agriculture

MDPI

Article

Identifying Tomato Growth Stages in Protected Agriculture with StyleGAN3–Synthetic Images and Vision Transformer

Yao Huo †, Yongbo Liu †, Peng He *, Liang Hu, Wenbo Gao and Le Gu

Institute of Agricultural Information and Rural Economy, Sichuan Academy of Agricultural Sciences, Chengdu 610011, China

* Correspondence: hepeng78@scsaas.cn

† These authors contributed equally to this work.

Abstract: In protected agriculture, accurately identifying the key growth stages of tomatoes plays a significant role in achieving efficient management and high-precision production. However, traditional approaches often face challenges like non-standardized data collection, unbalanced datasets, low recognition efficiency, and limited accuracy. This paper proposes an innovative solution combining generative adversarial networks (GANs) and deep learning techniques to address these challenges. Specifically, the StyleGAN3 model is employed to generate high-quality images of tomato growth stages, effectively augmenting the original dataset with a broader range of images. This augmented dataset is then processed using a Vision Transformer (ViT) model for intelligent recognition of tomato growth stages within a protected agricultural environment. The proposed method was tested on 2723 images, demonstrating that the generated images are nearly indistinguishable from real images. The combined training approach incorporating both generated and original images produced superior recognition results compared to training with only the original images. The validation set achieved an accuracy of 99.6%, while the test set achieved 98.39%, marking improvements of 22.85%, 3.57%, and 3.21% over AlexNet, DenseNet50, and VGG16, respectively. The average detection speed was 9.5 ms. This method provides a highly effective means of identifying tomato growth stages in protected environments and offers valuable insights for improving the efficiency and quality of protected crop production.

Keywords: StyleGAN3; ViT; deep learning; tomato

Academic Editor: Signe Marie Jensen

Received: 16 December 2024
Revised: 6 January 2025
Accepted: 6 January 2025
Published: 7 January 2025

Citation: Huo, Y.; Liu, Y.; He, P.; Hu, L.; Gao, W.; Gu, L. Identifying Tomato Growth Stages in Protected Agriculture with StyleGAN3–Synthetic Images and Vision Transformer. *Agriculture* **2025**, *15*, 120. https://doi.org/10.3390/agriculture15020120

1. Introduction

As one of the most important economic crops globally, the tomato's growth stages require careful monitoring and identification to safeguard both yield and quality [1]. The accurate recognition of critical developmental periods—such as the sapling, flowering, fruiting, and maturity stages—offers essential insights for precise management and decision-making regarding crop growth. This process is vital for improving agricultural efficiency and maximizing economic returns [2].

At present, the identification and monitoring of tomato growth stages primarily depend on manual observations based on the experience of practitioners. This method is not only inefficient and costly, but it is also highly vulnerable to the subjectivity of the observer and the influence of activities within the facility, which can compromise both the quality and yield of tomatoes. While technologies like inspection robots and smart imaging systems have emerged, their high costs and the difficulty of maintaining such equipment have hindered their widespread application, particularly in facilities with

limited financial resources. Consequently, there is an increasing need for research into digital technology-based methods for identifying tomato growth stages, which would offer real-time, convenient, and efficient solutions for monitoring.

Deep learning, an essential branch of machine learning, was introduced by Hinton et al. [3,4] in 2006, sparking a significant wave of research within the neural network domain. Due to its exceptional capabilities in feature learning and classification, deep learning has become a foundational tool in several fields, including object detection, natural language processing [5], text information matching [6], medical image segmentation and classification [7,8], and video semantic segmentation [9,10]. In recent years, deep learning has also seen growing applications in the agricultural sector, particularly in the identification and classification of crop diseases and pests [11,12]. Convolutional neural networks (CNNs) [13], which are among the most widely adopted and effective models in this context, were successfully applied to crop recognition and classification tasks, with well-known networks like AlexNet [14] and GoogLeNet [15] leading the way.

The introduction of the Vision Transformer (ViT) model [16] in 2020 marked a breakthrough in the field of crop monitoring and classification. Compared to traditional CNN-based approaches, ViT offers enhanced accuracy in tasks like crop disease recognition and classification. Bai Yupeng et al. [17] developed a wheat disease recognition algorithm based on ViT, which achieved an impressive average recognition accuracy of over 95% for three different types of wheat diseases. This performance surpassed that of the AlexNet and VGG16 [18] models by 6.68% and 4.94%, respectively. In another study, Wang Yang et al. [19] addressed the poor robustness of deep convolutional neural networks (DCNNs) to noise by incorporating additional modules, such as augmented block serialization and masked multi-head attention, into the standard ViT model for tomato disease classification. Their modified ViT model achieved a classification accuracy of 99.63% on a tomato dataset, improving over 6% compared to classic models like ResNet50 [20]. However, despite these advancements, the use of ViT models in crop growth stage recognition remains limited. Most researchers still rely on more established deep learning models for such tasks. Rasti et al. [21] explored three machine learning approaches for identifying the growth stages of 11 wheat and 10 barley species: a five-layer Conver Net self-training model, a transfer learning-based VGG19 model, and support vector machines. Their findings indicated that the transfer learning-based model was the most effective, with an accuracy rate exceeding 99% and a substantial reduction in training time. Meanwhile, Tan et al. [22] conducted a comparative experiment using traditional machine learning methods and deep learning approaches to recognize the growth stages of rice. Their results showed that deep learning methods outperformed traditional methods, with the EfficientNetB4 model achieving the highest recognition accuracy, surpassing 99% in both accuracy and average precision.

Due to practical constraints, the amount of manually collected image data is limited. The tomato dataset suffers from issues such as class imbalance. However, the Vision Transformer (ViT) model requires a sufficient amount of data to achieve optimal recognition performance [23]. Traditional data augmentation techniques, such as rotation, translation, and cropping, are inefficient and only generate images with low signal-to-noise ratios and limited diversity. To address overfitting and other challenges, generative adversarial networks (GANs) [24] can produce clearer and more realistic samples, but traditional GANs suffer from issues such as training instability and model collapse [25]. StyleGAN3 [26], an improved version of GAN introduced by NVIDIA in 2021 based on StyleGAN2 [27], resolves problems like feature entanglement and is capable of generating higher-resolution, higher-quality images. However, StyleGAN3 also requires substantial computational resources to achieve satisfactory training results, typically necessitating more than 5000 iterations. This

process imposes high demands on hardware performance and requires significant time and computational power to allow the model to adequately learn and capture the complex distributional characteristics of the data thereby generating high-quality images. Considering these factors, this work initially employs transfer learning by using a pre-trained model as the foundation for our custom StyleGAN model, which significantly reduces the training time. By further training the model on the existing dataset to generate more realistic synthetic images, we alleviate the data scarcity issue commonly faced by the ViT (Vision Transformer) model. This approach not only enhances the quality of the synthetic images but also expands the diversity of the training dataset, providing the model with a larger variety of learning samples. Ultimately, this contributes to improving the accuracy of tomato growth stage recognition. Through this optimization strategy, we aim to enhance the model's generalization ability and robustness, particularly in practical scenarios where data are limited.

The paper is organized into the following four sections: Section 1 provides an introduction, outlining the current state of tomato growth stage recognition in protected agriculture, along with a review of recent advancements in deep learning models for crop recognition and classification. Section 2 describes the dataset used in this work, as well as the StyleGAN3 and ViT models. Section 3 presents the experimental results and a comparative analysis with existing methods. Sections 4 and 5 are dedicated to the discussion and conclusion, respectively.

2. Materials and Methods

This section provides an overview of the dataset used in this work, along with a description of the models employed. The main contributions of this paper are divided into two parts: first, high-quality synthetic images are generated using the StyleGAN3 generative adversarial network; second, the ViT model is applied to tomato growth stage recognition and classification through transfer learning.

2.1. Dataset

Tomato images were collected from the Modern Agricultural Science and Technology Innovation Demonstration Park at the Sichuan Academy of Agricultural Sciences. This facility is located in the Xindu District of Chengdu, Sichuan Province, China, in the central depression of the Chengdu Plain, with an altitude of 510 m and a subtropical humid climate. The growth of tomatoes involves four stages, sapling, flowering, fruiting, and maturity, which occur throughout the year.

In this work, a total of 1200 images were collected across 4 growth stages, sapling, flowering, fruiting, and maturity, with 300 images for each stage. The collection time of the tomato images is in 2024. Based on the originally collected data, 2723 synthetic images were generated using GAN. All images were resized to a final resolution of 224 × 224. Example images for the four categories, along with specific details, are presented in Figure 1 and Table 1.

Table 1. Class-wise image distribution of growing period tomato dataset.

Category	Number of Images (Original)	Number of Images (Synthetic)
Sapling	300	450
Flower	300	396
Fructification	300	327
Mature	300	350

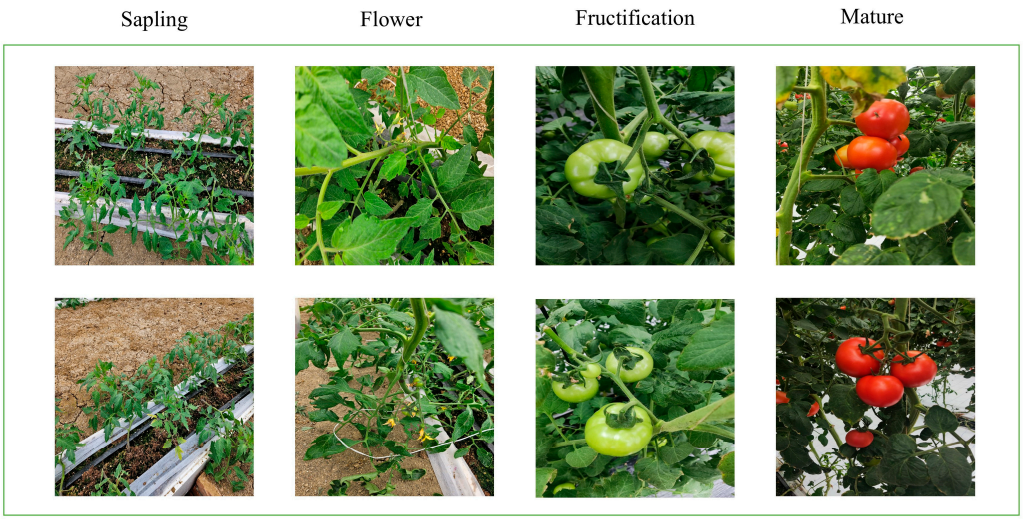

Figure 1. Four categories of tomato growth period.

2.2. StyleGAN3

Generative adversarial networks (GANs) operate by introducing two adversarial neural networks: the generator and the discriminator. These networks are optimized through adversarial training, where the generator creates realistic data samples, and the discriminator distinguishes between real and synthetic samples. Due to their robust generative capabilities and wide-ranging application potential, GANs have become a pivotal research area in fields such as computer vision, image processing, and machine learning. Over time, several GAN variants have been proposed, including Conditional GAN (CGAN) [28], Deep Convolutional GAN (DCGAN) [29], and BigGAN [30], each with its advantages and drawbacks. For instance, although CGAN and DCGAN are faster to train, they generate lower-resolution images with visible blurring at the edges, rendering them unsuitable for high-resolution tasks like generating tomato images. While BigGAN can produce high-resolution images, its considerable hardware requirements and long training times make it impractical for the present application. In contrast, StyleGAN3 offers a significant advantage: it can generate high-resolution images at 1024 $\times$ 1024 pixels and can be trained efficiently on a single GPU, thanks to its optimized CUDA cores, which enhance training speed and memory utilization.

As previously noted, StyleGAN3 effectively resolves the feature entanglement problem in StyleGAN2, where the fine details of generated images fail to move coherently with the object's shape, leading to inconsistencies in rotation and translation. The root cause of this issue lies in the generator architecture of existing models, which employs convolutions, nonlinear activations, and upsampling layers, failing to achieve adequate equivariance. In contrast, StyleGAN3 takes a signal processing approach, tracing the problem back to aliasing within the generator network, and proposes a solution. By treating all signals as continuous, StyleGAN3 introduces minor yet effective architectural modifications that prevent unintended information leakage during the hierarchical synthesis process, enhancing the model's overall performance.

StyleGAN3 achieves rotational invariance through two key modifications. First, all 3 $\times$ 3 convolution layers across the network are replaced with 1 $\times$ 1 convolutions, ensuring that only downsampling and upsampling operations propagate information between pixels thus preventing unnecessary information leakage into the hierarchical synthesis process. Second, the sinc downsampling filters are replaced with radial symmetric jinc filters, except for two critical layers. These changes significantly improve rotational invariance without compromising FID. Additionally, during the early stages of training, a Gaussian filter is

applied to all images seen by the discriminator to prevent early stage training collapse. The generator may occasionally introduce high-frequency signals with small delays, degrading the discriminator's performance. A schematic of the generator is shown in Figure 2.

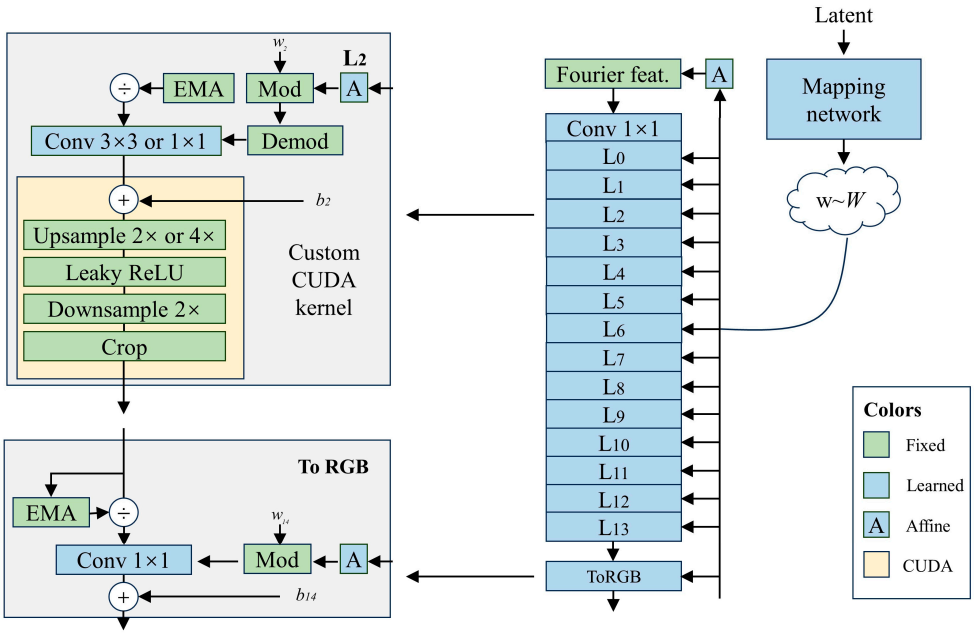

Figure 2. StyleGAN3 generator structure.

In this work, we employ StyleGAN3 to generate images representing different growth stages of tomatoes to address the limitations of the dataset. To generate these images, we first train a model on a tomato dataset. The training parameters are as follows: the configuration is set to stylegan-t (translation equiv); the total batch size is 32; gamma is set to 32 (R1 regularization weight); and the model is trained for 1000 kimg (total number of training iterations). Furthermore, to improve the monitoring and tracking of the training process, the snapshot frequency (snap) is set to 10.

2.3. Vision Transformer (ViT)

In this work, we employ the Vision Transformer (ViT) model to recognize the growth stages of tomatoes through transfer learning. The ViT architecture introduces the concept of image patches. Initially, the image is divided into non-overlapping patches of equal size, which are then encoded into sequence data using a positional encoding function. These sequence data are subsequently fed into a Transformer encoder. The output from the Transformer is processed through a fully connected layer followed by a softmax layer to produce the final classification result. The complete framework of the Vision Transformer is presented in Figure 3.

For tomato classification, we utilize the ViT-Base model, which segments the input image into 16 × 16 pixel patches. These patches are then processed through 12 encoder layers, each consisting of encoder blocks, stacked a total of 12 times. After passing through the embedding layer, each resulting vector has a dimension of 768. The fully connected layer includes 3072 units, and the multi-head self-attention mechanism incorporates 12 attention heads. The model contains 86 M parameters, and its architecture is detailed in Table 2.

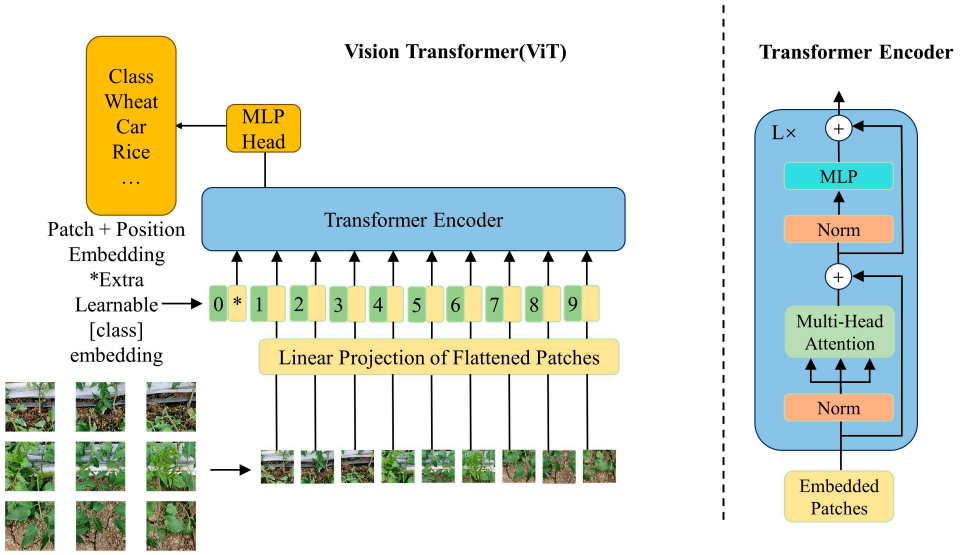

Figure 3. ViT model structure.

Table 2. Different versions of ViT model.

Model	Patch Size	Layers	Hidden Size D	MLP Size	Heads	Params
ViT-Base	16 × 16	12	768	3072	12	86 M
ViT-Large	16 × 16	24	1024	4096	16	307 M
ViT-Huge	14 × 14	32	1280	5120	16	632 M

2.4. Experimental Platform

The experiment utilized the PyCharm 2024.1.1 development environment alongside the PyTorch deep learning framework, with Windows 11 as the operating system. The hardware configuration included an Intel i5-12500 processor and an NVIDIA A5000 GPU with 24 GB of onboard memory. All comparative experiments were conducted using identical system specifications.

2.5. Experimental Setup and Evaluation Metrics

This work involved three experimental stages. In the first stage, we employed the StyleGAN3 model to train on each category of the tomato dataset for 1000 epochs, ultimately generating models corresponding to four tomato growth stages. Following training, we generated 1523 synthetic images representing the tomato growth stages and conducted a quality assessment of these images. The second stage involved training and validating the ViT-Base model on both the original tomato dataset and an augmented dataset that included the synthetic images, allowing us to compare the recognition performance of the two datasets. In the third stage, we selected a set of classic classification models and compared their recognition accuracy for tomato growth stages against the ViT-Base model.

To evaluate the performance of the models, we utilized a range of metrics, including accuracy, precision, recall, *F*1 score, and confusion matrix. The formulas for these evaluation metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

In the context of classification tasks, true positive (*TP*), false positive (*FP*), and false negative (*FN*) are essential metrics for evaluating the performance of predictive models. More precisely, *TP* represents the number of instances that are genuinely positive and are correctly classified as such by the model; *FP* denotes the number of instances that are actually negative but are mistakenly classified as positive; and *FN* indicates the number of actual positive instances that are erroneously classified as negative. These metrics are of paramount importance in performance analysis as they provide valuable insight into the model's classification accuracy, its ability to discriminate between classes, and the nature of its errors.

In this work, we use the Peak Signal-to-Noise Ratio (*PSNR*) to quantitatively evaluate the quality of the generated images. *PSNR* is a well established metric in the fields of image and video processing, primarily used to measure the level of noise in an image and to assess the quality of the image in comparison to a reference. This metric is also widely employed to evaluate the performance of image processing algorithms by providing an objective measure of the image's fidelity. *PSNR* is calculated based on the Mean Squared Error (*MSE*), which represents the average squared difference between the original and generated images. The following formulas define the calculations of *PSNR* and *MSE*:

$$PSNR = 20 \cdot \log_{10}\left(\frac{MAX_1}{\sqrt{MSE}}\right) \tag{5}$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2. \tag{6}$$

In Equation (5), *MSE* represents the Mean Squared Error, a measure of the average squared differences between the original and generated images. *MAX* refers to the maximum value of the pixel intensities in the image, which provides an upper bound for the image's dynamic range. In Equation (6), *I* and *K* are used to denote the original and generated images, respectively, while *m* and *n* correspond to the number of rows and columns in the images. These parameters are essential for computing the image quality metrics, which evaluate the fidelity of the generated image relative to the original.

3. Implementation and Result

This section begins by outlining the hardware, software, and evaluation metrics used in the experiments. It then evaluates the effects of using original and synthetic tomato images, analyzing the performance of tomato growth stage recognition with and without synthetic images. The section concludes by comparing the proposed model against several classical models.

3.1. Performance of StyleGAN3

The images generated using StyleGAN3, alongside the original tomato images, are presented in Figure 4. As illustrated, the generated images of the four tomato categories closely resemble the original images, demonstrating the model's ability to produce highly realistic results. However, it is important to highlight that StyleGAN3 requires significant computational resources and energy consumption, and the complexity of the images from the fruiting and maturity stages is considerably greater than that of the earlier stages. Achieving optimal training results for these stages necessitates over 20,000 epochs. Furthermore, the constraints of having only one high-performance GPU in our experimental setup will make the training time longer; the average training time per iteration is 4 min35 s, so to reduce the total training time, we sought to minimize the impact of background noise on

the model and enhance the quality of the generated images. To achieve this, we employed the Rembg tool, based on U2-Net architecture, to remove the background from the images thereby isolating the foreground object for improved output.

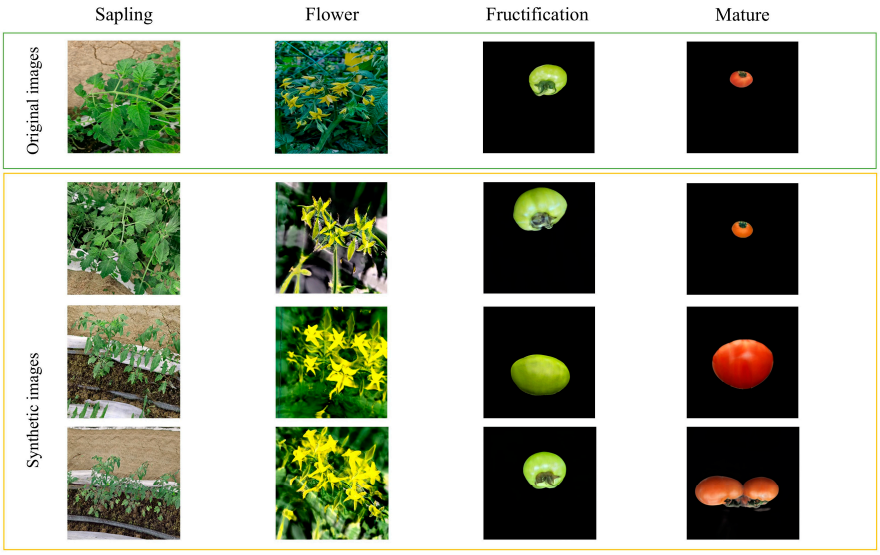

Figure 4. The original tomato images and the StyleGAN3-generated images: The first to fourth columns represent the four categories of tomato growth stages. The first row shows the original images of the four tomato growth stages, while the second to fourth rows display the generated images corresponding to each category.

Table 3 provides a detailed statistical analysis of the PSNR between the original and generated tomato images. The findings clearly demonstrate that the quality of the images generated using StyleGAN3 closely approximates the quality of the real, original images. StyleGAN3 produces images of varying quality depending on the category. For the sapling and flower categories, the PSNR values are approximately 28 dB, suggesting that the generated images are somewhat distorted compared to the originals. These distortions are likely subtle and may not be easily detectable by the human eye. In contrast, the fructification and mature categories show a notable enhancement in image quality, with PSNR values increasing to around 39 dB. This significant rise in PSNR indicates that the generated images preserve more fine details, with minimal distortion, and closely mirror the original images in terms of visual fidelity.

Table 3. PSNR values (in dB) between original images and GAN-generated images for four categories.

Category	PSNR Original Images	PSNR Original and StyleGAN3 Images
Sapling	27.903	28.018
Flower	27.925	27.932
Fructification	38.78	39.422
Mature	37.389	38.784

From Table 4, it is evident that StyleGAN3 exhibits varying levels of image quality in terms of SSIM (Structural Similarity Index) across different categories. SSIM is a metric that evaluates the structural, luminance, and contrast similarities between two images. A higher SSIM value, closer to 1, indicates greater structural similarity between the original and generated images. In the sapling and flower categories, the SSIM values are notably low, at 0.066 and 0.454, respectively. However, these low values do not necessarily indicate poor image quality, as SSIM primarily evaluates structural aspects of the images. The sapling and flower categories contain more complex elements, such as leaves and stems, which

may contribute to lower SSIM values. Therefore, the SSIM values of the sapling and flower categories should be interpreted in conjunction with the PSNR values and human visual judgment. On the other hand, in the fructification and mature categories, there is a marked increase in SSIM values. In the fructification category, the SSIM rises from 0.833 to 0.85, while in the mature category, it slightly decreases from 0.85 to 0.84. These values indicate that the generated images in both categories retain a high degree of structural similarity to the original images. Despite the slight decrease in SSIM in the mature category, the image quality remains high, with minimal distortion.

Table 4. SSIM values between original images and GAN-generated images for four categories.

Category	SSIM Original Images	SSIM Original and StyleGAN3 Images
Sapling	0.066	0.068
Flower	0.454	0.384
Fructification	0.833	0.85
Mature	0.85	0.84

3.2. Performance of ViT Model

As previously mentioned, to evaluate the classification performance of the tomato classification model, we selected accuracy (Acc), precision, recall, and F1 score as the evaluation metrics for the model's quality. Table 5 presents the classification performance of the ViT-Base model on both the original images and the combined dataset of original and generated images. The evaluation metrics for the ViT-Base model with the inclusion of generated images are 98.39%, 98.47%, 98.39%, and 98.39% for accuracy, precision, recall, and F1 score, respectively. These results represent improvements of 3.81%, 3.48%, 3.81%, and 3.98% over the performance achieved using only the original images. This demonstrates that incorporating generated images into the ViT model significantly enhances the model's classification ability and generalization, especially under data imbalance, while also mitigating the risk of overfitting.

Table 5. Impact of synthetic images on Vit-Base.

Model	Accuracy	Precision	Recall	F1 Score
ViT-Base	94.58%	94.99%	94.58%	94.41%
ViT-Base + Synthetic images	98.39%	98.47%	98.39%	98.39%

Figure 5 presents the confusion matrix results for the classification task, which was performed using both original and StyleGAN-generated images. The matrix indicates that classification performance—particularly in distinguishing true positives from false positives between the flowering and sapling stages—was notably reduced when only the original images were used. In contrast, a significant improvement in the model's ability to accurately classify these stages was observed with the inclusion of StyleGAN-generated images, resulting in a marked enhancement in overall classification performance. Specifically, the confusion matrix for the combined dataset, comprising both original and generated images, demonstrates a substantial increase in overall accuracy, as well as a more balanced distribution between true positives and false positives. These results emphasize the considerable advantages of augmenting the dataset with generated images, highlighting their critical role in enhancing the robustness and effectiveness of the classification model.

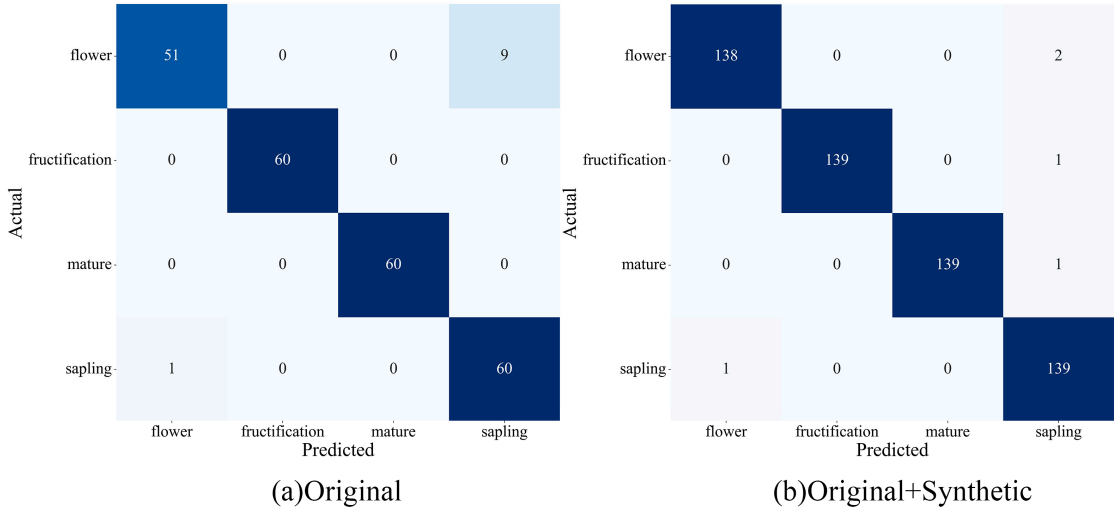

Figure 5. Confusion matrix of tomato growth classification using proposed method.

Figure 6 illustrates the AUC curves during training for ViT-Base, comparing the original dataset (Figure 6a) with the dataset augmented by generated images (Figure 6b). AUC, which assesses the model's performance at various classification thresholds, provides a more comprehensive evaluation by minimizing the effect of a single threshold. The results show that adding generated images significantly improves the performance of ViT-Base in the four-class classification task, with fewer fluctuations compared to using only the original dataset, and the VIT model can better perform its performance. This suggests that images generated by StyleGAN3 can effectively enhance the model's classification accuracy.

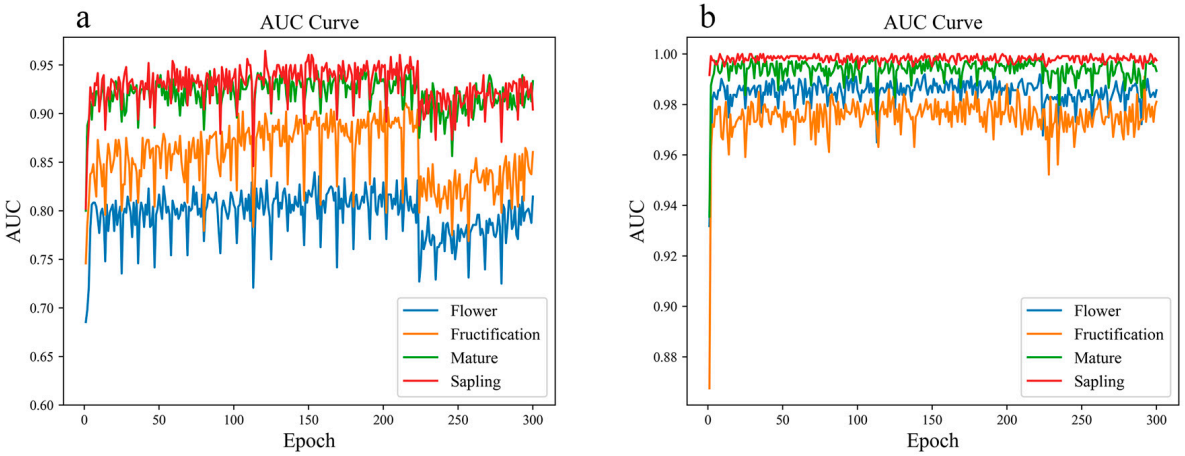

Figure 6. AUC comparison of ViT-Base performance with original dataset (**a**) and dataset augmented with generated image (**b**) during training.

3.3. Performance Comparison

In this subsection, we first perform a detailed curve analysis of four models, examining their training and validation performance regarding accuracy and loss. Specifically, we investigate the ViT-Base model, augmented with generated images, and compare its four-class classification performance against that of three widely used models—AlexNet, DenseNet50 [31], and VGG16—enhanced with generated images. As presented in Figure 7, the curve analysis reveals that the ViT-Base model achieves the most consistent and favorable performance. It demonstrates smooth convergence, the fastest reduction in loss, and an impressive validation accuracy of 99.6%. In contrast, the other models show varying levels of instability, with AlexNet exhibiting notable overfitting issues. These findings

suggest that the ViT-Base model not only outperforms the others in terms of accuracy but also maintains greater stability throughout the training process.

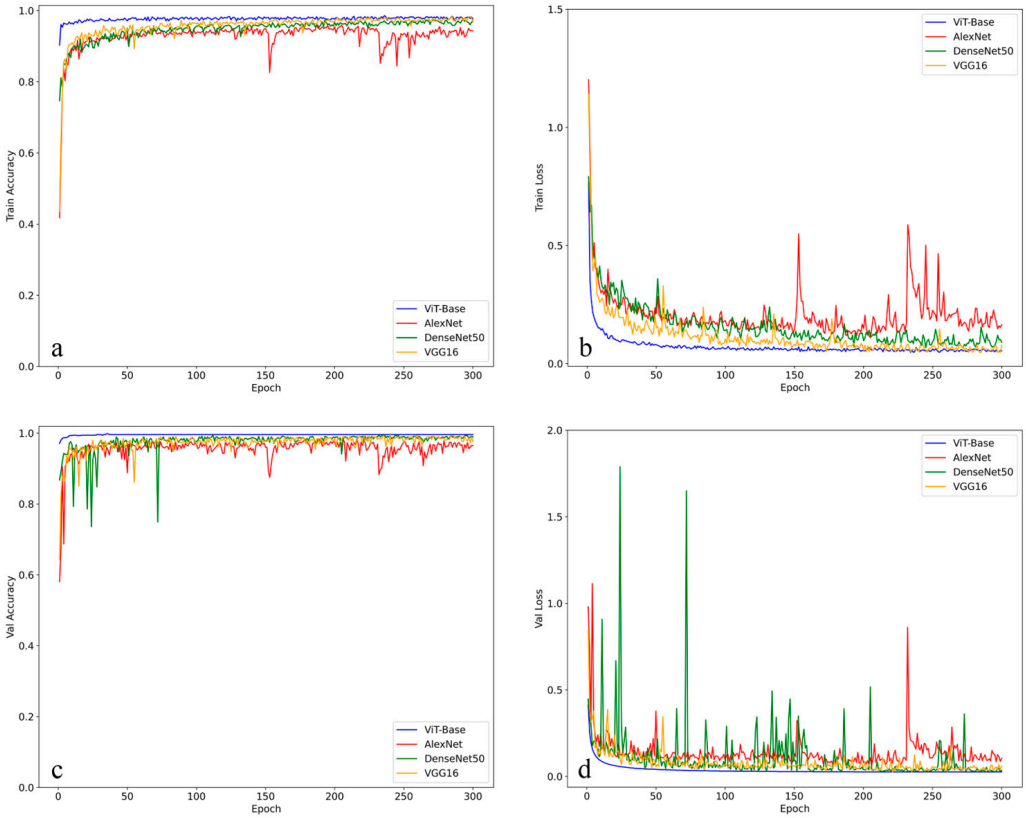

Figure 7. Performance of training parameters for four models. (**a**) Training accuracy, (**b**) training loss, (**c**) validation accuracy, (**d**) validation loss.

Table 6 demonstrates that the ViT-Base model achieves the highest performance, with accuracy improvements of 22.85%, 3.57%, and 3.21% over AlexNet, DenseNet50, and VGG16, respectively. This superior performance can be largely attributed to the ViT-Base model requirement for a large dataset to fully realize its classification potential. Notably, the extensive set of images generated by the StyleGAN3 model plays a crucial role in compensating for some of the inherent limitations of the ViT-Base model. By augmenting the training data with these synthetic images, we were able to address the model data requirements thus enabling it to achieve better results. Furthermore, a comparison with Table 4 reveals that, when augmented with generated images, the DenseNet50 and VGG16 models perform comparably to, or even outperform, the ViT-Base model trained solely on the original images. These results underscore the significant impact of data augmentation in the strategic use of StyleGAN3-generated images and in enhancing the performance of deep learning models.

Table 6. Performance comparison of different models.

Model	Accuracy	Precision	Recall	F1 Score
ViT-Base	98.39%	98.47%	98.39%	98.39%
AlexNet	75.54%	84.88%	75.54%	68.10%
DenseNet50	94.82%	95.61%	94.82%	94.77%
VGG16	95.18%	95.59%	95.18%	95.13%

4. Discussion

This work presents a method for recognizing the growth stages of tomatoes, leveraging the StyleGAN3 model to generate synthetic images in conjunction with the ViT-Base model. The experimental results in Section 3 demonstrate that the generated images closely resemble real images, providing a solid foundation for their use in the recognition task. By incorporating these generated images into the ViT-Base model, superior performance is achieved in identifying the various growth stages of tomatoes. Through comparative experiments, it can be seen that the performance of the ViT-Base model combined with generated images for the identification of the tomato growth stage is significantly better than that of AlexNet, DenseNet50, VGG16, and other classical models combined with generated images. ViT-Base was shown to outperform traditional CNN models in some computer vision tasks such as image classification, object detection, and segmentation, especially when large-scale datasets are available. ViT's self-attention mechanism is able to capture global dependencies, which is difficult to achieve with convolutional operations in some tasks. On the other hand, DenseNet50 and VGG16 benefit more from data augmentation primarily due to their deeper architectures and greater model complexities. Their ability to learn more detailed and generalized features, coupled with a larger number of parameters and layers, enables them to fully leverage the diversity introduced by augmentation. In contrast, AlexNet, being a shallower network with fewer parameters, is less capable of fully exploiting the increased data diversity, resulting in comparatively smaller performance gains.

When selecting the ViT model for tomato classification, there was initial uncertainty regarding whether to opt for the more recent and improved Swin model, which has demonstrated superior performance in certain tasks. However, following in-depth discussions among the authors, we determined that ViT was more suitable for our work's primary focus on classification. Unlike tasks such as segmentation and detection, which are prediction-intensive and in which the Swin model excels, our work did not involve such tasks. Although ViT requires longer training times compared to Swin, it maintains high accuracy in classification tasks, particularly when enhanced by data. Moreover, the ViT model's relatively simple architecture, coupled with the availability of extensive pretrained models, made it a more practical choice for training purposes thereby enabling more efficient model development for classification.

Although the StyleGAN3-generated images in this work are generally close to real images, their quality in noisy scenes is still insufficient, requiring preprocessing as is typical in many studies. For instance, the tomato fruiting and maturity stage datasets in this research were processed to remove background noise. Additionally, the long training times of StyleGAN3 pose challenges, especially with limited computational resources, leading to significant time and energy consumption. Transfer learning or cloud-based resources can be employed to mitigate these issues. Therefore, a critical area for future research will be developing methods to directly generate high-quality images quickly, without the need for preprocessing.

For an extended period, our research team has been leveraging computer vision technology to address practical problems in agricultural production. In the course of our work, we have also sought to extend our research to more complex production environments, such as open fields and orchards. As part of this effort, we utilized images collected from cameras in apple and citrus orchards to create a foundational dataset. By applying the StyleGAN3 model developed in this study for data augmentation, we successfully built growth stage recognition models tailored to specific apple and citrus varieties. While the controlled conditions in facility agriculture create a relatively stable environment, the complexity of orchards presents additional challenges. Although our approach showed some promising results in orchard environments and demonstrated its potential applicability in

other scenarios, the actual test outcomes were not as effective as those seen in the controlled tomato production environment. This gap indicates that further refinements are needed, and improving this aspect will be a primary focus of our future research efforts.

5. Conclusions

This work begins by leveraging the StyleGAN3 generative adversarial network to generate images, addressing issues such as limited training data and data imbalances. Following this, the ViT-Base model is employed to recognize the growth stages of greenhouse tomatoes, producing promising results in terms of classification performance. The primary conclusions are as follows:

(1) The quality of images generated by StyleGAN3 is nearly identical to that of real images, with an average generation time of 153 milliseconds per image. This method proves to be an effective data augmentation solution, especially for cases with limited training data, such as small sample datasets. However, when computational resources are limited, transfer learning or background denoising techniques are necessary to reduce training time.

(2) By generating images through a generative adversarial network (GAN) and then applying the ViT-Base model for tomato growth stage recognition, this method achieves superior performance compared to direct recognition using original images. The combination of ViT with generated images reached an accuracy of 98.39% on the test set and an average detection speed of 9.5 milliseconds. When compared to AlexNet, DenseNet50, and VGG16, this method showed improvements in accuracy by 22.85%, 3.57%, and 3.21%, respectively, demonstrating its enhanced effectiveness in classification tasks.

(3) In areas with limited access to intelligent devices, the use of images generated by a generative adversarial network (GAN) significantly reduces the labor demands and inconsistencies of manual image collection. Furthermore, applying the ViT-Base model for tomato growth stage recognition can provide crucial data for informed decision-making and the precise management of tomato growth conditions. This approach offers considerable economic value by improving the efficiency of tomato production. Additionally, it can be extended to other similar crop categories, such as apples and citrus, in addition to tomatoes.

Author Contributions: Conceptualization, Y.L., P.H. and Y.H.; methodology, P.H.; software, Y.H., Y.L., L.H. and L.G.; validation, Y.H. and W.G.; data curation, Y.H.; writing—original draft, Y.H. and Y.L.; writing—review and editing, Y.L. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Modern Agricultural Industrial Technology System Sichuan Grains Innovation Team special fund (SCCXTD-2024-20); the Special Fund for Independent Innovation of Sichuan Province—Research on the Application of Spatiotemporal Big Data Analysis in Agricultural Production Services (2022ZZCX034); the Key R&D Project of Sichuan Province Science and Technology Program; the 'Innovative Knowledge Service Platform for Crop and Livestock Breeding Research' (2021YFYZ0028-01); and the 'Tianfu Granary' Digital Agriculture Sichuan-Chongqing Joint Innovation Key Laboratory Project—Intelligent Planting Strategy Model and Application for Facility Tomatoes (TFLCSZ-JB3).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to our image datasets being self-built.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alajrami, M.A.; Abu-Naser, S.S. Type of Tomato Classification Using Deep Learning. *Int. J. Acad. Pedagog. Res.* **2020**, *3*, 21–25.
2. Wei, Y.; Qin, R.; Ding, D.; Li, Y.; Xie, Y.; Qu, D.; Zhao, T.; Yang, S. The impact of the digital economy on high-quality agricultural development--Based on the regulatory effects of financial development. *J. Huazhong Agric. Univ.* **2023**, *43*, 9–21.
3. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
4. Hinton, G.E.; Osindero, S.; Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neuralcomputation* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
5. Tomas, M.; Kai, C.; Greg, C.; Jeffrey, D. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
6. Jacob, D.; Ming, W.; Kenton, L.; Kristina, T. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
7. Mahesh, G.; Sweta, J. Stacked Convolutional Neural Network for Diagnosis of COVID-19 Disease from X-ray Images. *arXiv* **2020**, arXiv:2006.13871.
8. Mahesh, G.; Sweta, J.; Raghav, A. DeepRNNetSeg: Deep Residual Neural Network for Nuclei Segmentation on Breast Cancer Histopathological Images. In Proceedings of the International Conference on Computer Vision and Image Processing (CVIP), Singapore, 27–29 September 2019; pp. 243–253.
9. Xu, Y.S.; Fu, T.J.; Yang, H.K.; Lee, C.Y. Dynamic video segmentation network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6556–6565.
10. Tang, H.; Ding, L.; Wu, S.; Ren, B.; Sebe, N.; Rota, P. Deep Unsupervised Key Frame Extraction for Efficient Video Classification. *ACM Trans. Multimedia Comput. Commun. Appl.* **2023**, *119*, 1–17. [CrossRef]
11. Amreen, A.; Sweta, J.; Mahesh, G.; Swetha, V. Tomato plant disease detection using transfer learning with C-GAN synthetic images. *Comput. Electron. Agric.* **2021**, *187*, 106279.
12. Xu, J.; Wang, J.; Xu, X.; Ju, S. Image recognition for different developmental stages of rice by RAdam deep convolutional neural networks. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 143–150.
13. Zhang, X.; Hou, T.; Hao, Y.; Shangguan, H.; Wang, A.; Peng, S. Surface Defect Detection of Solar Cells Based on Multiscale Region Proposal Fusion Network. *IEEE Access* **2021**, *9*, 62093–62101. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
15. Al-Qizwini, M.; Barjasteh, I.; Al-Qassab, H.; Radha, H. Deep learning algorithm for autonomous driving using GoogLeNet. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium(IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 89–96.
16. Alexey, D.; Lucas, B.; Alexander, K.; Dirk, W.; Zhai, X.; Thomas, U.; Mostafa, D.; Matthias, M.; Georg, H.; Sylvain, G.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
17. Bai, P.; Feng, Y.; Li, G.; Zhao, M.; Zhou, H.; Hou, Z. Algorithm of wheat disease image identification based on Vision Transformer. *J. Chin. Agric. Mech.* **2024**, *45*, 267–274.
18. Karen, S.; Andrew, Z. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
19. Wang, Y.; Li, Y.; Xu, J.; Wang, A.; Ma, C.; Song, S.; Xie, F.; Zhao, C.; Hu, M. Crop Disease Recognition Method Based on Improved Vision Transformer Network. *J. Chin. Comput. Syst.* **2024**, *45*, 887–893.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
21. Rasti, S.; Bleakley, C.J.; Silvestre, G.C.M.; Holden, N.M.; Langton, D.; Gregory, M.P.; O'Hare, G.M. Crop growth stage estimation prior to canopy closure using deep learning algorithms. *Neural Comput. Appl.* **2021**, *33*, 1733–1743. [CrossRef]
22. Tan, S.; Liu, J.; Lu, H.; Lan, M.; Yu, J.; Liao, G.; Wang, Y.; Li, Z.; Qi, L.; Ma, X. Machine Learning Approaches for Rice Seedling Growth Stages Detection. *Front. Plant Sci.* **2022**, *13*, 914771. [CrossRef] [PubMed]
23. Liu, W.; Lu, X. Research Progress of Transformer Based on Computer Vision. *Comput. Eng. Appl.* **2022**, *58*, 1–16.
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farly, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *arXiv* **2014**, arXiv:1406.2661.
25. Han, X.; Li, Y.; Gao, A.; Ma, J.; Gong, Q.; Song, Y. Data Augmentation Method for Sweet Cherries Based on Improved Generative Adversarial Network. *Trans. Chin. Soc. Agric. Mach.* **2024**, *55*, 1–17.
26. Karras, T.; Aittala, A.; Laine, S.; Harkonen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-Free Generative Adversarial Networks. *arXiv* **2021**, arXiv:2106.12423.
27. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv* **2018**, arXiv:1812.04948.
28. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
29. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.

30. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv* **2018**, arXiv:1809.11096.
31. Huang, G.; Liu, Z.; Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993.