Deep Learning based Single Image Super Resolution

Harsh Vaghela
Computer Science Department
Lakehead University
Thunder Bay, Canada
hvaghela@lakeheadu.ca

Dr. Sabah Mohammed
Computer Science Department
Lakehead University
Thunde Bay, Canada.
mogammed@lakeheadu.ca

***Abstract*—The research in the field of image enhancement is particularly very important as it can boost up the progress in image recognition and images are needed to be in as high quality and high detail as possible. To address this problem there has been significant progress in the field of Image Processing and Image Enhancement. To advance this progress the concept of Super-resolution became very popular in last decade. Super-resolution is the process of enhancing or retaining the sharpness and quality of the image while converting it from Low Resolution (LR) to High resolution (HR) to make it highly detailed image. There are Multiple ways to achieve this task, one is to use multiple frames of same image with different resolution to generate one high resolution image which is Multi Image Super-resolution (MISR), and another is to use just one frame of the image to generate High resolution (HR) image which is Single Image Super-resolution (SISR). There are many solutions for this problem came into existence, but solution from deep learning algorithms have shown the most effective solution for the problem of Single Image Super-resolution (SISR). Most of the Deep Learning algorithms shows better result than conventional algorithms as they can learn the line and curve features of the input image and use them to generate the final output, that conventional algorithms are clearly not designed to do. In this paper, we will be discussing about a Deep Learning based algorithms that takes the problem of Single Image Super-resolution and shows very prominent results, in addition to that, we will be comparing that algorithm with other existing algorithm to see how it performs in its class with different sets of input.**

Keywords—Deep Learning, Single Image, Super-resolution, Image enhancement.

I. INTRODUCTION

For us, humans it is very easy to look at an image and identify what is in an image as we know that how an object looks like because of our experience. Computer doesn't see an image in its visual form, rather it sees an image as an array of numbers [1] (figure 1). It is very hard for computer to predict what is in an image from an array of numbers. Small change in those numbers which can be a huge difference for computer would be a minor change in visual effect, as small as human eye would not be able to recognize it. Humans can identify object as we know how that object looks like because we have seen and know that object from all the possible angles. To let computer have this capability, it needs to learn to understand how specific object looks like from all the possible sides.

Computer should have a memory of specific object, so it would be able to compare the object in the image with the memory of that object, and if it matches then it would be able to identify the object properly. That memory is called Feature Map or Activation Map in terms of Computer Vision [2]. To create memory for specific object, computer need to be train with images of that object from all the possible angles and lighting conditions. As computer sees image in numbers, it needs to do mathematical matrix multiplication and addition of numbers in specific region (called Filter) to create the Activation Map [3] but it needs more than one Activation Map to identify objects as it gradually learns the objects, starting from lines and curves to shapes, and finally to objects. If the similar object comes as an input, the matrix multiplication and addition will generate higher number as an output, this higher number predicts that object is similar to the Activation Map(s) and hence object identified [4].

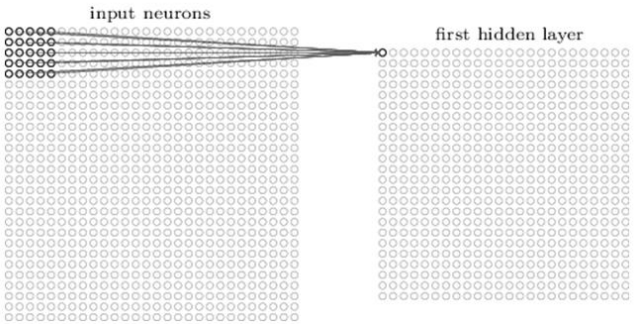

Figure 1: Filter and Feature map

This method can also be helpful to solve the problem of Super-resolution [5] as it is possible to retain the information of the image with the help of Activation Map (s). The process is called Deconvolution [6] which refers to generation the image from Activation Map(s).

The problem this paper discuss about is the problem of image super-resolution [7]. Image super-resolution can be described as converting the Low-Resolution (LR) of the image to High-Resolution (HR) image while retaining and enhancing the quality and the details of the image. There are many use cases of the solution of the problem. First and the most important use is in the field of image recognition [8]. Blurry and less detailed image is least preferred image to perform object recognition because computer outputs wrong outcome the most time and algorithm is not of use at all. With use of super resolution of the image, image recognition and object detection algorithm can preform much better to get the most promising results. Other than that, higher resolution imager is much better in quality than lower resolution image as it is pleasing for eyes to look at.

There were many attempts to solve this problem. Earlier attempts were using regular algorithms of the image enhancement and with that improving the quality and sharpness of the image [9]. Those conventional algorithms provided acceptable but not exceptional results. The main

reason for the not being state of the art solution is that conventional algorithms were not able to learn the characteristics of the images [10]. This problem we solved by deep learning algorithms. The main advantage being for deep learning algorithms is that they can learn from the training data and use that knowledge to improve on the test dataset [11]. In this paper, I will be explaining one of those deep leaning solutions for the problem of super resolution which uses the concept of residual and dense architecture of the deep learning algorithms [12].

II. Problem Defination

In the field of pattern and image recognition, resolution of input image plays vital role in feature extraction. It saves lot of effort and computing power if the input image has sharp details with large enough resolution. To tackle this problem super-resolution was proposed. There are two types of input that can be used for super-resolution: multiple images and single image. In multiple images method, more than one frame of the same image with different exposure and details are used to generate one output image that has maximum sharpness and details with high resolution. Whereas, in single image method, just one frame of Low Resolution (LR) images is used as an input to generate High Resolution (HR) as an output. LR image can be a possible crop of a HR image. The problem Single Image Super-Resolution (SISR) becomes difficult because the natural image space of HR image that algorithm aims to map to the LR image is mostly interactable.

In recent years, industry has started using the methods of Multi Image Super-Resolution [13] (MISR) as the commercial cameras are capable enough take multiple frames at a time with different configuration. This adversely affects the progression in the field of SISR. Even after these slow advancements because of advance computing power deep learning based SISR algorithms have achieved noticeable accomplishments but there are many hurdles to overcome and room for more research.

III. Related Research Work

The SISR problem has been studied in the literature using a wide categories of DL based techniques. I categorize existing methods into three groups based on the most distinctive features in their models.

A. Linear networks

Linear networks have very simple structure, mainly flowing from beginning to end in just singular path without having backtracking techniques, any other connections or more than one branch. In Linear networks design, multiple convolution network can be integrated on top of each other in single path where input flows from first to last layer. As there is just one path for flow of the data, convolution layers from the beginning cannot perform any task later in time in the execution of the algorithm. There are two types of Linear Networks on basis of when algorithm performs the task of sampling. In the network design features of images are extracted to add the details to the images, this process is called sampling. As I have mentioned two categories of Linear networks, Super-resolution Convolution Neural Network (SRCNN) [14] [15], VDSR [16], DnCNN [17] and IrCNN [18] are examples of early up-sampling design where they up-sample the given LR image into required resolution of HR image and then learn hierarchical feature representation to generate output and FSRCNN [19] and ESPCN [20] are examples of late up-sampling design where they learn hierarchical feature representation first and then up-sample the image to generate the output. Through experimental results it is apparent that late up-sampling is more efficient because feature extraction can be computationally expensive as the network structure grows and then algorithms must deal with large number of input pixel.

B. Recurcive networks

Method of recursion is very important for functions in algorithm because it can perform same operation multiple time without increasing complexity of the method and generate output in optimal time. So, the recursive networks or linked units have recursively connected take can take the control of the execution to previous methods. The main aim behind this network method is to break the problem into smaller units, that are easier to solve. The basic architecture is shown in Figure 2. Deep Recursive Convolution Network

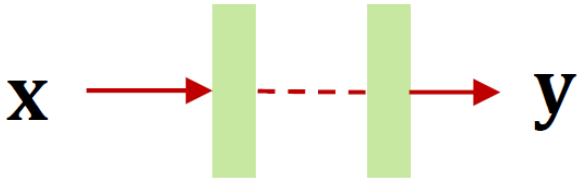

Figure 2: Linear Networks

(DRCN) [21] is one of the algorithms that works on recursive networks design as it applies the same convolution layers multiple times. In DRCN main advantage is that the number of parameters stays the same throughout the whole execution. DRCN has three smaller networks: embedding net, inference net and reconstruction net. Deep Recursive Residual Network (DRRN) [22] is another example of Recursive network design. It initializes a deep CNN model with limited parametric complexity. In comparison with another models DRRN has even deeper architecture with as high as convolution layers. This has been accomplished by joining residual image learning with connections among small parts of layers within the network. One other design of the same kind id memory network (MemNet) [23]. MemNet is consisting into three parts. First is feature extraction which extracts features from the input image. The second part is a group of memory blocks stacked together and it is consisting of recursive unit and gate unit.

C. Attention-based Networks

There is one thing common in both the types of network discussed before that, both the network type has same importance for each of the convolution layers for super-resolution. In many cases, letting separate convolution layer perform tasks with different priority can be a vital advantage. Attention based models [24] allow this flexibility of priority

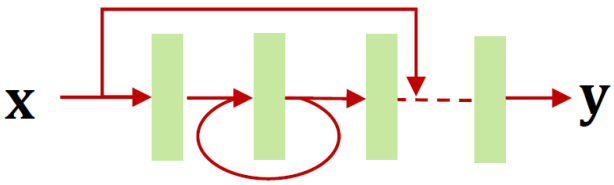

Figure 3: Recursive Networks

among convolution layers and it considers that not every feature has same importance for super-resolution. With deep networks, recent attention-based models have shown significant improvements for super-resolution. Selection unit for image Super Resolution Network (SelNet) [25] and Densely Residual Laplacian attention Network (DRLN) [26] are example of Attention-based Networks. In SelNet the selection unit is made of an identity mapping and a cascade of ReLU, 1x1 convolution and a sigmoid layer. SelNet has a total of convolutional layers, there are convolution layer followed by selection unit.

Whereas, DRLN structure has modular and hierarchal design, and the vital parts of the network are: modular architecture, densely connected residual units, Cascading connections, and Laplacian attention. DRLN shows different joints such as long-skips, medium-skips, local-skips with the cascaded ones. Same as, in each block, three residual units are densely connected to learn small representation. After that, the learned features are weighted using Laplacian attention in the same part. The structure is repeated throughout the network in each part.

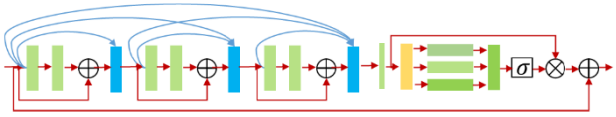

Figure 4: Attention-based Networks

All three of these lacks the interconnected architecture that one deep learning algorithm needs to provide state of the are results. In section 6, we will be comparing these algorithms with the algorithm I am reviewing.

IV. Methodology

To solve the explained problem, I review the concept of Residual network and the Dense network and using both of them for a deep learning network using concepts of both Residual network [27] and Dense network [28] (Figure 5). This network is used to fully make use of all the hierarchical features from the Low-Resolution input image to Residual and Dense block of the network.

A. Network Structure

The network is divided into four parts. The first part of the network will extract the shallow features from the Low - Resolution input image. The second part is the block consist of both Residual network and Dense network. This block performs the crucial part to extract the higher-level hierarchical features from the Low-Resolution input image. There will be more than one this block in the network structure and the number of this block would be dependent upon the performance of the architecture through experiment results.

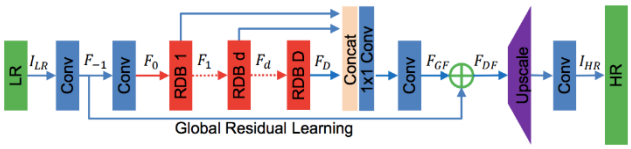

Figure 5: The Network

The Low-Resolution input image is denoted as I_{LR} and the High-Resolution output image from the network is denoted as I_{SR}. There are two Conv layers to extract the shallow features. The first two Conv layers F_{-1} and F_0 extracts the shallow features from Low-Resolution input image.

$$F_{-1} = H_{SFE1}(I_{LR})$$

$$F_0 = H_{SFE2}(F_{-1})$$

Where H_{SFE1} and H_{SFE2} denotes convolution operation. Assume that we have D number of residual network and dense network block and the output is Fd for the d-th block.

$$F_d = H_{,d}(F_{d-1})$$

$$Fd = H_{,d}(H_{,d}(H_{,d}(\ldots.H_{,d}(F_0)\ldots.))$$

Where $H_{,d}$ denotes the convolution operation of the d-th block or this can be a combined function of operation, as here it is composite function of convolution operation and activation layer of ReLu.

After extracting all the high-level features with set of blocks, we combine all of these features with a convolution operation.

$$F_{DF} = H(F_{-1}, F_0, F_1, \ldots, F_D)$$

Where FDF is a combination of feature maps from all the previous layers. The final High-Resolution image is generated by up-sampling net with the use of feature maps.

B. Residual and Dense Block

This part of architecture is consisting of Residual and Dense block with local fusion mechanism. The number of convolution operation layer in each block will be denoted as C. Local fusion is applied to adaptively add the states from the preceding blocks and all the convolution operation layers in current block. The feature maps of the (d-1)-th block are direct input to d-th block via concatenation way, and by doing this it reduces the feature numbers. To control the output information, 1 x 1 convolution layer is used.

$$F_d = H^d([F_{d-1}, F_{d,1}, \ldots, F_{d,c}, \ldots, F_{d,C}])$$

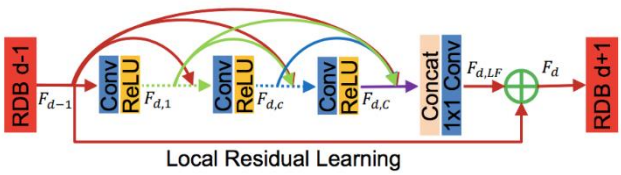

Figure 6: Residual and Dense Block

Where H^d denotes the 1 x 1 convolutional operation function in residual and dense block. The local fusion is used to improve the network workability and better performance.

C. *Global Fusion*

After getting local features with a set of residual and dense blocks, we need to combine the feature output from all the blocks. Global fusion is to perform that task.

$$F_{GF} = H_{GF}([F_1, F_2, F_3, \ldots, F_D])$$

Where F_{GF} denotes to the concatenation of features produced by each residual and dense block from 1 to D. H_{GF} is 1 x 1 and 3 x 3 convolutional layers. 1 x 1 layer is used to combine the features with different level, whereas, 3 x 3 convolutional operation layer is used to extract more higher-level features.

V. PROTOTYPING

The network has been implemented in Python 3.5 environment on Google's online Python execution platform Colab. The dataset which has been used for training the network is DIV2K [29]. DIV2K has 800 training images that consist of Low-Resolution and High-Resolution (2K resolution) images. DIV2K also has 100 validation images and 100 test images. The network is trained on 800 training images and 5 validation images in the process. For testing we have used five standard benchmark datasets: Set5 [30], Set14 [31], B100 [32] [33], Urban100 [34] and Manga109 [35]. Set5 and Set14 has random images from animals to human faces. B100 consists images from NTIRE 2017 competition. Urban 100 is made of images of human made structures and Manga 109 has images from professionally hand drawn pictures.

The important way to tweak the performance of the network to the best it can perform is via adjusting the values of number of residual and dense block in the network, number of convolutional operation layers in each residual and dense block and the learning rate of the network. After experimenting with the network, 20 residual and dense block, 6 convolution operation layers in each residual and dense block, and 32 learning rate seems to work the best. The network was trained on 200 epochs.

VI. RESULTS

In this section, we will be comparing the results of above explained network with reviewed networks in section 3. For the compression I have taken results for above mentioned algorithms from the datasets Set5, Set14, B100, Urban100 and Manga109. Set 5 and Set 14 has random images from animals to human faces. B100 consists images from NTIRE 2017 competition. Urban 100 is made of images of human made structures and Manga 109 has images from professionally hand drawn pictures. The settings for the explained network are 64 filters per convolution operation layer, D = 16, C = 8 and G = 64.

For the evaluation, I have considered two measurements. The first one is PSNR which is Pixel to Noise Ratio. PSNR generates a number, higher number suggests better results. Second one being SSIM, which stands for Structural Similarity Measurement Index. Higher the SSIM the better. Comparison between mentioned algorithms is in table 1 with evaluation measurements PSNR and SSIM.

Figure 7: Low-Resolution image from Urban 100 (256 x 161)

Figure 8: Part of Low-Resolution image (37 x 33)

Figure 9: Super Resolution image (1024 x 644)

Figure 10: Part of Super Resolution image (157 x 130)

Method	Set5		Set14		B100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRCNN	36.66	0.9545	32.45	0.9067	31.36	0.8879	29.51	0.8946	35.71	0.968
DRRN	37.74	0.9591	33.23	0.9136	32.05	0.8973	31.23	0.9188	37.92	0.976
SelNET	37.89	0.9598	33.61	0.9160	32.08	0.8984	32.55	0.9324	38.89	0.9775
Mentioned	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780

Table 1: Comparison between results of the algorithms

The results of explained network is shown in the figures. Figure 7 is the Low-Resolution image from the test dataset Urban 100 which goes as an input in the algorithm. Figure 8 is cut-out from the input image to see the quality of the image properly to compare between input and output. The output from the algorithm is figure 9 which is super resolution image or in simple terms high-resolution image of the image. Figure 10 is cut-out in the same position as figure 8. The results are apparent that the algorithm noticeably increases the quality of the image.

VII. Conclusion

From the paper we can conclude that the deep learning solutions provide much more promising results to the problem of super resolution. Among all the deep learning solutions for the problem, the network which has better architecture to cope up with the complex structure of the problem performs better then rest. The problem of super resolution is very complex problem. There are many solutions available for the problem, however, we are no where near the optimum results. Future enhancements in the field can put even better outcomes and can continue to improve in the area of image enhancement.

Refrences

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521, no. 7553 (2015): 436-444.

[2] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[3] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In *European conference on computer vision*, pp. 818-833. Springer, Cham, 2014.

[4] Donahue, Jeff, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. "A deep convolutional activation feature for generic visual recognition." *arXiv preprint arXiv:1310.1531* 1 (2013).

[5] Glasner, Daniel, Shai Bagon, and Michal Irani. "Super-resolution from a single image." In *2009 IEEE 12th international conference on computer vision*, pp. 349-356. IEEE, 2009.

[6] Jansson, Peter A., ed. *Deconvolution of images and spectra*. Courier Corporation, 2014.

[7] Yang, Wenming, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. "Deep learning for single image super-resolution: A brief review." *IEEE Transactions on Multimedia* 21, no. 12 (2019): 3106-3121.

[8] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[9] Kim, Kwang In, and Younghee Kwon. "Single-image super-resolution using sparse regression and natural image prior." *IEEE transactions on pattern analysis and machine intelligence* 32, no. 6 (2010): 1127-1133.

[10] Timofte, Radu, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. "Ntire 2017 challenge on single image super-resolution: Methods and results." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 114-125. 2017.

[11] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural networks* 61 (2015): 85-117.

[12] Zhang, Yulun, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. "Residual dense network for image super-resolution." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472-2481. 2018.

[13] Li, Xuelong, Yanting Hu, Xinbo Gao, Dacheng Tao, and Beijia Ning. "A multi-frame image super-resolution method." *Signal Processing* 90, no. 2 (2010): 405-414.

[14] Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Image super-resolution using deep convolutional networks." *IEEE transactions on pattern analysis and machine intelligence* 38, no. 2 (2015): 295-307.

[15] Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Learning a deep convolutional network for image super-resolution." In *European conference on computer vision*, pp. 184-199. Springer, Cham, 2014.

[16] Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pp. 1646-1654. 2016.

[17] Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising." *IEEE Transactions on Image Processing* 26, no. 7 (2017): 3142-3155.

[18] Zhang, Kai, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. "Learning deep CNN denoiser prior for image restoration." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3929-3938. 2017.

[19] Dong, Chao, Chen Change Loy, and Xiaoou Tang. "Accelerating the super-resolution convolutional neural network." In *European conference on computer vision*, pp. 391-407. Springer, Cham, 2016.

[20] Shi, Wenzhe, 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." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874-1883. 2016.

[21] Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Deeply-recursive convolutional network for image super-resolution." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637-1645. 2016.

[22] Tai, Ying, Jian Yang, and Xiaoming Liu. "Image super-resolution via deep recursive residual network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3147-3155. 2017.

[23] Tai, Ying, Jian Yang, Xiaoming Liu, and Chunyan Xu. "Memnet: A persistent memory network for image restoration." In *Proceedings of the IEEE international conference on computer vision*, pp. 4539-4547. 2017.

[24] Zhang, Yulun, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. "Image super-resolution using very deep residual channel attention networks." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286-301. 2018.

[25] Choi, Jae-Seok, and Munchurl Kim. "A deep convolutional neural network with selection units for super-resolution." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 154-160. 2017.

[26] Anwar, Saeed, and Nick Barnes. "Densely Residual Laplacian Super-Resolution." *arXiv preprint arXiv:1906.12021* (2019).

[27] Zhang, Yulun, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. "Residual dense network for image super-resolution." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472-2481. 2018.

[28] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708. 2017.

[29] Agustsson, Eirikur, and Radu Timofte. "Ntire 2017 challenge on single image super-resolution: Dataset and study." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126-135. 2017.

[30] Sheikh, Hamid R., Alan C. Bovik, and Gustavo De Veciana. "An information fidelity criterion for image quality assessment using natural scene statistics." *IEEE Transactions on image processing* 14, no. 12 (2005): 2117-2128.

[31] Zeyde, Roman, Michael Elad, and Matan Protter. "On single image scale-up using sparse-representations." In *International conference on curves and surfaces*, pp. 711-730. Springer, Berlin, Heidelberg, 2010.

[32] Martin, David, Charless Fowlkes, Doron Tal, and Jitendra Malik. "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, pp. 416-423. IEEE, 2001.

[33] Timofte, Radu, Vincent De Smet, and Luc Van Gool. "A+: Adjusted anchored neighborhood regression for fast super-resolution." In *Asian conference on computer vision*, pp. 111-126. Springer, Cham, 2014.

[34] Huang, Jia-Bin, Abhishek Singh, and Narendra Ahuja. "Single image super-resolution from transformed self-exemplars." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5197-5206. 2015.

[35] Fujimoto, Azuma, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. "Manga109 dataset and creation of metadata." In *Proceedings of the 1st international workshop on comics analysis, processing and understanding*, pp. 1-5. 2016.