data distribution $q(x_0)$, the forward process adds Gaussian noise step by step through $T=600$ steps, and its Markov chain process can be represented as follows:

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) \quad (1)$$

The single-step transfer is defined as a Gaussian distribution:

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

The noise scheduling strategy uses a modified cosine scheduling function squaredcos_cap_v2 with the expression:

$$\beta_t = clip\left(1 - \frac{\frac{t}{T} + 0.008}{0.008 + \cos\left(\frac{\pi t}{2T}\right)}, 0, 0.999\right) \quad (3)$$

where $s=0.008$ is the smoothing factor and $T=600$ is the total time steps. The strategy keeps β_t small in the early stages ($t < 0.1T$) to preserve semantic information, and rapidly increases the noise intensity in later stages ($t > 0.5T$).

The backward denoising process is modeled by a neural network that learns to gradually denoise the data by estimating the conditional distribution. $p_\theta(x_{t-1}|x_t)$.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, y), \sigma_t^2 I) \quad (4)$$

2.2 Multiple Attention Mechanism

The multinomial self-attention mechanism on an $h \times w$ characteristic graph $X \in \mathbb{R}, h \times w \times d$ is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (5)$$

Where Q, K, V are query, key, value matrix respectively. Each attention head is computed as:

$$\text{head}_i = \text{Softmax}\left(\frac{(XW_i^Q)(XW_i^K)^\top}{8}\right)XW_i^V \quad (6)$$

Class embedding:

$$E(y) = \begin{cases} W_0 & y = 0(\text{health}) \\ W_1 & y = 1(\text{blight}) \end{cases} \quad W_0, W_1 \in \mathbb{R}^{512} \quad (7)$$

In the sampling phase, the conversion from noise prediction to data reconstruction is realized by re parameterization Technology:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1 - \alpha_t}\epsilon_\theta}{\sqrt{\alpha_t}}\right) + \sqrt{1 - \alpha_{t-1}}\epsilon_\theta \quad (8)$$

2.3 Model Structure

In this research, two tomato disease synthetic models were implemented for comparative experiments, which are called the basic model and the enhanced model. Both two models use the torch. diffusers library's UNet2DModel as core framework. Both models use a symmetrical encoder and decoder structure, but the number of feature channels at each model are different. In the encoding process, the front layers use conventional convolutional models (DownBlock2D) and the multi head self-attention mechanism is introduced in the latter layers to strengthen the ability to capture the global contextual information. The decoder is designed in an opposite direction, with the front layers using an attention enhancement model (AttnUpBlock2D) and the latter layers using a conventional

deconvolution module (UpBlock2D), achieving progressive reconstruction of local details. Moreover, two models are also different in the loss function and pre-treating methods.

Model 1 is the basic model, using a standard U-net structure, containing 4 level downsampling and upsampling. The input consists of three parts: time step embedding, disease category embedding and noisy. The time step embedding maps diffusion steps to 512-dimension vectors through position coding, while the disease category embedding maps category labels to semantic vectors of the same dimension as the time step through a learnable Embedding layer. The input size of the image is fixed to $3 \times 256 \times 256$, which is progressively up-sampled and reconstructed after 5 layers of convolutional down-sampling to 16×16 resolution. In order to enhance the ability of the model to focus on the edge of the leaves a spatial attention module is introduced in the third level of downsampling (AttnDownBlock2D) and the first level of upsampling (AttnUpBlock2D), with channel dimensions of 256 and 512 respectively. See Figure 1.

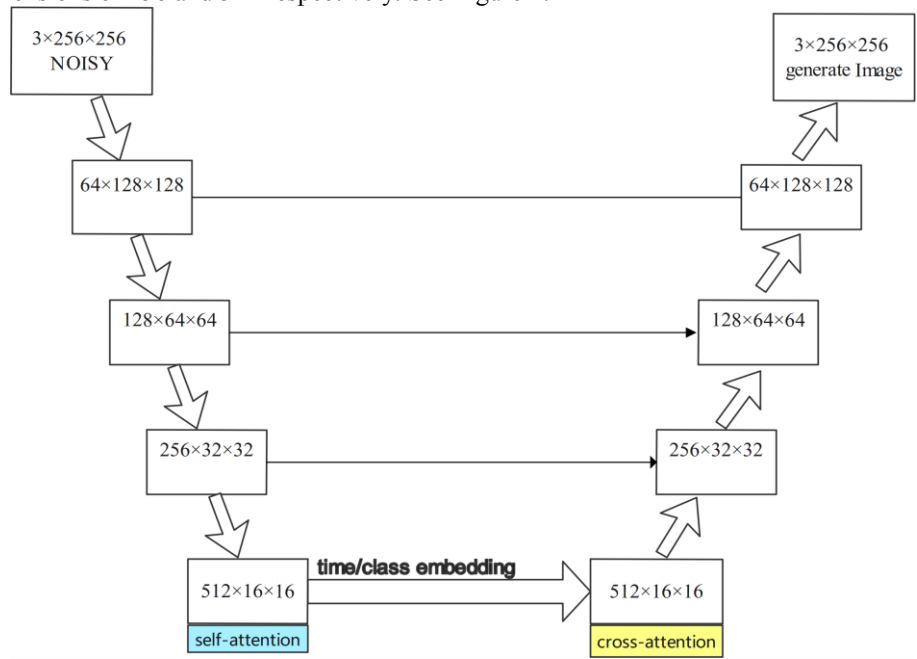


Figure 1 Model1 structure (Picture credit: Original)

An improved DDPM algorithm is used in the training phase, and the forward process adds Gaussian noise progressively over 600 timesteps, with the noise variance changing from $1e-4$ to 0.02 by a linear adjustment. During the reverse denoising process, the U-net network predicts the noise residuals of the current step, and the loss function is a weighted mean square error (MSE).

The enhanced model has some adjustments based on one basic model. First, the network depth expands to 5 levels of down sampling with 64, 128, 256, 512, 512 respectively. The spatial attention mechanism is introduced at the inverse two levels of all down sampling and up sampling layers. See Figure 2.

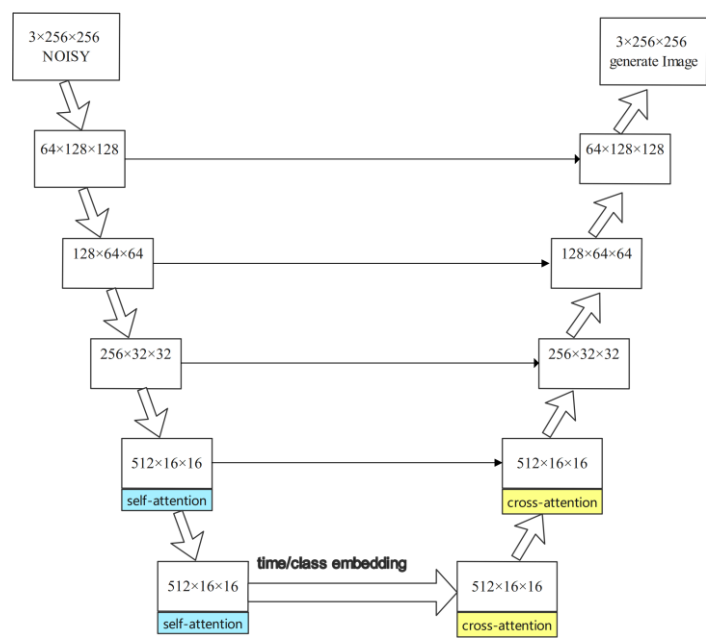


Figure 2 Model2 Structure (Picture credit: Original)

In addition, model 2 is designed with a multi-objective loss function, where the mean square error is combined with feature contrast loss to optimize the colour and texture details of the generated image. L1 regularity improves the edge of the leaves. The experimental settings $\lambda_1 = 0.65$, $\lambda_2 = 0.35$ and the multi-objective loss function is defined as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,\epsilon}[\lambda_1 \|\epsilon - \epsilon_{\theta}(x_t, t, y)\|_2^2 + \lambda_2 \|\epsilon - \epsilon_{\theta}(x_t, t, y)\|_1] \tag{9}$$

3 Experiment

3.1 Dataset

The experiment was using the Plant Village public dataset [9], which contains 3325 images of tomato images at 256×256 resolution, including 1590 healthy images and 1735 late blight images. All images were used for training. To measure the quality of the dataset, pictures from the same category were divided into two equal parts and calculate the Fréchet Inception Distance (FID) between two sets which were randomly divided and the tests were repeated for 10 times. After that, the cross-category FID was calculated as well. The FID values for the training set are shown in Table 1.

Table 1 Training Set FID

Category	Same Category	Cross	Category
	FID Mean (Std)	FID	
Healthy	12.99 (0.12)	92.84	
Blight	23.77 (0.20)		

In the data preprocessing stage, the base model is geometrically enhanced with random horizontal flipping and vertical flipping, and the pixel values are mapped to the $[-1,1]$ interval by channel normalization. In addition, since the category data in the training set is balanced, no additional processing is required. In the enhanced model, a pathology feature enhancement module for late blight was added to simulate lesion features by randomly adding brown spots and incorporating random colour dithering (brightness ± 0.1 , contrast ± 0.1 , saturation ± 0.1) to enhance data diversity.

3.2 Training Environment

Table 2 Training Environment

GPU	AutoDL: vGPU-32GB (similar to RTX 3090)
Crucial module	torch.diffusers, torch.vision
Version	Pytorch 2.3.0, cuda12.1, Python 3.12
Optimizer	AdamW, lr=1e-4, a=0.01
Batch size	Model1:30, Model2: 20

The environments of the experiment are shown in Table 2. Both models were implemented under Pytorch 2.3.0, cuda 12.1, and Python 3.12. The models were trained using the AdamW optimizer (learning rate 1e-4, weight decay 0.01) with a linear warm-up strategy to avoid early overfitting. To accelerate the training process, mixture-precision FP16 was used. The experiment was completed in vGPU-32GB which is provided by AutoDL platform. The batch size was set to 30 for the basic model and 20 for the enhanced model and the graphics memory footprints are around 25GB and 28GB, respectively. In the inference stage, both the models utilize the DDPM sampling algorithm to accelerate the generating process which contains 50 timesteps. In order to verify the category control capability, the models generate the samples of healthy and late blight disease under the same noise, and the generation results are visualized by 5×2 grid.

3.3 Evaluation method

FID is based on multivariate Gaussian distribution assumption. It uses the Inception-V3 model with the removal of the classification layer to extract the high-level features from the data. After that, generated pictures and real pictures are mapped as 2048-dimensional feature vectors respectively. Finally, the mean vectors of the two sets of features are computed and the distributional differences are quantified by the Fréchet distance equation:

$$FID = |\mu_r - \mu_g|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (10)$$

This distance combines mean differences and covariance differences, with smaller values indicating that the generated distribution is closer to the training set distribution.

KID also uses Inception-v3 to extract features and overcomes the Gaussian distribution assumption by using the kernel approach. Distributional differences are measured using Maximum Mean Discrepancy (MMD) [10]. Specifically, a polynomial kernel function is used to map the features to the Regenerative Kernel Hilbert Space (RKHS) and the kernel matrix difference from the true sample is computed for the generation:

$$KID = \mathbb{E}[k(x_r, x_r')] + \mathbb{E}[k(x_g, x_g')] - 2\mathbb{E}[k(x_r, x_g)] \quad (11)$$

x_r, x_g represent real and generated samples respectively. KID does not require parametric assumptions and reduces the bias through unbiased assumption, which is suitable for small datasets or non Gaussian distribution scenarios. On the other hand, t-SNE focuses on achieving visualization of high-dimensional features by maintaining local structure and displaying clusters in low-dimensional space. The closer the distance, the closer the distribution of the two categories.

3.4 Experiment Result

A portion of the model generates results that fail to show the internal textural structure of the leaf as well as the characteristics of a diseased leaves (Figure 3).



Figure3. Model1 healthy (first row), Late Blight (second row) (Picture credit: Original)

The FID and KID values of the results generated by model 2 are in table 3.

Table 3 Result of model 1

Generated number	500	1000	1500
Health FID	148	137	122
Late Blight FID	116	109	105
Cross Category	98	92	85

A portion of the enhanced model generates results. See figure 4 and 5:



Figure 4 Model2 Healthy (Picture credit: Original)



Figure 5 Model2 Late Blight disease (Picture credit: Original)

The FID and KID values of the results generated by model 2 are in table 4.

Table 4 Result of model 2

Generated number	500	1000	1500
Health FID	94	86	74
Late Blight FID	101	89	76
Cross Category	115	102	96
Healthy KID			0.21
Late Blight KID			0.18

The t-SNE plot of the generated image and the real image of the health class and late blight class of model 2 are shown in figures 6 and 7.

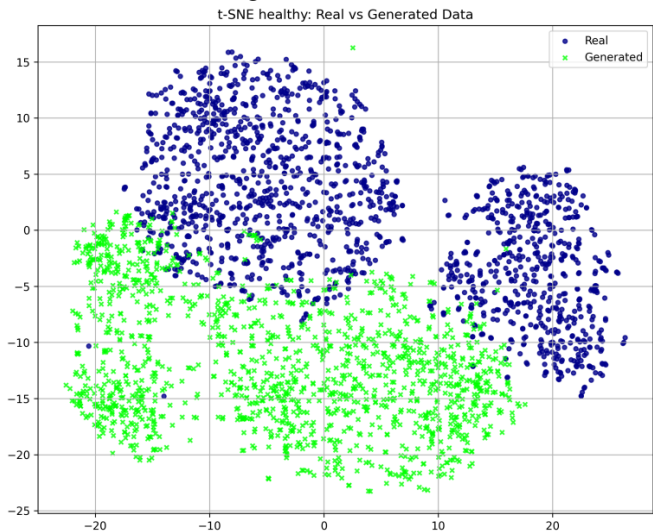


Figure 6 Scatter plot of t-SNE after reducing dimensions. Blue points represent the real dataset of the healthy category while the green points represent the generated images (Picture credit: Original)

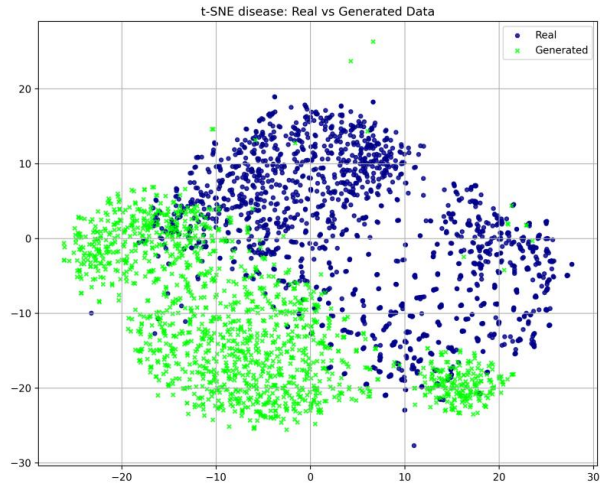


Figure7 Scatter plot of t-SNE after reducing dimensions. Blue points represent the real dataset of Late Blight category while the green points represent the generated images (Picture credit: Original)

4 Conclusion

This study presents a framework for tomato disease image generation using a conditional diffusion model. This method enhances data expending in tomato leaf disease. The results demonstrated that the diffusion model which combines categories embedding, multi-attention mechanism and pathology feature enhancement can solve the limitations of the traditional data enhancement methods in the generation of complex textures to a certain extent. It also provides a new way for the generation of agricultural pictures in the small sample condition.

Despite some progress, there are still limitations in this research: First, the efficiency of model training needs to be optimized, and the enhanced model takes about 30% longer to train a single batch due to the increase of network layer and the application of attention mechanism, with occupying about 28GB GPU memory, limiting its widely use. In addition, the images generated at this stage exhibit limited resolution, constrained to 256×256 pixels. Finally, the existing evaluation approach lacks sophistication, and future work should integrate plant expert evaluation for more reliable validation.

Future work contains the following directions: Initially, input dimensions can be reduced the through autoencoder, and the knowledge distillation strategy to compress the number of U-net parameters which can lightweight the model. Secondly, integrating the textual descriptions of the disease and multi-spectral image data, and constructing a cross-modal condition generation framework to improve the temporal consistency of the disease spot evolution. Finally, optimizing evaluation systems that contain agriculture and plant experts to assess the quality of generated pictures can build up. Furthermore, a comparative evaluation should be conducted between a classification model trained on a hybrid dataset (original and synthetic images) and one trained exclusively on the original dataset to assess performance differences.

In conclusion, this study provides technical support for the identification and control the plant disease by applying practice generative AI in the agriculture field.

References

1. C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
2. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.
3. T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401–4410.
4. Yan Congkuan, Zhu Dequan, Meng Fankai, Yang Yuqing, Tang Qixing, Zhang Aifang, Liao Juan, "Rice leaf disease image enhancement method based on improved CycleGAN," *Smart Agriculture*, vol. 6, no. 6, pp. 96–108, 2024.
5. Q. Dai, Y. Guo, Z. Li, S. Song, S. Lyu, D. Sun, et al., "Citrus disease image generation and classification based on improved FastGAN and EfficientNet-B5," *Agronomy*, vol. 13, no. 4, p. 988, 2023.
6. D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, Dec. 2013.
7. J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.
8. P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8780–8794, 2021.
9. D. P. Hughes and M. Salathé, "PlantVillage Dataset," *Mendeley Data*, V1, 2015. [Online]. Available: <https://data.mendeley.com/datasets/tywbtsjrjv/1>
10. M. Bińkowski, D. J. Sutherland, M. Arbel and A. Gretton, "Demystifying MMD GANs," *arXiv preprint arXiv:1801.01401*, Jan. 2018.