Image Super-Resolution using Deep Neural Networks

John Vadayattukunnel Lal
Department of Computer Science
University of Toronto
214 College St, Toronto, ON M5T 3A1
john.vadayattukunnellal@mail.utoronto.ca

Farhad Javid
Department of Computer Science
University of Toronto
214 College St, Toronto, ON M5T 3A1
fjavid@cs.toronto.edu

Abstract

Image super resolution (SR) is one of the most studied yet challenging problems in computer vision. Many learning-based approaches based on deep neural networks have been suggested for the image super resolution problem. Here, we implement a generative adversarial and a deep recursive neural network for image SR and study the effect of network architecture, loss function, and models parameters on the quality of generated SR images. Finally, we analyse our results and, inspired by our finding, provide suggestions to improve the models outputs and design new networks with state-of-the-art image SR performance.

1 Introduction

Image super resolution (SR), generating a high resolution (HR) image from a single low resolution (LR) one, is required in many computer vision tasks including image and video restoration, semantic segmentation, security, and even medical imaging [Yamashita and Markov, 2020a]. SR is an *ill-posed* inverse problem where several HR images can be reconstructed from a single LR image. The past couple of decades have seen several interpolation-based, reconstruction-based and learning-based methods to tackled the SR inverse problem [Zhang and Wu, 2006, Zhang et al., 2012, Dong et al., 2015]. Recent advances in high performance computer architectures such as GPUs and development of deep learning techniques, however, have made learning-based methods the best performing and consequently, the main-stream technique of SR. In this work, we study two of such techniques designed based on generative adversarial and recursive memory block networks under different network parameters, loss functions, and architectures.

2 Related Work

The first deep learning network for single image super resolution was developed based on the convolutional neural network (SRCNN) architecture by Dong et al. [2015]. SRCNN is a three-layer fully convolutional network, with pixel-wise mean square error (MSE) loss. The network's layers are specifically designed for patch extraction, nonlinear mapping and image reconstruction. To adjust the scale, the LR image is upsampled using bicubic interpolation before being fed to the network. Subsequent Deep learning neural networks designed for image SR has improved model performance through *loss functions*, *network architecture*, and *learning strategies* modifications [Yang et al., 2019].

The primary focus of learning-based image SR has been network architecture. The performance of the deep learning techniques for SR has significantly increased since the introduction of *very deep* neural networks for SR (VDSR) [Simonyan and Zisserman, 2014]. VDSR has a 20-layer deep VGG structure and uses a residual learning approach to improve the performance and the convergence of the method. For training, VDSR uses a hierarchy of upsampled LR images in different scales to help the network learn the mapping between the LR and and high resolution residuals.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

An additional focus of these techniques have been the loss function. Early image SR techniques rely on pixel-wise mean squared error[Yang et al., 2019]. It is, however, shown that the losses that measure the perceptual difference between LR and HR images, such as the one suggested for VSDR, lead to more visually acceptable results [Blau and Michaeli, 2018]. Such perceptual losses can be defined as a *Gibbs* energy model betweeen the LR and HR feature spaces.

3 Method

Memory network (MemNet) is a deep recursive neural network developed for image denoising, image restoration, and JPEG deblocking [Tai et al., 2017]. MemNet is a purely convolutional network composed of a feature extraction, a memory, and a reconstruction block. The feature extraction and the reconstruction blocks are made of single layer convolutions while the memory block has a residual (ResNet) architecture [He et al., 2016]. The memory block is divided into n memory units. Each memory unit is made of a recursive unit and a gate unit. The recursive unit consists of m residual layers aiming to generate short-term memory H at each state. The transformation function of the i^{th} residual layer within a recursive unit is $H_i = f\left(f(H_{i-1})\right) + H_{i-1}$, where H_i and H_{i-1} are respectively the short-term memories generated in the current and previous residual layers. The short-term memories $H = [H_1, H_2, \ldots, H_m]$ of each recursive along the long-term memories $M = [M_1, M_2, \ldots, M_{j-1}]$ generated generated in the previous blocks are fed into a gate unit to make the long-term memory, M_j, at the current memory bock. The output of the last memory unit is then fed into the reconstruction block to generate the super resolution image. The loss between the super-resolution output s and the high resolution ground truth, y is defined is $L = \frac{1}{2b} \left\| y - s \right\|^2$ where b is the batch size. Please see the Appendix for an illustration of the MemNet architecture.

The GAN approach to single image super resolution uses two separate CNN architectures to produce a high quality estimate of the LR images. The first network, a generator, is trained to take as input the LR and output an estimate of the HR. The second network, a discriminator, is trained to classify true high resolution images from an estimated HR made by the Generator network. Super Resolution architectures are generally trained on the MSE Loss. A consequence of this is that these models have very good peak-signal-to-noise ratio (PSNR). However, since MSE is defined on pixel wise image differences, it becomes difficult for these models to capture other details that may otherwise make an image look better to the human eye [Ledig et al., 2017]. A significant portion of SRGAN output quality has been attributed to the VGG loss introduced by Simonyan and Zisserman [2015]. The loss uses VGG-19 trained on ImageNet to to calculate the Euclidean distance between features of the source and estimated image. This approach aims to give the model access to information about image texture which in turn should make the image more acceptable visually. The detailed GAN architecture and loss can be found in the Appendix.

4 Experiments

MemNet is made of six memory units where each memory unit has six short-term recursive units. Batch normalization and ReLU activation are applied to the input data before each ResNet layer in the recursive units. All convolutions have 64 output features except the reconstruction block which has a single channel output. The kernel size of all convolution layers except the gate unit is 3×3 preserving the images input dimensions. The gate unit has a convolution layer with a kernel size of 1×1 (see Appendix for more details).

MemNet is trained on the training images of the Set14 Super Resolution Dataset including BSDS200 (Berkeley Segmentation Dataset), T91, and General100 [Bevilacqua et al., 2012a]. Due to the number of parameters of the model we cannot use the full size images for training. Patches of 31×31 with stride of 21 are cut from the training images and stored prior to training. To generate the low resolution images the high resolution patches are downsampeled and upsampled using bicubic technique. The scaling factor is chosen randomly during the training process. In total, 129,518 image patches are used for training. The training is done for 80 epochs. A stochastic gradient descent technique with momentum of 0.9 and weight dacay of 1E^{-4} were used for training. Finally, the initial learning rate was chosen to be 0.1 which is decays by 0.1 every 10 epochs, and the batch size is 128.

Generative Adversarial Networks have been established to be one of the more difficult architectures to train; showing great instability and even mode collapse in practice [Mescheder, 2018]. The model

was trained with the Adam Optimizer with β_1 = 0.9. Initial training iterations showed that the discriminator was much more effective than than the generator. Some quick testing with different learning rates proved to be of no effect, even with lower learning rates, the adversarial loss was quick to settle at 0. To combat this issue, some ideas from [Mordido et al., 2020] and their work on Dropout-GANs were incorporated into the SRGAN. Adding dropout layers in the discriminator is analogous to forcing the generator to satisfy an ensemble of discriminators which should help the generator produce results more widely acceptable and help prevent mode collapse. Additionally, this helps reduce how effective each individual discriminator is, which for the SRGAN enabled stable training. As already explored, the generator is trained on a combination of the MSE loss and the

Figure 1: Super resolution images generated by the SRGAN and MemNet models are compared with a bicubic upsampling technique and the high resolution ground truth. The upsampling scale factor is 4. Both SRGAN and MemNet outperform the interpolation-based bicubic technique.

perceptual VGG loss. A hyper-parameter of significant impact on the final model quality seemed to be the weight of the perceptual loss when training the generator. A deviation from the prescribed 0.006 from Ledig et al. [2017] was explored in this paper. The observation was that weighing the perceptual loss function lower created loss in texture detail as expected. Weighing the perceptual loss higher had an impact on the model's ability to learn basic information about image color & contrast resulting in faded images as shown in the appendix. An interesting property of SRGAN is that the generator is agnostic to image input size, Given any RGB input image, the Generator will return a 4x Super Resolution estimate. A consequence of this is that the generator can by trained on images of any size with some quick modifications to the discriminator. An exploration was made into the effect of patch size on the quality of the final model. The GAN was trained on both 64x64 patches and 96x96 patches and even though training became significantly more unstable with the larger patches, it resulted in higher quality HR estimates.

5 Discussion

The trained models were primarily tested on Set5 [Bevilacqua et al., 2012b] and Set14 [Yamashita and Markov, 2020b]. The main focus of the testing has been to judge the visual quality of the SR estimates. An examination of the results, primarily focused on edges and texture, confirms the fact that learning-based methods outperform interpolation-based methods. MemNet performs better than SRGAN which is mostly attributed to the size of Memnet, an 80 layer model with over 4.75M parameters. The SRGAN is, however, effective at creating smooth edges and strong textures, a consequence of the added VGG loss which is also led in predicting high texture in areas that may otherwise be low texture as shown in Fig. 2.

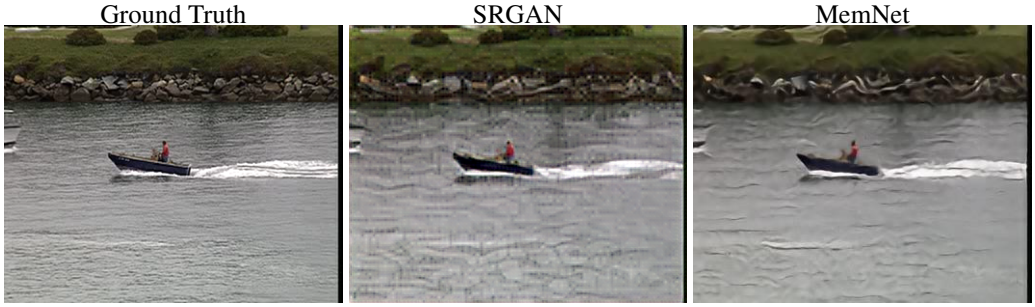

Figure 2: SRGAN tends to predict high texture even in areas with low texture.

An additional exploration into model quality beyond just visual was done in this project. The work done by Ledig et al. [2017] has proven that Peak Signal to Noise Ratio (PSNR), a metric based on the mean squared error, is not necessarily entirely indicative of the quality of the HR estimates produced by the model. To further explore this, partially inspired by Kavitha and Rao [2019], we decided to evaluate our models on a metric more focused on the features of the image. The metric uses the cosine similarity between the 9216 dimension feature vector obtained from AlexNet [Krizhevsky et al., 2012] to measure the perceptual similarity between the two images. The results are shown in Table 1.

Set5	Nearest	Bicubic	SRGAN	MemNet $\times 4$	MemNet $\times 3$	MemNet $\times 2$
PSNR	23.46	25.93	25.89	28.53	30.37	32.43
ALEX	0.702	0.835	0.932	0.938	0.964	0.984
Set14	Nearest	Bicubic	SRGAN	MemNet $\times 4$	MemNet $\times 3$	MemNet $\times 2$
PSNR	21.44	22.78	23.07	25.44	26.95	29.07
ALEX	0.709	0.797	0.899	0.902	0.942	0.976

Table 1: The PSNR and ALEX metric for all methods considered for sets 4 and 15.

If we rely entirely on PSNR, SRGAN and Bicubic seem to to be performing similarly. This seems to be in contradiction with the actual results of the model where SRGAN model significantly outperforms Bicubic. However, the low PSNR score is by design. The SRGAN sacrifices pixel wise error in favour of perceptual differences [Ledig et al., 2017]. For this reason, the AlexNet-based metric proves to be a much better measure of final image quality.

The MemNet model has 80 layers with more than 4.75M parameters. Considering the size of this model, over-fitting the training data is a potential problem of this model. In addition, training deep neural networks with many parameters is always challenging and requires excessive computational resources. So, it is necessary to find the optimum size of the MemNet for image SR tasks. To this end, in addition to the six memory block MemNet, we trained models with three and four memory blocks consisting of 23 and 38 convolutional layers, respectively. The PSNR of the SR images generated by these models are, respectively, 25.51 and 25.52 very close to the PSNR value reporter for the original model in Table 1. This implies that MemNet model with six memory blocks is probably too large for image super resolution.

6 Conclusion

In this work, we implemented two techniques, MemNet and SRGAN, for single image super resolution problem. MemNet is a deep convolutional network based on ResNet architecture which relies on a sequence of short- and long-term memories to extract the features and reconstruct the image. On the other hand, SRGAN relies on generative adversarial networks to generate a super resolution image in its generator unit. Both techniques showed superior results compared to the interpolation-based techniques such as linear and bicubic techniques.

Attributes

MemNet was developed by FJ and SRGAN was developed by JVL. The coauthors equally contributed in writing and scientific discussions.

References

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012a.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10. BMVA Press, 2012b. ISBN 1-901725-46-4. doi: http://dx.doi.org/10.5244/C.26.135.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015.

Kaiming He, 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*, pages 770–778, 2016.

K. Kavitha and B. Thirumala Rao. Evaluation of distance measures for feature based image registration using alexnet. *CoRR*, abs/1907.12921, 2019. URL `http://arxiv.org/abs/1907.12921`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012. doi: 10.1145/3065386.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

Lars M. Mescheder. On the convergence properties of GAN training. *CoRR*, abs/1801.04406, 2018. URL `http://arxiv.org/abs/1801.04406`.

Gonçalo Mordido, Haojin Yang, and Christoph Meinel. Dropout-gan: Learning from a dynamic ensemble of discriminators, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Ying Tai, 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*, pages 4539–4547, 2017.

Koki Yamashita and Konstantin Markov. Medical image enhancement using super resolution methods. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, pages 496–508, Cham, 2020a. Springer International Publishing.

Koki Yamashita and Konstantin Markov. Medical image enhancement using super resolution methods. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, pages 496–508, Cham, 2020b. Springer International Publishing. ISBN 978-3-030-50426-7.

Wenming Yang, 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(12):3106–3121, 2019.

Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(11):4544–4556, 2012.

Lei Zhang and Xiaolin Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE transactions on Image Processing*, 15(8):2226–2238, 2006.

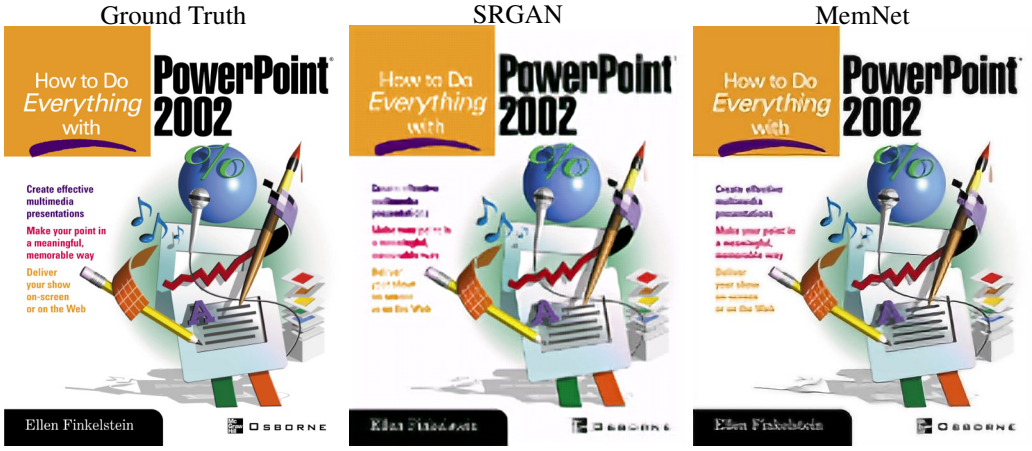

Figure 3: To no real surprise, these models are very ineffective with text.

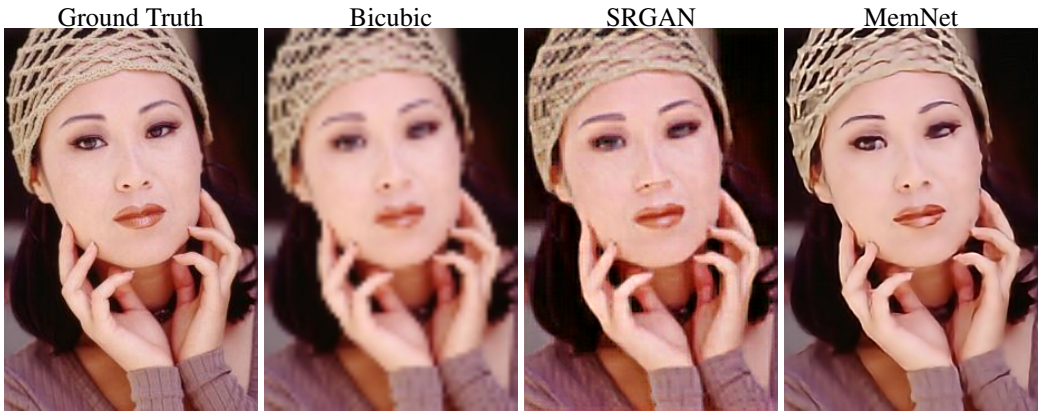

Figure 4: Generally, the models result in impressive improvement against all interpolation based methods.

7 Appendix

7.1 More Results

Some more results from the models are shown in Figs. 3 and 4.

7.2 Model Exploration

7.2.1 Image Quality with High Perceptual Loss

When the loss function gives higher priority to the perceptual loss, the GAN is limited in what it can learn as shown in Fig. 5. Significant detail about the color and contrast of the image seem to be entirely dependant on a very dominant MSE loss defining the GAN training.

7.2.2 Multi supervised MemNet

In a standard single-supervised MemNet, the last long-term memory M_M extracted by the last memory unit is used to generate the SR image in the reconstruction block. Alternatively, in a multi-supervised MemNet, all long-term memories $M = [M_1, M_2, \ldots, M_n]$ can be used to generate multiple SR outputs $s_m = [s_1, s_2, \ldots, s_n]$. These outputs can be then averaged by learnable weights to generate the final output. The loss function in this case will be defined as $L = \frac{\alpha}{2mb} \|y - s\|^2 + \sum_{i=1}^{n} \frac{1-\alpha}{2mb} \|y - s_i\|^2$, where $\alpha = 1/(1+m)$.

Figure 5: High perceptual loss.

The results of the multi-supervised MemNet model is compared to the single-supervised counterpart in Fig. 6. As can be seen here, the output of the single-supervised model as good as the the multi-supervised one although the peak signal to noise ratio (PSNR) is slightly higher for multi-supervised model.

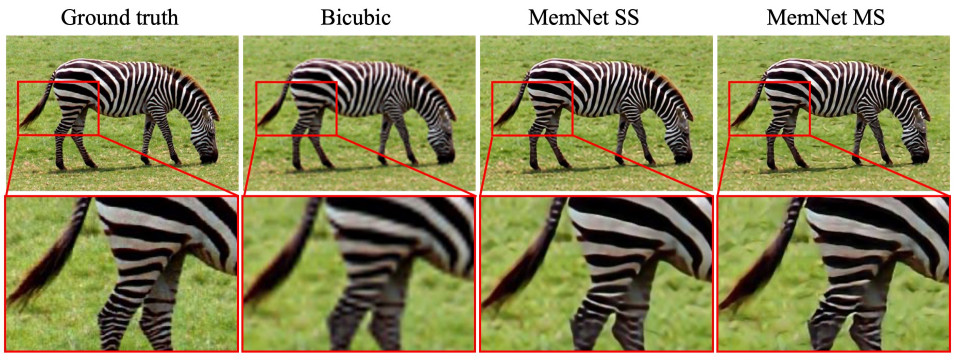

Figure 6: Single supervised versus multi supervised MemNet.

7.2.3 General Resampling

During training of MemNet, we used bicubic technique to downsample and upsample of the HR images to generate LR ones. The scaling factor for resampling was chosen randomly during the training process. Ideally, we should train our model for several scale factors on each image. Also, we should use multiple technique for resmapling to train our model for more general cases. However, due to the resource limitation we could not do this. As a consequence, our MemNet models does not perform well when different resampling techniques are used. However, we can retrain our model for a few (20) epochs on random resampling techniques to improve its quality in a general case. The results of our original and retrained MemNet models are shown in Fig. 7. As can be seen here, due to its deep network structure and many variables, MemNet is capable to generate high quality SR using different resampling techniques.

7.3 Model Architecture

The architecture of the MemNet model and the memory units are shown in Fig. 8

Also, the architecture of a multi-supervised MemNet is shown in Fig. 9

The GAN architecture is outlined in Fig. 10

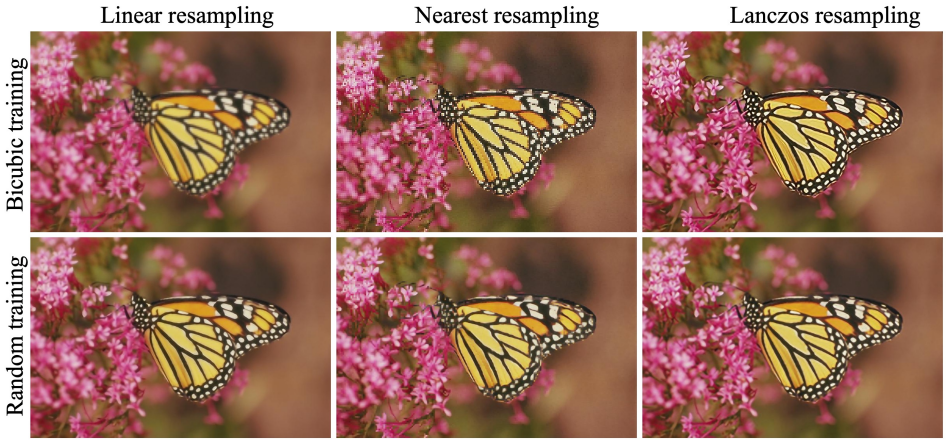

Figure 7: Effect of different resampling techniques techniques during training on the model performance; the first row are generated by a MemNet model trained on LR images made by bicubic resampling and the second row is generated by a model trained on LR images randomly generated by different resampling techniques.

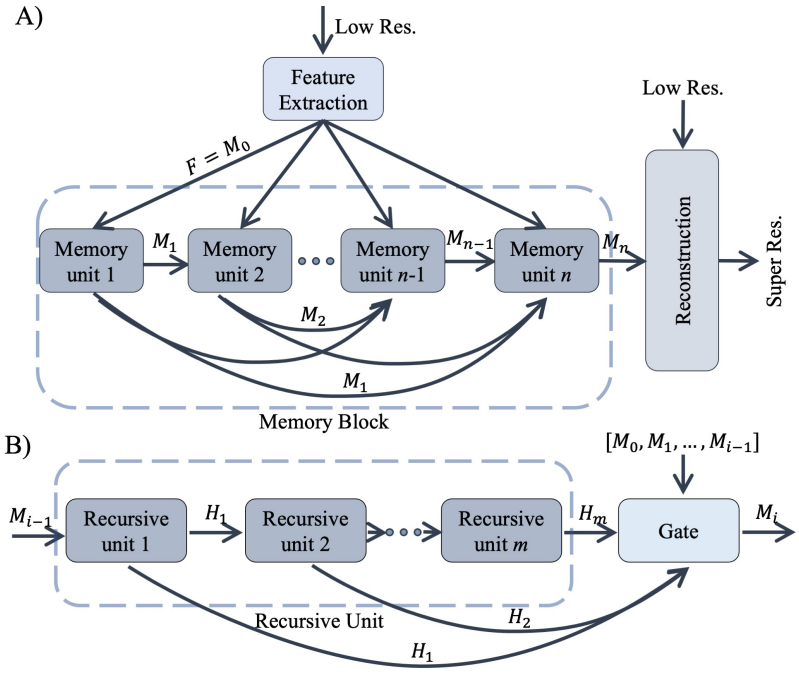

Figure 8: MemNet Architecture includes a feature extraction, a multi-layer memory, and a reconstruction block. All blocks are made of a two-layer convolutional residual unit.

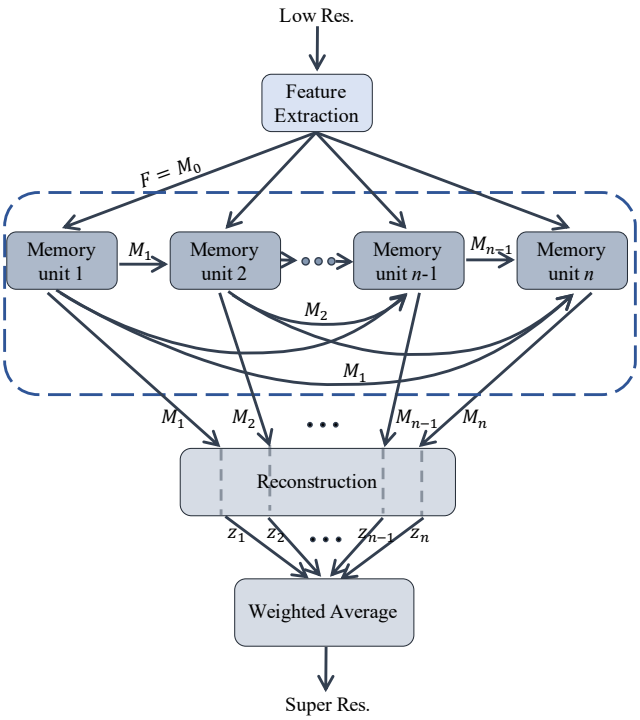

Figure 9: Multi-supervised MemNet architecture.

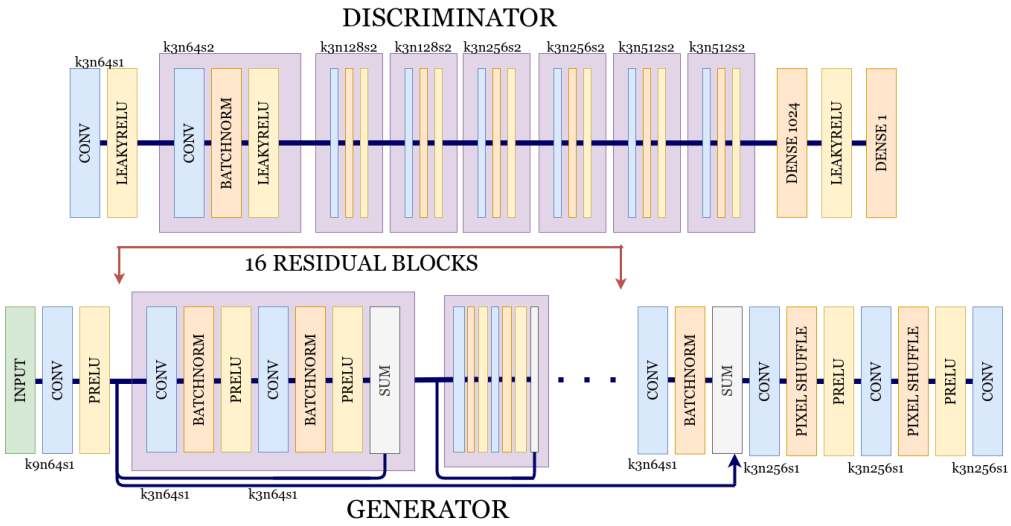

Figure 10: The generator / discriminator combination used in this project as described by Ledig et al. [2017]. As recommended, 16 residual blocks were used for this project.

$$\text{G}_{loss} = \frac{1}{WH}(I^{HR} - I^{HRE})^2 + \lambda_1 \cdot \frac{1}{W_{vgg}H_{vgg}C_{vgg}}(l_{vgg}(I^{HR}) - l_{vgg}(I^{HRE}))^2 + \lambda_2 \cdot (1 - D(I^{HRE}))$$

$$\text{D}_{loss} = 1 - D(I^{HR}) + D(I^{HRE})$$

A couple of things to note here. $I^{HRE} = G(I^{LR})$, HRE here stands for the estimate of the High resolution image. Furthermore, $l_{vgg}(I)$ are the features extracted from VGG. W_{vgg}, H_{vgg} & C_{vgg} are the height width and number of channels of the features from VGG.

7.4 Source Code

The full source code can be found in these repositories:

- Memnet
- SRGAN