





## Enhanced Image Super Resolution Using ResNet Generative Adversarial Networks

Shirina Samreen<sup>1\*</sup>, Vasantha Sandhya Venu<sup>2</sup>

<sup>1</sup> Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al Majmaah 15341, Saudi Arabia

<sup>2</sup> Department of Computer Science & Engineering, Vardhaman College of Engineering, JNTUH, Hyderabad 501218, India

Corresponding Author Email: [s.samreen@mu.edu.sa](mailto:s.samreen@mu.edu.sa)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.410432>

### ABSTRACT

**Received:** 2 November 2023

**Revised:** 26 March 2024

**Accepted:** 25 April 2024

**Available online:** 31 August 2024

#### **Keywords:**

*GAN, residual network, super resolution, ResNet-GAN*

Significant advancements in SISR have been achieved through the use of deeper CNNs, enhancing both speed and accuracy. However, a crucial challenge persists in restoring finer texturing details at higher up-scaling factors. Recent research efforts have focused on lowering Mean Square error of reconstruction to achieve high PSNR. However, these methods frequently fail to capture the high-frequency details necessary for preserving fidelity at higher resolutions. This paper introduces ResNet GAN, a GAN customized with residual learning for enhanced super resolution. Specifically, it excels in generating realistic images at a 4x upscaling factor. Notably, proposed perceptual loss function, encompassing both adversarial and content losses. A trained discriminator is employed to differentiate super-resolved and actual photos based on the computed adversarial loss. In contrast to traditional pixel space resemblance, the content loss relies on perceptual similarity. The results demonstrate that ResNet GAN with the proposed perceptual loss function outperforms Deep Residual Learning on Div2k. The framework exhibits superior metrics such as PSNR, SSIM, MOS, and MSE. By prioritizing perceptual details over pixel space on highly down-sampled images, the proposed approach successfully recovers photo-realistic features, addressing previous methods limitations. This advancement holds promising implications for applications requiring high-resolution image reconstruction.

## 1. INTRODUCTION

The acquired face pictures in surveillance systems are frequently quite tiny and vary between low-resolution and high-resolution photographs. As a result, face recognition ability suffers. This paper specifically addresses face recognition applications using surveillance cameras. Conventional approaches involve employing image pre-processing methods to mitigate lighting variations in raw video sequences captured by security cameras, thereby enhancing overall image quality. In the current computer era, optimal performance across a variety of applications hinges on the essential requirement for high-resolution images.

Image processing finds diverse applications in fields like medicine, defense, and more. Photographs are everywhere nowadays, more than ever, and thanks to improvements in digital technologies, it is relatively easy for anyone to generate a substantial quantity of images. Traditional image processing systems must deal with more difficult problems because of the abundance of images, as well as their adaptation to human eye requirement.

The availability of image datasets and benchmarks have contributed to the widespread applications of machine learning as well as image processing. The integration of machine learning into image processing is likely to bring significant advantages across various applications, enhancing the

comprehension of complex pictures. As the need for adaptation increases, the number of image processing algorithms incorporating learning components is expected to rise. However, an increase in adaptation is frequently associated with an increase in complexity, and any machine learning technique must be well trained in order to be successfully adapted to image processing difficulties. Handling substantial quantities of images requires the capability to deal with vast amounts of data, which is troublesome for most machine learning algorithms.

Super resolution (SR) fundamentally involves the complex task of generating a High-Resolution (HR) image from its Low-Resolution (LR) version. Super resolution has garnered significant interest in computer vision research because of its diverse range of applications. The SR task becomes particularly challenging when dealing with high up-scaling factors, as the reconstructed SR images often lack texture details.

Minimizing the MSE between the reconstructed HR image and the ground truth served as a common optimization objective for supervised SR algorithms. This approach was advantageous because reducing the MSE also improves the PSNR, which is a common metric for assessing and comparing SR techniques [1, 2]. However, both PSNR and MSE are based on pixel-wise differences and struggle to detect perceptually significant details, such as intricate textures. To address this

limitation, we propose using a GAN that incorporates a deep residual network (ResNet) with skip-connections as the primary optimization objective, moving away from reliance on MSE. By utilizing high-level feature maps and a discriminator, we create a novel perceptual loss function that helps generate results nearly indistinguishable from high-resolution reference images.

GANs (generative adversarial networks) [3-5] are computation networks which fit two neural networks against each other (hence the name "adversarial") to produce fresh, factitious samples of data which could pass for real information. Mostly, they were frequently used in photo, video as well as speech generation. The discriminator neural network model is utilized for categorizing whether the given image was real or the generated ones, and a generator model which utilizes inverse convolutional layers for transforming input to full two-dimensional picture with pixel values, are both required for developing a GAN for generation of pictures. Generative Adversarial Network GAN, is an architecture that trains the image generating framework through a picture discriminating framework using massive, unlabeled datasets. In some circumstances, the discriminating framework could be utilized as the starting point for creating classifier model. An SRGAN learns how to create upscaled pictures by combining the adversarial characteristics of GANs with deep neural networks (up to four times the resolution of the original). The resulting super resolution photographs are more accurate.

## 2. LITERATURE SURVEY

The researchers introduced an image SR method utilizing directional bicubic interpolation [6]. Bicubic interpolation is a common technique in image interpolation due to its low complexity and fairly good results. However, it typically operates only in horizontal and vertical directions, which can lead to issues such as blocking, blurring, and ringing along edges. Various methods for interpolating lost pixels were used depending upon local strengths as well as directions. The approach maintains sharp edges and details better than bicubic interpolation. The approach outperforms previous edge-governed interpolations with regard to subjective, objective measurements as well as it has a low computational complexity.

A lightweight hybrid dense with residual connections was used, evaluating the balance between network complexity and performance enhancement [7]. When compared to earlier approaches, the suggested methods can greatly lower both the memory requirement and the inference speed to hold parameters as well as intermediate feature maps while keeping equivalent picture quality. Even though increasingly deep neural network models outperform conventional approaches in terms of SR, however to the large network parameters and convolution operations of deeper and denser networks, it is hard to execute them on low computational complexity, limited power, and low-memory systems.

The researchers introduced a novel algorithm that combines traditional methods with deep learning, leveraging the autonomous feature extraction capability of deep learning to embark on more profound reconstructions of low-resolution images [8]. The utilization of deep learning techniques, in conjunction with traditional interpolation methods, facilitated the training and learning processes, resulting in the production of high-resolution reconstructed data. This proposed method

surpasses both standard interpolation algorithms and standalone deep learning approaches. Moreover, it excels in reconstructing intricate details, yielding sharper outlines and superior image quality. CDA's strong representational abilities and adaptable design ensure accurate matching of LR and HR image representations [9]. Its goal is to develop compact representation of input by keeping most crucial data.

The model forecasts the high frequency residuals in a fine way, and robust Charbonnier loss function providing close supervision [10]. The network architecture is generic; thus, it might be used to solve different image modification and synthesis issues. CNN designed with array of up-scaling filters has proved successful in displaying layer activations and creating semantic segmentations based on network high-level features [11].

It introduced RCAN with multiple residual groups interconnected by long skip connections for selective rescale of channel features while considering channel interdependencies [12].

Dense Nets inter layer connection sequence has been used for solving vanishing gradient problem, improve quality of feature, as well as promote feature reuse. Furthermore, it is suggested that DBPN be enhanced by incorporating dense connections in projection units, resulting in Dense DBPN. Dropout and Batch-Norm, that aren't ideal for SR because they lose range flexibility over features, were avoided, unlike in the original Dense Nets. Instead, before entering the projection unit, a  $1 \times 1$  convolution layer is applied for feature-pooling as well as dimensionality reduction. In D-DBPN, each part's input is concatenation of all preceding units' outputs. On huge scaling factors like 8X enlargement, the suggested network that surpasses state-of-art approaches such as SRCNN and DRRN [13].

Single scale and multi scale architectures were designed for reconstructing single SR and several scales of HR pictures in a model respectively [14]. Since Batch-Norm layers uses similar extent of memory as previous convolutional layers, GPU memory usage is likewise lowered. The proposed residual blocks are used to build a baseline (single scale) model. Although topology is similar to that of SRResNet, the model lacks ReLu activation external to the residual block. The multiscale representation adeptly manages various SR scales within a unified framework.

When interpolation methods are used to improve image resolution, they cause severe losses in their HF components. This may be seen in the smoothing that occurs as a result of the interpolation technique. The suggested image enhancement method aims to deliver higher resolution than existing SR-DWT and DASR. It forms necessary to preserve the borders and fine characteristics of an image in order to improve its resolution. The main difference among designed method and other standard algorithms is better edge and fine feature preservation, i.e., producing a sharp optimized image through intermediary stage in the HF sub-band interpolation procedure, and able to perform the disparity among the LL and the LR pictures [15].

To generate a high-quality photo, the researchers focused on developing and refining perceptual loss functions derived from extensive features of pre-trained networks [16]. By utilizing these loss functions, they trained forward networks for various image enhancement tasks, effectively merging the strengths of both methods. Their approach showcased results in photo style transfer, where a feed-forward network was trained to address enhancement challenges in real-time.

The researchers introduced SinGAN, a generative model capable of being trained unconditionally on a single natural image [17]. This model is designed to learn the internal distribution of patches within an image, allowing it to generate high-quality, diverse samples that maintain similar visual content. SinGAN consists of a pyramid of fully convolutional GANs, each responsible for capturing the patch distribution at different scales of the image. The authors employed a compact hourglass-shaped CNN architecture to accelerate and improve super-resolution (SR), enhancing the performance of the existing SRCNN model [18].

The SRCNN architecture was redesigned in three key aspects. First, a deconvolution layer was added at the end of the network to allow for direct learning of mappings from the low-resolution image to the high-resolution image. Second, the mapping layer was restructured by initially reducing the input feature dimension and then expanding it afterward. Finally, they adopted smaller filter sizes and increased the number of mapping layers in the third step.

The researchers introduced PULSE, which is a latent oriented up-sampling super-resolution method that generates high-resolution, realistic images by exploring the HR natural image manifold [19]. This self-supervised approach is not restricted to a single degrading operator used during training and focuses on creating images that downscale accurately to the original low-resolution input.

Even before training, the structure of the generator network could capture numerous low-level image details. In inverse tasks such as denoising, super-resolution, and inpainting, a randomly initialized neural network effectively serves as a prior, achieving remarkable results. The identical prior can even be utilized to analyze deep neural representations also restore images from flash-no-flash input pairs using the identical prior. Inductive bias is encapsulated within traditional generator network architectures [20].

The researchers proposed a self-supervised method with temporal self-containment, crucial for achieving temporally clear solutions without sacrificing spatial textures [21]. They introduced the Ping-Pong loss to enhance temporal consistency over time, preventing recurrent networks from accumulating artifacts while preserving detailed features. Additionally, they introduced a comprehensive set of metrics to objectively evaluate the correctness and perceptual quality of temporal evolution.

A novel architecture aims to maintain spatially accurate high-resolution representations while extracting significant contextual information from low-resolution representations. This approach features a multi-scale residual block with concurrent multiresolution convolution flows for multi-scale feature retrieval, cross-flow information exchange, spatial and channel attention mechanisms for context acquisition, and attention-based multi-scale feature aggregation. In essence, it learns an extensive set of features across various scales while preserving high-resolution spatial details [22].

The researchers demonstrated that models with wider features before ReLU activation significantly enhanced SISR performance without increasing parameters or computational costs [23]. A novel projection-based method to incorporate conditional information into the GAN discriminator without altering its functionality in the probabilistic model was proposed [24]. The researchers introduced EDVR, a Video Restoration framework using Enhanced Deformable Convolutions [25]. They developed the PCD alignment module, which handles large motions through pyramid,

cascading, and coarse-to-fine deformable convolutions for precise feature alignment.

### 3. EXISTING METHODS FOR IMAGE SUPER-RESOLUTION

#### 3.1 SISR

The SISR [26] is a technique focused on making a high-resolution (HR) image from a single low-resolution (LR) input. One approach within SISR, known as sample-dependent Super-Resolution, relies on example-based schemes. This technique involves constructing a dictionary that captures the relationships among LR and HR image patches. During the construction of this dictionary, irregular mappings are learned from LR to HR images. These mappings are crucial as they facilitate the generation of high-quality HR images during the super-resolution phase.

The process of learning irregular mappings involves analyzing a dataset of LR along with HR image pairs. Machine learning algorithms, such as neural networks, are often employed to learn the complex relationships between LR and HR images. These algorithms adaptively adjust the mapping functions to minimize the difference between the generated HR images and ground truth HR images. Through this iterative learning process, the model gradually improves its ability to generate HR images that closely resemble the ground truth. Irregular mapping is crucial for capturing intricate details present in HR images not directly perceptible in LR images. By learning these mappings, SISR methods can infer missing high-frequency information and spatial correlations between LR and HR images, resulting in HR images with enhanced resolution, sharpness, and fidelity to the original scene. Essentially, irregular mapping bridges the resolution gap between LR and HR images, advancing SISR techniques in improving image quality and visual perception.

In a typical SISR workflow, a conventional interpolation technique like Bicubic Interpolation is initially used to upscale the LR input image. Subsequently, this interpolated image is divided into smaller patches. Each patch is then compared against a dictionary in the SR processing unit, which contains pre-trained LR-HR patch pairs. These LR patches are matched with their corresponding HR patches in the dictionary. Additionally, missing high-frequency details in the LR image are estimated based on the overlaying of LR patches with appropriate HR patches.

However, SISR faces several challenges. Recovering the content of a high-frequency image from a low-resolution image is inherently difficult, often resulting in HR images with poor quality due to the lack of high-frequency information. Additionally, a single LR image can lead to multiple potential HR images, posing another challenge in the SISR process.

#### 3.2 CNN

There have been several sorts of Example-based Super-Resolution techniques created so far. The CNN-dependent SR has received significant attention in recent research due to its remarkable achievements. One notable approach within this category is the SRCNN, which consists of three-layer networks designed to ensure convergence by adjusting the learning rate. Additionally, Super-Resolution methods based on sparse coding have emerged over the past few decades,

including A+ and ScSR. The input LR blotch was calculated through a linear allegation of many basements contained in dictionary, HR blotches are substituted, as well as the picture is overlaid to generate a high-resolution image in these approaches.

The capabilities of Convolutional Neural Networks (CNNs) are already being harnessed across various applications, ranging from Facebook photo tagging to Amazon product recommendations, healthcare imaging, and autonomous vehicles. CNNs' success stems from their ability to require minimal preprocessing and effectively analyze 2D images using filters that other algorithms cannot.

The depth of the filter matches the depth of the input. The filter convolves across the input image, shifting by one unit at a time. Each convolution yields a single value obtained by summing the products of corresponding elements. This process is repeated across the entire image, resulting in a smaller matrix compared to the original input. The final output is represented by the feature map or activation map. Convolution serves various purposes such as edge detection, sharpening, and blurring, achieved by applying multiple filters to an image. The task only requires specifying characteristics like filter size, number, and network design.

The MSSR approach amalgamates A+, ScSR, and SRCNN, presenting a novel fusion of sparse coding and CNN-dependent methodologies to yield visually appealing images. This methodology operates on the premise that Sample-dependent Super-Resolution acts as a filter enhancing image quality solely through the SR processing unit, disregarding the enlargement process. The underlying principle relies on the uniform dimension conversion performed by the dictionary from LR to HR. Post standard super-resolution processing, MSSR bypasses the utilization of Bicubic interpolation, integrating SR processing units in series instead. A+ or ScSR was utilized in primary SR step, while SRCNN without bicubic interpolation is utilized in second SR step. After then, conventional super-resolution is carried out, followed by super-resolution without an expansion stage. The evaluations entailed manipulating the number of input images through a temporal analysis scheme, revealing that MSSR's performance is influenced by factors such as Signal-to-Noise Ratio (SNR). Although MSSR did not enhance much reconstruction accuracy, it yielded visually appealing images by combining sparse coding and CNN-dependent methodologies. This was achieved by bypassing Bicubic interpolation and integrating SR processing units, resulting in noticeably enhanced visual image quality.

The MSSR approach extends its capabilities with the introduction of MSSR-2, which integrates additional rotation and inversion processing procedures alongside a restoration step into the secondary stage of super-resolution (SR). Drawing insights from prior studies on MSSR, MSSR-2 aims to further enhance the reconstruction accuracy of high-resolution images.

Specifically, MSSR-2 conducts SR processing on groups of images generated through rotations or inversions of the input image. Subsequently, these images are restored to their original orientation through averaging. By subjecting the images to SR processing at different rotations and inversions and then averaging them, MSSR-2 aims to reduce errors in blotch selection and mapping within the dictionary, thereby enhancing reconstruction accuracy. This meticulous approach results in MSSR-2 surpassing MSSR in terms of reconstruction accuracy, marking a significant advancement in

the field of super-resolution techniques.

Most modern super-resolution algorithms use neural networks or patch-based methods to learn the mapping between low-resolution and high-resolution image spaces. Specifically, CNNs are extensively used for this purpose in computer vision. A notable example of this approach is Convolutional Super-Resolution, with the pioneering method being the Neural Network-based Super-Resolution (SRCNN) [27]. SRCNN has surpassed previous super-resolution techniques in performance. The SRCNN network learns the mapping through three main operations: non-linear mapping, patch extraction, and representation reconstruction.

Table 1 below offers a comparative analysis between SISR and CNN-based Super-Resolution techniques. SISR primarily aims at generating high-resolution (HR) images from single low-resolution (LR) inputs using example-based methods, while CNN-based approaches utilize Convolutional Neural Networks to learn the mapping between LR and HR image spaces. This comparison highlights the methodologies, learning approaches, advantages, and limitations of each technique, providing insights into their respective strengths and weaknesses in the context of super-resolution tasks [28].

**Table 1.** Comparison of SISR and CNN-based super-resolution techniques

Aspect	SISR	CNN-Based Super-Resolution
Focus	Generation of HR images from single LR input	Learning mapping of LR to HR image space
Methodology	Example-based, sample-dependent methods	CNNs, often with sparse coding
Learning Approach	Machine learning algorithms (e.g., NNs)	Convolutional Neural Networks (CNNs)
Advantages	Captures intricate details in HR images and improves visual image quality	Efficient analysis of 2D images and obtains promising results in generation of HR images
Limitations	Difficulty in accurately reconstructing HR images and struggles with recovering high-frequency information from LR images	Face challenges in accurately reconstructing fine details and textures, especially in complex scenes
Typical Techniques	Bicubic interpolation, example-based methods	SRCNN, A+, ScSR

#### 4. PROPOSED METHOD

ResNet-GAN is a technique used for super-resolution image generation, transforming low-resolution (LR) images into high-resolution (HR) ones. It leverages ResNet architecture for learning residual functions and generative adversarial networks (GANs) for generating realistic HR images. This approach enhances the quality and resolution of images, making them suitable for various applications requiring high-quality visuals.

##### 4.1 Generative adversarial network with residual learning

A GAN comprises a Generator and a Discriminator. The

Generator aims to craft realistic fake images, while the Discriminator discerns these fakes. This adversarial strategy is well-suited for generating high-quality SR images, as the iterative interaction between the Generator and Discriminator enhances results. GANs offer a robust framework for generating natural-looking, high-quality images. The GAN process guides reconstructions towards regions likely to contain photorealistic images, closely aligning them with the original image manifold. This paper introduces the first deep-ResNet framework that employs GANs to implement a perceptual loss function for photorealistic single image SISR. The Res-Net GAN, tailored for this purpose, integrates a novel perceptual loss by replacing the traditional MSE-based content loss with one computed using VGG network feature maps, which are more attuned to changes in pixel space.

#### 4.1.1 Generator

The Generator part in the GAN architecture here consists of Residual Network (ResNet) model for image processing. ResNet, a prominent neural network architecture in deep learning, is widely employed across various image processing tasks. As subsequent winning architectures incorporate more layers to reduce error rates in deep neural networks, they encounter a common challenge known as the vanishing/exploding gradient problem. This issue arises when

the gradient becomes either excessively small or large, leading to increased training and test error rates with an increasing number of layers. To address the vanishing/exploding gradient problem, ResNet introduces the concept of residual learning, which involves utilizing skip connections in the network architecture. These skip connections bypass certain layers during training and connect directly to the output.

Residual learning, implemented in ResNet, involves learning residual functions instead of direct mappings. Within residual learning, there are two types: global residual learning and local residual learning. In the suggested model, local residual learning is applied, which aids training by reducing task complexity and enhancing learning rates in deep learning. Shortcut connections between various layers within the network, of varying depths, are utilized in conjunction with element-wise addition to facilitate local residual learning.

#### 4.1.2 Discriminator

The discriminator in a GAN represents a classifier which distinguishes between original data and data fed by the generator. The selection of the network architecture can be based upon the type of data. The generator generates fake data instances, and these are used by the discriminator as negative examples during training. Sigmoid function to classify the image whether they are similar or not.

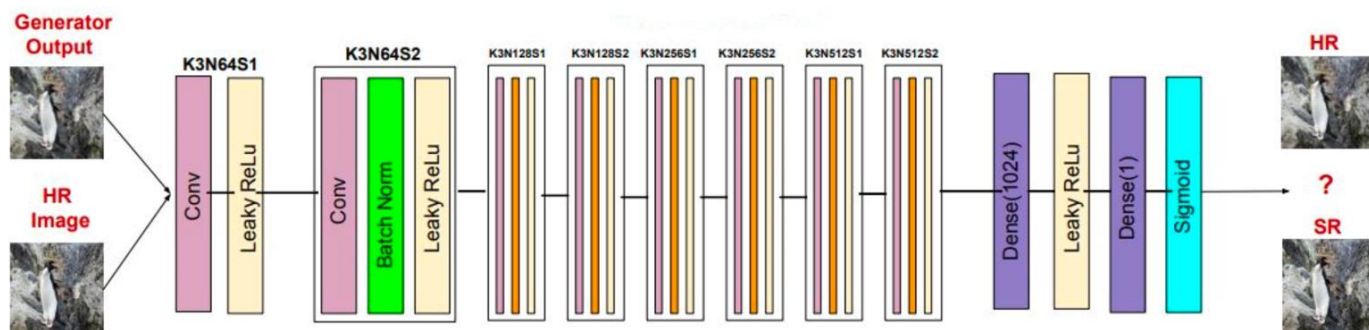


Figure 1. Discriminator architecture [29]

Because the purpose of discriminative models is to detect fraudulent data, the discriminative neural network is trained to reduce the final classification error. It learns to distinguish between the various classes by examining both genuine and fake samples generated by the generator and attempting to determine which are real and which are fake. Figure 1 shows how the discriminator of GAN architecture differentiates between HR and SR images.

The output of the discriminator, a binary value indicating real or fake, guides the generator in GANs to produce high-resolution images from low-resolution inputs. When the discriminator identifies a generated image as realistic, the generator refines its parameters to produce even higher-quality images. Conversely, if the discriminator detects flaws, the generator adjusts to enhance image quality. This iterative process enables the generation of more realistic high-resolution images from low-resolution inputs.

## 4.2 ResNet GAN model

Res-Net GAN is a state-of-the-art method for producing photo-realistic super-resolution(SR) images with high upscaling factors(x4), validated by exhaustive MOS analysis. The goal of SISR is to generate a high-resolution(HR) image

from a low-resolution(LR) input. During training, HR images are available, and LR equivalents are created by applying a Gaussian filter and downsampling.

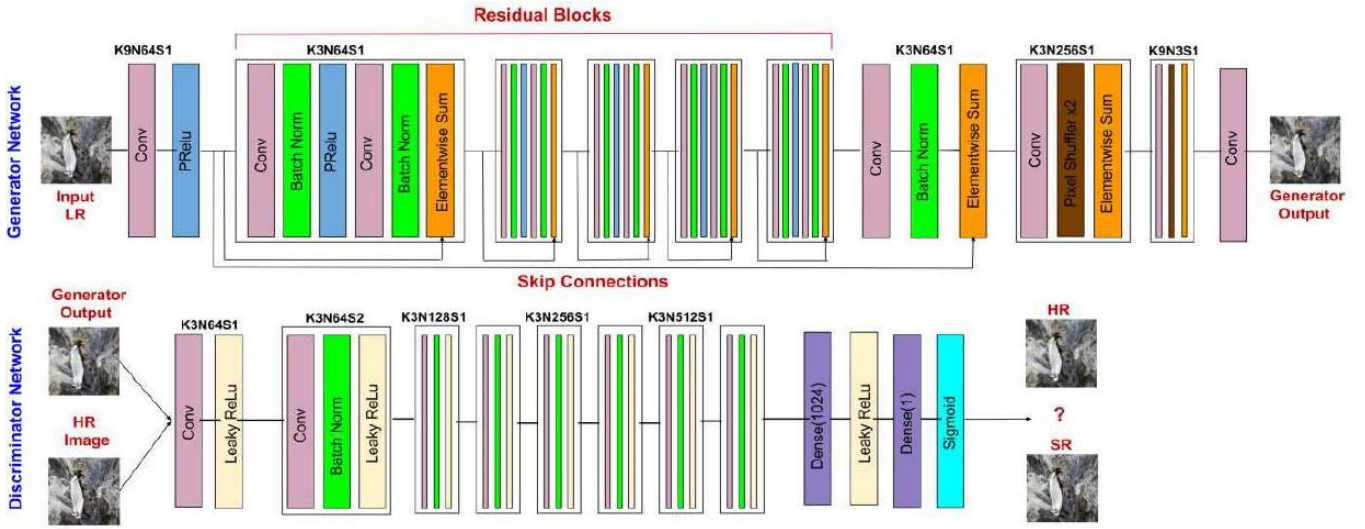
The primary objective is to train a generator function (G) that estimates the HR counterpart of a given LR image using a feedforward CNN. The generator network consists of multiple identical residual blocks, each with convolutional layers, batch normalization, and parametric ReLu. The input LR image is processed through these residual blocks, followed by pixel shufflers to increase the resolution, and a final convolutional layer.

The discriminator network is trained to distinguish real HR photos from generated SR images. It uses leaky ReLu activation and strided convolutions to reduce picture resolution, similar to the VGG network, and includes dense layers for classification.

The performance of the generator network relies on a perceptual loss function, combining Mean Square Error (MSE) and VGG feature maps. This approach encourages the generator to produce images that align closely with the original image manifold, aiming to deceive the discriminator model.

The architecture aims to generate high-quality SR images by promoting perceptually superior solutions found within the natural image subspace, as shown in Figure 2.





**Figure 2.** Schematic diagram of Res-Net GAN [29]

First, the dataset requiring higher resolution images is collected. Use the generator function to process them. This generator function boosts the image's resolution. The image is then sent to the discriminator function. The discriminator determines whether the image is genuine or not and returns a binary result. Return the picture to the generator function if the binary value is 0. Generator boosts the resolution once more and sends it to the discriminator, it checks the image once more. This procedure is repeated until a high-resolution image is obtained. Return the image if the discriminator's output is 1.

#### 4.2.1 Training details

The NVIDIA Tesla M40 GPU was utilized to train the networks using a random sample SR image set from Div2k [30]. These images differ from those used in testing. Low-resolution (LR) images were created by downsampling high-resolution (HR) images (BGR,  $C = 3$ ) with a bicubic kernel and a downsampling ratio of  $r = 4$ . Each mini-batch involved randomly cropping sixteen  $96 \times 96$  HR sub-images from various training photos. As the generator model is entirely convolutional, it can be applied to images of any size. LR input images were scaled to  $[0, 1]$ , while HR images were scaled to  $[1, 1]$ . MSE loss was computed on images with an intensity range of  $[1, 1]$ . The rescaling of VGG feature maps by a factor of 1 to 12.75 aims to align the scale of VGG losses with the scale of Mean Squared Error (MSE) loss. This rescaling factor ensures that the magnitudes of the losses are comparable, allowing for more balanced optimization during training. By using a rescaling factor of 0.006, this alignment is achieved, ensuring that both types of losses contribute effectively to the overall optimization process without one dominating over the other. Utilizing Adam with  $\beta = 0.9$  for optimization, provides a balance between adaptability and momentum making it effective for optimizing the training of deep neural networks when used in super-resolution tasks. SRResNet networks were trained with a learning rate of  $10^{-4}$  and  $10^{-6}$  update iterations. To avoid undesirable local optima, when training the GAN, MSE-based SRResNet was used as initialization for the generator. Various SRGAN variants were trained using  $10^5$  update iterations and learning rates of  $10^{-4}$  and  $10^{-5}$ . Generator and discriminator networks were alternately updated with  $k = 1$ . Remaining blocks in the generator network were identical ( $B = 16$ ). Batch normalization updates were disabled during testing to ensure deterministic results dependent only on the input.

### 4.3 Loss functions

It's a way of determining how well your algorithm models the data. At its most basic level, a loss function is a measurement of how well your prediction model predicts the expected outcome (or value). The learning problem is turned into an optimization problem, a loss function is established, and the algorithm is tuned to minimize the loss function. The difference between the expected output and the actual output of the machine learning model is computed which facilitates the gradient computation. Subsequently, the weights can be updated from the loss function. The cost is calculated as the average of all losses. There are three types of loss functions, they are: Perceptual loss, Content loss and Adversarial loss.

#### 4.3.1 Perceptual loss

Below, the perceptual loss function utilized is elaborated upon. The performance of our generator network relies on super-resolution (SR). While Mean Squared Error (MSE) is often utilized for low-resolution to high-resolution (I-SR) tasks, it lacks the ability to evaluate solutions based on perceptually significant qualities. To address this, we introduce a loss function that considers perceptual aspects. This comprises a weighted sum of content loss (I-SR-X) and an adversarial loss component to compute the perceptual loss.

In particular, our focus lies on image transformation problems, where an output image is generated from a modified input image. Recent methodologies tackling such challenges typically involve training feedforward convolutional neural networks using a loss measured in pixels per pixel, comparing the output to the ground-truth original images. We propose the utilization of perceptual loss functions for training feedforward networks designed for image transformation tasks. This approach amalgamates the benefits of both pixel-wise comparison and perceptual evaluation, potentially enhancing efficiency. Our feedforward network is trained to address optimization problems in real-time, demonstrated notably in the context of visual style transfer within this paper. Comparing our approach with optimization-based methods reveals qualitatively equivalent outcomes, achieved at a speed three orders of magnitude faster than the latter. As an additional experiment, we explore single-image super-resolution, substituting perceptual loss for per-pixel loss, resulting in visually pleasing results.

#### 4.3.2 Content loss

It is the prevailing optimization target in image super-resolution (SR), widely employed by state-of-the-art approaches. Nonetheless, solutions stemming from mean squared error (MSE) optimization often yield high peak signal-to-noise ratio (PSNR) but lack high-frequency content, leading to visually displeasing outcomes characterized by overly smooth textures. Instead of relying solely on pixel-wise losses, we opt for a loss function that prioritizes perceptual similarity, thereby mitigating the issue of excessively smooth textures and to measure how well the high-resolution output image preserves the structural content of the low-resolution input image.

#### 4.3.3 Adversarial loss

A continuously trained discriminator network defines the adversarial loss. It's a binary classifier that distinguishes between real-world data and data generated by a generative network. We add the generative block of our GAN algorithm to the perceptual loss function in addition to the losses from content discussed so far. By attempting to trick the discriminator network, this encourages the network to favor solutions which are based on a variety of natural images. The loss from generative function I-SR-Gen is calculated using the discriminator DD (GG (I-LR)) probabilities across all training samples.

#### 4.4 ResNetGAN for SR algorithm

```
Take image data set to enhance resolution.
Pass through Generator function.
def Gen():
    // code to increase resolution of the input image.
    //Return this to discriminator function.
Take input to discriminator function to check whether the
image is real or fake.
def Dis():
    // code to check real or fake.
    // if it is fake return that to generator to improve
quality.
    // if it is real, return to the output as SR images.
```

First take the data set for which you want to get the super resolution images. Pass them through generator function. This generator function increases the resolution of the image. Then it sends that image to discriminator function. Discriminator checks whether the image is real or fake and generates a binary value. If the binary value is 0 return that image to generator function. Generator again increases the resolution and sends it to discriminator and again discriminator checks the image.

This process continues until you get a high-resolution image. If the output of discriminator is 1 then return the image.

### 5. RESULT DISCUSSION

The input to the ResNet GAN network is a Low-Resolution image of size  $96 \times 96$  where the output image size turns out into  $4 \times$  up sampled one i.e.,  $384 \times 384$  from Div2K [30]. For a better observation, here we show the Low-Resolution image by up sampling it. The Res-Net GAN output is being compared with the ResNet without GAN and comparison is being made on 4 different parameters i.e., PSNR (Peak Signal to Noise Ratio), MSE (Mean Square Error), SSIM (Structure Similarity) as well as MOS (Mean Opinion Score) which are depicted in Figures 3-7.

Following are the definitions associated with the parameters used for result analysis:

**Peak Signal to Noise Ratio (PSNR):** It is computed as a fraction of the highest power of a signal and the power of noise affecting the accuracy of its representation. The metric is related to measuring the quality of transmission. In image processing applications, an individual pixel can be considered as a signal with 8-bit RGB values.

$$\text{PSNR} = 10 \cdot \log_{10} \frac{\text{MAX\_PIX}}{\text{MSE}} \quad (1)$$

In Eq. (1), MAX\_PIX is the highest value that can be given for a pixel and MSE is the Mean Squared error between the high-resolution image and the super-resolved image. As the pixel values have a wide range, we use a logarithmic scale.

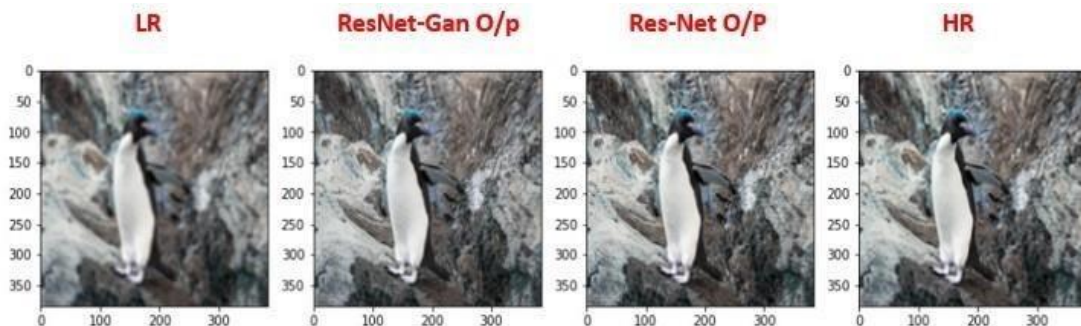
**Structure Similarity (SSIM):** This metric is analogous to human visual system (HVS color model) which works upon three parameters namely: correlation, luminance distortion, and contrast distortion. It is computed on various windows of the image rather than a comparison on the basis of pixel-by-pixel.

**Mean Opinion Score (MOS):** It represents a metric related to the perceived quality of image. It can be computed through observers' opinions about rating the images quality on a specific scale.

**Mean Square Error (MSE):** It represents the cumulative squared error between the super-resolution image (SR) and the high-resolution image (HR) wherein each pixel in SR image is compared against the opposing pixel in HR image.

Using NumPy package in python, we calculated MSE and from skimage metrics package, we had estimated SSIM. PSNR is formulated from MSE as follows:

$$\text{PSNR} = 20 \cdot \log_{10}(255.0 / \text{MSE}) \quad (2)$$

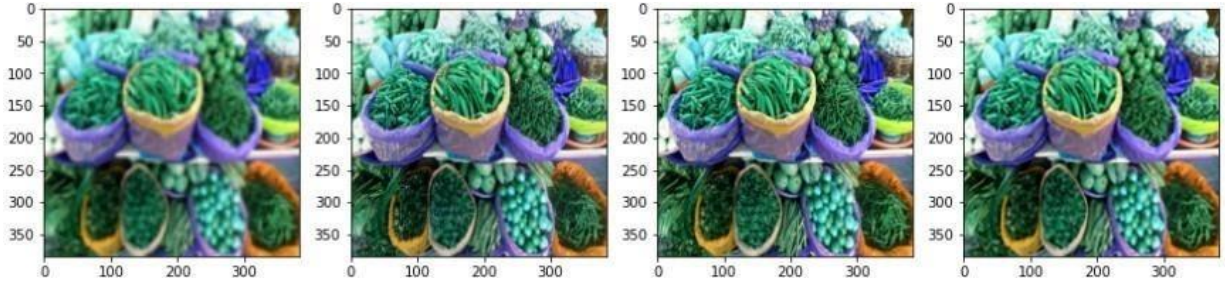


**Figure 3.** Results for input image 1 [29]

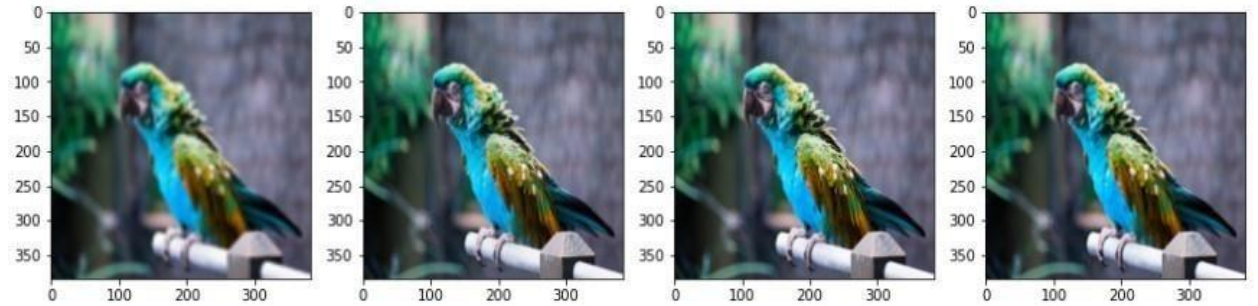




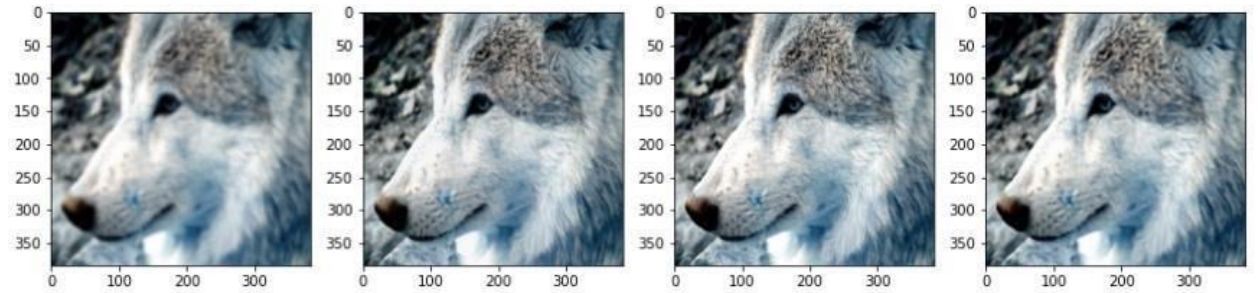
**Figure 4.** Results for input image 2 [29]



**Figure 5.** Results for input image 3 [29]



**Figure 6.** Results for input image 4 [29]



**Figure 7.** Results for input image 5 [29]

ResNet and ResNet-GAN generated photos performance is analyzed in terms four SR metrics, whose values are detailed in Table 2. It is clearly observed from Table 2, that ResNet GAN is more superior to ResNet without GAN model. The PSNR and MSE of HR images are represented as '-' when comparing the HR image against itself because the MSE becomes zero in this case, leading to a division by zero in the PSNR formula and resulting in an infinite value. So, it is here denoted by '-'. The ResNet Generative Adversarial Network (GAN) model surpasses the standard ResNet due to its use of adversarial training and perceptual loss functions. Adversarial training enables the generator to produce images indistinguishable from real data, while perceptual loss functions optimize for high-level features, resulting in visually appealing and realistic outputs with fine details. Mean squared

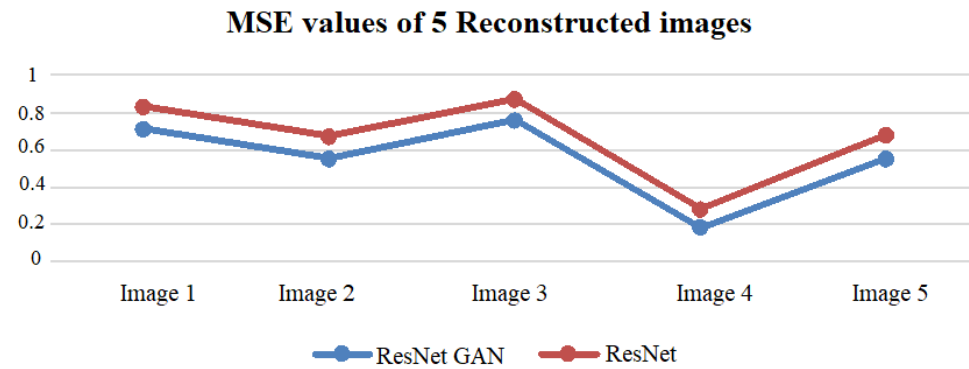
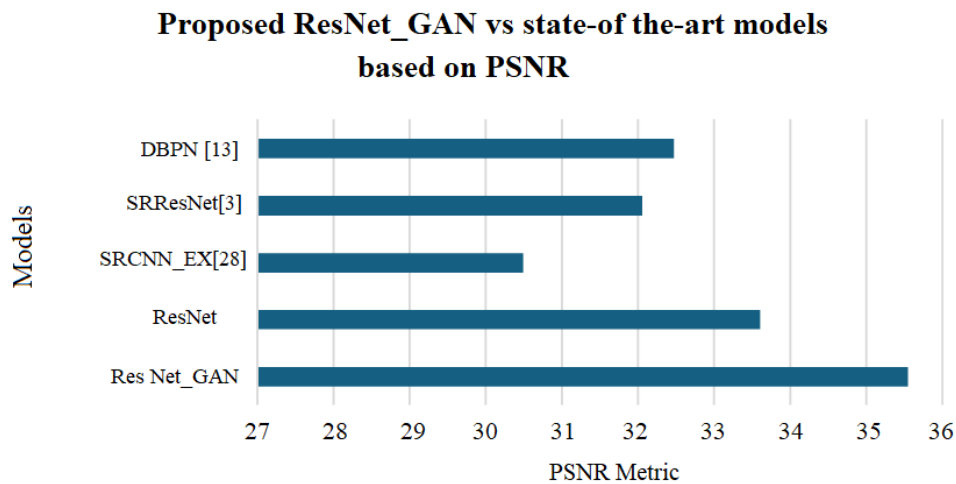
error (MSE) of reconstructed photographs of the ResNet GAN is lower when compared to ResNet, which is depicted in Figure 8. This indicates that reconstructed ones are closer to realistic photographs while maintaining low MSE values.

Figure 9 and Figure 10 illustrate a comparison between our proposed ResNet\_GAN model and other state-of-the-art models developed using the Div2K dataset, utilizing PSNR and SSIM metrics. The results reveal that our model outperforms the other models. Limitations of the proposed model includes, performance evaluation is constrained to images exclusively from the Div2K subset, thereby lacking validation on diverse datasets. It overlooks an examination of computational efficiency, which is essential aspect for assessing the model's real-world applicability.



**Table 2.** Performance analysis of five sample images based on ResNet GAN and ResNet models

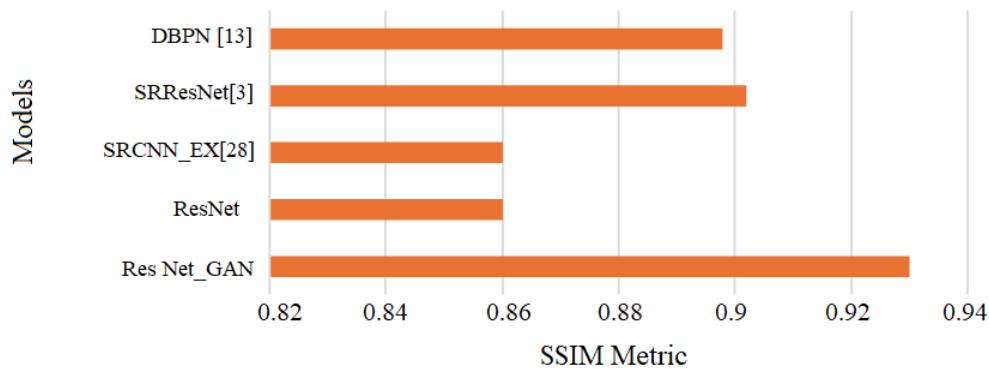
Image 1				Image 2		
Metrics	ResNet GAN	ResNet	HR	ResNet GAN	ResNet	HR
PSNR	29.63	28.94	-	30.74	29.88	-
SSIM	0.68	0.52	1	0.77	0.69	1
MOS	3.56	3.24	4.21	3.68	3.08	4.23
MSE	0.71	0.83	-	0.55	0.67	-
Image 3				Image 4		
Metrics	ResNet GAN	ResNet	HR	ResNet GAN	ResNet	HR
PSNR	29.33	28.75	-	35.54	33.61	-
SSIM	0.71	0.60	1	0.93	0.86	1
MOS	3.28	2.99	4.15	3.25	2.98	4.38
MSE	0.76	0.87	-	0.18	0.28	-
Image 5						
Metrics	ResNet GAN	ResNet	HR			
PSNR	30.77	29.82	-			
SSIM	0.75	0.65	1			
MOS	3.49	3.19	4.23			
MSE	0.55	0.68	-			

**Figure 8.** Performance of ResNet-GAN and ResNet in terms of MSE**Figure 9.** Performance of ResNet-GAN and state-of-the art models in terms of PSNR

Initially, we used a Gaussian filter with a state-of-the-art model architecture similar to the one [29], but found that Gaussian filtering blurred fine details and edges, producing less realistic and effective training data for super-resolution models. We then switched to bicubic filtering, which better preserved image details and edge sharpness, making it more

suitable for high-quality super-resolution. Additionally, we optimized the model by fine-tuning training configurations such as rescaling factors, mini-batch cropping, learning rates, and Adam beta values. This resulted in consistent improvement across all four metrics.

## Proposed ResNet GAN vs state-of-the-art models based on SSIM



**Figure 10.** Performance of ResNet-GAN and state-of-the art models in terms of SSIM

## 6. CONCLUSIONS AND FUTURE WORK

The Res-Net Generative Adversarial Network (GAN) has demonstrated highly effective in enhancing low-resolution images, establishing its superiority over traditional ResNet through comprehensive comparisons across four different parameters based on the perceptual quality of super-resolved images. The study focuses on the perceptual quality of super-resolved images over computational efficiency. The proposed approach plays a vital role across diverse applications in computer vision field. It can enhance the resolution of medical scans for precise diagnosis, improve satellite imagery for environmental monitoring, aid surveillance for better object identification, and enhance digital media content for superior visual experiences.

Furthermore, our study detailed SRResNet, achieving state-of-the-art performance in image super-resolution based on PSNR evaluation. We introduced SRGAN, combining content loss and adversarial loss for improved realism, particularly evident in large upscaling factors through MOS testing. We emphasized the limitations of PSNR and SSIM metrics in capturing perceptual image quality. While shorter networks offer efficient alternatives with minimal degradation, deeper designs benefit performance, influenced by ResNet architecture. Deeper networks enhance SRResNet performance, though longer training times are required. However, deeper SRGAN variations face challenges due to high-frequency aberrations in data. The optimal loss function varies based on application needs. Achieving perceptually convincing image reconstructions is a significant challenge for future research, involving the development of algorithms for real-time processing in fields like video streaming and surgery, handling dynamic environments in surveillance and autonomous vehicles, addressing limited annotated data in specialized fields like medical imaging, mitigating noise and artifacts in satellite imaging and microscopy, incorporating semantic understanding for scene understanding and medical diagnostics, and ensuring adaptability to diverse domains such as urban environments and industrial settings. Designing content loss functions that take into account the spatial content of the image while maintaining uniformity to changes in the pixel space will likely enhance the photorealistic results of

images even further in the future by employing variants of CNNs, along with regularization techniques.

## ACKNOWLEDGMENT

The authors would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project Number R-2024-1074.

## CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest to report regarding the present study.

## REFERENCES

- [1] Kim, J., Lee, J.K., Lee, K.M. (2016). Accurate image super-resolution using very deep convolutional networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 1646-1654. <https://doi.org/10.1109/CVPR.2016.182>
- [2] Kim, J., Lee, J.K., Lee, K.M. (2016). Deeply-recursive convolutional network for image super-resolution. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 1637-1645. <https://doi.org/10.1109/CVPR.2016.181>
- [3] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 4681-4690. <https://doi.org/10.1109/CVPR.2017.19>
- [4] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1): 53-65. <http://dx.doi.org/10.1109/MSP.2017.2765202>
- [5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B.,

- Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11): 139-144. <https://doi.org/10.1145/3422622>
- [6] Liu, J., Gan, Z., Zhu, X. (2013). Directional bicubic interpolation—A new method of image super-resolution. In *Proceedings of 3rd International Conference on Multimedia Technology (ICMT-13)*, pp. 463-470. <https://doi.org/10.2991/icmt-13.2013.57>
- [7] Esmailzadeh, A., Ahmad, M.O., Swamy, M.N.S. (2021). SRNHARB: A deep light-weight image super resolution network using hybrid activation residual blocks. *Signal Processing: Image Communication*, 99: 116509. <https://doi.org/10.1016/j.image.2021.116509>
- [8] Sun, N., Li, H. (2019). Super resolution reconstruction of images based on interpolation and full convolutional neural network and application in medical fields. *IEEE Access*, 7: 186470-186479. <https://doi.org/10.1109/ACCESS.2019.2960828>
- [9] Zeng, K., Yu, J., Wang, R., Li, C., Tao, D. (2015). Coupled deep autoencoder for single image super-resolution. *IEEE Transactions on Cybernetics*, 47(1): 27-37. <https://doi.org/10.1109/TCYB.2015.2501373>
- [10] Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 624-632. <https://doi.org/10.1109/CVPR.2017.618>
- [11] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 1874-1883. <https://doi.org/10.1109/CVPR.2016.207>
- [12] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, pp. 286-301. [https://doi.org/10.1007/978-3-030-01234-2\\_18](https://doi.org/10.1007/978-3-030-01234-2_18)
- [13] Haris, M., Shakhnarovich, G., Ukita, N. (2018). Deep back-projection networks for super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 1664-1673. <https://doi.org/10.1109/CVPR.2018.00179>
- [14] Kim, S.Y., Oh, J., Kim, M. (2020). Fsr: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss. *AAAI-20 Technical Tracks*, 34(7): 11278-11286. <https://doi.org/10.1609/AAAI.V34I07.6788>
- [15] Chavez-Roman, H., Ponomarev, V., Peralta-Fabi, R. (2012). Image super resolution using interpolation and edge extraction in wavelet transform space. In *2012 9th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Mexico City, Mexico, pp. 1-6. <https://doi.org/10.1109/ICEEE.2012.6421202>
- [16] Johnson, J., Alahi, A., Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *2016 European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, pp. 694-711.
- [17] Shaham, T.R., Dekel, T., Michaeli, T. (2019). SinGAN: Learning a generative model from a single natural image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 4569-4579. <https://doi.org/10.1109/ICCV.2019.00467>
- [18] Dong, C., Loy, C.C., Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 391-407. Amsterdam, The Netherlands. [https://doi.org/10.1007/978-3-319-46475-6\\_25](https://doi.org/10.1007/978-3-319-46475-6_25)
- [19] Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C. (2020). Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 2434-2442. <https://doi.org/10.1109/CVPR42600.2020.00251>
- [20] Ulyanov, D., Vedaldi, A., Lempitsky, V. (2018). It takes (only) two: Adversarial generator-encoder networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.11449>
- [21] Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thurey, N. (2020). Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4): 75:1-75:13. <https://doi.org/10.1145/3386569.3392457>
- [22] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L. (2020). Learning enriched features for real image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1934-1948. <https://doi.org/10.1109/TPAMI.2022.3167175>
- [23] Fan, Y., Yu, J., Liu, D., Huang, T.S. (2020). Scale-wise convolution for image restoration. *AAAI-20 Technical Tracks*, 7, 34(7): 10770-10777. <https://doi.org/10.1609/aaai.v34i07.6706>
- [24] Miyato, T., Koyama, M. (2018). cGANs with projection discriminator. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1802.05637>
- [25] Zhou, K., Li, W., Lu, L., Han, X., Lu, J. (2022). Revisiting temporal alignment for video restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 6043-6052. <https://doi.org/10.1109/CVPR52688.2022.00596>
- [26] Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q. (2019). Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12): 3106-3121. <https://doi.org/10.1109/TMM.2019.2919431>
- [27] Dong, C., Loy, C.C., Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, pp. 391-407. [http://dx.doi.org/10.1007/978-3-319-46475-6\\_25](http://dx.doi.org/10.1007/978-3-319-46475-6_25)
- [28] Dong, C., Loy, C.C., He, K., Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295-307. <https://doi.org/10.1109/TPAMI.2015.2439281>
- [29] Arun, S.N., Shalini, K., Hathiram, N. (2024). Low resolution image enhancement using Res-Net GAN. In: *Artificial Intelligence and Communication Technologies, SCRS*. Soft Computing Research Society, India, pp.

- 1143-1151. <https://doi.org/10.52458/978-81-955020-5-9>-108
- [30] Agustsson, E., Timofte, R. (2017). NTIRE 2017 challenge on single image super-resolution: Dataset and study. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, pp. 126-135. <https://doi.org/10.1109/CVPRW.2017.150>