GAN-BASED PHOTO REALISTIC SINGLE IMAGE SUPER RESOLUTION

Souhaiel BenSalem
ENS Paris Sacaly
souhaiel.ben_salem@ens-paris-saclay.fr

Abstract

Super resolution is a fundamental problem in computer vision, which aims at generating high-resolution images from low-resolution inputs. In recent years, deep learning approaches have shown promising results in addressing this challenge. In this project, we present a comprehensive review of deep learning-based super resolution methods, with a particular focus on GAN-based architectures such as SRGAN [15] and ESRGAN [25]. We discuss the key components of state-of-the-art super resolution networks, including feature extraction, feature fusion, and reconstruction. We also takcle the DIV2K dataset using these approaches while exploring the different training strategies and loss functions that have been proposed to enhance the performance of super resolution models. Furthermore, we provide an overview of benchmark datasets and evaluation metrics for our super resolution models.

1. Introduction

Single image super-resolution (SISR) has become an increasingly popular topic in both the research community and industry as it addresses a fundamental low-level vision problem. The primary objective of SISR is to recover a high-resolution (HR) image from a single, low-resolution (LR) input image and go beyond the limitations of the acquired data. The task of SR has various applications [30] [19] including medical imaging, remote sensing, and surveillance.

Figure 1. Original LR image and the SR image obtained using our model

Traditional methods for super resolution rely on handcrafted features and priors to estimate high-resolution images from low-resolution inputs. However, these methods are often limited in their ability to capture complex image structures and produce visually pleasing results. This is because when the upscaling factor is high, the ill-posed single image super-resolution (SR) problem becomes more difficult, and the texture detail in the reconstructed SR images is often missing.

Typically, the objective of supervised SR algorithms is to minimize the mean squared error (MSE) between the recovered high-resolution (HR) image and the ground truth. The practice of minimizing the MSE is convenient as it maximizes the peak signal-to-noise ratio (PSNR), which is a widely used measure for evaluating and comparing single image super-resolution (SR) algorithms [29]. However, the capability of MSE and PSNR to capture differences that are perceptually relevant, such as high texture detail, is limited. This limitation arises because they are defined using pixel-wise image differences [27].

In recent years, perceptual-driven [10] deep learning approaches such as SRGAN [15] and ESRGAN [25] have shown remarkable performance in super resolution by leveraging large-scale training data and powerful modeling capabilities. These methods laverage the power of genrative adversatial network (GANs) [7] to enforce a solutions that lies in the natural image maniforld.

We investigate the performance of these two architectures by tackling the challenging DIV2K dataset while also proposing a modified version of the loss function and tweaking the training process for better and more effecient results.

It is, however, important to note that our work and tests will was mostly carried using the SRGAN architecture. This is because as we will explain later, the ESRGAN is a much deeper and demanding network that requires computational ressources that we do not posess. We will nevertheless implement ESRGAN and use our implementation and the authors pre-trained weights to compare the performance of both architectures.

2. Related Work

2.1. Pre-deep-learning super resolution methods

Prior to the advent of deep learning, traditional methods for single image super-resolution (SISR) relied on hand-crafted features and priors to recover high-resolution (HR) images from low-resolution (LR) inputs. These methods can be broadly categorized into three groups: interpolation-based, reconstruction-based, and learning-based.

Interpolation-based methods, such as bicubic interpolation and Lanczos interpolation [3], are simple and fast approaches that estimate HR images by interpolating the LR images using predefined interpolation kernels. However, these methods often suffer from blurring and ringing artifacts.

Reconstruction-based methods, such as total variation (TV) [17] minimization and compressed sensing, formulate SISR as an optimization problem and use regularizers to promote sparsity or smoothness in the reconstructed HR images. These methods can produce visually pleasing results, but they are computationally expensive and may require complex optimization algorithms.

Learning-based methods, such as exemplar-based methods and patch-based methods [5] [4] [6] [9] , use training data to learn mappings between LR and HR image patches. Neighborhood embedding approaches [23] [24] are also a subcategory of learning-based methods that were popular prior to the emergence of deep learning. These approaches seek to learn the mapping between LR and HR image patches by embedding them into a common high-dimensional space. Then, the mapping is obtained by performing regression in this space. These methods can capture complex image structures and produce high-quality results. However, they may suffer from the limitations of the training data and may not generalize well to unseen data.

Overall, while pre-deep-learning SISR methods can achieve reasonable results, they often suffer from various limitations and may not be suitable for practical applications. With the emergence of deep learning, new approaches that leverage powerful modeling capabilities and large-scale training data have emerged, which have shown remarkable performance in SISR.

2.2. deep learning super resolution approaches

2.2.1 Convolutional Neural Networks

In the field of computer vision, the current state of the art for many problems is still achieved through the use of convolutional neural networks (CNNs), which have been specifically designed to address these tasks. The success of CNNs peaked especially after the introduction of ImageNet [14] in 2016. Deeper CNN architectures have been shown to have the potential to substantially improve accuracy by modeling mappings of high complexity. However, training these deep networks can prove challengin, which is why batch normalization is often employed to counteract the internal co-variate shift and ensure efficient training.

In light of these advancements, many researchers have explored the use of CNN-based techniques to address this problem. That is why, a lot of CNN based algorithms for SISR emerged during the past few years and have demonstrated remarkable performance. SISR-specific architectures such as [20] and [26] improve upon previous generic methods and can learn upscaling filters directly while producing more accurate results in less computation time. , Kim et al introduced deeply-recursive convolutional network (DRCN) [13] that was considered state-of-the-art at the time.

The current design choice for SISR specific netorks is ResNets [8]. Residual Networks (ResNets) have been shown to be effective in a range of computer vision tasks, including SISR. In a ResNet, the neural network's depth is increased by adding residual connections between the input and output feature maps, allowing the network to learn more complex and abstract features. The key idea is to learn a residual function, which estimates the difference between the HR and LR images, and adds it back to the LR image to obtain the HR image. This residual learning approach has been shown to be effective in addressing the degradation problem in SISR, where high-frequency information is lost in the LR image. This approach was first applied to SISR in [16]. This model achieved state-of-the-art performance on benchmark datasets, demonstrating the effectiveness of residual learning for the task of SISR. Since then, various ResNet-based models have been proposed, including the ones we study in this project (SRGAN and ESRGAN).

2.2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) have emerged as a powerful framework for learning to generate data with high quality and diversity. GANs consist of two neural networks: a generator network that synthesizes data samples from a random noise input, and a discriminator network that tries to distinguish between the synthesized data samples and real data samples. The generator and discriminator are trained simultaneously in an adversarial manner, where the generator tries to fool the discriminator by generating realistic data samples, while the discriminator tries to distinguish between the real and synthesized data samples. GANs have been applied to various computer vision tasks, including super resolution.

GANs have become a popular for SISR due to their ability to generate high-quality and visually realistic images. In the context of super resolution, the generator network synthesizes a super-resolved (SR) image from the low-resolution

(LR) input, and a discriminator network that tries to distinguish between the SR image and a ground-truth HR image. The generator and discriminator are trained simultaneously in an adversarial manner, where the generator tries to fool the discriminator by generating realistic SR images, while the discriminator tries to distinguish between the generated HR images and the ground-truth HR images. The first GAN-based SISR model was SRGAN, proposed in 2017. This model achieved state-of-the-art performance on benchmark datasets and produced visually pleasing SR images with sharp edges and rich textures. Since then, various GAN-based SISR models have been proposed such as the Enhanced SRGAN (ESRGAN)

2.3. Loss functions

Loss functions are an essential component of SISR algorithms as they help to guide the optimization process and determine the quality of the generated SR images. As discussed earlier, loss functions that operate at the pixel level, such as mean squared error (MSE), face difficulty in capturing the uncertainty associated with recovering high-frequency details, such as texture. This is because minimizing MSE tends to result in pixel-wise averages of possible solutions that are often overly smoothed and lack perceptual quality [18] [11]. Alternative such as PSNR also have their drawbacks as we discussed in the introduction.
To address these limitation, alternative loss functions such as perceptual loss [2] and adversarial loss have been proposed. Perceptual loss measures the distance between the high-level features extracted from the reconstructed and ground truth images using a pre-trained deep neural network such as VGG-19 [21]. Adversarial loss, which is used in generative adversarial networks (GANs), encourages the HR images generated by the generator network to be indistinguishable from the ground truth HR images by the discriminator network.

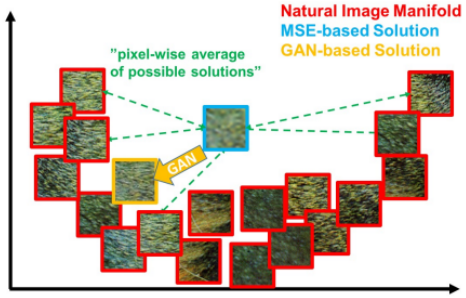

Figure 2. Effect of the used loss : patches from the natural image manifold (red) and super-resolved patches obtained with MSE (blue) and GAN (orange) [15]

These loss functions help to preserve high-frequency details and produce visually pleasing HR images. While both perceptual and adversarial loss functions have shown promising results, recent studies have shown that the combination of the two loss functions, also known as the VGG loss, leads to the best performance in terms of perceptual quality and numerical metrics.

3. Methodology

3.1. Datasets

The models were trained on the DIV2K dataset that contains:

- 1600 training images of different sizes divided into 800 high resolution images and their corresponding x4 down-scaled and bicubic-interpolated low resolution images.
- 200 test images divided into 100 high resolution images and their corresponding low resolution counterparts.

In addition to DIV2K's test set, we also evaluated our models on the Set5 and Set14 benchmark datasets.The performance of the models was evaluated using various quantitative metrics such as peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and visual quality.

3.2. models

The goal of single image super-resolution (SISR) is to produce a high-resolution, super-resolved image I^{SR} from a low-resolution input image I^{LR}. The image I^{LR} is the down-sampled and low-resolution version of its high-resolution counterpart I^{HR}. We harness the concepts of two photo realistic super resolution models: SRGAN [15] and ESRGAN [25].

3.2.1 SRGAN model

Architecture

The primary objective is to train a generative function G that can accurately estimate the corresponding high-resolution (HR) counterpart for any given low-resolution (LR) input image.
The authors use generative function G, which is a feed-forward convolutional neural network, specifically, a deep $ResNet$ (SRReset) parametrized by θ_G. Where, $\theta_G = \{W_{1:L}; b_{1:L}\}$ are the weights and biases of an L-layer deep network. We optimize a super-resolution specific perceptual loss function l_{SR} to obtain θ_G. The generator function is trained on a set of training images I_{HR_n}, I_{LR_n}, where

$n = 1, ..., N$ (N = 800 for DIV2K).
We aim to minimize the following objective function:

$$\hat{\theta_G} = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(I_n^{LR}), I_n^{HR}) \quad (1)$$

We introduce a discriminator network D_{θ_D}, which is optimized in an alternating manner with G_{θ_G} to solve the adversarial min-max problem. This problem aims to find the optimal values of θ_G and θ_D that can generate super-resolved images that are indistinguishable from the high-resolution images in the training set. The objective function for this problem is defined as follows:

$$\min_{\theta_G} \max_{\theta_D} E_{I^{HR} \sim p_{train}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + E_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))] \quad (2)$$

The main idea of using such a formulation is to train a generative model G to produce super-resolved images that can fool a differentiable discriminator D, which is trained to distinguish between real images and super-resolved images. By adopting this approach, the generator network can learn to generate solutions that are photo realistic and difficult for the discriminator network to classify, leading to perceptually superior results that lie in the manifold of natural images. This is in contrast to traditional super-resolution methods that minimize pixel-wise error measurements such as the MSE, which often results in overly smooth and unrealistic super-resolved images.
The architecture of the SRGAN model consists of a deep generator network G, which is composed of 16 residual blocks with identical layout. Each residual block has two convolutional layers with $3x3$ kernels and 64 feature maps, followed by $batch - normalization$ layers and $ParametricReLU$ as the activation function. The resolution of the input image is increased with two trained sub-pixel convolution layers ($PixelShuffle$ layers). We train the discriminator network to tell real HR images from generated SR samples. The discriminator network consist of eight convolutional layers with an increasing number of $3x3$ filters , increasing by a factor of 2 from 64 to 512 kernels, as in the VGG network and strided convolutions are used to reduce the image resolution each time the number of features is doubled. The resulting 512 feature maps are followed by two dense layers and a final $sigmoid$ activation function to enforce a probability for sample classification. We use a $LeakyReLU$ (α = 0.2) is used while avoiding max-pooling throughout the network.
The architecture of the model is summarized in Figure 3

Our Perceptual Loss and training process

The main contribution introduced by authors of the SRGAN paper is the use of the GAN architecture and the introduction of a novel perceptual loss function that consists of an adversarial loss and a content loss based on VGG's feature maps. This loss function is proved to be much better than MSE at capturing perceptually relevant differences.

$$l^{SR} = \underbrace{6.10^{-3} l^{SR}_{VGG/5,4}}_{content\ loss} + \underbrace{10^{-3} l^{SR}_{Gen}}_{adversarial\ loss}$$

Where The generative loss l^{Gen}_{SR} is defined based on the probabilities of the discriminator $D_{\theta_D}(G_{\theta_G}(I_{LR}))$ over all training samples as:

$$l^{SR}_{Gen} = -\sum_{n=1}^{N} \log D_{\theta_D}(G_{\theta_G}(I^{LR}))$$

Our approach consisted of implementing the SRGAN architecture proposed by the authors while making changes on the perceptual loss function and the training process.

Modified perceptual loss function for SRGAN:

through intuition and experimentation, we introduce the following perceptual loss function:

$$l^{SR}_{modified} = \underbrace{l^{SR}_{MSE}}_{pixel\ loss} + \underbrace{6.10^{-3} l^{SR}_{VGG/5,4}}_{content\ loss} + \underbrace{10^{-3} l^{SR}_{Gen}}_{adversarial\ loss} + \underbrace{2.10^{-8} l^{SR}_{TV}}_{total\ variation}$$

Intituition:

- We introduce MSE loss to penalize the differences in pixel space which ultimately leads to more accurate color fidelity between I^{SR} and I^{HR}. By minimizing this difference, the super-resolved image can achieve accurate color fidelity with the ground truth image. In addition to the adversarial loss and perceptual loss, MSE loss can provide a more direct control of the image quality and is computationally efficient to compute. Therefore, the introduction of MSE loss in our model can lead to improved image quality and greater fidelity between the super-resolved image and the ground truth image in terms of pixel-wise similarity.

- Inspired by style transfer GANs [28], we introduce TV loss to reduce noise inI^{SR}. TV loss acts as a regularization term that measures the overall variation of intensities in the image, promoting spatial smoothness and reducing noise. In the context of SISR, the introduction of TV loss can improve the quality of the

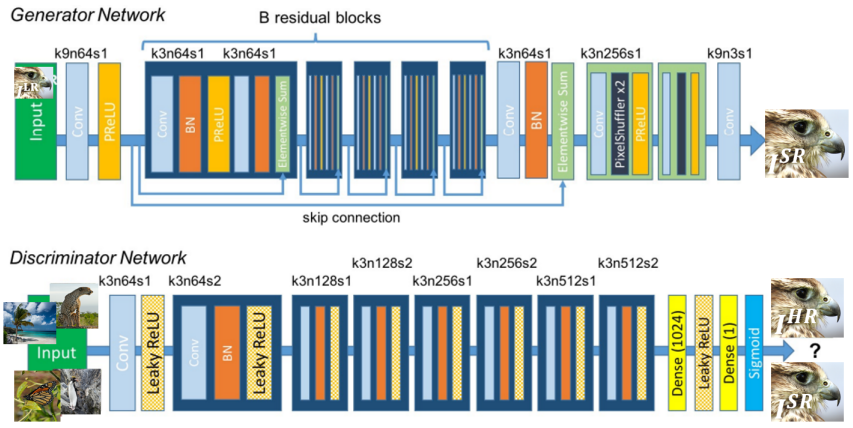

Figure 3. The Generator and Discriminator networks of our SRGAN model

super-resolved image by reducing the noise that may arise from the low-resolution input image. However, since TV loss can also lead to smoothing of textures in the image, we used a very low weight its loss component to avoid crashing textures during the super-resolution process. This allows for better preservation of the high-frequency details in the super-resolved image, while still benefiting from the noise reduction provided by the TV loss.

Tweaking the pre-training process:

Instead of pre-training the Generator network using MSE loss like the authors did, we experimented with a **weighted MSE** and $L1$ **loss**, which both yielded better results than MSE from a human viewer's perspective, and settled for the latter. Intuitively, this is because MSE is sensitive to outliers, which can result in overly-smoothed solutions and loss of fine details in the image. In contrast, MAE gives more weight to small differences between the predicted image and the ground truth image, allowing for better preservation of the high-frequency details in the SR image. In addition, MAE is a more robust loss function for handling the non-linear mapping from the low-resolution input image to the high-resolution output image. By minimizing the absolute differences between the predicted and ground truth images, the generator is able to learn more effectively and produce higher quality super-resolved images.
For the most part, we follow the same training process described by the authors in the paper.
We trained our model for 850 $epochs$ and used the following setting of hyperparameters:

- LR_CROPPED_SIZE = 24: The size of the low-resolution cropped input image used for training.
- UPSCALE = 4: The upscaling factor used to generate the high-resolution output image from the low-resolution input image.
- HR_CROPPED_SIZE = UPSCALE * LR_CROPPED_SIZE: The size of the high-resolution cropped output image used for training.
- BATCH_SIZE = 16: The number of image samples in each mini-batch during training.
- EPOCHS = 50: The number of epochs (i.e., complete passes through the training dataset) used for training.
- LR = 0.0001: The learning rate used for training the model.
- BETAS = (0.5, 0.9): The values of the beta1 and beta2 hyperparameters used by the Adam optimizer during training.
- adversarial_loss_coef = 0.001: The coefficient for the adversarial loss used during training.
- vgg_loss_coef = 0.006: The coefficient for the VGG loss used during training.

Observations:
During our training and experimentation with SRGAN, we observed that the super-resolved images produced by the model exhibited noticeable artifacts, particularly in dark areas and around edges. Upon further investigation, we discovered that these artifacts were caused by the fact that our generator was producing out-of-range values, specifically negative floating points, that upon projection to the interval $[0, 255]$ when reconstructing the image would cause these 'pixel burning' artifacts shown in Figure4
This is a known problem for SRGAN and it caused mainly because of the Batch Normalization layers. In our case, we

Figure 4. examples of the artifacts caused by batch normalization

used a simple $ReLU$ function to clip negative values and solve this problem.

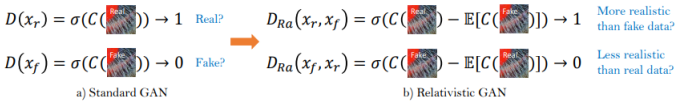

Figure 5. standard discriminator VS relativistic discriminator [25]

3.2.2 ESRGAN model

Architecture

ESRGAN (Enhanced Super-Resolution Generative Adversarial Networks) is an improved version of SRGAN that addresses some of the limitations of the original model. ESRGAN improves on the ideas introduced by SRGAN to achieve a better perceptual quality. The improvements are mainly technical and consist of modifying the architecture of the generator network as shown in Fig.7.
One major difference between ESRGAN and SRGAN is the use of the Residual-in-Residual Dense Block (RRDB) in ESRGAN, which replaces the residual blocks used in SRGAN. The RRDB contains multiple Residual Dense Blocks (RDBs), which are used to learn the features of the input image at different scales. The output of each RDB is then combined to form the final output of the RRDB. This approach allows for more efficient feature learning and greater preservation of details in the super-resolved image.
They also removed batch normalization layers which caused artifacts in the SR images produced by SRGAN and used nearest neighbor upsampling instead of pixelshuffling. The authors also modified the discriminator network based on the Relativistic GAN [12]. As opposed to the traditional discriminator in SRGAN, which determines the probability that an input image x is both real and natural, a relativistic discriminator focuses on predicting the probability that a real image x_r is comparatively more realistic than a fake image x_f.

This is illustrated by the following figure :

Enhanced Perceptual Loss and training

The loss function was also modified by introducing L_1 penalty and taking VGG features before activation for the perceptual loss component.

$$L_G = L_{percep} + \lambda L_G^{Ra} + \mu L_1$$

where $L_1 = E_{x_i}||G(x_i) - y||_1$ is the content loss that evaluates the 1-norm distance between the recovered image $G(x_i)$ and the ground-truth y, and λ and η are the coefficients used to balance different loss terms.

Due to the inherent large size of the ESRGAN network and the ambiguity of some implementation details in the original paper, we were not able to train our own ESRGAN model from scratch. However, we implemented the network architecture as described in the paper and converted the authors' pre-trained weights to be compatible with our implementation. We used our ESRGAN model to compare its performance with our SRGAN model on DIV2K validation set and the other benchmark datasets, in order to evaluate its effectiveness and generalization capability since the authors did also train their model on DIV2K.

4. Evaluation and Results

4.1. Quantitative results

Quality Measures

Our SRGAN model was trained for 850 epochs using the same parameters described before. We evaluated the two methods on the DIV2K validation set and the $Set5$ and

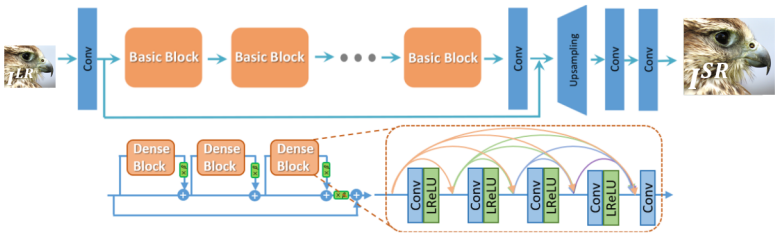

Figure 6. ESRGAN Generator network

$Set14$ benchmark sets. We use the $PSNR$ and $SSIM$ metrics for evaluation and confirm via experimentation that MSE is not reliable to measure perceptual accuracy.

- **PSNR** : or Peak Signal-to-Noise Ratio is a commonly used metric for evaluating the quality of a digital image or video signal. PSNR measures the ratio of the peak signal power to the noise power in the signal, and is often expressed in decibels (dB). In the context of image super-resolution, PSNR is often used as a quantitative measure to compare the quality of the super-resolved image to the ground truth high-resolution image. A higher PSNR value indicates a better quality image, as it implies that the amount of noise in the signal is relatively low compared to the strength of the signal itself. Mathematically, we can express it as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right)$$

where MAX_I is the maximum possible pixel value of the image.

- **SSIM** : or Structural Similarity Index is a method used to measure the similarity between two images by taking into account the structure of the images, rather than simply comparing the individual pixel values. SSIM is a full reference metric, meaning that it compares the processed image to the original image rather than comparing two processed images. In the context of image super-resolution, SSIM can be used as a quantitative measure to compare the quality of the super-resolved image to the ground truth high-resolution image. A higher SSIM value indicates a better quality image, as it implies that the processed image is more similar to the original image in terms of its structure, contrast, and luminance.
Mathematically, we can express it as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where x and y are the input images, μ_x and μ_y are the mean values of x and y, σ_x^2 and σ_y^2 are the variances of x and y, and σ_{xy} is the covariance of x and y. The constants c_1 and c_2 are small constants added to avoid division by zero.

Results

Using our own implementation of the discussed quality metrics, we evaluate and compare our SRGAN with the results of the original SRGAN paper and our ESRGAN. From the qualitative results displayed in the table below, we can see that our SRGAN model compares fairly well with the results of the authors of the original SRGAN paper who used a much larger training set (350k images sampled from ImageNet) and trained their model for much longer. We can also see the superiority of the ESRGAN model in each case. Indedd the enhanced architecture and the subtle but important technical modifications the authors introduced seem to be improve the results quantitatively. However, it is unclear to me if this improvement over previous methods is mostly thanks to the subtle architectural and theoretical changes or the sheer fact that the generator network of ESRGAN is much deeper and involves much more convolutions.
Note that for ground-truth HR images, $PSNR = \infty$ and $SSIM = 1$

.Performance of the two methods

DIV2K	SRGAN(ours)	SRGAN (Ledig et al)	ESRGAN
PSNR	25.373	-	28.174
SSIM	0.706	-	0.775
Set5			
PSNR	24.945	29.40	30.474
SSIM	0.718	0.8472	0.851
Set14			
PSNR	23.770	26.02	26.614
SSIM	0.636	0.7397	0.713

4.2. Qualitative results

Qualitative results are a pivotal component in the comprehensive evaluation of super-resolution methods as they offer an insightful and intuitive perspective to the

quantitative results obtained through objective metrics, such as PSNR and SSIM. The visual confirmation provided by qualitative analysis enables a more precise and elaborate understanding of the models' ability to recover the lost details and textures in the low-resolution images, and the perceptual quality of the reconstructed high-resolution images.

In our case, qualitative results are in accordance with the results we got using the analytical quality measures. In fact, our qualitative results demonstrate the ability of SRGAN in retrieving fine details and textures that were missing in the original low-resolution images. The high-resolution images produced by SRGAN show sharp and well-defined edges and feature relatively smooth transitions in color and texture. Overall, our results confirm the effectiveness of proposed perceptual loss and pre-training procedure. However, the results are not perfect or as good as ESRGAN's. This is expected since we only trained our model for 850 epochs on a relatively small dataset.

On the other hand, ESRGAN exhibits very good performance in producing super-resolved images with rich details and textures. The generated images have high perceptual quality (from a human's point of view) and surpass the ones produced by SRGAN. The enhanced performance of ESRGAN can be attributed to its more robust discriminator that utilizes a relativistic approach to evaluate the realism of the generated images, the subtle changes made in the loss function and architecture or the fact that it is a much more deeper network.

5. Discussion and Future work

In this project, we assessed the effectiveness of SRGAN and ESRGAN (mostly SRGAN) on the DIV2K dataset and other benchmarking sets. Both networks showed improvement over the traditional interpolation methods as well as previous state-of-the-art super-resolution methods. The use of perceptual loss and deep residual learning in both SRGAN and ESRGAN have proven to be effective in super-resolving images, and the use of pre-trained VGG network in SRGAN has shown to be effective in capturing the high-level feature similarities between the super-resolved and ground truth images. However, the comparison between the two networks has shown that ESRGAN performs better than SRGAN in terms of both quantitative metrics (PSNR, SSIM) and visual quality.

Moreover, using a modified version of the perceptual loss function proposed by Ledig et al, our SRGAN model performs reasonably well on the benchmark datasets and so does ESRGAN, especially given the relatively low number of epochs that we used given that SR methods and GAN-based models in generals require a lot of training time . We intend to continue with the training of ESRGAN on DIV2K and re-traing SRGAN for longer time and with a heavy augmentation pipeline as for this project, no augmentations were applied except a center crop. Allso exploring different datasets can prove to benifical as explained in [22].

We also intend to explore multi-frame super resolution methods like the one presented in [1]. These methods rely on fusing images of the same sceneg enerated by hand movements or intentional camera hardware move- ments, and merge their information to obtain a higher reso- lution image.

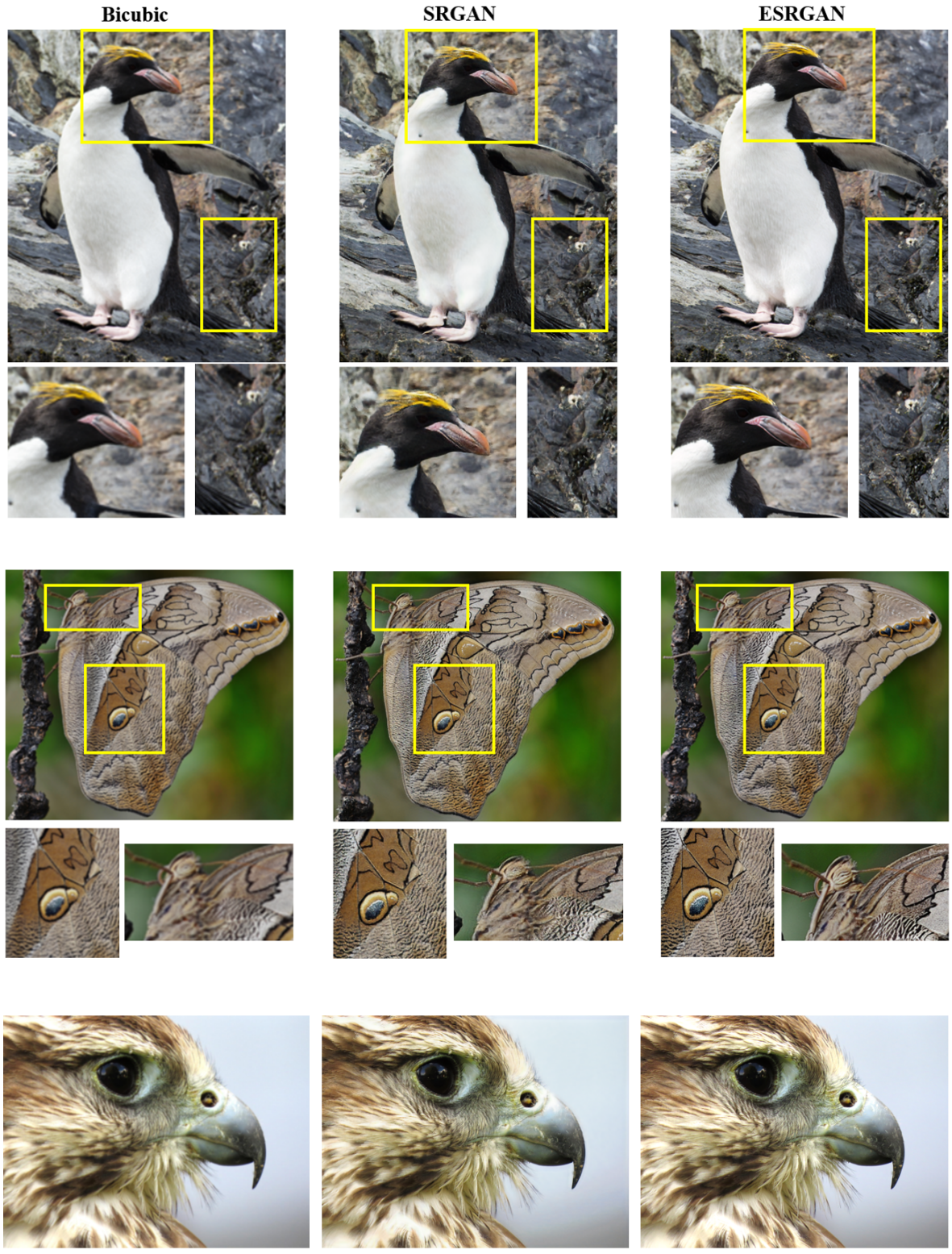

Figure 7. Bicubic (left) VS SRGAN (middle) VS ESRGAN (right) reconstruction results

References

[1] Goutam Bhat, Martin Danelljan, L. Gool, and R. Timofte. *: Deep burst super-resolution*. CVPR 2021. 8

[2] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics, 2015. 3

[3] Claude E. Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology and Climatology*, 18(8):1016 – 1022, 1979. 2

[4] W.T. Freeman, T.R. Jones, and E.C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002. 2

[5] W.T. Freeman and E.C. Pasztor. Learning low-level vision. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1182–1189 vol.2, 1999. 2

[6] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 349–356, 2009. 2

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[9] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 2

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016. 1

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 3

[12] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan, 2018. 6

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution, 2015. 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 2

[15] C. Ledig, Theis, L., Husz´ar, F.and Caballero, A.and Acosta J.and Cunningham, A., Aitken andA., Tejani, A., J. Totz, Wang, Z., and et al. *: Photo-realistic single image super resolution using a generative adversarial network*. In CVPR 2017. 1, 3

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017. 2

[17] Jiaming Liu, Yu Sun, Xiaojian Xu, and Ulugbek S. Kamilov. Image restoration using total variation regularized deep image prior. *CoRR*, abs/1810.12864, 2018. 2

[18] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error, 2015. 3

[19] Kamal Nasrollahi and Thomas B. Moeslund. Super-resolution: A comprehensive survey. *Machine Vision Applications*, 25(6):1423–1468, June 2014. 1

[20] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, 2016. 2

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 3

[22] Nao Takano and Gita Alaghband. *: SRGAN: Training Dataset Matters*. 8

[23] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013. 2

[24] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, pages 111–126, Cham, 2015. Springer International Publishing. 2

[25] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. *: ESRGAN: enhanced super-resolution generative adversarial networks*. 1, 3, 6

[26] Yifan Wang, Lijun Wang, Hongyu Wang, and Peihua Li. End-to-end image super-resolution via deep and shallow convolutional networks, 2016. 2

[27] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[28] Zheng Xu, Michael Wilber, Chen Fang, Aaron Hertzmann, and Hailin Jin. Learning from multi-domain artistic images for arbitrary style transfer, 2018. 4

[29] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 372–386, Cham, 2014. Springer International Publishing. 1

[30] Qingxiong Yang, Ruigang Yang, James Davis, and David Nister. Spatial-depth super resolution for range images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 1