WAEP: Wide Activation with Enhanced Perception Super-Resolution General Adversarial Networks

Shashank Agarwal, Devikalyan Das, and Nobel Jacob Varghese

Universität des Saarlandes, 66123 Saarbrücken, Germany
{shag00001,deda00002,noja00001}@stud.uni-saarland.de

Abstract. Over the last decades, convolutional neural networks have provided remarkable improvement in single image super-resolution (SISR) as compared to classical super resolution algorithms. Among recent advances, GAN based networks focusing on perceptual quality provides photo-realistic SR results. However, visual perception is a subjective matter and there is still room for improvements. Even though recent approaches like ESRGAN provides perceptually enhanced SR images, it suffers from discolored artifacts. Moreover, super resolution is an ill posed problem but many state-of-the-art methods instead use a deterministic mapping approach and ignore the stochastic variation. Hence, we propose a novel GAN based network architecture with wider activation channels, regularization in the network and a novel loss function based on LPIPS. Benefiting from these improvements the proposed WAEP-SRGAN produces more realistic images with better visual quality and reduced artefacts. The performance gains of our method has been quantified using MSE, perceptual and no reference based metrices.

Keywords: Single Image Super-Resolution · Generative Adversarial Networks · Perceptual Loss Function ·

1 Introduction

Single image super resolution (SISR) has received increased attention from the computer vision community. SISR task aims at estimating high resolution image from its corresponding single low-resolution image. SRCNN [3] brought breakthrough development by employing CNN based approach for solving the task of SISR. Various network designs which came after this continuously improved SR performance, but failed to achieve superior perceptual SR result, as these were mainly optimized by minimizing MSE in order to achieve high PSNR. They failed to capture the perceptually important differences such as high frequency details.

Perceptually oriented approach with generative adversarial network such as SRGAN [10] which tried to provide a more realistic and pleasing SR results. These methods used a perceptual loss which minimizes the error in the feature

space instead of pixel space as used in PSNR based approach. SRGAN uses residual blocks and perceptual loss for optimization in a GAN [4] framework to improve visual quality of the SR image. ESRGAN [18] enhanced the SRGAN by using an Residual-in-Residual Dense Blocks (RRDB) without batch normalization and used residual scaling for training a very deep network. Also, it used Relativistic average GAN (RaGAN) [7] based improved discriminator and VGG features based improved perceptual loss before activation.

In this work we propose WAEP-SRGAN, a generative adversarial network based super resolution architecture, which tries to achieve enhanced SR results compared to the baseline ESRGAN .

2 Related Work

Single image super resolution (SISR) have been tackled by various approaches such as interpolation, learning based approaches. Among the deep learning-based approach, SRCNN [3] made the first breakthrough providing superior performance compared to previous classical approaches using convolutional neural network in an end-to-end manner. Then, there had been many approaches around this network such as VDSR [8] which made the network deeper, DRCN [9] introduced recursive networks, SRResNet [10] with residual parts, EDSR [11] provided superior performance by removing batch normalization. RCAN [20] achieved superior PSNR performance with residual connections and channel attention. Other architectures such as networks with residual dense blocks [21], Memnet [17] with densely connected network have also been introduced for addressing the task of super resolution. Also, architectures with channel attention [20] have been explored for SR.

There has also been many GAN [4] based methods for achieving photo-realistic SR results such as WGAN [2] which minimized the Wasserstein distance and used weight clipping to regularize the discriminator, SRGAN [10] which proposed to improve the perceptual quality by introducing a perceptual loss in a GAN based framework. SRGAN uses perceptual loss along with adversarial loss for achieving photo-realistic output SR images. Recent works such as ESRGAN [18] enhanced the SRGAN by using a better architecture with RRDB blocks along with relativistic average GAN based discriminator. It also removed batch normalization layers, introduced residual scaling [16] and used smaller initialization which helped training a deeper network. Also, the perceptual loss used in ESRGAN, has the VGG features used before activation while it was used after activation in SRGAN.

There are still exists room for improvement in ESRGAN results if we compare with ground-truth images. ESRGAN produces discoloured artefacts at certain regions of the image [12] and it is mostly deterministic in its approach whereas

super resolution is fundamentally an ill-posed problem. Our proposed WAEP-SRGAN tries to address these issues. Our main contributions are:

- We propose a new GAN based-architecture with wider activation maps and noise in the discriminator. This ensured more information available to deeper layers and introduced stochasticity to the network as the previous GAN based approaches were mainly deterministic.

- We propose a modified weighted loss function which helped to address the discoloured artefacts and further enhance the perceptual quality.

this report, we describe about the proposed methods including network architecture and loss function in section 3. Experimental details such as train and test data, training details and detailed discussion about results with visual illustrations are provided in section 4. The future work and concluding remarks are in section 5.

3 Proposed Method

3.1 Network Architecture

ESRGAN offers a basic block in the generator with high capacity, and the same has been used in our proposal, with a modification by using wider activation maps. We used 128 activation maps instead of 64 in each convolution layer. This helped in providing higher amount of data flow in the network and thereby using more information for extraction of better features. We have also used the same number of wider activation maps in the discriminator. Fig. 1 depicts the overall architecture of generator used in our proposed model and fig. 2 shows the internal architecture of each Wide Activation Block (WAB) of the generator.

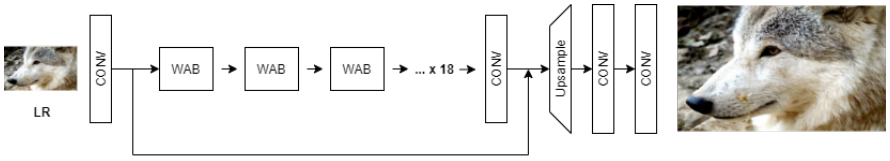

Fig. 1. Architecture of Generator in WAEP-SRGAN using **18** Wide Activation Blocks (WAB)

Recent works[15,5,6] have proven deterministic networks often suffer from the problem of model convergence failure. This is due to the fact that the real and generated outputs have disjoint support regions and makes the discriminator ideal. This ultimately leads to the problem of vanishing gradient and makes

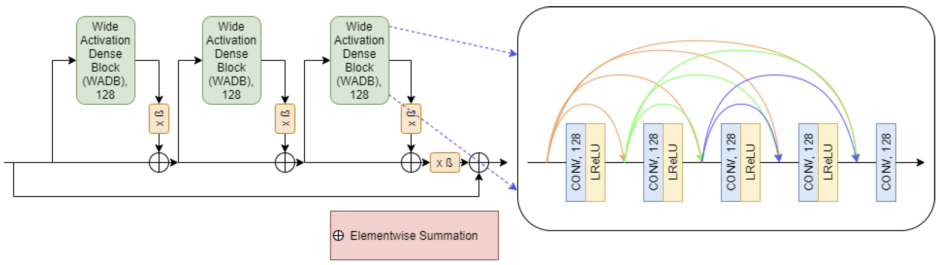

Fig. 2. Schematic of each 'Wide Activation Block (WAB)' using activation maps of size **128**

the model unstable. In order to avoid such issues, introducing regularization in the discriminator would make its job harder and force it to learn better. To implement this, we added instance noise to the inputs (HR label image and SR generated image) of the discriminator. This noise decreases linearly by a factor of 0.1 with a patience level of 7 over the iterations as shown in fig 3

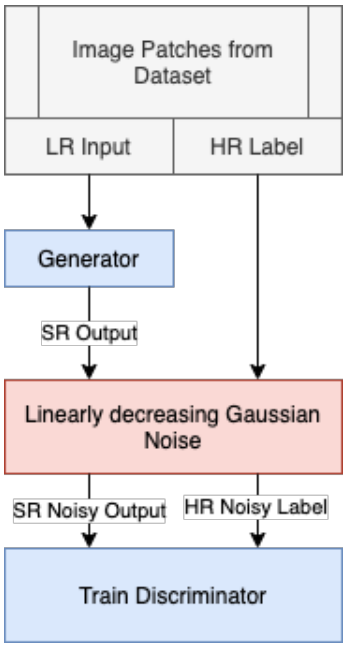

Fig. 3. Schematic of the inputs to the discriminator. Instance noise has been added to SR output and HR label images to train the discriminator

3.2 Loss Function

Furthermore, ESRGAN [18] uses a weighted loss function using content loss ($L_{content}$) based on L1 weighted perceptual loss using VGG extractor, adversarial loss (L_G) using the output from the discriminator, and L1 pixel loss (L_{pixel}) based on pixel to pixel differences. In our model, we use a novel loss function (shown in fig. 4 and eq. (1)) which introduces a new weighted loss (L_{LPIPS}) using L2 perceptual loss based on Learned Perceptual Image Perception (LPIPS) [19] loss function. LPIPS loss uses weights of Alexnet pre-trained model for extracting the feature between the SR generated image and HR image, and uses an L2 loss which helps in making the generated images smoother and removing the discolored artifacts present in ESRGAN.

$$L_G = L_{content} + \lambda L_G^{Ra} + \eta L_{pixel} + \underbrace{\alpha L_{LPIPS}}_{\text{new loss term}} \tag{1}$$

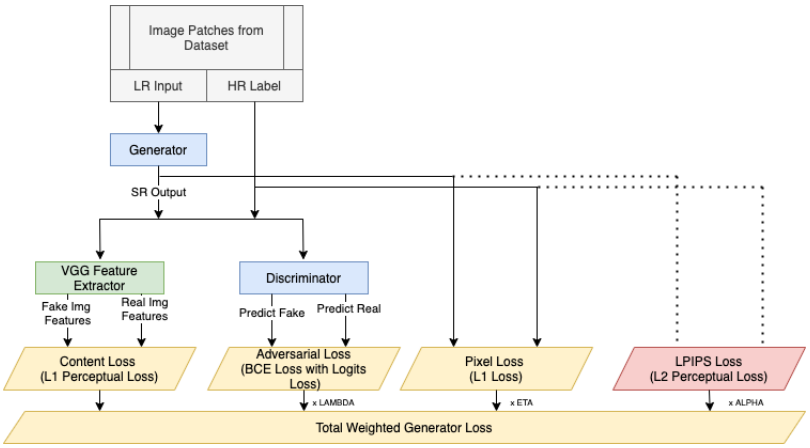

Fig. 4. Schematic of computing the new loss to train the generator.

4 Experimental Results

4.1 Data

We wanted to use the same dataset used in ESRGAN for training and so we chose DIV2K[1] data set for this project. It provides high quality (2k) images for image restoration tasks along with their corresponding low resolution images. The low resolution images are a down-scaled (x4) version of the high quality images using bicubic interpolation.

We used 800 images from the data set and divided them randomly into two subsets: training (640 images) and validation (120 images). For training set, we

used a total of 40 images. Training on such a large dataset also helped the model to get better trained in extracting a variety of different features.

We have trained the model using RGB channels (same as done in ESRGAN) and performed data augmentation including random vertical and horizontal flip, random rotation and normalisation.

One of the challenge we faced using this dataset was the varied size of the images and to address this issue, we uniformly resized them to identical dimensions: nearest multiple of 64 and 256 dimensions for LR and HR images, respectively. This further helped in using a patch based approach for training the model. Fig. 5 shows the usage of patch based approach during the test phase.

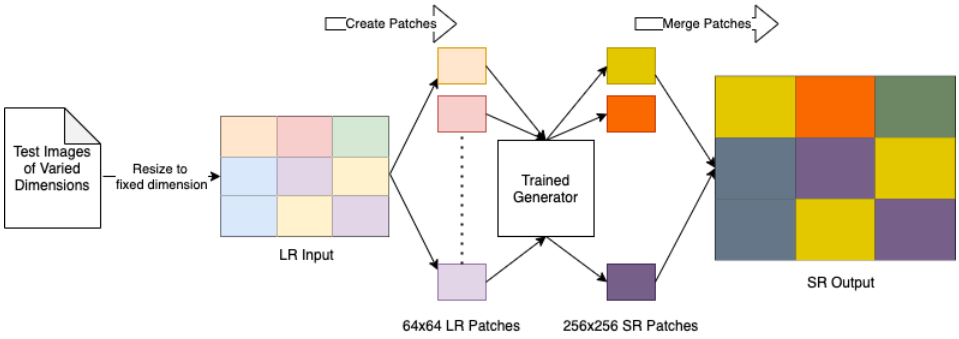

Fig. 5. Data flow during test phase using patch-based approach to generate super resolution images

4.2 Training Process

Training GANs for image super resolution requires a large number of iterations, for example ESRGAN used 500k iterations to train the model. Keeping the time constraints in mind, we trained WAEP-SRGAN and the baseline (ESRGAN) [14] for 22k iterations. Details on time and space consumption are available in table 3 in Appendix A.

ESRGAN also uses a patch based approach while training the network and since the deeper networks benefit from larger patches[18], we increased the patch size from 32x32 to 64x64 while training the model. We also noticed that decreasing the number of Wide Activation Dense Blocks in the generator from 23 to 18 does not degrade the performance much, instead saves the computational effort. All models were trained using a NVIDIA-RTX3090 GPU with 24GB DDR6x memory. Further training details have been provided in table 4 in Appendix B.

4.3 Results

We have compared the performance of our model both qualitatively and quantitatively against the baseline (ESRGAN). As mentioned in SRFLOW[12], ES-

RGAN suffers from discolored artifacts in some locations. Visually, it can be inferred from fig. 6 that WAEP-SRGAN is able to address this issue with the help our novel loss function, achieving better optimization. Moreover, it can be seen that even though bicubic interpolation gives better quantitative results, results from our model look perceptually better in terms of realistic looking images.

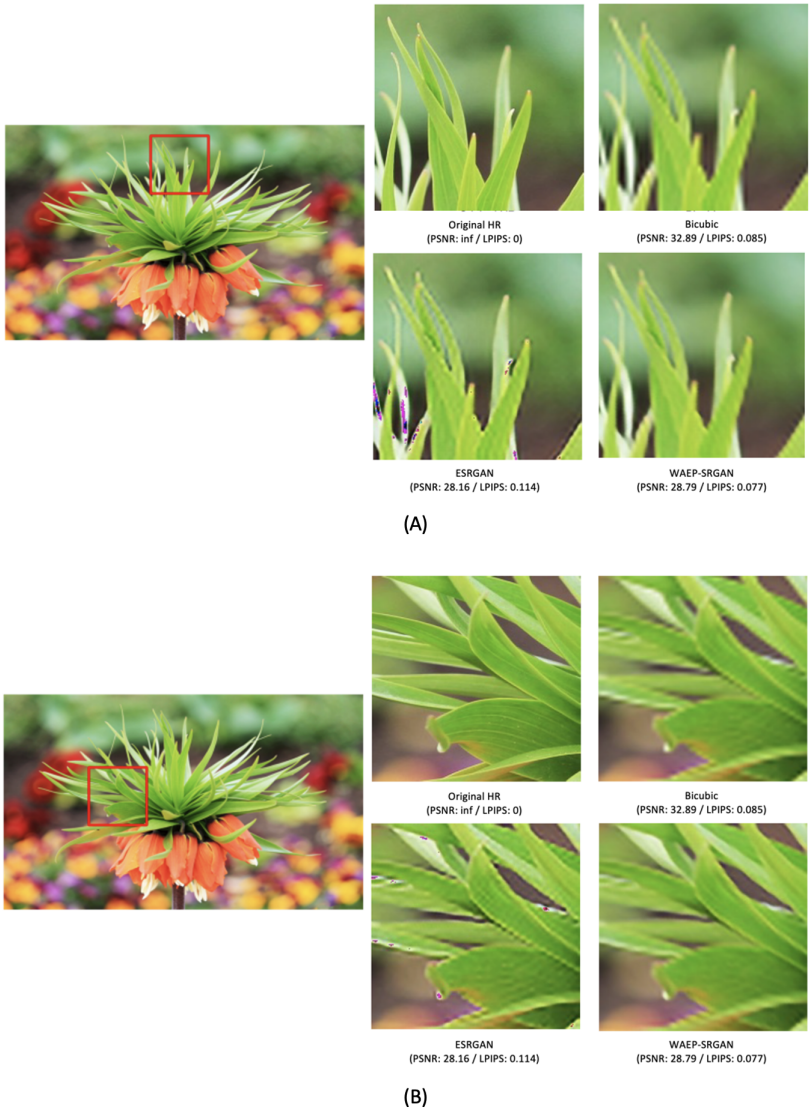

Fig. 6. Overall visual comparisons focusing on two different patches of the flower image

Fig. 7 further shows that our model is also able to make the SR images more realistic looking as compared to ESRGAN, which appears like an oil painted image.

Fig. 7. Overall visual comparisons focusing on two different patches of the building image

Quantitatively, we have evaluated the performance of our model using a combination of mean squared error as well as perceptual based metrics - PSNR (Peak Signal to Noise Ratio), SSIM (Structural Similarity Index), LPIPS (Learned Perceptual Image Patch Similarity)[19], BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator)[13]. PSNR has been measured on the luminance channel in the YCbCr color space which focuses more on the pixel wise grey value instead of perceptual similarity. On the other hand, SSIM quantified image quality degradation based on luminance, contrast and structure. LPIPS is a learnable metrics which helped in evaluating the distance between the feature maps of the SR(output) and HR (label) images. Lastly, BRISQUE being a no reference quality metrics helped in quantifying the results in a different way using point wise statistics and deviations from natural image model. As a whole, we tried to check the quality of the results in different aspects using a variety of different quality metrices.

Table 1 shows the metric scores comparison of our model with the baseline and bicubic interpolation. Even though bicubic interpolation achieves the best PSNR score but it is not a good indicator for measuring the realistic quality of the images. Both ESRGAN and WAEP-SRGAN produced images with better textures and details as compared to bicubic interpolation. Results of these models are based on 22k iterations of training and it is evident that our proposed method (WAEP-SRGAN) outperforms ESRGAN in terms of PSNR, LPIPS as well as BRISQUE with a good margin. However, it can be observed that our model could not improve SSIM score, which is due to the smoothing introduced by our modified loss function. Since, there is a trade off between the sharpness and discolored artifacts in the image, addressing this will be the focus of our future work.

Table 1. Quantitative evaluation of Bicubic interpolation, ESRGAN (baseline) and WAEP-SRGAN (our proposal) SR algorithms: average PSNR (on Y channel), LPIPS, BRISQUE, SSIM. The best results between ESRGAN and WAEP-SRGAN are **highlighted**. Arrows denote if the values should be higher or lower for the metrics

Models	**PSNR↑**	**LPIPS↓**	**BRISQUE↓**	**SSIM↑**
Bicubic	24.94	0.2439	48.93	0.9237
ESRGAN	22.60	0.2196	50.92	**0.9091**
WAEP-SRGAN	**22.82**	**0.2113**	**44.52**	0.9005

4.4 Ablation Study

We first developed ESRGAN baseline and ran it for 22k iterations, followed by a modified version of it wherein we mainly added wider activation maps in the generator and discriminator (128 instead of 64) along with the addition of instance noise in the discriminator. This helped in checking the model behaviour and the amount of improvement in the performance. As shown in table 2, the

metrics performance improved by a small but considerable margin. Then, we added our new novel loss function on top of this model along with some minor changes in the hyper-parameter settings, which outperformed ESRGAN metrics results by a good margin.

Table 2. Ablation study for ESRGAN (baseline), WAEP-SRGAN Lite is the model with wide activation and noise in discriminator, WAEP-SRGAN (our proposal). Values in bracket indicate the change in scores w.r.t. ESRGAN (blue values are improvements and red values are degradation) (All the models have been run for **22k** iterations)

Models	**PSNR↑**	**LPIPS↓**	**BRISQUE↓**	**SSIM↑**
ERGAN	22.60	0.2196	50.92	0.9091
WAEP-SRGAN (Lite)	22.68 (+0.08)	0.2141 (-0.0055)	49.82 (-1.1)	0.9094 (+0.0003)
WAEP-SRGAN	22.82 (+0.20)	0.2113 (-0.0083)	44.52 (-6.4)	0.9005 (-0.0087)

5 Conclusions and Future Work

We aimed to enhance the perceptual quality of SR images by removing high frequency artefacts producing better results both qualitatively and quantitatively, but there is still room for improvements. We found through our work that the better content loss we have, the more enhanced feature maps will be available for better reconstruction. The addition of an extra weighted L2 perceptual loss (LPIPS loss) addressed the high frequency artefacts, but it introduced slightly smooth results. This proves the trade-off between sharpness in the image and removal of high frequency artifacts, and our goal for the future would be to better tune the weight in the loss function for this newly added perceptual loss. This would help in achieving the optimal SR output images in terms of sharpness and reduction of artifacts.

Importantly, number of training iterations greatly influences the performance of the model and due to high amount of training time it was not possible for us to train the model for long duration. Thus, as a next step we would propose to increase the number of iterations and check how well our proposed model performs for 500k iterations in comparison to ESRGAN.

We also speculate that increasing the stochasticity in the network can further improve the results for a wider variety of images. This work can also be further explored to produce super resolved images on various domains like text recognition, object recognition etc.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)

2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (2017)
3. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks (2015)
4. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)
5. Guan, J., Pan, C., Li, S., Yu, D.: Srdgan: learning the noise prior for super resolution with dual generative adversarial networks (2019)
6. Jenni, S., Favaro, P.: On stabilizing generative adversarial training with noise (2019)
7. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard gan (2018)
8. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks (2016)
9. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution (2016)
10. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network (2017)
11. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution (2017)
12. Lugmayr, A., Danelljan, M., Gool, L.V., Timofte, R.: Srflow: Learning the super-resolution space with normalizing flow (2020)
13. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050
14. Norén, E.L.: Pytorch-gan. https://github.com/eriklindernoren/PyTorch-GAN/tree/master/implementations/esrgan (2019)
15. Rakotonirina, N.C., Rasoanaivo, A.: Esrgan+: Further improving enhanced super-resolution generative adversarial network. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (May 2020). https://doi.org/10.1109/icassp40776.2020.9054071, http://dx.doi.org/10.1109/ICASSP40776.2020.9054071
16. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning (2016)
17. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration (2017)
18. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y., Tang, X.: Esrgan: Enhanced super-resolution generative adversarial networks (2018)
19. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018)
20. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks (2018)
21. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution (2018)

A Comparison of time and memory consumption between WAEP-SRGAN and ESRGAN

Table 3. Time and memory requirements for ESRGAN and WAEP-SRGAN **(for 22k iterations)**

Requirements	**ESRGAN**	**WAEP-SRGAN**
No. of Trainable parameters	38,549,123	120,956,803
Total memory (in MB)	1762.60	3205.32
Total time (in hours)	7	8

B Comparison of hyper-parameters used between WAEP-SRGAN (our proposal) and ESRGAN (baseline)

Table 4. Various hyper parameters used in WAEP-SRGAN with their quantitative comparison with ESRGAN

Parameters	**WAEP-ESRGAN**	**ESRGAN**
Upscale Factor	x4	x4
Batch Size	3	16
Image Patch Size	64x64	32x32
Base no. of features (G)	128	64
Base no. of features (D)	128	64
Residual Dense Blocks (G)	18 WAB	23 RRDB
Residual Scale (G)	0.2	0.2
Optimizer	Adam	Adam
No. of iterations	22k	500k
LR Scheduling	Decrease by factor of 0.1 at patience 7	Decrease by 0.5 at [50k, 100k, 200k, 300k]
Pixel Loss Weight	1e-2	1e-2
Adversarial Loss Weight	5e-3	5e-3
Content Loss Weight (Perceptual L1 Loss)	1	1
LPIPS Loss Weight (Perceptual L2 Loss)	1e-1	Not Used

Contribution of team members

We would like to thank our tutors for the valuable feedback. We faced quite a few challenges which helped us to manage time in the best way possible. Given below are the major contribution of the team members in this project :

Shashank Agarwal (ID : 7009562)

- Code
- Presentation
- Report

Devikalyan Das (ID : 7007352)

- Code
- Presentation
- Report

Nobel Jacob Varghese (ID : 7002401)

- Presentation
- Docker Setup