From Synthesis to Discrimination: Deep Learning Techniques for Analyzing AI-Generated SEM Images

A thesis submitted in fulfilment of the requirements

for the degree of Master of Technology in Computer Science & Engineering

by

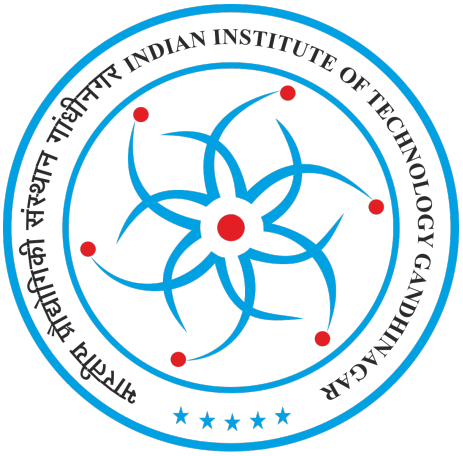
Ashish Kumar Sah

under the supervision of

Dr. Shanmuganathan Raman

and co-supervision of

Dr. Karthik Subramaniam Pushpavanam



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GANDHINAGAR

May 20, 2024

Certificate

It is certified that the work contained in this thesis entitled "From Synthesis to Discrimination: Deep Learning Techniques for Analyzing AI-Generated SEM Images" by "Ashish Kumar Sah" has been carried out under my supervision and that it has not been submitted elsewhere for a degree.

Dr. Shanmuganathan Raman
Associate Professor
Jibaben Patel Chair in Artificial Intelligence
Electrical & Computer Science and Engineering
Indian Institute of Technology Gandhinagar

Dr. Karthik Subramaniam Pushpavanam
Assistant Professor
Chemical Engineering
Indian Institute of Technology Gandhinagar

Abstract

In the field of materials sciences, the integrity of scanning electron microscope (SEM) images holds great importance. SEM images are used by researchers across multiple disciplines, such as material sciences, geology and many others and industries like semiconductor manufacturing, automotive, etc, heavily rely on these images to maintain the quality of their products. With recent instances of forgery detected in research papers utilizing generated SEM images, robust discrimination between authentic and artificially generated images has become necessary. This thesis presents a pivotal advancement in combating such forgery by leveraging deep learning techniques to distinguish between real images and fake images generated by GANs. Leveraging FastGAN and StyleGAN ADA, two state-of-the-art GAN models, synthetic images were generated alongside authentic SEM images. Subsequently, a ResNet-based classification model, augmented by nearest pixel relationship analysis, was employed to classify between the two image categories. Notably, the model achieved a commendable test accuracy of 96%, marking a significant stride towards mitigating forgery in materials sciences research. This work signifies the potential of deep learning methodologies in safeguarding the integrity of scientific imaging data, thereby fortifying the credibility of research outcomes in the materials sciences domain.

Acknowledgements

I am immensely thankful to my advisor, Dr. Shanmuganathan Raman & co-advisor Dr. Karthik Subramaniam Pushpavanam, who provided me with guidance, support, and motivation through all the phases of my thesis work. His insightful comments and approaches helped me finalize the thesis work, from beyond the level of unostentatious into something concrete. I thank my lab members and friends Prajwal Singh, Ashish Tiwari, Akbar Ali, Rushali R, Soumyaratna Debnath, Pooja Goel, Vipul Baghel, Satyam Bhardwaj, & Shivam Mishra for the discussion and all the helps they did for my thesis work. I would also like to thank my family and friends for their love, belief and moral support.

Contents

List of Figures

List of Tables

Abbreviations

SEM	**S**canning **E**lectron **M**icroscope
GAN	**G**enerative **A**dversarial **N**etwork
MSE	**M**ean **S**qaure **E**rror
AI	**A**rtificial **I**ntelligence
NPR	**N**eighboring **P**ixel **R**elationship
ADA	**A**daptive **D**iscriminator **A**ugmentation
GPU	**G**raphics **P**rocessing **U**nit
BCE	**B**inary **C**ross **E**ntropy
CNN	**C**onvolutional **N**eural **N**etwork

Symbols

$\mathbb{R}$	Real Number
V_x	Image grid of x'th patch
$l(x, y)$	Loss between x and y
$\sum_{k=1}^{n}$	Sum of elements from 1 to n

Dedicated to my Family...

Chapter 1

Introduction

Counterfeiting presents a significant obstacle across various sectors, from financial systems and luxury markets to the digital landscape and identity verification processes. Counterfeiting has been a persistent challenge throughout history, impacting economies and industries worldwide. The circulation of counterfeit currency, for instance, has long troubled governments and financial institutions due to its potential to erode trust in monetary systems and facilitate illicit activities. Similarly, the production of counterfeit luxury goods not only violates intellectual property rights but also deceives consumers and undermines legitimate businesses, posing significant ethical and economic concerns.

In the digital era, counterfeiting has evolved significantly, driven by the widespread use of digital media and the accessibility of technology to the general population. The emergence of deep fakes, leveraging AI and machine learning algorithms to generate convincing yet fabricated audio and video content, is of particular concern. These sophisticated manipulations raise questions about media authenticity and trustworthiness, with far-reaching implications across various sectors, including politics, entertainment, and cybersecurity. Recent advancements in image synthesis techniques, such as diffusion and GANs, have sparked public concern, prompting fears that discerning between real and fake imagery becomes increasingly challenging[4]. According to Hong Kong police, in a recent incident, a financial employee at a multinational corporation was deceived into transferring $25 million to scammers who employed deepfake technology to impersonate the company's chief financial officer during a video conference.[3]. Therefore, the ability to discriminate between the real and fake becomes very important.

In the field of material sciences, Scanning Electron Microscope (SEM) images play a crucial role in quality control processes, enabling the examination of material surface quality and

integrity. Industries like semiconductor manufacturing, aerospace, and automotive heavily rely on SEM analysis to uphold stringent quality standards. SEM images also serve as indispensable tools for scientific research across various disciplines, including materials science, nanotechnology, biology, and geology. However, the integrity of SEM imagery has come into question, with instances of fake SEM images surfacing in research publications[7]. The Retraction Watch database contains over 51,000 recorded retractions, corrections, or expressions of concern. Approximately 4% of these cases raise concerns regarding images [1]. Renee Hoch from the PLOS Publication Ethics team in San Francisco, California, suggests that the heightened reporting of image-related problems may stem from an increase in whistle-blowing, attributed to the global community's heightened awareness of integrity issues.

Sadly, not all researchers take the time to verify whether the images they use are genuine. If fake SEM images are used in research, it can lead to disastrous outcomes. Imagine if further work is based on these false images— the entire research could be headed in the wrong direction. The end result might not only be inaccurate but could also have harmful effects. So, it's not just about keeping scientific papers honest; it's about ensuring the safety and reliability of the research that impacts our lives.

Such forgery has raised the critical need for robust methodologies to ascertain the authenticity of such imagery. Therefore, addressing this problem is not only imperative for upholding the integrity of scientific publications but also for ensuring the reliability and trustworthiness of research findings in the broader scientific community.

In this thesis work, we have attempted to address the critical challenge of discerning between authentic scanning electron microscope (SEM) images and those artificially generated by generative adversarial networks (GANs). Leveraging deep learning methodologies, we developed and trained a classification model capable of accurately distinguishing between real SEM images and synthetic counterparts. To achieve this, we used a diverse dataset[2] comprising authentic SEM images alongside artificially synthesized ones. Utilizing state-of-the-art techniques such as FastGAN and StyleGAN-ADA, we synthesized SEM-like images to augment the dataset. Furthermore, we integrated a novel approach leveraging neighboring pixel relationships (NPR) as artifacts representations to enhance the classifier's performance in detecting deepfake SEM images. Through extensive experimentation and evaluation, we demonstrated the efficacy and robustness of the developed methodology, paving the way for enhanced authenticity verification in the field of materials science and beyond.

Chapter 2

Literature Review

In this chapter, we will explore techniques for generating new images from existing ones and explore methodologies for distinguishing between generated and authentic images.

2.1 Generative Networks

2.1.1 Introduction

Generative networks represent a significant advancement in artificial intelligence, especially for creating synthetic data like images. They've become widely recognized in various fields because they can mimic real data well, helping with tasks such as making new images, adding more data to a dataset, and finding anomalies. This section aims to thoroughly explain generative networks, covering their basic principles, structures and uses. It also discusses recent progress and challenges researchers have faced.

Generative networks are a type of artificial neural network designed to understand and copy the patterns in a given dataset. Unlike other models that focus on categorizing data, generative models aim to capture the overall structure and traits of the dataset, making it possible to create new samples that look like the original data. One of the most well-known types of generative networks is called Generative Adversarial Networks (GANs). These networks work by having two parts: a generator, whose work is creating new samples and a discriminator that checks if the samples are real or fake. They compete against each other in an adversarial setup during the training process aiming to improve the quality of the generated samples of image.

2.1.2 Generative Adversarial Networks

Introduced by Goodfellow et al. in 2014, GANs represent a class of generative models. This comprises two interconnected neural networks: a generator and a discriminator [6]. The generator network is tasked with mapping random noise vectors sampled from a prior distribution to synthetic data samples, whereas the discriminator network is trained to distinguish between real and fake images. Throughout the training process, the generator strives to generate image samples that closely resemble real images, while the discriminator works to correctly classify between real and synthetic image samples. Through adversarial training, GANs continuously improve the generator's capacity to produce realistic data distributions that follow the image distribution, resulting in synthetic samples of high-quality images that closely emulate the attributes of the real images. This iterative refinement process has led to significant success in diverse applications, including image generation, style transfer, and image-to-image translation.

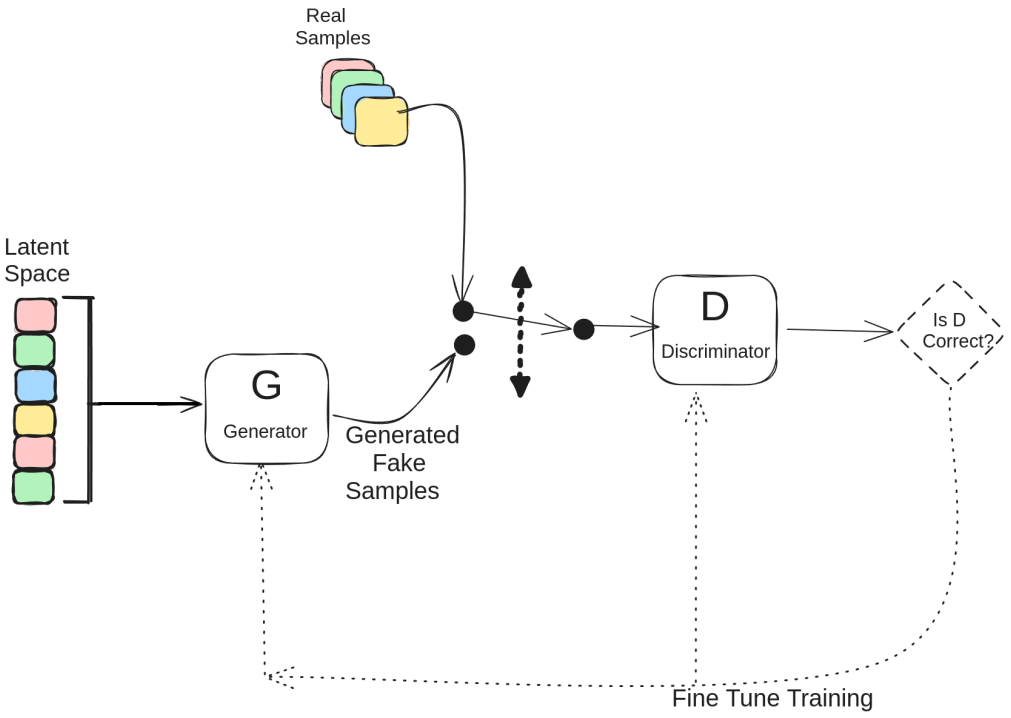


FIGURE 2.1: General GAN Architecture

In their nascent development stages, Generative Adversarial Networks (GANs) encountered hurdles such as training instability and mode collapse, where the generator repetitively produced similar outputs. Moreover, the assessment of generated image quality

lacked standardized metrics. In 2016, Alec Radford et al.[15] introduced Deep Convolutional Generative Adversarial Networks (DCGAN), a GAN variant that mitigated training instability and enhanced image fidelity. However, while DCGAN addressed certain challenges, the controllability of generated outputs remained constrained, posing difficulties in steering the generator toward specific desired features. Subsequent efforts focused on diversifying architectures, leading to innovations like StyleGAN[10] and ProGAN[8], renowned for their remarkable photorealism in image generation. Conditional GANs, incorporating controllable inputs such as captions or sketches, also emerged. GANs found applications across diverse domains, including medical imaging for synthetic data generation to train diagnostic models and time series forecasting. Despite these strides, the demand for extensive training datasets persisted, hindering the generation of diverse images with limited input.

Training a GAN on a small dataset presents a unique challenge. The discriminator, tasked with distinguishing real from generated images, can easily overfit the limited training data. This overfitting leads to the discriminator becoming overly confident in its classifications, causing the training process to diverge and fail to converge on a good solution.

While traditional GAN training struggles with data scarcity, resulting in discriminator overfitting and training divergence, recent advancements offer promising solutions for realistic image generation, even with limited input examples.

StyleGAN2-ADA (Adaptive Discriminator Augmentation)[9] represents a recent breakthrough addressing the overfitting challenge in few-shot image synthesis. Building upon the successful StyleGAN2 architecture known for its exceptional image quality, StyleGAN2-ADA introduces an adaptive augmentation mechanism to the discriminator during training. This mechanism dynamically adjusts data augmentation based on the current training stage.

In the research paper by Bingchen Liu et al.[11], the focus is on tackling the challenge of training GANs on limited datasets to generate high-fidelity images with minimal computational resources. This architecture is commonly known as fastGAN, and it introduces a very specialized module called the “Skip-Layer Channel-wise Excitation Module”. This module enhances information flow within the generator network. This helps to enable the model to effectively leverage limited training data by capturing essential features in the scant available samples.

2.2 Classification Network

Traditional methods for image authenticity verification include manual inspection, where experts visually examine images for signs of tampering or manipulation. Watermarking involves embedding digital signatures or identifiers into images to track their origin and authenticity, while image forensics techniques analyze metadata, compression artifacts, and pixel inconsistencies to detect alterations or forgeries. Despite their usefulness, these methods are time-consuming, prone to human error, and somewhat limited in detecting sophisticated manipulations.

In recent years, Generative AI has introduced a new challenge, as synthesized images exhibit remarkable realism, rendering them exceedingly difficult to identify using traditional verification methods. Consequently, there has been a surge in research efforts to enhance the capability of deep learning techniques to address this emerging concern. Rossler et al.[16] extensively evaluated methods designed to detect facial manipulation techniques, employing CNN-based approaches tailored explicitly for detecting facial and mouth replacements. Similarly, Marra et al.[12] demonstrated the efficacy of employing simple classifiers to identify images generated by image translation networks, although they did not explore cross-model transfer.

Building upon this foundation, Sheng-Yu Wang et al. developed a model capable of discerning between real and AI-generated images, regardless of variations in architecture or dataset[20]. Subsequent research endeavours have focused on refining Wang et al.'s methodology, with Chuanghuang Tan et al. augmenting their approach by training over feature vectors derived from both real and synthetic images. Over time, performance enhancements were achieved by mitigating upsampling artifacts present in both generated and authentic images[19].

However, it is essential to note that while these models initially showed promise when trained on conventional CNN image datasets, they encountered challenges in accurately discerning Scanning Electron Microscope (SEM) images. As a result, modifications were necessary to adapt the existing architecture to accommodate the specific distribution and characteristics of SEM images, highlighting the ongoing need for tailored solutions to address diverse image types and modalities.

Chapter 3

Methodology

3.1 Dataset and Pre-processing

Our dataset comprises two class image folders: "real" and "fake." The "real" folder contains scanning electron microscope (SEM) images obtained from the publication at https://www.nature.com/articles/s41597-020-0439-1. These images were selected as they provide authentic representations of SEM imagery. However, as the original images included metadata at the bottom, they underwent cropping and resizing to remove this information, ensuring consistency and focusing solely on the image content. Despite this preprocessing step, no further adjustments were deemed necessary for the real images to maintain their integrity. It is worth noting that the total number of real images in our dataset was limited to 750, reflecting the scarcity of publicly available SEM images suitable for training purposes.

In contrast, the "fake" folder contains synthetic images generated using two advanced unconditional Generative Adversarial Networks (GANs): StyleGAN-ADA and FastGAN. These GANs were explicitly chosen for their efficacy in producing synthetic images that closely resemble SEM images, making them ideal for our dataset. Unlike traditional GANs, StyleGAN-ADA and FastGAN work well even in scenarios where training data is less, as is often the case with SEM imagery. By leveraging these state-of-the-art GANs, we aimed to augment our dataset with diverse synthetic images, enhancing its utility for training and evaluation.

The synthetic images generated by StyleGAN-ADA and FastGAN undergo a rigorous validation process to ensure their fidelity and relevance to SEM imagery. Despite their

artificial nature, these images exhibit characteristics and features akin to authentic SEM images, thereby enriching the diversity of our dataset. Researchers and practitioners can access the final dataset at SEM Detection Final Dataset, for download and further utilization in their respective projects. This dataset represents a essential resource for advancing research in SEM image detection and related domains, offering a comprehensive collection of both real and synthetic SEM images for training and evaluation purposes.

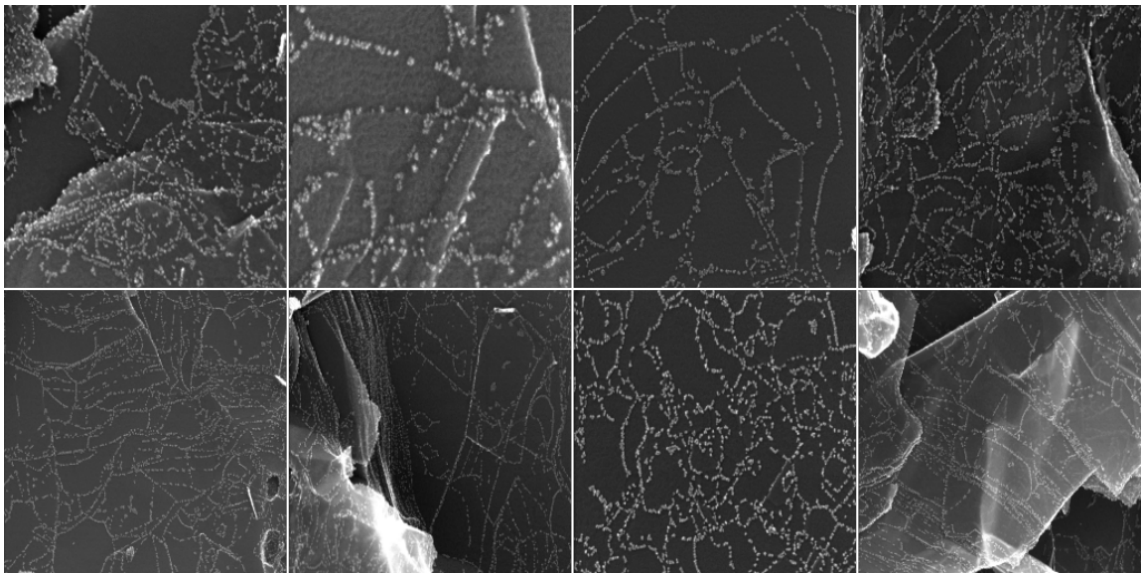


FIGURE 3.1: Real SEM Images Sample

The samples in Figure 3.1 represent a subset of real SEM image data. These eight samples were randomly chosen from a pool of 750 images to provide a representative illustration.

3.2 Fake Image Synthesis

This section delves into the intricacies of the two types of Generative Adversarial Networks (GANs) employed in our study to generate images for the fake folders. Our focus will primarily be on their architectural frameworks and components, as well as the functioning of their generator and discriminator networks.

3.2.1 StyleGAN-ADA

StyleGAN-ADA marks an advancement from the original StyleGAN model, introducing the concept of Adaptive Discriminator Augmentation (ADA) to enhance training stability

and improve image quality. The main architecture is the same as any other GAN consisting of generator and discriminator networks. During adversarial training, the generator network iteratively enhances its image-generation skills by attempting to outwit the discriminator. Simultaneously, the discriminator endeavours to classify accurately between real and generated images in response to the generator's improvements.

The Generator Network (G) within StyleGAN2-ADA encompasses a sophisticated multi-stage architecture.

1. **Mapping Network:** This initial network transforms a latent vector, a random vector encapsulating the essence of the desired image, into an intermediate latent space.

2. **Style Encoder:** This network further processes the latent code to extract style information that governs various image attributes like facial features, hair colour, etc (in our case, the image attribute can be the relative positions and structure of materials). This style information is crucial for generating image variations.

3. **Synthesis Network:** The final stage employs the processed latent code and extracted style information to progressively construct the image, pixel by pixel, across escalating resolutions. This network leverages techniques like transposed convolutions to upscale the image and progressively add details.

On the other hand, the **Discriminator Network (D)** functions as a critic, discerning between real and generated images. It accepts an image as input and outputs a probability score indicating the likelihood that the image is real. Unlike the original StyleGAN, StyleGAN2-ADA incorporates Adaptive Discriminator Augmentation (ADA) during training.

Adaptive Discriminator Augmentation (ADA) represents a pivotal innovation within StyleGAN2-ADA. During each training step, the input images presented to the discriminator undergo minor, random transformations such as cropping, flipping, or colour jittering. This strategy compels the discriminator to become more robust against such transformations, ultimately leading to the generation of images that are less susceptible to artifacts caused by these common data augmentation techniques.

3.2.2 Fast-GAN

FastGAN is a specialized variant of GAN tailored to produce high-fidelity images with minimal training data and computational resources. In contrast to conventional GANs, which often demand extensive training data, FastGAN exhibits the capability to achieve satisfactory outcomes with as few as 100 images[11]. This feature makes it particularly advantageous in scenarios characterized by data scarcity. Additionally, the architecture of FastGAN is lightweight, allowing for training on a single Graphics Processing Unit (GPU) within a short timeframe, typically a few hours. This represents a significant acceleration compared to other GAN models, making FastGAN an efficient solution for image generation tasks.

FastGAN incorporates two crucial techniques aimed at enhancing both efficiency and image quality. Firstly, the SLE plays a pivotal role in facilitating the network to capture long-range dependencies within the data. This module enhances the network's ability to learn effectively, even when trained on limited data. Secondly, the self-improving discriminator serves as a feature encoder, thereby improving the quality of the generated images.

The Generator Network (G) within the Fast-GAN incorporates two crucial techniques to enhance efficiency and image quality.

1. **Skip-layer Channel-wise Excitation (SLE) module:** This module facilitates the flow of gradients across different resolutions within the network. This improved gradient flow allows the network to learn effectively even with limited training data.
2. **Autoencoder Regularization:** The generator is trained with the additional objective of reconstructing the input noise vector from the generated image. This enforces a form of self-supervision, encouraging the network to learn a more meaningful latent space representation, even with few training images.

The **Discriminator Network** in Fast-GAN is designed to be lightweight. It employs a standard convolutional neural network architecture to classify between real and generated images. However, unlike StyleGAN2-ADA, FastGAN does not incorporate techniques like Adaptive Discriminator Augmentation (ADA), focusing instead on its lightweight architecture and efficient training process.

3.3 Training the Classifier

3.3.1 The Artifact

In the context of a GAN framework, the generator function assumes a central role by establishing a mapping from a lower-dimensional latent space to the image space. This mapping process incorporates two fundamental components within the generator's architecture: convolutional layers and up-sampling layers. Convolutional layers are tasked with extracting features from the input latent space while up-sampling layers are instrumental in enhancing the resolution of the generated images. Specifically, up-sampling layers function by taking low-resolution features as input and outputting high-resolution features, facilitating the generation of images with increased detail and fidelity.

Despite the varied architectural configurations observed across different GAN models, the incorporation of up-sampling modules persists as a consistent practice due to their efficacy in enhancing image resolution. However, it is imperative to recognize that the utilization of up-sampling operations can inadvertently introduce artifacts that may impact the quality of the generated images. These artifacts, which encompass phenomena like blurriness or pixelation, possess the potential to be exploited by classifiers to discern between authentic and synthesized images. A comprehensive understanding of the ramifications associated with up-sampling operations is paramount for the development of robust GAN models capable of generating high-fidelity images while mitigating the occurrence of artifacts.

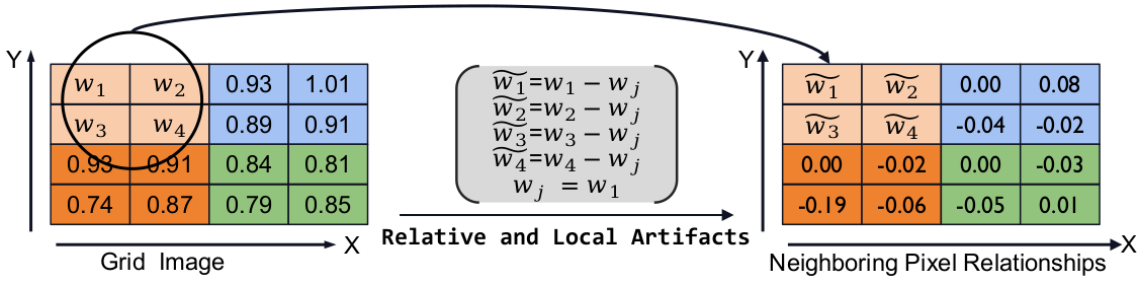


FIGURE 3.2: The Overview of Neighboring Pixel Relationships

Figure 3.2 depicts the overview of relative and local artifacts present in an image. The image was adopted from Rethink Upsampling by Tan et al.[19]. This method focuses on the portion of the generator near the output images, consisting of an up-sampling layer up with l scale, convolutional layers $conv$ with activate functions, input feature maps $x \in \mathbb{R}^{W*H*C}$, and the output images $I \in \mathbb{R}^{(l*W)*(l*H)*3}$.

$$\bar{X} = up(X) \tag{3.1}$$

$$I = conv(\bar{X}) \tag{3.2}$$

where $\bar{X} \in \mathbb{R}^{(I*W)*(I*H)*C}$ is the up-scaled feature map. The Image I is divided into $H * W$ grid where each grid is $L * L$ patch. Let V_x denote the grid set of the up-scaled featured map. The elements of each grid of V_x exhibit a strong correlation generated by the up-sampling layer. For instance, when adopting nearest neighbour interpolation as the up-sampling layer, the elements in a grid of V_x share the same value. This method captures the correlation of local pixels in a grid as the up-sampling artifacts.

This approach leverages the concept of neighbouring pixel relationships (NPR) as representations of artifacts to train the classifier for detecting fake images. The NPR captures the local, relative correlations between pixels within local patches of an image. These correlations, which are evident in the image domain, originate from the upsampling layer and take advantage of the translation invariance property of the convolutional layer. By focusing on the relative and local characteristics of the proposed upsampling artifacts, this approach aims to facilitate the generalization of neighbouring pixel relationships to unfamiliar sources.

In essence, the utilization of NPR as representations of artifacts provides a structured framework for training the classifier to discern between real and fake images. By analyzing the correlations between neighboring pixels within local patches, the classifier can effectively identify patterns indicative of up-sampling artifacts commonly found in fake images. This approach capitalizes on the inherent structure and characteristics of images, allowing for the detection of subtle cues that may indicate image manipulation or generation.

3.3.2 Network Architecture

In developing our network architecture, we crafted a highly efficient Convolutional Neural Network (CNN) model augmented with convolution layers and ResNet, boasting approximately 5 thousand parameters. This architecture was meticulously designed to strike a balance between computational efficiency and performance, ensuring optimal utilization of available resources while maintaining competitive accuracy levels.

The process begins with the input image (x) traversing through the network, where it undergoes a series of operations aimed at enhancing its features and representations. One notable step involves upsampling the image using the nearest neighbour's interpolation method, followed by downsampling to restore it to its original size. This upsampling and downsampling process helps to refine the image's details while preserving its overall structure and characteristics.

After this preprocessing stage, the resulting image $\bar{x}$ is subjected to a subtraction operation, wherein it is subtracted from the original input image (x). This operation enables the extraction of meaningful information and features from the image, which are crucial for subsequent classification tasks. Finally, the processed image is fed into the proposed classification model, where it undergoes further analysis and evaluation to determine its class label or characteristics.

By employing this lightweight CNN architecture coupled with upsampling, downsampling, and subtraction operations, we aim to harness the power of deep learning for effective image classification. This approach not only ensures computational efficiency but also enhances the model's ability to extract relevant features and make accurate predictions. Through rigorous experimentation and optimization, we strive to achieve superior performance and reliability in various image classification tasks, thereby contributing to advancements in the field of computer vision and machine learning.

Loss Function The BCEWithLogitsLoss function serves as a prevalent loss function frequently employed in binary classification tasks, particularly within deep learning frameworks such as PyTorch. This function amalgamates the sigmoid function, which compresses output values within the range of 0 to 1 with binary cross-entropy loss. The term "logits" pertains to the raw output values generated by the model prior to the application of the sigmoid function. BCEWithLogitsLoss effectively applies the sigmoid function to these logits to derive predicted probabilities, subsequently calculating the binary cross-entropy loss between these probabilities and the target labels. This approach facilitates the model's ability to assess the probability distribution of the predicted outcomes and determine the extent of deviation from the actual labels, thereby enabling effective optimization and enhancement of model performance in binary classification tasks.

$$l(x, y) = L = \{l_1, ..., l_N\}^T \tag{3.3}$$

$$l_n = -w_n[y_n log\sigma(x_n) + (1 - y_n)log(1 - \sigma(x_n))] \tag{3.4}$$

where N is the batch size.

$l(x, y)$ denotes the loss of a batch, and the loss of individual elements in a batch is denoted by l_n in equation 3.4.

x_n is the ground truth for n^th item and,

y_n is the predicted value for the same item.

Chapter 4

Experiments and Results

4.1 Generating Images using Fast-GAN

In this section, we will discuss the training methods and the qualitative and quantitative results of generated images using Fast-GAN.

4.1.1 Network Training

In the training process of both the generator and discriminator components, we utilized the entirety of our real Scanning Electron Microscope (SEM) dataset, resizing each image to dimensions of 512 * 512 square pixels to ensure uniformity and consistency. Subsequently, a batch comprising eight images was fed into the network, with both the generator and discriminator featuring 64 parameters each, while the latent vector size was set to 256. To enhance the dataset's variability and augment the training process, we applied Random Horizontal Flip transformation to the images. The training regimen spanned a total of 100,000 iterations aimed at refining the model's performance and capabilities. Execution of this training process was conducted on a Quadro RTX 5000 GPU, consuming approximately 12 hours to complete, thus underscoring the computational resources required for robust model training.

4.1.2 Results

The samples depicted in Figure 4.1 represent a subset of images generated using the Fast-GAN method using the training parameters mentioned in section 4.1.1.

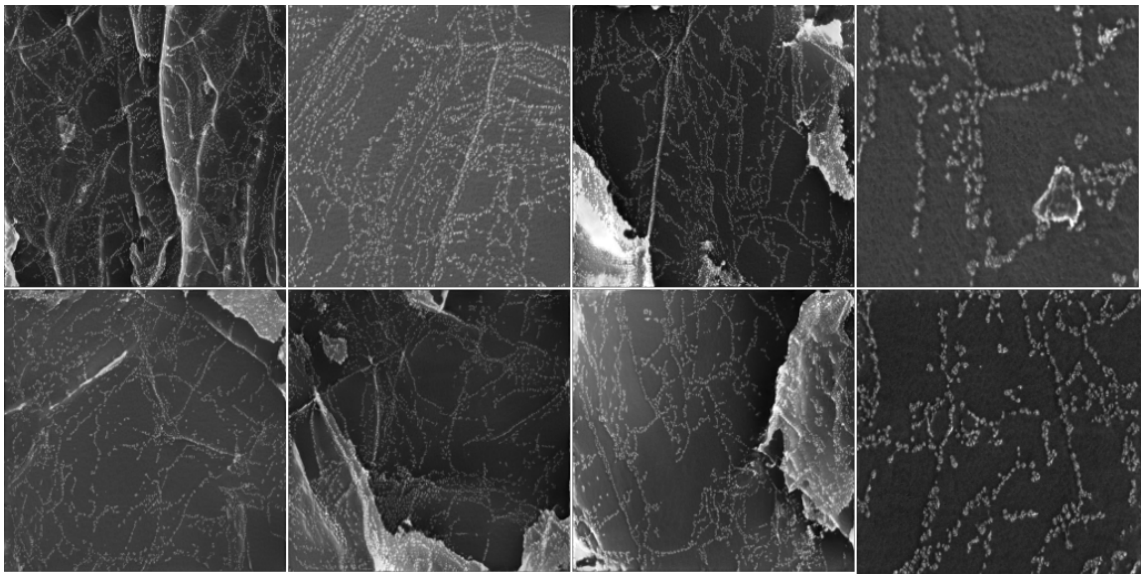


FIGURE 4.1: Sample of Generated Images using Fast-GAN

4.1.3 Quantitative Analysis

The efficacy of this model in producing synthetic images resembling real Scanning Electron Microscope (SEM) images underwent rigorous evaluation through qualitative & quantitative analyses. Two distinct quantitative methods, namely MSE and LPIPS, were employed to assess the fidelity and perceptual quality of the generated images.

Mean Squared Error serves as a fundamental metric to quantify the average squared differences between the pixel values of the synthetic images and their corresponding real counterparts. A lower MSE value indicates a closer resemblance between the synthetic and real images, reflecting superior fidelity and accuracy in image generation. However, it is essential to note that MSE solely measures pixel-wise differences and may not fully capture perceptual discrepancies or nuances in image quality.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

In our experimental procedure, we generated a dataset comprising 2000 synthetic images using our trained model, subsequently computing the Mean Squared Error (MSE) between each generated image and all 750 original images from our dataset.

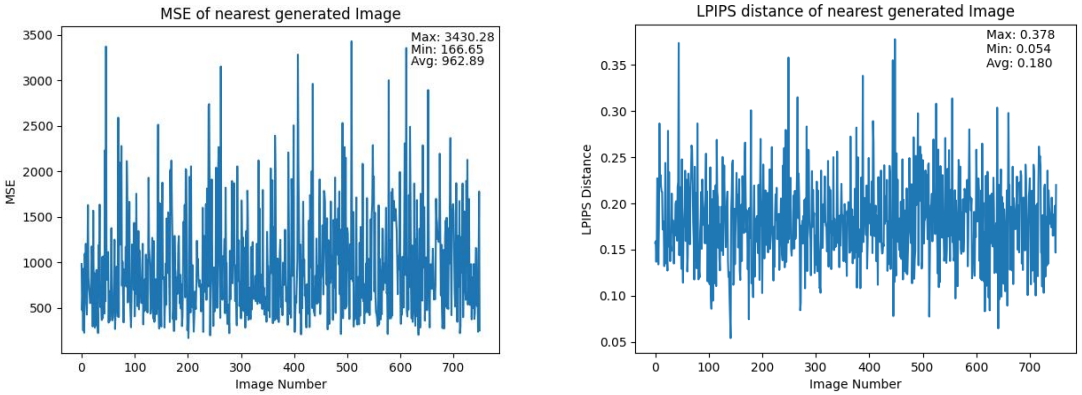


FIGURE 4.2: Minimum MSE(left) & LPIPS(right) Distance between Real and Generated Images

In our analysis, we used MSE to compute the distance between each real image and the set of fake images. We plotted the minimum MSE distance observed for each of the real images. This calculation was used to evaluate the similarity between the generated and real images, as illustrated in Figure 4.2. The minimum MSE distance recorded was 166.65, indicating that none of the generated images precisely replicated any of the real images. Moreover, the average MSE value calculated at 962 suggests a consistent resemblance between the generated images and their real counterparts, implying that the generated images conform to the same underlying distribution as the real images. This analysis provides valuable insights into the fidelity and accuracy of our model in generating synthetic images resembling real SEM images.

Learned Perceptual Image Patch Similarity (LPIPS) offers a more nuanced evaluation by leveraging a learned perceptual distance metric to assess the similarity between image patches. Unlike MSE, LPIPS considers perceptual features and structural characteristics, providing a more comprehensive assessment of image quality and similarity[21]. By considering perceptual factors such as texture, colour, and spatial arrangement, LPIPS offers insights into the perceptual fidelity and realism of the generated images, thereby complementing the quantitative assessment provided by MSE.

In addition to assessing MSE, we evaluated LPIPS between the real and generated images. We computed the LPIPS distance with all generated images for each real image and plotted the minimum distance observed in Figure 4.2. The minimum LPIPS value recorded

was 0.054, indicating a close perceptual match between some of the real and generated images. Furthermore, the average LPIPS value calculated at 0.180 suggests a consistent resemblance between the generated images and their real counterparts, implying that the generated images adhere to the same underlying distribution as the real images. In numerous state-of-the-art image synthesis methodologies, the reported LPIPS distances for dataset variations typically fall within the 0.2 to 0.6[21][17]. The images generated by these Generative Adversarial Networks adhere to the original image distribution while exhibiting increased diversity. Hence, the LPIPS distance observed for our model falls within a reasonable range, indicating that the images generated by our model are of high quality. The fact that the minimum LPIPS distance is not zero suggests that no exact copies were generated, underscoring the diversity and uniqueness of the synthetic images. This observation, coupled with the average LPIPS distance, reinforces the notion that the generated images closely resemble the real ones and adhere to the distribution of the original dataset.

Methods	Mean Distance
MSE	166.65
LPIPS	0.180

TABLE 4.1: Mean distance for Fast-GAN generated images

4.2 Generating Images using StyleGAN-ADA

This section will discuss the training methods and the qualitative and quantitative results of generated images using StyleGAN-ADA.

4.2.1 Network Training

The training methodology employed for generating photorealistic images using StyleGAN-ADA closely resembled that of Fast-GAN, as elucidated in section 4.1.1. However, an additional image augmentation technique, namely image mirroring, was incorporated in this process. The model underwent training with a batch size of 16 over a span of 1000 iterations. Notably, the training duration was approximately 65 hours, conducted on a Quadro RTX 5000 GPU, showcasing the computational resources required for robust model training.

4.2.2 Results

The samples depicted in Figure4.3 represent a subset of images generated using the StyleGAN-ADA method using the training parameters mentioned in section 4.2.1.

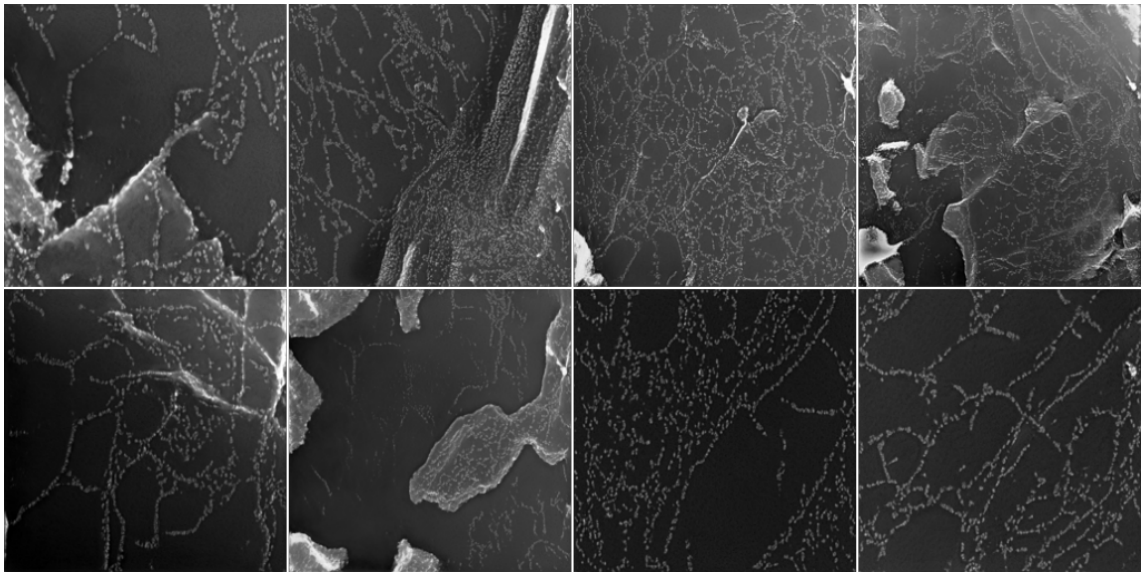


FIGURE 4.3: Sample of Generated Images using StyleGAN-ADA

4.2.3 Quantitative Analysis

In our experimentation, we generated a dataset comprising 1000 synthetic images utilizing StyleGAN-ADA. We conducted a comprehensive analysis by calculating both MSE and LPIPS distances for each original image. The plot depicted in Figure 4.4 illustrates the minimum distance observed between the real and generated images, with the Minimum MSE value recorded at approximately 174. Additionally, the minimum LPIPS value obtained was 0.06, indicating that the generated images are not exact replicas of the real images. Moreover, the average LPIPS values corroborate the quality of the generated images, underscoring their fidelity and perceptual similarity to the real images.

The average MSE and LPIPS distance using StyleGAN-ADA is mentioned in table 4.2

Methods	Mean Distance
MSE	174.87
LPIPS	0.178

TABLE 4.2: Mean distance for StyleGAN-ADA generated images

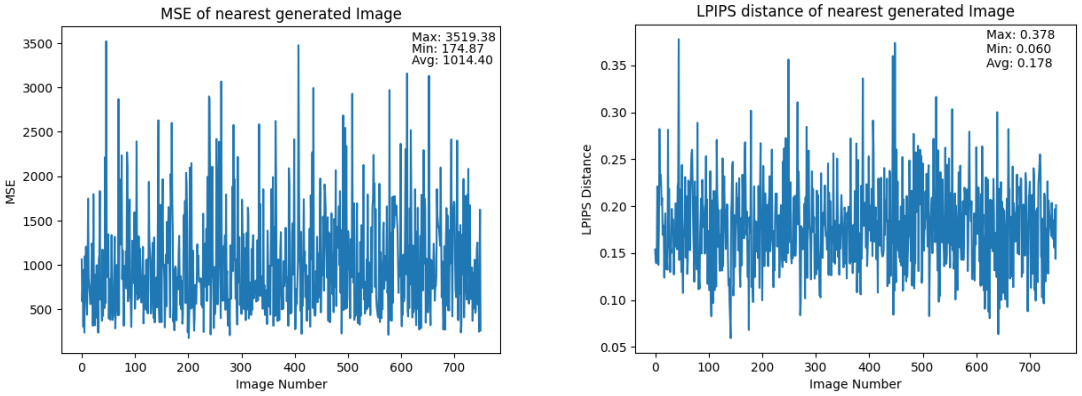


FIGURE 4.4: Minimum MSE(left) & LPIPS(right) Distance between Real and Generated Images

4.3 Classifying Real Vs Generated

This section outlines our classification model's training methodology, offering a concise overview of the techniques and parameters employed to optimize performance and accuracy.

4.3.1 Network Training

We employed the SEM dataset to train the classification model, comprising 750 real images and an equal number of generated images from FastGAN and StyleGAN, totalling 1500 images. This dataset was partitioned into training, testing, and validation sets, with 60%, 30%, and 10% of the data allocated, respectively. The training process involved 100 epochs, utilizing a batch size of 128 and a learning rate 0.0002, conducted on a Quadro RTX 5000 GPU.

4.3.2 Results

The mean accuracy of the model while testing on 30% of the data(test data) was 95.8 %, along with an average precision of 99.5 %.

Class	Accuracy	Average Presion	Accuracy (Real)	Accuracy (Fake)
StyleGAN-ADA	95.1	99.1	96.4	93.8
FastGAN	96.4	99.8	93.8	99.1

TABLE 4.3: Accuracy of Classification Model

4.4 Qualitative Analysis of Classification Model

Grad-CAM, stands for Gradient-weighted Class Activation Mapping. It is one of the techniques used in deep learning for visualizing and understanding the work of a model. It tells about which parts of an input image contribute most to the predictions made by a CNN. It addresses the challenge of interpreting the decisions made by these complex models, especially in tasks like image classification and object detection. Grad-CAM works by utilizing the gradients of the desired class concerning the feature maps of the network's final convolutional layer. By analyzing these gradients, it highlights the regions in the input image that are most influential in the model's decision-making process.

One of the key advantages of Grad-CAM is its ability to provide interpretable and localized visual explanations for CNN predictions without requiring modifications to the model architecture or additional training[18]. Unlike other visualization techniques that rely on complex post-processing or perturbations of input data, Grad-CAM directly utilizes the gradients computed during the backpropagation process, making it efficient and straightforward to implement. Additionally, Grad-CAM produces high-resolution heatmaps that effectively highlight the relevant image regions, aiding in both qualitative analysis and model debugging.

The operational mechanism of Grad-CAM entails calculating the gradients of the target class score concerning the feature maps of the final convolutional layer. These gradients act as significance measures, determining the influence of each feature map on the ultimate prediction. Through a weighted linear amalgamation of the feature maps using these significance measures and the application of a ReLU activation, Grad-CAM generates a heatmap. This heatmap visually highlights the discriminative regions within the input image, offering insights into the model's decision-making process and enhancing transparency and confidence in deep learning systems.

In our analysis, we applied Grad-CAM to visualize the last convolutional layer of our classification model. This allowed us to gain insights into how the model distinguishes between real and fake images based on the features it focuses on. Upon examination

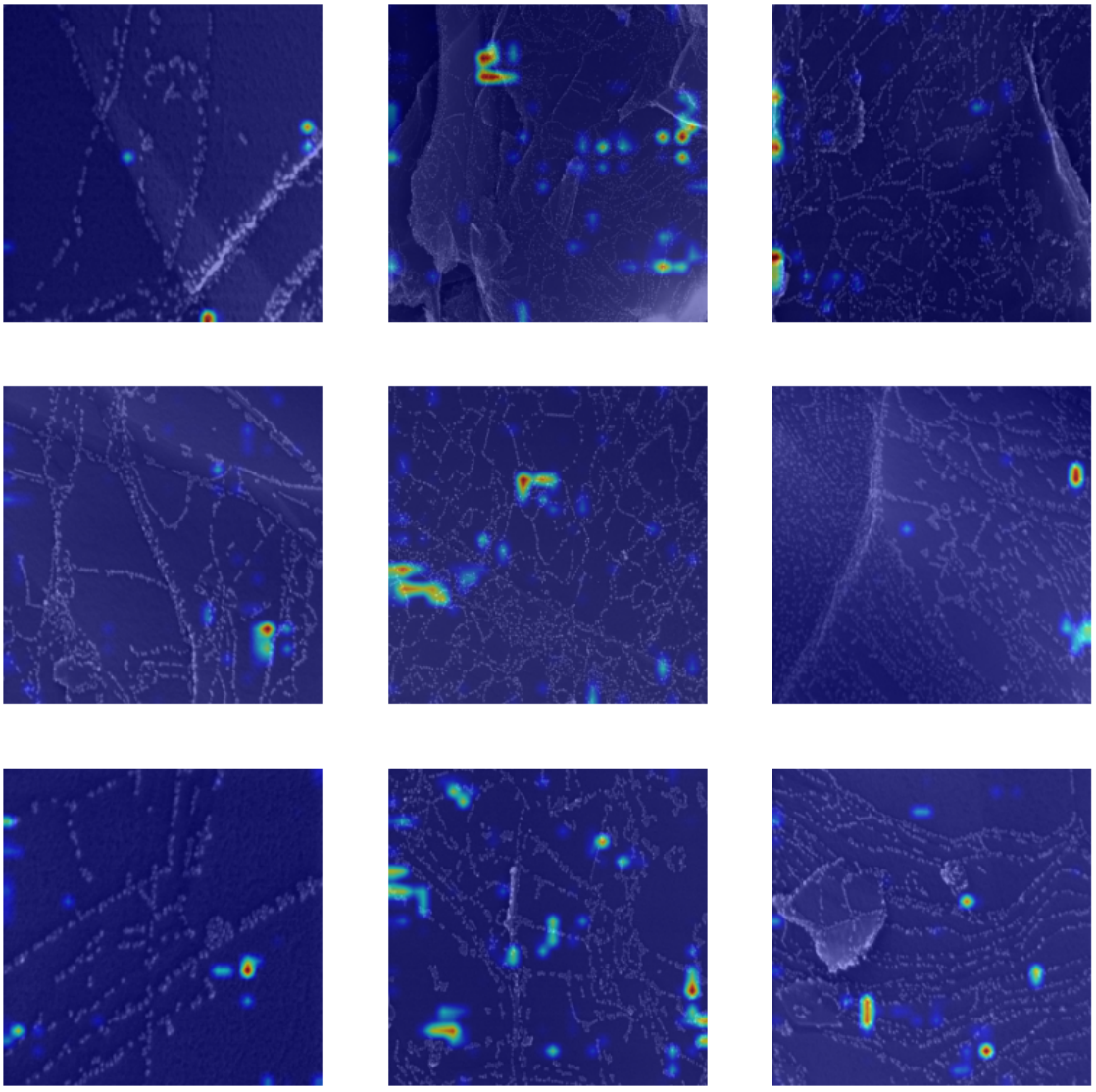


FIGURE 4.5: Grad-CAM Visualization of Real SEM Images

of the Grad-CAM visualizations, notable differences were observed between real and fake images. In the case of real images, the heatmaps focused on specific and localized regions, indicating that the model relied on fine details and characteristic features to make accurate classifications. Conversely, the visualization heatmaps displayed a broader spread for fake images, suggesting that the model struggled to identify distinct features and instead relied on more generalized patterns for classification. These observations underscore the discriminative power of our classification model, as evidenced by the different patterns highlighted in the Grad-CAM visualizations. Figures 4.5 and 4.6 provide visual representations of these findings, showcasing the differences in attentional focus between real and fake images captured by Grad-CAM.

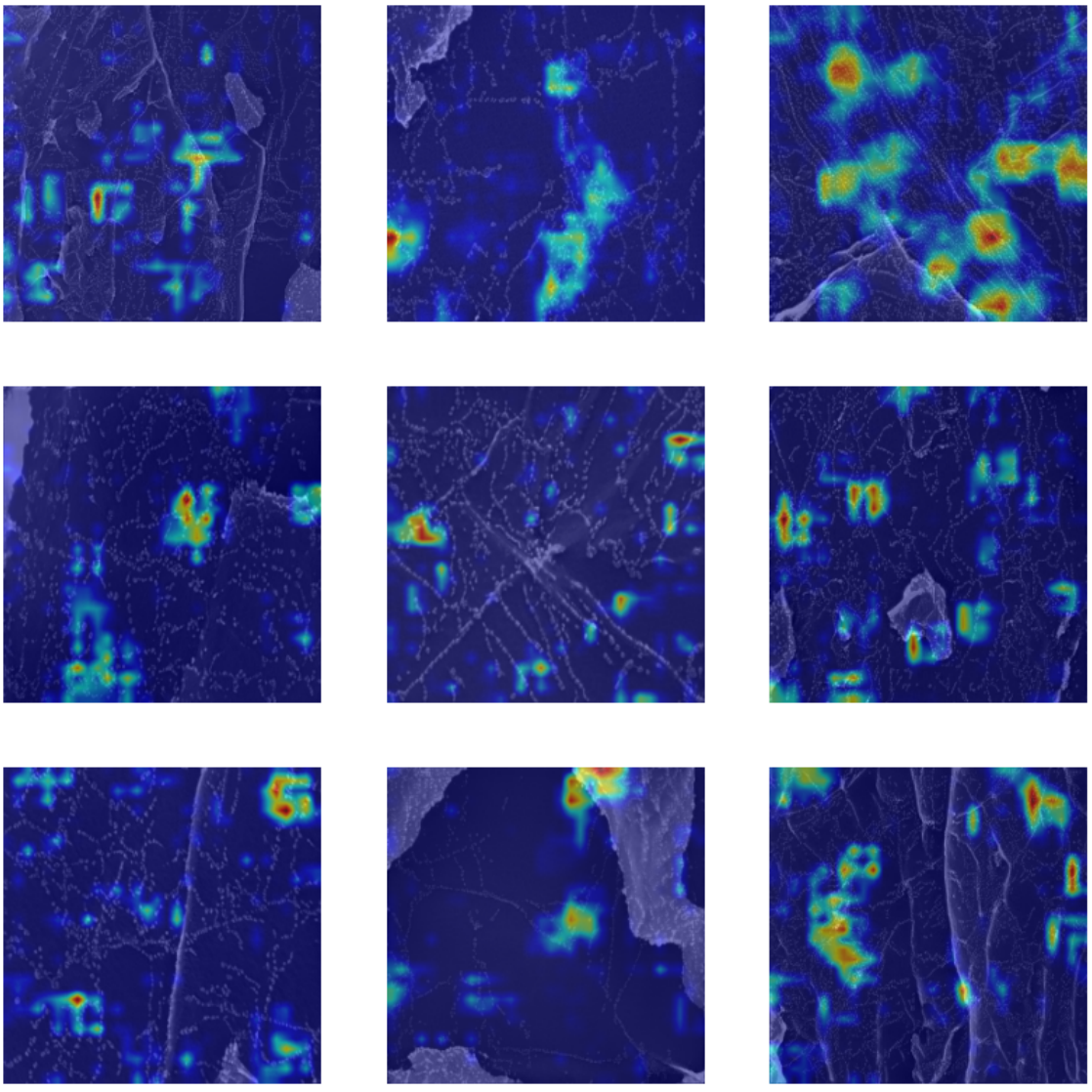


FIGURE 4.6: Grad-CAM Visualization of Generated SEM Images

Chapter 5

Discussion & Future Work

The proliferation of duplicated Scanning Electron Microscope (SEM) images presents a growing concern within academic publications, often resulting in retractions upon detecting fraudulent practices. Instances of such misconduct are documented on platforms like retractionwatch.com[5][14][13]. Images can be tampered with by altering their content through editing tools or by manipulating their appearance to deceive viewers. Copying, cloning, or reusing existing images without authorisation is also one of the problems. People may get away with such image tampering and duplication for a small number of images. However, doing so for a large number of images will be very difficult, and people are most likely to use AI tools for good-quality fake images that human eyes cannot detect. Our results suggest that fastgan and stylegan-generated images retain some detectable fingerprints that distinguish them from real images. However, this work is not robust to images generated using some different models. Modern diffusion models can produce high-quality images, which might be undetectable utilising this model, and we need a robust model to detect images generated from other sources.

Given the resource-intensive nature of training GANs and diffusion models, the availability of ample SEM images becomes paramount for future research endeavours. In our case, the number of images required to train different GANs and diffusion models was limited. Hence, we limited our work to just two GAN models. Therefore, future work mainly depends upon more real SEM images so that we can train various models and then use them to synthesise fake images. This future work will help us create a large dataset; hence, we can train a robust classification model.

Despite advancements in current techniques, several practical concerns persist. Firstly, even highly effective forensic detection systems may face challenges balancing between

accurately detecting fake images and minimising false positives. Malicious users often aim to create individual fake images, making it easier for them to bypass detection thresholds by selectively choosing images that evade detection. Secondly, extensive evaluation is necessary to assess the ability of existing detectors to handle various transformations, including compression, resizing, and re-sampling. These transformations can alter the characteristics of images, posing challenges for detection algorithms and highlighting the need for further research and refinement.

It is essential to recognise that detecting fake images represents only a fraction of the broader challenge of combating visual disinformation. Effective solutions require a holistic approach that spans technical, social, and legal domains to mitigate the proliferation of fraudulent visual content. By leveraging a combination of strategies, including enhanced detection techniques and robust regulatory frameworks, stakeholders can collaboratively address the multifaceted nature of this evolving threat landscape.

Chapter 6

Conclusion

Throughout this thesis, our primary objective was to develop and evaluate deep learning-based methodologies for discriminating between natural images and those generated by generative adversarial networks (GANs) using scanning electron microscope (SEM) imagery. Our experiments demonstrated the effectiveness of GAN models, specifically FastGAN and StyleGAN ADA, in generating synthetic SEM images that closely resemble real-world counterparts. Furthermore, our classification model, based on a ResNet architecture and nearest pixel relationship, exhibited high accuracy in distinguishing between real and generated images, achieving a test accuracy of 96%. This research contributes to advancing the intersection of deep learning and materials science by providing novel methodologies for analyzing SEM imagery and detecting potential instances of forgery or manipulation. By demonstrating the feasibility of leveraging GANs and deep learning techniques for authenticity verification, our work addresses a critical need to ensure the integrity and trustworthiness of scientific publications. The ability to detect fake or manipulated SEM images can enhance the credibility of scientific publications and prevent the dissemination of erroneous or misleading information.

While our research has yielded promising results, it is not without limitations. Future research efforts could explore the use of more advanced GAN architectures and training techniques to further improve the quality of generated images. Additionally, expanding the dataset and incorporating additional modalities, such as elemental mapping data, could enhance the generalization capability and robustness of the classification model.

We hope our findings will inspire further research and collaboration in this rapidly evolving field.

Bibliography

[1] Bik, E. M., Casadevall, A., and Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *MBio*, 7(3):10–1128.

[2] Boiko, D., Pentsak, E., Cherepanova, V., and Ananikov, V. (2020). Electron microscopy dataset for the recognition of nanoscale ordering effects and location of nanoparticles — Dataset 1 (ordered).

[3] Chen, H. (2024). Finance worker pays out $25 million after video call with deep fake 'chief financial officer'. *CNN*.

[4] Farid, H. (2016). *Photo forensics*. MIT press.

[5] Ferguson, A. (2015). Nothing gold can stay: gold nanoparticle paper retracted for figure theft. In *Materials Research Express*. https://retractionwatch.com/2015/03/30/nothing-gold-can-stay-gold-nanoparticle-paper-retracted-for-figure-theft/.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

[7] Jones, N. (2024). How journals are fighting back against a wave of questionable images. *Nature News*.

[8] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196.

[9] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020a). Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114.

[10] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020b). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

[11] Liu, B., Zhu, Y., Song, K., and Elgammal, A. (2020). Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*.

[12] Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. (2018). Detection of gan-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 384–389. IEEE.

[13] Oransky, I. (2010). Journal of the american chemical society retracts gold nanoparticle paper. In *Journal of American Chemical Society*. https://retractionwatch.com/2010/12/17/journal-of-the-american-chemical-society-retracts-gold-nanoparticle-paper/.

[14] Oransky, I. (2013). Nano letters retracts chopstick nanorod paper questioned this week on chemistry blogs. https://retractionwatch.com/2013/08/16/nano-letters-retracts-chopstick-nanorod-paper-questioned-this-week-on-chemistry-blogs/.

[15] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[16] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

[17] Sauer, A., Schwarz, K., and Geiger, A. (2022). Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10.

[18] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

[19] Tan, C., Zhao, Y., Wei, S., Gu, G., and Wei, Y. (2023). Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114.

[20] Wang, S.-Y., Wang, O., Owens, A., Zhang, R., and Efros, A. A. (2019). Detecting photoshopped faces by scripting photoshop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10072–10081.

[21] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.