APPLICATION OF PRE-TRAINED REAL-WORLD SUPER RESOLUTION MODELS TO OPTICAL SATELLITE IMAGE

Takayuki Shinohara, Riho Ito, Yohei Kobayashi, Toshiaki Satoh, Yasunobu Shimazaki, Sho Nakamura

PASCO Corporation, Tokyo, Japan

Commission III, WG III/1

KEY WORDS: Satellite Image, Super Resolution, BSRGAN, Real ESRGAN, SWIN IR, Image Restoration.

ABSTRACT:

The single image super-resolution (SISR) technique refers to improving the resolution over the original image. In recent years, we use deep learning-based convolutional neural networks to improve the spatial resolution of images more reasonably. To train such deep learning models, we use training samples consisting of the original HR images and LR images obtained by bicubic downsampling. However, this method of training data using downsampling has a negative impact when applying the trained model to real images. That is because the downsampling function that occurs in the real image is unknown, and the hypothetically created LR image does not represent the resolution degradation that can realistically occur. Therefore, SISR methods that use realistic degradation called real-world super-resolution (RWSR) have been proposed. In this paper, we investigate how such RWSR methods using realistic degradation affect the SISR performance of satellite images. The results of applying the trained model to optical satellite images show that the RWSR method is not the most effective way to handle optical satellite images when compared to the deep learning method without modeling the degradation. In particular, we showed that the effect of RWSR with not only upsampling but also noise and blur removal is significant in the visibility of optical satellite images. Moreover, pre-trained RWSR models can be an aid in visually deciphering ground objects.

1. INTRODUCTION

The Single Image Super-resolution (SISR) is the process of applying an algorithm to a low resolution (LR) image to derive a higher resolution (HR) image. The SISR approach does this using a single LR image and is considered a classic problem in computer vision (Dong et al., 2015). In parallel with many other computer vision problems, SR approaches employing deep convolutional neural networks (SRCNNs) have outperformed other techniques in the last few years. Generalizing the SISR task, we can define that the task is to restore the high-resolution (HR) image from a low-resolution (LR) image. In this case, LR images are generated by applying a degradation function f to the HR image like below:

$$\mathrm{LR} = f(\mathrm{HR}). \tag{1}$$

In other words, the definition of the degradation function f is important to train a SISR model. In general, in training a naive SR model, an LR image is created by bicubic down-sampling as the degradation function. However, since the degradation of the actual image and the degradation of the image obtained by bicubic downsampling are far apart, applying the trained model to another actual image will cause problems. Therefore, the real-world super-resolution (RWSR) method (Figure 1) was proposed. RWSR is a super-resolution method that proposes a degradation function that models the degradation that can occur in real situations in more detail. As a degradation function, the RWSR method use not only reduces the resolution of the images but also adds blur and noise. By training to restore the original HR image from the LR image created by such a degradation function, the prediction performance was improved for any input LR image.

The single image super-resolution (SISR) technique using deep learning has been applied to remote sensing (Benecki et al., 2018, Lu et al., 2019, Lanaras et al., 2018). There is a wide

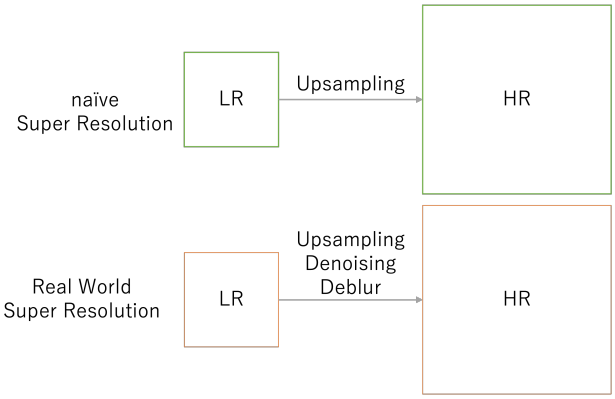

Figure 1. Overview of Real-World Super Resolution.

range of applications of SISR technology for satellite imagery, such as further increasing the resolution of sub-meter imagery like WorldView-2/3, and increasing the resolution of free imagery such as Landsat and Sentinel. The same problem of degradation function occurs when applying the SISR technique to such optical satellite images. Consider the case where an optical satellite image is upsampled to a resolution equivalent to that of an aerial photo, for example. Since optical satellite images have similar observation conditions to each other, it is possible to directly pair an LR satellite image with an HR satellite image and train SR task (Pouliot et al., 2018, Shin et al., 2020, Salgueiro Romero et al., 2020). However, optical satellite images and aerial photographs do not correspond pixel by pixel due to misalignment issues and distortions, and thus cannot be directly trained for SISR.

Therefore, it is practical to apply the trained model to optical satellite images. In this case, it is necessary to remove various degradations contained in the optical satellite image when

applying SISR. Specifically, optical satellite images include atmospheric effects and sensor noise as well as resolution. Therefore, training a SISR model by simply pairing the downsampled image with the original image is difficult to apply to optical satellite images. Thus, we decided to apply a method called the *real-world super-resolution (RWSR)* method (Figure 1). RWSR not only reduces the resolution of the image but also removes blur and noise. In this paper, to confirm the effectiveness of applying the RWSR method to satellite images, we apply a deep learning model trained not on satellite images but natural images.

2. SUPER-RESOLUTION

The single image super-resolution (SISR) task is the problem of improving the spatial resolution of an image by considering only the information available from the input image itself and the knowledge obtained in the past in the form of algorithms or trained models. Since the publication of SRCNN (Dong et al., 2015), there has been a lot of research on this SISR technique using deep learning. SRCNN is an approach to obtain HR images from LR images by extracting the relationship between LR images and HR images using a Convolutional Neural Network (CNN). By incorporating deep learning, it is characterized by its ability to estimate HR images with higher accuracy than conventional methods. In addition, EnhanceNet (Sajjadi et al., 2017) proposed a new SISR method using adversarial training in addition to considering the perceptibility and texture consistency of images as loss functions. Moreover, a new SISR method (Ledig et al., 2017) is proposed using adversarial training, which enables us to obtain clearer images than other SISR methods such as SRCNN.

To train a CNN for SISR, we need image pairs that represent the same scene at different resolutions. The LR image is the input to the network, while the HR image is the desired output of the network for the corresponding input image. Therefore, the training process involves adjusting the weights of the CNN to minimize the error between the output image produced by the network and the corresponding ground truth HR image. When creating training data, it is common practice to obtain pairs of images by artificially downsampling the target HR image to obtain the corresponding LR source image. The ratio of the downsampling depends on the magnitude of the expected SISR effect and is usually between two and four times the original resolution. The appropriate choice of loss function during training is an important issue when training CNNs, and the mean square error (MSE) or mean absolute error (MAE) between the predicted HR image and the ground truth HR image is used. Generative Adversarial Networks (GANs) have also received a lot of attention from researchers and can be understood as a different training method closely related to the loss function. SRGAN (Ledig et al., 2017) is an example of using GAN training for SISR, where the discriminator network that should classify between predicted HR image and ground truth HR images is also trained at the same time. This adversarial approach yielded good visual results but was not directly reflected in the numerical metrics in pixel-wise similarity. In addition, an ESRGAN (Wang et al., 2018) based on the SRGAN was proposed. ESRGAN improved the network architecture, adversarial loss, and perceptual loss of SRGAN. ESRGAN used a deeper model using Residual in Residual Dense Block (RRDB) without batch normalization layers in the generator and Relativistic Discriminator in the discriminator.

The ultimate goal of SISR is real-world applications (Chen et al., 2021). However, these methods do not work well on real images because they are primarily designed for bicubic downsampling. This stems from the fact that they consider the inverse mapping of an ideal HR image to an LR image with a degradation function applied. This is because, in general, the blur kernel plays an important role in the success of SISR methods, and the bicubic kernel is too simple for a degradation function (Zhou and Süsstrunk, 2019). To solve this problem, some researchers use real-world degradation functions with the blur kernel and noise models (Zhang et al., 2021, Wang et al., 2021a, Liang et al., 2021). These studies are called real-world super-resolution(RWSR). Since both LR and HR images may be noisy or blurred, it is not necessary to adopt the order of blurring, downsampling, noise addition as in the conventional degradation function when generating LR images. Since the blur kernel space of the conventional degradation model should vary with scale, it is difficult to determine a very large scale factor in practice. Bicubic decomposition is hardly suitable for real LR images, but it can be used for data augmentation, and indeed for super-resolution of clear and sharp images. To solve this problem, Zhang et al. proposed a practical degradation model BSRGAN for RWSR and showed its effectiveness on real images (Zhang et al., 2021). Subsequently, some methods for RWSR such as Real ESRGAN (Wang et al., 2021a) and Swin IR (Liang et al., 2021) have been proposed. It is now possible to construct realistic degradation functions by randomly recombining blur, downsampling, and noise, or by applying each several times. The trained model can then be applied to real images without any degradation in performance. In this study, we confirmed the effectiveness of this RWSR technique by applying it to optical satellite images.

3. PRETRAINED REAL-WORLD SR MODELS

In this study, we use four pre-trained super-resolution models and apply them to optical satellite images(Table 1).

3.1 ESRGAN(baseline)

3.1.1 Deterioration data One of the famous non RWSR methods is ESRGAN(Wang et al., 2018). We used pretrained ESRGAN as the baseline method. In ESRGAN, the input LR image is created by downsampling the HR image by 1/4. The downsampling method uses bicubic. Therefore, only the degradation of resolution is considered as the degradation of the image.

3.1.2 Network ESRGAN is based on SRGAN(Ledig et al., 2017) architecture. ESRGAN adopts the Residual in Residual Dense Block (RRDB) architecture and removes Batch Normalization (BN). By adopting the Relativistic average GAN (RaGAN) discriminator, ESRGAN achieves the expressive power to generate realistic HR images. ESRGAN is the model that won the PIRM 2018 SR Challenge, a super-resolution competition.

3.2 BSRGAN

3.2.1 Deterioration In BSRGAN (Zhang et al., 2021), blur, downsampling, and noise were identified as important elements of the degradation function. A direct way to improve the utility of the degradation function is to make the space that the degradation function of the three important elements can take a large and realistic as possible. By adopting a random shuffling

Table 1. Pre-trained super-resolution models are used in this paper.

		ESRGAN	BSRGAN	Real ESRGAN	SWIN IR
Deterioration	Down sample	✔	✔	✔	✔
	Blur	-	✔	✔	✔
	Noise	-	✔	✔	✔
	Method	-	Random	2nd order	Random
Network	Base method	SRGAN	ESRGAN	ESRGAN	u-shaped transformer
	Originality	•Removing BN •RRDB •RaGAN	-	•U Net Discriminator •Spectral Normalization	•SWIN Transformer

strategy for the three critical elements, we further enlarge the degradation function space. In other words, an LR image could be a "noise→ downsample→blurred version of an HR image" or a "downsample →noise → blurred version of an HR image".

3.2.2 Network The network design used the same architecture as ESRGAN(Wang et al., 2018). Experiments on synthetic and real image datasets show that the trained model performs favorably on images corrupted by a variety of degradations.

3.3 Real ESRGAN

3.3.1 Deterioration In Real ESRGAN (Wang et al., 2021a), they claimed that the classical first-order degradation function could not be modeled well in the real situations proposed in BSRGAN (Zhang et al., 2021). Real ESRGAN proposed a new second-order degradation model. A second-order model is one that repeats n degradation processes, and each degradation process adopted a classical degradation model with the same procedure but different hyperparameters. The term "second-order" here mainly refers to the implementation time of the same operation, as opposed to that used in mathematical functions. In Real ESRGAN, the degradation function was defined using the method of iterative application of blurring, downsampling, and noise.

3.3.2 Network The generator uses the same network architecture as ESRGAN (Wang et al., 2018), but a few changes are proposed in Real ESRGAN. First, they changed the discriminator from a simple CNN to UNet (Schonfeld et al., 2020) to improve the discriminator capability. In addition, they used spectral normalization in the discriminator to avoid large gradients and stabilize the training.

3.4 SWIN IR

3.4.1 Deterioration The data preparation method of BSRGAN (Zhang et al., 2021) is used to train SWIN IR (Liang et al., 2021).

3.4.2 Network SWIN IR is based on a u-shaped transformer (Wang et al., 2021b). Additionally, SWIN IR (Liang et al., 2021) is a Swin Transformer (Liu et al., 2021)-based image restoration model. This model is composed of three parts described as below.

1. shallow feature extraction
2. deep feature extraction
3. HR reconstruction module

In particular, for deep feature extraction, a stack of Residential Swin Transformer Blocks (RSTBs) is used, where each RSTB consists of a Swin Transformer (Liu et al., 2021) layer, a convolutional layer, and a residual connection.

4. EXPERIMENTAL RESULTS

In order to verify the effectiveness of the method on optical satellite images, we conducted a validation experiment using actual optical satellite images. In this section, we describe the dataset, the experimental setup, and the super-resolution results of satellite images.

4.1 Dataset

The WorldView-2 from MAXAR (Panchromatic) image shown in Table 2 was prepared as the input image for the experimental evaluation. The ground sample distance (GSD) was 30 cm and the images were taken from the typical area in Japan. The original satellite images were divided into small patches, which cannot be input into the deep learning model in its original size, so we spilled it into patches of a certain size. The patch size was 256×256. These patches were used as the input images for experimental evaluation.

Table 2. Experimental Dataset. These images have 30 cm GSD and were collected with the WorldView-2 sensor from MAXAR.

satellite	ch	GSD	#patches
Worldview-2	RGB	30 cm	100

4.2 Experimental Setup

For the experiments, only the trained model was modified while the inference code and data preprocessing were kept in common in order to compare equal super-resolution results. The implementation used in this experiment was KAIR[1]. We used ABCI of the National Institute of Advanced Industrial Science and Technology(AIST) was used as the computing environment.

4.3 Results

4.3.1 Comparison of pre-trained models We qualitatively evaluated the characteristics of each pretrained SISR model. The results super-resolution were performed on the image for experimental evaluation using each pretrained model with scale factor 4. Since there was no HR image to be ground-truth, we have visually confirmed the results to evaluate. The super-resolved results are shown in Figure 2.

Since the GSD of the input image of WorldView-2 is 30 cm, the output of the trained model increases the resolution by a scale factor of 4, resulting in a ground resolution equivalent to 7.5 cm. As a comparison, Figure 2 shows the results of super-resolution for the input image using five different methods. Figure 2 shows the results of bilinear upsampling, the results of ESRGAN as an

[1] `https://github.com/cszn/KAIR`

Figure 2. The results of super-resolution with scale factor are 4. From left to right: input image, upsampled with bilinear interpolation, ESRGAN (Wang et al., 2018), BSRGAN (Zhang et al., 2021), Real ESRGAN (Wang et al., 2021a), SWIN IR (Liang et al., 2021). BSRGAN, Real ESRGAN, SWIN IR are Real World SR methods. The images are obtained WorldView-2 sensor from MAXAR.

example of a method that was naive SISR, and the results of the RWSR methods BSRGAN, Real ESRGAN, and SWIN IR. The overall trend was that both the bilinear upsampling and ESRGAN methods perform super-resolution while retaining the blur in the input image. In other words, they simply improve the resolution of the input image without departing from the original image. However, by using real-world super-resolution methods such as BSRGAN, Real ESRGAN, and SWIN IR, we were able to improve the resolution while removing the blur in the input image, making it easier to visually read the edges of buildings. This improvement invisibility can be attributed to the fact that the method of creating the training data incorporates not only resolution degradation but also various degradations of blur and noise that can occur in reality. However, the distortions of the original satellite image are still emphasized, so care must be taken when visually reading the image.

We compared the results of the RWSR methods BSRGAN, Real ESRGAN, and SWIN IR. We show in Figure 3 the results of a typical ground object. First, we discussed the results for buildings. For buildings, all methods improved the visibility of the edges of the buildings. Next, the white line of the parking lot is shown. The visibility of the white line was improved by all methods, but especially by using Real ESRGAN, we could obtain an image with enhanced edges. This is because a more realistic method of creating the degradation function, which was not used in BSRGAN and SWIN IR, was incorporated in the creation of the dataset. Generally, the results of pretrained RWSR model are expected that will assist in visual reading.

However, some cases of the ground objects, which are unique to satellite images, could not be sufficiently obtained because the RWSR models were trained on only natural images. For example, artificial objects were observed in the white lines of the road. This is thought to reflect the distortion of satellite images when they are converted to orthoimages. In the super-resolution results of vegetation, the images tended to be flat in general. This may be because the leaves of the trees were treated as high-frequency noise and were not emphasized during the super-resolution process.

4.3.2 Effect of scale factor Next, we investigated the effect of the scale factor of RWSR. We applied two times pretrained Real ESRGAN with scale factor 2(Real ESRGAN x2x2 in Figure 4) and a pretrained Real ESRGAN with scale factor 4(Real ESRGAN x4 in Figure 4)).

A visual comparison of these results showed no significant differences. This is because the training process of Real ESRGAN removes image degradation (noise, blur, etc.) that may occur in the real world, and thus removes the products generated by the second time 2x super-resolution.

4.3.3 Poll-based evaluation In addition, the results of the visual evaluation are shown in the Table 3. The evaluation method was based on the number of votes that selected the most natural super-resolution result in the form of a questionnaire for ten images in Figure 2. The polling was limited to the RWSR method (BSRGAN, Real ESRGAN, SWIN IR) in ten different situations, the Real ESRGAN x2x2 images from the previous section were also included. For this evaluation, we focused on the textures, edges such as building roofs, and natural objects such as vegetation. The overall tendency was that the SWIN IR method and the BSRGAN method were selected. Thus, the RWSR model trained on natural images was applied to optical satellite images for super-resolution, and good results were obtained in the visual evaluation. The reason why Real ESRGAN was not polled was the stronger deterioration function than BSRGAN and SWIN IR. Real ESRGAN defined a degradation function that can handle a variety of degradations that occur in the real world, so it was thought to have worked to eliminate the texture of the satellite image.

Pretrained RWSR models can be attributed to the presence of features such as building edges, road edges, and roof textures that appear in the satellite image within the domain of the natural image used for training. These results suggest that deep learning models trained on natural images are expected to be effective for various tasks in satellite imagery.

5. CONCLUSION

In this paper, single-image super-resolution (SISR) of satellite images is achieved by applying a trained model of a deep learning method aimed at real-world super-resolution. SISR of satellite images is a task that not only restores resolution but also removes degradation functions that include various factors such as standby effects and sensor effects. Therefore, applying a SISR model trained on a pair of pseudo-LR images (e.g., bicubic downsampling) and a base image will not be able to cope with the degradation of satellite images that occurs in reality. Therefore, we adapted the real-world super-resolution (RWSR) method to satellite images by making the degradation function correspond to the real world. To verify the effectiveness of such deep learning models for RWSR on satellite images, we used the trained models of BSRGAN, Real ESRGAN, and SWIN IR, which were trained on natural images. By applying these RWSR trained models, we were able to achieve super-resolution of the satellite images and improve the visibility by removing noise. This means that even if a model trained on natural images is used, the trained model has already acquired universal super-resolution for the image itself, and therefore, the edge enhancement and noise removal required for super-resolution of satellite images can be achieved. This suggests the effectiveness of applying the trained model of natural images to satellite images as well.

However, since the RWSR model used in this paper was trained on natural images, it may not provide sufficient super-resolution results due to the difference in the domain from satellite images. This problem can be solved by applying the degradation function not only to natural images but also to satellite images and aerial photos while training RWSR.

ACKNOWLEDGEMENTS

We would like to thank the WorldView-2 data were obtained from the Maxar Technologies. Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

REFERENCES

Benecki, P., Kawulok, M., Kostrzewa, D., Skonieczny, L., 2018. Evaluating super-resolution reconstruction of satellite images. *Acta Astronautica*, 153, 15–25.

Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R. E., Zhu, C., 2021. Real-world single image super-resolution: A brief review. *Information Fusion*.

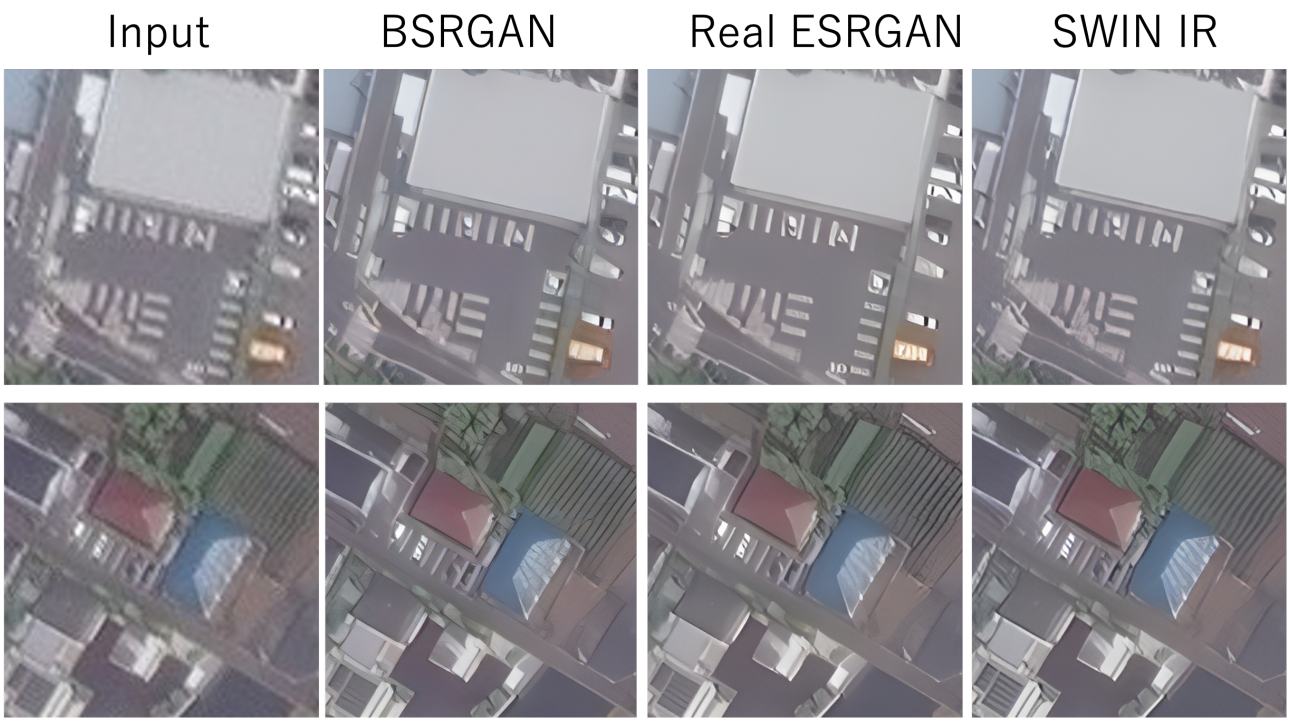

Figure 3. The typical case of real-world super-resolution results at scale factor 4. From left to right: input image, BSRGAN (Zhang et al., 2021), Real ESRGAN (Wang et al., 2021a), SWIN IR (Liang et al., 2021). The images are obtained WorldView-2 sensor from MAXAR.

Figure 4. The comparison of the scale factor. The left figure shows the result of the trained model with a scale factor of 4. The right figure shows the results of two iterations of the trained model with a scale factor of 2. The images are obtained WorldView-2 sensor from MAXAR.

Dong, C., Loy, C. C., He, K., Tang, X., 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2), 295–307.

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., Schindler, K., 2018. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146, 305–

Table 3. The poll-based evaluation of legibility improvement in the typical RWSR results2. We valuated ten different situations.

	BSRGAN	Real ESRGAN x4	Real ESRGAN x2x2	SWIN IR
Number of top votes	1	1	0	9
Percentage of votes received	22%	8%	4%	66%

319.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. Swinir: Image restoration using swin transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.

Lu, T., Wang, J., Zhang, Y., Wang, Z., Jiang, J., 2019. Satellite image super-resolution via multi-scale residual deep neural network. *Remote Sensing*, 11(13), 1588.

Pouliot, D., Latifovic, R., Pasher, J., Duffe, J., 2018. Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training. *Remote Sensing*, 10(3), 394.

Sajjadi, M. S. M., Schölkopf, B., Hirsch, M., 2017. EnhanceNet: Single Image Super-Resolution through Automated Texture Synthesis. *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, 4501–4510.

Salgueiro Romero, L., Marcello, J., Vilaplana, V., 2020. Super-resolution of sentinel-2 imagery using generative adversarial networks. *Remote Sensing*, 12(15), 2424.

Schonfeld, E., Schiele, B., Khoreva, A., 2020. A u-net based discriminator for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8207–8216.

Shin, C., Kim, S., Kim, Y., 2020. From planetscope to worldview: Micro-satellite image super-resolution with optimal transport distance. *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 898–902.

Wang, X., Xie, L., Dong, C., Shan, Y., 2021a. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1905–1914.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C., 2018. Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wang, Z., Cun, X., Bao, J., Liu, J., 2021b. Uformer: A General U-Shaped Transformer for Image Restoration. *arXiv preprint 2106.03106*.

Zhang, K., Liang, J., Van Gool, L., Timofte, R., 2021. Designing a practical degradation model for deep blind image super-resolution. *arXiv preprint arXiv:2103.14006*.

Zhou, R., Süsstrunk, S., 2019. Kernel modeling super-resolution on real low-resolution images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2433–2443.