RAT-SRGAN: RESIDUAL-AWARE TRANSFORMER-BASED SUPER-RESOLUTION GENERATIVE ADVERSARIAL NETWORK FOR SATELLITE IMAGERY

Yeseok Lee[1] and Yongil Kim*[2]
[1]Graduate Student, Dept. of Civil and Environmental Engineering, Seoul National University,
35-318, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea
Email: ysl4065@snu.ac.kr

[2]Professor, Dept. of Civil and Environmental Engineering, Seoul National University,
35-410, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea
Email: yik@snu.ac.kr

KEY WORDS: Super-resolution; Transformer; Generative adversarial network; Residual-aware loss

ABSTRACT: Nowadays the growing utilization of satellite imagery across various fields has resulted in an increased demand for high-resolution (HR) satellite imagery. Single Image Super Resolution (SISR), which is the method for the restoration of HR from low-resolution (LR) ones, has gained importance in remote sensing as a cost-effective method for acquiring HR satellite imagery. Recently, super-resolution using Generative Adversarial Networks (GANs) has shown promising results in generating more realistic images. Nevertheless, GAN-based methods are known to suffer from generation of visual artifacts which are particularly problematic in satellite imagery where small objects and high-frequency information are scattered diversely. To overcome this issue, we propose a Residual-aware Transformer-based Super-Resolution Generative Adversarial Network (RAT-SRGAN) for satellite imagery. We employed the Efficient Super-Resolution Transformer (ESRT) as the generator for proposed model. We made simple adjustments to the ESRT's architecture to enable the consideration of multiscale information during training and we replaced the ESRT's SELayer with Convolutional Block Attention Module (CBAM) to incorporate not only channel information but also spatial information into the model. Moreover, we adopted a novel loss function called residual-aware loss to address the challenges of visual artifact generation. The experimental results demonstrated that the modifications applied to ESRT were effective, and the proposed model outperformed existing state-of-the-art method, particularly in terms of quantitative evaluation metrics such as the Structural Similarity Index (SSIM) and the Learned Perceptual Image Patch Similarity (LPIPS). It was confirmed through additional experiments that the residual-aware loss can be generally applied to enhance the performance of GAN-based SISR methods.

1. INTRODUCTION

With recent advancements in remote sensing technology, the utilization of High-Resolution (HR) satellite imagery is expanding across various applications, including building extraction, object detection, change detection, land cover classification, disaster management (Chen et al., 2021; Dong et al., 2019; Peng et al., 2020; Tong et al., 2020; Xing et al., 2023), and so on. However, acquiring imagery of a desired area at a specific time can be challenging due to technical constraints inherent in satellite imagery acquisition methods, such as limitations in satellite orbit, cloud cover, and sensor capabilities. Single Image Super-Resolution (SISR) is a method for restoring HR images from Low-Resolution (LR) ones, which is a cost-effective and efficient method for addressing challenges related to image acquisition, and there is a growing recognition of its importance in high-resolution satellite imagery acquisition.

Traditionally, SISR techniques have relied on example-based methods (Bevilacqua et al., 2012; Cui et al., 2014; Dai et al., 2015; Glasner et al., 2009) and sparse-coding-based methods (Lu et al., 2012; Wang et al., 2010). In example-based methods, statistical techniques are used to learn the relationship between LR images and HR images for similar examples, and the learned examples are used to super-resolution. On the other hand, in sparse-coding-based methods, LR image patches are encoded into a series of basis vectors related to HR images, establishing a foundational structure for reconstructing HR images through linear combinations of their counterparts, thereby enabling super-resolution. Recently, with the rapid developments in deep learning technology, deep learning approaches have found widespread applications in various computer vision tasks, and in recent years, many researchers have been investigating deep learning-based SISR. Various deep learning methods using Convolutional Neural Networks (CNNs) have been proposed and have demonstrated improved performance compared to conventional methods (Dong et al., 2015; Dong et al., 2016; Kim et al., 2016; Lim et al., 2017; Shi et al., 2016). However, CNN-based SISR methods have typically employed Mean Squared Error (MSE) loss or Cross-Entropy loss as their loss function, resulting in the issue of not capturing texture and high-frequency details when restoring HR images, leading to blurry outcomes.

In order to address these issues, subsequent studies have proposed Generative Adversarial Networks (GANs) based SISR methods to generate more realistic images. (Dong et al., 2021; Ledig et al., 2017; Wang et al., 2018; Zhang et al., 2019). GANs consist of two components which are called generator and discriminator. The generator creates images from the latent space of input images, while the discriminator compares the generated images with real ones to determine their authenticity. They engage in competitive training until reaching a Nash equilibrium, making it increasingly challenging to distinguish real images from fake images. Moreover, GAN-based methods have achieved greater realism by incorporating perceptual loss. Nevertheless, GAN-based methods are known to suffer from unstable training and the generation of visual artifacts which are undesirable distortions that appear in generated images. These limitations are particularly problematic in satellite imagery where small objects and high-frequency information are scattered diversely.

Recently, the Transformer architecture has achieved significant success in the field of Natural Language Processing (NLP). As a result, there is a growing interest in utilizing it for visual tasks, a trend known as the Vision Transformer (ViT; Dosovitskiy et al., 2020). Moreover, various attempts have been made to apply this approach to SISR (Conde et al., 2022; Liang et al., 2021). Unlike CNNs, which inherently incorporate the inductive bias of locality and translation invariance, Transformers integrate all information using positional embedding and self-attention mechanisms. This can potentially result in the neglect of local information. Consequently, achieving high performance with Transformers often requires large-scale datasets for generalization, extensive computational resources, and huge GPU memory. To address these problems, hybrid methods that combine CNN and Transformer have also been proposed (Fang et al., 2022; Lu et al., 2022). However, these hybrid approaches sometimes produce results that are smooth and lack high-frequency details.

In this study, we proposed the Residual-Aware Transformer-Based Super-Resolution Generative Adversarial Network (RAT-SRGAN), which incorporates a hybrid approach into the GAN architecture. We adopted a hybrid approach due to the characteristics of satellite imagery, which covers extensive regions and contains a multitude of objects, requiring the consideration of both global and local features when utilizing satellite imagery. The generator within RAT-SRGAN is designed with a hybrid structure allowing for the simultaneous consideration of locality and globality. Furthermore, RAT-SRGAN leverages the GAN architecture to preserve high-frequency information and generate more realistic Super-Resolved (SR) images. To address the visual artifact generation issue associated with GANs, we have introduced the residual-aware loss during the training process. Through these efforts, we aimed to develop a more robust SISR method for satellite imagery. The main contributions of this study are as follows:

1. We introduced a novel loss function for mitigating visual artifact generation in GANs, and our experiments have demonstrated its applicability in enhancing the performance of other GANs.
2. We proposed a SISR model for satellite imagery, incorporating a hybrid model with GAN, and we have confirmed its competitiveness when compared to existing state-of-the-art methods.

2. RELATED WORK

2.1 GAN-based SISR methods

GAN was proposed by Goodfellow et al. (2014) and has become a state-of-the-art method in many vision tasks since its appearance. The Super-Resolution Generative Adversarial Network (SRGAN; Ledig et al., 2017) utilizes GAN for SISR and introduces perceptual loss using high-level feature maps from the VGG network (Simonyan & Zisserman., 2014) to overcome the limitations of existing methods, which lacked texture detail. The Enhanced Super-Resolution Generative Adversarial Network (ESRGAN; Wang et al., 2018) improves upon SRGAN's performance by using a residual-in-residual dense block instead of SRGAN's residual blocks. It also aims to generate more realistic images by employing relativistic GAN, achieving results that appear even more true to real than the originals. The Reference-based Remote Sensing Generative Adversarial Network (RRSGAN; Dong et al., 2021) restores LR images' details by leveraging the rich texture information from HR reference images, compensating for the lack of information in LR images. Additionally, it uses a relevance attention module and demonstrates robust and outstanding performance. The Second order Attention Generator Adversarial Network (SAGAN; Zhang et al., 2019) generates images by maximizing the utilization of prior information in LR images using second-order channel attention mechanisms and region-level non-local modules. To suppress visual artifact generation, it introduces region-aware loss, resulting in visually superior outcomes compared to existing methods.

2.2 ViT-based SISR methods

Since the emergence of Transformers, they have attracted significant attention and demonstrated outstanding performance in various fields. Particularly in the field of computer vision, there have been ongoing efforts to apply Transformers, and these efforts have seen some success. ViT was proposed by Dosovitskiy et al. (2020), demonstrating that pure Transformers can replace CNNs by inputting sequences of image patches into a Transformer model. Subsequently, not only ViT but also research on hybrid models combining ViT and CNN has been actively conducted. SwinIR (Liang et al., 2021) leverages the Swin Transformer to prove the practicality of pure Transformer architectures in SISR, a field where most state-of-the-art methods are CNN-based. Swin2SR (Conde et al., 2022), on the other hand, utilizes Swin Transformer V2 to address various issues that Transformer models faced, such as training instability, resolution gaps between pre-training and fine-tuning, and data hunger. Most related to ours is the model of Lu et al. (2022), which called Efficient Super-Resolution Transformer (ESRT). ESRT is a hybrid model that combines ViT and CNN. As shown in Figure 1, it consists of a Lightweight CNN Backbone (LCB) for extracting deep features using a CNN-based High Preserving Block (HPB) and a Lightweight Transformer Backbone (LTB) that utilizes Efficient Transformers (ET), incorporating Efficient Multi-Head Attention (EMHA) to make the model lighter. This structure allows it to deliver excellent performance with a reduced computational cost. As we will discuss later, we employ the basic structure of ESRT with some modifications as the generator of our model.

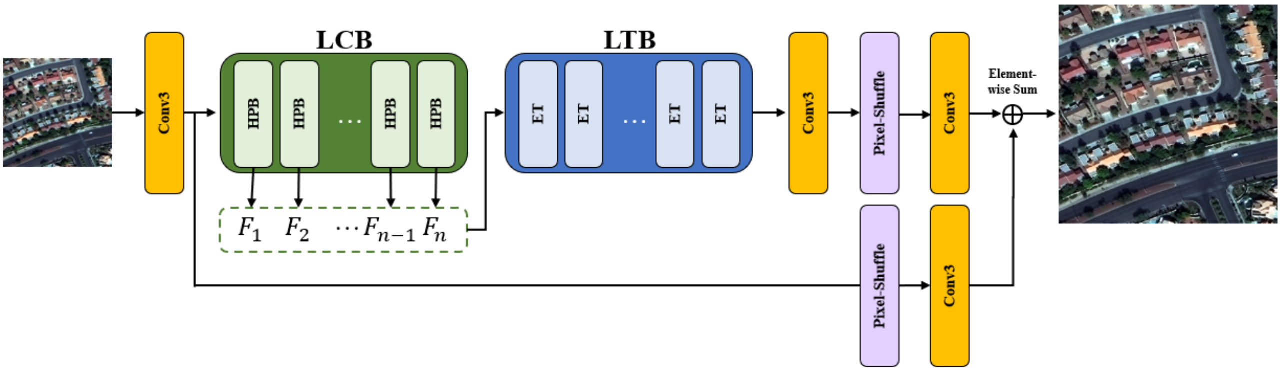


Figure 1. The architecture of the ESRT (Lu et al., 2022). LCB, HPB, LTB and ET stand for Lightweight CNN Backbone, High Preserving Block, Lightweight Transformer Backbone and Efficient Transformers respectively.

3. METHODOLOGY

Our primary objective is to propose a SISR method that can be applied to satellite imagery, which is distinct from widely used images as it captures extensive regions while including small objects. To achieve this, we considered a hybrid approach combining Transformer and CNN, allowing us to simultaneously capture global and local information. Additionally, to overcome challenges such as blurry results and the loss of high-frequency details, common in hybrid approaches, we adopted GAN architecture. However, the problem of visual artifact generation inherent to GANs is particularly critical in satellite imagery given the prevalence of small objects. To address this, we propose a residual-aware loss based on the residual between HR images and SR images.

As shown in Figure 2, the proposed RAT-SRGAN follows the simplest GAN structure consisting of a generator and a discriminator. Preprocessed LR images pass through the generator to produce SR images, during which the residual-aware loss calculated from SR and HR images is employed as a part of the generator loss. The generator and discriminator proceed with adversarial training: the generator is trained to produce SR images in a way that the discriminator cannot distinguish them from HR images, while the discriminator is trained to differentiate between SR images and HR images.

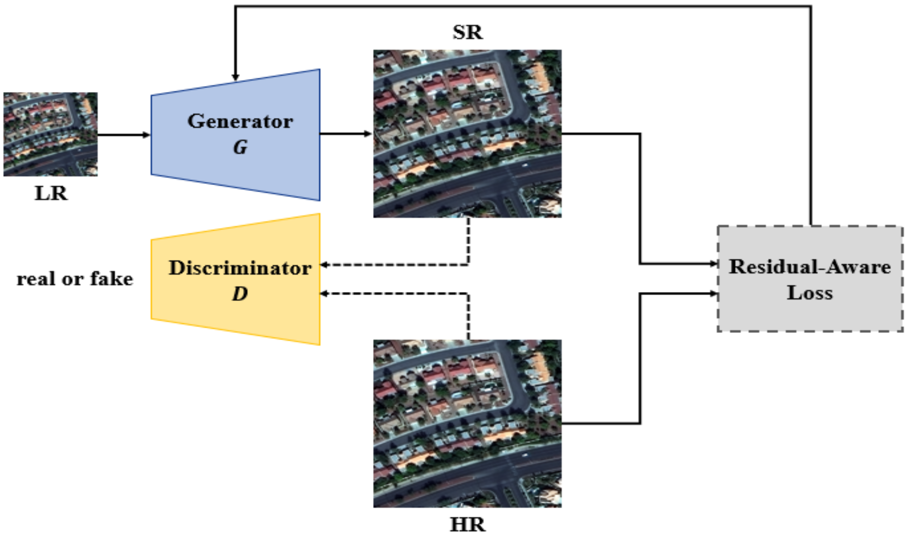


Figure 2. The overall framework of the proposed RAT-SRGAN.

3.1 Preprocessing

The purpose of SISR is to restore HR images from their corresponding LR images. While it would be ideal to train the model to increase the resolution of the available original images, obtaining higher-resolution images than the originals is difficult. Therefore, in most cases, the original images are assumed to be HR images and are often degraded to LR images. In this study, we employ Equation 1 to generate LR images.

$$I_{LR} = \downarrow_s (I_{HR} \otimes k) + n \quad \textbf{(1)}$$

where I_{HR} and I_{LR} represent the LR and HR images, $\downarrow_s$ indicates bilinear down-sampling with factor of s and $\otimes, k, n$ denote convolution operation, blurring kernel and gaussian noise respectively. A scale factor of 4 was employed, with a blurring kernel size of 5x5, and gaussian noise with a mean of 0 and a standard deviation of 0.02 was used.

3.2 Network Architecture

3.2.1 Generator: The generator of RAT-SRGAN adopts ESRT and introduces two simple modifications to ESRT's HPB. As shown in Figure 3(a), HPB consists of an Adaptive Residual Feature Block (ARFB) that addresses the gradient vanishing problem using a residual structure while reducing computational complexity through the reduction and expansion module. It also includes a High-Frequency Module (HFM) to preserve high-frequency information. We trained the model with multiscale features to better capture spatial diversity, thus enhancing the model's robustness. Furthermore, we replaced the SELayer in HPB with the Convolutional Block Attention Module (CBAM; Woo et al., 2018) to consider not only channel information but also spatial information. These modifications are depicted in Figure 3(b), and their effect is discussed in an ablation study.

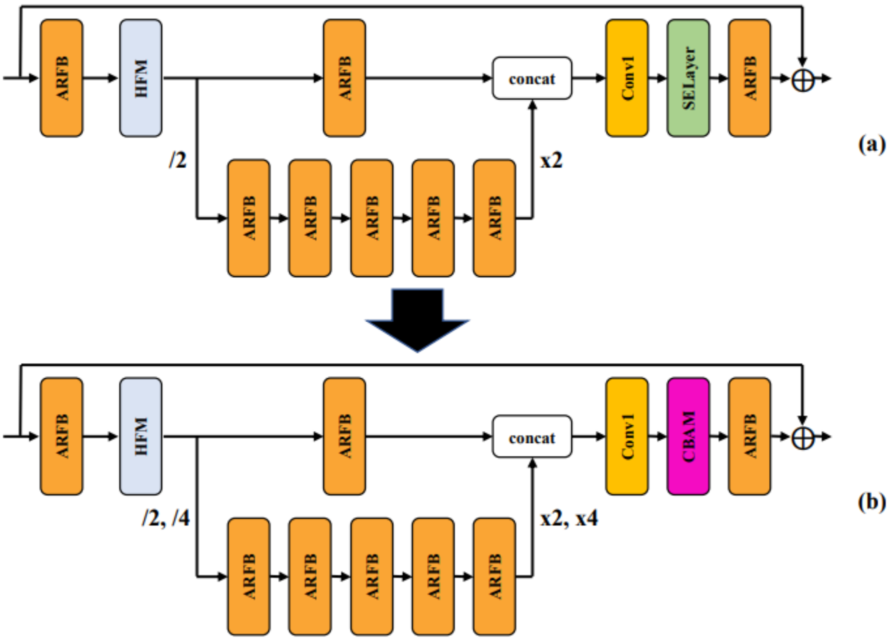


Figure 3. The architectures of HPB. (a) The architecture of ESRT's HPB and (b) The architecture of HPB used in RAT-SRGAN's generator with two modifications applied to (a).

3.2.2 Discriminator: The structure of SRGAN's discriminator (Ledig et al., 2017) was employed for our discriminator (Figure 4). The output of the generator generates 64 features through the convolution layer and passes through seven blocks consisting of convolution layers, batch normalization, and Leaky-ReLU activation function, increasing to 512 features. Subsequently, the output features pass through an average pooling layer instead of a dense layer for computational efficiency, and then go through a sigmoid activation function to indicate whether it's real or fake. Discriminator's loss function is based on the idea proposed in Goodfellow et al. (2014) with the omission of the logarithm for a more intuitive implementation. The loss function of the discriminator is formulated as follows.

$$L_D = 1 - \mathbb{E}_{I_{HR} \sim p_{train(I_{HR})}}[\mathrm{D}(I_{HR})] + \mathbb{E}_{I_{LR} \sim p_{G(I_{LR})}}[\mathrm{D}(\mathrm{G}(I_{LR}))] \quad \textbf{(2)}$$

where 𝔼 represents the averaging operator. To decrease the loss function, the discriminator is trained to output 1 when it receives HR images as input and 0 when it receives SR images as input.

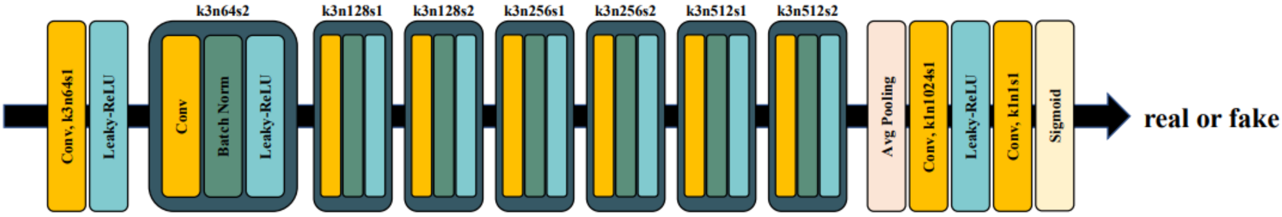


Figure 4. The architecture of discriminator.

3.3 Residual-Aware Loss

We have developed a residual-aware loss to address the visual artifact generation problem widely known in GANs. The fundamental assumption behind the residual-aware loss is that visual artifacts do not appear in all regions of the generated images but instead occur as outliers in regions where training is less stable. We hypothesized that these regions correlate with areas where residuals, derived from features of various scales, exhibit large values and by focusing more on learning in these specific regions, we could suppress visual artifacts. If we define f_s as the shallow feature extractor to obtain feature maps from HR and SR images, and f_e as the feature extractor to obtain feature maps from various scales, then the R^n, which is residual between the n-th features obtained from HR and SR images, is computed as follows.

$$\forall \mathrm{n} \in \mathbb{N}, \begin{cases} R^n = f_s(I_{HR}) - f_s(I_{SR}), where\ n = 1 \\ R^n = f_e^{n-1}\big(f_s(I_{HR})\big) - f_e^{n-1}\big(f_s(I_{SR})\big), otherwise \end{cases} \tag{3}$$

Subsequently, the residuals at multiple scales go through pixel shuffle and feature selection, resulting in vectors of the same size and dimensionality. These individual vectors are concatenated to the Residual Vector ℝ. This process can be represented by Equation 4 and 5.

$$\delta^n = \mathrm{LeakyReLu}(Conv_{k3n8s1}(PixelShuffle(R^n, 2^{n-1}))) \tag{4}$$

$$\mathbb{R} = [\delta^1, \delta^2, \cdots, \delta^N] \tag{5}$$

As seen in Equation 6, $\mathrm{\rho(r,c)}$ is computed, which is defined as the residual magnitude at a specific row r and column c derived from the generated ℝ. A higher value of $\mathrm{\rho(r,c)}$ indicates inaccurate predictions during the training process, and to suppress the generation of visual artifacts appearing as outliers, we set a threshold to the top 5% of values across all rows and columns. Values of ρ exceeding this threshold were preserved as they are, while the rest were set to 0.

$$\mathrm{\rho(r,c)} = \begin{cases} \sqrt{\sum_{d=1}^{D} \mathbb{R}_d(r,c)^2}\,, if\ \sqrt{\sum_{d=1}^{D} \mathbb{R}_d(r,c)^2} > the\ top\ 5th\ percentile\ of\ \rho \\ 0, otherwise \end{cases} \tag{6}$$

where $\mathbb{R}_d$ denotes the d-th dimension of residual vector and D equals $\dim(\delta^1) + \dim(\delta^2) + \cdots + \dim(\delta^N)$. We adjusted the training process to emphasize areas with a higher likelihood of visual artifact generation by using inner product between residual magnitude and the residuals of HR and SR images. To ensure more stable application of this loss during training, we normalized it to have values between 0 and 1. The proposed residual-aware loss is as shown in Equation 7, and the overall structure of the residual-aware loss can be observed in Figure 5. The comprehensive generator loss used in this study is given by Equation 8.

$$\mathrm{L}_{RA} = 1 - e^{-(\rho \bullet |I_{HR} - I_{SR}|)} \tag{7}$$

$$\mathrm{L}_{G} = \mathrm{L}_{MSE} + \lambda \mathrm{L}_{Percep} + \xi \mathrm{L}_{adv} + \eta \mathrm{L}_{RA} \tag{8}$$

where L_{MSE} represents the MSE loss between HR and SR, L_{Percep} is the perceptual loss, which is the L1 norm between HR and SR features passed through the 35th layer of VGG 19 and L_{adv} is the adversarial loss, which is defined as $1 - \mathbb{E}_{I_{LR} \sim p_{G(I_{LR})}}[\mathrm{D}(\mathrm{G}(I_{LR}))]$. λ, ξ, η are coefficients and we have set their values to 0.006, 0.001, and 0.1, respectively.

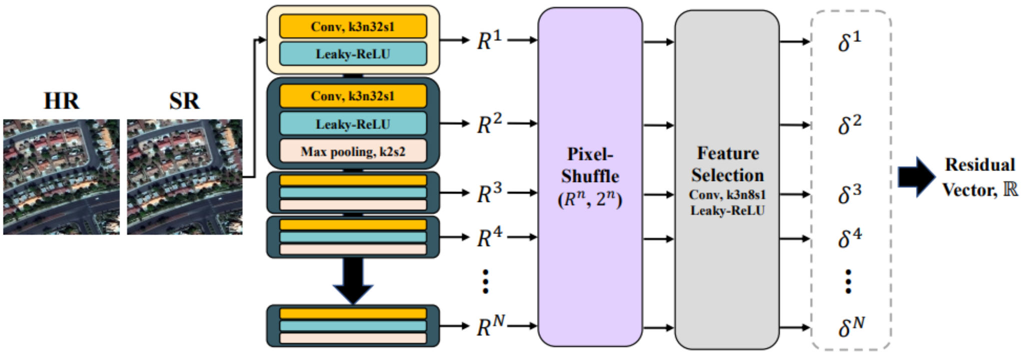


Figure 5. The overall architecture of residual-aware loss.

4. EXPERIMENTS

4.1 Training Details

The SpaceNet 2 Dataset (Van Etten et al., 2018) provides satellite images of various regions, captured by the WorldView-3 with a spatial resolution of 0.3 meters, which have been pan-sharpened. For our experiment, we specifically used images from the Las Vegas area. We utilized 120 images of size 650 x 650 as training data and 40 images as test data. During the training process, we patched the images to a size of 384 x 384. For each epoch, we randomly cropped 256 x 256 regions within this range, which served as the training input. Batch size was set to 8, and a learning rate of 1e-4 was used. The optimizer employed was Adam with β_1= 0.9, β_2= 0.999 and We trained both the generator and discriminator alternately until 200 epochs were completed. At the end of each epoch we calculated the Peak Signal-to-Noise Ratio (PSNR) of the training data for that epoch. The weight values from the epoch with the highest PSNR were saved and used for testing. All experiments were conducted using PyTorch 1.9.1 on a system with 16-GB RAM and one NVIDIA GeForce RTX 2070 GPU.

4.2 Result and Discussion

In Table 1, we quantitatively compared the results of RAT-SRGAN with existing state-of-the-art SISR methods. As shown in Table 1, GAN-based methods generally exhibit worse outcomes compared to CNN-based methods, which is suspected to be influenced by the loss function during training. Indeed, metrics like PSNR or SSIM, which are calculated based on MSE, show that CNN-based models perform significantly better, while considering perceptual aspects such as FID, GAN-based models demonstrate superior results. However, it is difficult to explain the results of LPIPS or NIQE based on the influence of the loss function only. These outcomes are likely due to the instability introduced by GAN-based models during training on satellite imagery, which contains a rich information compared to general images. The proposed RAT-SRGAN, despite utilizing a GAN, demonstrated stable results across most metrics, particularly in SSIM and LPIPS metrics, it shows best results. This indicates that the proposed model is competitive when compared to existing models and aligns with the goal of this study, which aimed to develop more visually realistic and stable super-resolution model for satellite imagery.

We have also visually provided the results for comparative experiments in Figure 6 and Figure 7. Overall, it appears that GAN-based methods preserve high-frequency information better than CNN-based methods. However, it is observed that GAN-based methods exhibit a distinctive grid-like pattern. Figure 7 is an enlarged version of Figure 6, and it confirms that RAT-SRGAN performs relatively well in restoring the central vehicle in the image compared to other models. Particularly, it effectively suppresses the blurriness along building boundaries, as opposed to SRGAN and ESRGAN, providing stable image restoration. These results, as seen in the earlier quantitative evaluations, demonstrate the impressive performance of the proposed model and are consistent with the objectives of this study.

Table 1. Quantitative results compared to state-of-the-art methods. ↑ indicates better results with higher values, while ↓ indicates better results with lower values, and the best performing method for each metric is highlighted in bold.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	NIQE↓
Bicubic	29.6361	0.7306	0.4635	11.7875	7.963
SRCNN	29.4045	0.7574	0.4900	12.8078	6.326
VDSR	**29.7072**	0.7386	0.4617	11.7663	8.004
EDSR	28.8229	0.7746	0.4423	9.4471	6.351
RCAN	29.2572	0.7817	0.4402	10.1487	**6.322**
SRGAN	28.5933	0.7182	0.5061	8.0229	8.309
ESRGAN	28.9392	0.7260	0.4505	**3.7232**	6.498
RAT-SRGAN	28.4565	**0.7660**	**0.4198**	4.4491	7.900

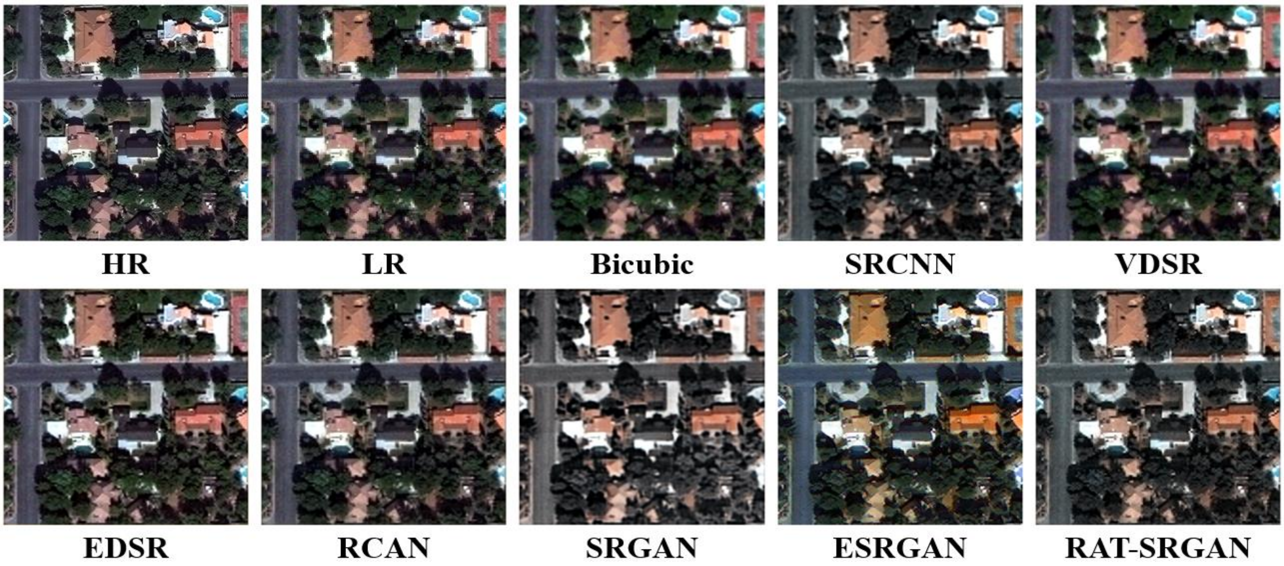


Figure 6. Qualitative comparison with state-of-the-art methods on one of test images.

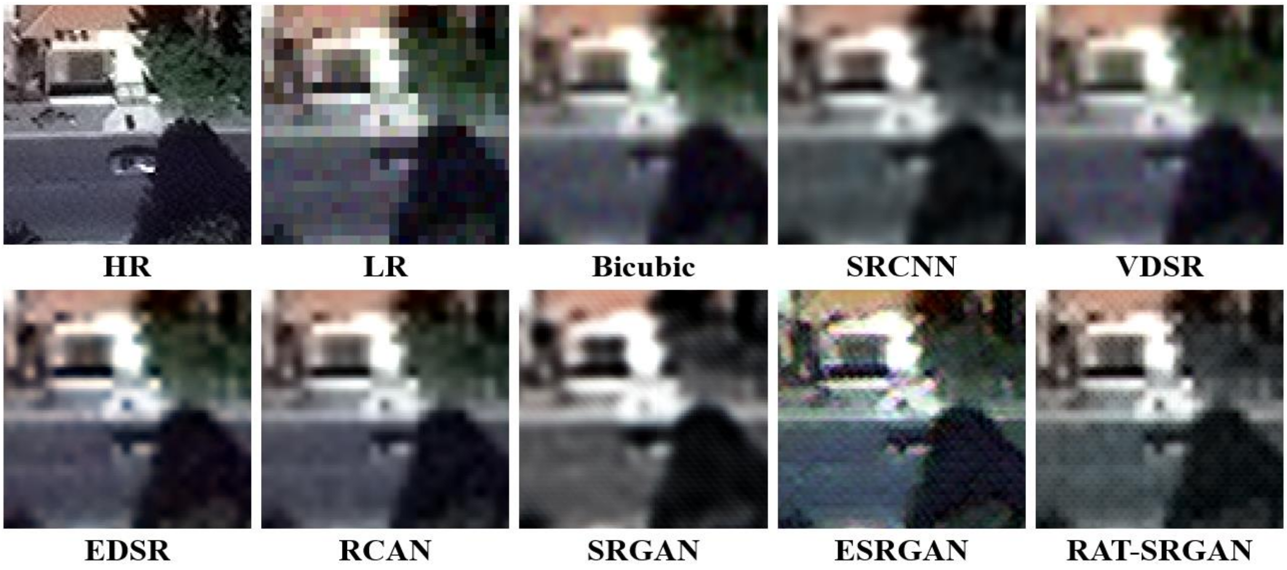


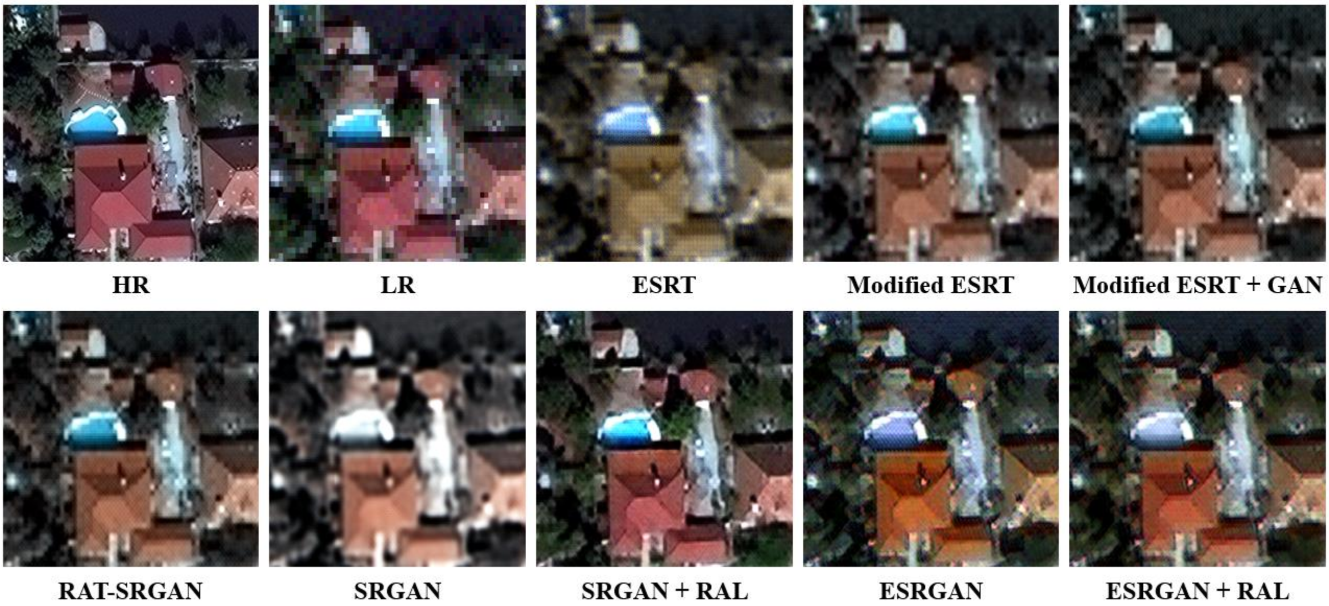
Figure 7. Qualitative comparison of the enlarged image in Figure 6.

4.3 Ablation Study

Additional researches were conducted to demonstrate the effectiveness of modifications applied to ESRT and the impact of the proposed residual-aware loss. As shown in Table 2, overall quantitative performance improved when modifications were applied to ESRT. Furthermore, when both GAN and modifications were simultaneously applied to ESRT, four out of the five metrics showed better results than when ESRT was applied with modifications alone. Additionally, it was observed that the addition of the residual-aware loss resulted in performance improvements across all metrics compared to when it was not applied. Additionally, to assess the general applicability of the residual-aware loss, we conducted comparative experiments by adding the residual-aware loss to both SRGAN and ESRGAN. It was observed that overall metrics showed improvements. Notably, when the residual loss was applied to SRGAN, there was a remarkable performance enhancement. In comparison to all state-of-the-art models tested in this paper, SRGAN with the residual loss outperformed others in all quantitative metrics. The qualitative results for the ablation study are presented in Figure 8. As observed in the quantitative evaluations, introducing the residual-aware loss to SRGAN leads to a significant visual improvement. These results confirmed the effectiveness of the modifications applied to ESRT and the residual-aware loss. Specifically, the residual-aware loss demonstrated its applicability, not only to the proposed model but also to GAN-based SISR methods.

Table 2. Quantitative results for ablation study. RAL means residual-aware loss and the best-performing method in each evaluation metric, for each backbone method, is highlighted in bold.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	NIQE↓
ESRT	27.7397	0.6851	0.5616	9.5864	8.526
ESRT + Modification	28.0278	0.7547	0.4702	9.0549	**6.848**
ESRT + Modification + GAN	28.3851	0.7662	0.4222	4.8234	11.110
ESRT + Modification + GAN + RAL (RAT-SRGAN)	**28.4128**	**0.7680**	**0.4198**	**4.4491**	7.900
SRGAN	28.5933	0.7182	0.5061	8.0229	8.309
SRGAN + RAL	**29.8880**	**0.7822**	**0.3605**	**2.3154**	**5.099**
ESRGAN	28.9392	0.7260	0.4505	**3.7232**	**6.498**
ESRGAN + RAL	**29.0689**	**0.7401**	**0.4467**	3.8156	6.762



HR LR ESRT Modified ESRT Modified ESRT + GAN

RAT-SRGAN SRGAN SRGAN + RAL ESRGAN ESRGAN + RAL

Figure 8. Qualitative results for the ablation study. It can be observed that the result of applying RAL to SRGAN is closest to HR image.

5. CONCLUSION

In this work, we propose the Residual-Aware Transformer-based Super-Resolution Generative Adversarial Network (RAT-SRGAN) for satellite imagery super-resolution. RATSRGAN leverages the Efficient Super-Resolution Transformer (ESRT) as its generator and incorporates two simple modifications. Additionally, we propose a novel residual-aware loss function. Experimental results demonstrate the quantitative and qualitative competitiveness of proposed model when compared to state-of-the-art methods. Furthermore, we validate the effectiveness of modifications through supplementary experiments. Notably, the residual-aware loss not only enhances the performance of proposed model but also improves the performance of other GAN-based models. In the future work, it is necessary to conduct experiments for super-resolution at various scales to evaluate the applicability of RAT-SRGAN and investigate its compatibility with diverse satellite sensor datasets. Furthermore, it is required to develop a more suitable algorithm for setting the threshold of the residual-aware loss.

ACKNOWLEDGMENTS

This research was financially supported by the Institute for Peace and Unification Studies (IPUS) at Seoul National University under the project of “Laying the Groundwork for Unification and Peace.”

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (NRF-2023R1A2C2005548)

REFERENCES

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

Chen, M., Wu, J., Liu, L., Zhao, W., Tian, F., Shen, Q., ... & Du, R., 2021. DR-Net: An improved network for building extraction from high resolution remote sensing image. *Remote Sensing*, *13*(2), 294.

Conde, M. V., Choi, U. J., Burchi, M., & Timofte, R., 2022. Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pp. 669-687.

Cui, Z., Chang, H., Shan, S., Zhong, B., & Chen, X., 2014. Deep network cascade for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Proceedings, Part V 13*, pp. 49-64.

Dai, D., Timofte, R., & Van Gool, L., 2015. Jointly optimized regressors for image super-resolution. In *Computer Graphics Forum*, *34*(2), pp. 95-104.

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), pp. 295-307.

Dong, C., Loy, C. C., & Tang, X., 2016. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Proceedings, Part II 14*, pp. 391-407.

Dong, R., Zhang, L., & Fu, H., 2021. RRSGAN: Reference-based super-resolution for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, pp. 1-17.

Dong, Z., Wang, M., Wang, Y., Zhu, Y., & Zhang, Z., 2019. Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(3), pp. 2104-2114.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fang, J., Lin, H., Chen, X., & Zeng, K., 2022. A hybrid network of cnn and transformer for lightweight image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1103-1112.

Glasner, D., Bagon, S., & Irani, M., 2009. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pp. 349-356.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y., 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672-2680.

Kim, J., Lee, J. K., & Lee, K. M., 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646-1654.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681-4690.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R., 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833-1844.

Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K., 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136-144.

Lu, X., Yuan, H., Yan, P., Yuan, Y., & Li, X., 2012. Geometry constrained sparse coding for single image super-resolution. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1648-1655.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., & Zeng, T., 2022. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 457-466.

Peng, D., Bruzzone, L., Zhang, Y., Guan, H., Ding, H., & Huang, X., 2020. SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(7), pp. 5891-5906.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... & Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874-1883.

Simonyan, K., & Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tong, X. Y., Xia, G. S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, *237*, 111322.

Van Etten, A., Lindenbaum, D., & Bacastow, T.M., 2018. SpaceNet: A Remote Sensing Dataset and Challenge Series. ArXiv, abs/1807.01232.

Wang, J., Zhu, S., & Gong, Y., 2010. Resolution enhancement based on learning the sparse association of image patches. *Pattern Recognition Letters*, *31*(1), pp. 1-10.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., … & Change Loy, C., 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision workshops,* pp. 701-710.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S., 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision,* pp. 3-19.

Xing, Z., Yang, S., Zan, X., Dong, X., Yao, Y., Liu, Z., & Zhang, X., 2023. Flood vulnerability assessment of urban buildings based on integrating high-resolution remote sensing and street view images. *Sustainable Cities and Society*, *92*, 104467.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A., 2019. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354-7363.