Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Exploring the Effectiveness of SRGAN As a Video Upsampler

[1]Lakshay Vashisth, [2]Aayush Bajaj, [3]Abhishek Jha, [4]Prof. Ajay Kaushik
Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India;
[1]lakshaylv97@gmail.com,[2] dev.aayushbajaj@gmail.com,
[3]jhabhishek3797@gmail.com, [4]ajaykaushik@mait.ac.in

Article Info
Page Number: 1172-1177
Publication Issue:
Vol. 71 No. 3s (2022)

Article History
Article Received: 22 April 2022
Revised: 10 May 2022
Accepted: 15 June 2022
Publication: 19 July 2022

Abstract — SRGAN, a generative adversarial network (GAN) has been one of the state-of-the-art techniques for image super-resolution. The proposed perceptual loss function, which is further composed of an adversarial loss and a content loss, is able to upscale images effectively by a factor of 4. While enough research and experiments have been carried out to evaluate the performance of SRGAN on images, its potential on videos is still obscure. In this paper, our main objective is to generate high-resolution counterparts of low-resolution videos via SRGAN and evaluate the network's capabilities on the same. To quantify the model's performance, we have used two image quality assessment metrics, Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM). While PSNR has been in use for a long time and is widely considered to be a tried and tested approach, SSIM is a newer metric that is designed based on three factors i.e. luminance, contrast, and structure, in order to relate it more to the mechanics of the human visual system. Through these two metrics, we aim to gauge the effectiveness of SRGAN as a video upsampler, from the perspective of reconstructed pixel quality as well as human perception.
Index Terms—video upsampling, generative adversarial networks, super-resolution, quality assessment.

I.INTRODUCTION

Upsampling or super-resolution refers to enhancing the coarse quality of *low-resolution (LR)* images and videos to produce their much-more refined *high-resolution (HR)* counterparts. While multiple solutions exist to achieve this, Deep Learning methods have certainly gained an edge over others due to their ability to learn the relationship between input and output on being provided with enough data [1]. Super-Resolution Generative Adversarial Network (SRGAN), a deep residual network (ResNet) with skip connection works on a similar principle. With its perceptual loss, it is able to extract photo-realistic features from heavily downsampled data [2].

The motivation for writing this paper primarily arises because of two reasons. 1) Due to the prevalence of high-definition displays and fast internet connections, there is avid demand for consuming content in high resolution. However, lately, many social media websites have suffered outages because of traffic overload due to such high-quality media. While downsampling an image or a video is a fairly easy task, reconstructing that downsampled data to obtain the original high-quality media is extremely tricky. If we can create a solution for this

task, the network load can decrease and viewers can still enjoy high-resolution content. 2) In domains such as surveillance and astronomical data, where the availability of high-resolution videos can be extremely critical to the task at hand, an efficient super-resolution algorithm can find many applications and prove critical for success.

While the niche of SRGAN lies in processing images, in this work, we apply this network, as proposed in the original paper, to videos and compare the obtained one with its low-resolution counterpart. This evaluation is done with the help of Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM), to see if SRGAN can be a panacea to video super-resolution problems.

This paper is organized into 5 sections. Section 2 discusses the related works done in this domain. In section 3 we explain our approach, the data, the network architecture, and the training part of it. Section 4 contains the final results obtained. Finally, we conclude and discuss the future steps in section 5.

II.RELATED WORK

The super-resolution of images and videos has been a topic of research for a long time. Methods like the nearest neighbour, bilinear interpolation and bicubic interpolation [3] were developed for this purpose and were the algorithms of choice for some time. But with the advent of deep learning, we have seen a sharp rise in the capability of machines at the task of upscaling images and videos. Convolutional Neural Network (CNN), Deep CNN, Enhanced Deep Super Resolution network (EDSR) [4] and other deep learning methods have shown much better results than the previously mentioned conventional methods and are able to add details to the upscaled media to an extent to which the traditional methods can't, thus making them highly suitable for upsampling. GAN is a relatively new type of deep learning technique which has shown great results as an image up-sampler than previously mentioned deep learning techniques thus making it a potential candidate for the task of video SR as well.

III.METHOD

A.Data

The dataset used for training the SRGAN is the DIV2K dataset [5]. It contains 1000 2K resolution images out of which 800 images are designated for training, 100 for validation and rest 100 for testing. The images in this dataset are highly diverse consisting of humans, objects, animals, monuments, buildings and pristine nature shots and they are set in various habitats around the world such as deserts, mountains, forests, cities, and villages among many others.

The dataset also consists of corresponding LR images for all the HR images. The available LR images are already downscaled by x2, x3, and x4 factors using Bicubic interpolation. We used the HR images and the corresponding x4 downscaled LR images for our training.

The diversity of the dataset and the availability of already downscaled LR images in organized folders make the DIV2K dataset a highly suitable and therefore popular choice for the task of training Deep Learning models for Super-Resolution of real-life images and videos

of all kinds.

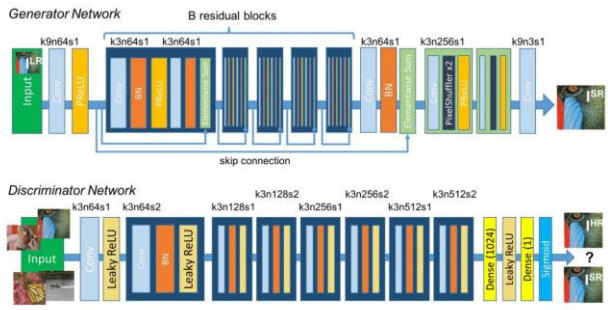

Fig. 1
Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.[2]

B.Network

Inspired by [2] we use two convolutional layers with small 3×3 kernels and 64 feature maps followed by batch-normalization layers [6] and ParametricReLU [7] as the activation function. We increase the resolution of the input image with two trained sub-pixel convolution layers as proposed by [8].

To discriminate real HR images from generated SR samples we train a discriminator network. The architecture is shown in fig. 1. We used LeakyReLU activation ($\alpha = 0.2$) and avoid max-pooling throughout the network. It contains eight convolutional layers with an increasing number of 3 × 3 filter kernels, increasing by a factor of 2 from 64 to 512 kernels as in the VGG network [9]. 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 obtain a probability for sample

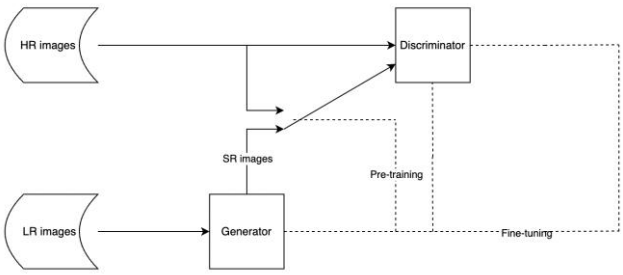

Fig. 2

classification.

C.Training

The network's training happens in two stages:

1)Generator pre-training

First the generator is trained alone on 800 transformed LR images, the training dataset, for 100,000 steps. Ideally, we wanted this number to be close to a million but due to hardware constraints we had to limit ourselves.

Corresponding HR images are used to calculate MSE loss to train our generator. This step is primarily done to increase mode coverage and enhance the capabilities of our generator.

2)Generator fine-tuning

Once the 1st step gets completed, we tune these pre-trained weights by training the generator along with the discriminator i.e. the entire GAN for around 200,000 steps. This time the discriminator gets fed with the HR counterparts of LR images in the training dataset to calculate the perceptual loss and enable the generator to get better at producing realistic outputs. Fig. 2 demonstrates the entire training workflow.

D.Processing Video

Rather than working on the entire video as a whole, we decompose the LR video into its constituent frames. These LR frames are then used as input for the trained generator to obtain corresponding SR images. We then compile all the SR frames into an SR video. It can be said that video upsampling is equivalent to upsampling a series of images.

IV.RESULTS

As discussed in the last section, we evaluated our trained model on individual frames of the input videos. A decent improvement in quality was observed in the obtained frames. A few such comparisons can be seen in fig. 3 and fig. 4. LR column contains the actual frame while the SR (GAN) column contains its upsampled counterpart.

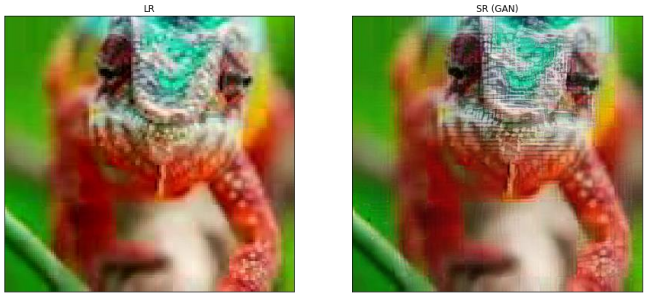

Fig. 3

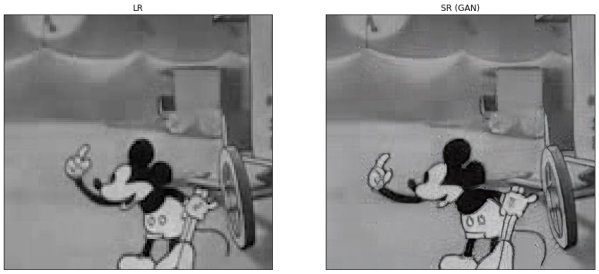

Fig. 4

To quantify the improvements observed above, we downsampled a high-resolution video by 67% and generated its SR counterpart using our trained network (fig. 5). Subsequently, we calculated the PSNR and SSIM score for each pair of the original and corresponding upsampled frame (fig. 6).

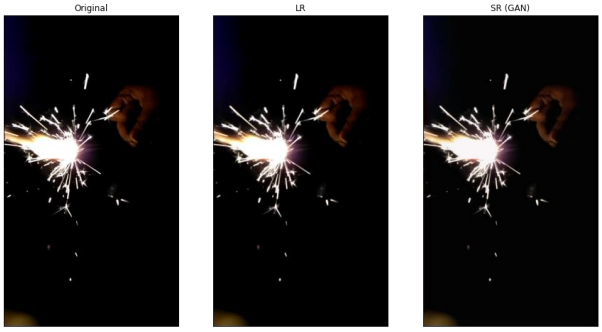

Fig. 5

We get an average PSNR score of 24.24 and an average SSIM score of 0.678. While these scores are not exceptionally high, they demonstrate that if trained appropriately, SRGAN can be very well used as a video up-sampler by collating the generated high-quality frames.

V.CONCLUSION

We have described a deep residual network to extend SRGAN capabilities to a video sample using a looping mechanism. The baseline PSNR and SSIM scores of 24.24 and 0.678 are significant enough to prove the model's capabilities on video data. More sophisticated training on a compute-heavy machine can yield better scores on the same dataset. We have confirmed that SRGAN reconstructions for large upscaling factors (4×) on a video dataset. Furthermore, a frame reference methodology proposed in [10] provides notably better results than our methods.

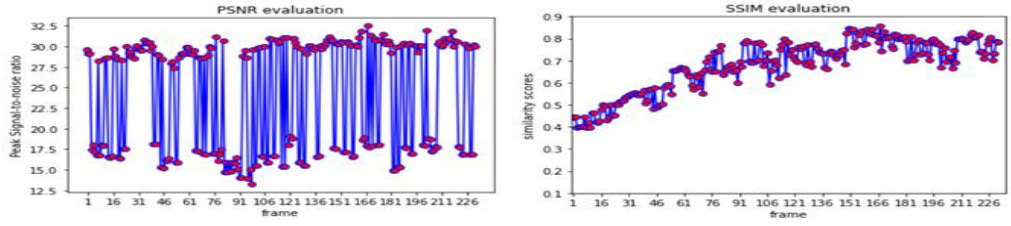

Fig. 6

REFERENCES

1. Anwar, S., Khan, S. and Barnes, N., 2021. A Deep Journey into Super-resolution. *ACM Computing Surveys*, 53(3), pp.1-34.
2. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
3. Savagave, A. and Patil, A., 2014. *Study of Image Interpolation*. [online] Ijiset.com. Available at: <https://www.ijiset.com/v1s10/IJISET_V1_I10_71.pdf>.
4. Lim, B., Son, S., Kim, H., Nah, S. and Lee, K., 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*
5. Data.vision.ee.ethz.ch. 2022. *DIV2K Dataset*. [online] Available at: <https://data.vision.ee.ethz.ch/cvl/DIV2K/>.
6. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of The 32nd International Conference on Machine Learning (ICML), pages 448–456, 2015.
7. Varun, B. N. ., S. . Vasavi, and S. . Basu. “Python Implementation of Intelligent System for Quality Control of Argo Floats Using Alpha Convex Hull”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 5, May 2022, pp. 60-64, doi:10.17762/ijritcc.v10i5.5554.
8. K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In IEEE International Conference on Computer Vision (ICCV), pages 1026–1034, 2015.
9. W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1874–1883, 2016.
10. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR), 2015.
11. Chadha, A., Britto, J. and Roja, M., 2020. iSeeBetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks. *Computational Visual Media*, 6(3), pp.307-317.
12. Tume-Bruce, B. A. A. ., A. . Delgado, and E. L. . Huamaní. “Implementation of a Web System for the Improvement in Sales and in the Application of Digital Marketing in the Company Selcom”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 5, May 2022, pp. 48-59, doi:10.17762/ijritcc.v10i5.5553.