# A Comparative Analysis of GANs and Diffusion Models for Synthesizing Electron Microscopy Images

Ni Li Nora Tseng Songxi Yang Maria Elisa Montes

## Abstract

*Electron Microscopy (EM) is an essential tool for materials characterization, but the analysis of EM images often relies on machine learning models, which are limited by the availability of labeled training data. Generative models such as Generative Adversarial Networks (GANs) and diffusion models have demonstrated remarkable success in synthesizing realistic images for augmentation and training. However, their relative performance for EM image synthesis remains unexplored. In this study, we compare StyleGAN, a state-of-the-art GAN, and Latent Diffusion Models (LDMs), a recent advancement in diffusion models, to determine their ability to generate realistic EM images. Our results indicate that LDM-generated images exhibit better fidelity, achieving a lower Fréchet Inception Distance (FID) score of 36.1 compared to 58.96 for StyleGAN. This findings highlight the potential of LDMs for EM data augmentation and provide valuable insights into the application of generative models in electron microscopy analysis.*

## 1. Introduction

Electron Microscopy (EM) is a critical technique for characterizing materials at the micrometer scale. Machine learning models, such as object detection and segmentation, have been employed to analyze EM images, achieving automated defect detection with human-like accuracy [17]. However, limited training data often necessitates data augmentation to generate synthetic samples for better generalization [3, 14].

Generative Adversarial Networks (GANs) and diffusion models are two of the most popular generative models, both achieving remarkable success in generating realistic images on standard datasets like ImageNet [4]. GANs, such as StyleGAN, are known for their ability to generate high-resolution images with detailed features, while diffusion models, such as Latent Diffusion Models (LDMs), have shown superiority in modeling complex data distributions and stability during training. Despite their success with regular images, it is unclear which approach performs better for the unique characteristics of EM images—grayscale, low signal-to-noise ratio, and subtle features.

To address this question, we selected StyleGAN and LDM to compare their performance in generating realistic EM images. This project aims to evaluate the effectiveness of these two models for EM image synthesis, providing valuable insights for researchers leveraging generative models in the field of electron microscopy.

## 2. Related work

GANs and diffusion models are state-of-the-art image synthesis techniques that have been used for data augmentation and denoising of EM images [9, 11] in the material sciences and biomedical fields. In previous work, Diffusion models have demonstrated superior image quality compared to GANs on standard datasets such as ImageNet [4]. In the material sciences field, other researchers have attempted using GANs to produce synthetic microscopy images [2, 18] and automated labeling of EM images [19]. However, a comparison of traditional GANs and diffusion models in producing EM-specific images has not been conducted, from the best of our knowledge. Given the unique characteristics of EM images—grayscale, low signal-to-noise ratio, and subtle features—it is unclear whether diffusion models will outperform GANs for EM image synthesis. This project aims to answer this question by comparing the performance of these deep learning models on EM images specifically of irradiated FeCrAl alloys.

## 3. Technical approach

To identify whether StyleGAN or Latent Diffusion Model (LDM) had a better performance on EM image synthesis, we trained the best performing implementation of each model using 62,400 images cropped and augmented from the Damaged Alloys augmented dataset. After training both StyleGAN and LDM models to convergence on the full dataset, we then generated 50,000 images using each model and compared their performance quantitatively calculating Fréchet Inception Distance (FID). We also compared the performance of both models qualitatively by iden-

tifying key features of EM metal alloy images on a subset of the generated images. We used one GPU NVIDIA A100 32 GB for training and image generation. The scripts that we used in this project are in the in the GitHub repository [1]

### 3.1. Dataset

We used the publicly-available Damaged Alloys Dataset [2] [10]. This dataset consists of EM images of irradiated iron-chromium-aluminum alloy (FeCrAl) for the detection of loop defects. We used a total of 494 images. In this dataset, 442 images were size 1024 x 1024 pixels, and 52 images were of size 2048 x 2048 pixels. In order to increase the number of images for training the diffusion and GAN models and to reduce the dimensionality of the data we cropped the images into sections of 256 x 256 pixels. To further increase the number of images we rotated and flipped the crops. We rotated the crops by 90, 180, and 270 degrees, and flipped them vertically and horizontally. After performing the augmentation, the dataset we used for training consisted of 62,400 total images with size 256 x 256.

### 3.2. StyleGAN Models

StyleGAN is a family of GAN models known for generating high-quality images and controlling generated content, mainly including StyleGAN [7], StyleGAN2 [8], and StyleGAN3 [6], as shown in Figure 1.

StyleGAN's main contributions include: 1) instead of feeding the latent vector z directly into the generator, StyleGAN maps z to an intermediate latent space w via a mapping network. This enables disentangled control over generated features; 2) StyleGAN introduces styles into the generator through adaptive instance normalization (AdaIN) in the synthesis network [7]. StyleGAN2 makes some improvements over StyleGAN, including removing artifacts like droplet patterns, replacing AdaIN with weight demodulation, adding path length regularization to improve smoothness in latent space interpolation [8]. StyleGAN3 further improves the model architecture by ensuring translation and rotation equivariance of generated images, such as maintaining consistency when shifted or rotated [6]. Overall, StyleGAN family ensures model latent space control and high fidelity image synthesis.

This study leverages the official PyTorch implementations of StyleGAN2 and StyleGAN3 models[3]. To assess the effectiveness of StyleGAN models in generating EM images, we first analyze the impact of generator mapping depth (2, 4, and 8) and the maximum feature size (256, 512, and 1024) of the mapping network across StyleGAN2, StyleGAN3-T, and StyleGAN3-R. Each configuration was trained for 800kimg, where kimg represents 1,000 training images, a standardized unit often used in GAN research to measure training progress. The best-performing model was then trained to convergence and compared with diffusion models to draw conclusions.

### 3.3. Latent diffusion model

The latent diffusion model is known for its computational efficiency yet good performance [15]. As shown in Figure 2, The model contains two parts: Autoencoder (AE) and diffusion model (with U-Net architecture). The encoder of the AE can compress the images into the latent space and thus decrease the dimension of the input data. The diffusion model is trained on the latent space, therefore the computational cost is much lower compared to those diffusion models based on pixel space. The decoder then reconstruct those diffusion model generated latent space vectors back into images. To achieve the best image quality, we first selected a pretrained AE from a list of candidates [13] that were trained on ImageNet or OpenImages. We had each AE, including DALL-E dVAE and various VQGANs, reconstruct 1000 randomlly selected images from our augmented dataset. With the reconstructed and original images, we then calculated the average Fréchet Inception Distance (FID) reported in Table 1. Selecting the best first stage model for our context, we use VQGAN (f8, 8192, OpenImages; f8 = 8x compression, 8192 = number of codebook entries).

The U-Net architecture in the Latent Diffusion Model (LDM) is designed for efficient latent space processing, taking $32 \times 32$ inputs with 4 channels and producing 4-channel outputs. It features a base channel width of 256 and employs a hierarchical structure with channel multipliers $[1, 2, 2, 4, 4]$ to progressively increase feature channels across resolution levels. The network uses 2 residual blocks per resolution stage and incorporates multi-head attention with 8 heads at resolutions $8 \times 8$, $4 \times 4$, $2 \times 2$, and $1 \times 1$ to capture long-range dependencies. This architecture is optimized for latent space diffusion, balancing computational efficiency with the ability to generate high-fidelity images aligned to input conditions.

| Autoencoder | Average FID |
|---|---|
| VAE DALL-E | 116.64 |
| VQGAN (f8, 8192, OpenImages) | **31.15** |
| VQGAN (f16, 1024, ImageNet) | 82.19 |
| VQGAN (f16, 16384, ImageNet) | 60.26 |

Table 1. Autoencoder Reconstruction Average FID Scores.

Using the selected pretrained AE, we trained the LDM model on 90% of the augmented dataset, reserving the remaining 10% for evaluation. We experimented with batch sizes ranging from 8 to 64, as well as U-Net architectures

[1] https://github.com/nytseng/839_microscopy_project
[2] https://www.materialsdatafacility.org/detail/pub_100_li_detection_v1.2
[3] https://github.com/NVlabs/stylegan3

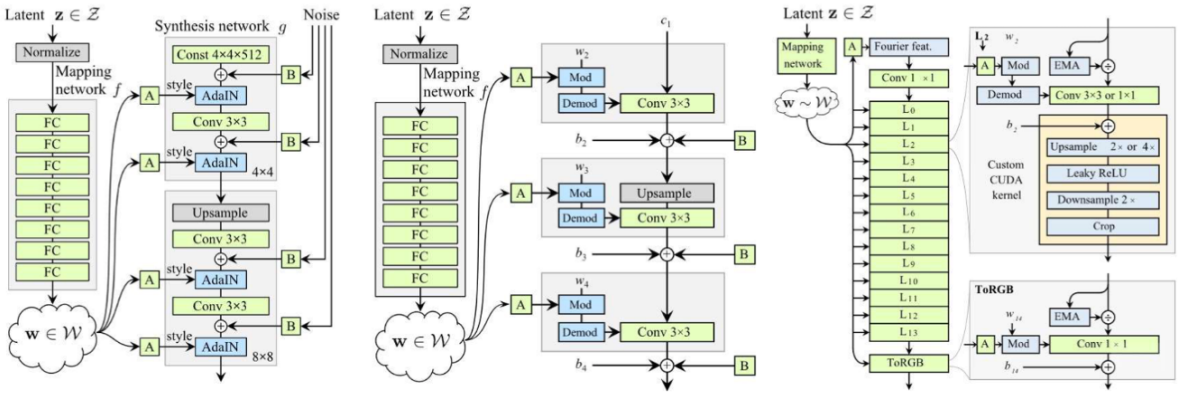


Figure 1. Architectures of StyleGAN Generators (Adapted from [12]): (a) StyleGAN, (b) StyleGAN2, (c)StyleGAN3.

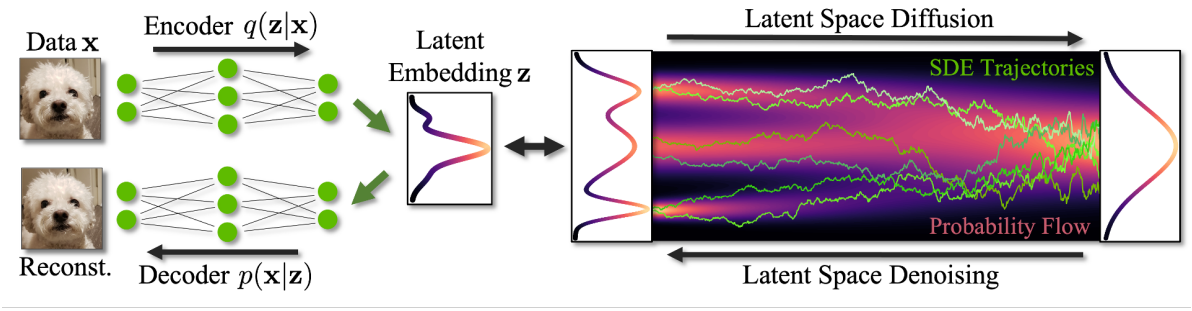


Figure 2. Overview of the Latent Diffusion Model (LDM) pipeline.

with 3 and 4 layers. The 4-layer U-Net demonstrated superior performance compared to the 3-layer configuration. The impact of batch size on the quality of the generated images is further discussed in the Experiments and Results section.

# 4. Experiments and Results

## 4.1. Evaluation metrics

We evaluated the performance of the models on image synthesis using the Fréchet Inception Distance (FID). We chose to use FID because it indicates the similarity between two sets of images, in this case, the generated EM images and the real EM images. We did not use the Inception Score (IS) because this metric may not represent the quality and diversity of EM images. IS uses an Inception network that is pre-trained on ImageNet [1]. This metric assigns higher scores when the generated images contain a single object and are diverse across the classes in ImageNet [1]. Even though the features of the Inception model in FID were optimized on ImageNet, this metric estimates the distance between the distributions of the feature representations of the two sets being compared [5]. In our experiments, we calculate the FID between the EM images in the dataset and the generated images using Pytorch_fid package [16].

## 4.2. Performance of StyleGAN models

### 4.2.1 Impact of Generator Mapping Depth

Figure 3 reveals the impact of generator mapping depth across StyleGAN3-T, StyleGAN3-R, and StyleGAN2, with depths of 2, 4, and 8 at 800kimg.

For StyleGAN3-T, depth 2 achieved the best FID (90.54), while deeper configurations degraded performance (FID 147.42 and 163.28). Similarly, StyleGAN-R showed optimal results with depth 2 (FID 86.92), while depths 4 and 8 led to poorer FID and slower convergence. However, StyleGAN2 performed best with depth 4 (FID 75.44), followed by depth 8 (FID 76.82). Depth 2 of StyleGAN2 achieved comparable results.

Across the three models, both StyleGAN3-T and StyleGAN3-R benefit from a shallow mapping network (depth 2), but face overfitting, convergence problems and consume expensive training resources. Thus, we chose StyleGAN2 model with depth 4 to analyze the impact of maximum feature sizes.

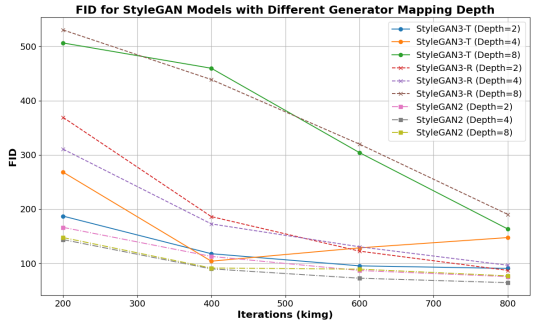


Figure 3. Impact of Generator Mapping Depth on StyleGAN Models.

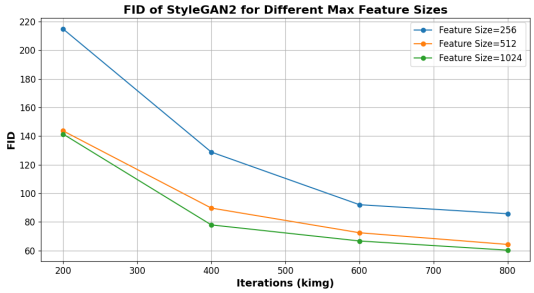


Figure 4. Impact of Maximum Feature Size on StyleGAN Models.

#### 4.2.2 Impact of Maximum Feature Size

Figure 4 summarizes the effect of maximum feature size (256, 512, and 1024) on StyleGAN2 with a generator mapping depth of 4, evaluated at the end of 800kimg.

A feature size of 256 shows the weakest performance (FID 214.87 at 200kimg, improving to 85.73 at 800kimg but still inferior). The default feature size of 512 achieves balanced results (FID 64.28 at 800kimg). Finally, increasing the size to 1024 has the best performance (FID 60.26 at 800kimg), indicating that larger feature sizez enhance the generator's ability to model details and improve realism.

#### 4.2.3 Best StyleGAN Model

Table 2 presents the metrics for the best-performing StyleGAN2 configuration (generator mapping depth 4, maximum feature size 1024).

Specifically, FID improves consistently with more iterations, starting at 141.41 (200kimg) and reaching the best FID of 58.96 (1200kimg). This demonstrates convergence improvements with extended training. In this experiment, sampling 50k images using StyleGAN2 for evaluation takes approximately 8 minutes.

Figure 5 shows sample generated images from best StyleGAN2 model configuration. These results indicate that

| Iterations (kimg) | FID |
|---|---|
| 200 | 141.41 |
| 400 | 77.93 |
| 600 | 66.69 |
| 800 | 60.26 |
| 1000 | 61.60 |
| 1200 | **58.96** |
| Model: StyleGAN2 | |
| Generator Mapping Depth: 4, Max Feature Size: 1024 | |

Table 2. Performance Metrics of the Best 256 x 256 StyleGAN model.

the StyleGAN2 model with a generator mapping depth of 4 and a maximum feature size of 1024 achieves an effective balance between realism and diversity. However, the arrows in the real samples indicate loops that were not well-followed in the generated samples, highlighting limitations in the model's ability to capture intricate structural details.

### 4.3. Performance of Diffusion models

The LDM model converges after 5 epochs. The generated EM images from optimized LDM configuration are compared to real images and shown in Figure 6. The images were generated using 200 sampling time steps. The real images exhibit sharp details and well-defined features, such as distinct loop structures and clear boundaries between features and the background. ALthough the LDM-generated images exhibit some subtle differences, such as slightly distorted loops (as indicated by the red arrow). the general structures and textures resemble those of real EM images, and it's difficult for human eyes to tell the difference between the generated images and the real images. The FID score calculated comparing 50k of generated images with all the real images in the training dataset is 36.1, which is much lower than the images generated from the StyleGAN2 model.

#### 4.3.1 Impact of layers in UNet

Figure 7 illustrates the effect of U-Net architecture depth on the quality of EM images generated by Latent Diffusion Models (LDMs). The left panel displays images generated using a 3-layer U-Net, while the right panel showcases images generated with a 4-layer U-Net. The 3-layer U-Net produces reasonably detailed images but shows noticeable artifacts and lacks fine structural details in certain regions. In contrast, the 4-layer U-Net generates images with improved fidelity and clearer feature boundaries, exhibiting fewer distortions and enhanced representation of complex structures. This comparison highlights the importance of network depth in capturing intricate details, with the 4-layer

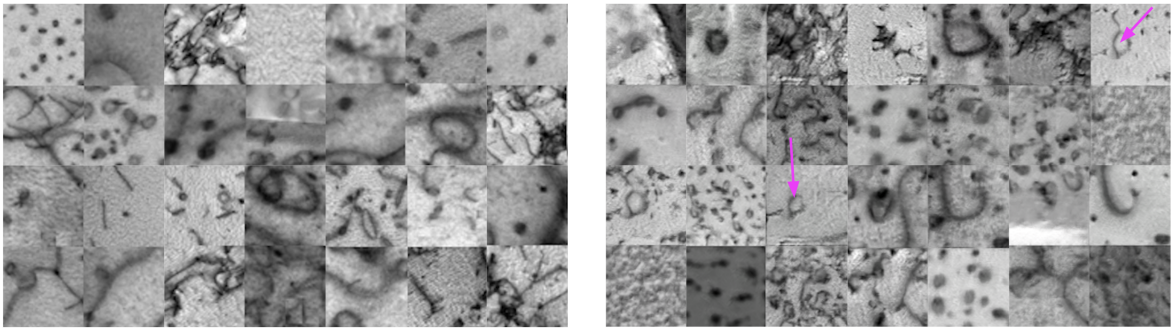

Figure 5. Comparison of StyleGAN2 Generated and Real Samples. (left: real images; right: generated images)

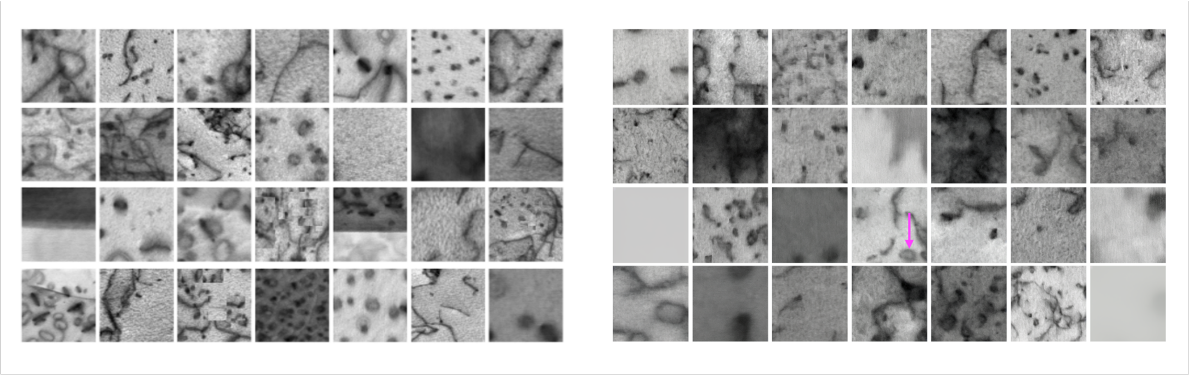

Figure 6. Comparison of LDM Generated and Real Samples. (left: real images; right: generated images)

U-Net outperforming the 3-layer configuration in producing high-quality EM images.

### 4.4. Comparison of StyleGAN and Diffusion Models

As shown in Figure 3 and 6, both StyleGAN2 and LDM can generate pretty realistic images. Table 3 compares the performance of StyleGAN and Latent Diffusion Models (LDM) in terms of FID score, training time, and sampling time for generating 50,000 images. LDM outperforms StyleGAN in fidelity, achieving a lower FID score of 36.1 compared to StyleGAN's 58.96, indicating better quality of generated images. Furthermore, LDM demonstrates significantly faster training convergence, requiring only 6 hours on a single GPU compared to StyleGAN's 18 hours. However, LDM is considerably slower in sampling, taking 30 hours to generate 50,000 images, whereas StyleGAN completes the same task in just 8 minutes. This highlights a trade-off between training efficiency, image quality, and sampling speed when selecting models for specific applications.

## 5. Conclusions

Our results reveal that while both StyleGANs and LDMs are capable of producing convincing Damaged Alloy EM images, LDMs produce higher quality images (FID score of 36.1) on our dataset. However, researchers may opt to use GANs instead as they require much less sampling time due to Diffusion models' iterative nature. Additionally, with the human eye, we note that GAN-produced/LDM-produced images contain inaccurate features such as incomplete, distorted loops and features blended into background images. Nonetheless, vision models have made substantial improvements over time and demonstrate great potential in the EM-imaging field.

### 5.1. Limitations

The main limitation of this project is its scope. Our training data consisted only of metal alloy EM images from one source. In the future, this project could be extended using different datasets, including EM images from other materials and biological matter such as tissues and cells. Models with EM images of different types should incorporate guidance so that users can generate images from the class of their interest. Another limitation is that we only evaluate two models: StyleGAN and LDM. A future extension would be to evaluate more types of models, such as CycleGan and Stable Diffusion. Finally, we only used one evaluation metric to compare model performance. In the future, using more metrics or perhaps developing a metric specific to electron microscopy images could lead to a more infor-

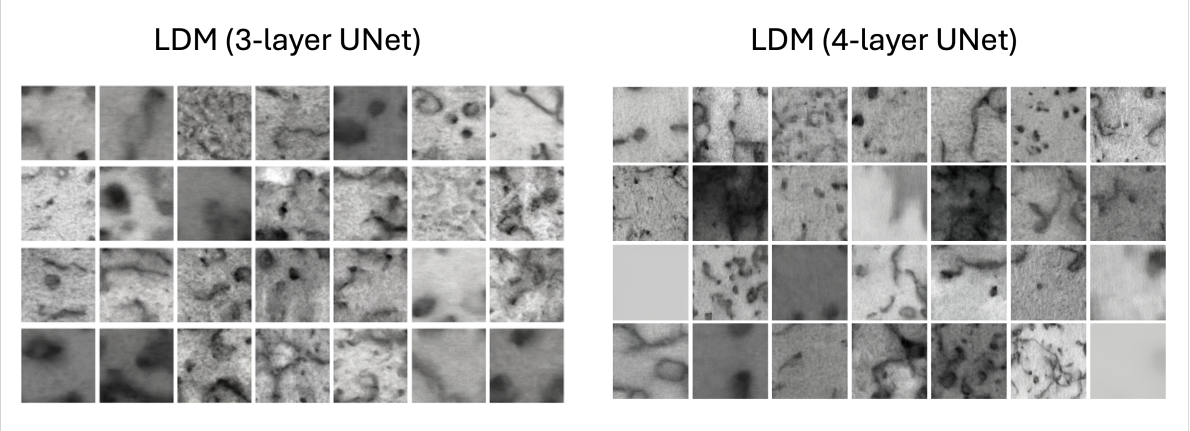


Figure 7. Comparison of EM images generated by Latent Diffusion Models (LDMs) with 3-layer and 4-layer U-Net architectures. The left panel shows images generated with the 3-layer U-Net, while the right panel presents images generated with the 4-layer U-Net.

| Model | FID | Training time to converge (1 GPU) | Sampling time (50k images) |
|---|---|---|---|
| StyleGAN | 58.96 | 18 hr | **8 min** |
| LDM | **36.1** | **6 hr** | 30 hr |

Table 3. Comparison of StyleGAN and LDM in terms of FID, training time, and sampling time for generating 50k images.

mative comparison of the generated images.

## References


[1] Shane Barratt and Rishi Sharma. A note on the inception score, 2018. 3

[2] Xiaoyu Cheng, Chenxue Xie, Yulun Liu, Ruixue Bai, Nanhai Xiao, Yanbo Ren, Xilin Zhang, Hui Ma, and Chongyun Jiang. Image segmentation of exfoliated two-dimensional materials by generative adversarial network-based data augmentation. *Chinese Physics B*, 33(3):030703, mar 2024. 1

[3] Sejune Cheon, Hankang Lee, Chang Ouk Kim, and Seok Hyung Lee. Convolutional neural network for wafer surface defect classification and the detection of unknown defect class. *IEEE Transactions on Semiconductor Manufacturing*, 32(2):163–170, May 2019. 1

[4] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 1

[5] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024. 3

[6] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 2

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 2

[9] Abid Khan, Chia-Hao Lee, Pinshane Y. Huang, and Bryan K. Clark. Leveraging generative adversarial networks to create realistic scanning transmission electron microscopy images. *npj Computational Materials*, 9(1):85, May 2023. 1

[10] Wei Li, Kevin G. Field, and Dane Morgan. Detection of open loop defects in stem images of irradiation-damaged alloys – source code for detection and image dataset, 2018. 2

[11] Chixiang Lu, Kai Chen, Heng Qiu, Xiaojun Chen, Gu Chen, Xiaojuan Qi, and Haibo Jiang. Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nature Communications*, 15(1):4677, Jun 2024. 1

[12] Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makaravets, Dzianis Pirshtuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter. Face generation and editing with stylegan: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[13] Björn Ommer Patrick Esser, Robin Rombach. Taming transformers for high-resolution image synthesis, 2021. 2

[14] Graham Roberts, Simon Y. Haile, Rajat Sainju, Danny J. Edwards, Brian Hutchinson, and Yuanyuan Zhu. Deep learning for semantic segmentation of defects in advanced stem images of steels. *Scientific Reports*, 9(1), Sept. 2019. 1

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[16] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0. 3

[17] Mingren Shen, Guanzhao Li, Dongxia Wu, Yuhan Liu, Jacob R.C. Greaves, Wei Hao, Nathaniel J. Krakauer, Leah Krudy, Jacob Perez, Varun Sreenivasan, Bryan Sanchez,

Oigimer Torres-Velázquez, Wei Li, Kevin G. Field, and Dane Morgan. Multi defect detection and analysis of electron microscopy images with deep learning. *Computational Materials Science*, 199:110576, 2021. 1

[18] Ervin Tasnadi, Alex Sliz-Nagy, and Peter Horvath. Structure preserving adversarial generation of labeled training samples for single-cell segmentation. *Cell Reports Methods*, 3(9):100592, 2023. 1

[19] Wenhao Yuan, Bingqing Yao, Shengdong Tan, Fengqi You, and Qian He. Deep generative models-assisted automated labeling for electron microscopy images segmentation, 2024. 1