Re-Learning ShiftIR for Super-Resolution of Carbon Nanotube Images

Yoshiki Kakamu[a], Takahiro Maruyama[b] and Kazuhiro Hotta[c]
Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan

Keywords: Super-Resolution, Carbon Nanotube, Shift, SwinIR, Re-Learning.

Abstract: In this study, we perform super-resolution of carbon nanotube images using Deep Learning. In order to achieve super-resolution with higher accuracy than conventional SwinIR, we introduce an encoder-decoder structure to input an image of larger size and a Shift mechanism for local feature extraction. In addition, we propose super-resolution method by re-training to perform super-resolution with high accuracy even with a small number of images. Experiments were conducted on DIV2K, General100, Set5, and carbon nanotube image dataset for evaluation. Experimental results confirmed that the proposed method provides higher accuracy than the conventional SwinIR, and showed that the proposed method can super-resolve carbon nanotube images. The main contribution is the proposal of a network model with better performance for super-resolution of carbon nanotube images even if there is no crisp supervised images. The proposed method is suitable for such images. Effectiveness of our method was demonstrated by experimental results on a general super-resolution dataset and a carbon nanotube image dataset.

1 INTRODUCTION

Deep learning technology for image processing is currently applied not only to the IT field but also to various fields such as civil engineering, and materials engineering. In materials engineering, this study focuses on carbon nanotube (CNT). CNT is nanometer-sized tubular materials composed of carbon, which can be used in a wide variety of applications, such as fuel cells and medical materials. The problem of blurred images often occurs in transmission electron microscope (TEM) observation, which is the main method for CNT characterization. A solution for this problem is to use deep learning to obtain super-resolution images.

In this study, we use a super-resolution method called SwinIR as a baseline. This model is based on the Swin Transformer, which has achieved the highest accuracy in many super-resolution tasks. However, the pre-trained SwinIR was trained on datasets such as plants, animals, and buildings, and not on material images. Therefore, the pre-trained models may fail to super-resolution of material images. Although it is desirable to train on a large number of clear CNT images, it is difficult to collect such data. Therefore, we propose a re-training method in which fine tuning is performed using the images that have already been successfully super-resolved by ShiftIR as teacher images. This method can incorporate CNT-specific features while utilizing generic knowledge from pre-trained models.

In order to achieve super-resolution with higher accuracy than SwinIR, we incorporated an encoder-decoder structure that enables us to input an image of larger size. We also incorporated the Shift mechanism of ShiftViT, which performs better than Swin Transformer.

To evaluate the proposed method, we used the DIV2K, General100 and Set5 datasets, which are commonly used in super-resolution experiments. We also use the CNT image dataset which is the subject of this study. In our experiments, we first evaluated the proposed method on the DIV2K, General100, Set5 datasets to evaluate its super-resolution performance. The results show that the proposed method can achieve higher PSNR accuracy than conventional SwinIR on those general datasets. Next, we conducted experiments on CNT image datasets to

[a] https://orcid.org/ 0000-0001-6698-3634
[b] https://orcid.org/0000-0003-4559-348X
[c] https://orcid.org/0000-0002-5675-8713

Kakamu, Y., Maruyama, T. and Hotta, K.
Re-Learning ShiftIR for Super-Resolution of Carbon Nanotube Images.
DOI: 10.5220/0011708200003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 371-377
ISBN: 978-989-758-634-7; ISSN: 2184-4321

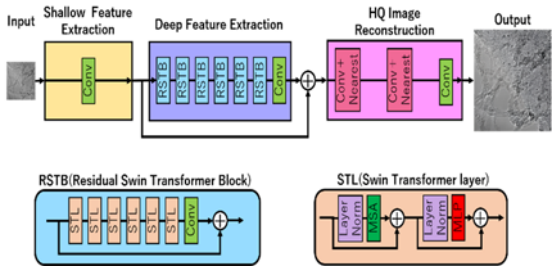

Figure 1: Overview of SwinIR.

evaluate the effectiveness of the proposed method on CNT images and the usefulness of the re-learning based super-resolution. As a result, we confirmed that the proposed method can achieve higher PSNR accuracy than the conventional SwinIR on CNT image datasets. In addition, an ablation study was conducted without the encoder-decoder structure and the Shift mechanism in the proposed method to demonstrate the usefulness of the proposed method, and we show the effectiveness of both devices in our method.

The structure of this paper is as follows. Section 2 describes related works. Section 3 describes the proposed method. Section 4 presents experimental results. Section 5 concludes and discusses future works.

2 RELATED WORKS

We discuss related works of this study. We discuss SwinIR in Section 2.1 and Shift ViT in Section 2.2.

2.1 SwinIR

SwinIR is a super-resolution method based on Swin Transformer. Figure 1 shows the overview of SwinIR, which consists of three parts: a shallow feature extraction part, a deep feature extraction part, and a high-quality image reconstruction part. The shallow feature extraction part extracts shallow features from an input image using convolution layers. The deep feature extraction part consists of Residual Swin Transformer Blocks (RSTB). Each block utilizes multiple Swin Transformer Layers (STLs). In addition, a convolutional layer is added at the end of the block to enhance functionality, and residual connections are used. The high-quality image reconstruction section reconstructs the sum of features from the shallow and deep feature extraction sections into a quadruple-sized image using convolution layers and the Nearest Neighbor method. These mechanisms make SwinIR high-quality super-resolution method by taking advantage of the Swin Transformer. Conventional SwinIR consumes a lot of memory and cannot treat input images of large sizes. Therefore, only small size images can be input, and only a small amount of information on the input image can be utilized. To solve this problem, this paper introduces an encoder-decoder structure that enables the input of large images without increasing memory usage.

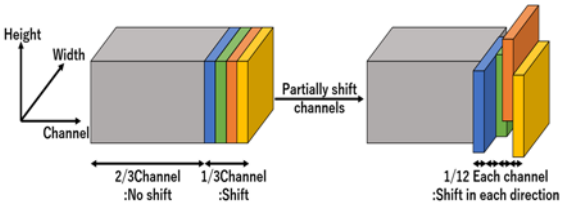

Figure 2: Shift Mechanism.

2.2 ShiftViT

ShiftViT is a variant of Vision Transformer that uses the Shift mechanism instead of Attention in the Swin Transformer, an operation that swaps some channels between adjacent features. It has the advantage that no arithmetic operations or parameters are required. Figure 2 shows the Shift mechanism in the ShiftViT. The Shift mechanism first divides a portion of the input channel into four equal parts. Next, one of divided part is shifted by one pixel to the left, right, top, and bottom.

This operation is a substitute for Attention in Swin Transformer. As a result, ShiftViT achieved better performance than Swin Transformer on the tasks such as image classification, object detection, and segmentation. By replacing the Swin Transformer Layer (STL) in SwinIR with the ShiftViT Layer (SVL), this study enables super-resolution with high accuracy and low memory consumption.

3 PROPOSED METHOD

In this section, we describe the proposed method, Section 3.1 describes super-resolution method by re-learning, and Section 3.2 describes the structure of the proposed ShiftIR.

3.1 Re-Learning Based Super-Resolution

In this study, we propose a re-training based super-resolution method. This method is based on fine-tuning using images that have been successfully

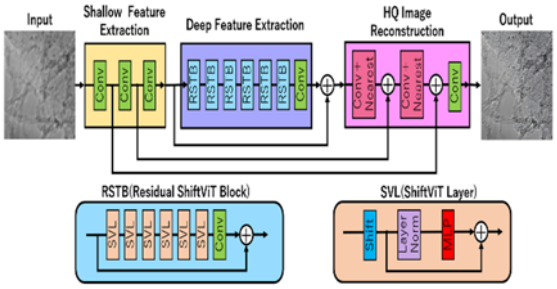

Figure 3: Structural diagram of ShiftIR.

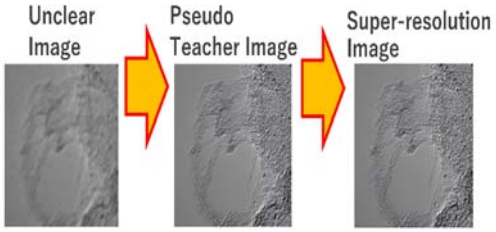

Figure 4: Re-learning based Super-resolution process.

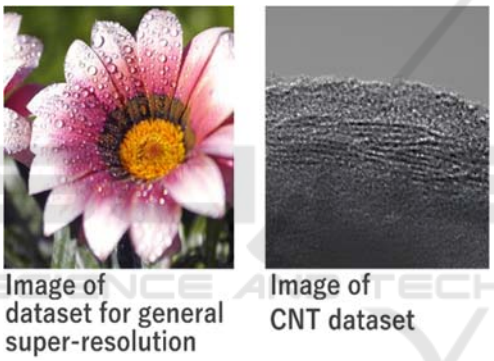

Figure 5: Images in datasets.

Table 1: Results on general super-resolution datasets.

Dataset	Network	Loss		PSNR(dB)	
		BEST	AVERAGE	BEST	AVERAGE
DIV2K	SwinIR	0.00379	0.00928	43.468	40.658
	ShiftIR	0.00127	0.00492	**46.990**	41.273
General100	SwinIR	0.00307	0.00693	46.576	43.393
	ShiftIR	0.00110	0.00435	**50.936**	45.547
Set5	SwinIR	0.00418	0.00882	45.528	43.223
	ShiftIR	0.000226	0.00388	**60.000**	52.514

super-resolved images by a pre-trained model as a teacher image. The process of the super-resolution method based on relearning is shown in Figure 4. CNT images available for this study are blurred images. To successfully conduct experiments under the harsh conditions, we adopted ShiftIR which has been pre-trained in DIV2K to the blurred CNT images. We can obtain better CNT images than the original ones. Thus, the output images are used as pseudo teacher images. However, since the CNT images are very different from the dataset used for pre-training the ShiftIR, some images would fail to be super-resolved. Therefore, fine tuning of ShiftIR is required. This allows us to super-resolve CNT images.The CNT images used in this study are difficult to collect clear images, and it is difficult to create a sufficient dataset for training. In addition, the datasets used for super-resolution are generally images of plants, animals, and buildings, which are not similar to CNT images. Therefore, the usage of only pre-trained models may fail to super-resolve material images. Therefore, by using learned images that have been successfully super-resolved by ShiftIR as teacher images, we can create pseudo-clear teacher images, resulting in high-quality super-resolution. The method can also incorporate CNT-specific features while utilizing generic knowledge from pre-trained models.

3.2 ShiftIR

Figure 3 shows the overview of the proposed ShiftIR. In conventional methods, feature extraction is performed by a single-layer convolution in the shallow feature extraction section. However, only one layer has poor feature extraction capability and cannot make good use of information in the input image. Therefore, encoder-decoder structure is introduced into the shallow feature extraction part and the high-quality image reconstruction part. The input image size is therefore four times larger than that of the conventional method. The proposed method adds a layer of residual connections and convolution of information from the shallow feature extraction section between each layer, while the conventional method only has two layers of convolution and the nearest neighbor as high-quality image reconstruction sections. The proposed method, however, adds a residual connection and a convolution between each layer. This allows us to take advantage of features in large images without increasing memory consumption.

Since the conventional Swin Transformer used self-attention in a window. However, attention between far points in a window may not be useful for super-resolution. We consider that local relationship is more important for super-resolution. For example, if we try to super-resolve the petals in the left image in Figure 5, what is important is the shape of the surrounding petals and the border with the center of the flower. On the other hand, the relationship with the background and leaves is not so important. Thus,

Figure 6: Super-resolution results by ShiftIR.

local relationships are more important for super-resolution. Local relationships are also important in the CNT image shown in Figure 5 because of the importance of separating CNTs from the background and the nature of the CNTs being formed in a series. Thus, we use Shift layer which excels in more localized processing to achieve more accurate super-resolution processing. Figure 3 shows the proposed ShiftIR that we use local Shift of feature maps.

By using an encoder-decoder structure, ShiftIR enhances the shallow feature extraction part of the image by inputting a large size image and effectively utilizing its information. In addition, by using the Shift mechanism, local features are extracted and the deep feature extraction part is enhanced.

Table 2: Result of CNT image dataset.

Network	Loss		PSNR(dB)	
	BEST	AVERAGE	BEST	AVERAGE
SwinIR	0.00418	0.00882	**45.528**	43.223
ShiftIR	0.000135	0.000483	**59.874**	57.317

4 EXPERIMENTS

In this section, we present the experimental results. Section 4.1 describes the carbon nanotube images. Section 4.2 presents the experimental results on general super-resolution dataset. Section 4.3 shows the experimental results for the CNT image dataset. Section 4.4 shows ablation study.

4.1 Carbon Nanotube (CNT) Image

Carbon nanotube (CNT) is nanometer-sized tubular materials composed of carbon, which are used in a wide variety of applications, including fuel cells and medical materials. However, transmission electron microscopy, which is the primary method for evaluating CNTs, often results in blurred images. The solution to this problem is to use ShiftIR to obtain super high resolution images of CNTs. In addition, the CNT images used in this study are not suitable for using as teacher images because they are unclear. Therefore, the proposed super-resolution method based on retraining was used to create pseudo teacher images and use them for training.

4.2 Results on General Super-Resolution Datasets

The experimental results on the general super-resolution dataset (DIV2K, General100, Set5) are shown in Figure 4. In these experiments, we trained each method till 50,000 epochs. Figure 6 shows that the overall quality of the images by ShiftIR is improved.

Loss and PSNR values for SwinIR and ShiftIR for the general super-resolution dataset are shown in Table 1. Table 1 shows that the proposed ShiftIR method and the conventional SwinIR method have 40 dB or higher values. Since 30 dB or higher is generally considered to be high image quality, all of them can be said to have high image quality. In addition, the PSNR values of the proposed method are approximately 3.5 dB higher on the DIV2K dataset, 4.3 dB higher on the General100 dataset, and 14.5dB higher on the Set5 dataset compared to the conventional method.

From these results, we see that the proposed method is able to achieve super-resolution with better accuracy. This is due to the fact that the encoder-decoder structure of ShiftIR extracts the overall features of larger input image and utilizes them for image reconstruction, thereby enhancing the shallow feature extraction and image reconstruction parts. The deep feature extraction part was also enhanced by changing from the Swin Transformer to the Shift mechanism.

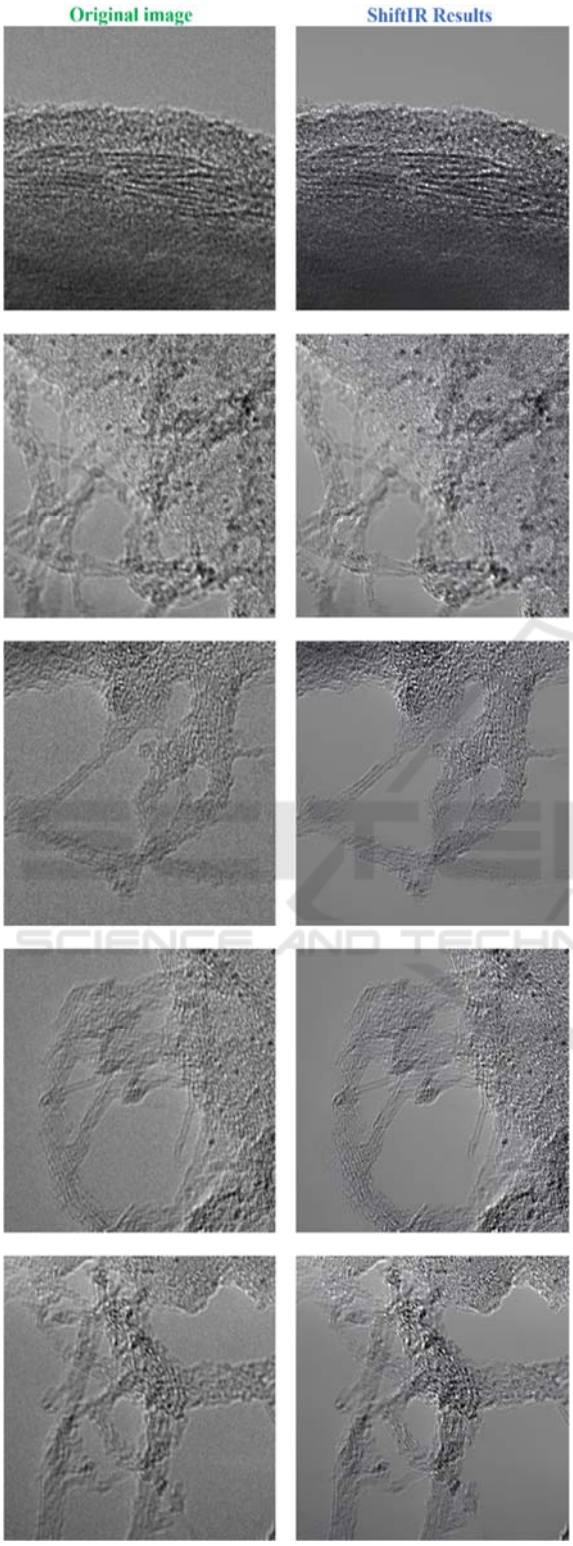

Figure 7: Super-resolution of CNT images by ShiftIR.

Table 3: Ablation Studies.

Ablation	Loss		PSNR(dB)	
	BEST	AVERAGE	BEST	AVERAGE
Encoder-Decoder	0.000138	0.00161	**56.990**	51.307
Shift	0.000166	0.000870	**56.990**	52.216

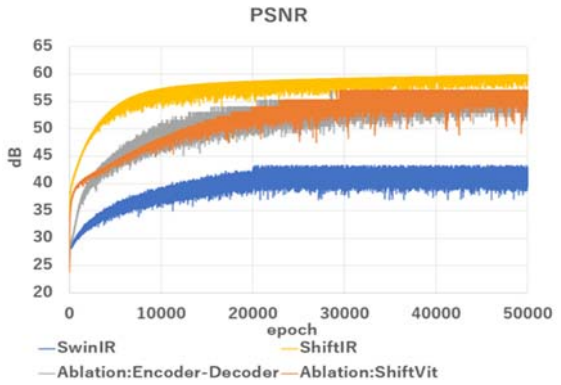

Figure 8: PSNR for each method on CNT dataset.

4.3 Results on CNT Image Dataset

The usefulness of the proposed method was confirmed in the experiments on the datasets for general super-resolution in section 4.2. Therefore, we adopt the proposed method to CNT images, which are the main subject of this study. We also adopt the re-learning based super-resolution method to the CNT image dataset in order to confirm the usefulness. The experimental results on the CNT image dataset are shown in Figure 7, and the Loss and PSNR values of SwinIR and ShiftIR are shown in Table 2. We trained each method till 50,000 epochs. Figure 6 shows that the image with ShiftIR has higher quality, and the background noise disappears. In addition, as shown in Table 2 and Section 4.2, both ShiftIR and SwinIR are higher than 40 dB. The PSNR value of the proposed method is about 16.7 dB higher than that of the conventional SwinIR.

From these results, we see that the proposed method is able to achieve super-resolution with higher quality. This indicates that the proposed method is more useful than conventional methods on CNT image datasets. In addition, the higher improvement on CNT image dataset than general super-resolution datasets though no clear teacher images exist. This indicates that the re-training based super-resolution method is useful.

4.4 Ablation Study

Ablation study of the proposed ShiftIR is performed to verify which mechanism is effective. The CNT

image dataset is used in this experiment. One of the two mechanisms in the proposed method, the encoder-decoder structure and the Shift mechanism, was removed in this experiment. The experimental results are shown in Table 3 and the graphs of each PSNR value are shown in Figure 8.

We also trained each method till 50,000 epochs. Table 3 and Figure 8 show that the results excluding each factor are higher than the conventional method but less accurate than the proposed method.

The PSNR values for the encoder-decoder structure alone and the Shift mechanism alone are nearly same. These results indicate that both the encoder-decoder structure and the shift mechanism are important for ShiftIR. These experiments confirm the effectiveness of the encoder-decoder structure and the shift mechanism.

The encoder-decoder structure improved accuracy because it was able to take advantage of the features of large images. When the height and width of a square image is halved, the number of pixels in the image is reduced to one-fourth. This means that the image information is also reduced to one-fourth, and conversely, the image information is increased by a factor of four when the image size is doubled. Therefore, an encoder-decoder structure that can handle four times as many input images as conventional methods is considered an important element for ShiftIR.

The reason for the improved accuracy with the Shift mechanism can be attributed to the improved performance of local feature extraction. Local relationships are more important than the relationships between distant locations in an image, because features around objects are important for super-resolution. Especially, local relationships are especially important due to the characteristics of CNTs, so the Shift mechanism with its high local feature extraction capability is quite effective.

ShiftViT's Shift mechanism has fewer parameters than Swin Transformer's Attention mechanism and can build deeper models within a limited amount of computation, making it superior for spatial feature extraction. Therefore, the Shift mechanism is considered as an important element for ShiftIR because it can extract more features than conventional methods. Based on these factors, we believe that ShiftIR can achieve higher resolution than conventional methods due to the synergistic effect of the encoder-decoder structure and the Shift mechanism.

5 CONCLUSION

In this paper, we proposed ShiftIR, which introduces an encoder-decoder structure and the Shift mechanism to the conventional SwinIR for super-resolution of carbon nanotube images. In addition, we proposed a re-training based learning method to perform super-resolution with high accuracy even with a small amount of low quality images. Experimental results show that ShiftIR can perform super-resolution with a maximum accuracy of 59dB on both general super-resolution datasets and CNT image datasets, which is higher than the accuracy of conventional methods. Through the ablation study, we see that both the encoder-decoder structure and the shift mechanism are effective for super-resolution. In addition to CNTs, there are the other fields in materials engineering such as catalyst images that require super-resolution, and we would like to develop models for those fields.

REFERENCES

Vaswani,A.,Shazeer,N.,Parmar,N.,Uszkoreit,J.,Jones,L.,Gomez,A. N.,Polosukhin,I.,"Attention is all you need.", International Conference on Neural Information Processing Systems, pp. 6000-6010, 2017

Dong,C.,Loy,C.C.,He,K.,Tang,X.,"Learning a deep convolutional network for image super-resolution. ", European Conference on Computer Vision, pp. 184-199 ,2014

Yu,C.,Xiao,B.,Gao,C.,Yuan,L.,Sang,N.,Wang,J.,"Lite-hrnet: A lightweight high-resolution network.", IEEE/CVF Conference on Computer Vision and Pattern Recognition,pp. 10440-10450 ,2021

Chun,P.J.,Shota,I.,Tatsuro,Y., "Automatic detection method of cracks from concrete surface imagery using two-step light gradient boosting machine.", Computer-Aided Civil and Infrastructure Engineering, pp. 61-72 ,2021

Dung,C.V.,"Autonomous concrete crack detection using deep fully convolutional neural network.",Automation in Construction, pp. 52-58, 2019

Ramprasad,R., Batra,R., Pilania,G., Mannodi-Kanakkithodi,A., Kim,C., "Machine learning in materials informatics: recent applications and prospects.", NPJ Computational Materials, pp. 1-13, 2017

Kim,C., Chandrasekaran,A., Huan,T.D., Das,D., Ramprasad,R., "Polymer genome: a data-powered polymer informatics platform for property predictions.", The Journal of Physical Chemistry C, pp. 17575-17585, 2018

Sumio,I.,"Carbon Nanotube",Electron microscope 34-2, pp.103-105, 1999

Takahiro,K., Junji,N., "Specificity of electrocatalysts using carbon nanotubes as supports.", Surface science 32-11, pp.704-709, 2011

Heister,E., Brunner,E.W., Dieckmann,G.R., Jurewicz,I., Dalton,A.B., "Are carbon nanotubes a natural solution? Applications in biology and medicine. ", ACS Applied Materials & Anterfaces, pp.1870-1891, 2013

Liang J., Cao,J., Sun,G., Zhang,K., VanGool,L., Timofte,R., "Swinir: Image restoration using swin transformer", IEEE International Conference on Computer Vision, pp.1833-1844, 2021.

Liu,Z., Lin,Y., Cao,Y., Hu,H., Wei,Y., Zhang,Z., Lin,S., Guo,B., "Swin transformer: Hierarchical vision transformer using shifted windows", IEEE International Conference on Computer Vision, pp.10012-10022, 2021

Mao,X.,Shen,C.,Yang,Y.B.,"Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections.", Advances in Neural Information Processing Systems, pp. 2810–2818 ,2016.

Wang,G., Zhao,Y., Tang,C., Luo,C., Zeng,W., "When Shift Operation Meets Vision Transformer An Extremely Simple Alternative to Attention Mechanism.", arXiv preprint arXiv:2201.10801, 2022

Agustsson,E.,Timofte,R.,"Ntire 2017 challenge on single image super-resolution: Dataset and study.", IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126-135 , 2017

Dong,C.,Loy,C. C.,Tang,X.,"Accelerating the super-resolution convolutional neural network.", European Conference on Computer Vision, pp. 391-407, 2016

Bevilacqua,M., Roumy,A., Guillemot,C., Alberi-Morel,M.L., "Low-complexity single-image super-resolution based on nonnegative neighbor embedding.", British Machine Vision Conference, pp.135.1-135.10, 2012

Dosovitskiy,A., Beyer,L., Kolesnikov,A., Weissenborn,D., Zhai,X., Unterthiner,T., Houlsby,N., "An image is worth 16x16 words: Transformers for image recognition at scale.", International Conference on Learning Representations, 2021

Kobako,M., "Electronic Documents Image Compression Guidelines", IM Monthly Vol.50 No.5, pp.21-24, 2011.