A REPORT

ON

UPSCALING BLURRY IMAGES FOR ENHANCED WEB CONFERENCING

UNDERTAKEN AT

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

APRIL, 2021

UPSCALING BLURRY IMAGES
FOR ENHANCED WEB CONFERENCING

This report has been submitted by:

Name: Rohit Jain

ID: 2017A7PS0122P

Course No.: CS F376 Design Oriented Project

Name: Atharva Chandak

ID: 2019A7PS0062P

Course No.: CS F266 Study Oriented Project

In partial fulfilment of the project type courses, Second Semester 2020-21 under Prof. J. Jennifer Ranjani, CSIS Department, BITS Pilani.

ABSTRACT

Amidst the Covid-19 pandemic, both educational institutes and the corporates have seeked for virtual technology and resources to continue meetings and lectures. This calls for a dire need of video conferencing with promising video quality and user experience. However, due to the high bandwidth requirements of video conferencing, it is common to experience blurry images and video lags. We propose to build an AI model that refines the video at the receiver's end to give best video resolutions. We model this problem as a super resolution task on videos and perform a detailed comparative study of popular Single Image Super Resolution (SISR) models both qualitatively and quantitatively. We also performed a detailed analysis of image assessment metrics and loss functions popularly used in literature.

ACKNOWLEDGMENT

We wish to express our gratitude towards Prof. J. Jennifer Ranjani, CSIS Department, BITS Pilani for her enthusiastic support, cooperation, and help. Her constant mentorship and useful critiques have helped our progress in the project and provided us with a great learning experience.

We also wish to thank the CSIS Department of BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI for facilitating the course work which gave us an opportunity to have an exposure in the domain of academic research.

TABLE OF CONTENTS

1. SUPER RESOLUTION

Super-resolution is the process of obtaining high-resolution (HR) images/videos from the corresponding low-resolution (LR) images/videos. It is in general an ill posed problem as there can exist numerous HR outputs of a given LR input.

Fig. 1 : Single Image super resolution task

Traditionally, the task was achieved using various image processing interpolation techniques such as Nearest Neighbors, Bilinear interpolation, Bicubic interpolation etc. However these techniques were lacking and couldn't produce visually appealing images. With the evolution of Deep Learning for the task, the images/videos obtained are more visually pleasing than that of these traditional interpolation techniques.

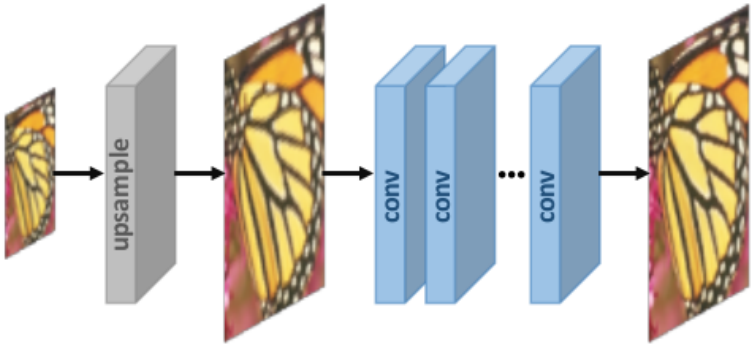


Fig 2: Using Deep Learning Architectures for Super Resolution

2. EDSR & MDSR: Enhanced Deep Residual Networks

The EDSR architecture proved to be a great leap forward in super resolution. Unlike the previously existing networks, it removed batch normalization which saves 40% memory. The authors of this paper also introduced the use of self ensemble techniques (EDSR+ & MDSR+) for better generalization. Multi-scale image upscaling using a single model.

2.1. EDSR

The EDSR architecture built upon the previously existing SRResNet architecture, which used residual blocks with Batch Norm in every block, which EDSR got rid of. The architecture originally had 32 residual blocks each having 2 convolution layers with 256 feature channels and relu activation function. It also had a global skip connection to ensure that information loss is minimized. Also, a scaling factor (of value 0.1) was added to each residual block to ensure training stability. Finally, the upsampling is done using the pixel shuffling technique.

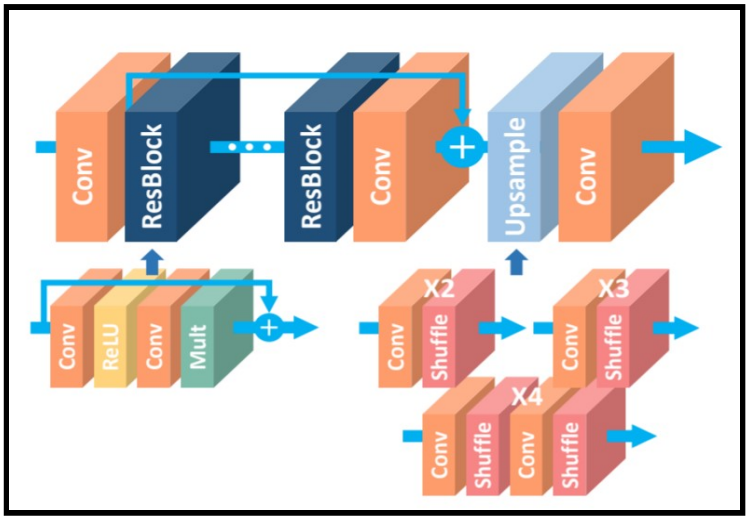


Fig. 3 : The architecture of EDSR single-scale SR-network

2.2. MDSR

The architecture for MDSR is a slight variant of EDSR and uses a 3-module structure

1. Pre-processing (scaling factor dependent)
2. Main processing
3. Upscaling (scaling factor dependent)

The pre-processing and upscaling unit trains the model for each of the scales simultaneously while training. The other implementation details remain similar to EDSR.

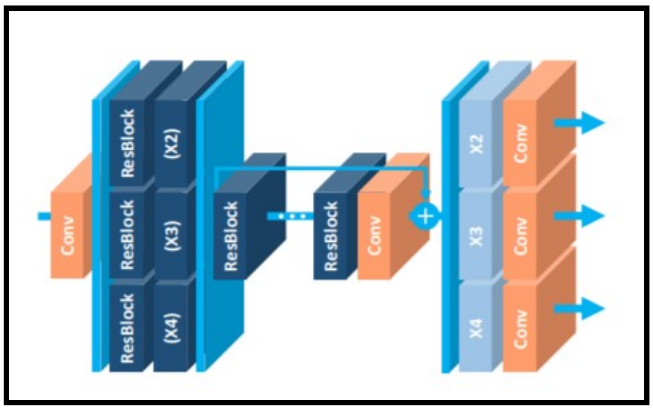


Fig. 4 : The architecture of MDSR multi-scale SR-network

2.3. LOSS FUNCTION

Both EDSR and MDSR use the L1 loss function to optimize the network parameters. The authors were among the first ones to replace the then prevalent L2 loss with L1 loss.

2.4. LIMITATIONS

Even though it was a great advancement in image super resolution, the perceptual quality of images obtained is not impressive. The images obtained are very smoothened out which is understandable as it tries to minimise the L1 loss. Another drawback is that the architecture is very large and has more than 43 million parameters (for EDSR) which makes it very slow and practically infeasible for use in video super resolution.

3. RCAN: Residual Channel Attention Networks

The RCAN model was built on top of the EDSR model. Its most important contribution is the introduction of Channel Attention to capture the more informative features and exploiting the interdependencies among feature channels. RCAN produces images with high PSNR values and is one of the state of the art architectures for PSNR maximizing networks.

3.1. RESIDUAL-IN-RESIDUAL STRUCTURE OF RCAN

RCAN has a Residual in Residual structure, meaning it has Residual Groups (RG units) which themselves have residual blocks inside(RCAB). The authors originally trained the model with 10 Residual Groups (RG), each containing 20 RCAB units each.

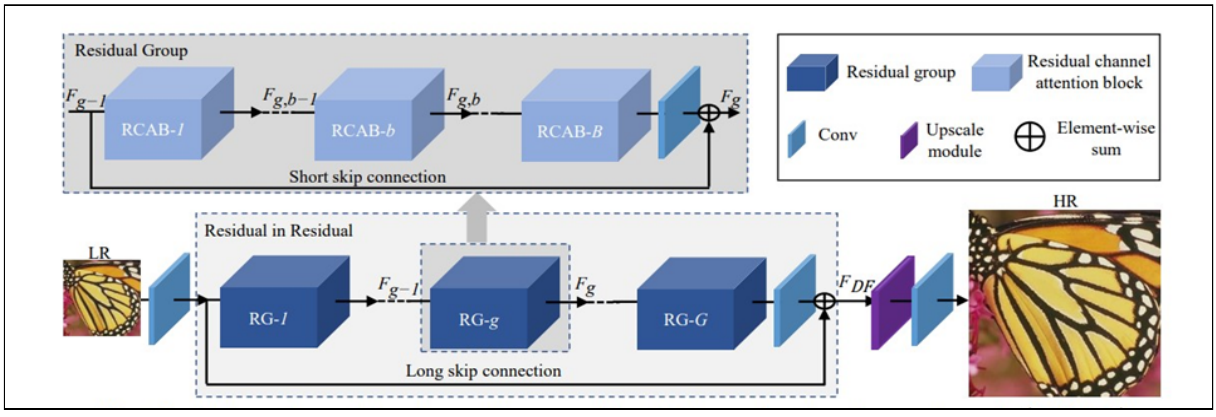


Fig. 5 : The architecture of RCAN network

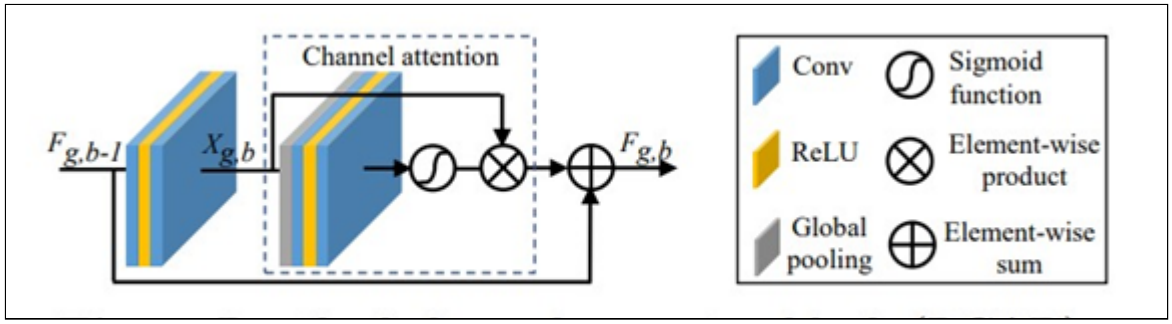


Fig. 6 : Residual Channel Attention Blocks (RCAB) structure

3.2. RESIDUAL CHANNEL ATTENTION BLOCKS (RCAB)

Each RCAB unit consists of 2 convolutional layers followed by Channel Attention. The authors argued that such channel attention allowed the network to transfer information between the channels. The authors empirically showed that simply stacking residual blocks to increase the depth of the architecture wasn't helpful for performance improvement, and to achieve better results, Channel Attention technique was helpful.

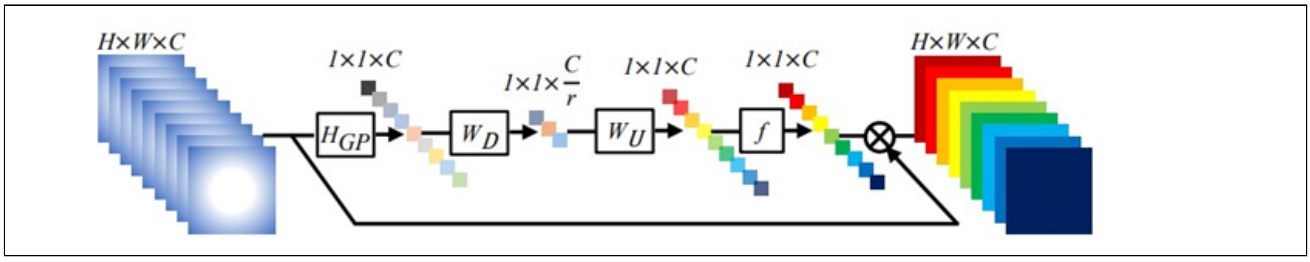


Fig. 7 : Channel Attention

3.3. LOSS FUNCTION

The RCAN model too uses the L1 loss defined as:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \left\| H_{RCAN}\left(I_{LR}^{i}\right) - I_{HR}^{i} \right\|_1 ,$$

where H_{RCAN} is the network output and I_{HR} is the ground truth HR image.

3.4. LIMITATIONS

Even though images produced by RCAN have high PSNR values, the images obtained are not perceptually appealing.

4. GAN Based Approach to Super Resolution

GANs are a class of AI algorithms used in Unsupervised Machine Learning. GANs are deep neural network architectures composed of two networks (Generator and Discriminator) pitting one against the other (thus the "adversarial").

While deep convolutional networks have been quite popular for computer vision tasks, GANs have recently been proven to solve tasks like the super resolution by looking at it as an Image generation problem. The GAN framework is integrated to denote if the patch created by the generator is similar to a ground truth set of high-resolution patches. The error of the generator network is then calculated through the adversarial loss, as well as the perceptual loss.

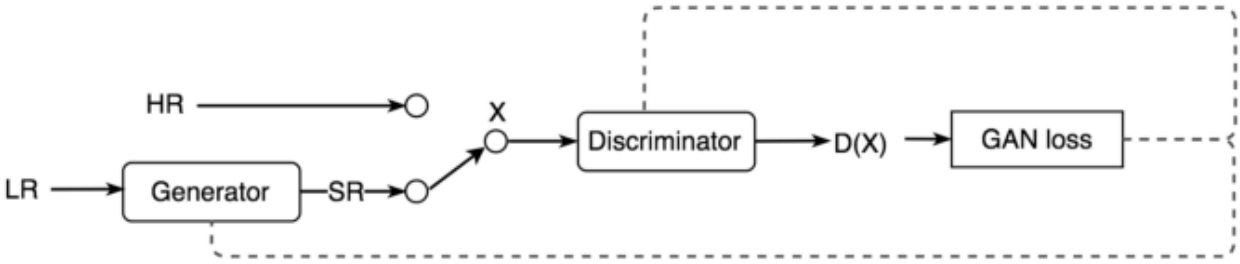


Fig 8. GAN based Approach to Super Resolution

As shown in Fig. 8 above, in a GAN based model for super resolution, during the training process, A high-resolution image (HR) is downsampled to a low-resolution image (LR). A GAN generator upsamples LR images to super-resolution images (SR). We use a discriminator to distinguish the HR images and backpropagate the GAN loss to train the discriminator and the generator.

5. SRGAN: Super Resolution GAN

SRGAN was the first GAN based approach to Single Image Super Resolution (SISR) task. While the traditional DCNN approaches were both fast and accurate, they failed to recover finer texture details from the low resolution images. Previous work had largely focused on minimizing the mean squared reconstruction error resulting in high peak signal-to-noise ratios(PSNR) means we have good image quality results, but they were often lacking high-frequency details and were perceptually unsatisfying as they were not able to match the fidelity expected in high resolution images.

5.1. MODEL ARCHITECTURE

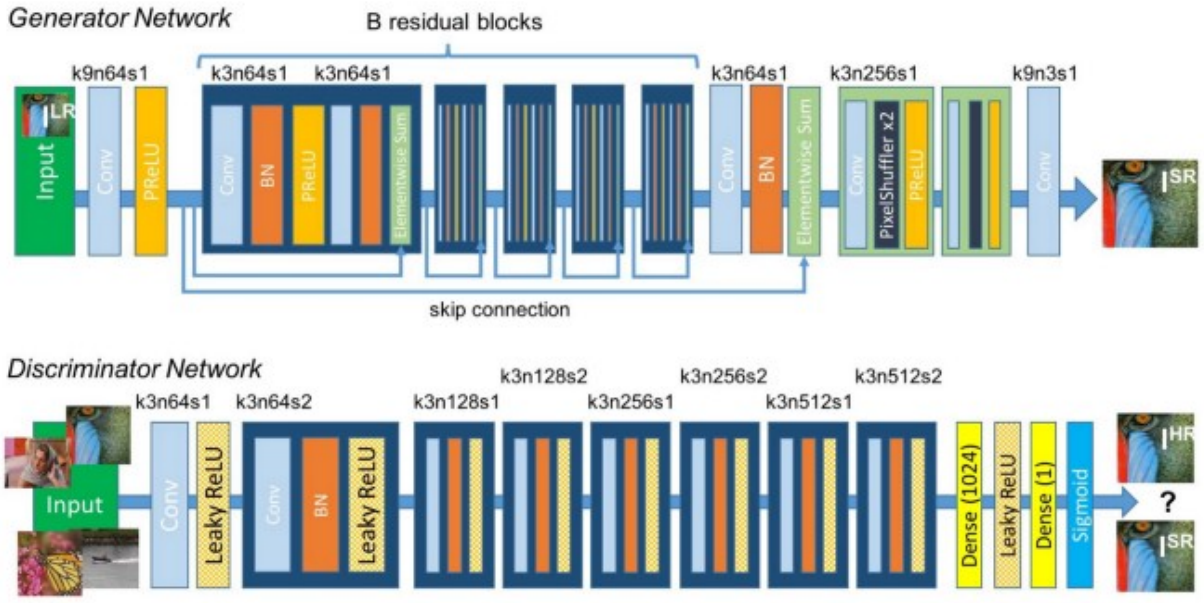


Fig 9: SRGAN model architecture

Fig. 9 above shows the network design for the generator and the discriminator. It mostly consists of convolution layers, batch normalization and parameterized ReLU (PReLU). The generator also implements skip connections similar to ResNet.

5.2. LOSS FUNCTIONS

SRGAN introduces a novel perceptual loss to ensure the finer details in the super resolved image and perceptual similarity between the generated image and ground truth. The perceptual loss consists of content loss and adversarial loss.

$$l^{SR} = \underbrace{\underbrace{l_{\mathrm{X}}^{SR}}_{\text{content loss}} + \underbrace{10^{-3} l_{Gen}^{SR}}_{\text{adversarial loss}}}_{\text{perceptual loss (for VGG based content losses)}}$$

Adversarial loss pushes the solution to the natural image manifold using a discriminator network that is trained to differentiate between the super-resolved images and original photo-realistic images.

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

Content loss keeps perceptual similarity instead of pixel wise similarity. This will allow us to recover photo-realistic textures from heavily down sampled images. It defines content loss as the VGG loss based on the ReLU activation layers of the per-trained 19 layer VGG network. VGG loss is calculated using the euclidean distance between the feature representations of a reconstructed image and the reference image.

$$l_{VGG/i.j}^{SR} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2$$

5.3. LIMITATIONS

While the SRGAN provided state-of-the-art results, due to the batch normalisation technique and the proposed model architecture, the reconstructed images suffered from artifacts.

6. ESRGAN: Enhanced Super Resolution GAN

ESRGAN is the enhanced version of the SRGAN where the authors of the ESRGAN tried to enhance the SRGAN by modifying the model architecture and loss functions. The main architecture of the ESRGAN is the same as the SRGAN with some modifications. ESRGAN has Residual in Residual Dense Block(RRDB) which combines multi-level residual network and dense connection without Batch Normalization.

6.1. RESIDUAL-IN-RESIDUAL BLOCK

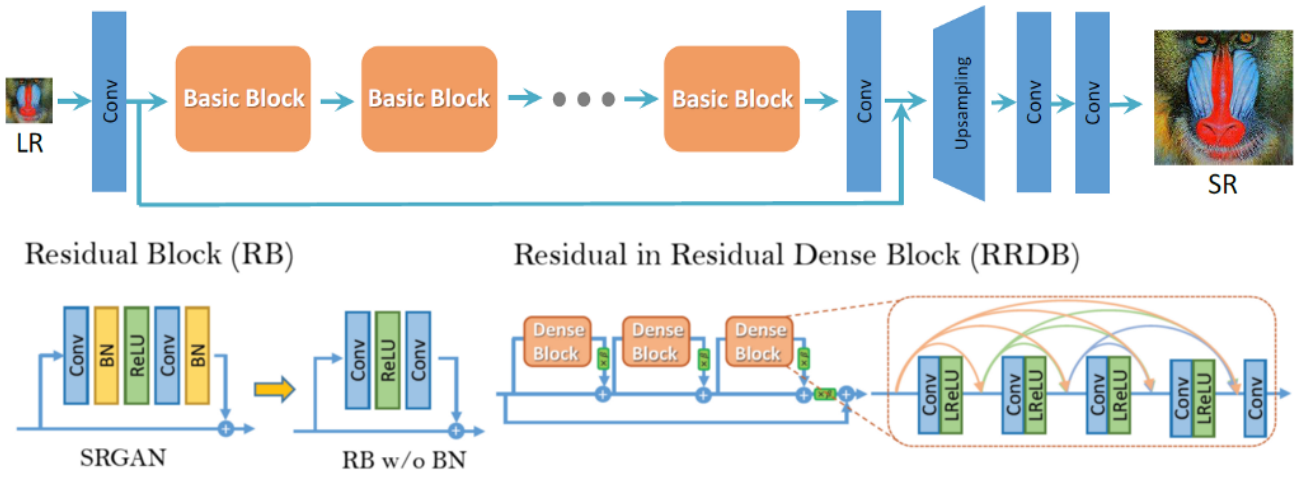


Fig 10: RRDB Architecture

Unlike SRGAN which uses Residual Block as the basic block unit in the model architecture, ESRGAN uses Residual-In-Residual Blocks as the basic unit of the architecture. An RRDB is a series of Dense Blocks (Residual Blocks without batch normalisation) with short and long skip connections and is used to capture the complex features of the low resolution images. While removing batch normalisation from the model architecture has proven to increase performance and reduce the computational complexity in different PSNR-oriented tasks, including Super resolution, it also resolves the problem of artifacts in the reconstructed images.

6.2. RELATIVISTIC GAN

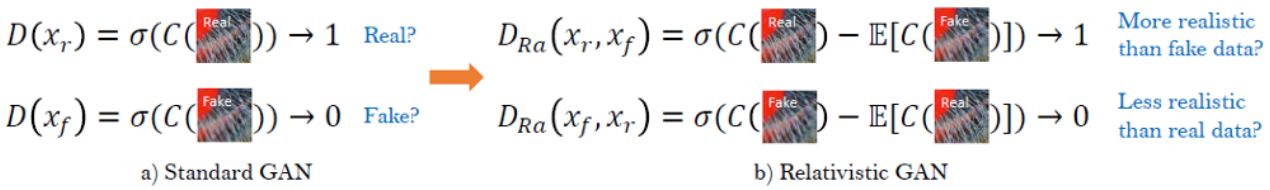


Fig 11: Standard discriminator vs Relativistic discriminator

The standard discriminator D in SRGAN estimates the probability that one input image x is real and natural. In contrast, a relativistic discriminator as used in ESRGAN tries to predict the probability that a real image xr is relatively more realistic than a fake one xf. Thus, the new discriminator and the generator loss are defined as:

$$L_D^{Ra} = -\mathbb{E}_{x_r}[\log(D_{Ra}(x_r, x_f))] - \mathbb{E}_{x_f}[\log(1 - D_{Ra}(x_f, x_r))].$$

$$L_G^{Ra} = -\mathbb{E}_{x_r}[\log(1 - D_{Ra}(x_r, x_f))] - \mathbb{E}_{x_f}[\log(D_{Ra}(x_f, x_r))],$$

6.3. LOSS FUNCTIONS

ESRGAN uses the same perceptual loss as the SRGAN however with the use of the features before the activation layers for calculating VGG loss. Since, the sparse activation provides weak supervision and thus leads to inferior performance and using features after activation also causes inconsistent reconstructed brightness compared with the ground-truth image, this approach helps in more accurate images and without artifacts.

Thus, the loss function is given as:

$$L_G = L_{\text{percep}} + \lambda L_G^{Ra} + \eta L_1,$$

7. COMPARATIVE STUDY OF SR Models

7.1. DATASETS

We train the models on DIV2k dataset consisting of 1000 2K resolution high quality images, generally split as 800 (training) + 100 (validation) + 100 (testing) and compare by testing on Set5, Set14, BSD100, Urban100 datasets.

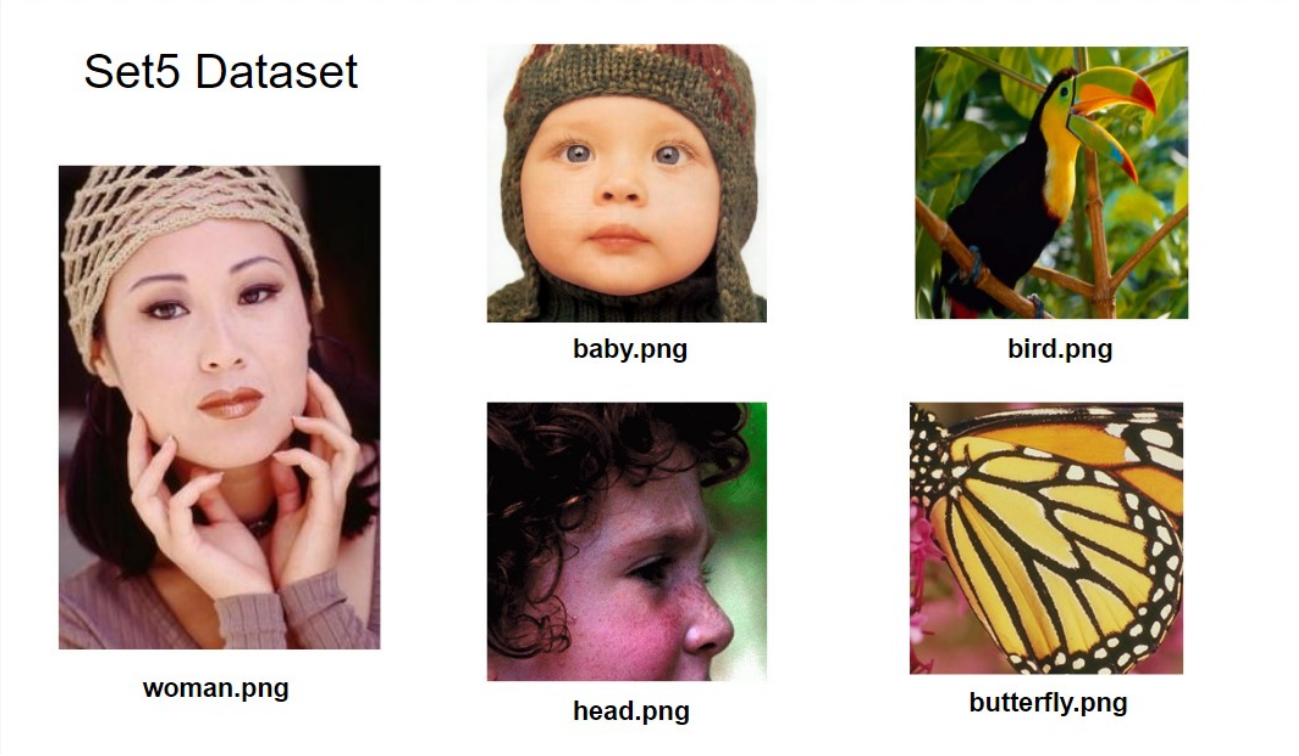


Fig 12. Set5 Dataset (Used for Testing)

7.2. MODEL COMPARISONS

We perform a detailed comparison of quantitative and qualitative results of the four most popular models in SR tasks:

1. EDSR
2. RCAN
3. SRGAN
4. ESRGAN

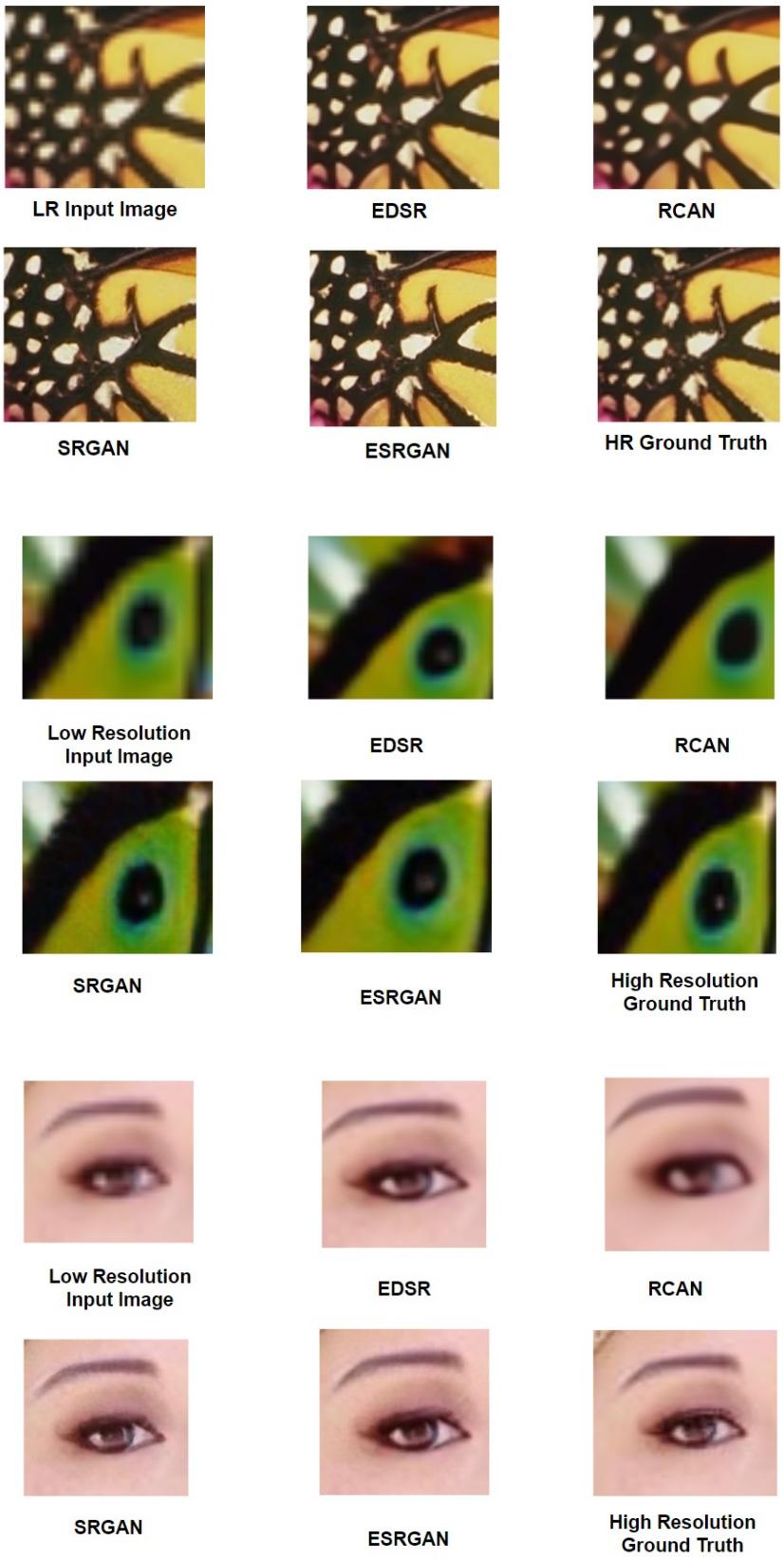
	EDSR	RCAN	SRGAN	ESRGAN
Model	Used ResNets without BN to get SR image of the same distribution as input.	Used channel attention to learn inter-channel feature dependencies	GAN based approach with Standard Discriminator	Uses RRDB as basic units and relativistic discriminator
Batch Normalization	No	No	Yes	No
Loss Function	L1 Loss function	L1 Loss function	Perceptual Loss for generator, Min Max Loss for discriminator	Perceptual Loss (before activation layers)
Author's results on Set5 dataset	PSNR: 32.46 SSIM: 0.8968	PSNR: 32.73 SSIM: 0.9013	PSNR: 29.40 SSIM: 0.84	PSNR: 32.73 SSIM: 0.9011

Table 1: Brief review of the four models

Hyper Param	EDSR	RCAN	SRGAN	ESRGAN
Training Dataset	DIV2K	DIV2K	ImageNet	DIV2K
Depth	32 ResBlocks	10 RGs (20 RCABs in each)	16 ResBlocks	23
Upscaling Factor	2x/3x/4x/8x	2x/3x/4x/8x	4x	4x

Table 2: Comparison of Hyperparameters

7.3. QUALITATIVE COMPARISON



7.4. QUANTITATIVE RESULTS

Model Name	PSNR (Ours)	PSNR (Author)	SSIM (Ours)	SSIM (Author)
EDSR	32.095	32.46	0.892	0.8968
RCAN+	32.738152	32.73	0.900286	0.9013
SRGAN	29.373820	29.40	0.841682	0.8472
ESRGAN	32.730021	32.73	0.901131	0.9011

Table 3: Quantitative Results

7.5. CONCLUSION

1. While SRGAN argues that batch normalization leads to faster convergence, ESDR argues that BN leads to inflexibility in the output range with more parameters to learn and ESRGAN avoids BN due to artifacts in the output images. Thus, we may conclude removing batch normalisation in SR tasks resolves artifacts and makes the model simpler.
2. From all the models, we have concluded that using residual connections have outperformed the benchmarks. So it is a good idea to keep residual/dense nets in an SR architecture along with short and long skip connections.
3. It has been suggested that deeper networks perform better than wider ones (for Super Resolution), so residual nets would help train better.
4. Using a combination of PSNR oriented loss function like L1 loss and perceptual loss helps in optimising both PSNR & perceptual quality of the SR image.

8. ANALYSIS OF IMAGE QUALITY ASSESSMENT (IQA) METRICS

8.1. FULL REFERENCE METRICS

Full-reference algorithms compare the input image against a pristine reference image with no distortion. These metrics try to capture the similarity of the reconstructed image with the ground truth in pixel space. Some of the popular metrics include: PSNR, SSIM, IFC, VIF.

8.2. NO REFERENCE METRICS

No-reference algorithms use statistical features of the input image to evaluate the image quality. These metrics have been recently proposed to capture the perceptual quality of the reconstructed image. Some of the metrics include NIQE, BRISQUE and PIQE.

8.3. ANALYSIS OF IQA METRICS

	VIF	GMSD	PIQE	NIQE	BRISQUE
EDSR	0.516	0.050	77.066	5.793	42.2498
RCAN	0.535	0.046	82.952	6.1899	43.181
SRGAN	0.422	0.069	24.857	4.362	21.620
ESRGAN	0.45	0.061	35.938	4.362	25.828
HR	-	-	41.049	4.850	34.079

Table 4: Numerical Values of IQA Metrics for different models

As seen from the results in table 4 and comparing with corresponding qualitative results obtained, we can conclude that the no reference metrics are more suitable for evaluating images having better perceptual quality and feel more natural.

9. ANALYSIS OF SR LOSS FUNCTIONS

Loss Function	Conclusion
MSE Loss	Maximises PSNR however restricts models to pixel space only and results in low quality images.
L1 Loss	Results in good quality images while also restricting the model in pixel space thus maximising PSNR values. Since, our focus is on constraining to perceptual quality than pixel quality, we can use L1 only to instantiate our model in a 2 step training process or as a part of our loss function
VGG Loss	A good placeholder to capture the perceptual quality loss and used extensively in GANs. Constraints the model on feature space rather than pixel space
LPIPS Loss	A recently introduced perceptual loss function that minimises the LPIPS score for the images. Outperforms VGG Loss for extreme SR (16x or more upscaling)

Table 5: Popular Loss Functions in literature

10. CONCLUSION

We proposed to resolve the bandwidth limitations in video conferencing by providing super resolution of videos at the receiver's end. We performed a detailed qualitative and quantitative analysis of different popular architectures in single image super resolution task namely EDSR, RCAN, SRGAN and ESRGAN and concluded that GAN based approaches perform better than the DCNN based approaches. Models without batch normalisation and with residual networks as part of their architectures give faster convergence and more accurate results with lesser architects. We also performed an analysis of different IQA metrics and concluded that the No-Reference IQAs seem to have a correlation with the human perception of good quality images. Thus, loss functions that would try to optimise the NR-IQAs might give better super resolved images. We also analysed different loss functions as used by many authors in the super resolution literature and concluded that a combination of pixel wise loss like PSNR oriented loss functions and perceptual loss functions result in both pixel wise similarity perceptually good images.

11. REFERENCES

1. C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In ECCV 2014.
2. Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
3. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPRW (2017)
4. Mittal, A., Soundararajan, R., Bovik, A.C.: ‘ Making a “completely blind” image quality analyzer’, IEEE Signal Process. Lett., 2013,
5. Wang, Xintao, et al. "Esrgan: Enhanced super-resolution generative adversarial networks." Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018.
6. WANG, Z., CHEN, J., AND HOI, S. C. Deep learning for image super-resolution: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).
7. Wanjie Sun and Zhenzhong Chen. Learned image downscaling for upscaling using content adaptive resampler. IEEE Transactions on Image Processing, 29:4027–4040, 2020.
8. Yuan, Yuan, et al. "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018.
9. Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In ECCV 2018
10. Zhang, Wenlong, et al. "Ranksrgan: Generative adversarial networks with ranker for image super-resolution." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.