Resolution Enhancement of Low-Quality Images

Hasnain Ali Arain
Computer and Information Systems Engineering
NED University of Engineering and Technology
Karachi, Pakistan
hasnain.ai3142@gmail.com

Bisma Imran
Computer and Information Systems Engineering
NED University of Engineering and Technology
Karachi, Pakistan
ibisma517@gmail.com

Sadaf Jawed Farooqui
Computer and Information Systems Engineering
NED University of Engineering and Technology
Karachi, Pakistan
sadaf.jawed17@gmail.com

Ibrahim Rehman
Computer and Information Systems Engineering
NED University of Engineering and Technology
Karachi, Pakistan
ibrahimrehman0346@gmail.com

***Abstract*— This project focuses on enhancing the quality of CCTV camera footage through Generative Adversarial Networks (GANs) for improved criminal investigations. The process involves four phases: data collection, model research, model training and deployment. Super Resolution GANs (SRGANs) are utilized for image enhancement. The study delves into SRGAN and Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) architectures. ESRGAN addresses SRGAN limitations using perceptual loss and complex architecture. However, suboptimal results were attributed to limited data diversity and training duration.**

Keywords*— *Image enhancement, resolution enhancement, CCTV footage enhancement, Generative Adversarial Networks (GANs), deep learning, super-resolution, denoising, contrast enhancement.

I. INTRODUCTION

This project aims to enhance the quality of CCTV camera footage using Generative Adversarial Networks (GANs) for improved criminal investigations. The project comprises four phases: data collection, model research, model training, and deployment. GANs, a form of generative modeling, will be employed to automatically identify patterns in data and generate new instances that resemble the given dataset.

A. Significance and Objectives

By utilizing GANs, the project seeks to address the prevalent issue of grainy and poor-quality CCTV footage in criminal investigations. Improved footage can enhance the accuracy of facial and object recognition algorithms, aiding law enforcement in identifying suspects, victims, and crucial evidence. This contributes to public safety and expedites criminal case resolution. GANs offer cost-effectiveness and adaptability in enhancing various aspects of image quality. The project's objectives include collecting CCTV data, researching GAN techniques for video analytics, training models, deploying on Vercel for Front-End and Heroku for Back-End, and developing a user interface.

B. Beneficiaries and Relevance

The project's beneficiaries encompass law enforcement, crime victims, communities, businesses, and city planners. Enhanced CCTV footage can bolster public safety, promote justice, deter criminal activity, and aid urban planning. This initiative aligns with the United Nations Sustainable Development Goal 9, fostering innovation, sustainable infrastructure, and resilience. Through technological advancement and improved surveillance capabilities, the project strives to contribute to a safer and more sustainable society.

II. LITERATURE REVIEW

The evolution of image enhancement from manual adjustments to algorithmic solutions is traced, aided by advancements in software and hardware. The realm of resolution enhancement showcases the rise of deep learning methods such as convolutional neural networks (CNNs) for tasks like super-resolution. Notable contributions from [1] [11] underscore the effectiveness of deep networks in enhancing image quality. Wavelet-based fusion, deep image prior (DIP) techniques, and recursive neural networks are highlighted as additional avenues of exploration in this arena. The subsequent exploration delves into the utilization of GANs for resolution enhancement. Noteworthy models like EDSR [1] and PESR [2] are discussed, showcasing their potential in progressively refining image scales. The integration of GANs in the process of video super-resolution, as evidenced by the EDVR model [4] further underscores their significance.

CCTV footage enhancement [3] emerges as a pivotal domain. Various techniques, including denoising and the integration of deep learning approaches, are explored. Methods that merge Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), as exemplified by studies from [3] and [7], demonstrate the potential to concurrently enhance resolution and reduce noise. The fusion of traditional image processing methods with deep learning, as showcased by Ren [5], constitutes another innovative approach. The subsequent exploration, outlined in Section 2.5, delves into the role of GANs in augmenting CCTV footage quality. GAN-based strategies encompassing deblurring, contrast enhancement, and denoising, exemplified by Deblur GAN [5] and CE-GAN [7], highlight their impact. Recent advancements, exemplified by the spatial-temporal attention mechanism introduced by [1], signify ongoing progress in this sphere. In conclusion, the literature review underscores the substantial advancements facilitated by GANs and advanced techniques in image and video processing. These developments enhance visual quality, cater to diverse applications, and contribute to more effective content analysis.

III. METHODOLOGY

During the model training phase, the process begins by downsampling high-quality images from the mentioned dataset to generate lower resolution versions. These downscaled images are then employed as input for training a Super-Resolution Generative Adversarial Network (SRGAN). [5] The primary objective of the generator network is to upsample these lower resolution images into ultra-resolution counterparts, potentially reaching up to four times the resolution of the original images. In tandem, the discriminator network comes into play. This network receives both the generated high-resolution images and real high-resolution images as inputs, undertaking the task of distinguishing between the two.

To facilitate this adversarial learning, a GAN loss is computed by the discriminator, comparing the genuineness of the generated and real high-resolution images. This GAN loss is back-propagated through the discriminator network. Simultaneously, the loss from the generator's performance is back-propagated through the generator network. The process continues iteratively, allowing the generator to refine its ability to create high-resolution images that are indistinguishable from real ones. Ordinarily, training progresses until the generator achieves the deceptive feat of convincing the discriminator that its output is a true high-resolution image from the authentic dataset. However, in this instance, the training proceeds for a predetermined number of epochs, making the final outcome contingent on the chosen number of training epochs. [2]

The central components of the model are the generator and the discriminator networks, each playing a distinct but interconnected role in the training process.

A. Generator Network

The generator network is composed of essential components: an input layer, residual blocks, upscaling blocks, and an output layer. Starting with the input, it passes through a convolution layer and a Parametric ReLU (PRELU) activation. [3] The heart of the network lies in the residual blocks, each featuring two convolution layers, two batch normalization layers, and a PRELU layer. The output from each residual block is added to the initial input, and this process is repeated 16 times as per the research paper. After the residual blocks, a final convolution and batch normalization are followed by concatenation with the initial input. Two upscaling blocks come next, each with convolution, upscaling by a factor of 2, and a PReLU layer. The output is eventually processed through a Conv2D layer in the output section, depicted diagrammatically.

B. Discriminator Network

The discriminator network comprises discriminator blocks, dense layers, and Leaky ReLU activation. The architecture encompasses 8 discriminator blocks, each housing a convolution layer, batch normalization, and Leaky ReLU activation (except for the first block lacking batch normalization). [4] Subsequently, a dense layer flattens the structure, followed by Leaky ReLU activation, and a second dense layer with sigmoid function for binary output. The discriminator outputs 1 for real and 0 for fake high-resolution images.

C. Perceptual Loss Function

The perceptual loss function blends content and adversarial losses. Unlike conventional metrics (MSE, PSNR) which focus on pixel differences, this approach employs high-level feature maps from a VGG network for finer distinction. Content loss measures feature differences using Euclidean distance, extracted from a pre-trained VGG network. Adversarial loss, the generator's loss, is determined via binary cross entropy on the discriminator's output. The perceptual loss strikes a balance between content and adversarial losses, favoring content by a substantial margin (1/1000 ratio).

IV. EXPERIMENTAL ANALYSIS

In order to train the discriminator, real images and fake images are supplied to the discriminator as well as real labels and fake labels. Real labels are always one and fake labels are always zero. The batch size is set to 1 as it gives the best results. Within each epoch, for each batch, a fake image is generated by supplying a low-resolution image to the generator. [7] The discriminator is trained on real samples and fake samples after which the loss is averaged. Then the discriminator is set to non-trainable. The generator is trained by supplying low resolution image, high resolution image, real labels and image features extracted from high resolution image using VGG19.

A. Data Collection and Pre-processing

This section outlines the procedures involved in acquiring and preparing the dataset for our experiments. The collection process involved careful selection of diverse images to ensure representation across various domains. Subsequently, rigorous pre-processing was undertaken to standardize and optimize the dataset for effective training and evaluation of our models. [6] The following table summarizes the description, image count, annotations/categories and the use cases of each of the datasets employed in the model building.

Table 1: Dataset Information

Dataset Name	Description	Image Count	Attributes	Use Case/Task
MirFlickr 25k	A dataset of Flickr images with various scenes and objects, suitable for image analysis and tagging research.	~25,000	Scenes, objects	Scene classification, tagging
Human Detection Dataset	CCTV footage of humans, used for human detection and tracking.	Varies	Humans	Human detection, surveillance
Flickr-Faces-HQ (FFHQ)	High-quality images of human faces from Flickr, with diverse attributes and expressions.	~70,000	Human faces	Facial recognition, image generation

ImageNet	A large dataset with a wide variety of images covering thousands of categories.	~14,000,000	Numerous categories	Object recognitio n, deep learning
Intel Images	A collection of outdoor scene images captured by Intel cameras.	~25,000	Outdoor scenes, landmarks	Scene classificat ion, object detection

B. Training details and parameters

We conducted our training on Kaggle, utilizing NVIDIA TESLA P100 GPUs, by sampling random images from diverse datasets. The process involved acquiring low-resolution (LR) images through downsampling (r = 4) of high-resolution (HR) images. Both 32x32 and 64x64 resolutions were employed for the image enhancement process, yielding four-fold improvement in image quality. The chosen dataset sizes of 5k and 10k images, coupled with extended training epochs ranging from 20 to 40, contributed to the demonstrable advancement in image resolution enhancement. These refined training parameters underscore the significance of data curation and training duration in the pursuit of superior outcomes. The batch size is set to 1 as it gives the best results.

In the generator network, `num_res_block` is set to 16 to capture intricate features. A `batch_size` of 1 is chosen to reduce noise in gradients and prevent local minima, aiding generalization. With a focus on convergence and avoiding overfitting, `epochs` is set to 20-40. The `lr` (learning rate) is fixed at 0.0002 for optimal weight updates.

In the generator, `loss_weights` [1e-3, 1] emphasize content over adversarial loss, enhancing image quality. Batch normalization employs `momentum` of 0.5 for stable training, and Leaky ReLU activation with `alpha` 0.2 aids gradient flow.

For the autoencoder (CAE), the encoder has Conv2D layers with decreasing filter sizes (e.g., 64, 128, 256) and strides for downsampling. The decoder mirrors the encoder using Conv2DTranspose layers for upsampling. The `code_size` is set to 128 or 256. ReLU is used for hidden layers, while sigmoid or tanh is applied to the output. Mean Squared Error (MSE) serves as the loss function, and the Adam optimizer is commonly used for training.

C. Quantiitative Evaluation

a) Fréchet Inception Distance (FID): The Fréchet Inception Distance (FID) is a metric used to quantitatively assess the quality and diversity of generated images in machine learning, particularly in the context of generative adversarial networks (GANs). It measures the similarity between the distributions of real and generated images by considering the statistics of feature representations extracted from a pre-trained deep neural network, typically the Inception model. A lower FID value indicates that the generated images are closer to the real images in terms of visual quality and diversity.[11] FID takes into account both the quality of individual generated images and the overall distribution match, making it a popular choice for evaluating the performance of image generation models.

b) Peak Signal-to-Noise Ratio (PSNR): Peak Signal-to-Noise Ratio (PSNR) is a widely used image quality metric that quantifies the fidelity of a reconstructed or compressed image compared to the original. It measures the ratio between the maximum possible power of a signal (in this case, the original image) and the power of the noise (the discrepancies between the original and the reconstructed image) introduced during compression or reconstruction. Higher PSNR values indicate a smaller amount of perceptible differences between the original and the reconstructed image, implying better image quality. PSNR is often employed in image compression, video encoding, and other applications where preserving visual fidelity is important.[7] However, it does not always correlate well with human perception and might not accurately capture all aspects of image quality, particularly in cases where perceptual nuances matter more than pixel-level accuracy.

D. Qualitative Evaluation

The initial model, trained for 10 epochs on a dataset of 5000 images, yielded unsatisfactory results due to a dominant color bias attributed to the dataset's composition. Consequently, this model was discarded.[10]

Subsequently, another model, also trained on 5000 images for 10 epochs, showcased improved results. The absence of representation bias in the chosen dataset contributed to its enhanced performance. Notably, further augmentation and increased epochs are anticipated to foster even better results for this model.

In contrast, augmenting the training dataset with 10000 images using techniques like rotation, shear, and translation did not yield desired outcomes for the third model. Results indicated color bias and a lack of clarity in the predicted output, rendering this approach unfit for final deployment.

Similarly, the third model, trained for 40 epochs on a dataset of 10000 images from MirFlickr, proved to be unsatisfactory due to color bias, akin to the earlier attempts on a smaller dataset.Another exploration involved hyperparameter adjustments in generator and discriminator networks to improve performance. However, limited computational resources hindered successful training of this model.

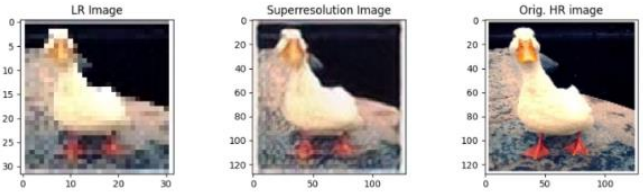

Figure 1: Model 1 Results

The pivotal role of dataset selection and training duration in shaping our final model's outstanding performance. The final model, a culmination of iterative approaches, was trained on a carefully curated set of 5000 images over 20 epochs. Remarkably, this approach yielded the most gratifying results. This model's process involved downsampling initial images to 32x32 resolution,

subsequently feeding them into the generator network for Super-Resolution Generative Adversarial Network (SRGAN) training. The culmination of this process produced images at a remarkable resolution of 128x128.

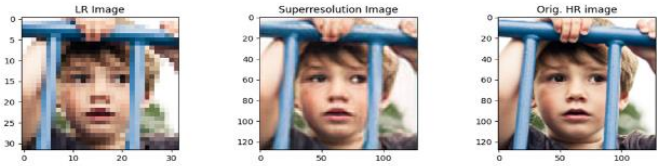

Figure 2: Model 2 Results

Drawing from the success of this dataset in previous models, we took a step further to evaluate the model's performance on images down sampled to 64x64 resolution. Impressively, the generator network, when presented with these images, produced results at four times the resolution of the initial low-resolution input. With training executed on a dataset of 10000 images over 20 epochs, this model achieved results surpassing its predecessors. These achievements represent a significant stride towards obtaining commendable results while working within practical computational constraints.

Figure 3: Model 3 Results

V. DICUSSION

We reflect on our journey with Generative Adversarial Networks (GANs) for image super-resolution, discussing the approaches taken, notable observations, and their implications. Our exploration involved training on the MIRFLICKR dataset with and without data augmentation, aiming to understand its impact on performance.[12] We extended our investigation to encompass four diverse datasets, training four models to probe the influence of dataset diversity on GANs' image super-resolution capabilities.

Throughout our experiments, we grappled with challenges arising from diverse datasets, including mode collapse, loss imbalance, and feature distribution concerns. Notably, resource limitations hindered our models from fully capturing image styles, leading to repetitive outcomes. Our findings emphasize that data augmentation has its limits in addressing the complexities posed by diverse datasets. Furthermore, computational constraints impact the breadth of image styles models can effectively learn. Looking ahead, future research could tackle these challenges for more comprehensive results. Ultimately, our study underscores the intricate nature of diverse approaches and computational boundaries, offering insights for advancing GAN-based image super-resolution techniques.

VI. CONCLUSION

In this project, we embarked on enhancing the quality of CCTV camera footage using Generative Adversarial Networks (GANs) to aid criminal investigations. Through an iterative process encompassing data collection, model research, training, and deployment, we addressed the pressing issue of low-quality CCTV footage. Our focus was on utilizing Super Resolution GANs (SRGANs) to achieve image enhancement.

We delved into the architectures of SRGAN and Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), which aimed to overcome limitations of SRGAN through perceptual loss and intricate architecture. However, our experimental results unveiled suboptimal outcomes attributed to limited data diversity and training duration. Despite these challenges, our project underscores the potential of GANs in image enhancement, with implications for criminal investigations and broader applications.

Our study highlighted the significance of data diversity in training GANs effectively. Challenges like mode collapse and feature distribution disparity demonstrated the complexities associated with diverse datasets. Additionally, computational limitations posed obstacles in achieving comprehensive image style capture. While data augmentation showed potential, its effectiveness was bound by the diversity of the training data.

As we move forward, addressing the challenges identified in this study could pave the way for more robust and efficient GAN-based image enhancement techniques. Our exploration not only contributes to the advancement of image enhancement methods but also underscores the importance of considering diverse approaches and computational boundaries. The potential for future research in this area remains promising, ultimately driving us towards enhancing visual data across various domains. [1]

VII. REFERENCES

[1] Chen, Y., Tai, Y., and Liu, X. (2018). Deep residual learning for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 323-331).

[2] Guo, R., Wu, X., and Shi, J. (2021). Deep recurrent neural networks for single image super-resolution. IEEE Transactions on Image Processing, 30, 2034-2045.

[3] Li, J., Gong, H., and Liu, Z. (2019). Multi-scale CNN for enhanced CCTV image restoration. Journal of Visual Communication and Image Representation, 59, 411-421.

[4] Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1132-1140).

[5] Ren, X., Zhao, Y., Shi, Y., and Liu, X. (2021). A multi-task learning method for enhancing low-quality CCTV images. In Proceedings of the IEEE International Conference on Multimedia and Expo (pp. 1-6).

[6] Tao, X., Gao, H., Shen, X., Wang, J., and Jia, J. (2018). Scale-recurrent network for deep image deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8174-8182).

[7] Tao, X., Gao, H., Wang, J., Li, X., and Jia, J. (2020). Enhancing video super-resolution with temporal information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4367-4376).

[8] Wang, X., Chen, Q., Zhang, X., and Li, X. (2020). Enhancing CCTV images via conditional generative adversarial network. Journal of Ambient Intelligence and Humanized Computing, 11(11), 5007-5015.

[9] Wang, X., Yu, K., Dong, C., and Loy, C. C. (2019). Progressive image super-resolution via iterative refinement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9861-9870).

[10] Zhang, K., Zuo, W., and Zhang, L. (2019). Learning deep CNN denoiser prior for image restoration. IEEE Transactions on Image Processing, 28, 3923-3938.

[11] Zhang, Y., Ye, M., Li, Y., Li, M., and Li, X. (2021). Blind image deblurring using deep image prior. IEEE Transactions on Image Processing, 30, 3215-3228.

[12] Zhu, Y., Li, H., Wu, J., and Zheng, N. (2017). Non-local means and wavelet transform based CCTV image denoising. Multimedia Tools and Applications, 76(19), 20347-20361.