Low-Resolution Data Preprocessing for Image Super Resolution

Hieu Nguyen, Jiangwei Wang

Image Super-Resolution (SR) is a well-studied class of image processing techniques to enhance the resolution of images and videos in computer vision. Previous deep learning techniques have been proposed to improve the image super-resolution performance, including EDSR, SRGAN, SRCAN. However, image preprocessing, which also plays important role in super-resolution, has not been fully investigated yet. In this project, we apply wavelet transformation to images before the network, we apply SRGAN as our network structure and EDSR as the generator. Experiment results show that the wavelet transformation largely improves the visual performance of the low-res image, especially when the image quality is very low.

I. Introduction

Image super-resolution (SR), which refers to the process of recovering high-resolution (HR) images from low resolution (LR) images are an important class of image processing techniques in computer vision and image processing. image super-resolution has a wide range of real-world applications, including security and surveillance, medical imaging, and other computer vision and image processing tasks like object detection. Image super-resolution in general is a hard problem because a single low-resolution image can have multiple corresponding high-resolution images. Traditional image super-resolution techniques include edge-based method, prediction-based method, statistical method, etc [1]–[4].

The learning-based super-resolution has become the mainstream With the rapid development of deep learning techniques these years. Convolutional neural network (CNN) based method and recently developed Generative Adversarial Nets (GAN) have been improving the performance of SR from different aspects, like peak signal to noise ratio(PSNR) and perceptual difference with ground truth images. In general, the main difference of the existing approaches can be categorized into (1) network architectures (2) loss functions (3) learning principles and strategies.

Though different deep learning techniques have been proposed to improve the super-resolution, low-resolution image preprocessing has not been fully investigated yet. Several data augmentation methods like image cropping and adding noise have been shown to improve the performance of SR. In this work, we apply the wavelet transformation to images, by decomposing the image signal into high-frequency sub-bands denoting texture details and low-frequency sub-bands containing global topological information, and using GAN network structure with EDSR as a generator, the perceptual performance of the output is largely improved compared to EDSR without wavelet transformation. Our contributions are twofold:

- We apply the wavelet transformation to the images into a high-frequency band and low-frequency band.
- By adopting GAN structure and EDSR with the generator, and using wavelet transformation to preprocess the images, the experiment shows obvious perceptual improvement on the super-resolution when applying to real low-resolution images.

The remaining part of the report is organized as follows: Section II provides a literature review of the deep learning-based SR techniques, Section III illustrates proposed wavelet transformation based SR methods, experiment results and comparisons are shown in section IV, Section V gives conclusion and future works.

II. Literature Review

To give a thorough review of the development and state-of-art of SR, this section provides the problem definition, image quality assessment metrics, SR frameworks, upsampling methods, network structure design, and loss functions.

A. Problem definition

Generally, the LR image I_x is modeled as the output of the following degradation:

$$I_x = D(I_y, \delta) \tag{1}$$

where D denotes a degradation mapping function, I_y is the corresponding HR image and δ is the parameters of the degradation process (e.g., the scaling factor or noise). The degradation process is unknown and only LR images are provided. In this case, also known as blind SR, researchers are required to recover an HR approximation $\hat{I_y}$ of the ground truth HDR image I_y from the LR image I_x, as follow:

$$\hat{I_y} = F(I_x, \theta) \tag{2}$$

where F is the model used and θ is the parameter of the model.

Most works directly model the degradation as a single downsampling operation, as follows:

$$D(I_y, \delta) = (I_y) \downarrow_s, s \subset \delta \tag{3}$$

where $\downarrow_s$ is a downsampling operation with the scaling factor s. most datasets for generic SR are built based on this pattern, and the most commonly used downsampling operation is bicubic interpolation with anti aliasing.

the objective of SR is as follows:

$$\hat{\theta} = \arg\max_{\theta} L(\hat{I_y}, I_y) + \lambda\phi(\theta) \tag{4}$$

where $L(\hat{I_y}, I_y)$ represents the loss function between the generated HR image, $\phi(\theta)$is the regularization term and λ is the tradeoff parameter.

B. Image Quality assessment

Image quality refers to the visual attributes of images and focuses on the perceptual assessments of viewers. In general, image quality assessment (IQA) methods include subjective methods based on humans' perception (i.e., how realistic the image looks) and objective computational methods. The former is more in line with our need but often timeconsuming and expensive, thus the latter is currently the mainstream. However, these methods aren't necessarily consistent with each other, because objective methods are often unable to capture the human visual perception very accurately, which may lead to a large difference in IQA results. We illustrate three evaluation metrics here:

(1) Peak Signal-to-Noise Ratio

Peak signal-to-noise ratio (PSNR) is one of the most popular reconstruction quality measurement of lossy transformation (e.g., image compression, image inpainting). For image super-resolution, PSNR is defined via the maximum pixel value (denoted as L) and the mean squared error (MSE) between images. Given the ground truth image I with N pixels and pixels and the reconstruction $\hat{I}$, the PSNR between I and $\hat{I}$ are defined as follows:

$$PSNR = 10log_{10}(\frac{L^2}{\frac{1}{N}\sum((I - \hat{I})^2)}) \tag{5}$$

(2)Structural Similarity

The structural similarity index (SSIM) is proposed for measuring the structural similarity between images, based on independent comparisons in terms of luminance, contrast, and structures. Since the SSIM evaluates the reconstruction quality from the perspective of the HVS, it better meets the requirements of perceptual assessment and is also widely used.

(3)Mean Opinion Score

Mean opinion score (MOS) testing is a commonly used subjective IQA method, where human raters are asked to assign perceptual quality scores to tested images. Typically, the scores are from 1 (bad) to 5 (good). And the final MOS is calculated as the arithmetic mean overall ratings.

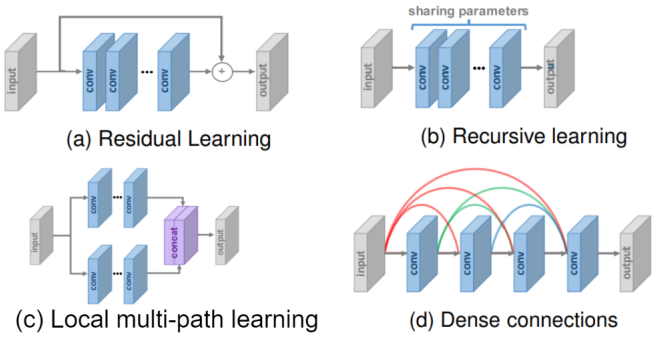

Fig. 1: Different network design

C. network design

We illustrate four different network architectures that are mostly being adopted and analyze their advantages and limitations, as shown in Fig. 1.

(1) Residual Learning

The local residual learning is similar to the residual learning in ResNet [5] and used to alleviate the degradation problem caused by increasing network depths, reduce training difficulty, and improve the learning ability. It is widely used in SR Problems [6]–[9].

(2) Recursive Learning

By applying the same module several times recursively, recursive learning has less amount of parameters compared to other networks. In general, recursive learning can learn more advanced representations without an extremely large amount of parameters. However, it still has high computational cost and gradient vanishing problems, therefore they need to be combined with residual learning sometimes [8], [10]–[12].

(3) Multi-path Learning

Multi-path learning refers to passing features through multiple paths, which perform different operations, and fusing them back for providing better modeling capabilities. Through such multi-path learning, the SR models can better extract image features from multiple scales and further improve performance [13]–[16].

(4) Dense Connections

For each layer in a dense block, the feature maps of all preceding layers are used as inputs, and its feature maps are used as inputs into all subsequent layers. The dense connections not only help alleviate gradient vanishing, enhance signal propagation and encourage feature reuse, but also substantially reduce the model size by employing a small growth rate (i.e., number of channels in dense blocks) and squeezing channels after concatenating all input feature maps [17]–[19].

D. loss functions

We will cover two loss functions in this subsection, including pixel losses and adversarial losses.

(1) Pixel losses Pixel loss measures pixel-wise difference between two images and mainly includes L1 loss (i.e., mean absolute error) and L2 loss

$$L_{Pixel_L1} = \frac{1}{hwc}\sum|\hat{I} - I| \tag{6}$$

$$L_{Pixel_L2} = \frac{1}{hwc}\sum(|\hat{I} - I|)^2 \tag{7}$$

where h, w, c are the height, width, and number of channels of the evaluated images

(2) Adversarial losses Generative adversarial network (GAN) has been adopted to super-resolution (SR) tasks [20]. We only need to treat the SR model as a generator, and define an extra discriminator to judge whether the input image is generated or not.

$$L_{gan-ce-g}(\hat{I}; D) = -logD(\hat{I}) \tag{8}$$

$$L_{gan-ce-d}(\hat{I}, I_s; D) = -logD(I_s) - log(1 - D(\hat{I})) \tag{9}$$

III. METHODOLOGY

A. M-Band Wavelet

One of many signal processing methods is the Fourier Transform. However, when applying Fourier Transform to get into the frequency-domain, we get the magnitude of the frequency but lose out on the location information. It is difficult to visualize and interpret frequency information after Fourier Transform.

Wavelet Transform is a solution for the aforementioned challenge as it fulfills both conditions of decomposing the signal into the frequency domain and space information. [21] [22]. For our experiment, we construct a 4-Band wavelet so that we can break down our features data. Below is the filter banks for our 4-Band Wavelet:

$$\alpha = [-0.067371, 0.094195, 0.405805, 0.567372, \\ 0.567372, 0.405805, 0.094195, -0.067372] \tag{10}$$

$$\beta = [-0.094195, 0.067372, 0.567372, 0.405805, \\ -0.405805, -0.567372, -0.067372, 0.094195] \tag{11}$$

$$\gamma = [-0.094195, -0.067372, 0.56737, -0.405805, \\ -0.405805, -0.56737, -0.067372, -0.094195] \tag{12}$$

$$\delta = [-0.067372, -0.094195, 0.405805, -0.567372, \\ 0.567372, -0.405805, 0.094195, 0.067372] \tag{13}$$

where α is the low pass filter bank, and β, γ, δ are the high pass filter banks such that they satisfy the following conditions:

$$\sum_{i=1}^{8} \alpha_i = 2 \tag{14}$$

$$\sum_{i=1}^{8} \beta_i = \sum_{i=1}^{8} \gamma_i = \sum_{i=1}^{8} \delta_i = 0 \tag{15}$$

$$|\alpha| = |\beta| = |\gamma| = |\delta| \tag{16}$$

$$\alpha \cdot \beta = \alpha \cdot \gamma = \alpha \cdot \delta = \beta \cdot \gamma = \beta \cdot \delta = \gamma \cdot \delta = 0 \tag{17}$$

An example where $S \in R^{4^k} (k \in N, k \geq 2)$, a 4-Band Wavelet Transform matrix T_1 (See Fig. 11) is constructed by shifting and wrapping around the filter banks. Let's call our wavelet matrix $W \in R^{4^k} \times R^{4^k}$ and it is an orthonormal

$$\begin{bmatrix}
\alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 \\
\alpha_5 & \alpha_6 & \alpha_7 & \alpha_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\
\beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 \\
\beta_5 & \beta_6 & \beta_7 & \beta_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 \\
\gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 \\
\gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \\
\delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 & \delta_7 & \delta_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 & \delta_7 & \delta_8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 & \delta_7 & \delta_8 \\
\delta_5 & \delta_6 & \delta_7 & \delta_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta_1 & \delta_2 & \delta_3 & \delta_4
\end{bmatrix}$$

Fig. 2: An example of $4^2 \times 4^2$ Wavelet Transform matrix

matrix since the column and row vector of W form a set of the orthonormal basis for R^{4k}. To get into the frequency domain (F), we apply wavelet transformation to the signal by simple perform a matrix multiplication:

$$F = W \cdot S \cdot W^T = \begin{bmatrix} a & d_1 & d_2 & d_3 \\ d_4 & d_5 & d_6 & d_7 \\ d_8 & d_9 & d_{10} & d_{11} \\ d_{12} & d_{13} & d_{14} & d_{15} \end{bmatrix} \in R^{4^k} \times 4^k$$

where $a \in R^{4^{k-1}} \times R^{4^{k-1}}$ is the low frequency component, and $d_i \in R^{4^{k-1}} \times R^{4^{k-1}}$, are the high frequency components.

Fig. 3: Wavelet Domain of Sample Bicubic LR image

From the figure above, it is a clear indicator that the bicubic down sampled data still contains a lot of details. We can compare this with when we try to apply wavelet transform to a real low resolution image.

Fig. 4: Wavelet Domain of Sample Real LR image

Using signal processing method to decompose the sample images above, it is clear that real low resolution image contains a lot less high frequency information. Thus, for our experiment, we would like to remove the details from the original down-sampling image, but we do not want to remove all of the details since each image is different. Thus, to enhance the capability of our deep learning model, we choose to randomly remove the detail information of each high frequency band by setting a conditional random number generation such that:

$$d_i = \begin{cases} 0 & r_i = 0 \\ d_i & r_i = 1 \end{cases}$$

where r_i is a random integer number between $[0, 1]$ for each i. Example below is the image after wavelet transform: We apply the inverse of wavelet transform to the frequency component to get back to the signal domain.

$$S^* = W^T \cdot F \cdot W$$

The process of detail removal is an End-to-End process since we employ the method within the data-loader portion of ESDR model. Below is an example of Before and After Wavelet Transformation of a training data batch.

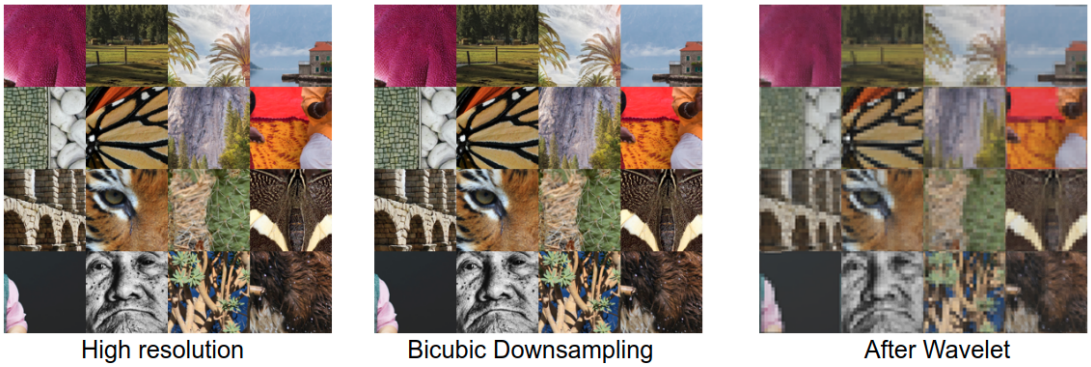

Fig. 5: Training batch sample before and after removing details

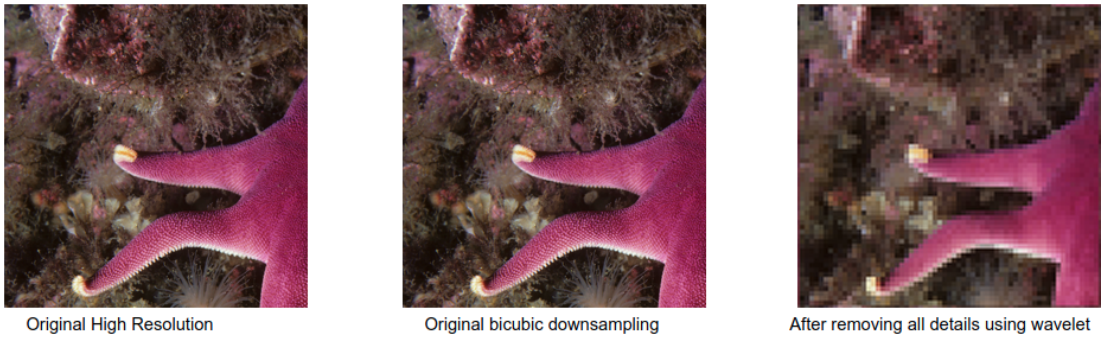

Fig. 6: A closer look comparison before and after removing details

We first train EDSR with the Wavelet Transformed training data to obtain SR images. Once we obtain the initial result, we enhance the SR images to using GAN to generate the details so that it could match the ground truth by minimizing the loss function. Below is the architecture of our proposed model.

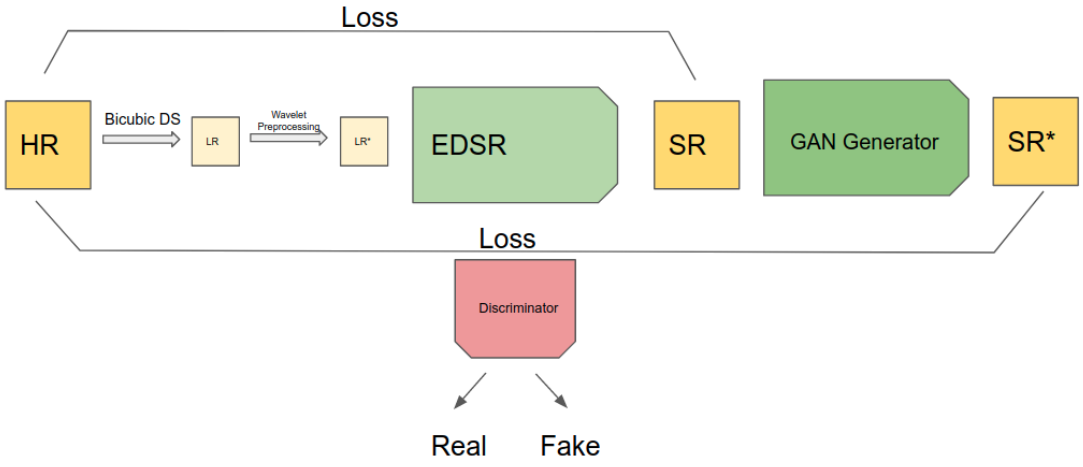

Fig. 7: Proposed Model

IV. Experiment Results

We use DIV2K data as our training input. We load the pre-trained weight of the EDSR model and fine-tuning it with our generate the low-resolution dataset. Below is the sample test between using the original LR vs using Wavelet-generated LR data. The performance of our proposed method on this test data is slightly inferior compared to the EDSR model due to obvious reason; the data we trained with has the details randomly removed from the high-frequency bands, so the EDSR is fine-tuning differently to adjust with the new information.

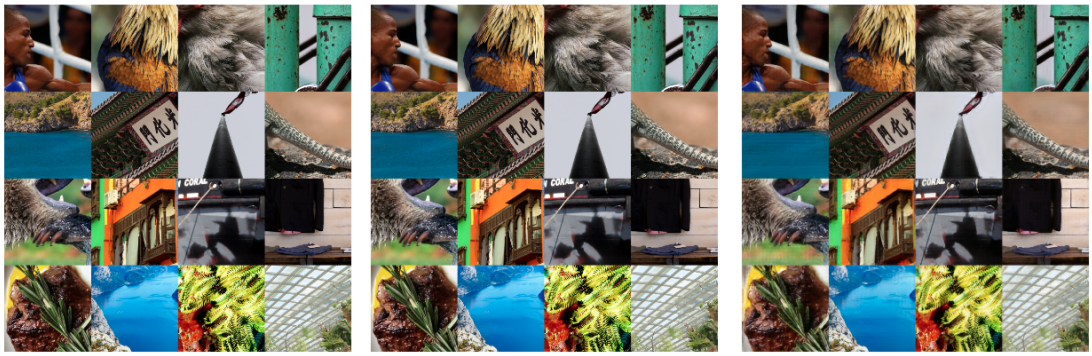

Fig. 8: Testing batch comparision between EDSR vs proposed method

However, as we suggested earlier, the EDSR model works well for bicubic down-sampling data, when applying the model to the real-world low-resolution data set, it would perform poorly since it hasn't learned to generate the "deleted" high-frequency information effectively. Below is an example:

Fig. 9: Comparison between LR vs Wavelet processed-LR

To compare how the high-frequency bands (details) have been generated from both models, we run a sample wavelet decomposition experiment from the results above. It is clear that using our proposed model, the network

can generate detailed information much better which leads to high perceptual performance for real low-resolution images.

Fig. 10: EDSR-GAN result in wavelet domain

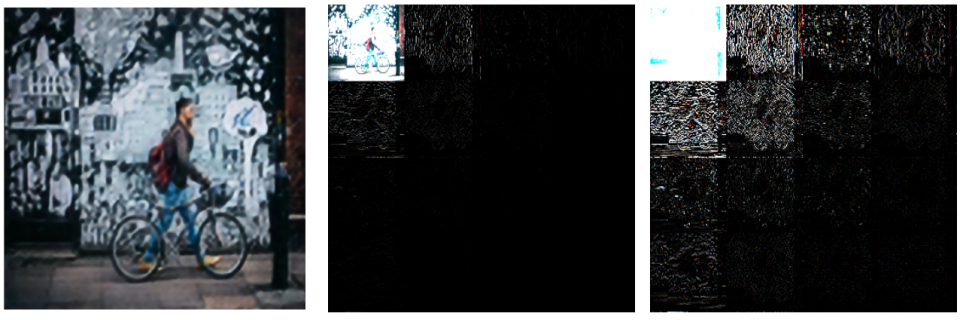

Fig. 11: Wavelet-EDSR-GAN result in wavelet domain

Below is the comparison of our PSNR results for DV2K test set:

EDSR	Wavelet + EDSR	EDSR + GAN	Wavelet + EDSR + GAN
37.305	33.894	30.828	25.025

TABLE I: PSRN comparision

From our experiment, our the PSRN's results are inferior compare to when using bicubic down-sample low-res input. As already mentioned earlier, since our input data has fewer details, it is more difficult for the model to generate all the details back to compare with the ground truth. At the moment of this research, we could not find other data set that has real low-resolution with a HR correspondence, thus we cannot quantitatively compare the results.

V. Conclusion

In this project, we apply the wavelet transformation for super-resolution image preprocessing. We adopt the GAN network structure and EDSR as the generator. Experiments results show that wavelet transformation can improve the perceptual performance of the image compared to the network without wavelet transformation.

For future research consideration, we would like to directly apply the EDSR model to the frequency domain. Instead of completely remove all of the detailed information of the high-frequency bands, we would apply a random threshold to remove some details and train the model to regenerate them back. We assume by training the model this way, the network can learn the mapping easier since it has a direction of which details to generate instead of learning the image as a whole.

REFERENCES

[1] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE transactions on acoustics, speech, and signal processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[2] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, pp. 1–11, 2011.

[3] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[4] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 349–356.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[7] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, 2016, pp. 2802–2810.

[8] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1654–1663.

[9] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3791–3799.

[10] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[11] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4539–4547.

[12] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[14] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 517–532.

[15] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 252–268.

[16] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 864–873.

[17] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1664–1673.

[18] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[19] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. 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*, 2017, pp. 4681–4690.

[21] P. Steffen, P. N. Heller, R. A. Gopinath, and C. S. Burrus, "Theory of regular m-band wavelet bases," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3497–3511, 1993.

[22] Z. Pan, X. Wang, *et al.*, "A wavelet-based non-parametric estimator of the variance function," *Computational Economics*, vol. 15, no. 1/2, pp. 79–87, 2000.