C.A.R.R.O.L
Context Aware Super Resolution

Andrew Eurdejian*, Ankur Gupta†, Revant Mahajan‡, and Akash Shaji§

*Worcester Polytechnic Institute
Email: ageuredjian@wpi.edu
†Worcester Polytechnic Institute
Email: agupta4@wpi.edu
‡Worcester Polytechnic Institute
Email: rmahajan@wpi.edu
§Worcester Polytechnic Institute
Email: ashaji@wpi.edu

***Abstract*—We addressed the classic computer vision problem of super-resolution. Most modern methods of super-resolution have achieved very impressive results with scaling factors of around 2-4x when using Generative Adversarial Networks (GAN) on the entirety of single images. Other methods have been known to produce high Pixels-to-Signals Noise Ratio (PSNR) values. However, these methods dont necessarily account for the perceptual features of the objects. This leads to certain distortions while converting these images to super-resolution. This paper proposes an alternative approach to conducting super resolution on images. The algorithm is set up to focus on the important parts of the image that are deemed valuable to a human viewer. This is done by extracting those subjects from the image and running a class-conditional Super-Resolution GAN (SRCGAN) on it. Our model, trained on a custom data set generated from source mp4 videos, was able to produce images where the important subjects were more emphasized in the image. We generated our own dataset in order to account for the fact that the segmentation model we are using is limited to certain types of classes. The fidelity of the generated images is tested by using PSNR against the ground truth.**

I. Introduction

Super resolution is a classic problem in computer vision involving upscaling a low resolution image in order to obtain a clearer or more detailed image which has seen a wide range of use from raw image enhancement to feature detection within images [1]–[3]. In general, super resolution has been quite successful when implemented with a Generative Adversarial Network (GAN) [1]. Our goal in this paper is to apply this type of super resolution algorithm with an added class-conditional feature to various individual subjects in an image to increase their visibility.

A problem with current methods of super resolution is that normal GANs and deep learning algorithms have a lot of trouble handling images with multiple subjects. The output of these algorithms have the possibility of being entirely unrecognizable compared to the original image [4]. Another problem with current super resolution approaches is that they operate on an entire image, the issue being that the difference in resolution between the foreground and background doesnt necessarily change that much. This means that there will still be the same lack of focus on the background in the new image as there was in the original. The last problem this paper will address is that common super resolution algorithms only use the original images pixels as input data to run super resolution. While this method has been shown to work quite well, we believe it can be improved by adding a class condition.

One of the main decisions we had to make was how to go about building the super resolution algorithm. Since our focus isnt so much the super resolution algorithm itself and there has already been extensive research on the performance of various algorithms, we did some research to find an algorithm that is accepted by the majority of the community as the best for our purposes which is a conditional GAN [1]. We believe that in employing an SRCGAN with the supporting processes described, the product of the complete process will be a more meaningful image that puts emphasis on the more important parts of the original image.

II. Background

Super-resolution aims at generating a high-resolution (HR) image from a low-resolution (LR) image. Many different approaches have been tried out to carry out this task. Some of them are - minimizing mean squared error(MSE), Super Resolution Convolutional Neural Network (SRCNN) and Generative Adversarial Network (GAN). A Supervised Learning Algorithm is used to minimize the cost function which is the MSE between the generated image and ground truth. This itself is a convenient way to maximize the peak signal to noise ratio (PSNR). As shown by Ledig et al. [3], this is not the most accurate way of comparison. There is a limitation to MSE capturing image features and perceptual relevant details as they are based on pixel-wise image difference. Interpolation is one other simpler technique that is used for super resolution [5]. An image is divided into multiple mathematical subspaces. Within these subspaces, interpolation is performed to fill in the missing pixel data and the final image is reconstructed from these subspaces. These approaches lack to account for

a fundamental property of images. Pixels on a standalone basis dont represent much. On the other hand, they represent features of the objects in the image when viewed with their neighbors. These approach lacks the acknowledgment of these features and does not produce photo-realistic images [6].

A better approach then this is to use a modified version of Convolutional Neural Networks(CNN). CNN is the state of the art methods for image detection and classification [7], [8]. CNNs comprise more than one convolutional layers followed by one or more fully connected layers to classify images. The architecture is designed to take advantage of the 2D structure of the image and the fact that images can be classified by features which are represented by the neighboring group of pixels. A CNN consists of multiple convolutional layers with pooling layers separating them. One convolutional layer is formed from the previous layer by running a patch on the complete image and taking the dot product between the patch values and the image values covered. Each convolution layer consists of filters. The number of filters grows as the network grows. As the data is processed through the network, the convolutional layers get reduced in length and width but their depth increases. First few layers are able to identify low-level features like lines. Moving forward, mid-level features like arcs, curves, etc. are recognized. Finally, deep layers are able to recognize complex features. The approach of Super Resolution Convolutional Neural Network(SRCNN) [9] considers a convolutional neural network with an end-to-end mapping between low- and high-resolution images. It consists of three main steps: patch extraction and representation, non-linear mapping, and image reconstruction [9].

Another widely accepted approach is to use Generative Adversarial Network for super-resolution [3]. For Super Resolution Generative Adversarial Network(SRGAN), the goal is to train a generating function G that estimates HR for the given LR image. The generative function is trained with a discriminator D that is trained to distinguish super-resolved images from real images. The goal is to train them so that D is no longer able to distinguish between the generated super-resolution images and the ground truth. Eventually, the generator learns to create images that are similar to the ground truth.

The problem with the above-mentioned methods is that they look at the image as a whole. The intuition behind our idea is that each object in the image has its own unique features and those features should be maintained when we run super-resolution algorithms on them. We take sample images and run them through a Fully Convolutional Network(FCN). FCNs are the state of the art technique used for image segmentation. The output of the FCN is the segmented image. Using this output, images for each of the objects belonging to different classes are formed. These images are then given input to a Conditional GAN(CGAN). CGAN are a form of GAN which creates images based on the context provided. For example, if the input to this CGAN is 9, it will produce images that it thinks resembles 9. Each super-resolution image is then stitched back together to produce the complete super-resolution image. This process is explained in detail below in the methods.

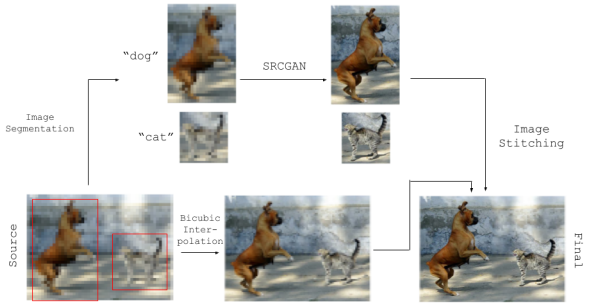

Fig. 1: An example of how our technique would work end-to-end. The source image (bottom right) would first be put through an Image Segmentation algorithm to determine the different subjects in the image. Once the subjects and their classes have been determined, they are then run through the SRCGAN. In the meanwhile, the whole image is run through a bicubic interpolation. Once the higher quality images have been formed, they are merged together through the image stitching algorithm. Once the images have been stitched together, the final image will be formed.

III. Methods

A. Overview

Most approaches to the Super Resolution problem (especially GANs) are run on datasets such as MNIST and CelebA. While those datasets excel at showing the differences in image quality, they are not very representative of the real world. In this problem, we attempt to tackle the super resolution of complex images by breaking down the process into multiple steps:

1) Segment the images into known classes
2) Run the segmented images through a Super Resolution Conditional Generative Adversarial Network
3) Run the non-segmented parts of the image into a standard bicubic interpolation function
4) Merge the images together through image stitching

This process is illustrated by Figure 1. By breaking the process up into smaller, more manageable chunks, one is able to take advantages of the different top-performing models. The following sections will go into greater detail on the reasons behind choosing the various methods and how they were implemented.

B. Dataset Selection

We opted to select datasets with one to three subjects in it in order present where our proposed solution would shine. The datasets we selected for training were created by splitting open source HD footage frame by frame. This allowed us to create large easy to label datasets with ease.

C. Segmentation

Fully Convolutional Network: Fully Convolutional Network (FCN) is one of the states of the art techniques for semantic

segmentation. Semantic segmentation [10] refers to the process of associating each pixel in an image with a class label such as animals, roads, buildings, etc. FCNs build up on CNN. CNNs are very good for image classification data but they do not retain any spacial information with convolutions. All the features are detected but their locations in the image are not maintained. FCNs improve on this by just having convolutional layers. A typical FCN network consists of an encoder block, 1x1 convolutional layer, decoder block. The encoder block is pretty much the same as a CNN without the fully connected layers. It downsamples the image with each convolution and increases the feature maps. 1x1 convolutions are used to change the filter dimensionality (either increase it or decrease it) before sending it to the decoder block. The decoder block consists of transposed convolutional layers often called deconvolutional layers. This part of the network deals with upsampling the image to its original size. The output at the end is a image associating each pixel with a class.

We decided to use Googles pre-trained DeepLab FCN [11]. It is able to perform this network to get a semantic segmentation output for the backgrounds and object of interest. These objects of interest are then cropped out as individual images to be fed into the super-resolution algorithm. The intact background is passed forwards as well to the further methods.

D. Super Resolution

1) General Adversarial Networks: General Adversarial Networks (GANs) were first introduced by Goodfellow et al. in 2014 [12]. GANs were a new framework that trained generative models using an adversarial process. Under this framework, there are two models: a generative model G, which trains off of the data distribution, and a discriminative model D, which determines whether a sample is real or has been generated. Both of the models are trained simultaneously, resulting in a two-player game of sorts. In this game, G generates a sample and D tries to guess whether or not the sample is real or not. In order to beat D, G has to create samples that are closer to the data set. Conversely, in order to beat G, D has to become more selective on what is a truth or not. This back and forth process can intuitively be expressed as a generic 2-player minmax game:

$$\min_G \max_D (D, G) = E_{x p_{data}}[\log(D(x))] + E_{z p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

2) Conditional GANs: Mirza and Osindero improved upon Goodfellow et al.s vanilla GAN by simply adding auxiliary features to inputted data [13]. These conditioning features are added as inputs to both the discriminator and generator functions as well as labels to the data. These modifications can be seen in the following objective function:

$$\min_G \max_D (D, G) = E_{x p_{data}}[\log(D(x|\mathbf{y}))] + E_{z p_z(z)}[\log(1 - D(G(z|\mathbf{y})))] \quad (2)$$

In which the auxiliary feature vector is represented by $\mathbf{y}$.

Mirza and Osindero were able to successfully generated MNIST digits conditioned on both the MNIST dataset and the respective class labels. [13]

3) Super Resolution with CGANs: GANs are currently some of the top methods for super resolution and CGANs have been shown to work well in combining GANs and auxiliary information. Chen et al. combined the two ideas by using CGANs for image super resolution [1]. In their report, Chen et al. introduce two methods of introducing new features into their Super Resolution GAN (SRGAN) [1]:

1) (SRCGAN) - Add the class information as another input feature.
2) (SRGAN + Class Loss) - Create an independent classifier whose purpose is to determine if the generated image is of the correct class and factor that into the objective function.

In this project, an SRCGAN was chosen due to its demonstrated accuracy improvements over a vanilla GAN [1]. As per Chen et Al.s findings, the objective function for this model is as follows:

$$\min_{\Theta_G} \max_{\Theta_D} (D, G) = E_{I^{HR} \sim p_{train}}(I^{HR})[\log(D_{\Theta_D}(I^{HR}, \mathbf{c}))] + E_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\Theta_G}(G_{\Theta_G}(I^{LR}, \mathbf{c})), \mathbf{c})] \quad (3)$$

In which $\mathbf{c}$ is the conditional information, D and G are the updated discriminator and generator functions, and I^{HR} and I^{LR} are the high-resolution and low-resolution images, respectively.

For the creation of the SRCGAN, Linder-Norns implementation of the SRGAN [14] from Ledig et Al was used as a bootstrap due to its success in photo-realistic super resolution [3]. The class information of the sample subject was then added as another feature to the input vector and the objective function was modified to account for the change in the feature space.

While Chen et Al. discuss that using class data as auxiliary inputs in the SRCGAN is trivial, they were only working on the MNIST and CelebA datasets [1]. In this case, it is trivial to see how class information (such as person) would be a trivial feature to add, as all of the samples would share the same information [1]. However, this application hosts a variety of different classes for subjects to be in, so including the class information plays a non-trivial role.

4) Bicubic Interpolation: While GANs have made impressive advances in image super resolution, they are not without fault. As Goodfellow discusses [4], GANs tend to have issues when dealing with:

- Counting
- Perspective

- Global Structures

While these drawbacks may be a bit more trivial when working on more restrictive datasets such as MNIST or CelebA, they become more of an ordeal when dealing with complex, multi-subject images. For that reason, the SR model chosen for the base images was Lukin et al.s Bicubic Interpolation SR model from 2006 [15]. The Bicubic Interpolation SR model has seen impressive results in a scaling factor of 2x and is now considered a baseline SR model. For this project, OpenCVs bicubic interpolation implementation was used [16].

E. Image Stitching

One side-effect of the processes described in this paper is that the subjects of the image and the background are separated. As a result, the images need to be stitched back together into one final product image. Pythons OpenCV library and PIL (Python Imaging Library) are utilized to accomplish this task. Provided the (x,y) coordinate and (width, height) of the extracted subject image from the original image, PIL inserts the new subject images back into the main image at their original locations. The data for this operation is retrieved during the image segmentation process.

However, a simple copy/paste procedure wouldnt be enough to truly recreate an image. Since the background image and the subject images have different resolutions, the combined image has the possibility of looking somewhat awkward on the edge of the background and subject image from the sudden change in resolution. OpenCV has functionality that can be implemented to blur parts of an image with a certain intensity, which is used here to blur along the shared edges with different intensity such that the transition from the background image to subject image looks more natural.

IV. Results

A. Overview

In theory, this project is designed to overcome the shortcomings of different methods in order to have a more optimal outcome. To reiterate, the primary shortcomings of current methods in relation to complex (multi-subject images) are:

- Bicubic Interpolation & Non Deep Learning approaches - generally have been to be found less effective than DL solutions
- GANs (and other DL approaches) - are known to have problems with counting, perspective, and global structures. It is theorized that these problems could be overcome with a more complex model architecture, but we do not have the hardware to properly test that [4].

By segmenting the subjects of the images out, this method effectively minimizes both drawbacks. Segmenting the images gets rid of the counting, perspective, and global structures issues for GANs. Additionally, because GANs are known to produce better results than bicubic interpolation, having the subjects of interest be enhanced by GANs will produce a stronger image overall.

Our team had two main issues in the execution of this project:

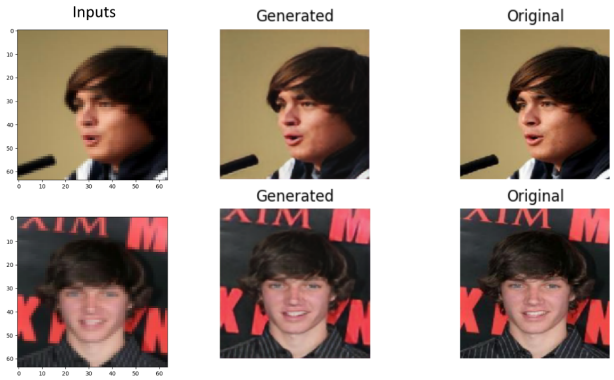

Fig. 2: Results of the SRCGAN when run on the CelebA dataset after 6900 iterations. The images on the left are the inputted images, alongside class information, the images in the middle are generated images, and the image on the right are the ground truth

1) We didnt have the hardware to train all of the models from scratch, so we tried to use as many off the shelf parts as possible.
2) We tried to parallelize the development of this project as much as possible.

While these methods work for working on each component in seclusion, they are unoptimized for creating the final end-to-end solution. As of the time of writing this paper, we have different parts of this technique working, but currently, dont have any results for the combined model.

These results will be available in the final revision of this report.

B. Super Resolution

The results for the SRCGAN are shown in 2. The specific model was trained for 6900 iterations. As shown in the figure, the GANs tends to perform well in creating an image which has similar semantics to the ground truth. In fact, if one didnt know about the existence of the ground truth, it is entirely possible to believe that the generated image isnt fake.

PSNR is what is planned to be used in the measuring the final performance of the super-resolution algorithm. Unfortunately, at the time of writing this paper, the technique wasnt in a state to present the PSNR results. These results will be available in the final iteration of this paper.

C. Semantic Segmentation

Our team used Googles DeepLab FCN to perform semantic segmentation. In Figure 3, subfigure 3a is the sample input image fed into the network. Subfigure 3b is the output of the FCN. Using this output, the objects of interest are extracted out to be used as input for the SRCGAN. This process is carried out for all the images in the dataset. Along with the images, the top left coordinates of the cropped images are passed as well to aid in imaging stitching later on.

(a) The source image fed into Google's DeepLab FCN.

(b) FCN's classification results. All of the subjects of interest are in purple, while the background is black.

(c) The FCN results overlayed on the original image.

Fig. 3: Google's DeepLab Fully Convolutional Network.

D. Stitching

Before undergoing image stitching, the subject images and background image are all separated as individual images. An example of this separation is shown below:

[BACKGROUND IMAGE] [SUBJECT IMAGE X] [SUBJECT IMAGE X] [SUBJECT IMAGE X] ...

From here, PIL is used to take the subject images and place them in their appropriate places on the background image. Their positions are determined from data provided by the image segmentation algorithm that gives the (x,y) coordinate

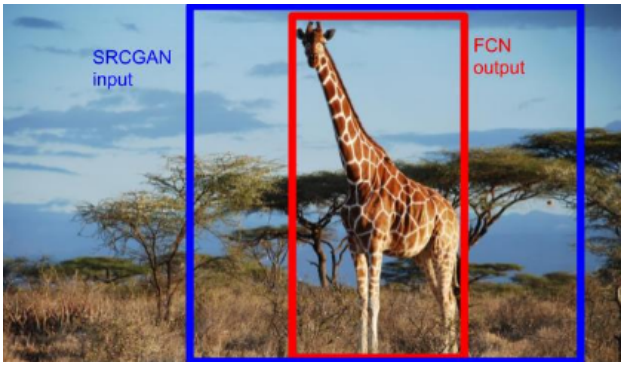

Fig. 4: An example of an inefficiency with our method. Although the FCN knows the giraffe is bounded within the red box, due to the nature of SRCGAN's we are forced to pass the blue box, essentially wasting computation.

of the specified subject image in the background image and the original subject images height and width dimensions. This reconstruction results in a single image, an example of which is demonstrated below:

[RECONSTRUCTED IMAGE]

Now that all the images are merged into one image, blurring along the shared edges between the subject images and background image needs to be done to make the entirety of the image look more natural. This is done using OpenCVs image blurring functionality with a blur radius of (WIDTH, HEIGHT). The blurring is done in such a way as to make a smooth transition from the resolution of the subject images to the resolution of the background image. An example of the final product can be seen below:

[IMAGE]

More details about the results of the image stitching process will be available in the final iteration of this paper.

V. Conclusion

In theory, with our method we achieve the best of both worlds: having access to the power of SRCGANs for the more important parts of the images we superresolution while saving on computation costs by not training on as much noise. However, due to not having the hardware to train an FCN and using third party libraries, our current end product has to re-calculate various values. Our method is also restricted by the limitations of SRCGANs. Since SRCGANs have a fixed nxn sized input, we have to either upscale or downscale our FCN output in order to pass it to the SRCGAN. Additionally, when we have long or tall subjects, like a giraffe, we need to pass additional background information that we know isnt our subject into the SRCGAN (as seen in Figure 4), resulting in noise when both training and running the SRCGAN.

Another issue with our method is when multiple subjects overlap. Since we run our SRCGAN for each image we classify, any overlap is enhanced multiple times, causing unnecessary calculation as seen in Figure 5. As such, our method would not work well when the FCN classifiers more area as subjects than the size of the original image.

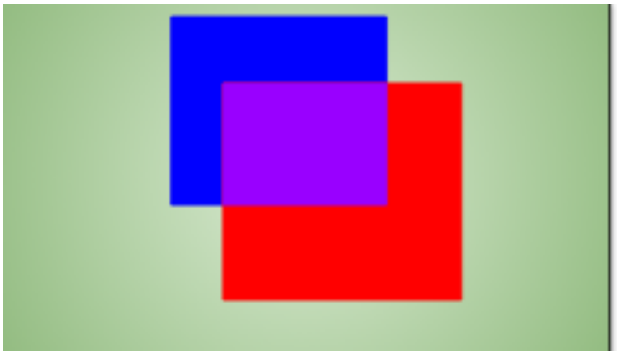

Fig. 5: If the FCN classifies the red area as one class and the blue are as another, the purple area is run through the SRCGAN twice.

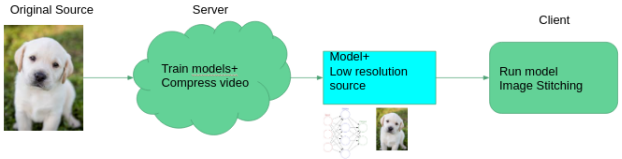

Fig. 6: A potential server-client model for video compression using Context Aware Super Resolution

VI. Future Work

Our goal for Context Aware Super Resolution is to one day use it for video compression. The end goal would be to have . While current research has shown potential for video compression using super resolution, nothing pratical has yet to be created [17], [18]. We hope to create a refined version of our method that could potentially be used for compression videos with a low subject count (Figure 6)

Acknowledgment

The authors would like to thank...

Prof. Micheal A. Gennert for helping us with image stitching.

Luis Corona for letting us borrow his GTX 1080 for training.

Google DeepLab for providing the FCN we use in our methods.

References

[1] V. Chen, L. Puzon, and C. Wadsworth, "Class-Conditional Superresolution with GANs," Tech. Rep. [Online]. Available: http://cs231n.stanford.edu/reports/2017/pdfs/314.pdf

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[3] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi Twitter, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," Tech. Rep. [Online]. Available: https://arxiv.org/pdf/1609.04802.pdf

[4] I. Goodfellow, "NIPS 2016 Tutorial: Generative Adversarial Networks," Tech. Rep., 2017. [Online]. Available: http://www.iangoodfellow.com/slides/2016-12-04-NIPS.pdf

[5] W. Siu and K. Hung, "Review of image interpolation and super-resolution," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec 2012, pp. 1–10.

[6] J. A. Ferwerda, "Three Varieties of Realism in Computer Graphics," Tech. Rep. [Online]. Available: http://cin.ufpe.br/ in1123/material/vor_hvei03_v20.pdf

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Tech. Rep. [Online]. Available: http://code.google.com/p/cuda-convnet/

[8] D. Jaswal and K. P. Soman, "Image Classification Using Convolutional Neural Networks," *International Journal of Advancements in Research & Technology*, vol. 3, no. 6, 2014. [Online]. Available: http://www.ijser.org

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Tech. Rep. [Online]. Available: https://people.eecs.berkeley.edu/ jonlong/long_shelhamer_fcn.pdf

[10] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, June 2018. [Online]. Available: http://link.springer.com/10.1007/s10462-018-9641-3

[11] L.-C. Chen, G. Papandreou, S. Member, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," Tech. Rep. [Online]. Available: http://liangchiehchen.com/projects/

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," jun 2014. [Online]. Available: http://arxiv.org/abs/1406.2661

[13] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," nov 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[14] E. Linder-Norén, "Keras implementations of Generative Adversarial Networks." [Online]. Available: https://github.com/eriklindernoren/Keras-GAN

[15] A. Lukin, A. S. Krylov, and A. Nasonov, "Image Interpolation by Super-Resolution," Tech. Rep., 2006.

[16] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[17] R. Molina, A. K. Katsaggelos, L. D. Alvarez, and J. Mateos, "Towards a new video compression scheme using super-resolution," Tech. Rep. [Online]. Available: https://pdfs.semanticscholar.org/4524/6d41c45fda222f733660bdac49093d8f859d.pdf

[18] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Super-resolution of compressed videos using convolutional neural networks," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2016, pp. 1150–1154. [Online]. Available: http://ieeexplore.ieee.org/document/7532538/