

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**SINGLE-FRAME AND MULTI-FRAME SUPER-RESOLUTION
ON REMOTE SENSING IMAGES VIA DEEP LEARNING APPROACHES**



Ph.D. THESIS

Peijuan WANG

Department of Communication Systems

Satellite Communication and Remote Sensing Programme

JULY 2022

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**SINGLE-FRAME AND MULTI-FRAME SUPER-RESOLUTION
ON REMOTE SENSING IMAGES VIA DEEP LEARNING APPROACHES**



Ph.D. THESIS

**Peijuan WANG
(705172003)**

Department of Communication Systems

Satellite Communication and Remote Sensing Programme

Thesis Advisor: Prof. Dr. Elif SERTEL

JULY 2022

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**DERİN ÖĞRENME YAKLAŞIMLARIYLA UZAKTAN ALGILAMA
GÖRÜNTÜLERİNDE TEK ÇERÇEVE VE ÇOK ÇERÇEVE SÜPER
ÇÖZÜNÜRLÜK**

DOKTORA TEZİ

**Peijuan WANG
(705172003)**

İletişim Sistemleri Anabilim Dalı

Uydu Haberleşmesi ve Uzaktan Algılama Doktora Programı

Tez Danışmanı: Prof. Dr. Elif SERTEL

TEMMUZ 2022

Peijuan WANG, a Ph.D. student of ITU Graduate School student ID 705172003, successfully defended the thesis/dissertation entitled “SINGLE-FRAME AND MULTI-FRAME SUPER-RESOLUTION ON REMOTE SENSING IMAGES VIA DEEP LEARNING APPROACHES”, which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Prof. Dr. Elif SERTEL**
Istanbul Technical University

Jury Members : **Prof. Dr. Çiğdem GÖKSEL**
Istanbul Technical University

Doç. Dr. Uğur ALGANCI
Istanbul Technical University

Prof.Dr. Bülent BAYRAM
Yildiz Technical University

Dr.Öğr.Üye. Tolga BAKIRMAN
Yildiz Technical University

Date of Submission : 22 June 2022
Date of Defense : 29 July 2022





To my spouse and children,



FOREWORD

Firstly, I would like to express my great thanks to my advisor Prof. Dr. Elif SERTEL for her guidance during my Ph.D. study and for providing me with scholarships to support my education expenses over the last five years. Also, after the pandemic, she provided me with a computer for my research at home, so that I can continue my research smoothly and finish my study on time. I would like to thank my advisor again for her kind support when it's most needed. Besides, I would like to thank Prof. Dr. Bulent BAYRAM and Assoc. Prof. Ugur ALGANCI for their valuable advice on my research and my thesis.

Moreover, I would like to thank all the professors and researchers working at ITU Research and Application Center for Satellite Communications and Remote Sensing (CSCRS) for their significant guidance and kind help during my study. I would like to thank CSCRS for providing the Pleiades satellite images freely for my research.

Finally, I would like to express my sincere thanks to my family, especially my husband and my son for their encouragement, support, and company during my Ph.D. study. They are a great motivation for my study and hard work.

July 2022

Peijuan WANG
(Research Assistant)



TABLE OF CONTENTS

	<u>Page</u>
FOREWORD	ix
TABLE OF CONTENTS	xi
ABBREVIATIONS	xiii
SYMBOLS	xv
LIST OF TABLES	xvii
LIST OF FIGURES	xix
SUMMARY	xxi
ÖZET	xxiii
1. INTRODUCTION	1
1.1 Motivation	1
1.2 Definition of Problem	2
1.3 Contributions Of The Dissertation	3
2. LITERATURE REVIEW	5
2.1 Review of the Single-Frame Remote Sensing Image Super-resolution	5
2.1.1 Deep learning-based SISR	5
2.1.1.1 Deep CNN-based SR methods	5
2.1.1.2 Attention-based methods	6
2.1.1.3 Cascading networks	8
2.1.1.4 GAN-based methods	8
2.1.2 Supervised methods for SISR on remote sensing images	9
2.1.2.1 CNN-based methods	9
2.1.2.2 GAN-based methods	14
2.1.2.3 Attention-based methods	16
2.1.2.4 Back-projection based methods	18
2.1.3 Unsupervised methods for SISR on remote sensing images	19
2.2 Review of the Multi-Frame Remote Sensing Image Super-Resolution	21
2.2.1 Deep-learning based video super-resolution	21
2.2.2 Multi-frame RSISR	22
2.3 Attention Mechanisms	24
2.3.1 Self-attention	25
2.3.2 Channel attention, spatial attention, and pixel attention	26
2.3.3 Non-local attention	27
2.3.4 Mixed high order attention	28
2.4 Loss functions	29
2.4.1 Pixel-wise Loss	29
2.4.2 Perceptual Loss	30
2.4.3 Adversarial Loss	31
2.4.4 Cycle Consistency Loss	32
3. DATASETS AND METHODOLOGY	33
3.1 Datasets	33
3.1.1 Dataset for single RSISR	33
3.1.2 Dataset for MFSR remote sensing images	34

3.2 Attention CNN-based SFSR On Remote Sensing Images	34
3.2.1 Architecture of the Attention CNN-based SFSR on RS Images	34
3.2.1.1 The overall architecture	34
3.2.1.2 Layer Attention Module and Spatial Attention module	37
3.2.1.3 Random downsampling	38
3.2.2 Model Training	38
3.3 Attention GAN-based SFSR On Remote Sensing Images	39
3.3.1 Architecture of the attention GAN-based SFSR on RS images	39
3.3.2 Model Training	42
3.4 Multi-frame Remote Sensing Image Super-resolution	42
3.4.1 Architecture of the MFSR on RS Images	42
3.4.2 Model Training	47
3.5 Evaluation Criteria	47
4. RESULTS AND DISCUSSIONS OF SINGLE-FRAME REMOTE SENSING IMAGE SUPER-RESOLUTION	51
4.1 Results of the Attention-based SFSR on Remote Sensing Images	51
4.1.1 Experiment settings	51
4.1.2 Results on downsampled images	52
4.1.2.1 Quantitative and qualitative results of DFC (2019) Dataset	52
4.1.2.2 Quantitative and qualitative results on Pleiades images	61
4.1.3 Results on original DFC images	68
4.1.4 Results on multispectral images	71
4.1.5 Discussion	75
4.2 Results of the Attention GAN-based SFSR on Remote Sensing Images	75
4.2.1 Experiment settings	75
4.2.2 Quantitative results	75
4.2.3 Qualitative results	77
4.2.4 Discussion	79
5. RESULTS AND DISCUSSIONS OF MULTI-FRAME REMOTE SENSING IMAGE SUPER-RESOLUTION	81
5.1 Experiment Settings	81
5.2 Results Of The MFSR On SpaceNet7 Images	81
5.2.1 Quantitative results	81
5.2.2 Qualitative results	82
5.3 Test results on Real-world Jilin-1 images	88
5.4 Discussions	89
6. CONCLUSIONS AND RECOMMENDATIONS	93
6.1 Conclusions	93
6.2 Recommendations	94
REFERENCES	97
CURRICULUM VITAE	113

ABBREVIATIONS

CNN	: Convolutional Neural Network
RS	: Remote Sensing
SR	: Super-resolution
HR	: High Resolution
LR	: Low Resolution
GAN	: Generative Adversarial Network
RaGAN	: Relativistic Generative Adversarial Network
RSISR	: Remote Sensing Image Super-Resolution
MS	: Multispectral
DL	: Deep Learning
RMSE	: Root Mean Square Error
SSIM	: Structural Similarity Index Measure
PSNR	: Peak Signal-to-Noise Ratio
AG	: Average Gradient
NIQE	: Natural Image Quality Evaluator
ERGAS	: Erreur Relative Globale Adimensionnelle de Synthese
LPIPS	: Learned Perceptual Image Patch Similarity
PI	: Perceptual Index



SYMBOLS

Ψ	: Degradation model
δ	: parameters of the downsampling operations
$\mathbf{I}_{LR}, \mathbf{I}_{HR}, \mathbf{I}_{SR}$	: LR, HR and SR image
$\mathbf{F}$	: Super-resolution model
θ	: Parameters of the Super-resolution model
$\mathbf{L}$	: Loss function
s	: Scale factor of the Super-resolution model
$\mathbf{k}$	: Blurring kernel
$\mathbf{n}$	: Noise
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	: Query, Keys, and Values
γ	: Scaling factor
$\mathbf{H}, \mathbf{W}, \mathbf{C}$	: Hight, Width, and Channels of the input
$\mathbf{X}$	: Input Feature Maps
$\mathbf{A}(\mathbf{X})$	: Attention Maps of $\mathbf{X}$
σ	: Non-linear Activation Function
$\mathbf{R}$	: Number of the Order
$\mathbf{G}$	: Generator
$\mathbf{D}$	: Discriminator
concat	: Concatenation Operation
f	: Convolutional Operation
β	: Trainable Parameter



LIST OF TABLES

	<u>Page</u>
Table 4.1 : Quantitative values of DFC dataset with a scaling factor of $\times 2$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	52
Table 4.2 : Quantitative values of DFC dataset with a scaling factor of $\times 4$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	53
Table 4.3 : Quantitative values of DFC dataset with a scaling factor of $\times 8$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	54
Table 4.4 : Quantitative values of Pleiades images with a scaling factor of $\times 2$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	61
Table 4.5 : Quantitative values of Pleiades images with a scaling factor of $\times 4$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	62
Table 4.6 : Quantitative values of Pleiades images with a scaling factor of $\times 8$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	62
Table 4.7 : Quantitative results of Attention GAN-based model on DFC dataset with a scaling factor of $\times 4$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	76
Table 5.1 : Quantitative results of SpaceNet7 images. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.....	82
Table 5.2 : The results of different combinations of Attention modules. Red color and blue color represent the metric value ranked as the 1st and the 2nd, respectively.	90
Table 5.3 : The effects of the numbers of the Residual Blocks. Red color represents the metric value ranked as the 1st.	91
Table 5.4 : The effectiveness of the Attention Discriminator. Red color represents the metric value ranked as the 1st.	91
Table 5.5 : The effects of the loss function. Red color represents the metric value ranked as the 1st.....	92



LIST OF FIGURES

	<u>Page</u>
Figure 2.1 : Self-attention architecture.	25
Figure 2.2 : Channel attention module.....	27
Figure 2.3 : Spatial attention module.....	27
Figure 2.4 : Pixel attention module.....	27
Figure 2.5 : Non-local attention.	27
Figure 2.6 : Mixed high order attention.	28
Figure 3.1 : The overall architecture of the proposed HANRS.	35
Figure 3.2 : The architecture of the multi-layer Attention module (LAM).	35
Figure 3.3 : The architecture of the spatial Attention module (SAM).	35
Figure 3.4 : The architecture of the generator in ESRGAN.....	39
Figure 3.5 : Basic block in ESRGAN.	39
Figure 3.6 : Dense block proposed in ESRGAN.	39
Figure 3.7 : Our proposed Dense block for RSISR.	40
Figure 3.8 : Proposed MHOA Attention GAN-Discriminator.....	42
Figure 3.9 : Overview of the generator of the proposed MFSR method.	43
Figure 3.10 : Proposed attention to the MFSR method.....	43
Figure 3.11 : Discriminator of the proposed MFSR.	47
Figure 4.1 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on railway..	55
Figure 4.2 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on parking lot.	55
Figure 4.3 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on road region.	56
Figure 4.4 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on roof region.	56
Figure 4.5 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on railway..	57
Figure 4.6 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on parking lot.	57
Figure 4.7 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on road region.	58
Figure 4.8 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on roof region.	58
Figure 4.9 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on railway..	59
Figure 4.10 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on parking lot.	59
Figure 4.11 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on road region.	60
Figure 4.12 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on the roof region.	60
Figure 4.13 : Visualizations of Pleiades images with a scaling factor of $\times 2$	63
Figure 4.14 : Visualizations of Pleiades images with a scaling factor of $\times 2$	64

Figure 4.15 : Visualizations of Pleiades images with a scaling factor of $\times 2$	64
Figure 4.16 : Visualizations of Pleiades images with a scaling factor of $\times 4$	65
Figure 4.17 : Visualizations of Pleiades images with a scaling factor of $\times 4$	65
Figure 4.18 : Visualizations of Pleiades images with a scaling factor of $\times 4$	66
Figure 4.19 : Visualizations of Pleiades images with a scaling factor of $\times 8$	67
Figure 4.20 : Visualizations of Pleiades images with a scaling factor of $\times 8$	67
Figure 4.21 : Visualizations of Pleiades images with a scaling factor of $\times 8$	68
Figure 4.22 : Visual results on original DFC images on railway area.	69
Figure 4.23 : Visual results on original DFC images on roof.	69
Figure 4.24 : Visual results on original DFC images on parking lot.	70
Figure 4.25 : Visual results on original DFC images on road region.	70
Figure 4.26 : Visual results on original DFC images on roof and buildings.	71
Figure 4.27 : Visual Results on Pleiades multispectral images.	72
Figure 4.28 : Visual Results on Pleiades multispectral images.	72
Figure 4.29 : Visual Results on Pleiades multispectral images.	73
Figure 4.30 : Visual Results on Worldview-3 multispectral images.	73
Figure 4.31 : Visual Results on Worldview-3 multispectral images.	74
Figure 4.32 : Visual Results on Worldview-3 multispectral images.	74
Figure 4.33 : Visual results on DFC images.	77
Figure 4.34 : Visual results on DFC images.	77
Figure 4.35 : Visual results on DFC images.	77
Figure 4.36 : Visual results on DFC images.	78
Figure 4.37 : Visual results on DFC images.	78
Figure 4.38 : Visual results on DFC images.	78
Figure 4.39 : Visual results on DFC images.	79
Figure 5.1 : Visual Results on SpaceNet7 images.	83
Figure 5.2 : Visual Results on SpaceNet7 images.	83
Figure 5.3 : Visual Results on SpaceNet7 images.	84
Figure 5.4 : Visual Results on SpaceNet7 images.	85
Figure 5.5 : Visual Results on SpaceNet7 images.	85
Figure 5.6 : Visual Results on SpaceNet7 images.	86
Figure 5.7 : Visual Results on SpaceNet7 images.	86
Figure 5.8 : Visual Results on SpaceNet7 images.	87
Figure 5.9 : Visual Results on SpaceNet7 images.	87
Figure 5.10 : Visualization of Jilin-1 images.	88
Figure 5.11 : Visualization of Jilin-1 images.	89
Figure 5.12 : Visualization of Jilin-1 images.	89

SINGLE-FRAME AND MULTI-FRAME SUPER-RESOLUTION ON REMOTE SENSING IMAGES VIA DEEP LEARNING APPROACHES

SUMMARY

As a quite significant computer vision task, image super-resolution (SR) has been widely applied in remote sensing (RS), medical imaging, video surveillance, and biometrics. Image SR aims to restore high-resolution (HR) images by enhancing the spatial, spectral, or temporal resolution of the low-resolution (LR) inputs. In recent years, great efforts have been made for improving the SR approaches. One of the approaches is to design deeper networks. Nevertheless, this greatly increases computation and memory consumption. As a result, some mechanisms (such as cascading networks, attention mechanisms, and back projection) are proposed to improve the performance and training process of the complex networks.

Satellite imagery can be seen in various fields, namely Land cover/Land use classification, road and building extraction, observation of climate, and earthquake prediction. However, in some cases, the resolution of satellite images can not meet the application requirements due to the technology and cost limitations during the satellite design; therefore, the improvement of image resolution might be necessary. Since it is not possible to upgrade the equipment onboard the launched satellite, the software-based SR algorithms are deserved to be explored in RS fields. This thesis aims to strengthen the spatial resolution of optical satellite imageries by using deep learning (DL) methods.

Generally, image SR algorithms can be categorized as single-frame image SR (SFSR) and multi-frame image SR (MFSR). The inputs of the SFSR can be a single LR image. While multi-frame image SR aims to restore HR image by using multiple LR images, which can be obtained under different conditions and different angles.

Recently, great contributions have been made to improve the SR methods including two aspects: (1) increasing the value of PSNR (Peak-Signal-Noise-Ratio); (2) improving the image quality perceptually. Nevertheless, some algorithms obtain a high PSNR but with a low perceptual quality which is more important to human perception. Therefore, this thesis has the following objectives: (1) Explore a perceptual-driven approach to enhance the SR image quality visually on single-frame and multi-frame RS imageries; (2) Explore Generative Adversarial Network (GAN)-based models for single-frame and multi-frame RS imagery SR task to fulfill the multi-scale problem and blind to the degradation model; (3) Explore an image fusion method that can generate an arbitrary size of the super-resolved image rather than a small patch.

This thesis firstly gives an overview of single-frame and multi-frame RS image SR methods. The single-frame RS image SR methods are briefly classified into supervised and unsupervised methods. The former mainly includes Convolutional Neural Networks (CNN)-based, GAN-based, attention-based, and Back-projection based methods. In addition, the commonly used attention mechanisms including self-

attention, channel attention, spatial attention, mixed high-order attention (MHOA), non-local attention (NLA), and non-local sparse attention (NLSA) are also introduced. Moreover, loss functions including pixel-wise loss, perceptual loss, adversarial loss, and cycle consistency loss are presented.

For the single-frame RSISR, firstly, an attention CNN-based SR method is proposed. Although CNN-based algorithms have made outstanding achievements in computer vision tasks, the traditional CNNs methods treat the abundant low-frequency information included in the LR inputs equally across channels. Attention-guided algorithms play a vital role in the informative features extraction in various tasks including image SR. With the application of the attention mechanism, the proposed CNN-based method can further learn the deeper relationships among the different channels. Instead of simply integrating the attention module with the residual blocks, a Layer Attention Module (LAM) and Spatial Attention Module (SAM) are proposed to further learn the relationships among the Residual Groups (RG). Moreover, the perceptual loss function is adopted in the training process to enhance the generated image quality perceptually, and Random down-sampling is applied to strengthen the model's generalization ability.

Secondly, an attention GAN-based SR method is explored for the single-frame RS images. CNN-based methods have made great contributions to increasing the value of PSNR/SSIM. Nevertheless, the generated outputs tend to be overly smooth and blurry. GAN can generate more realistic images than normal CNN-based methods and has been introduced to single image SR (SRGAN, ESRGAN, EEGAN). Standard GANs only function on spatially local points in LR feature maps. The attention mechanism can directly learn the long-range dependencies in the feature maps both in the generator and discriminator in a powerful way. By applying the attention mechanism, the network allocates attention based on the similarity of color and texture. Therefore, based on ESRGAN, an attention GAN-based method is for the single-frame RS image SR. The ESRGAN was mainly improved from two aspects: (1) we further improved the architecture of the residual blocks by adding more skip connections; (2) we add attention modules to the residual blocks for further feature extraction.

Moreover, instead of working on aerial photographs or low-resolution and medium-resolution satellite images, we are focusing on Very High-Resolution (VHR) satellite imageries, such as the Pleiades, and Worldview-3. The spatial resolutions of the multispectral images for the Pleiades, and Worldview-3 are 2m, and 1.24m, respectively. Furthermore, for the attention CNN-based method, we evaluated the method on the Pleiades and Worldview-3 datasets with scaling factors of $\times 2$, $\times 4$, and $\times 8$. For the attention GAN-based method, we evaluated the method on the Pleiades and Worldview-3 datasets with a scaling factor of $\times 4$. The experimental results show the attention-based method can provide better perceptual results both quantitatively and qualitatively.

At last, we proposed an attention GAN-based method for the multi-frame RS image SR. Firstly, we introduced an attention mechanism to the Generator and proposed a space-based network that worked on every single frame for better temporal information extraction. Secondly, we proposed a novel attention module for better spatial and spectral information extraction. Thirdly, we applied an attention-based discriminator to enhance the discriminator's discriminative ability. Finally, the experimental results on the SpaceNet7 dataset and Jilin-1 dataset exhibit the superior of the proposed model both quantitatively and qualitatively.

DERİN ÖĞRENME YAKLAŞIMLARIYLA UZAKTAN ALGILAMA GÖRÜNTÜLERİNDE TEK ÇERÇEVE VE ÇOK ÇERÇEVE SÜPER ÇÖZÜNÜRLÜK

ÖZET

Önemli bir bilgisayarla görüş görevi olarak, süper görüntü çözünürlüğü (SR), uzaktan algılama (RS), tıbbi görüntüleme, video gözetimi ve biyometride yaygın olarak uygulanmaktadır. SR Görüntü, düşük çözünürlüklü (LR) bir görüntü kümesinden yüksek çözünürlüklü (HR) bir görüntü oluşturma tekniğidir. SR Görüntü, düşük çözünürlüklü görüntülerin mekansal, spektral veya zamansal çözünürlüğünü geliştirmek amacıyla hizmet etmektedir. Son yıllarda, birçok araştırmacı, SR algoritmalarının performansının iyileştirilmesine büyük katkılarda bulunmuşlardır. Bu tekniklerden biri ağların derinliğini artırmaktır. Ancak, bu durum büyük miktarda hesaplama ve bellek tüketimini ortaya çıkarmaktadır. Bu nedenle, karmaşık ağların performansını ve eğitim sürecini iyileştirmek için bazı mekanizmalar (örn. basamaklı ağlar, “attention” mekanizması, geri yansıtma) önerilmektedir.

Uydu görüntüleri tarım, arazi örtüsü sınıflandırması, bina çıkarımı, deprem tahmini ve nesne tespiti gibi farklı uygulamalar için yaygın olarak kullanılmaktadır. Ancak, optik ve sensör teknolojisindeki sınırlamalar ve uydulardaki ekipmanların daha da iyileştirilmesi için yüksek maliyetler o söz konusu olmaktadır. Dünya gözlem uydularının mekansal ve spektral çözünürlüğü genellikle istenen yüksek seviyede olmadığından uydu görüntülerinin mekansal ve spektral kalitesini iyileştirmek için yazılım tabanlı algoritmalar geliştirmek önemlidir. Bu tez çalışmasında, derin öğrenme (DL) yöntemlerini kullanarak optik uydu görüntülerinin mekansal çözünürlüğünün artırılması amaçlanmıştır.

Genel olarak, SR görüntüleri tek çerçeveli SR görüntü ve SR çok çerçeveli görüntü olmak üzere iki kategoride sınıflandırılmaktadır.: Bir LR görüntü setinden bir HR görüntüsü oluşturmayı amaçlayan tek çerçeveli SR görüntülerden, orijinal RS görüntülerinden daha iyi mekansal ayrıntılar sağlamak için çözünürlükleri iyileştirebilir. Çok çerçeveli SR görüntü ile , farklı koşullar ve farklı açılar altında elde edilebilen birden çok LR görüntüsünden bir HR görüntüsü oluşturmaya çalışılır.

Son yıllarda, bu konuda çalışan birçok araştırmacı, iki yönü içeren SR görüntü algoritmalarının performansının iyileştirilmesine (1) PSNR'nin (Peak-Signal-Noise-Ratio) performansının iyileştirilmesi; (2) görüntülerin algısal kalitesinin iyileştirilmesi bağlamında büyük katkılarda bulunmuşlardır. Bazı algoritmalar ile yüksek bir PSNR elde edilmiştir. Ancak, bunlar insan algısı için çok daha önemli olan algısal kalite bakımından düşük bir seviyeye sahiptir. Bu nedenle, bu tez çalışmasında aşağıda

sıranan hedefler için çalışılmıştır. (1) Tek çerçeve ve çok çerçeve RS görüntülerinde SR'nin görsel kalitesini iyileştirmek için algısal odaklı bir yaklaşımı keşfetmek; (2) Tek çerçeveli ve çok çerçeveli RS görüntüleri süper çözünürlük için, Üretken Çelişki Ağı (GAN) tabanlı bir modeli, paternin bozulmasını “attention”e almaksızın çok ölçekli çalışmak geliştirmek; (3) SR görüntünün küçük bir yama yerine rastgele bir boyutunu oluşturabilen bir görüntü birleştirme yöntemini geliştirmek.

Bu tez ile önceliklemek çerçeveli ve çok çerçeveli RS görüntüleri için SR yöntemlerine genel bir bakış sunmaktadır. Tek çerçeveli RS görüntüsü SR yöntemleri, kontrollü ve kontrolsüz yöntemler olarak sınıflandırılır. Kontrollü yöntemler temel olarak Evrişimli Sinir Ağları (CNN) tabanlı yöntemleri, GAN tabanlı yöntemleri, “attention” tabanlı yöntemleri ve Geri projeksiyon tabanlı yöntemleri içerir. Ayrıca, kendi kendine “attention”, kanal “attention”i, mekansal “attention”, karma yüksek sıralı “attention” (MHOA), yerel olmayan “attention” (NLA) ve yerel olmayan seyrek “attention” (NLSA) gibi yaygın olarak kullanılan “attention” mekanizmaları da tanıtılmaktadır. Ayrıca derin öğrenme modelinin eğitiminin önemli bir parçası olarak, piksel bazında kayıp, algısal kayıp, çekişmeli kayıp ve döngü tutarlılık kaybını içeren kayıp fonksiyonu sunulmaktadır.

Tek çerçeveli RS görüntüsü SR için, bu tezde “attention” bazlı CNN tabanlı bir SR yöntemi önerilmiştir. CNN tabanlı algoritmalar bilgisayarlı görüş görevlerinde olağanüstü başarılar elde etmiş olsa da, düşük çözünürlüklü girdilerde ve özelliklerde bulunan çokça düşük frekanslı bilgi, geleneksel CNN yöntemlerinde kanallar arasında eşit olarak ele alınır. “attention” temelli algoritmalar, SR görüntüsü dahil olmak üzere zorlu görevlerde bilgilendirici özelliklerin çıkarılmasında hayati bir rol oynamaktadır. “attention” mekanizmasının uygulanmasıyla, önerilen CNN tabanlı yöntem, farklı kanallar arasındaki daha derin ilişkileri daha fazla öğrenebilir. “attention” modülünü, artık bloklarla basitçe entegre etmek yerine, Artık Gruplar (RG) arasındaki ilişkileri daha fazla öğrenmek için bir Katman “attention” Modülü (LAM) ve Mekansal “attention” Modülü (SAM) önerilmiştir. Ayrıca, süper çözölmüş görüntülerin algısal kalitesini iyileştirmek için eğitim sürecinde algısal kayıp işlevi benimsenmiştir ve modelin genelleme yeteneğini geliştirmek için Rastgele aşağı örnekleme kullanılmıştır.

İkinci olarak, tek çerçeve RS görüntüleri için “attention” GAN tabanlı süper çözünürlük yöntemi araştırılmıştır. CNN tabanlı yöntemler, görüntü süper çözünürlüklü görevinde PSNR/SSIM performansının iyileştirilmesine büyük katkı sağlamıştır. Bununla birlikte, bu yöntemler, gerçekçi ayrıntılardan yoksun, aşırı düzgün ve bulanık görüntüler oluşturmaya eğilimlidir. GAN ile, normal CNN tabanlı yöntemlerden daha gerçekçi görüntüler üretebilir ve tek görüntü süper çözünürlüklü (SRGAN, ESRGAN, EEGAN) ile tanıştırılmıştır. Geleneksel evrişimli GAN'lar, düşük çözünürlüklü özellik haritalarında yalnızca mekansal olarak yerel noktaların bir işlevi olarak yüksek çözünürlüklü ayrıntılar üretir. “attention” mekanizması, özellik haritalarındaki uzun menzilli bağımlılıkları doğrudan modellemek için hem üreticiye hem de ayırıcıya daha fazla güç verir. “attention” mekanizmasını uygulayarak, ağ, “attention”i sadece mekansal komşuluktan ziyade renk ve doku benzerliğine göre tahsis edebilir. Bu nedenle, ESRGAN'a dayalı olarak, tek çerçeveli RS görüntüsü SR

için GAN “attention” tabanlı bir yöntem kullanılır. ESRGAN temel olarak iki yönden geliştirilmiştir: (1) daha fazla atlama bağlantısı ekleyerek artık blokların mimarisini daha da geliştirmek; (2) daha fazla özellik çıkarımı için artık bloklara “attention” modülleri eklemek.

Ayrıca, hava fotoğrafları veya düşük çözünürlüklü ve orta çözünürlüklü uydu görüntüleri üzerinde çalışmak yerine, Pleiades tarafından elde edilen çok yüksek çözünürlüklü (VHR) uydu görüntüleri ve Worldview-3'e odaklanılmıştır. Pleiades için, Worldview-3, sırasıyla 2m ve 1.24m'dir. Ayrıca, “attention” CNN tabanlı yöntem için, yöntemi $\times 2$, $\times 4$ ve $\times 8$ ölçek faktörü ile Pleiades ve Worldview-3 veri setleri üzerinde değerlendirilmiştir. Dikkat tabanlı GAN yöntemi için, yöntemi $\times 4$ ölçek faktörü ile Pleiades ve Worldview-3 veri kümeleri üzerinde değerlendirme yapılmıştır. Deneysel sonuçlar, “attention” temelli yöntemin hem nicel hem de nitel olarak algısal olarak daha iyi sonuçlar sağlayabileceğini göstermektedir.

Son olarak, çok çerçeveli RS görüntüsü SR için “attention” tabanlı GAN bir yöntem önerilmiştir. İlk olarak, generatöre bir “attention” mekanizması sunulmuş ve daha iyi zamansal bilgi çıkarımı için her bir çerçeve üzerinde çalışan uzay tabanlı bir ağ önerilmiştir. İkinci olarak, daha iyi mekansal ve spektral bilgi çıkarımı için yeni bir “attention” modülü önerilmiştir. Üçüncü olarak, ayırmacının ayırt etme yeteneğini geliştirmek için “attention” temelli bir ayırmacı uygulanmıştır. Son olarak, SpaceNet7 veri kümesi ve Jilin-1 veri kümesinin deneysel sonuçları ile, önerilen yöntemin hem nicel hem de nitel olarak etkinliği ortaya konmuştur.



1. INTRODUCTION

1.1 Motivation

Image super-resolution (SR) which is a significant task of computer vision has been widely applied in various applications, including remote sensing (RS), biometrics, and video surveillance. SR aims to synthesize high-resolution (HR) output images by enhancing the resolution of the low-resolution (LR) input images.

In recent years, extensive efforts have been made to improve the SR approaches. Satellite imageries are an important geoinformation source for different applications such as environmental monitoring, agricultural crop type identification, land use/cover mapping, disaster management, decision support systems, weather prediction, climate modeling, geographic object detection, biophysical parameter estimation that have been used by a variety of Earth Science disciplines. However, due to the technical limitations and high costs to upgrade the equipment on the satellites, the resolution of satellite images may not meet some application requirements even if they are in the high spatial resolution category; therefore, the software-based SR algorithms are deserved to be explored in RS fields. Remote Sensing Image Super-resolution (RSISR) provides effective solutions to enhance the resolution of LR satellite image inputs, enhancing the image processing applications and denoising. Therefore, RSISR has drawn widespread attention and become more and more popular in RS (Tuna, Unal, & Sertel, 2018).

With the rapid development and the remarkable progress of the deep learning (DL) techniques in a variety of computer vision tasks namely scene classification, object detection, image segmentation, and image SR, many researchers have paid attention and contributed greatly to the DL-based RSISR methods in last several years.

1.2 Definition of Problem

RSISR aims to restore the HR Remote Sensing images from the given LR Remote Sensing images. The degradation model Ψ between HR and LR can be mathematically formulated as:

$$I_{LR} = \Psi(I_{HR}, \delta) \quad (1.1)$$

where I_{LR} and I_{HR} represent the LR and HR RS images, respectively. δ represents the parameters of the downsampling operations. RSISR is the inverse process of the degradation model and mainly focuses on recovering the I_{HR} from the given I_{LR} , which can be written as:

$$I_{SR} = \Psi^{-1}(I_{LR}, \theta) = F(I_{LR}, \theta) \quad (1.2)$$

where θ represents the parameters of the Super-resolution model F , and I_{SR} indicates the super-resolved image.

In general, the learning process is to optimize parameters θ for the Super-resolution model F to make the generated super-resolved result I_{SR} as similar as the HR image I_{HR} . The optimization process can be described as:

$$\hat{\theta} = \operatorname{argmin}(L(I_{SR}, I_{HR})) \quad (1.3)$$

Where $\hat{\theta}$ represents the optimized parameters.

As the supervised methods dominate the DL-based SR in recent years, scholars tend to design RSISR methods by adopting supervised learning. However, in the real world, HR images are usually unavailable. To generate the HR-LR pairs for the training, the bicubic interpolation approach is widely adopted. Researchers usually take the given images as the HR and generate the corresponding LR by applying the Bicubic method with a scaling factor of s , which can be described as:

$$I_{LR} = \downarrow_s(I_{HR}) \quad (1.4)$$

where $\downarrow_s$ denotes downsampling with a factor of s .

Nevertheless, the bicubic interpolation method is not suitable for the real degradation model. Thus, some researchers downsample the HR by using Gaussian Kernels and noise which is described as:

$$I_{LR} = \downarrow_s (I_{HR} \otimes k) + n \quad (1.5)$$

where k and n denote the blurring kernel and noise and $\otimes$ represent the convolution operation, respectively.

For the multi-frame Remote sensing image super-resolution, the input images are LR frames. Thus, equation 1.2 can be rewritten when the input frames are set to 5:

$$\begin{aligned} I_{SR} &= \Psi^{-1}(\{I_{LR}^{t-2}, I_{LR}^{t-1}, I_{LR}^t, I_{LR}^{t+1}, I_{LR}^{t+2}\}, \theta) \\ &= F(\{I_{LR}^{t-2}, I_{LR}^{t-1}, I_{LR}^t, I_{LR}^{t+1}, I_{LR}^{t+2}\}, \theta) \end{aligned} \quad (1.6)$$

where t denotes the temporal index of the input frames.

1.3 Contributions Of The Dissertation

The objectives of the thesis mainly contain three aspects:

- (1) Explore a perceptual-driven approach to perceptually enhance the SR image quality on single-frame and multi-frame remote sensing imagery;
- (2) Explore GAN-based models for single-frame and multi-frame RS imagery super-resolution tasks to fulfill the multi-scale problem and blind to the degradation model;
- (3) Explore an image fusion method that can generate an arbitrary size of the super-resolved image rather than a small patch;

This thesis contains the following chapters: In Chapter 2, the review of the image super-resolution on natural images and remote sensing images is given. Chapter 3 outlines the datasets and methodology of the single-frame and multi-frame RSISR utilized in this study. Then, the results and discussions of the single-frame and multi-frame RSISR are presented in Chapters 4 and 5, respectively. Finally, Chapter 6 provides the conclusions and recommendations for the future in this field.



2. LITERATURE REVIEW

2.1 Review of the Single-Frame Remote Sensing Image Super-resolution

2.1.1 Deep learning-based SISR

2.1.1.1 Deep CNN-based SR methods

Generally speaking, the Deep CNN-based SR method learns the feature maps of the images automatically by stacking the convolutional layers and activation layers together. A Super-resolution Convolutional Neural Network (SRCNN) that utilizes just three convolutional layers for the image SR task and outperformed the traditional bicubic interpolation method was proposed by Dong et al. (C. Dong, Loy, He, & Tang, 2014). SRCNN also was applied in a pan-sharpening and remote sensing RS task and showed outstanding performance than the conventional algorithms. The SRCNN architecture is too shallow to learn the deep features of the images. A very deep SR for image SR (VDSR) approach that contains 20 convolutional layers was proposed by Kim et al. (Hara & Tanaka, 2016). By learning the residuals of the output and the input, the proposed residual learning speeds up the training convergence. Nevertheless, as SRCNN and VDSR are trained in high-dimensional space, they bring high complexity and computational time. To solve this problem, a group of methods (post-processing) is proposed to train the models directly on LR images. In (Shi et al., 2016), an efficient sub-pixel convolutional neural network (ESPCN) upscales the LR feature maps by proposing a sub-pixel convolutional module and significantly improves the reconstruction speed. Fast super-resolution (C. Dong, Loy, & Tang, 2016) (FSRCNN) is also directly trained on the LR images and it up-samples the feature maps by using a deconvolutional layer. Due to the noticeable advantages of the post-processing structure, it has become mainstream in the DL-based SR methods.

Increasing the depth of the model is an effective solution to better extract the hierarchical feature information from LR inputs, and further enhance the reconstruction image quality. However, the difficulty of the model training and the parameter of the model increase as well. To address these issues, a deeply-recursive convolutional network (DRCN) (Kim, Lee, & Lee, 2016) uses a deep recursive convolutional network which is up to 16 recursions. By adding a residual architecture to DRCN, Deep Recursive Residual Network (DRRN) (Tai, Yang, & Liu, 2017) designed a deeper (up to 52 convolutional layers) yet concise network. In (Ahn, Kang, & Sohn, 2018b), the authors proposed an accurate and lightweight deep network named Cascading Residual Network (CARN) for image SR by designing a network that combines a residual network with the cascading mechanism. In (Mao, Shen, & Yang, 2016), the authors proposed a Residual Encoder-Decoder Network (RED-Net) for image enhancement tasks such as denoising, JPEG deblocking, image deblurring, image inpainting, and SR. A Laplacian Pyramid Super-Resolution (LapSRN) (Lai, Huang, Ahuja, & Yang, 2017) was proposed for the progressive reconstruction of the sub-band residuals of HR images. Tong et al. (Tong, Li, Liu, & Gao, 2017) introduced dense skip connections (SRDenseNet) in a very deep network which efficiently combined the low-level and high-level features to accelerate the reconstruction performance. Lim et al. (Lim, Son, Kim, Nah, & Lee, 2017) developed an enhanced deep super-resolution network (EDSR) which boosted the performance by eliminating the unnecessary modules in conventional residual networks. In (Yu et al., 2018), a Wide Activation for Efficient and Accurate Image Super-Resolution (WDSR) was proposed which simply expanded features before ReLU activation instead of using various shortcut connections resulted in significant improvements for SISR.

2.1.1.2 Attention-based methods

Most current SR studies attempt to enhance the performance by increasing the depth of the model. However, deeper layers require extensive computation and memory resources. Despite CNN-based algorithms having been extensively implemented in diverse assignments and exhibiting remarkable achievements, the traditional CNNs

methods treat the abundant low-frequency information included in the LR inputs equally across channels. The attention mechanism successfully addresses these issues by learning further correlations among channels. Attention-guided algorithms play a vital role in the informative features extraction in various tasks including image super-resolution. Many researchers made outstanding contributions to the proposal of different kinds of attention-based methods for the more abundant informative feature extraction (Peijuan Wang & Sertel, 2021)(J. Chen, Zhang, Hu, & Chen, 2021)(Gao, Sun, Cui, Du, & Sun, 2021)(Dai, Cai, Zhang, Xia, & Zhang, 2019).

Channel attention, spatial attention, and pixel attention are three significant modules among these methods. Channel attention (Hu, Shen, Albanie, Sun, & Wu, 2020) intends to learn a 1-Dimension ($C \times 1 \times 1$) vector feature map. Spatial attention (Woo, Park, Lee, & Kweon, 2018) aims to learn a 2-Dimension ($1 \times H \times W$) attention feature map. Pixel attention (Zhao, Kong, He, Qiao, & Dong, 2020) aims to learn a 3-Dimension ($C \times H \times W$) feature map. Cheng et al. (Cheng, Li, & Yang, 2018) proposed Recursive Squeeze and Excitation Networks (SESR) for image SR. The SESR unit is designed to learn the channel interdependencies and considerably improves the performance. By taking advantage of the attention module, Zhang et al. (Y. Zhang, Li, Wang, & Fu, 2018) presented a residual channel attention network (RCAN) for image SR. Moreover, a residual in residual (RIR) network was proposed to concentrate on learning the high-frequency information. Later, Dai et al. (Dai et al., 2019) proposed a deep second-order attention network (SAN) by exploiting second-order feature statistics instead of first-order ones as channel attention does. Moreover, Zhao et al. (Zhao et al., 2020) designed a lightweight CNN for efficient image SR (namely PAN) by proposing a pixel attention mechanism that can generate 3D attention maps. Later on, in (Niu et al., 2020), the authors proposed a Holistic Attention network for image SR (HAN) which includes a LAM (Layer attention module) and a CSAM (Channel-spatial attention module) for the further learning of the correlations among different layers rather than a single layer.

2.1.1.3 Cascading networks

Cascading is another solution to ameliorate the performance of the SR algorithms. Zhang et al. (F. Zhang et al., 2018) addressed a novel cascaded CNN framework to implement a multi-scale image SR task with a single trained model. There are two subnetworks. Different subnetworks use different convolutional kernels with different sizes to achieve multi-scale feature mapping. Wei et al. (Wei et al., 2019) constructed the cascading dense network to fully make use of the hierarchical features from all the convolutional layers for an accurate SISR method. Lan et al. (Lan et al., 2020) proposed a cascading residual network (CRN) and local wider residual blocks (LWRBs) for the SISR. Another enhanced residual network (ERN) which employs a dual global pathway structure and multiscale block (MSB) was also proposed. Ahn et al. (Ahn, Kang, & Sohn, 2018a) applied a progressive learning scheme to the deep convolutional neural network by multiple-stage learning resulting in a gradual increase of the output image resolution.

2.1.1.4 GAN-based methods

With fast evolution and remarkable achievements of DL, Generative Adversarial Networks (Goodfellow & Pouget-Abadie, 2014) (GAN) have swiftly become the hot topic of academic research and industry application soon after its emergence in various fields and computer vision tasks. Ledig et al. (Christian Ledig Ferenc Huszar, 2017) presented an image SR GAN model (SRGAN) in which a deep residual network and a perceptual loss using high-level feature maps of the pre-trained VGG network were employed to generate photo-realistic images. However, there were still artifacts available on the images. To enhance the image quality visually, Enhanced Super-Resolution Generative Adversarial Networks (Xintao Wang, Yu, Wu, Gu, & Liu, 2018) (ESRGAN) was designed by proposing a Residual-in-Residual Dense Block (RRDB) in the generator, removing the Batch Normalization from the Basic block, and training the network with relativistic GAN. ESRGAN exhibited perceptually realistic and natural textures compared to CNNs.

Recently, as the attention mechanism has been shown effective in extracting the feature and deep interdependencies among the channels, it is also applied to the GANs. By introducing the self-attention mechanism into the GANs, Zhang et al. (Han Zhang, Goodfellow, Metaxas, & Odena, 2019) proposed a Self-attention Generative Adversarial Network (SAGAN) which can model long-range, multi-level dependencies across image regions and produce fine details at every location.

2.1.2 Supervised methods for RSISR on remote sensing images

The RSISR focuses on restoring HR outputs from its given LR inputs. RS images have special characteristics: (a) The images are obtained from a very long distance, several hundred kilometers away from the Earth's surface; (b) The images contain a lot of small objects and diverse distribution, including various land surface scenes and ground objects of different sizes; in addition, objects always share a joint distribution with their environment; (c) The scene images have rather wide differences due to the images being always captured under changing seasons, varying atmospheric conditions, and different sensor viewing geometry; (d) The currently proposed methods for RSISR issue mainly include two categories: Supervised and Unsupervised RSISR methods. A detailed review of these methods will be provided. Moreover, we also present the dataset, applied methodology, and accuracy assessment from previous studies in Appendix 1. The majority of the existing RSISR approaches are supervised learning-based methods which are trained with the HR-LR pairs. In this subsection, we will discuss the CNN-based, GAN-based, Attention-based, and Back-projection-based RSISR methods.

2.1.2.1 CNN-based methods

Due to the remarkable achievement of DL in computer tasks and image SR, a group of researchers started to work on the DL-based RSISR.

The study proposed by Leibel et al. (Liebel & Körner, 2016) in 2016 is the first CNN-based work applied to the RSISR task. In this study, the authors explored the CNN method on Sentinel-2 images and showed that the CNN methods can be successfully

applied to RS data (Sentinel-2 images) both in the Y channel (which is luma components in YCbCr color space) and multispectral (MS) remote sensing images by fine-tuning the SRCNN with the particular dataset. However, they focused only on the third band of the multispectral images for the performance observation. Tuna et al. (Tuna et al., 2018) applied the SRCNN (C. Dong, Loy, He, & Tang, 2016) and VDSR (Hara & Tanaka, 2016) models incorporating IHS transformation on the satellite images acquired by VHR SPOT6 and 7 and the Pleiades 1A and 1B satellites with a scale factor of 2, 3, and 4. The experimental results demonstrated that VDSR provided superior performance than SRCNN both on PAN and MS images. In 2017, Huang et al. (N. Huang, Yang, Liu, Gu, & Cai, 2017) reported that directly applying the VDSR on Sentinel-2A cannot generate satisfactory results. Thus, they proposed a remote sensing deep residual-learning (RS-DRL) network and proved to outperform VDSR (Hara & Tanaka, 2016) on Sentinel-2A images. However, due to the feature complexity of the RS images, it is deserved to propose and design specified algorithms for RSISR.

Lei et al. in (Lei, Shi, & Zou, 2017) developed a local-global combined network (LGCNet) which contains a local-global combination with a multiforked structure by concatenating the results from different convolutional layers to learn multi-level representations of RS images. Such a design can not only learn the local details including the edges and contours of an object but also learn the global features including environment types. Moreover, in (Xu et al., 2018), the authors proposed a deep memory connected network (DMCN) which can generate high-quality outputs by applying local and global memory connections. To reduce memory, the authors proposed down-sampling and up-sampling units to conduct an hourglass structure. Experimental results showed that DMCN outperforms LGCNet, SRCNN, and VDSR on NWPU-RESISC45, UC Merced dataset, and Gaofen-1 testing images with scales of 2, 3, and 4.

To boost the performance of the DL-based models, one group of scholars put forward dense residual blocks. To enhance the performance of the CNN-based network, in

(Jiang, Wang, Yi, Jiang, et al., 2018), the authors presented a novel deep distillation recursive network (DDRN) which includes a group of ultra-dense residual blocks UDB, a multi-scale purification unit (MSPU), and a distillation mechanism for RS images. The proposed MSPU module can effectively restore the high-frequency components. The proposed DDRN showed better performance on Kaggle open dataset and Jilin-1 satellite images. Deeba et al. (Deeba et al., 2020) proposed a wide RS residual network for SR(WRSR). The proposed WRSR improved the training loss performance and the quality of the super-resolved results with a wider but shallower residual network, and performing the weight normalization,

Recently, Ren et al. (Ren, He, Qing, Wu, & Pu, 2021) proposed an enhanced residual convolutional neural network (ERCNN) with a dual-luminance scheme (DLS) for RSISR which enhanced the feature flow module and ability to distinguish learning across feature maps.

To release the computation burden and reduce the number of parameters, Wang et al. (Jin Wang et al., 2021) proposed a feedback ghost residual dense network (FGRDN) for RSISR. The proposed FGRDN adopts the feedback mechanism and ghost module to obtain the lower features through high-level refining with lower parameters and applied the spatial and channel attention module (SCM) at the end of the RDB to learn more useful information from features. In addition, Chang et al. (Chang & Luo, 2019) proposed a Bidirectional Convolutional LSTM Neural Network for RSISR. The proposed method consists of a recursive inference block based on dense connection and a bidirectional convolutional LSTM (BiConvLSTM) layer to understand the feature correlations from each recursion and implement an adaptive selection for the complementary information of the reconstruction layer.

Most of the SR algorithms adopt only a single up-scaler to up-sample the LR feature, such as preprocessing methods (SRCNN) and post-processing (using sub-pixel convolutional or deconvolutional layer). This causes the loss of the LR image information. To emphasize this issue, Huan et al. (Huan et al., 2021) proposed a new pyramidal multi-scale residual network (PMSRN) by fusing the hierarchical features

and forming a multi-scale dilation residual block (MSDRB). The proposed PMSRN can enhance the ability to extract the image features and its parameters are only 43.33% of that of EDSR.

Zhang et al. (J. Zhang, Liu, Peng, & Li, 2020) implemented a progressive residual depth neural network (PRDNN) for RSISR. The proposed progressive residual structure can provide more detailed features by gradually learning the feature maps of satellite images at different levels and different receptive fields. The proposed PRDNN can generate more accurate edges and textures information

In (Shao, Wang, Wang, & Deng, 2019), a novel coupled sparse autoencoder (CSAE) that combines the sparse learning and autoencoder was proposed for RSISR. The CSAE benefits from a set of sparse coefficients as the prior knowledge and can effectively and accurately model the relation between the LR and HR pairs.

To enhance the representation power of the network to further extract the features contained in RS images, some research aims at extracting the features by applying multi-scale networks.

Fu et al. (Fu, Liu, & Li, 2018) implemented a CNN-based SR by applying different convolution filters for a group of features. An end-to-end training was implemented on Jilin-1 and exhibited superior to FSRCNN.

Lu et al. (Lu, Wang, Zhang, Wang, & Jiang, 2019) also considered a multi-scale residual neural network (MRNN) to exploit the multi-scale properties of objects in RS imageries for accurate SR reconstruction. By extracting three scales of the image at large-, middle- and small-scale features for the multi-scale information representation, and fusing the multi-scale information, the proposed can generate visually better images. In (Xinying Wang, Wu, Ming, & Lv, 2020), the authors applied an adaptive multi-scale feature fusion network (AMFFN) for the RSISR task. By applying adaptive multi-scale feature extraction (AMFE) modules and the squeeze-and-excited and adaptive gating mechanisms, the proposed AMFFN can adaptively extract multi-scale features and adaptively reserve the feature information to enhance the efficiency

of the information usage. Multi-scale feature mapping was also addressed in (Keshk & Yin, 2020), by using a multi-scale feature mapping structure that allows the network to apply different kernels sizes in a single-layer convolutional layer, an enhancement deep convolutional neural network algorithm (EDCNN) was proposed to better learn the information in each convolutional layer to obtain the SR satellite images. More recently, in (Deeba et al., 2021), the authors proposed a transferred wide residual SIRSSR deep neural network model (WRSR) which considerably reduced memory costs and improved the accuracy with the increase of width and decrease in the residual network depth. The previous work usually processes images in the original domain; some researchers trained the model in multiscale frequency bands. Wang et al. (T. Wang, Sun, Qi, & Ren, 2018) developed a CNN-based SR algorithm together with wavelets on aerial images. The authors trained multiple models at different frequency bands for SR to restore frequency features in different directions. The proposed method successfully captures both high-frequency local changes and low-frequency global layouts. Moreover, in (W. Ma, Pan, Guo, & Lei, 2019), a method that incorporates the wavelet transform (WT) and the recursive Res-Net was introduced to restore HR images in frequency realm. Meanwhile, to alleviate the training process, recursive and residual modules are widely utilized. In addition, it is proven that it can ease the network and save memory by removing of the batch normalization.

Generally, the majority of these methods are trained with RGB images (8-bit) in the same way as natural image SR methods. However, the radiometric resolution of satellite images is usually 11-bit or 12-bit and includes multi-spectral bands that can provide more abundant radiometric and spectral information. Thus, Wagner et al. (Wagner, Liebel, & Körner, 2019) adapted a generic SISR algorithm-VDSR for dealing with the various spatial resolutions in Sentinel-2 images. By training the four 10 m bands of Sentinel-2 images with their down-sampled 20 m images, the model can super-resolve the images from 20 m to 10 m. Experiments showed that the proposed method outperformed the original VDSR. The work conducted by Müller et al. (Müller, Ekhtiari, Almeida, & Rieke, 2020) also applied satellite images. By

applying the SFIM pan-sharpening method to generate the pan-sharpened as the HR images, the authors proposed an RSISR method that takes advantage of Panchromatic images and super-resolves all bands in the multi-spectral images.

Based on our analysis, how to fully utilize the abundant information contained in multispectral images has been limitedly explored and deserves detailed further research in the future.

Another challenge in the RSISR task is how to super-resolve the real satellite images. Zhu et al. (X. Zhu, Talebi, Shi, Yang, & Milanfar, 2020) proposed a data generation model more practically. Moreover, by proposing a residual convolutional neural network and a combined perceptual loss function, the proposed approach showed the ability to improve the quality of real satellite images.

2.1.2.2 GAN-based methods

The Generative Adversarial Network (GAN) was recently applied to RSISR. A transferred GAN-based method TGAN was explored in (W. Ma, Pan, Guo, & Lei, 2018). In TGAN, batch normalization was deleted for calculation reduction and accuracy improvement. The model was firstly implemented on DIV2K natural dataset, then transferred with the RS images. In (Jiang et al., 2019), another EEGAN architecture was proposed. In particular, the generator of EEGAN contains an ultradense subnetwork (UDSN) and an edge-enhancement subnetwork (EESN). The EEGAN surpasses other SR approaches with the experimental results on several datasets. Wang et al. (Z. Wang, Jiang, & Yi, 2019) proposed an ultra-dense GAN (udGAN) by integrating the ultra-dense connections and GAN training strategy.

Lanaras et al. (Lanaras, Bioucas-dias, Galliani, & Baltsavias, 2018) proposed an improved Enhanced SR Generative Adversarial Network (ESRGAN) trained on WorldView-Sentinel image sets to create super-resolved multispectral Sentinel-2 output. In (Pineda, Ayma, & Beltran, 2020), a GAN-based SR model was presented to enhance the Sentinel-2 imageries by taking PeruSat1 images as the training dataset. Gong et al. (Gong et al., 2021) proposed an Enlighten-GAN for SR task by designing

the enlightened blocks, using the Self-Supervised Hierarchical Perceptual Loss for the performance improvement, and employing the internal inconsistency loss and cropping-and-clipping strategy for seam line avoidance. These methods were carried out on mid-resolution satellite imageries. Nevertheless, how to super-resolve a VHR image (e.g. 50 cm or 30 cm) to generate a higher resolution image is still practical and challenging work.

In (Sustika, Suksmono, Danudirdjo, & Wikantika, 2020), by applying residual dense network (RDN) in the generator network, the authors proposed a generative adversarial network (GAN) based approach which provides high performance on classical (objective) evaluation metrics along with high perceptual quality for RSISR. In (B. Liu et al., 2020), the authors proposed a cascaded conditional Wasserstein generative adversarial network (CCWGAN) by designing the residual dense block to obtain HR RS images. . (Guo, Xia, Xu, Li, & Luo, 2021) developed a Cascade Generative Adversarial Nets (CGAN) by designing an edge enhancement module conserving edge details and suppressing the noise. In (W. Ma, Pan, Yuan, & Lei, 2019), the authors implemented a dense residual GAN (DRGAN)-based Single RSISR model which can harness the hierarchical features and stably train the model by applying Wasserstein GAN with gradient penalty (WGAN-GP) to strengthen RS imageries.

Apart from the enhancement of generator capability, the designing of the discriminator is also crucial for the improvements of the GAN-based results. In (Lei et al., 2019), the authors implemented coupled-discriminated GANs (CDGANs) for RSISR, which is a GAN-based super-resolution algorithm, to solve the 'discrimination-ambiguity' problem by introducing a dual pathway network architecture, a random gate, and a coupled adversarial loss.

Despite the excellent achievements of GANs in generating more perceptually realistic images, attention GAN-based methods are proved to be effective in enhancing the representation power of the generator. Therefore, some attention GAN-based networks are proposed for the RSISR. Moustafa et al. (Moustafa & Sayed, 2021) proposed an SCSE-GAN for satellite image SR by simply stacking Spatial and Channel Squeeze-

and-Excitation (SCSE) block after each residual block and introducing Wasserstein distance with gradient penalty (WGAN-GP). In (Gao et al., 2021), a GAN-based SR approach uniting residual channel attention (CA) was introduced to enhance deep feature extraction ability and produce sharper edges for RS images. Recently, a multi-attention GAN framework that consists of a pyramidal convolution residual-dense block (PCRDB) and the attention-based up-sample (AUP) block was proposed in (Jia, Wang, Jia, & Xu, 2022) for RSISR to achieve up-sampling at any scale. Moreover, local attention and global attention-based GAN (SRAGAN) was presented by (Yadong Li et al., 2021) for RSISR to learn the structural components and long-range correlations in the channel and spatial domains.

Although GAN-based methods can generate more realistic results, it is also experimentally found that the original image does not include some reconstructed textures. Therefore, in (Z. X. Huang & Jing, 2020), the authors proposed an SR reconstruction method combining wavelet transform and generative adversarial network for RSISR reconstruction. The experiments showed that training wavelet decomposition coefficients in the GAN can effectively improve the local detail information of the reconstructed image.

2.1.2.3 Attention-based methods

Due to the remarkable success of the attention mechanism in the image SR task, many researchers apply the attention mechanism to the RS fields, including pan-sharpening (Peijuan Wang & Sertel, 2021) and image SR, and improve the performance effectively. Deep Residual Squeeze and Excitation Network (DRSEN) (Gu, Sun, Zhang, Fu, & Wang, 2019) was presented for RSISR by proposing a residual squeeze and excitation block (RSEB) which includes a local feature fusion (LFF) unit and squeeze and excitation (SE) unit to model the interdependencies among channels and strengthen the representation ability. A multi-perception attention network (MPSR) (X. Dong, Xi, & Sun, 2019) consisting of the enhanced residual network (ERB) and residual channel attention group (RCAG) was introduced to RSISR. The MPSR exhibited superior results by being pre-trained on DIV2K. A novel RSISR algorithm

(Haut, Fernandez-beltran, Paoletti, Plaza, & Plaza, 2019) combining attention strategy with residual nets was designed. The authors utilized a similar architecture with RCAN (Y. Zhang et al., 2018) and proved to intensely improve the performance. Recently, Wang et al. (Peijuan Wang, Bayram, & Sertel, 2021) applied CBAM to the deep dense residual network for the performance enhancement of the SISR solution while focusing on the VHR Satellite Images. Experimental results on the VHR satellite image dataset, including WV-3 (DFC 2019) and Pleiades images, showed the usefulness of the attention mechanism on performance improvement. Another similar work was demonstrated in (Peng, Wang, Zhang, & Liu, 2021), by designing several residual blocks which contain an additional well-designed gated convolution unit with long skip connections, a pre-training of gated CNN (PGCNN) that can focus on learning high-frequency information was proposed for RSISR. By applying transfer learning, the proposed method can produce more texture and detailed information. However, these methods utilize a simple and direct way to just add the attention module inside the residual blocks, and cannot show the deeper features of the LR. Recently, a second-order multi-scale SR network (SMSR) was (X. Dong, Wang, et al., 2020) presented for RSISR. In (D. Zhang, Shao, Li, & Shen, 2020), a mixed high-order attention network (MHAN) was introduced for RSISR. The proposed high-order attention module (when $R=3$) is proved to be more effective than the first-order attention module (channel attention). Li et al. (X. Li, Zhang, Liang, Ouyang, & Shao, 2020) presented a fused recurrent network via channel attention (CA-FRN) by retaining and merging hierarchical local information of LR and HR for RSISR. a multiscale attention network (S. Zhang et al., 2020) for the extraction of the multi-level feature from different scenes of RS images was proposed. In (B. Huang, He, Wu, & Guo, 2021), by developing a residual dual-attention block (RDAB) including a local multi-level fusion module and a dual-attention mechanism (DAM) enabling the network to focus on high-frequency information areas, the authors proposed a deep residual dual-attention network (DRDAN) to achieve the global and local feature combination for RSISR. Wang et al. (H. Wang, Hu, Wu, Chi, & Yu, 2020) proposed a non-locally up-down convolutional attention network (NLASR) which includes a

non-local features enhancement module (NLEB), an enhanced up-sampling channel-wise attention (EUCA) module, and an enhanced down-sampling spatial-wise attention (EDSA) module for RSISR. Integration of the channel-wise and multi-scale spatial information resulted in more precise mapping from LR to HR.

To enhance the representation ability with fewer parameters, a parallel-connected residual channel attention network (PCRCAN) (Yinhao Li, Iwamoto, Lin, & Chen, 2020) was presented for RSISR. PCRCAN consists of a parallel module with feature aggregation modules based on group convolution and uses a residual channel attention block as the basic feature mapping unit. More recently, Chen et al. (L. Chen et al., 2021) proposed a residual aggregation and split attentional fusion network (RASAF) to generate high-quality RS images. The RASAF contains a split attentional fusion block to attain cross-channel feature group interplay for various land surface scene reconstructions and a residual aggregation mechanism for higher local residual feature generation. Meantime, a hierarchical loss function is used for multi-level feature learning.

Due to the rather restricted priors in LR space, RSISR with large-scale (e.g. $\times 8$) is a challenging task that remains to explore. In (X. Dong, Sun, et al., 2020) designed a dense-sampling mechanism and wide attention block (WAB). Their approach, dense-sampling SR network (DSSR), explored the large-scale RSISR. The proposed DSSR was proved to provide prominent performance even under large scaling factors $\times 4$ and $\times 8$. Ma et al (Y. Ma, Lv, Liu, Sun, & Zhong, 2021), implemented a dense channel attention network (DCAN) for RSISR in which they designed a dense channel attention mechanism (DCAM) to harness multi-level feature maps.

2.1.2.4 Back-projection based methods

Back projection is another method utilized for the performance enhancement of DL-based image SR. Deep Back-Projection Networks (DBPN) (Haris, Shakhnarovich, & Ukita, 2018) that could iteratively leverage down- and up-sampling layers and provide a stage-based error feedback mechanism for projection errors was presented. Up- and

down-sampling stages were mutually connected to define different kinds of image degradation and HR elements. A Residual Dense Back-projection Network (RDBPN) (Pan, Ma, Guo, & Lei, 2019) was proposed for RSISR. Global and local residual learning was employed with the presented residual dense back-projection block. Then fast RDBPN (FRDBPN) was implemented to boost the speed of reconstruction but changing the down-projection module to the downscaling module. Moreover, by introducing attention mechanisms, Dong et al. (X. Dong, Xi, Sun, & Yang, 2020) proposed an Enhanced Back-Projection (EBPN) for RSISR which improves the DDBPN (Haris et al., 2018) for channel-wise feature extraction.

2.1.3 Unsupervised methods for SISR on remote sensing images

Although supervised RSISR methods have produced successful results, they suffer from the lack of paired HR and LW images. The traditional supervised algorithms cannot meet the requirement in the application of generating real RS images with complex data acquisition processes. Therefore, it is of paramount significance to develop unsupervised methods for RSISR. In the past years, substantive contributions have been devoted to unsupervised RSISR algorithms.

In (Sheikholeslami, Nadi, Naeini, & Ghamisi, 2020), the authors proposed an efficient unsupervised SR (EUSR) DL model for RSISR. The proposed EUSR boosts the reconstruction performance and alleviates the computational burden by using densely connected convolutional layers and skip connections. In (Choi & Kim, 2020), the authors proposed a new degradation model that emphasizes naturalness and a modified SRResNet to enhance the KOMPSAT-3 image quality. A method that transforms the reference into the latent space as the migration image prior to unsupervised high temporal RSISR (Jiaming Wang, Shao, Lu, et al., 2021) was proposed. Later on, they (Jiaming Wang, Shao, Huang, et al., 2021) implemented an unsupervised learning net-Enhanced Image Prior (EIP) to fulfill SR tasks by advancing the previous work with an adaptation of a recurrent update strategy for transforming the reference into a latent space as the enhanced image prior.

Another group of researchers benefitted from the idea of CycleGAN. In (Pengrui Wang, Zhang, Zhou, & Jiang, 2019), the authors proposed an unsupervised Cycle-CNN model including two generative CNNs for down-sampling and SR separately for RSISR. More recently, by proposing two-cycle modules - The first module is used for mapping LR to HR, while the second module maps HR images back to LR images, like down-sampling, the authors (Haopeng Zhang, Wang, & Jiang, 2020) proposed a Cycle-CNN which can be trained with unpaired images for RSISR.

Moreover, GAN is also applied to unsupervised learning-based RSSR. Zhang et al. (N. Zhang, Wang, Zhang, Xu, & Wang, 2020) employed an unsupervised generative adversarial network (GAN)-based RSISR method that does not need HR RS images in the training process. Haut et al. (Haut et al., 2018) implemented an unsupervised model which can primarily learn relationships between LRs and HRs via various convolutional, down-sampling, batch normalization, and activation layers to super-resolve LR RS data.

In summary, according to our analysis above, we have some findings regarding the problem of SISR:

- (a) In general, GAN-based approaches can generate perceptually high-quality images but with low PSNR values. While CNN-based methods usually achieve high PSNR values but with low perceptual quality.
- (b) By applying the attention and back-projection mechanisms, we can extract deeper information and improve the performance and image quality.
- (c) With the increasing number of Earth Observation satellites for various purposes in recent years, and wider applications of satellite imagery; a common benchmark dataset including multi-sensor VHR satellite images is needed to ensure generalizability and transferability of the proposed approaches for different satellite images, locations and landscape characteristics.

2.2 Review of the Multi-Frame Remote Sensing Image Super-Resolution

2.2.1 Deep-learning based video super-resolution

Recently, plenty of DL-based video SR algorithms have blossomed with the speedy advancement of Artificial Intelligence (AI) and Deep Learning (DL) techniques. As the video SR usually was processed frame by frame, therefore, it provides us with great ideas for the multi-frame Remote Sensing SR. As demonstrated in (H. Liu et al., 2020), the typical paradigm of the deep learning methods for video SR includes three modules: Inter-frame alignment module, Feature extraction & fusion module, and Reconstruction module. Inter-frame alignment modules usually are traditional methods or deep CNNs. The feature extraction & fusion module and the reconstruction module usually adopt deep CNNs. The inter-frame alignment module is optional. Therefore, according to if the methods utilize the alignment module, the deep learning-based VSR algorithms are categorized into methods with alignment and methods without alignment. The methods with alignment include methods using motion estimation and motion compensation (MEMC) (eg. VESPCN, DRVSR, FRVSR, MMCNN), and Deformable Convolution (DC) methods (eg. EDVR (Xintao Wang, Chan, Yu, Dong, & Loy, 2019), TDAN (Tian, Zhang, Fu, & Xu, 2020), D3Dnet (Ying et al., 2020)). The methods without alignment include 2D convolution-based methods (FFCVSR), 3D convolution-based methods (DUF (Jo, Oh, Kang, & Kim, 2018), 3DSRnet), Recurrent convolutional neural network (RCNN) based methods (STCN), and Non-local attention-based methods (PFNL (Yi, Wang, Jiang, Jiang, & Ma, 2019), MuCNN, MSHPFNL (Yi et al., 2020)).

From another perspective, the VSR methods were categorized into time-based and space-based methods based on how to utilize temporal information (Yi et al., 2019). Time-based methods regard frames as time-series data and deal with frames in sequence. Space-based methods consider multiple frames as supplementary materials to assist the reference frame regeneration. Space-based methods include Direct fusion, Slow fusion, and 3D convolution methods.

Based on SRCNN (C. Dong, Loy, He, et al., 2016), VSRnet (Kappeler, Yoo, Dai, & Katsaggelos, 2016) mainly consisting of MEMC and convolutional layers were proposed. A speedy and precise VSR algorithm (VESPCN) (Caballero et al., 2017) by combining the sub-pixel convolution with the Spatio-temporal networks and MC was proposed. In the detail-revealing deep video superresolution (DRVSR) (Tao, Gao, Liao, Wang, & Jia, 2017) method, the authors proposed a sub-pixel MC layer (SPMC) performing the up-sampling and MC concurrently and adopted the ConvLSTM module to deal with Spatio-temporal information.

A VSR algorithm containing MC and residual net (MCResNet) (Dingyi Li & Wang, 2017) which is able to handle large and complex motions adaptively was proposed. Robust VSR with Learned Temporal Dynamics (LTDVSR) (D. Liu et al., 2017) that includes a temporal adaptive network and spatial alignment network to utilize temporal information efficiently and effectively was proposed. Motivated by the dynamic filter network, dynamic upsampling filters (DUF) (Jo et al., 2018) were proposed. The DUF avoids applying MEMC by combining the dynamic up-sampling filter with the Spatio-temporal information learned by 3D convolution.

The progressive fusion network via Non-local attention (PFNL) for VSR proposed by Yi et al. (Yi et al., 2019) follows the space-based approach. Non-local attention was applied before the feature extraction module to replace the complex MEMC procedures. Later, Yi et al. (Yi et al., 2020) developed a Multi-scale structure and hybrid convolutions-based progressive fusion net via Non-local attention (MSHPPFNL) for VSR. The authors also introduce the GAN to the network for more realistic and temporally consistent results.

2.2.2 Multi-frame RSISR

Recently, thanks to the launch of the multi-temporal and video satellite, such as “Jilin-1” (‘Jilin-1 Smart Video Satellites’, 2017), “Gaofen-4”, “OVS-1” and the Planet, the studies on the multi-frame super-resolution became possible and some researchers also

made some contributions to the technical development of multi-frame SR. Thus, research on multi-frame SR has a quite promising future.

In (Luo, Zhou, Wang, & Wang, 2017), a modified VDSR is trained on the Gaofen-2 dataset and tested on Jilin-1 videos to prove the effectiveness of the model. In (A. Xiao, Wang, Wang, & Ren, 2018), the authors proposed a five-layer end-to-end network by carrying out a reshape or deconvolution operation to retain the distribution of the ground objects. The proposed network is trained with extracted Jilin-1 video satellite frames. A progressively enhanced network (PECNN) (Jiang, Wang, Yi, & Jiang, 2018) which is composed of a pre-training CNN-based network to reconstruct a basic HR image and an enhanced dense connection network to enhance the previous output was proposed. The proposed methods were tested on Jilin-1 video images and Kaggle Open Dataset. Later on, deep distillation recursive network (DDRN) (Jiang, Wang, Yi, Jiang, et al., 2018) which includes a set of ultra-dense residual blocks (UDB), a multi-scale purification unit (MSPU), and a reconstruction module was presented. Distillation and compensation technique was introduced to the network. However, these methods are trained on a single image and just test on the video satellite images but cannot fully utilize the temporal information contained in the video satellite images.

Recently, a recurrent refinement network with an UNet structure for the complementary information enhancement for Jilin-1 satellite video data SR is proposed by Xiao et al. (Y. Xiao, Su, & Yuan, 2021). By coupling MFSR and edge-SFSR (E-SFSR), Shen et al. (Shen, Qiu, Yue, & Zhang, 2021) proposed an edge-guided VSR (EGVSR) framework for Jilin-1 and OVS-1 video satellite image SR. By applying an optical flow estimation subnetwork (OFEnet), an upscaling subnetwork (Upnet), and an attention-based residual learning subnetwork (ARLnet), He et al. (He, Li, Liu, He, & Xiao, 2022) proposed a multi-frame VSR network (MVSRnet) for Jilin-1 and OVS-1 video satellite images SR reconstruction.

A DL network for Satellite VSR (Y. Xiao, Su, Yuan, et al., 2021) was proposed based on Multiscale Deformable Convolution Alignment and Temporal Grouping Projection.

Furthermore, to boost the research of the multi-frame Remote Sensing image SR, more recently, a PROBA-V super-resolution challenge was issued by the European Space Agency (ESA) to real-world super-resolved multi-temporal PROBA-V satellite imagery. In this context, a Deep neural network for SR of Unregistered Multitemporal images (DeepSUM) which can exploit both spatial and temporal correlations was proposed (Bordone Molini, Valsesia, Fracastoro, & Magli, 2020). HighRes-net which was to super-resolve PROBA-V satellite imagery was presented by Deudon et al. (Deudon et al., 2020). An end-to-end mechanism including co-registration, fusion, upsampling, and registration-at-the-loss was introduced to HighRes-net. A real end-to-end Residual Attention network for Multi-image SR (RAMS) (Salveti, Mazzia, Khaliq, & Chiaberge, 2020) was proposed for RS data. RAMS consists of a Residual Temporal Attention Network by applying 2D convolution, a Residual Feature Attention Network by applying 3D convolution, and a Temporal Reduction block. RAMS can exploit both spatial and temporal relationships to merge multiple LR inputs effectively. Most recently, a 3D Wide Activation Super Resolution (3DWDSR) (Dorr, 2020) contains a 3D CNN residual unit with WDSR blocks and a frame permutation technique as the data augmentation was presented. These methods provide good performance and satisfactory super-resolved results. Nevertheless, the aforementioned research adopts a direct fusion way while neglecting the temporal information, and focusing on the PROBA-V dataset whose resolution is low.

2.3 Attention Mechanisms

There is no doubt that CNN-based algorithms exhibit superior performance compared to other traditional methods in various tasks. Nevertheless, the bountiful high-frequency information involved in the LR images cannot be fully utilized since the equal treatment across the channels by the CNN kernels. The attention mechanism successfully addresses this issue due to its ability to trench deeper correlations across channels.

The attention mechanism was first proposed and applied to the machine translation assignment (Vaswani, 2017) and its remarkable performance on WMT 2014 English to-German translation task brought it extensive attention and research in academia. Briefly, Scaled Dot-Product Attention is written as:

$$Attention(Q, K, V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

where Q , K , and V denote vectors of query, keys of dimension d_k , and values of dimension d_v . The attention is calculated by computing the dot products of the query with all keys, dividing each by $\sqrt{d_k}$, and carry out a softmax function to produce the weights on the values.

Recently, the attention mechanism has exhibited more and more successful applications in various tasks. In this research, we integrate the attention mechanism to the CNN-based network and GAN-based network and explore the effects and effectiveness of the attention mechanism on the CNN and GAN networks. Therefore, in this subsection, several common attention modules are illustrated.

2.3.1 Self-attention

Here, we will explain the self-attention mechanism as illustrated in Figure 2.1 used in the Self-attention Generative Adversarial Networks (SAGAN) (Han Zhang et al., 2019).

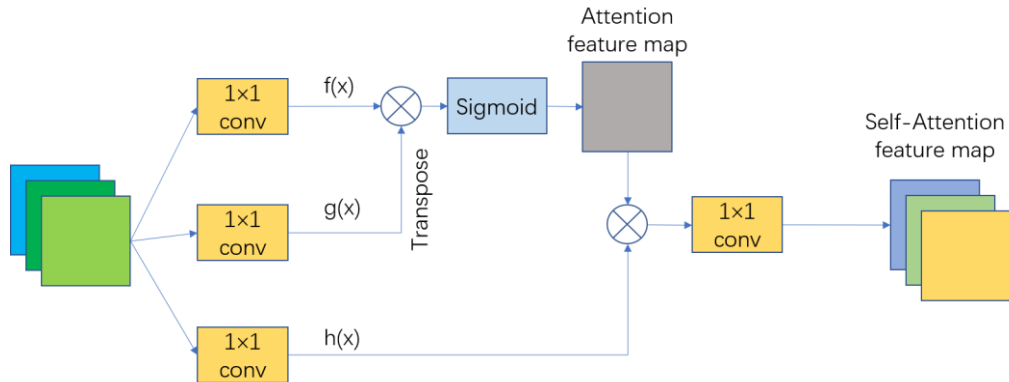


Figure 2.1 : Self-attention architecture.

Let x be the input image feature maps. First, the input X is branched into three copies by using three 1×1 convolutional kernels.

$$f(x) = W_f(x) \quad (2.2)$$

$$g(x) = W_g(x) \quad (2.3)$$

$$h(x) = W_h(x) \quad (2.4)$$

Then, we output the self-attention feature maps by applying the dot-product attention.

Thus, we have

$$\alpha_{i,j} = \text{softmax}\left(g(x_i)^T f(x_j)\right) \quad (2.5)$$

$$o_j = W_v\left(\sum_{i=1}^N \alpha_{i,j} h(x_i)\right) \quad (2.6)$$

At last, the extracted attention maps are multiplied by a scale parameter and added to the original input feature map:

$$y = x_i + \gamma o_i \quad (2.7)$$

The scaling factor γ is initialized to 0 and gradually increased during the training.

2.3.2 Channel attention, spatial attention, and pixel attention

Figure 2.2 demonstrates Channel Attention Module (CA) as presented in (Y. Zhang et al., 2018). The CA is composed of one average pooling and two 1×1 convolutional layers. Then, after the sigmoid function, the output of the activation function does an element-wise product with the original $H \times W \times C$ input. The Spatial Attention Block (SA) (Woo et al., 2018) is depicted in Figure 2.3. SA concatenates the Max pooling and Average pooling along the channel axis and then sends them to a convolution layer. The final attention map is computed in a similar way to CA. Figure 2.4 illustrates the Pixel attention (PA) (Zhao et al., 2020) module. Pixel attention aims to learn a 3D ($C \times H \times W$) feature map. PA obtains attention coefficients by considering all feature map pixels. PA adopts a 1×1 convolution layer and a sigmoid function to generate the attention maps and then multiplies the generated attention maps with the input features.

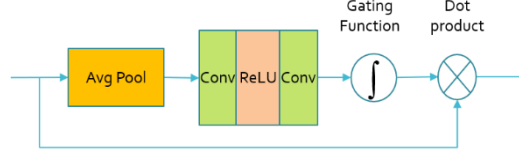


Figure 2.2 : Channel attention module.

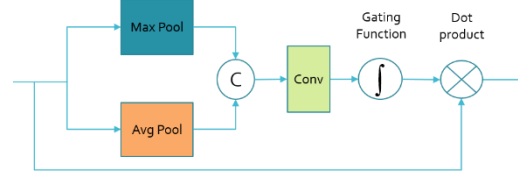


Figure 2.3 : Spatial attention module.

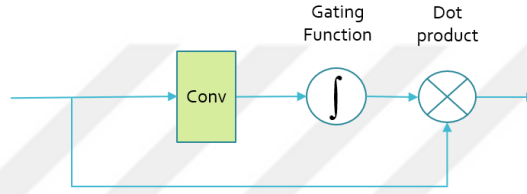


Figure 2.4 : Pixel attention module.

2.3.3 Non-local attention

The aforementioned attention modules mainly process on the local neighborhood at once. Non-local (NL) operation (Xiaolong Wang, Girshick, Gupta1, & He, 2018) is expanded from the self-attention module and designed to capture the long-range correlations. The NL operation is depicted in Figure 2.5.

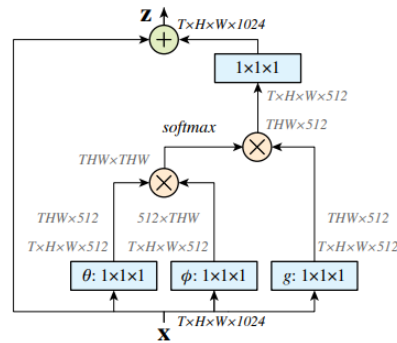


Figure 2.5 : Non-local attention.

Given the tensor X with the shape $T \times H \times W \times 1024$, which means the tensor has 1024 channels. $\otimes$ and $\oplus$ present matrix multiplication and element-wise addition, respectively. The blue boxes illustrate the $1 \times 1 \times 1$ convolutional operation. The

softmax function is applied on each row. The channels of the intermediate layers are set to 512. At last, the attention maps are reshaped to $T \times H \times W \times 1024$ and added to the original tensor X .

Moreover, in (Mei, Fan, & Zhou, 2021), Non-local sparse attention (NLSA) was proposed for the image SR. The design of NLSA can benefit from the long-range modeling capability of the NL process and the robustness and high efficiency of sparse representation.

2.3.4 Mixed high order attention

CA (Hu et al., 2020) learns a 1-Dimension ($C \times 1 \times 1$) vector attention map and can be regarded as first-order attention. Mixed high order attention (D. Zhang et al., 2020)(B. Chen, Deng, & Hu, 2019)(Du, Jiang, Si, Xu, & Liu, 2021) is depicted in Figure 2.6.

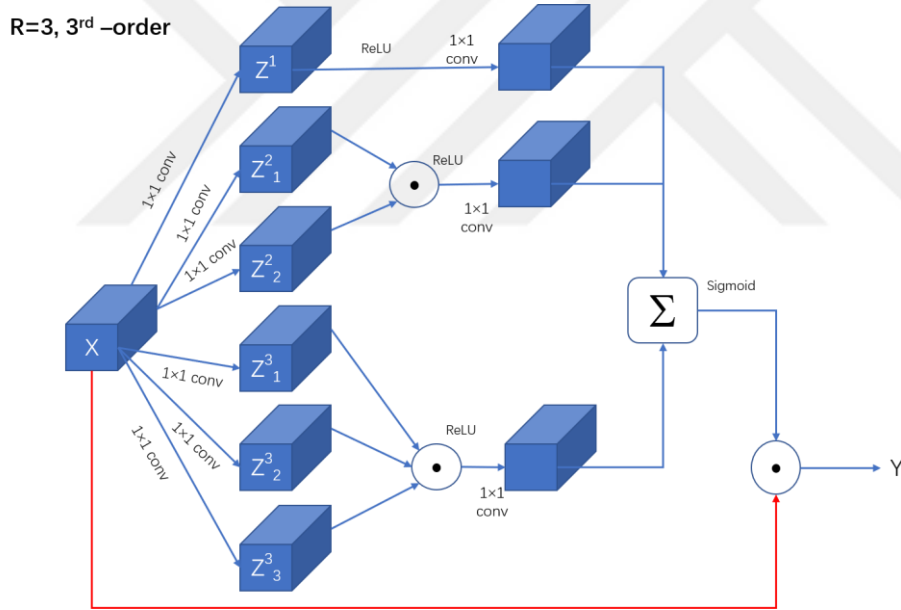


Figure 2.6 : Mixed high order attention.

Let $X \in R^{C \times H \times W}$ be the input image feature maps, the channel and spatial attention can be simply formulated as:

$$y = A(X) \otimes X \quad (2.8)$$

where $\otimes$ represents the Hadamard product. $A(X)$ represents the attention map. If

$$A(X) = \text{rep}[V]|^{H,W} \quad (2.9)$$

where $V \in R^C$ is a scale vector and $rep[V]|^{H,W}$ means replicate this scale vector along with height and weight dimensions by H and W times, respectively. Thus, Equation 1 will be the channel attention. If

$$A(X) = rep[M]|^C \quad (2.10)$$

where $V \in R^{H,W}$ is a spatial mask and $rep[M]|^C$ means replicate this spatial mask along channel dimension by C times. Thus, Equation 1 will be the spatial attention.

To model the high-order feature representation, we can define $A(X)$ as:

$$A(X) = sigmoid(\sum_{r=1}^R \hat{a}^{rT} \sigma(z^r)) \quad (2.11)$$

where σ depicts the non-linear activation function and $\hat{a}^r$ represents the weight vector. R denotes the number of the order. Thus, Equation 1 will be the Mixed high-order attention. The mixed high-order attention with $R = 3$ is depicted in Figure 1 which is easily implemented by applying 1×1 convolutional layers.

2.4 Loss functions

Appropriate selection of the loss function is one of the most critical learning strategies in DL applications. In the image SR task, the loss function aims to estimate differences between super-resolved results I_{SR} and the reference HR images I_{HR} , and optimize the SR models. L2 loss function (also known as MSE loss function) is widely used by researchers at its early stages. Later, it is found that the L2 loss function is not able to precisely measure the reconstruction quality. Therefore, researchers have proposed many loss functions (including Chabonnier Loss, Perceptual Loss, Adversarial Loss, etc.) to generate the reconstruction results more accurately. Now, we will overview the loss functions utilized in SR tasks.

2.4.1 Pixel-wise Loss

L1 (MAE) and L2 (MSE) loss functions are the simplest and most widely used pixel-wise Loss functions. Given super-resolved result I_{SR} and its target image I_{HR} , L1 loss function can be defined as:

$$\mathcal{L}_1(I_{SR}, I_{HR}) = \sqrt{(I_{HR} - I_{SR})^2} \quad (2.12)$$

L2 loss function can be formulated as:

$$\mathcal{L}_2(I_{SR}, I_{HR}) = (I_{HR} - I_{SR})^2 \quad (2.13)$$

Nowadays, the L1 loss function is more popular than the L2 loss function, because the L1 loss function has been proved to show improved performance and faster convergence. Moreover, as the differentiable alternative to the L1 loss function, the Chabonnier Loss function, which is robust and can be written as:

$$\mathcal{L}_{Cha}(I_{SR}, I_{HR}) = \sqrt{(I_{HR} - I_{SR})^2 + \varepsilon^2} \quad (2.14)$$

where ε is empirically set to a small constant 10^{-3} .

Due to the less sensitiveness to the outlier in the data, Huber Loss is also used for SR tasks and can be defined as:

$$\mathcal{L}_\delta(I_{SR}, I_{HR}) = \begin{cases} \frac{1}{2}(I_{HR} - I_{SR})^2 & \text{for } |I_{HR} - I_{SR}| \leq \delta \\ \delta(|I_{HR} - I_{SR}| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \quad (2.15)$$

Since the PSNR is exceeding associated with pixel-wise differences, the pixel-wise Loss function is rather effective especially when the target is to further improve the PSNR value. However, experimental results showed that the images with high PSNR values are always providing perceptually unsatisfactory results with overly smooth textures.

2.4.2 Perceptual Loss

To overcome the shortcomings of the pixel-wise loss function and produce perceptually satisfying super-resolved results, perceptual loss (Goodfellow & Pouget-Abadie, 2014)(Johnson, Alahi, & Fei-Fei, 2016) which aims to measure the distance between the high-level features generated from the pre-trained network Ψ was proposed. The perceptual loss can be defined as:

$$\mathcal{L}_{perceptual}(I_{SR}, I_{HR}, \Psi, l) = \|\Psi^l(I_{HR}) - \Psi^l(I_{SR})\|_2 \quad (2.16)$$

where l indicates the l -th layer of the pre-trained network Ψ .

The models trained with the perceptual loss usually provide visually finer details and textures but with lower PSNR compared to the pixel-wise loss-driven models.

Another perceptual loss function is LPIPS (R. Zhang, Isola, Efros, Shechtman, & Wang, 2018) which is trained with a dataset based on human perceptual similarity judgments and is more in conformity with the human perception preferences than the VGG perceptual loss. The LPIPS loss is defined as:

$$\mathcal{L}_{LPIPS}(I_{SR}, I_{HR}) = \sum_k \tau^k (\phi^k(I_{SR}) - \phi^k(I_{HR})) \quad (2.17)$$

where ϕ represents a feature extractor, τ transforms deep embedding to scalar LPIPS score, and the score is computed and averaged from k layers.

2.4.3 Adversarial Loss

Adversarial Loss is proposed and widely adopted in the training of the GANs (Goodfellow & Pouget-Abadie, 2014). Generally speaking, the GAN network is made up of generators and discriminators. For GAN-based RSISR, the SR model acts as the generator to generate Super-resolved results with the LR RS images as the input, and a discriminator acts as a classifier that determines whether the given image is generated or real.

Specifically, in the training process, the generator and discriminator have trained alternatively with cross entropy-based adversarial loss described as follows:

$$\mathcal{L}_G(I_{SR}) = -\log(\mathcal{D}(I_{SR})) \quad (2.18)$$

$$\mathcal{L}_D(I_{SR}, I_{HR}) = -\log(\mathcal{D}(I_{HR})) - \log(1 - \mathcal{D}(I_{SR})) \quad (2.19)$$

where $\mathcal{D}$ is the Discriminator network, $\mathcal{L}_G$ and $\mathcal{L}_D$ represents the loss of generator and discriminator, respectively. Although GAN-based networks can generate more realistic results than CNN-based methods, it is difficult to train. Therefore, Wasserstein GAN [50], WGAN-GP [51], and Relativistic Adversarial Loss are proposed to ease up the training difficulties and to improve the image quality.

More recently, Relativistic GAN (RaGAN) (Jolicoeur-Martineau, 2018) in which the discriminator assesses the probability that the given real data is more realistic than a randomly sampled generated data was proposed. The discriminator loss of the relativistic GANs is defined as:

$$\mathcal{L}_D^{Ra} = -E_{x_r} [\log(\mathcal{D}_{Ra}(x_r, x_f))] - E_{x_f} [\log(1 - \mathcal{D}_{Ra}(x_f, x_r))] \quad (2.20)$$

The adversarial loss for the generator is described as:

$$\mathcal{L}_G^{Ra} = -E_{x_r} [\log(1 - \mathcal{D}_{Ra}(x_r, x_f))] - E_{x_f} [\log(\mathcal{D}_{Ra}(x_f, x_r))] \quad (2.21)$$

where the x_r and x_f denotes the real data and generated fake data, respectively. $\mathcal{D}$ is the discriminator network. Experimental results have shown that the models with RaGAN loss provide finer details while with more stable training. However, the adversarial loss-based approaches tend to produce pleasant visual quality results but with worse PSNR and SSIM.

2.4.4 Cycle Consistency Loss

CycleGAN (J. Zhu, Park, Efros, Ai, & Berkeley, 2017) is a useful approach for image-to-image translation tasks. In (Pengrui Wang et al., 2019), the authors applied it to RSISR tasks. For cycle-CNN-based approaches, the Cycle Consistency Loss is composed of forwarding and backward Cycle Consistency Losses. In the forward part, an up-sampling model super-resolves the I_{LR} to produce the I_{SR} , and a degradation model downsamples the I_{SR} back to $I_{LR'}$. In turn, in the backward part, the degradation model down-samples the HR image to obtain I_{HLR} , and then the up-sampling model generates $I_{HR'}$ from I_{HLR} . The cycle consistent loss is proposed to keep the consistency between the I_{LR} , I_{HR} and their generated results $I_{LR'}$, $I_{HR'}$.

$$\mathcal{L}_{Cycle}(I_{LR}, I_{LR'}, I_{HR}, I_{HR'}) = \|I_{LR} - I_{LR'}\|_2 + \|I_{HR} - I_{HR'}\|_2 \quad (2.22)$$

3. DATASETS AND METHODOLOGY

3.1 Datasets

In this subsection, the information of the utilized datasets for single-frame and multi-frame RSISR was provided.

3.1.1 Dataset for single RSISR

For single RSISR, two datasets- the IEEE Data fusion contest (DFC 2019) dataset ('Data Fusion Contest', 2019), and the Pleiades dataset are utilized for the experimental assessment.

DFC 2019 is an open dataset provided by the IEEE data port and contains 2783 images for training and 50 images for testing with 1024×1024 pixels for each image acquired by the Worldview-3 satellite. The Worldview-3 satellite provides the panchromatic and 8-band multispectral image with a spatial resolution of 0.31m and 1.24m, respectively. The dataset covers buildings, elevated roads, bridges, high vegetation, ground, water, etc.

Moreover, VHR satellite images provided by the Pleiades satellite were also utilized to evaluate the proposed method. Pleiades satellite provides the panchromatic and 4-band multispectral image with product spatial resolution of 0.5m and 2.0m, respectively. The Pleiades dataset covers agriculture, residence, water, forest, and so on. Five full-size images were utilized for training and 30 images with 720×720 pixels cropped from another image were utilized for testing.

For the DFC dataset, the training and test patches are 44083 and 445 with a size of 256×256 pixels, respectively. For the Pleiades dataset, the training and test patches are 890 and 50 with a size of 720×720 pixels.

3.1.2 Dataset for MFSR remote sensing images

The SpaceNet7 dataset ('SpaceNet7', 2020) which contains ~ 100 locations, spread out across the globe was chosen for the MFSR. The dataset contains 60 groups of Planet I imageries which are of 4-meter resolution as training examples and 20 groups of images as validation examples. Each of the data cubes contains 19 to 25 (temporal dimensions) images with a size of around 1024×1024 pixels. In the training process, the inputs are clipped as $32 \times 32 \times 3 \times 5$, and the reference images are $128 \times 128 \times 3$. We use 13500 patches for the training, and 10 images with a size of 1024×1024 for the testing. Moreover, the Jilin-1 video satellite images are also used for the testing of the proposed MFSR as the real-world MFSR evaluation. First, 4 Jilin-1 videos were downloaded from ('Jilin-1 Smart Video Satellites', 2017). Then 9 frames from every 20 frames were selected and cropped to 1024×1024 pixels images. As a result, with a scale factor of 4, we generate the SR images with 4096×4096 pixels.

3.2 Attention CNN-based SFSR On Remote Sensing Images

In this subsection, the Attention CNN-based SFSR on RS images is illustrated.

3.2.1 Architecture of the Attention CNN-based SFSR on RS Images

3.2.1.1 The overall architecture

Figure 3.1 depicts the overall architecture of our proposed HANRS network. The proposed network consists of the shallow feature extraction part, deep feature extraction, layer attention module (LAM, illustrated in Figure 3.2), spatial attention module (SAM, illustrated in Figure 3.3), and reconstruction part.

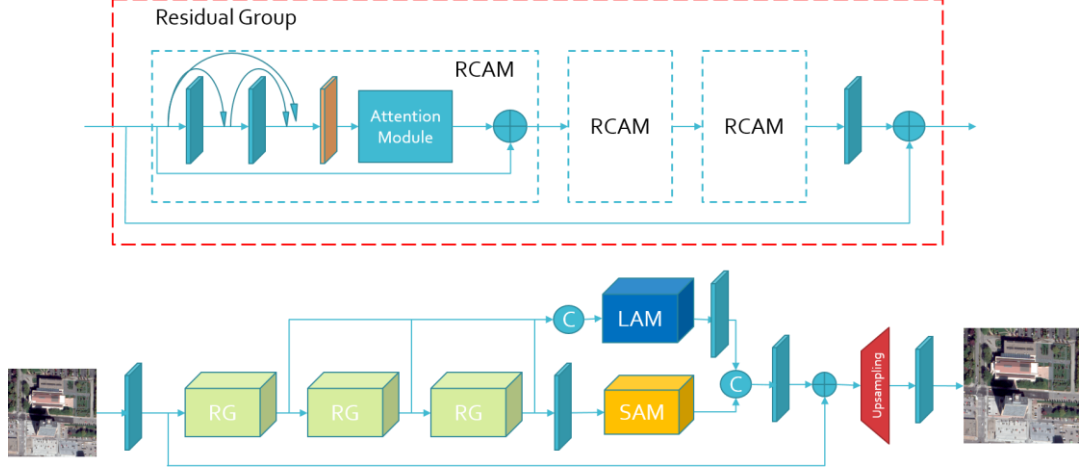


Figure 3.1 : The overall architecture of the proposed HANRS.

Shallow feature extraction. Given LR input I_{LR} , the first convolutional layer of the proposed network extracts a set of shallow feature maps I_0 .

$$I_0 = f_0(I_{LR}) \quad (3.1)$$

where f_0 denotes the convolutional operator of the shallow feature extraction.

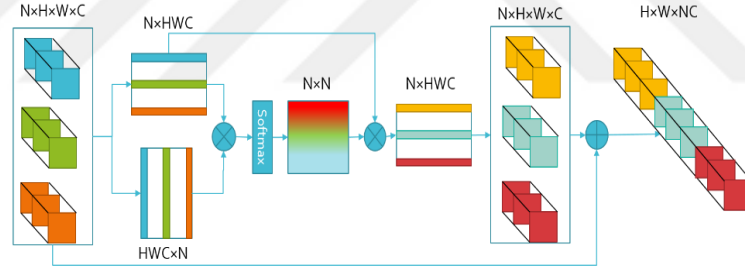


Figure 3.2 : The architecture of the multi-layer Attention module (LAM).

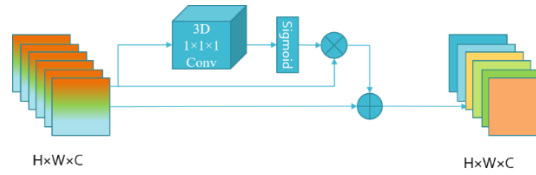


Figure 3.3 : The architecture of the spatial Attention module (SAM).

Deep feature extraction. After the shallow feature extraction, a series of residual groups (RGs) further extract deeper feature representations of the low-resolution input. In addition to the RGs, the deep feature extraction part includes a layer attention module (LAM) and a spatial attention module (SAM). LAM, as depicted in Figure 3.2,

learns the correlations of each output from RGs and the spatial attention module (SAM) as depicted in Figure 3.3 learns the interdependencies between channels and pixels.

Assume we have N residual groups, the intermediate features I_i is calculated as:

$$I_i = f_{RG_i}(I_{i-1}), \quad i = 1, 2, \dots, N \quad (3.2)$$

where I_{i-1} and I_i denote the $i - 1$ -th and i -th residual group, respectively. f_{RG_i} denotes the weights of the i -th residual group.

The output of the LAM I_{LAM} is calculated by weighing the concatenation of all the intermediate feature maps and can be written as:

$$I_{LAM} = f_{LAM}(\text{concat}(I_1, I_2, \dots, I_N)) \quad (3.3)$$

where concat depicts the concatenation operation, f_{LAM} presents the weights of the proposed layer attention module.

The spatial attention module takes the last feature layer of the Residual Group I_N . The output of the spatial attention module is:

$$I_{SA} = f_{SA}(I_N) \quad (3.4)$$

where f_{SA} denotes the weights of the proposed spatial attention module.

Then, after a convolutional layer, the outputs of the LAM and SAM are concatenated together, then added to the shallow feature maps I_0 . Thus, we have

$$I_{out} = \text{concat}(f_{conv}(I_{LAM}), I_{SA}) + I_0 \quad (3.5)$$

where f_{conv} denotes the weights of the convolutional layer.

Reconstruction module. Finally, the reconstruction part consists of an up-sampling block and a convolutional layer. We applied a post-upsampling architecture in our proposed method. Following (Shi et al., 2016), we applied pixel shuffle as the upsampling part. Then, a convolutional layer is utilized for the reconstruction part to generate the super-resolved image.

$$I_{up} = f_{PS}(I_{out}) \quad (3.6)$$

$$I_{SR} = f_{rec}(I_{up}) \quad (3.7)$$

where I_{up} and I_{SR} denote the output of the up-sampling module and the final super-resolved image, respectively. f_{PS} and f_{rec} represent the process of the up-sampling module and the reconstruction convolutional layer, respectively.

3.2.1.2 Layer Attention Module and Spatial Attention module

The layer attention module (LAM) aims to further learn the feature representation among the multilayers. As shown in Figure 2, the layer attention module takes the intermediate feature maps $I_{FG} \in R^{N \times H \times W \times C}$ from the N Residual Groups as the input. Then, the feature maps I_{FG} is reshaped to a 2D matrix with the dimension of $N \times HWC$, and the attention map between different layers is calculated by applying a matrix multiplication with its corresponding transpose $HWC \times N$. Thus, we obtain a $N \times N$ attention map. We can write the correlation

$$w_{j,i} = \delta(\varphi(I_{FG})_i \cdot \varphi(I_{FG})_j^T), \quad i, j = 1, 2, \dots, N \quad (3.8)$$

where δ and φ represent the softmax activation function and reshape operation, respectively. $w_{i,j}$ denotes the correlation between the i -th and j -th feature groups. At last, we multiply the reshaped feature maps $\varphi(I_{FG})$ by the generated $N \times N$ attention map with a scale factor of γ , and add the input feature maps I_{FG} ,

$$F_{L_j} = \gamma \sum_{i=1}^N w_{i,j} I_{FG_i} + I_{FG_j} \quad (3.9)$$

where γ is a trainable parameter and initialized to 0.

The proposed channel-spatial attention module (CSAM) is depicted in Figure 3. As shown in Figure 3, the proposed CSAM contains a 3D convolution operation and a skip connection. The CSAM aims to make full use of the scale dimension information and the channel dimension feature.

Given the feature maps of the last layer $I_N \in R^{H \times W \times C}$. We feed the I_N to a 3D convolution layer to produce the attention weights W_{CS} which can capture the joint channel and spatial features. In our experiment, we choose a 3D convolution layer with

a kernel size of $1 \times 1 \times 1$ and stride of 1. Then, we perform an element-wise multiplication with the activated attention map and the input feature map I_N . At last, the weighted maps are produced by multiplying the weighted results by a learnable scale factor of β , and adding the input feature I_N . The weighted feature can be described as:

$$I_{csa} = \beta \sigma(W_{csa}) \otimes I_N + I_N \quad (3.10)$$

where σ denotes the sigmoid function, $\otimes$ denotes the element-wise product, and β is a trainable parameter and initialized to 0.

3.2.1.3 Random downsampling

Despite that the scaling factor of satellite images between panchromatic and multispectral is usually 4, it restricts the generalization of the DL networks since it is not always the case. Therefore, we obtained the LR images from its HR images by applying random sampling to enhance the generalization ability of the networks. As in the experiments, our patch size of the image is 256×256 in high resolution. Instead of generating its corresponding LR image with 64×64 pixels, we randomly down-sample the image to a selected size of $s \times s$. Here, s is an integer from a uniform random distribution between (a, b) . Specifically, we set a and b to 20 and 80 in the experiment, respectively.

3.2.2 Model Training

Similar to (X. Zhu et al., 2020), we use a perceptual loss consisting of a fidelity loss, and a perceptual loss to optimize our model. The loss function is defined as:

$$L = f(I_{HR}, w(I_{SR})) + \alpha q(w(I_{SR})) \quad (3.11)$$

where I_{HR} and I_{SR} represent the ground truth image and the super-resolved image, respectively. w denotes the weights of the model. f represents the SmoothL1 loss, and q represents the perceptual loss function generated by LPIPS (R. Zhang et al., 2018). α is the scaling factor set to 0.001.

Moreover, to increase the performance of the PSNR, we also train our model by using the L1 loss function. The L1 loss function is defined as:

$$L = \sqrt{(I_{SR} - I_{HR})^2} \quad (3.12)$$

3.3 Attention GAN-based SFSR On Remote Sensing Images

3.3.1 Architecture of the attention GAN-based SFSR on RS images

In this subsection, the architecture of the attention GAN-based SFSR on RS images is described as demonstrated in Figures 3.4 to 3.7.

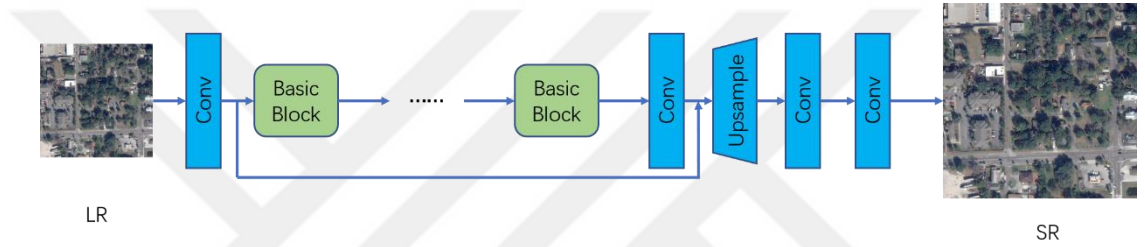


Figure 3.4 : The architecture of the generator in ESRGAN.

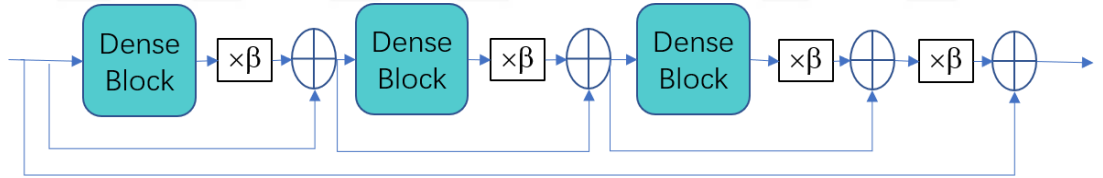


Figure 3.5 : Basic block in ESRGAN.

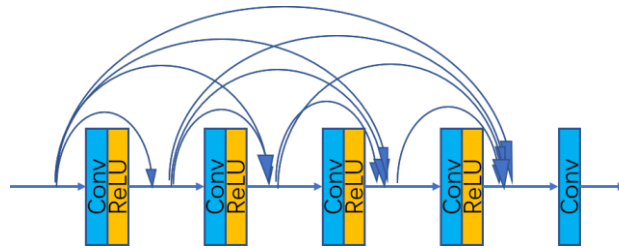


Figure 3.6 : Dense block proposed in ESRGAN.

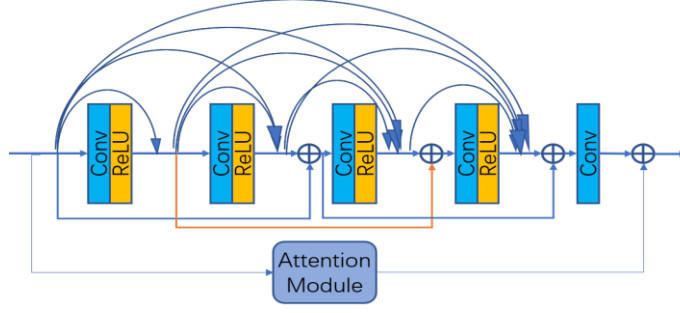


Figure 3.7 : Our proposed Dense block for RSISR.

Generator. The generator of the proposed Attention GAN-based for the Remote Sensing Image Super-resolution is strongly improved from the ESRGAN (Xintao Wang et al., 2018). Figure 3.4 illustrates the overall architecture of the proposed Attention GAN-based method. Given the input LR, first, a convolutional layer is conducted for the shallow feature extraction, which is written as:

$$I_0 = f_0(I_{LR}) \quad (3.13)$$

where I_{LR} denotes the input LR image, f_0 denotes the first convolutional operation and I_0 is the output of the shallow feature maps.

Then, the extracted feature maps I_0 is fed to a series of basic blocks for deep feature extraction. Assume we have N basic blocks, the intermediate feature maps I_i can be described as:

$$I_i = f_{RG_i}(I_{i-1}), \quad i = 1, 2, \dots, N \quad (3.14)$$

where f_{RG_i} depicts the weights of the i -th basic blocks, and I_i and I_{i-1} denote the feature maps of the i -th and $(i - 1)$ -th basic blocks, respectively.

At last, an upsampler and two more convolutional layers are applied for the reconstruction of the expected SR images. The reconstruction process can be written as:

$$I_{up} = f_{PS}(I_0 + I_N) \quad (3.15)$$

$$I_{SR} = f_{rec}(I_{up}) \quad (3.16)$$

where I_{up} and I_{SR} denote the output of the up-sampling module and the final super-resolved image, respectively. f_{PS} and f_{rec} represent the process of the up-sampling module and the reconstruction convolutional layer, respectively.

Similar to ESRGAN (Xintao Wang et al., 2018), a Residual in Residual module which is depicted in Figure 3.5 is applied in the basic block. The basic block is composed of three dense blocks and several skip connections with a scale factor β .

Figure 3.6 and 3.7 demonstrated the dense block applied in ESRGAN and proposed for the attention GAN-based SRSISR, respectively. The improvements of the proposed attention GAN-based mainly include two aspects: (1) we improve the architecture of the RRDB block by adding three more skip connections for further feature representation; (2) we add an attention module in the RRDB blocks in a parallel way (note that we also try to add the attention module just behind the RRDB block, however, it does not work and cannot generate the SR images).

Discriminator. The structure of the discriminator consists of ESRGAN (Xintao Wang et al., 2018) as illustrated in Figure 3.8. As the standard discriminator estimates the probability that the inputs are real or fake, while the relativistic discriminator evaluates the probability that the inputs are more realistic than fake ones. The same as ESRGAN, a relativistic discriminator is utilized in our experiments to optimize our proposed models. The discriminator loss is defined as:

$$L_D^{Ra} = -E_{x_r}[\log(D_{Ra}(x_r, x_f))] - E_{x_f}[\log(1 - D_{Ra}(x_f, x_r))] \quad (3.17)$$

And, the adversarial loss of the Generator is defined as:

$$L_G^{Ra} = -E_{x_r}[\log(1 - D_{Ra}(x_r, x_f))] - E_{x_f}[\log(D_{Ra}(x_f, x_r))] \quad (3.18)$$

where $x_f = G(x_i)$ and x_i is the input LR image.

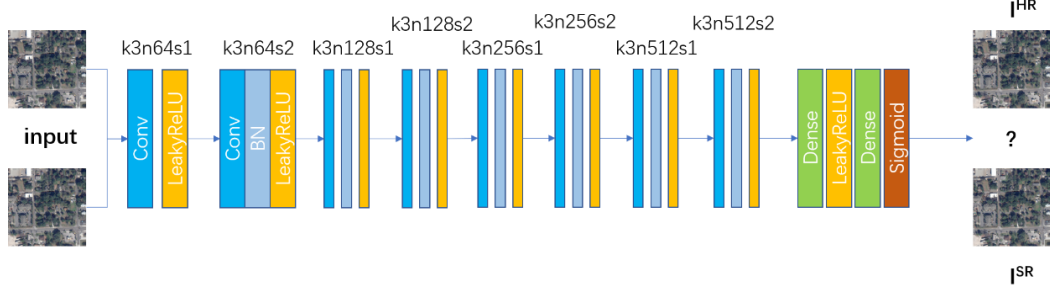


Figure 3.8 : Proposed MHOA Attention GAN-Discriminator.

3.3.2 Model Training

The training of the GAN is composed of two parts: (1) the training of the Generator and (2) the training of the Discriminator. The discriminator loss is defined in equation 3.17.

The generator loss is defined as:

$$L_G = L_{percep} + \lambda L_G^{Ra} + \eta L_{content} \quad (3.19)$$

where L_{percep} is the perceptual loss which aims to enhance the perceptual quality of generated outputs. The perceptual L_G^{Ra} is the adversarial loss for the generator which is defined in equation 3.18, and $L_{content}$ is the content loss for the training of the generator which is usually the L_1 loss function. λ and η are the coefficients to balance the different loss terms, which usually are very small content and are set empirically.

3.4 Multi-frame Remote Sensing Image Super-resolution

3.4.1 Architecture of the MFSR on RS Images

With the excellent achievements of the GAN-based methods in recent years, for the multi-frame RSISR, an attention GAN-based model was explored. The overview of the generator of the proposed MFSR method and the proposed attention module are illustrated in Figures 3.9 and 3.10, respectively.

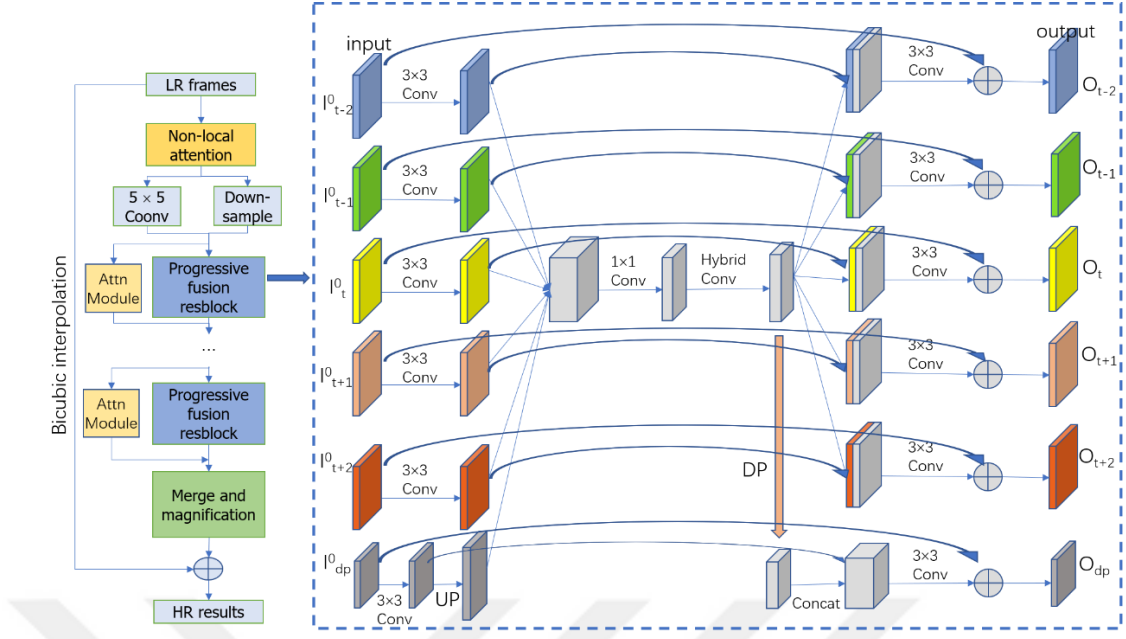


Figure 3.9 : Overview of the generator of the proposed MFSR method.

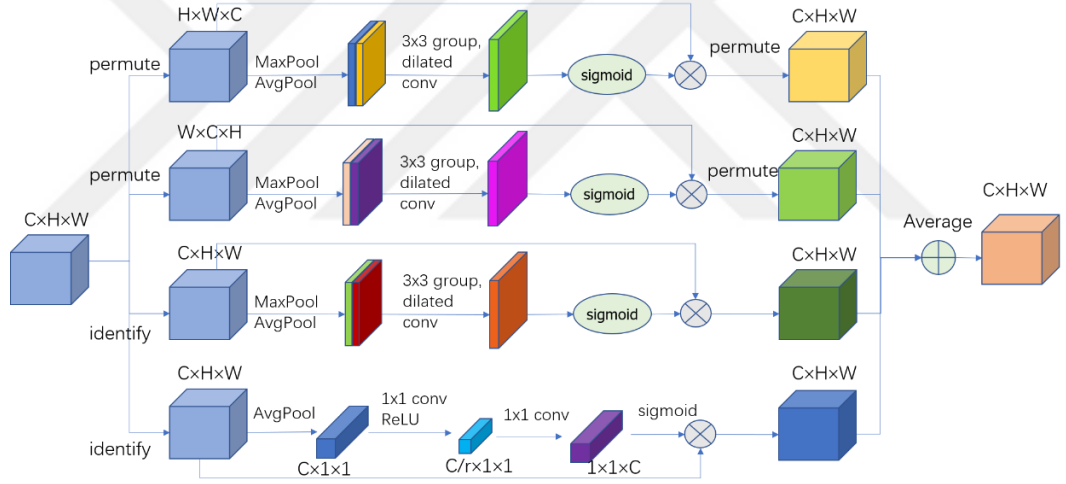


Figure 3.10 : Proposed attention to the MFSR method.

Generator. The design of the generator is strongly based on MSHPFNL(Yi et al., 2020) which is proposed by Yi et al. The overview of the generator of the proposed MFSR method is depicted in Figure 3.9. The left part describes the pipeline of the proposed attention-based MFSR method. Similarly, a non-local operation was adopted for learning the input frames' consistency instead of using MEMC. Before the non-local attention module, the input LR frames are concatenated together. Then, after the non-local attention module, one 5×5 convolutional layers and a down-sampling module were applied for the basic feature extraction. After that, a series of attention

PFRBs were applied to fully extract the intraframe spatial relationships and inter-frame temporal relationships among LR frames. At last, the extracted information from all channels in attention PFRBs was merged and enlarged to generate residual HR components, by obtaining its addition to the bicubically magnified image of the LR frame, the final SR output is generated. Finally, the merge and magnification module applies a sub-pixel magnification layer as in (Shi et al., 2016).

Assume that taking the five frames $\{I_{t-2}, I_{t-1}, I_t, I_{t+1}, I_{t+2}\}$ as the input, after the non-local attention module and 5×5 convolutional layers, the input of the PFRBs contains five feature maps $\{I_{t-2}^0, I_{t-1}^0, I_t^0, I_{t+1}^0, I_{t+2}^0\}$ from five consecutive frames from the input and one feature map I_{dp}^0 which is down-sampled from the five concatenated frames after the NLRB. The factor of the down-sampling was set to 2. In addition, for each PFRB block, an attention module was also added to further learn the correlations among the channels.

As can be seen from Figure 3.9, a 3×3 convolutional layer was first conducted with the same filters on the input of the PFRBs, which can be written as:

$$I_t^1 = F_t^1(I_t^0) \quad (3.20)$$

$$I_{dp}^1 = F_{dp}^1(I_{dp}^0) \quad (3.21)$$

where t represents the index of the temporal dimension, F_t^1 represents the convolutional operation, and I_t^0 denotes the input frames. I_t^1 denotes the feature maps extracted from each input frame. I_{dp}^1 and I_{dp}^0 denote the feature maps at a smaller scale. F_{dp}^1 denotes the convolution conducted on I_{dp}^0 . Then, I_{dp}^1 is further upsampled to the original size I_{dp}^{up} :

$$I_{dp}^{up} = UP(I_{dp}^1) \quad (3.22)$$

where UP is an up-sampling function that consists of 1×1 convolution layer and a sub-pixel convolution layer.

The feature maps $I_{t-2}^1, I_{t-1}^1, I_t^1, I_{t+1}^1, I_{t+2}^1, I_{dp}^{up}$ are concatenated together into one part I^1 and the depth of the merged feature maps is $(5 + 1) \times N$ when the input frames are set to 5. Then one 1×1 convolution and a hybrid convolution that consists of a normal 3×3 convolutional layer with $N/2$ filters and a 3×3 dilated convolutional layer (dilation rate is 2) with $N/2$ filters are conducted, which can be described as:

$$I^2 = F^2(I^1) \quad (3.23)$$

$$I^3 = HC(I^2) \quad (3.24)$$

where F^2 denotes the 1×1 convolution, and I^2 presents the feature maps after the 1×1 convolutional layer. HC depicts the hybrid convolution, and I^3 depicts the feature maps after hybrid convolution.

Then, I^3 is downsampled to I_{dp}^3 which is written as:

$$I_{dp}^3 = DP(I^3) \quad (3.25)$$

where DP is the down-sampling operation which contains one 1×1 convolution and one reshape operation. The factor of the down-sampling operation is set to 2. Then, I^3 and I_{dp}^3 are concatenated to the previous feature maps I^1 and I_{dp}^1 , respectively, which become I^4 and I_{dp}^4 . Later, a 3×3 convolution layer is adopted for spatial and temporal information extraction:

$$O_t = F_t^4(I_t^4) + I_t^0 \quad (3.26)$$

$$O_{dp} = F_{dp}^4(I_{dp}^4) + I_{dp}^0 \quad (3.27)$$

where I_t^4 represents the merged feature maps of $\{I^3, I_t^1\}$, F_t^4 denotes the fourth depth of the convolutional layer, O_t denotes the outputs of the PFRBs module.

Moreover, we have added an attention module to each PFRB module to learn the complicated correlations among the channels, thus we have

$$O'_t = F_t^{attn}(I_t^0) + O_t \quad (3.28)$$

$$O'_{dp} = F_{dp}^{attn}(I_{dp}^0) + O_{dp} \quad (3.29)$$

where F_t^{attn} and F_{dp}^{attn} represent the attention module applied on input frames I_t^0 and I_{dp}^0 , respectively. O'_t and O'_{dp} are the corresponding outputs of the attention-based PFRBs.

Proposed attention module. In addition to the exploration of the effects of the Mixed High Order Attention (MHOA) and self-attention module on the PFRBs, a new attention module is also proposed to extract cross-channel correlations for discriminative feature representation enhancement. As illustrated in Figure 3.10, the proposed attention module contains four sections: two of them try model CA ($C \times 1 \times 1$) and SA (H, W). The other two try to extract the interplay between the channel space and spatial space ((W, C) or (C, H)). Given an input $X \in R^{C \times H \times W}$, in the CA section, the CA weights are utilized to extract the feature $\hat{X}^C$. In the SA section, we first obtain two 2-Dimension maps ($1 \times H \times W$) by applying average-pooling and max-pooling operations along the channel axis. Then, a 3×3 dilated convolution with dilation rate 3 and a group convolution with group 2 followed by a sigmoid function is applied to generate the attention maps ($1 \times H \times W$). At last, the SA weights are multiplied with the input tensor X to produce the recalibrated feature $\hat{X}^{H \times W}$. For the rest two sections, the input tensor X is first transposed and forwarded to the average-pooling and max-pooling modules to produce the corresponding attention maps ($1 \times W \times C$ or $1 \times C \times H$), respectively. Afterward, the obtained attention maps are carried out on the transposed maps and then rotated 90° along the width axis or the height axis to maintain the original input shape ($C \times H \times W$). At last, the extracted maps of four sections ($\hat{X}^C$, $\hat{X}^{H \times W}$, $\hat{X}^{H \times C}$, $\hat{X}^{C \times W}$) are accumulated by the averaging operation for the discriminative representation ability enhancement of the network, which is described as:

$$\hat{X} = \frac{(\hat{X}^C + \hat{X}^{H \times W} + \hat{X}^{H \times C} + \hat{X}^{C \times W})}{4} \quad (3.30)$$

where $\hat{X}$ is the output recalibrated feature.

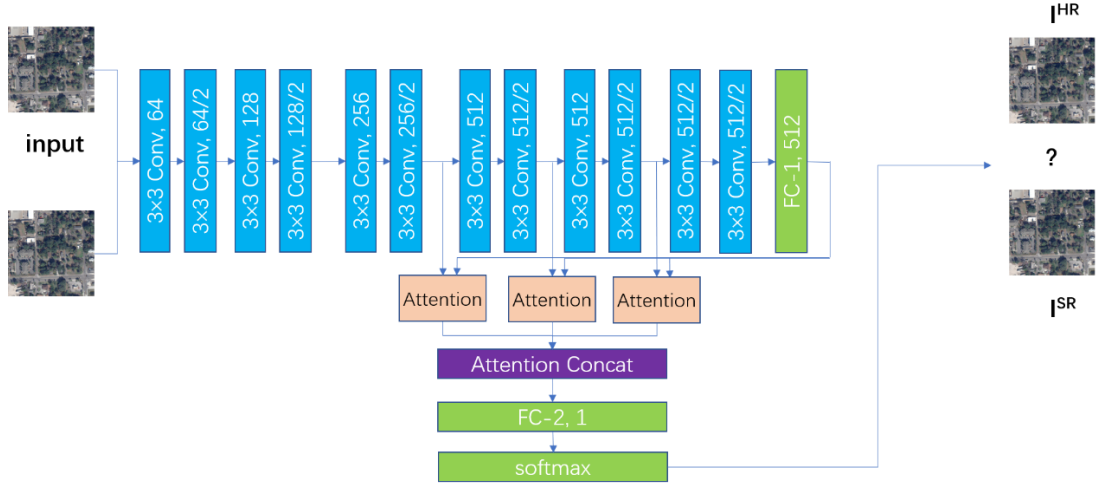


Figure 3.11 : Discriminator of the proposed MFSR.

Discriminator. Following (Jetley, Lord, Lee, & Torr, 2018), we designed an attention-based discriminator for our proposed MFSR as shown in Figure 3.11. Instead of a fully connected layer (FC-2), we add an attention concatenation layer by learning the attention maps from different levels of the basic VGG net.

3.4.2 Model Training

To optimize the proposed algorithm, first, we train the model with L1 loss, which is written as:

$$L_1 = \sqrt{(I_{SR} - I_{HR})^2} \quad (3.31)$$

Then, we train the model with the GAN strategy. The GAN loss includes Perceptual, adversarial, and content loss as defined in equation 3.19. Apart from the widely applied perceptual loss which utilizes the VGG framework, a different perceptual loss that benefits from LPIPS is also explored. And the results are discussed in Section 5.4.

3.5 Evaluation Criteria

(1) Peak-Signal-Noise-Ratio (PSNR): PSNR aims to assess the image quality by calculating the distance between images at the pixel level. For RSISR, assume that we have generated a super-resolved result I_{SR} and target image I_{HR} which are both with N pixels, we first calculate the mean squared error (MSE) between images as

$$MSE = \frac{1}{N} \sum_{i=1}^N (I_{HR}(i) - I_{SR}(i))^2 \quad (3.32)$$

$$PSNR = 10 \log_{10} \left(\frac{M^2}{MSE} \right) \quad (3.33)$$

where M is the maximum possible pixel value directly related to the radiometric resolution of the images and dynamic range. Higher PSNR values indicate better visual quality.

(2) Structural Similarity Index Measure (SSIM): SSIM measures the structural similarity between two images. It is a perception-based model and uses luminance, contrast, and structure (W. Zhou, Alan, Hamid, & Eero, 2004). Given the super-resolved image I_{SR} and target image I_{HR} , SSIM is calculated by:

$$SSIM = l(I_{SR}, I_{HR}) * C(I_{SR}, I_{HR}) * S(I_{SR}, I_{HR}) \quad (3.34)$$

where $l(I_{SR}, I_{HR})$, $C(I_{SR}, I_{HR})$, and $S(I_{SR}, I_{HR})$ represent the similarity of the luminance, contrast, and structure, respectively. Due to the possible uneven distribution of the images, there are limitations to using SSIM in practice. Thus, Mean Structural Similarity (MSSIM) and Multi-Scale Structural Similarity (MS-SSIM) are used to assess the SSIM by dividing the image into multiple windows, calculating the SSIM of each window, and finally obtaining the MSSIM or MS-SSIM. Higher SSIM values indicate more similarity to the HR image.

(3) Learned Perceptual Image Patch Similarity (LPIPS): LPIPS estimates the perceptual quality and is used to evaluate the distance between two image patches (R. Zhang et al., 2018). Smaller LPIPS values indicate more similarity; whereas higher LPIPS values indicate becoming different.

(4) Natural Image Quality Evaluator (NIQE): NIQE does not require a reference ground truth image. It fits two multivariate Gaussian models to images and calculates the distance between these two models (Mittal, Soundararajan, & Bovik, 2013). Quality-aware features deduced from the original scene statistic model are used while fitting the multivariate Gaussian models. Smaller NIQE values indicate better visual quality.

(5) Perceptual Index (PI): PI is proposed in the 2018 PIRM Challenge on Perceptual Image Super-resolution (Blau, Mechrez, Timofte, Michaeli, & Zelnik-Manor, 2018). The PI combines the no-reference image quality measures of Ma et al. (C. Ma, Yang, Yang, & Yang, 2017) and NIQE (Mittal et al., 2013) as:

$$PI = \frac{1}{2}((10 - Ma) + NIQE) \quad (3.35)$$

Smaller PI values indicate better visual quality.

(6) Average Gradient (AG): AG measures the gradient magnitude of the image (A. A. Chen, Chai, Chen, Bian, & Chen, 2018) and is defined as:

$$AG = \frac{1}{(h-1)(w-1)} \sum_x \sum_y \frac{|G(x,y)|}{\sqrt{2}} \quad (3.36)$$

In addition, there are some other assessment metrics used for the quantitative evaluation of the RSISR results, such as SAM (Spectral Angle Mapper) (Yuhas, Goetz, & Boardman, 1992), Q-index (W. Zhou & Bovik, 2002), ERGAS (Erreur Relative Globale Adimensionnelle de Synthèse) (J. G. Liu, Wald, & Ranchin, 2000), QNR (Quality with No Reference) (Alparone et al., 2008), VIF (R. & Bovik, 2006), FSIM (L. Zhang, Zhang, Mou, & Zhang, 2011). Here, we do not introduce them in detail.

In this research, we use three reference-based metrics - PSNR, SSIM, LPIPS, and three No-reference-based metrics - NIQE, PI, and AG as the evaluation metrics to assess the quality of our SRSISR results. All these metrics are estimated with RGB images.



4. RESULTS AND DISCUSSIONS OF SINGLE-FRAME REMOTE SENSING IMAGE SUPER-RESOLUTION

In this section, the quantitative and qualitative results and discussions of the single-frameRSISR are demonstrated. First, the results of the attention CNN-based SFSR on RS images are illustrated. Then, the results of the attention GAN-based SFSR on remote sensing images are described.

4.1 Results of the Attention-based SFSR on Remote Sensing Images

4.1.1 Experiment settings

The experiments are conducted in PyTorch 1.0 ('pytorch', n.d.) with NVIDIA GeForce 1060 GPU. All models are optimized by Adam optimizer with $\beta_1=0.9$, $\beta_2=0.999$ with RGB images.

The ratio between LR and HR images of $\times 2$, $\times 4$, and $\times 8$ are explored. The LR inputs are degraded with bicubic interpolation from the corresponding HR images.

In the training part, 10 RGs and 15 RCAMs are applied. 64 (3×3) filters are adopted and Leaky-ReLU with $\alpha = 0.2$ is utilized except the conv layers in CAB. For the DFC dataset, the training and test patches are 44083 and 445 with a size of 256×256 pixels, respectively. For the Pleiades dataset, the training and test patches are 890 and 50 with a size of 720×720 pixels.

In the training part, the LR inputs are clipped to 64×64 pixels. The learning rate is 2×10^{-4} and decayed 0.5 every 3000 iterations.

4.1.2 Results on downsampled images

4.1.2.1 Quantitative and qualitative results of DFC (2019) Dataset

The quantitative results of the attention CNN-based SFSR methods on the DFC dataset with scaling factors of $\times 2$, $\times 4$, and $\times 8$ are presented in Table 4.1, Table 4.2, and Table 4.3, respectively. To better interpret the quantitative results, the top three best values of each metric were ranked and represented with three different colors-Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Table 4.1 : Quantitative values of DFC dataset with a scaling factor of $\times 2$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	AG $\uparrow$	NIQE $\downarrow$	LPIPS $\downarrow$
Bic	35.6420	0.9939	5.7035	16.8375	0.1502
LGCNet	38.2751	0.9978	6.6383	15.9451	0.0936
RDN	39.4546	0.9982	6.8032	15.1104	0.0793
SAN	39.0940	0.9981	6.7800	15.4520	0.0822
MHAN	38.3671	0.9980	6.7180	15.5853	0.0823
CAFRN	38.7754	0.9980	6.7055	15.7495	0.0864
RCAN	39.3158	0.9982	6.7833	15.3407	0.0791
HAN	39.0963	0.9982	6.7636	15.0983	0.0790
HANRS_L1	39.5721	0.9982	6.8026	15.0114	0.0778
HPANRS_L1	39.3668	0.9982	6.7981	15.1516	0.0785
HANRS_per	39.1004	0.9979	6.9434	14.2242	0.0542
HPANRS_per	39.0658	0.9979	6.9104	14.0881	0.0551

To estimate the effectiveness of the proposed attention CNN-based SFSR algorithms, several state-of-the-art approaches, including the bicubic interpolation method, LGCNet, RDN, SAN, MHAN, CAFRN, RCAN, and HAN methods are compared with the proposed method. HANRS_L1 and HANRS_per represent the Holistic attention (HAN) method trained with L1 loss and perceptual loss which is illustrated in equation 3.11 on the DFC dataset, respectively. HPANRS_L1 and HPANRS_per represent the HAN model with pixel attention trained with L1 loss function and perceptual loss function, respectively.

As observed in Table 4.1, the HANRS_L1 surpasses other algorithms concerning PSNR and SSIM. RDN and HPANRS_L1 are ranked as the 2nd and 3rd in terms of PSNR, but for SSIM, HPANRS_L1 is ranked as the 2nd and RDN is ranked as the 3rd. Concerning both AG and LPIPS, HANRS_per and HPANRS_per are ranked 1st and 2nd, respectively. RDN is ranked as the 3rd regarding AG and HANRS_L1 is ranked as the 3rd regarding LPIPS. NIQE metric value shows HPANRS_per, HANRS_per, and HANRS_L1 are ranked as the 1st, 2nd, and 3rd, respectively.

Table 4.2 : Quantitative values of DFC dataset with a scaling factor of $\times 4$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	AG $\uparrow$	NIQE $\downarrow$	LPIPS $\downarrow$
Bic	32.0850	0.9441	4.0430	18.2822	0.3517
LGCNet	33.0725	0.9677	4.9836	17.6345	0.3019
RDN	33.8564	0.9769	5.5065	15.8714	0.2471
SAN	33.2988	0.9754	5.3347	16.6242	0.2582
MHAN	33.7318	0.9761	5.3746	16.7748	0.2554
CAFRN	33.5843	0.9742	5.3073	16.6709	0.2656
RCAN	33.9150	0.9774	5.5393	16.1878	0.2452
HAN	33.9408	0.9778	5.5655	15.9538	0.2407
HANRS_L1	34.0223	0.9781	5.5576	15.8333	0.2379
HPANRS_L1	34.0388	0.9783	5.6006	15.7236	0.2381
HANRS_per	33.8033	0.9770	5.9738	15.9424	0.1772
HPANRS_per	33.7406	0.9768	5.9162	14.7277	0.1826
HANRS_RD	33.9539	0.9779	5.6228	15.9715	0.2364
HPANRS_RD	33.8739	0.9777	5.5986	16.3296	0.2413

Table 4.2 demonstrates the quantitative values of the DFC dataset with a scaling factor of $\times 4$. HANRS_RD and HPANRS_RD represent the models with random sampling. As shown in Table 4.2, HPANRS_L1 performs the best concerning PSNR and SSIM. HANRS_L1 and HANRS_RD rank second and third concerning PSNR and SSIM, respectively. Regarding AG and LPIPS, HANRS_per, HPANRS_per, and HANRS_RD rank as first, second, and third, respectively. Finally, HPANRS_per, HPANRS_L1, and HANRS_L1 as the first, second, and third, respectively in terms of NIQE. Therefore, generally speaking, the models with the perceptual loss provides better perceptual quality compared to the L1 loss function.

Table 4.3 : Quantitative values of DFC dataset with a scaling factor of $\times 8$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	AG $\uparrow$	NIQE $\downarrow$	LPIPS $\downarrow$
Bic	30.3893	0.8159	2.7232	16.8577	0.5391
LGCNet	30.8539	0.8586	3.3590	15.7403	0.4929
RDN	31.2332	0.8868	3.8078	15.5104	0.4392
SAN	31.1745	0.8842	3.6516	15.8759	0.4516
MHAN	31.1701	0.8853	3.7021	15.9469	0.4561
CAFRN	31.2287	0.8868	3.6963	15.5309	0.4449
RCAN	31.2545	0.8882	3.8576	15.5488	0.4344
HAN	31.2195	0.8867	3.8610	15.7322	0.4367
HANRS_L1	31.2699	0.8895	3.9280	15.3569	0.4308
HPANRS_L1	31.2337	0.8887	3.9477	15.5291	0.4293
HANRS_per	31.1164	0.8869	4.4041	15.1472	0.3511
HPANRS_per	31.0998	0.8859	4.4324	15.1720	0.3530

The quantitative values on the DFC dataset when the scaling factor equals 8 are depicted in Table 4.3. It is observed that the proposed HANRS_L1 outperforms other models regarding PSNR and SSIM. RCAN and HPANRS_L1 are ranked as the 2nd and 3rd in terms of PSNR, but for SSIM, HPANRS_L1 is ranked as the 2nd and RCAN is ranked as the 3rd. NIQE metric shows HANRS_per performs the best, HPANRS_per and HANRS_L1 are ranked the 2nd and 3rd.

Moreover, the visual results on the DFC dataset with scaling factors of $\times 2$ are also demonstrated in Figure 4.1 to Figure 4.4. We choose 4 patches that cover the railway, parking lot, road region, and roof region. Generally speaking, the qualitative results of the methods on DFC datasets with scaling factors of $\times 2$ and $\times 4$ do not have too much difference and are very close to the reference images as the inputs contain more details and information.

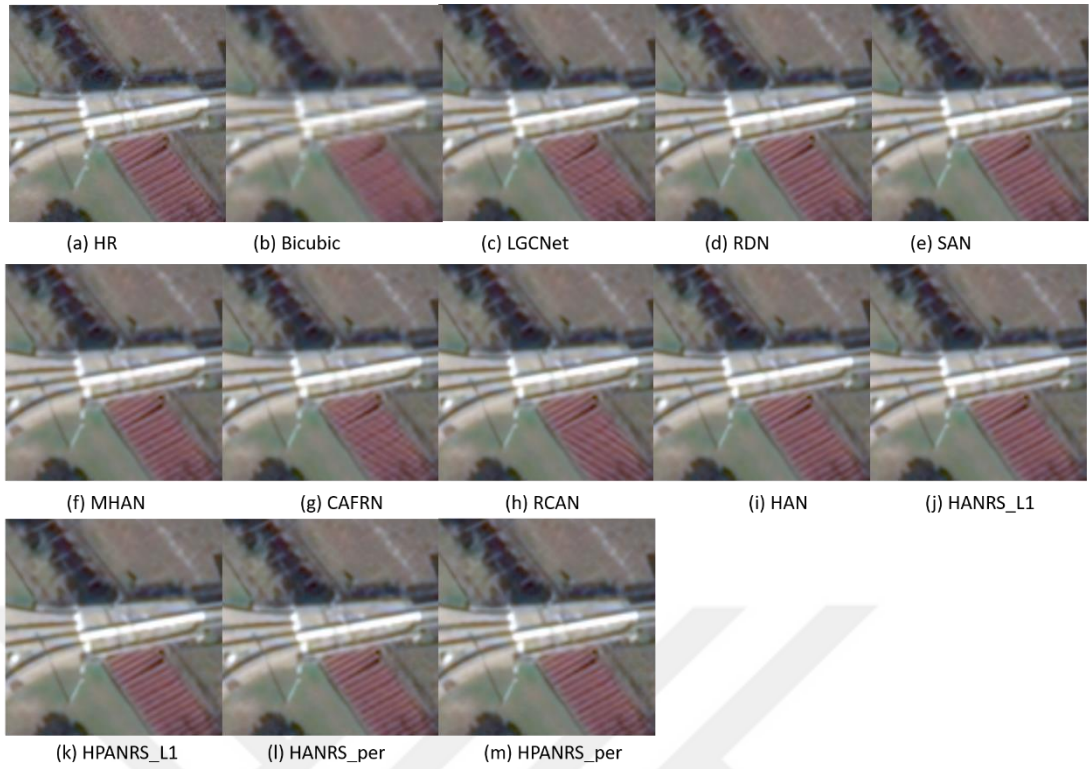


Figure 4.1 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on railway.

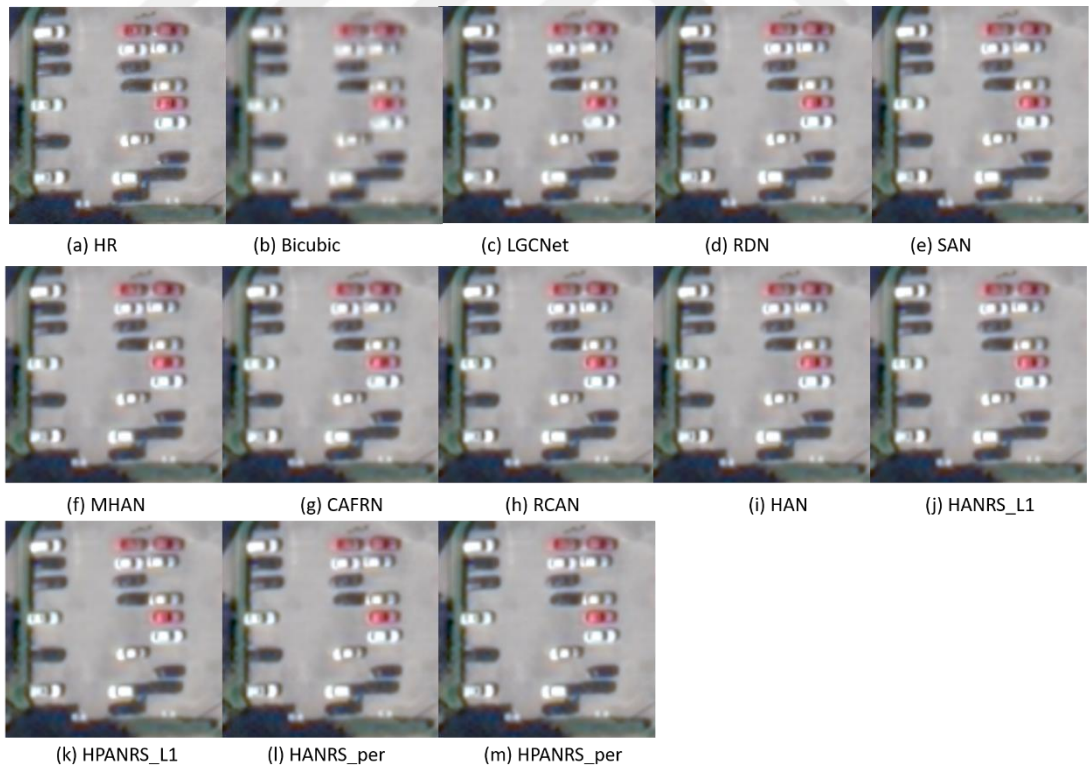


Figure 4.2 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on parking lot.



Figure 4.3 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on road region.

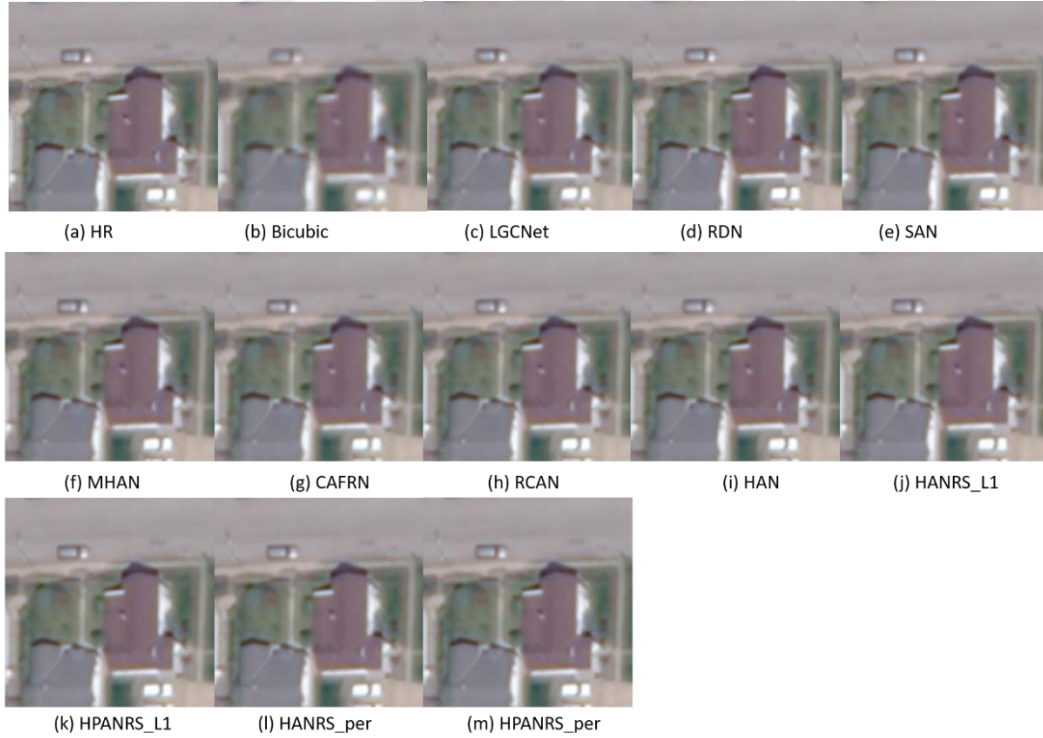


Figure 4.4 : Visualizations of DFC dataset with a scaling factor of $\times 2$ on roof region.

The visual results on the DFC dataset with scaling factors of $\times 4$ are demonstrated in Figure 4.5 to Figure 4.8.

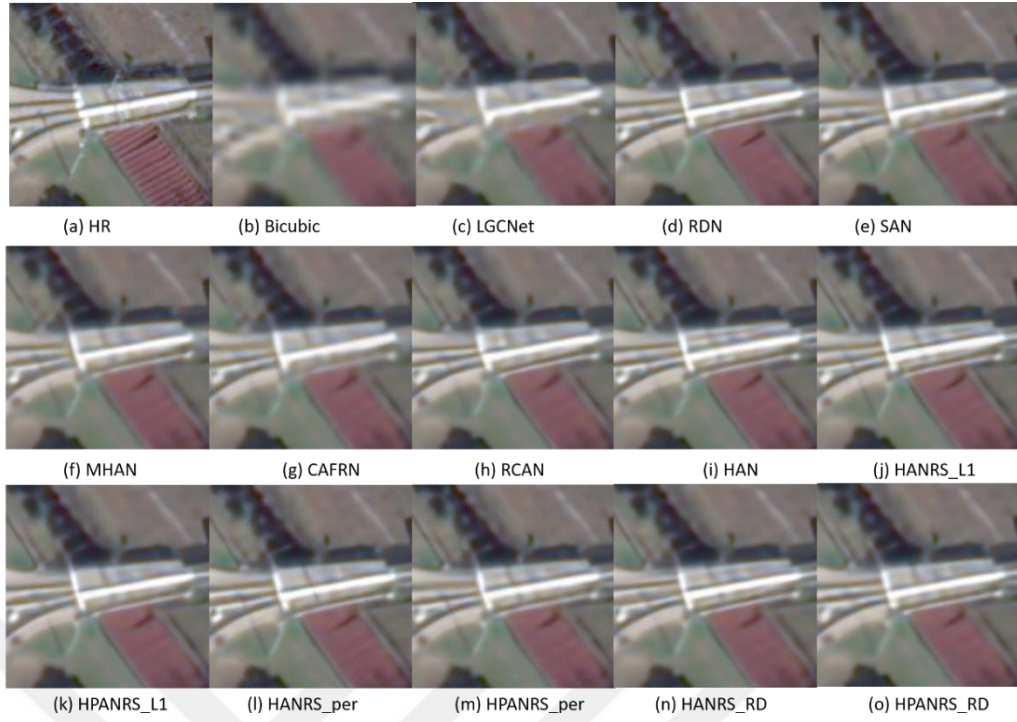


Figure 4.5 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on railway.

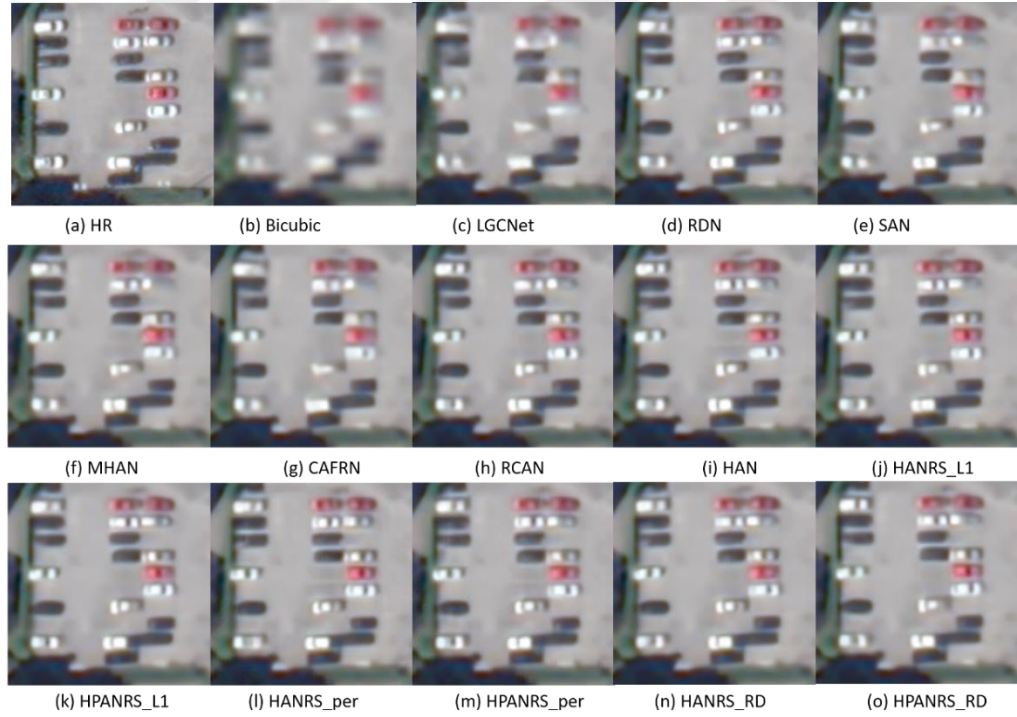


Figure 4.6 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on parking lot.



Figure 4.7 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on road region.

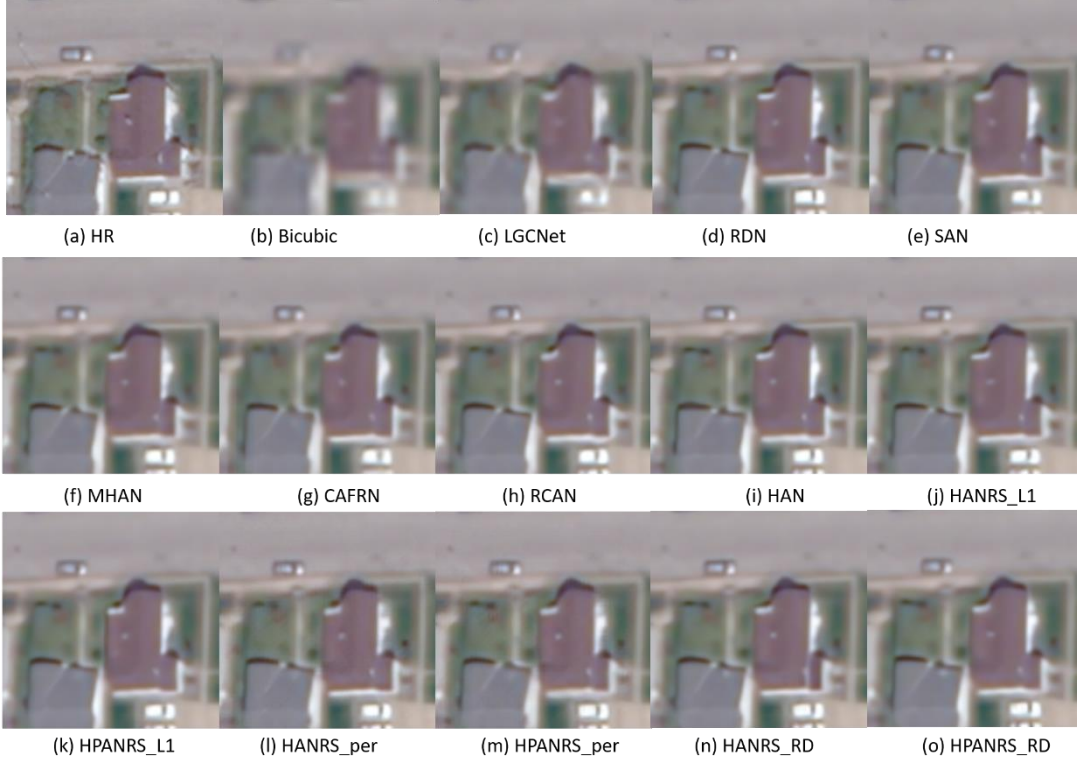


Figure 4.8 : Visualizations of DFC dataset with a scaling factor of $\times 4$ on roof region.

The visual results on the DFC dataset with scaling factors of $\times 8$ are also demonstrated

in Figure 4.9 to Figure 4.12.

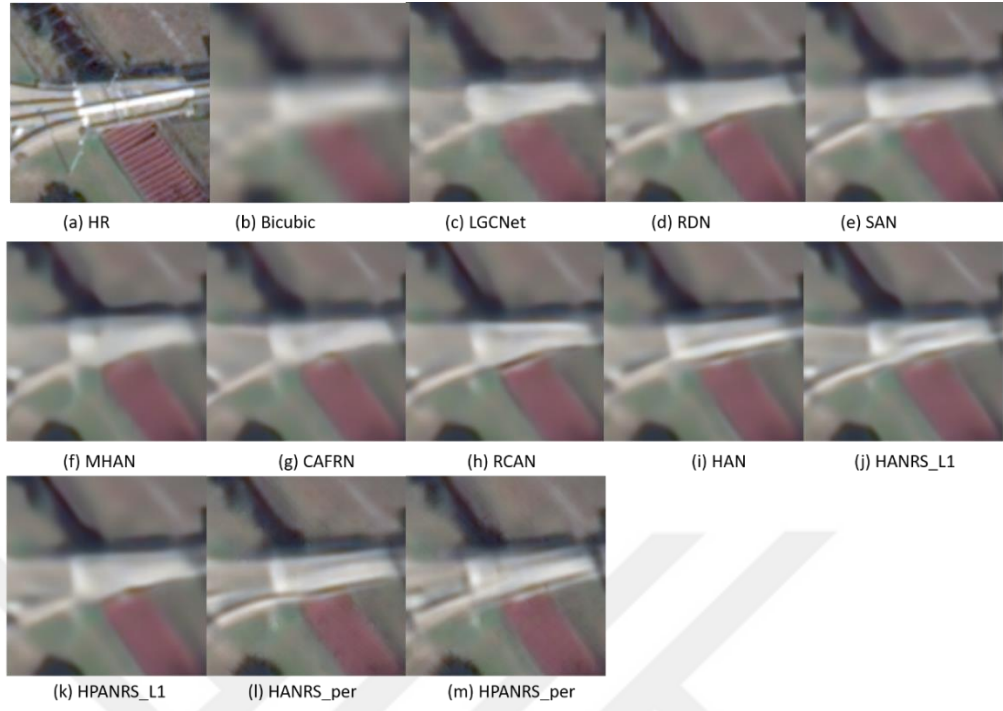


Figure 4.9 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on railway.

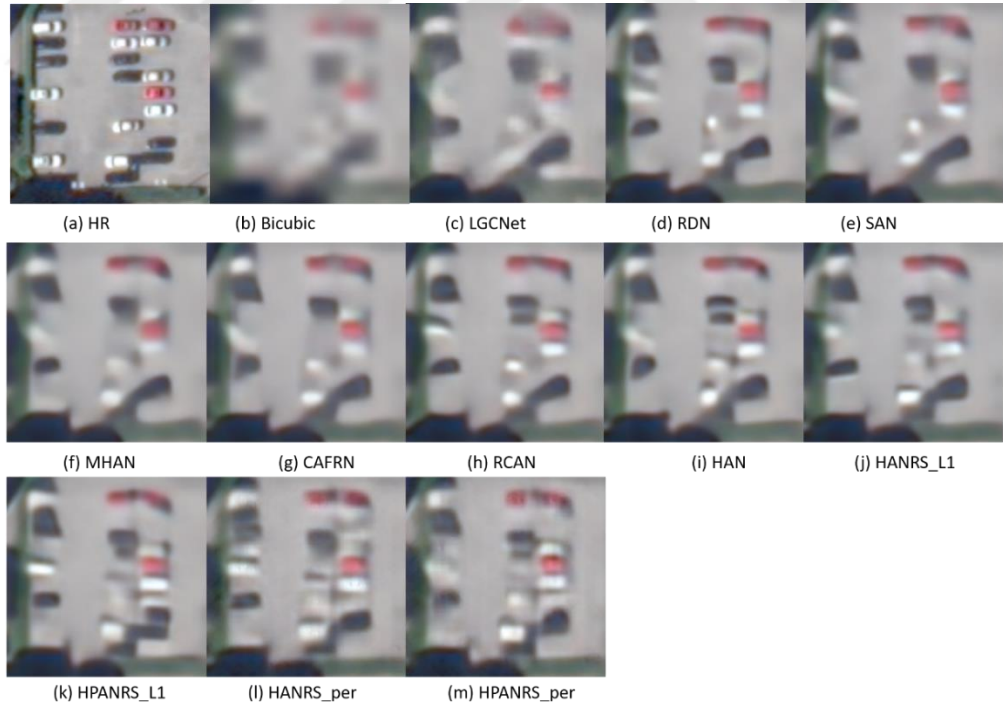


Figure 4.10 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on parking lot.

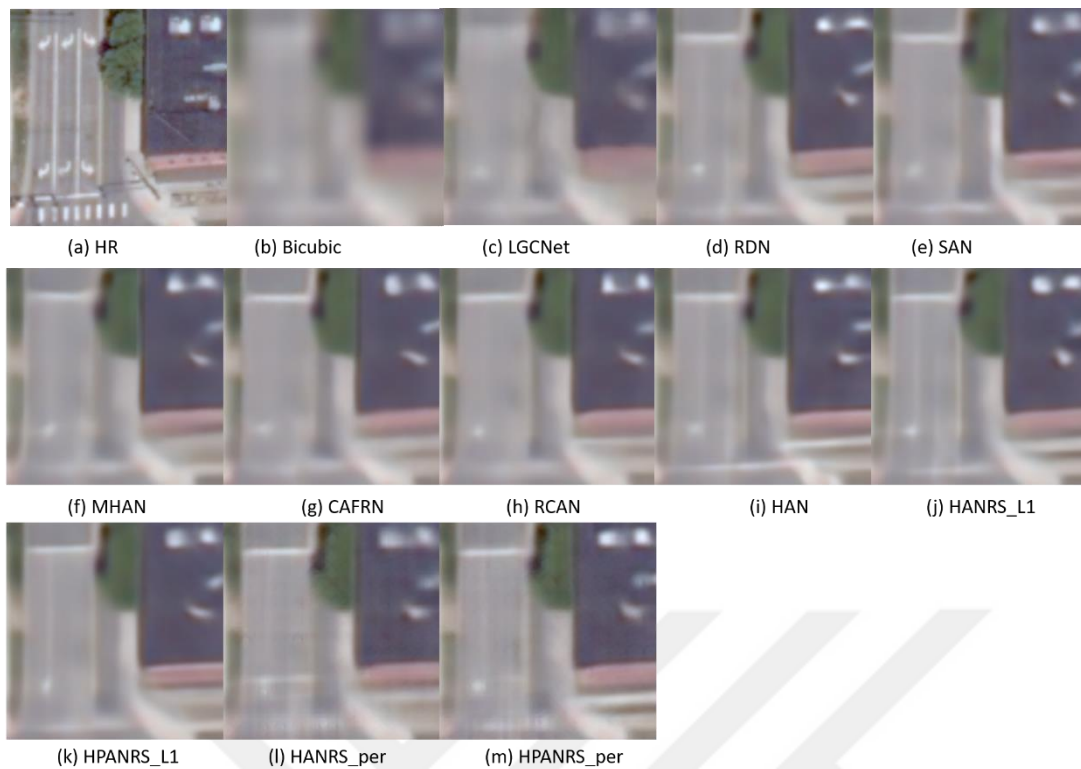


Figure 4.11 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on road region.

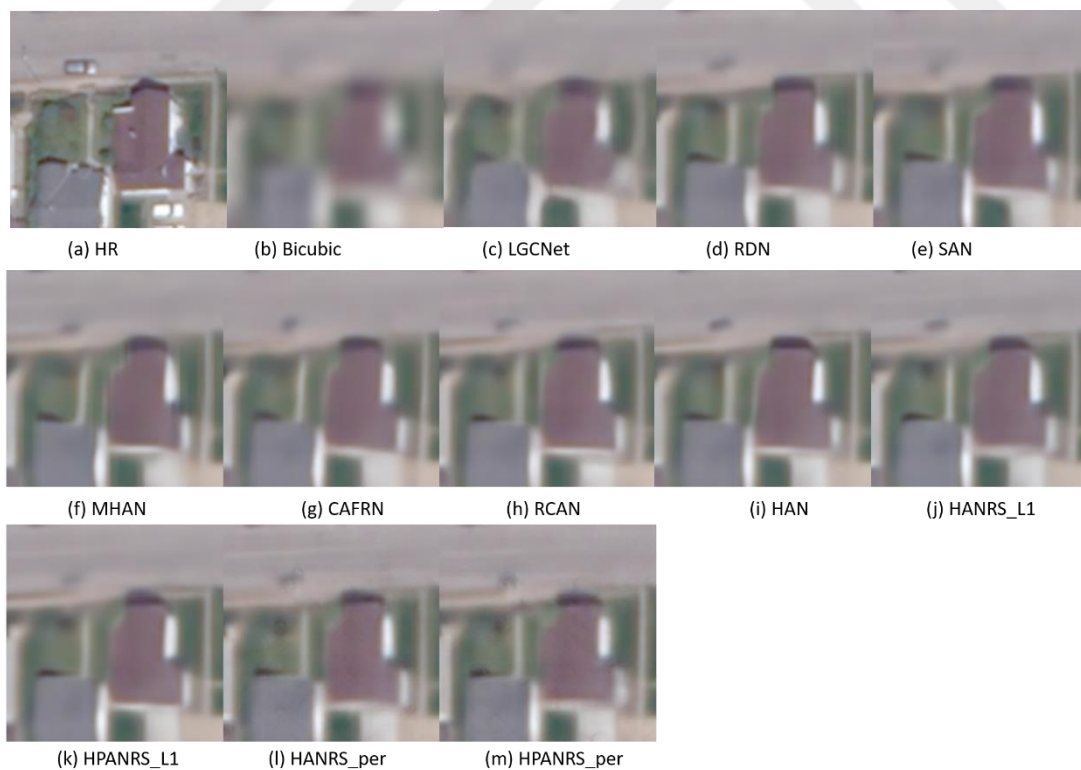


Figure 4.12 : Visualizations of DFC dataset with a scaling factor of $\times 8$ on the roof region.

The qualitative results of the methods on DFC datasets with scaling factors of $\times 8$ are extremely blurry as the input images provide less information and more challenging for the SR task.

4.1.2.2 Quantitative and qualitative results on Pleiades images

The quantitative and visual results of the attention CNN-based method on Pleiades images are presented. The quantitative results of the attention CNN-based method on Pleiades images with scaling factors $\times 2$, $\times 4$, and $\times 8$ are illustrated in Tables 4.4, 4.5, and 4.6, respectively.

As shown in Table 4.4, the RDN model outperforms the remaining algorithms regarding PSNR and SSIM. The HANRS_per and HPANRS_per ranked 2nd and 3rd in terms of PSNR. HANRS_L1 and SAN have the same value as the RDN model in terms of SSIM. The proposed HANRS_per and HPANRS_per show the best and second-best in terms of the AG, NIQE, and LPIPS. However, HANRS_L1, SAN, and RDN ranked as the 3rd in terms of AG, NIQE, and LPIPS, respectively.

Table 4.4 : Quantitative values of Pleiades images with a scaling factor of $\times 2$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	AG $\uparrow$	NIQE $\downarrow$	LPIPS $\downarrow$
Bic	34.0234	0.9904	5.9309	16.6224	0.1541
LGCNet	35.7670	0.9964	7.6347	16.4340	0.0993
RDN	36.1882	0.9967	7.6442	16.2332	0.0937
SAN	36.0466	0.9967	7.6983	16.0825	0.0945
MHAN	35.8010	0.9964	7.6678	16.8455	0.0984
CAFRN	35.8543	0.9963	7.7105	16.5986	0.0971
RCAN	35.8780	0.9966	7.6507	16.6521	0.0960
HAN	35.7627	0.9965	7.6668	17.1244	0.0976
HANRS_L1	35.9760	0.9967	7.7237	16.1815	0.0957
HPANRS_L1	35.9025	0.9966	7.5953	16.4138	0.0979
HANRS_per	36.0801	0.9966	7.7426	15.8762	0.0856
HPANRS_per	36.0566	0.9965	7.7316	16.0035	0.0864

As can be seen from Table 4.5, different metrics are pointing out different models as the best. PSNR metric shows RDN as the 1st, HANRS_L1 as the 2nd, and HPANRS_per as the 3rd. SSIM values are very close to each other and are observed

when the scale factor is set to 2. The same scaling factor $\times 2$, according to AG and LPIPS, the HANRS_per and HPANRS_per ranked as 1st and 2nd, respectively. But HANRS_L1 performs best concerning NIQE. HPANRS_L1 is ranked as the 2nd and RDN as the 3rd regards to NIQE.

Table 4.5 : Quantitative values of Pleiades images with a scaling factor of $\times 4$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	AG $\uparrow$	NIQE $\downarrow$	LPIPS $\downarrow$
Bic	31.5056	0.9302	3.7584	17.8552	0.3573
LGCNet	32.0637	0.9549	4.9689	17.6781	0.3132
RDN	32.2679	0.9588	5.2894	16.8582	0.2929
SAN	32.2132	0.9582	5.2347	17.0487	0.2963
MHAN	32.1921	0.9575	5.1498	17.4237	0.3005
CAFRN	32.1553	0.9573	5.1567	17.6394	0.3013
RCAN	32.1805	0.9577	5.1864	17.2590	0.2979
HAN	32.0670	0.9563	5.2517	16.9982	0.3002
HANRS_L1	32.2642	0.9589	5.3023	16.6872	0.2937
HPANRS_L1	32.2641	0.9590	5.2804	16.7023	0.2916
HANRS_per	32.1631	0.9574	5.6574	17.6197	0.2537
HPANRS_per	32.1926	0.9577	5.5764	17.1224	0.2575

Table 4.6 : Quantitative values of Pleiades images with a scaling factor of $\times 8$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	AG $\uparrow$	NIQE $\downarrow$	LPIPS $\downarrow$
Bic	30.3858	0.7997	2.3091	16.6409	0.5392
LGCNet	30.6280	0.8460	3.3402	16.5041	0.4809
RDN	30.7888	0.8595	3.7761	16.2057	0.4552
SAN	30.7542	0.8580	3.5465	16.9370	0.4625
MHAN	30.7695	0.8585	3.4048	16.6367	0.4719
CAFRN	30.7785	0.8586	3.3633	16.6678	0.4666
RCAN	30.7016	0.8550	3.7539	17.0807	0.4616
HAN	30.7223	0.8560	3.6220	16.6447	0.4621
HANRS_L1	30.7727	0.8591	3.7444	16.4158	0.4541
HPANRS_L1	30.7602	0.8586	3.8165	16.4020	0.4539
HANRS_per	30.6423	0.8561	4.0934	16.8355	0.4230
HPANRS_per	30.6746	0.8551	4.2662	16.7158	0.4211

It is observed from Table 4.6, that the PSNR and SSIM values are very close to each other at the scaling factor of $\times 8$. RDN received the highest score concerning PSNR,

SSIM, and NIQE. With regards to AG and LPIPS, HPANRS_per, HANRS_per, and HPANRS_L1 are ranked as the 1st, 2nd, and 3rd respectively.

The qualitative results of the attention CNN-based on Pleiades images with a scaling factor of $\times 2$ are presented from Figures 4.13 to 4.15. Generally speaking, the qualitative results of the methods on Pleiades images with a scaling factor of $\times 2$ do not have too much difference and are very close to the reference images as the inputs contain more details and information. This is consistent with what is observed in Table 4.4 concerning SSIM.

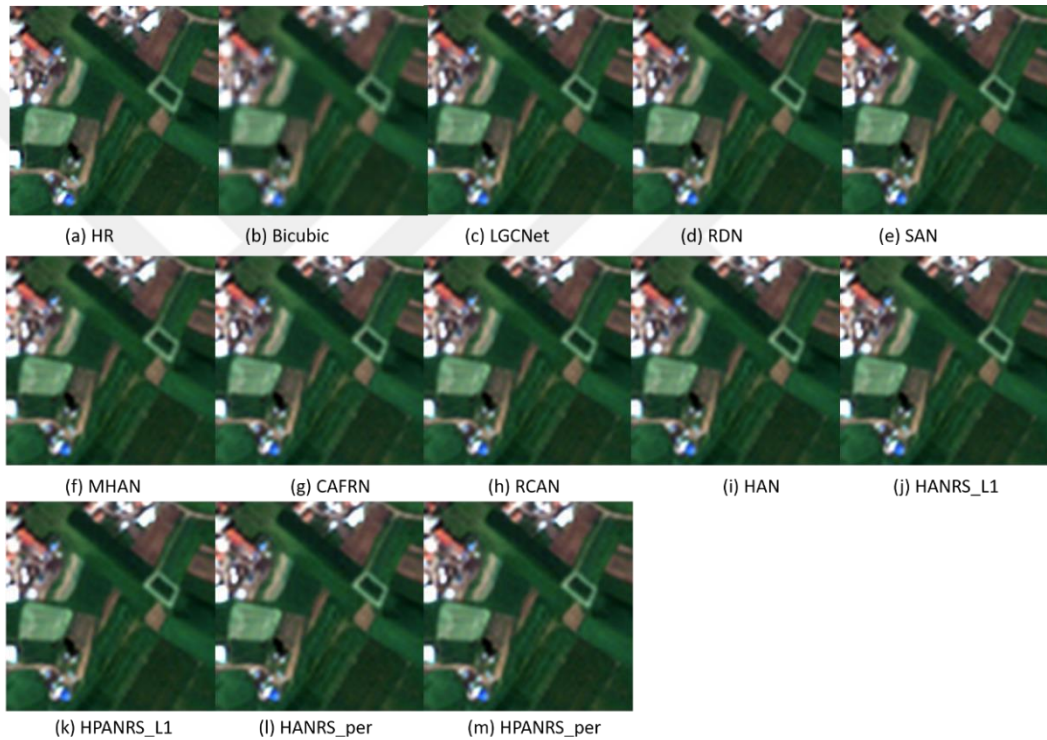


Figure 4.13 : Visualizations of Pleiades images with a scaling factor of $\times 2$.

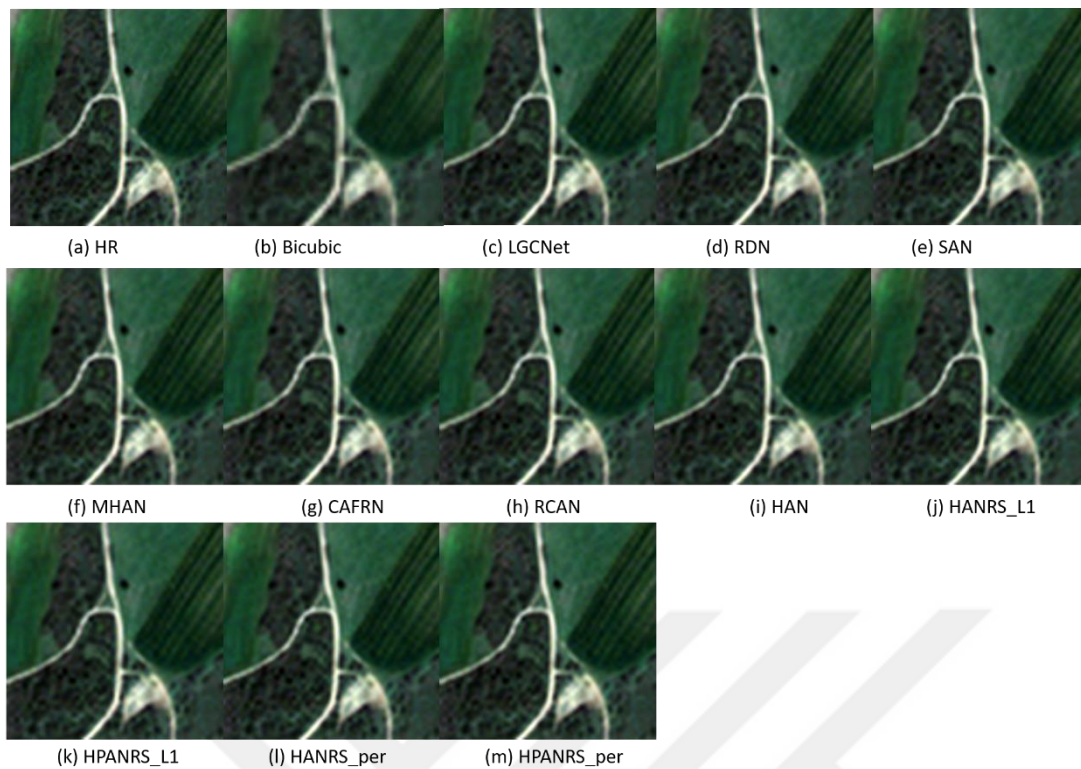


Figure 4.14 : Visualizations of Pleiades images with a scaling factor of $\times 2$.

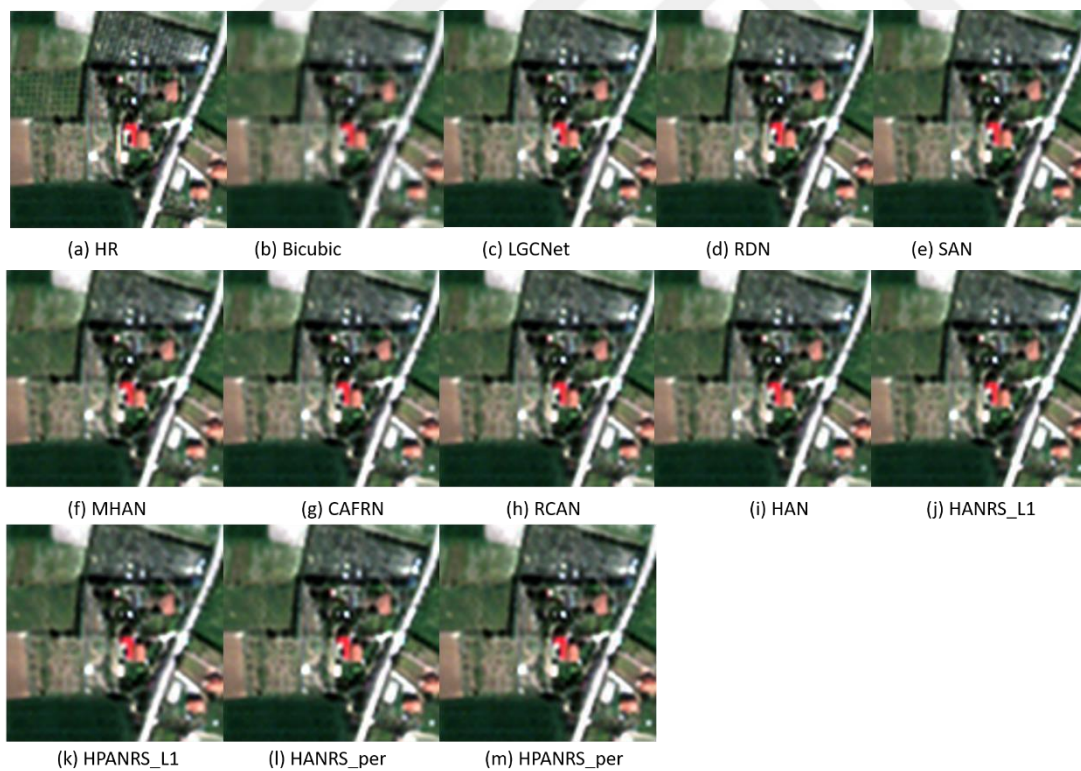


Figure 4.15 : Visualizations of Pleiades images with a scaling factor of $\times 2$.

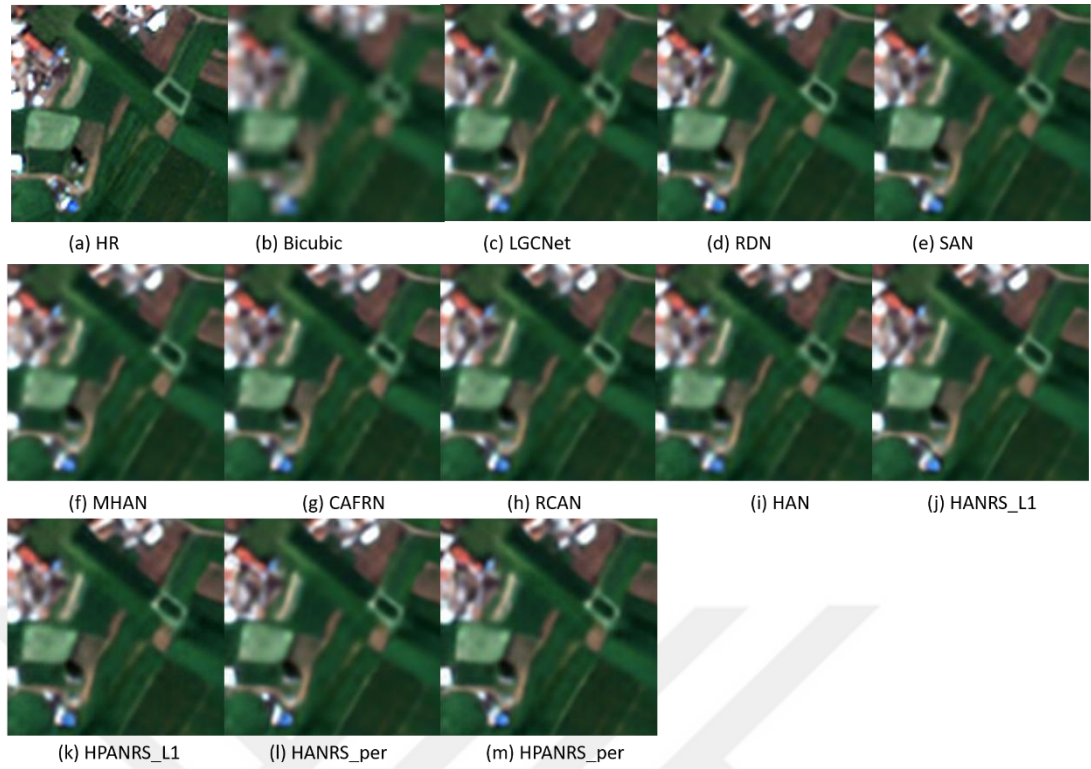


Figure 4.16 : Visualizations of Pleiades images with a scaling factor of $\times 4$.

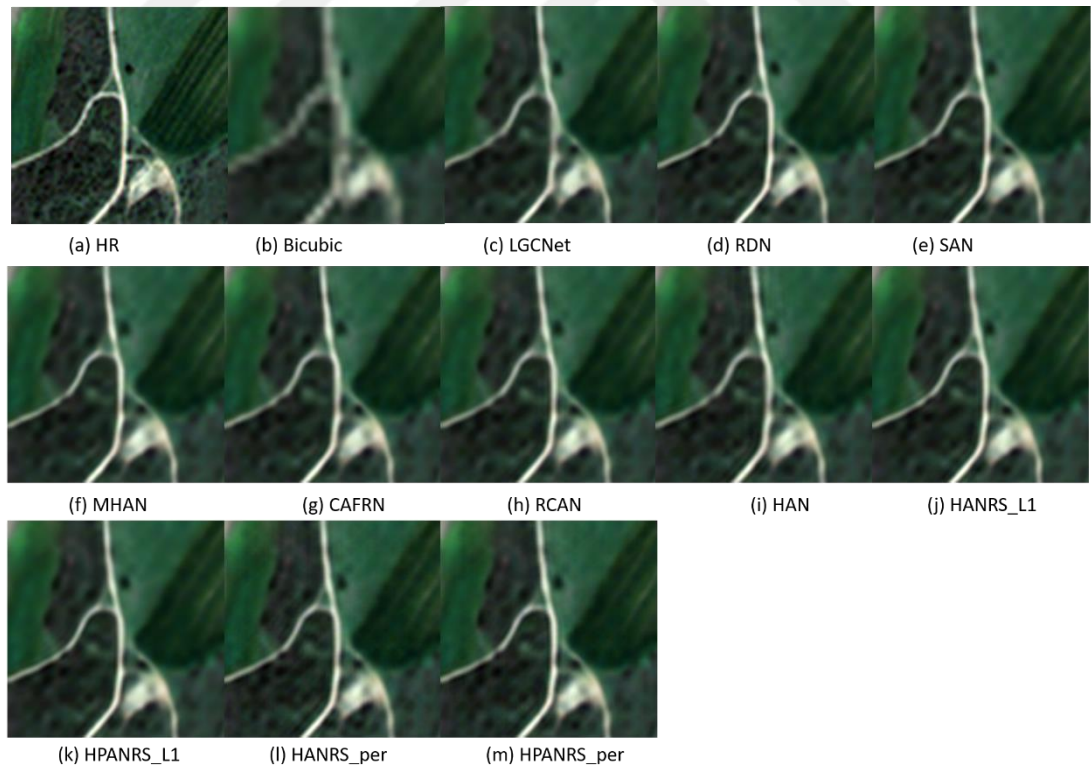


Figure 4.17 : Visualizations of Pleiades images with a scaling factor of $\times 4$.

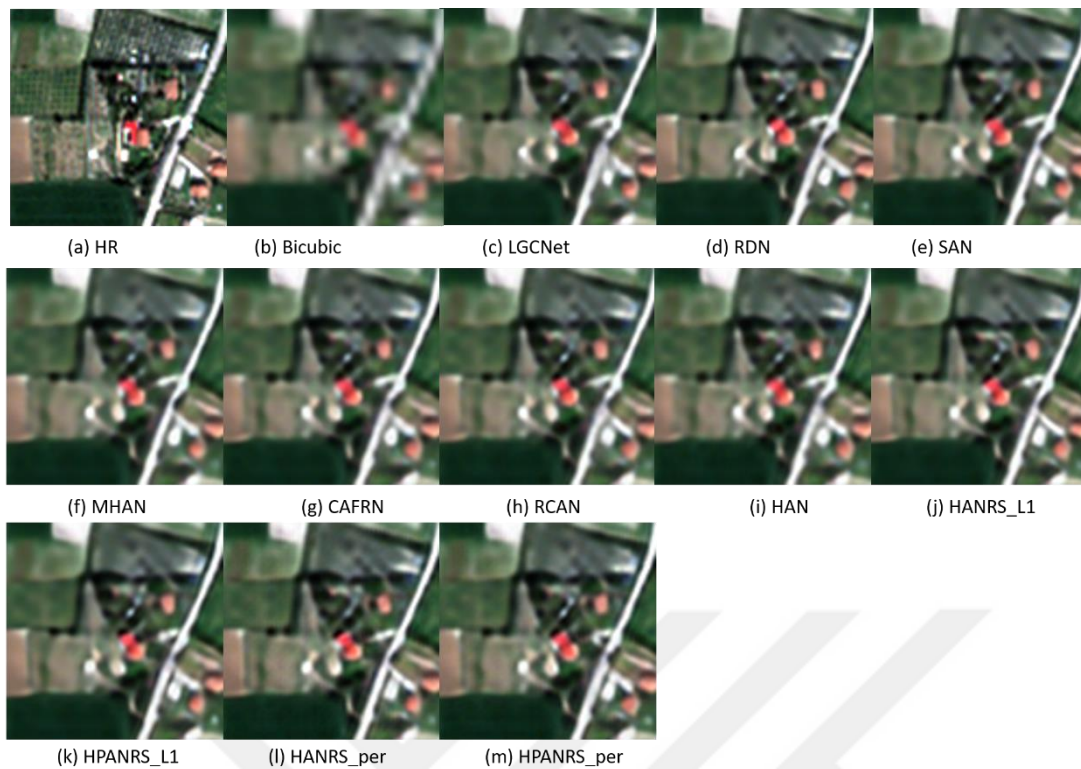


Figure 4.18 : Visualizations of Pleiades images with a scaling factor of $\times 4$.

The qualitative results of the attention CNN-based on Pleiades images with a scaling factor of $\times 4$ are presented from Figures 4.16 to 4.18. Same with the results of the scaling factor of $\times 2$, the deep learning methods outperform the bicubic method but seem very close to each other.

The qualitative results of the attention CNN-based on Pleiades images with a scaling factor of $\times 8$ are presented in Figures 4.19 to 4.21. As observed from Figures 4.19 to 4.21, as the input images contain less information when the scale factor is set to 8, thus, the super-resolved images are seriously blurry. It is observed that the proposed method trained with perceptual loss still shows good performance in the roads, edges, and straight lines.

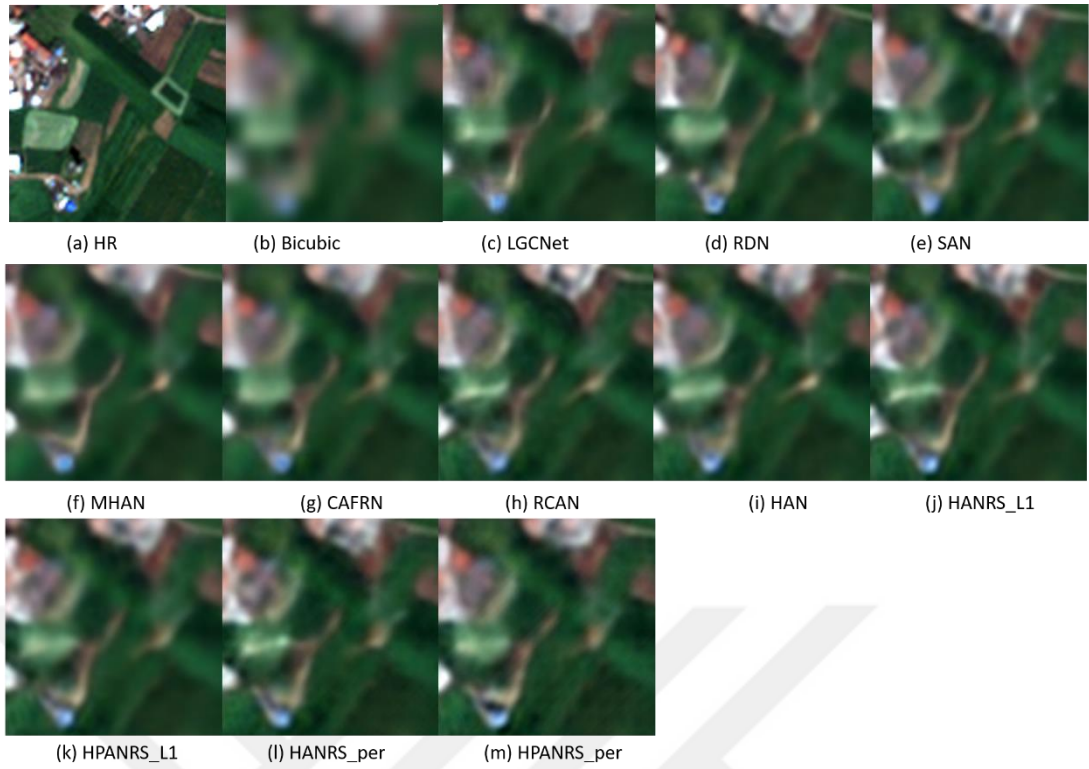


Figure 4.19 : Visualizations of Pleiades images with a scaling factor of $\times 8$.

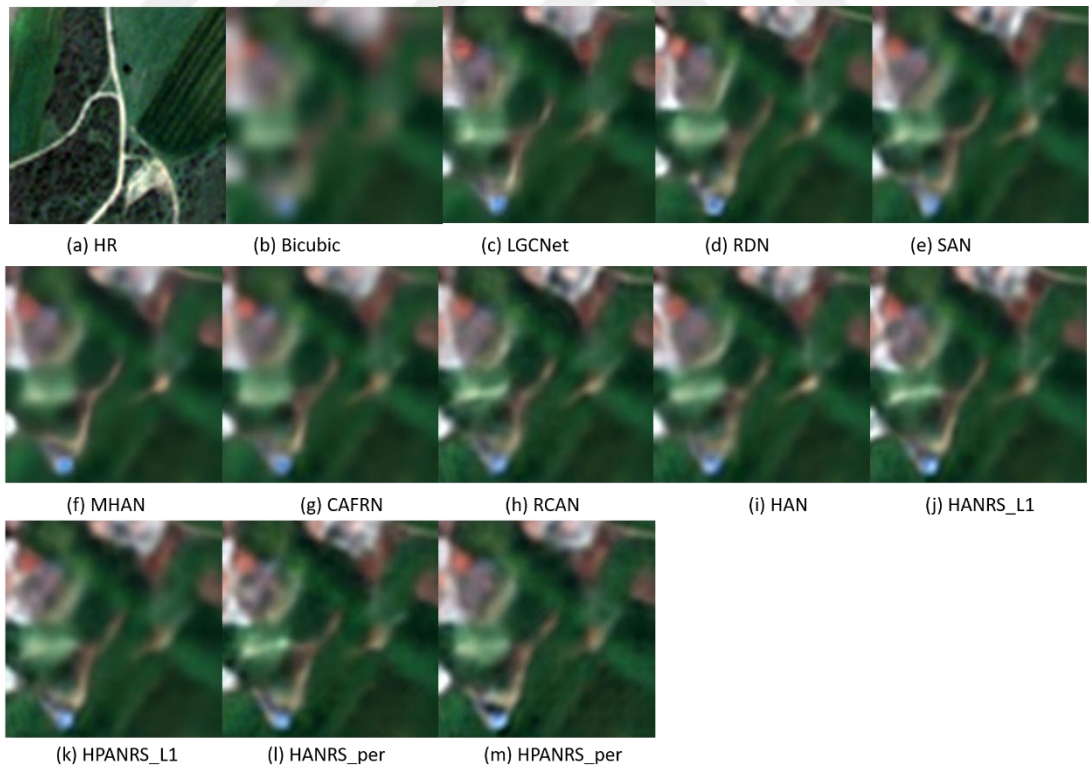


Figure 4.20 : Visualizations of Pleiades images with a scaling factor of $\times 8$.

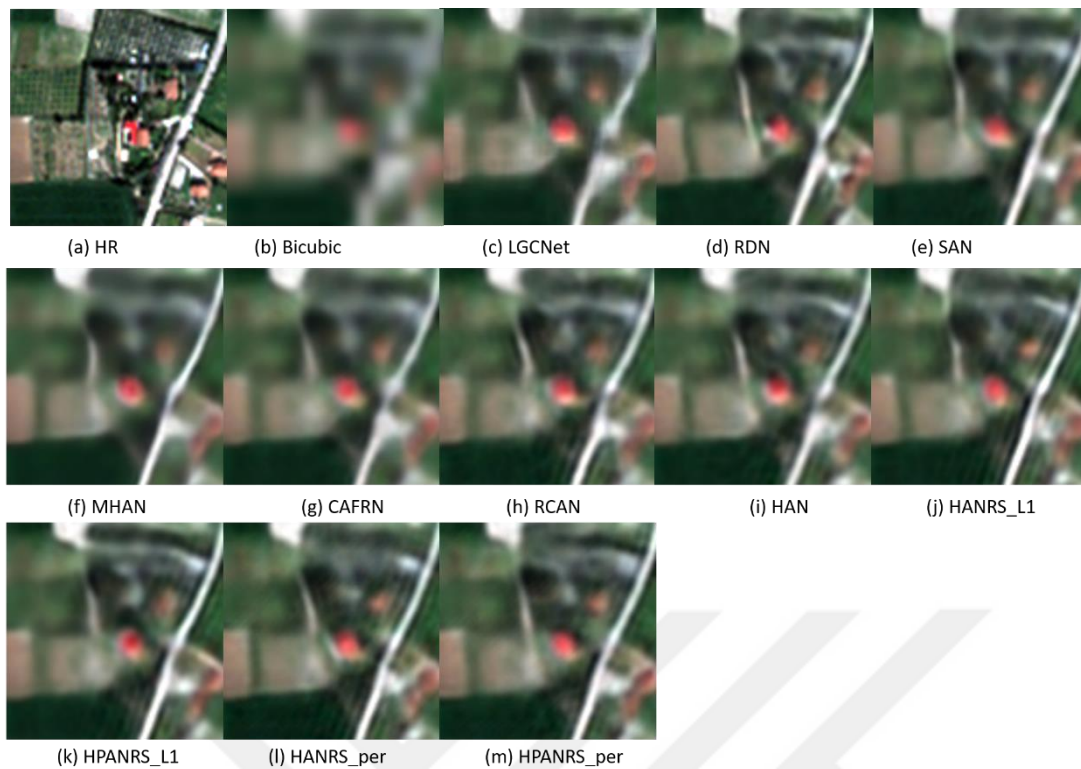


Figure 4.21 : Visualizations of Pleiades images with a scaling factor of $\times 8$.

4.1.3 Results on original DFC images

The goal of the SR is to super-resolve the resolution of the RS imageries. In this subsection, visualizations of the presented methods on the original DFC dataset (in full resolution with a scaling factor of $\times 4$) are evaluated and presented in Figures 4.13 to 4.17.

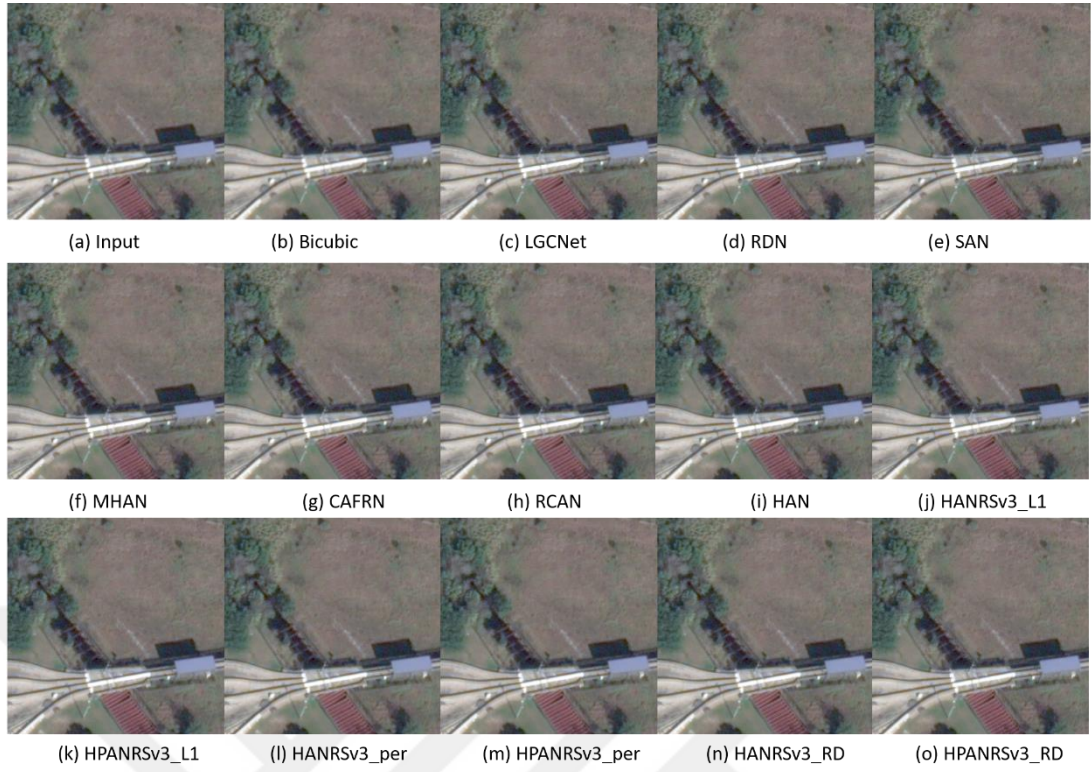


Figure 4.22 : Visual results on original DFC images on railway area.

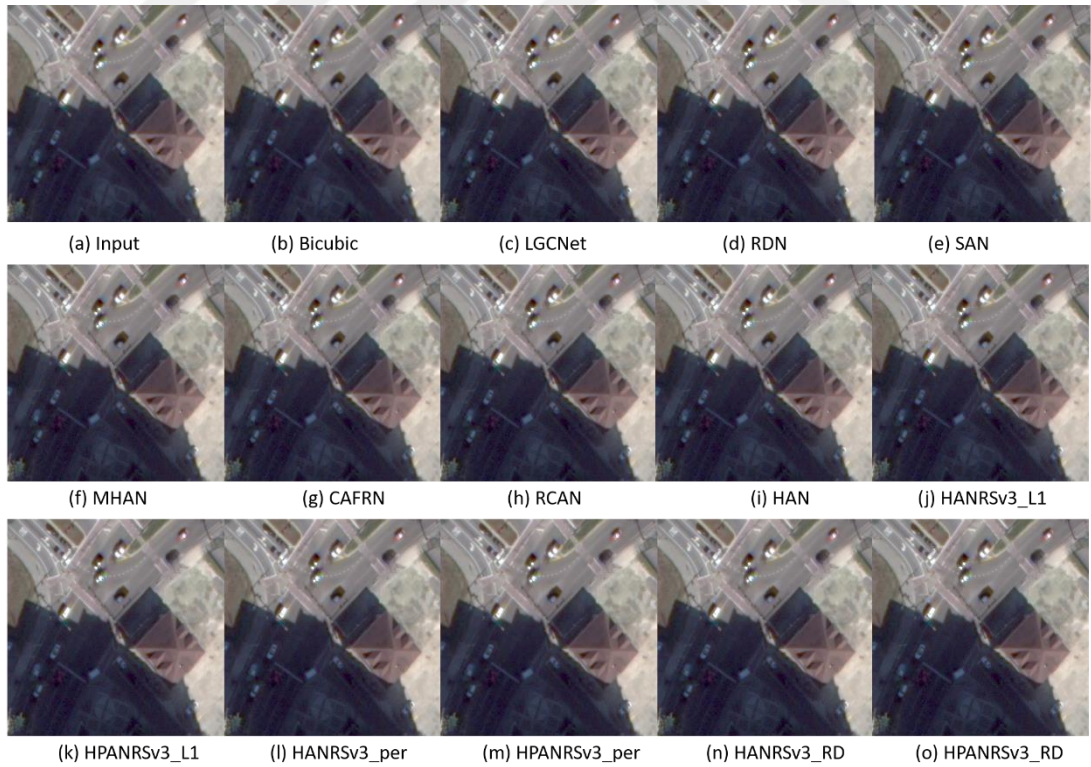


Figure 4.23 : Visual results on original DFC images on roof.

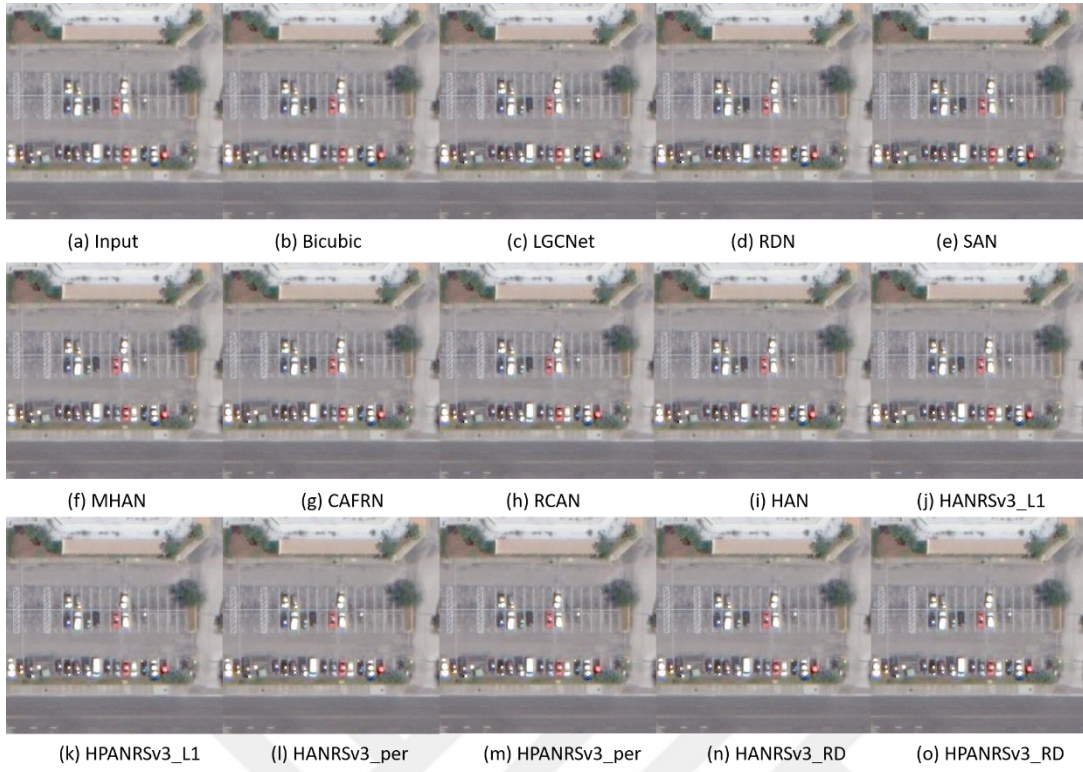


Figure 4.24 : Visual results on original DFC images on parking lot.

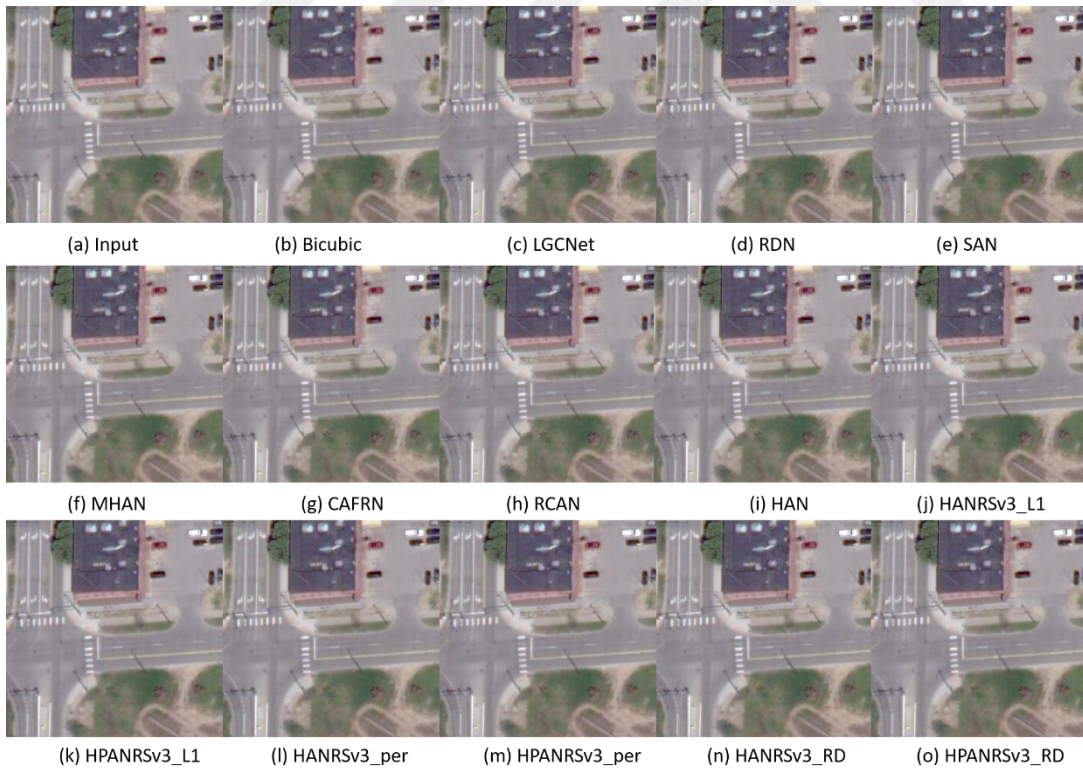


Figure 4.25 : Visual results on original DFC images on road region.



Figure 4.26 : Visual results on original DFC images on roof and buildings.

4.1.4 Results on multispectral images

Moreover, the visual results of the proposed methods on multispectral images are also evaluated and presented. Normally, the super-resolution methods are trained on RGB channels with 8-bit images. However, the original optical remote sensing images usually contain 4 bands or 8 bands, and the radiometric resolution of the images is 11-bit or 12-bit. Therefore, it is meaningful to evaluate the proposed methods on multispectral images.

The visual results of the proposed methods on Pleiades images with a scale factor of $\times 4$ are presented in Figures 4.18 to 4.20. The visual results of the proposed methods on Worldview-3 images with a scale factor of $\times 4$ are presented in Figures 4.21 to 4.23. Generally, although the proposed methods are trained on RGB channels, they are also effective in preserving the spectral and spatial information on the multispectral images.

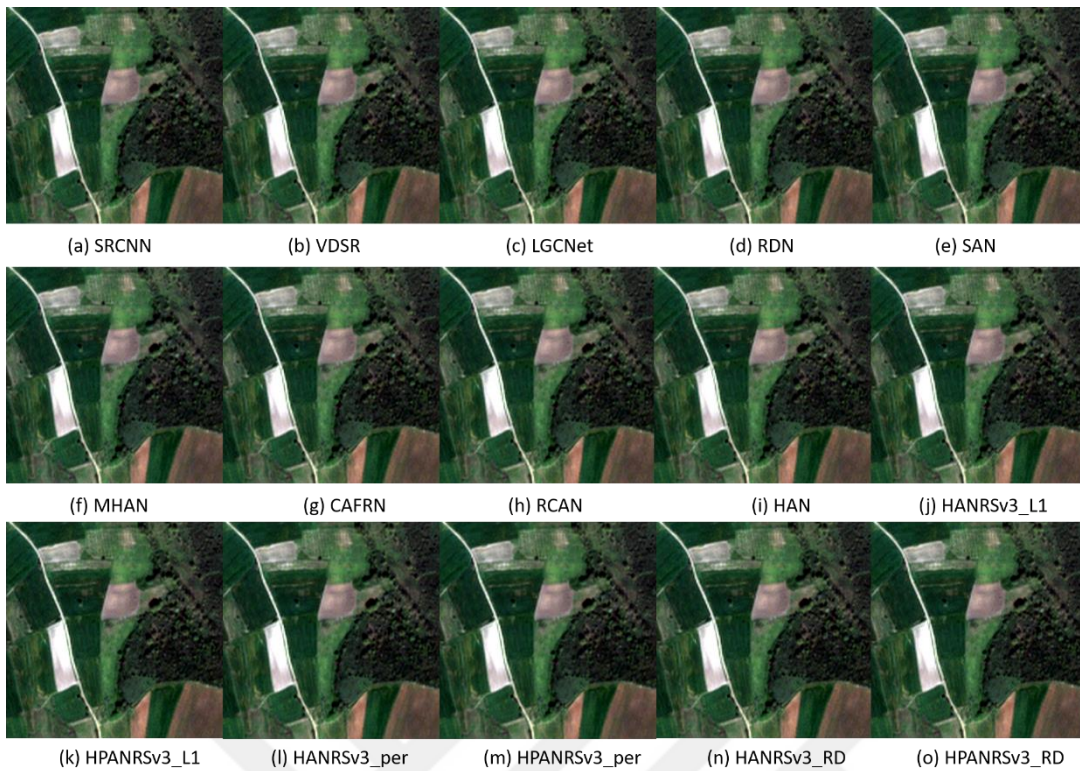


Figure 4.27 : Visual Results on Pleiades multispectral images.

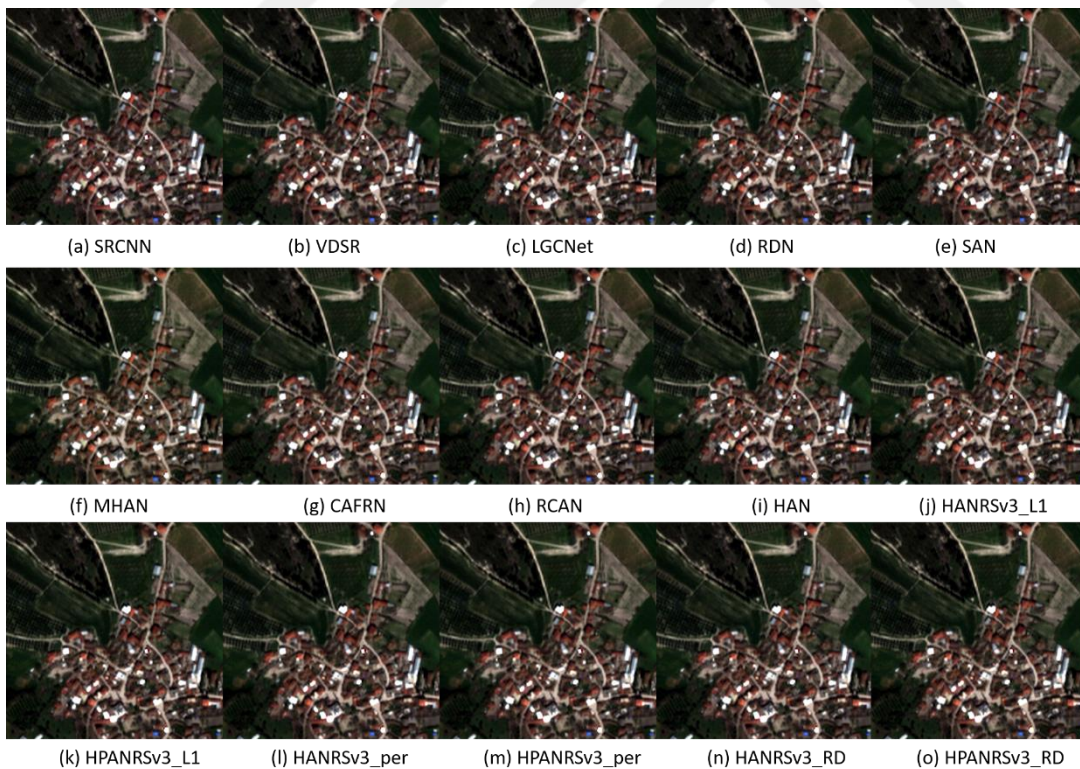


Figure 4.28 : Visual Results on Pleiades multispectral images.

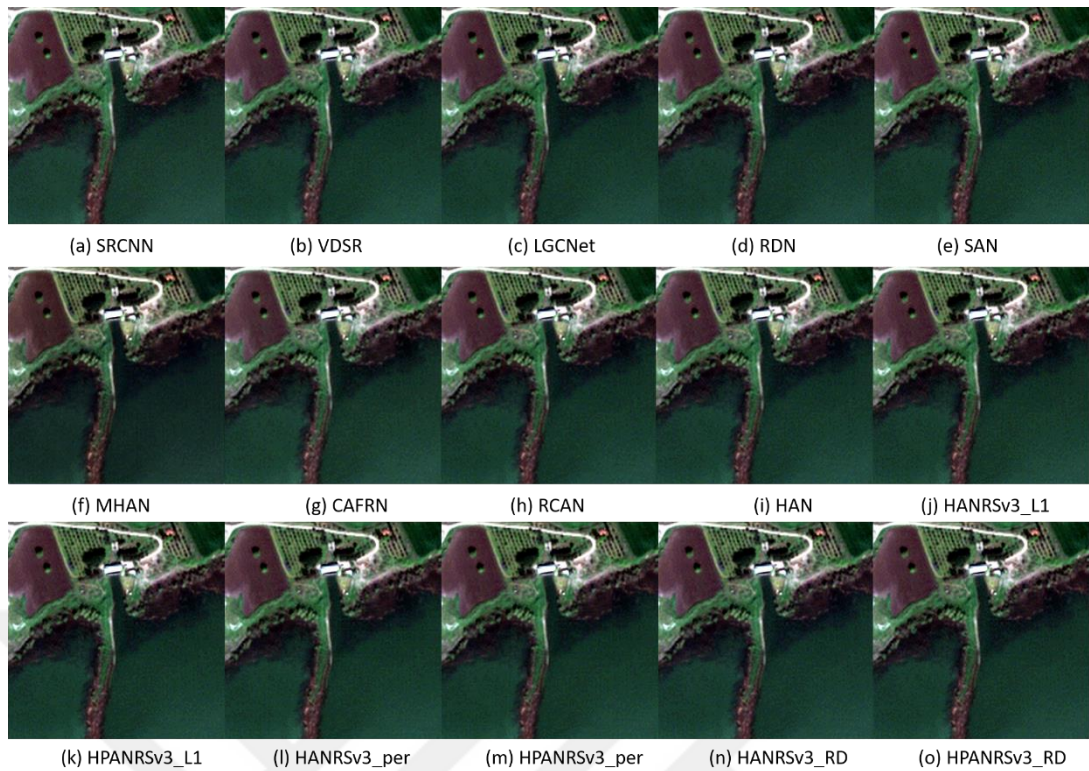


Figure 4.29 : Visual Results on Pleiades multispectral images.

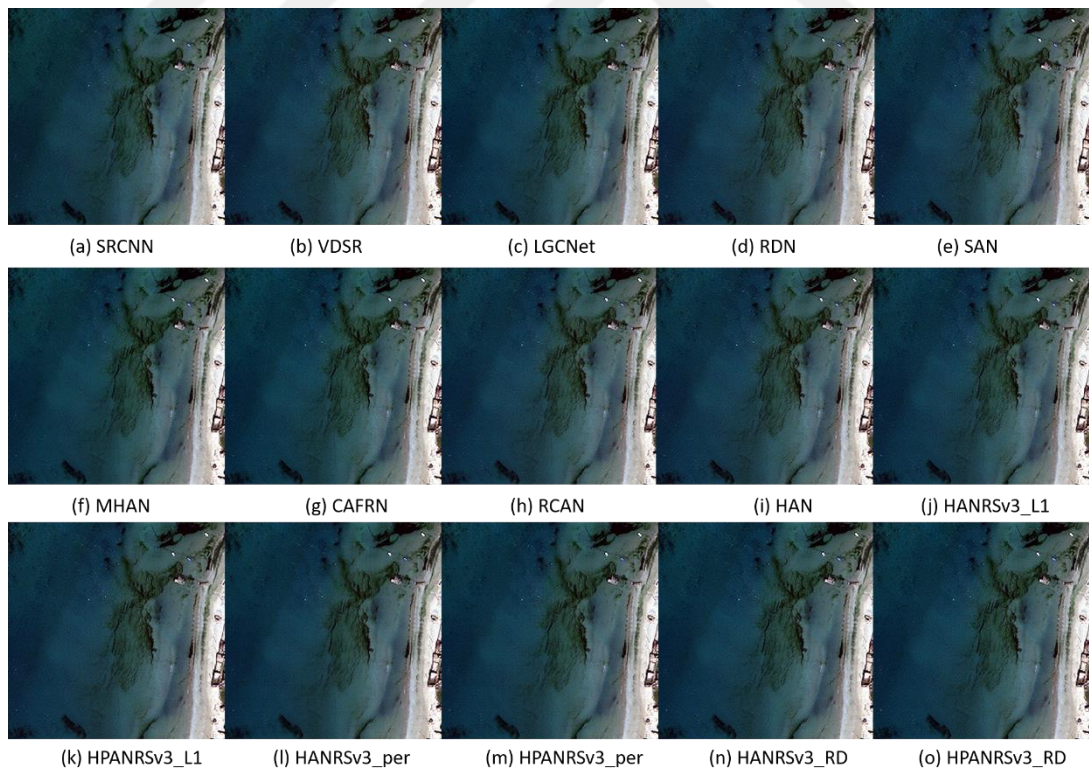


Figure 4.30 : Visual Results on Worldview-3 multispectral images.

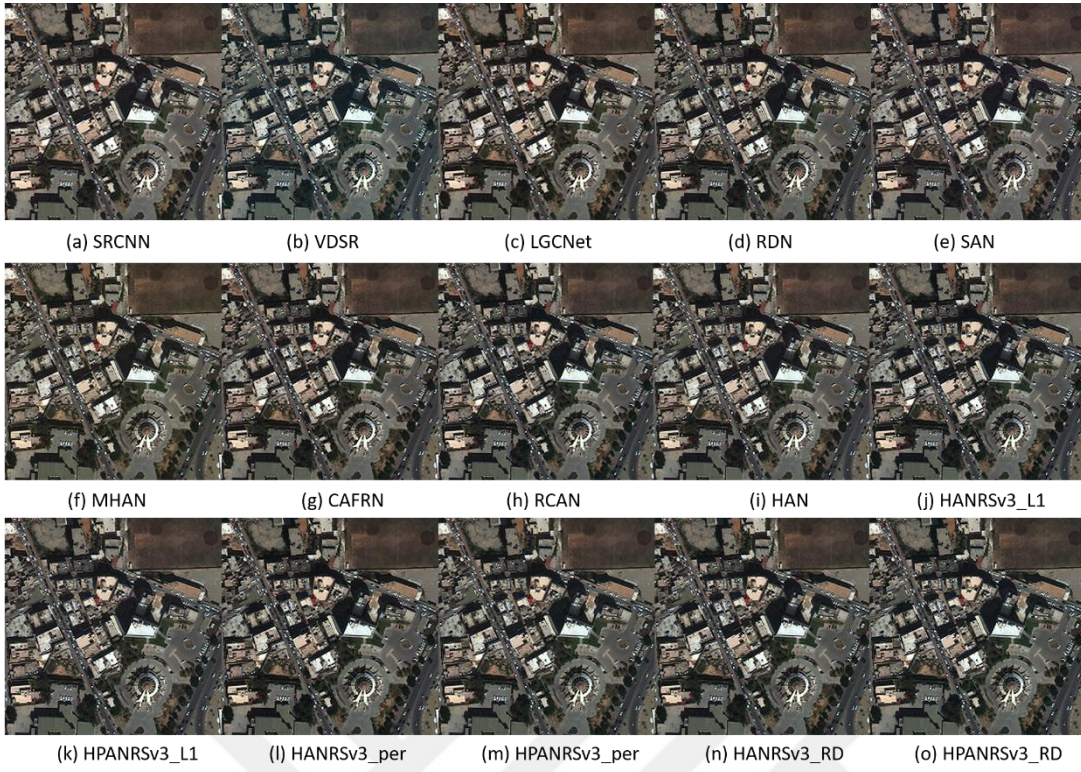


Figure 4.31 : Visual Results on Worldview-3 multispectral images.

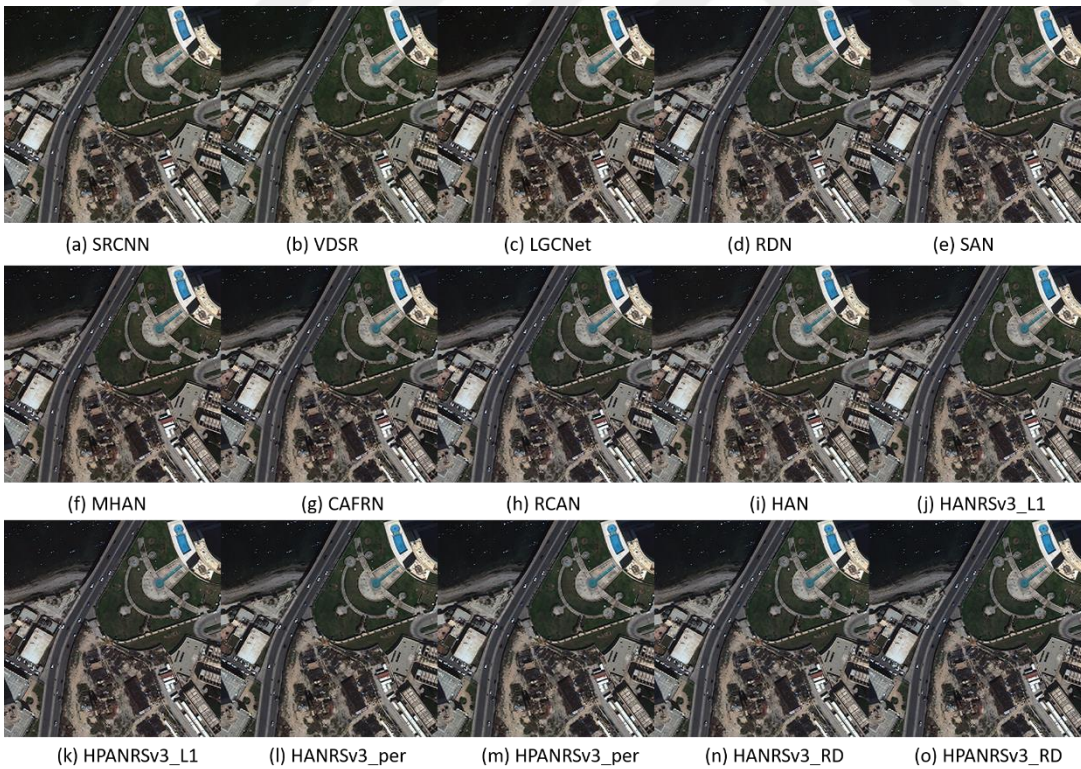


Figure 4.32 : Visual Results on Worldview-3 multispectral images.

4.1.5 Discussion

The results of the attention CNN-based methods on DFC datasets and Pleiades dataset indicate the proposed method can generate satisfactory results both quantitatively and perceptually. Generally speaking, the attention mechanism is effective when combined with the CNN modules as the attention mechanism is able to extract the deeper relationships among channels. Apart from the experiments on the downsampled images, the results on the original and multispectral images are also evaluated. The results illustrate that although the training process is conducted in RGB channels, the results of the trained model show promising results in super-resolving the images in original and multispectral images.

4.2 Results of the Attention GAN-based SFSR on Remote Sensing Images

4.2.1 Experiment settings

For the attention GAN-based SFSR on RS images, the models are trained in PyTorch 1.0 ('pytorch', n.d.) with NVIDIA GeForce GTX 1060 GPU. All models are optimized by Adam optimizer with $\beta_1=0.9$, $\beta_2=0.999$ with RGB images.

The scaling factor between LR and HR images is set to $\times 4$. The LR inputs are degraded with bicubic interpolation from the corresponding HR images.

In the training phase, 16 RGs are applied. 64 (3×3) filters are utilized in each layer. Leaky-ReLU with $\alpha = 0.2$ is adopted except the conv layers in RIR block.

In the training process, the LR inputs are clipped to 64×64 pixels. The learning rate is 1×10^{-4} and decayed 0.5 every 3000 iterations.

4.2.2 Quantitative results

The effectiveness of the proposed attention GAN-based methods is evaluated by using the PSNR, SSIM, AG, LPIPS, PI, and NIQE, and the numerical assessment values are depicted in Table 4.7. For the comparison, the SRGAN and ESRGAN are retrained on DFC 2019 dataset. RRDB16 and RRDB20 in the model name indicate the model with

16 and 20 RRDB blocks, respectively. ESRGAN_CA2_RRDB16_L1 and ESRGAN_CA2_RRDB16_per denote the ESRGAN method with a Channel attention (CA) module trained with L1 loss function and GAN loss function, respectively. ESRGAN_selfA and ESRGAN_MHOA represent the ESRGAN model with the Self-attention module and Mixed high order attention module, respectively. ESRGAN_MHOA4 represents the ESRGAN model with the proposed RIR block as shown in Figure 3.7. As can be seen from Table 4.7, ESRGAN_MHOA4_RRDB16_L1, ESRGAN_MHOA4_RRDB20_L1, and ESRGAN_MHOA4_RRDB16_L1 are ranked as the 1st, 2nd, 3rd with regards to PSNR and SSIM, respectively. Regarding AG, ESRGAN_RRDB16_per performs best, ESRGAN_MHOA_RRDB16_per is ranked 2nd, and SRGAN_per is ranked 3rd. According to the LPIPS metric, ESRGAN_MHOA4_RRDB16_per, ESRGAN_selfA_RRDB16_per, and ESRGAN_MHOA4_RRDB20_per are ranked as the 1st, 2nd, and 3rd, respectively. Finally, with respect to PI and NIQE, the ESRGAN with the proposed MHOA model with 20 RRDBs, 16 RRDBs, and ESRGAN with the Self-attention module are ranked as the 1st, 2nd, and 3rd, respectively.

Table 4.7 : Quantitative results of Attention GAN-based model on DFC dataset with a scaling factor of $\times 4$. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	AG $\uparrow$	LPIPS $\downarrow$	PI $\downarrow$	NIQE $\downarrow$
SRGAN16_L1	33.7744	0.9757	4.5508	0.2608	5.7406	5.6861
SRGAN16_per	31.3964	0.9675	5.5091	0.2263	5.7728	7.4661
ESRGAN_RRDB16_L1	33.6692	0.9762	4.5131	0.2573	5.7459	5.7264
ESRGAN_RRDB16_per	32.1699	0.9671	6.0494	0.2075	4.4209	4.9926
ESRGAN_CA2_RRDB16_L1	33.8541	0.9767	4.5925	0.2506	5.6466	5.4904
ESRGAN_CA2_RRDB16_per	32.4771	0.9683	5.4551	0.2081	4.2385	4.6328
ESRGAN_selfA_RRDB16_L1	33.6137	0.9769	4.5840	0.2455	5.5948	5.3723
ESRGAN_selfA_RRDB16_per	32.3944	0.9692	5.4418	0.1935	3.9230	4.0425
ESRGAN_MHOA_RRDB16_L1	33.9408	0.9769	4.5936	0.2507	5.6675	5.5342
ESRGAN_MHOA_RRDB16_per	32.5334	0.9683	5.5158	0.2069	4.2720	4.7143
ESRGAN_MHOA4_RRDB16_L1	34.0352	0.9777	4.6554	0.2436	5.6011	5.4208
ESRGAN_MHOA4_RRDB16_per	32.7795	0.9695	5.4665	0.1930	3.9195	3.9018
ESRGAN_MHOA4_RRDB20_L1	33.9983	0.9773	4.6678	0.2448	5.5705	5.3492
ESRGAN_MHOA4_RRDB20_per	32.8474	0.9685	5.4714	0.1969	3.7344	3.6581

4.2.3 Qualitative results

In this subsection, visual results of the proposed attention GAN-based method on single-frame RSISR (DFC 2019 dataset) with a scale factor of $\times 4$ are illustrated as shown in Figures 4.33 to 4.39.

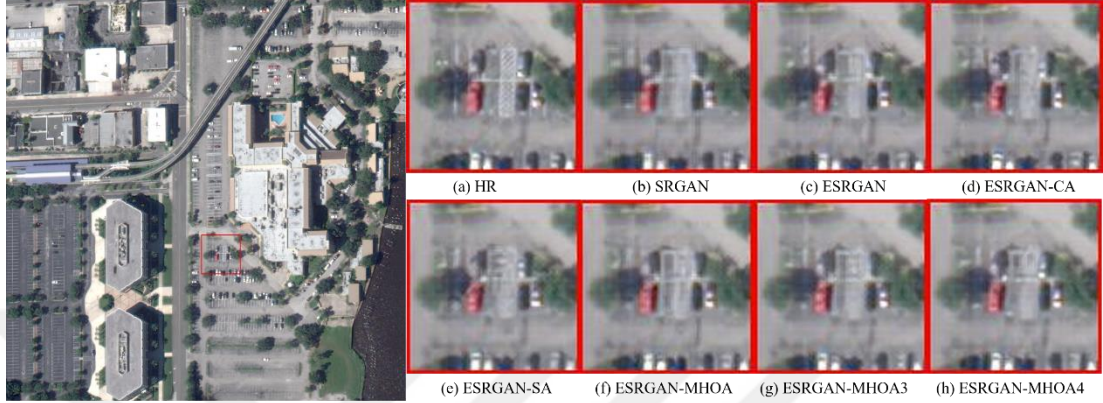


Figure 4.33 : Visual results on DFC images.

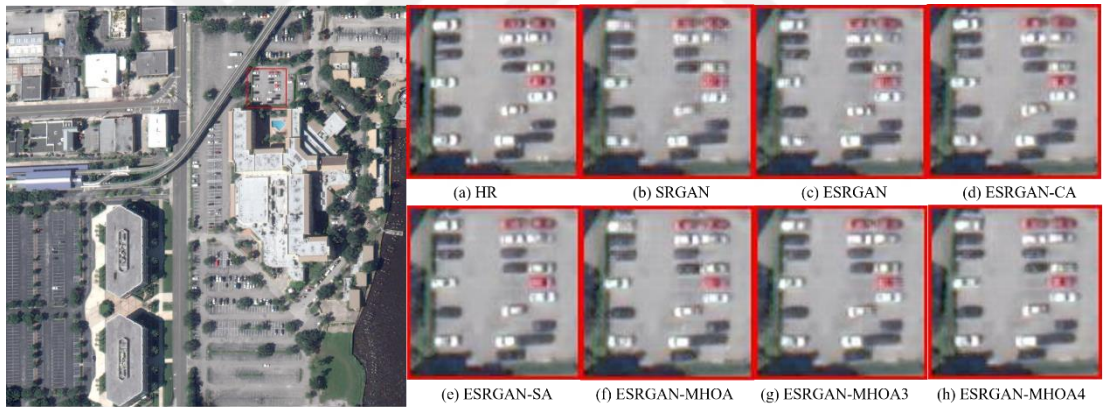


Figure 4.34 : Visual results on DFC images.

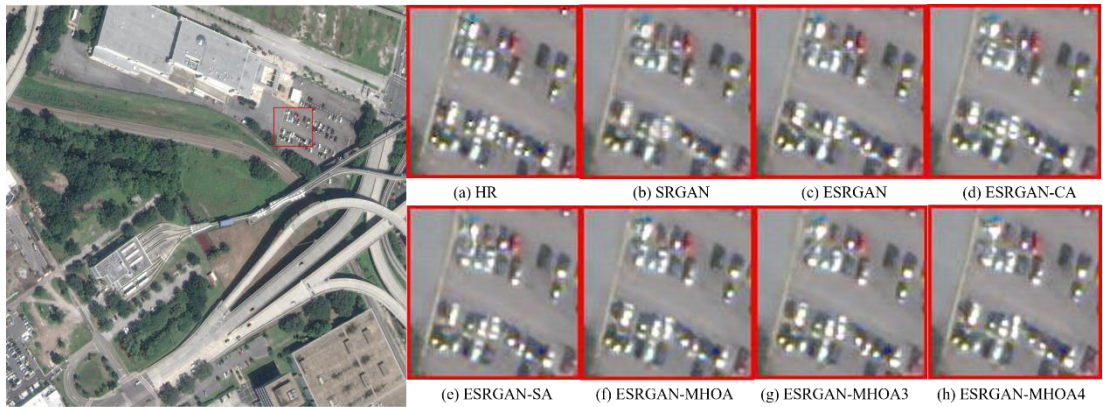


Figure 4.35 : Visual results on DFC images.



Figure 4.36 : Visual results on DFC images.

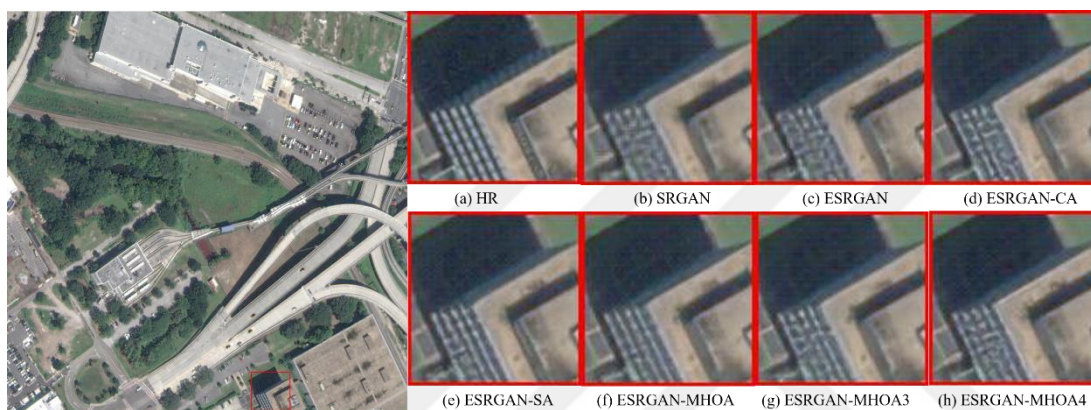


Figure 4.37 : Visual results on DFC images.



Figure 4.38 : Visual results on DFC images.

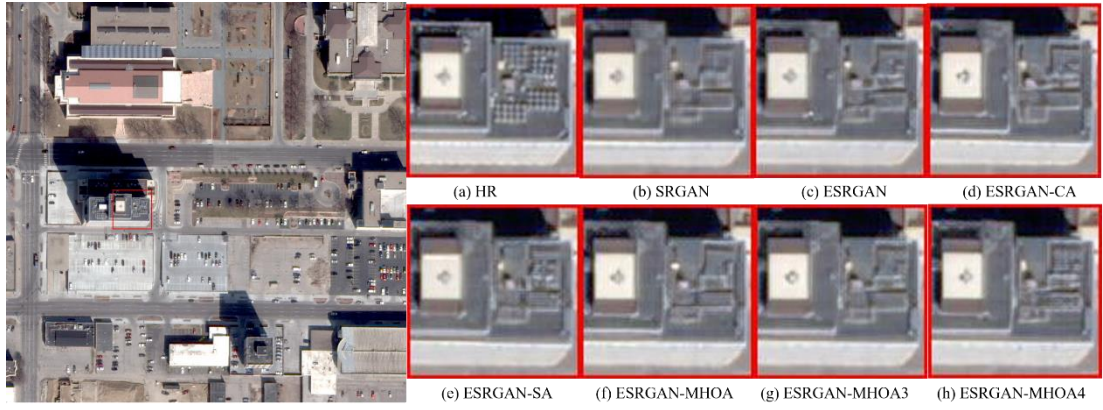


Figure 4.39 : Visual results on DFC images.

As shown in Figures 4.33 to 4.35, the proposed GAN-based method with MHOA attention module (h) shows perceptually good quality in the cars and parking places than other GAN-based methods. Figure 4.36 to 4.39 present the visual results on buildings and roof. It is noted that the presented MHOA attention-based GAN method can provide more texture details of the roofs and lines of the buildings.

4.2.4 Discussion

The results of the attention GAN-based methods indicate that the improved ESRGAN can provide more balanced results both quantitatively and qualitatively by integrating the MHOA attention module in the residual in residual blocks in the ESRGAN. The performance is improved compared to the attention CNN-based methods. This is exactly what we expected. On the other hand, the performance of the ESRGAN is improved when the complexity of the residual in residual blocks is increased.

It is observed that when the attention module is integrated into the dense block in a cascading way, the model doesn't work. But when the attention module is integrated into the dense block in a parallel way, the model can generate the expected results.



5. RESULTS AND DISCUSSIONS OF MULTI-FRAME REMOTE SENSING IMAGE SUPER-RESOLUTION

In this chapter, the results and discussions of the multi-frame RSISR are presented. In detail, the experiment setting, quantitative results, and qualitative results are presented. Finally, the discussions are illustrated.

5.1 Experiment Settings

For the experiments of the MFSR on Remote sensing images, all the experiments are trained on Pytorch 1.0 with NVIDIA GeForce GTX 1060 GPU. The proposed methods are optimized by Adam optimizer with $\beta_1=0.9$, $\beta_2=0.999$ on RGB images.

In the training process, the inputs are clipped as $32*32*3*5$, and the reference images are $128*128*3$. We use 13500 patches for the training, and 10 images with a size of $1024*1024$ for the testing. The images are first trained with L1 loss function for $1e6$ iterations and then trained with GAN for $4e5$ iterations. The learning rate is set to $1e-4$.

5.2 Results Of The MFSR On SpaceNet7 Images

5.2.1 Quantitative results

The numerical assessment values of the proposed methods are illustrated in Table 5.1. For the comparison, we compare the proposed methods with VESPCN, MCRESNET, DRVSR, LTDVSR, DUF, PFNL, MSHPFNL, 3DWDSR, and RAMS. MFESRGAN represents the ESRGAN method trained on the SpaceNet7 dataset. MHOA_MFESRGAN represents the ESRGAN method with the MHOA attention module trained on the SpaceNet7 dataset. MHOA_MSHPFNLGAN denotes the

MSHPPFNLGAN with the MHOA attention module trained on the SpaceNet7 dataset. And, NA_MSHPPFNLGAN denotes the MSHPPFNLGAN with the proposed new attention (NA) module and attention discriminator trained on the SpaceNet7 dataset.

Table 5.1 : Quantitative results of SpaceNet7 images. Red color, blue color, and green color represent the metric value ranked as the 1st, the 2nd and the 3rd, respectively.

	PSNR↑	SSIM↑	AG↑	NIQE↓	PI↓
VESPCN	28.5061	0.9316	4.2631	7.0933	6.5748
MCRESNET	28.6142	0.9331	4.2935	6.8878	6.4354
DRVSR	28.3082	0.9296	4.3400	7.0420	6.5457
LTDVSR	28.4703	0.9310	4.2367	7.1710	6.6163
DUF	29.5899	0.9471	4.4032	6.2014	5.9638
PFNL	30.4158	0.9488	4.4614	6.7528	6.4857
MSHPPFNL	30.2918	0.9471	4.6095	6.2796	6.0827
3DWDSR	29.0106	0.9398	4.3364	6.8518	6.4324
RAMS	28.8917	0.9385	4.3670	6.5724	6.2804
MFESRGAN	27.5740	0.9122	5.1742	5.5933	4.7659
MHOA_MFESRGAN	28.0313	0.9218	4.9740	5.2870	4.6732
MHOA_MSHPPFNLGAN	26.8320	0.9062	5.1787	5.7458	5.1040
NA_MSHPPFNLGAN	27.4735	0.9090	5.4272	4.7496	4.5365

As can be seen from Table 5.1, PFNL, MSHPPFNL, and DUF ranked first, second, and third concerning the PSNR and SSIM, respectively. The presented NA_MSHPPFNLGAN approach outperforms the remaining models regarding AG, NIQE, and PI. MHOA_MSHPPFNLGAN ranked second in terms of AG, and MHOA_MFESRGAN ranked second in terms of NIQE and PI. MFESRGAN ranked third in terms of AG, NIQE, and PI.

5.2.2 Qualitative results

In this subsection, the qualitative results on the SapceNet7 dataset are illustrated as shown in Figures 5.1 to 5.9. As can be seen from Figure 5.1, (g) PFNL and (h) MSHPPFNL are good at preserving the information of the rectangular buildings but with less information on the circle part and the roof structure of the green buildings. However, the proposed GAN-based methods from (j) to (n) provide more details in

the circle areas and the roof structure of the buildings but with low quality of the rectangular buildings.

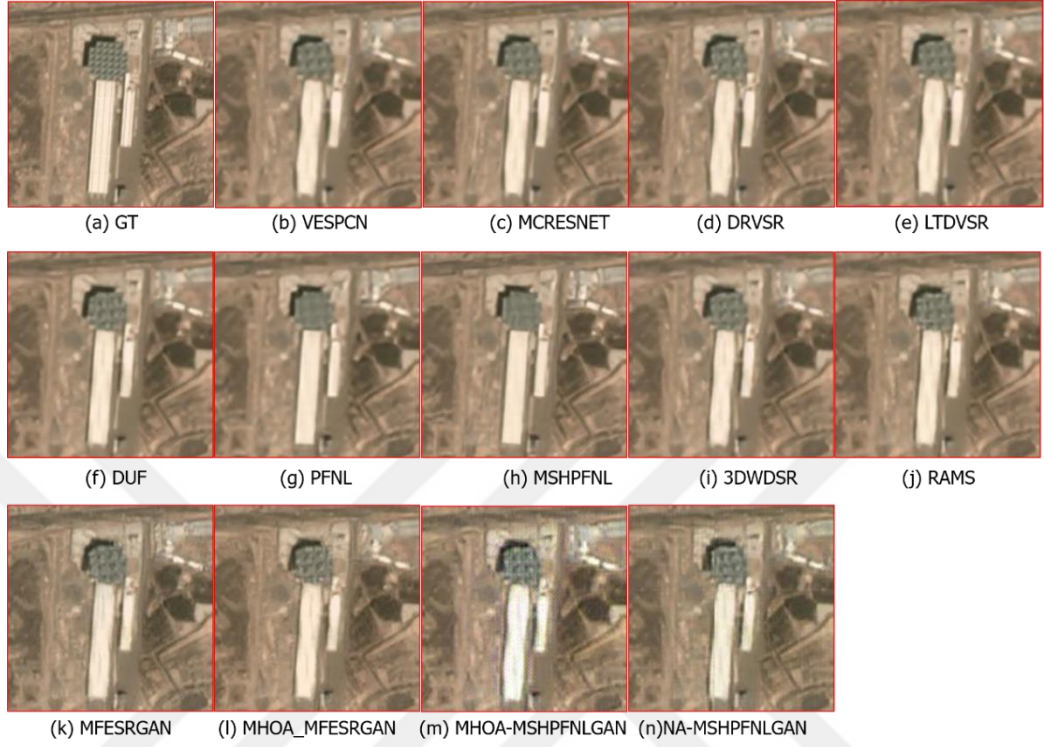


Figure 5.1 : Visual Results on SpaceNet7 images.

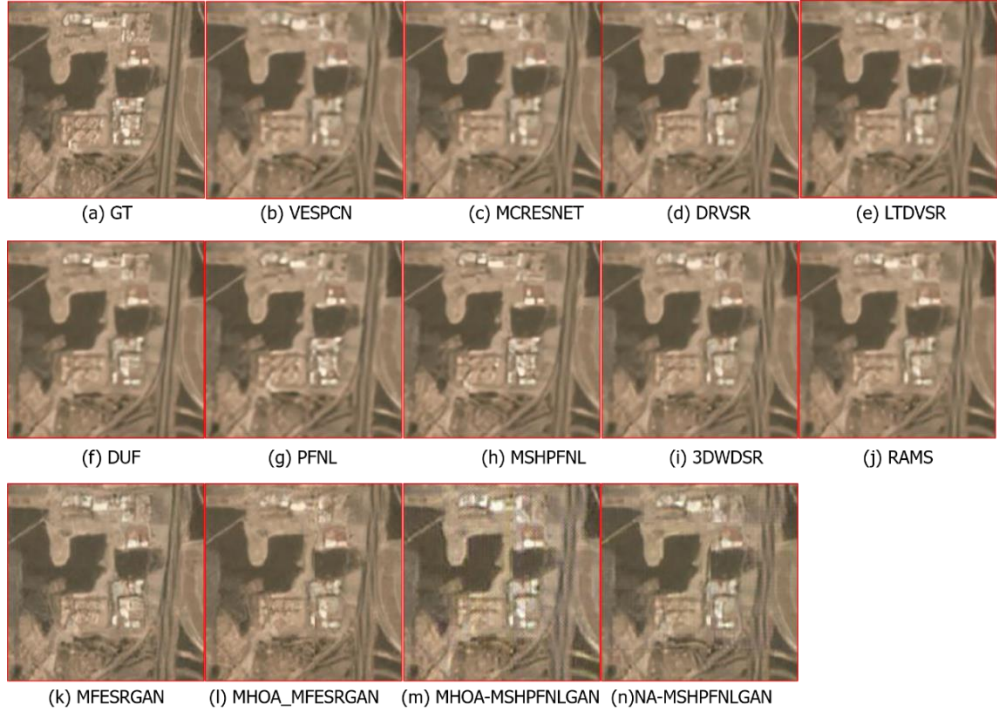


Figure 5.2 : Visual Results on SpaceNet7 images.

As shown in Figure 5.2, (f)-(h) show good ability at preserving the straight lines emphasized by the highway, the GAN-based methods (k)-(n) can provide more information on the building but with artifacts. The same with Figure 5.2, Figure 5.3 to Figure 5.9 have the same observation.

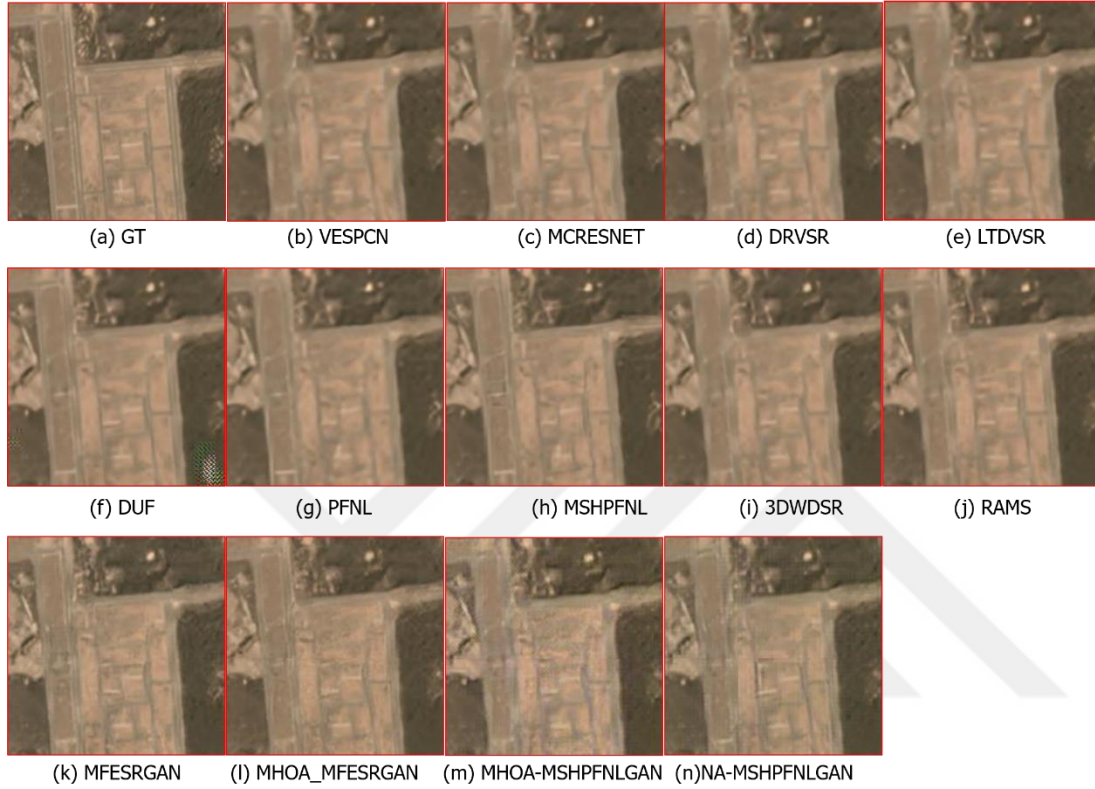


Figure 5.3 : Visual Results on SpaceNet7 images.

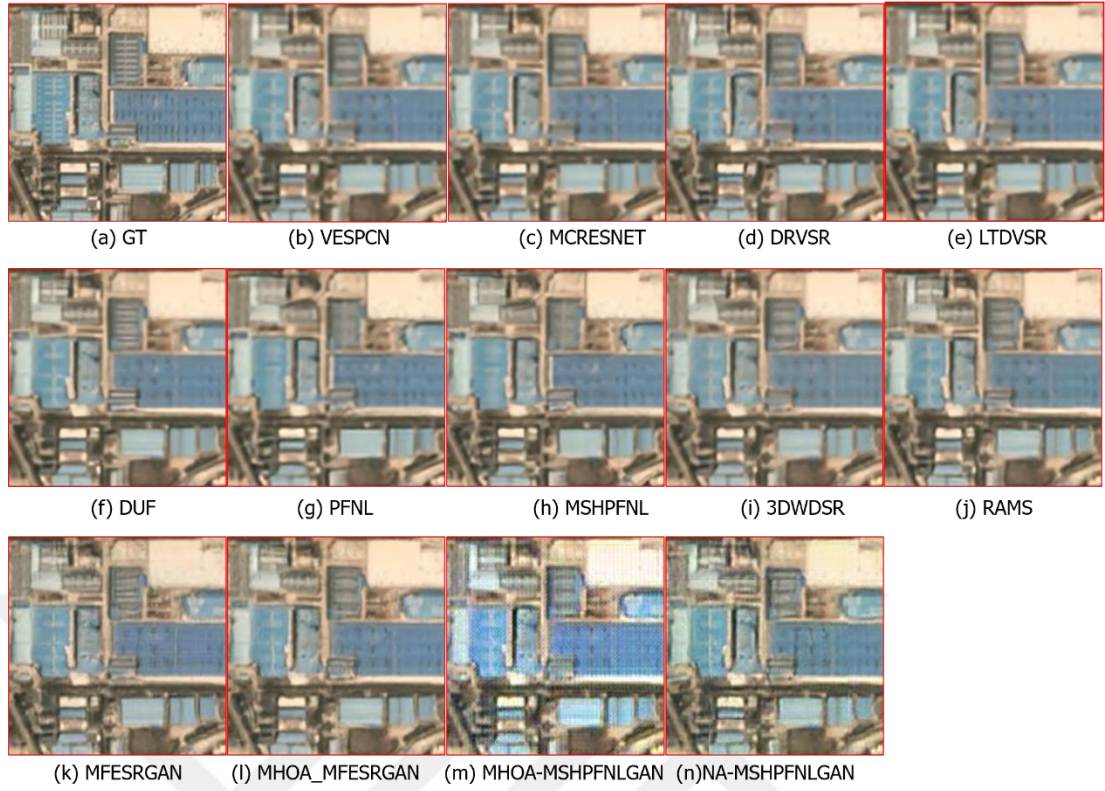


Figure 5.4 : Visual Results on SpaceNet7 images.

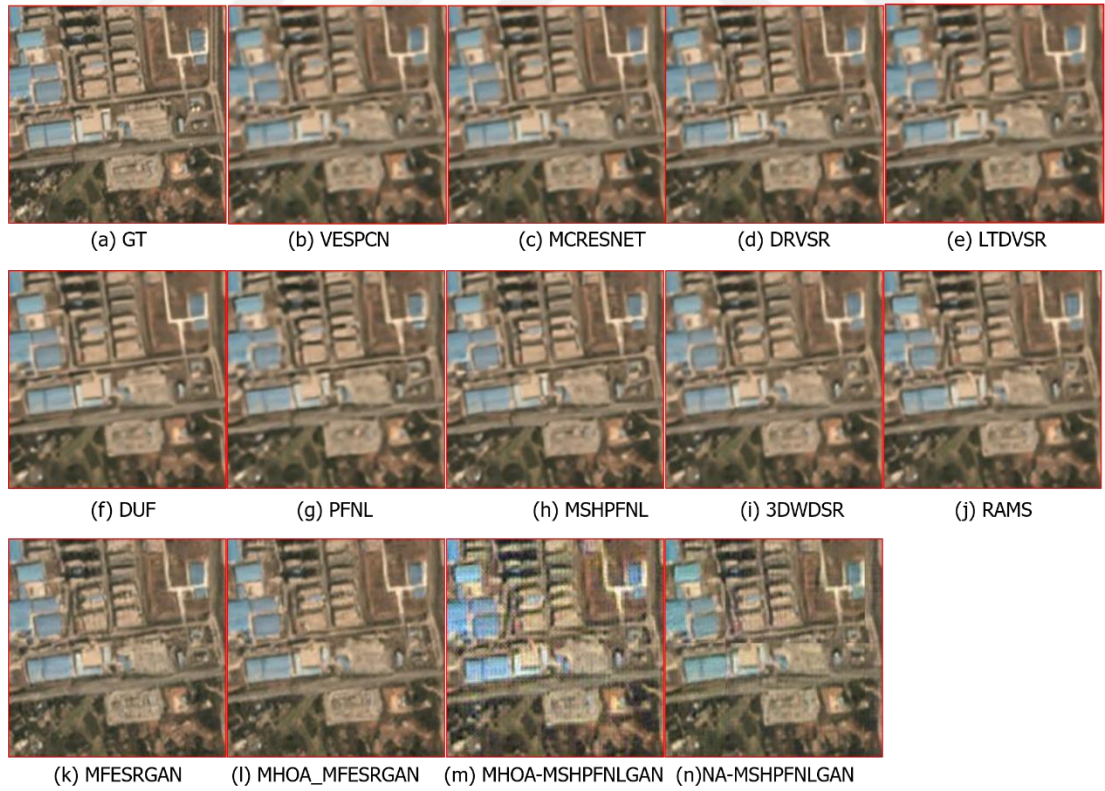


Figure 5.5 : Visual Results on SpaceNet7 images.

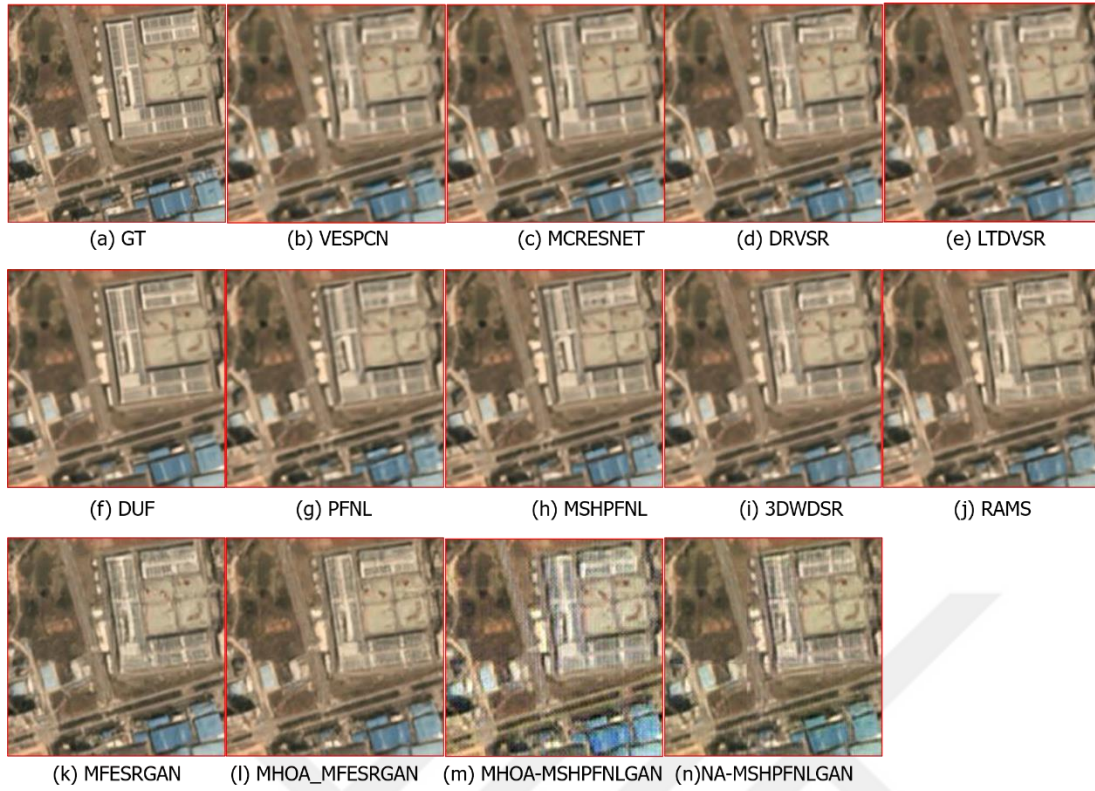


Figure 5.6 : Visual Results on SpaceNet7 images.

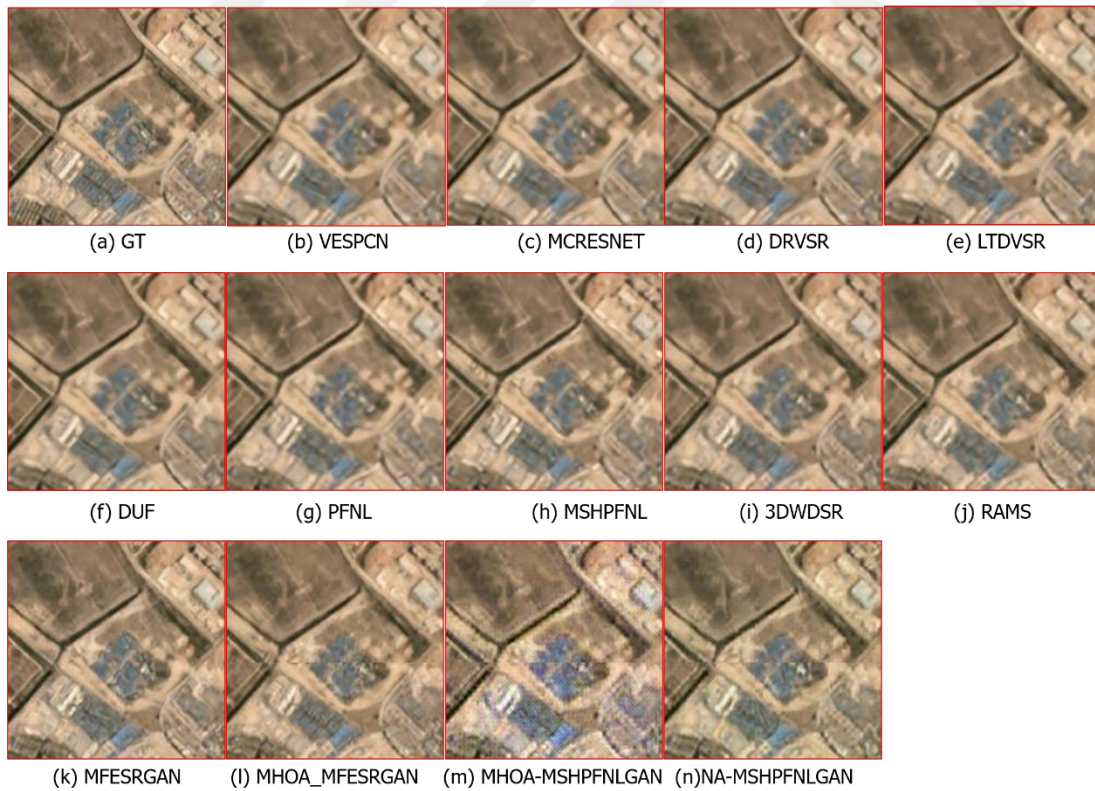


Figure 5.7 : Visual Results on SpaceNet7 images.

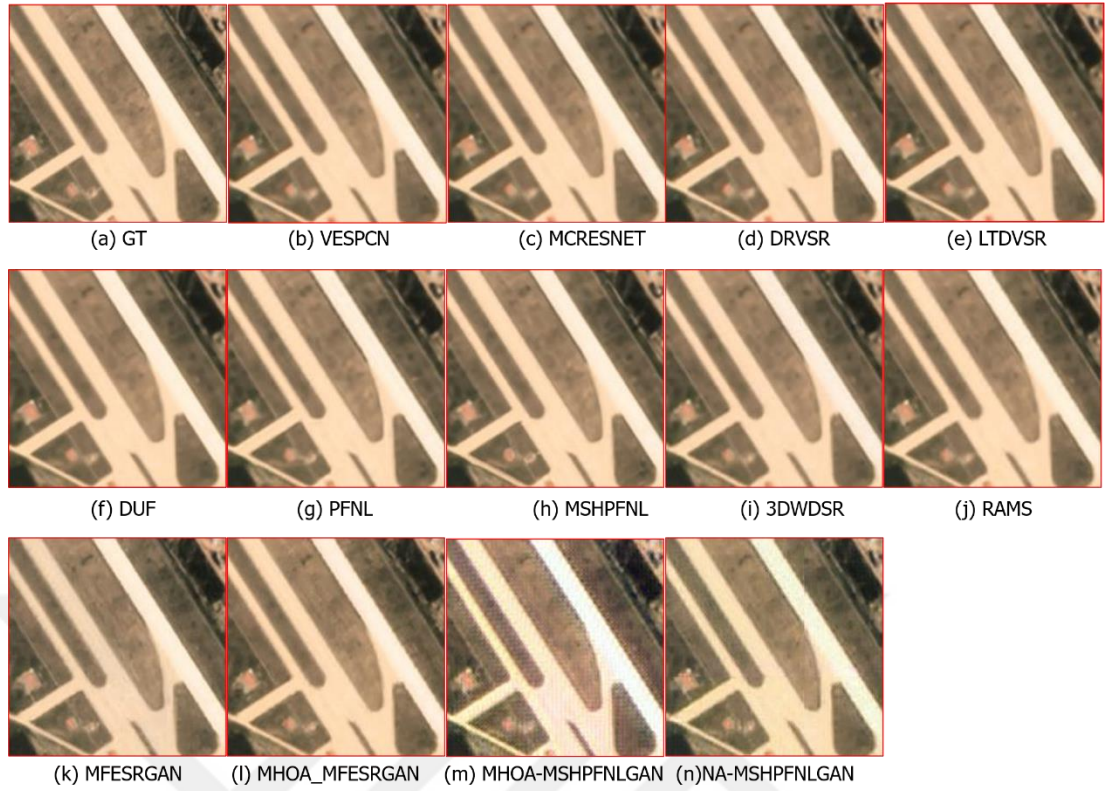


Figure 5.8 : Visual Results on SpaceNet7 images.

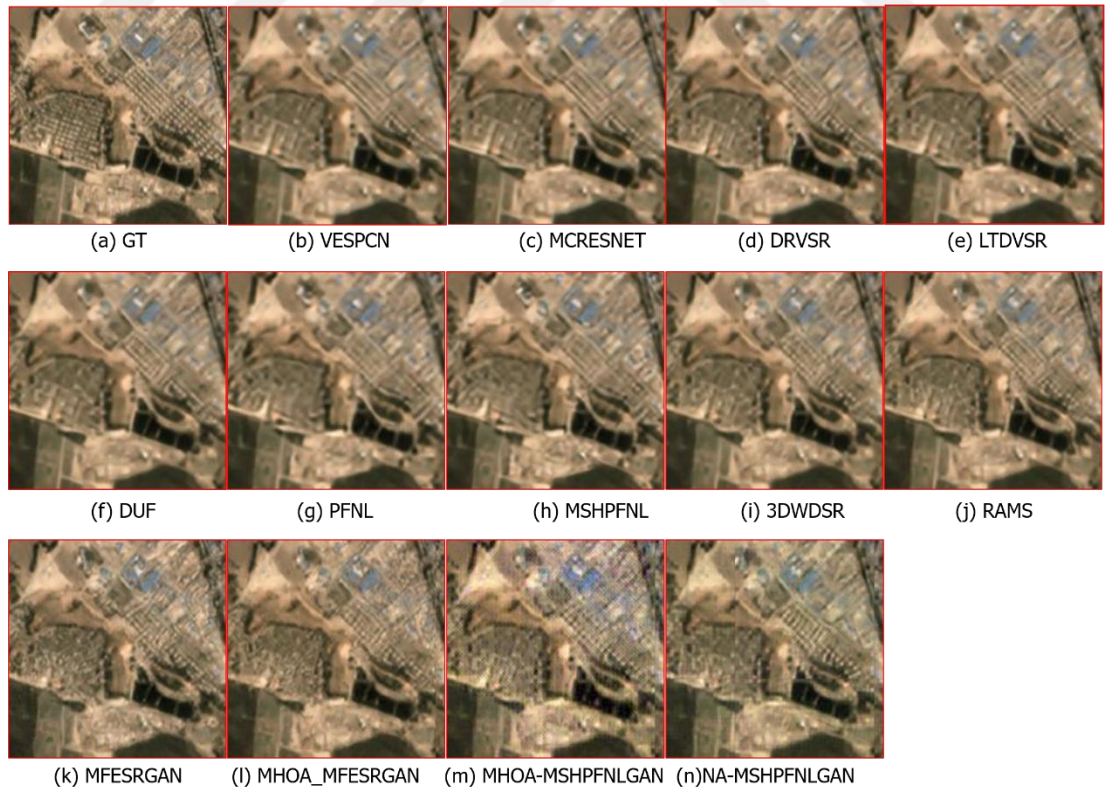


Figure 5.9 : Visual Results on SpaceNet7 images.

5.3 Test results on Real-world Jilin-1 images

In addition, for the evaluation of the GAN methods, we downloaded 4 Jilin-1 videos from ('Jilin-1 Smart Video Satellites', 2017), including Jeddah, Dubai, Boston, and Macao. Then I chose 9 frames from every 20 frames, and cropped the images to 1024×1024 pixels. With a scale factor of 4, we generate the SR images with 4096×4096 pixels. We used the Jilin-1 data set for testing purposes to evaluate the generalization and transferability of the implemented MFRSISR models which were trained with RS images from different satellites (Figures 5.10, 5.11, and 5.12). Since we could obtain only a limited number of Jilin-1 video images, we could not train the model with this data set but used it for testing. This is also important for the evaluation of the MFRSISR models with RS images of different sensors. The reconstructed results illustrate that the proposed NAMSHPFNLGAN method achieves good performance with respect to perceptual quality. MHOAMSHPFNLGAN method produced some slight differences in spectral values.



Figure 5.10 : Visualization of Jilin-1 images.



Figure 5.11 : Visualization of Jilin-1 images.

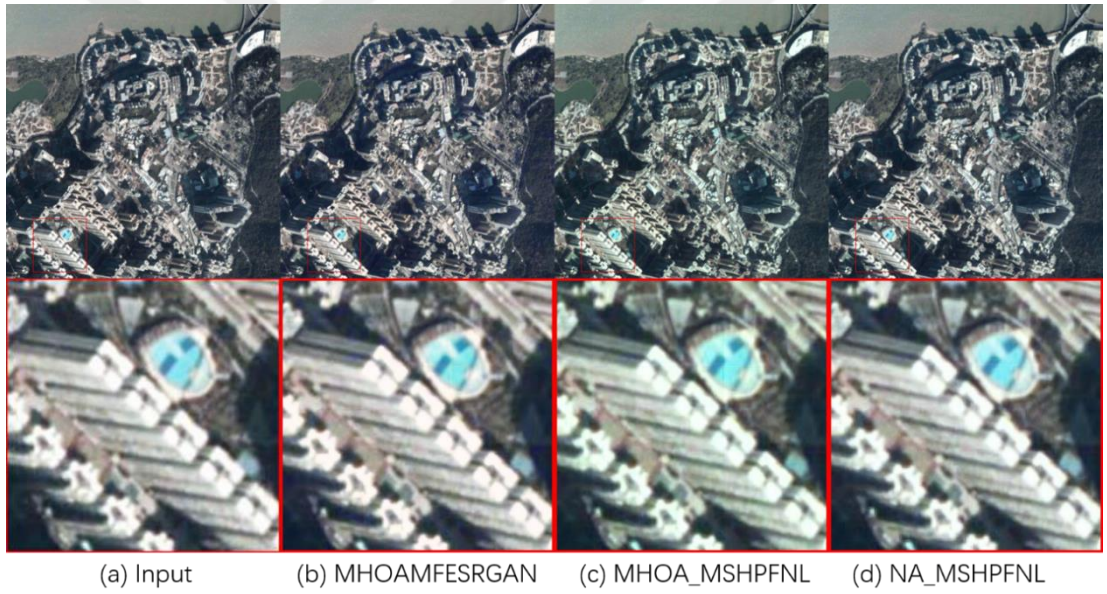


Figure 5.12 : Visualization of Jilin-1 images.

5.4 Discussions

The results of different combinations of Attention modules. In this thesis, the effects of different combinations of different attention modules are explored, and the results are demonstrated in Table 5.2.

Table 5.2 : The results of different combinations of Attention modules. Red color and blue color represent the metric value ranked as the 1st and the 2nd, respectively.

NLSA	NL	MHOA	SA	NA+AD	PSNR↑	SSIM↑	AG↑	NIQE↓	PI↓
	√	√			26.8320	0.9062	5.1787	5.7458	5.1040
	√		√		26.4871	0.8954	5.4315	5.6582	4.9919
	√			√	27.4735	0.9090	5.4272	4.7496	4.5365
√		√			26.5476	0.8999	5.2441	5.4275	5.0615
√			√		26.1108	0.8894	5.5282	5.2815	4.8206
√				√	27.4037	0.9069	5.3702	4.9087	4.6018

As shown in Table 5.2, NLSA and NL represent the non-local sparse attention and non-local attention module, respectively. MHOA is the mixed high order attention. SA is the self-attention module. And, NA+AD denotes the proposed novel attention module plus the attention-based discriminator. As presented in Table 5.2, the NL+NA+AD model outperforms the rest models as to PSNR, SSIM, NIQE, and PI. The NLSA+NA+AD model ranked second with regards to PSNR, SSIM, NIQE, and PI. However, in terms of AG, NLSA+SA and NL+SA ranked first and second, respectively.

The effects of the numbers of the Residual Blocks. The effects of the numbers of the residual blocks are presented in Table 5.3. It is obvious when the number of the residual blocks is set to 10 with the NL attention module, the model achieves the best scores as to NIQE and PI. The model with 40 residual blocks and NL module outperforms the rest approaches with regard to PSNR and SSIM. Regards to AG, the model with 30 residual blocks and NL module performs best.

Table 5.3 : The effects of the numbers of the Residual Blocks. Red color represents the metric value ranked as the 1st.

	Nums of Residual Blocks	PSNR↑	SSIM↑	AG↑	NIQE↓	PI↓
NLSA	10	27.3830	0.9071	5.4128	5.2937	4.8509
	20	27.6303	0.9130	5.2720	5.1289	4.7605
	30	27.4037	0.9069	5.3702	4.9087	4.6018
	40	27.4159	0.9066	5.3954	5.1738	4.7367
NL	10	27.4735	0.9090	5.4272	4.7496	4.5365
	20	27.8003	0.9168	5.1369	5.0579	4.7774
	30	27.3459	0.9042	5.4325	5.0999	4.6708
	40	27.9899	0.9228	5.0689	5.4441	5.1996

The effectiveness of the Attention Discriminator. The effectiveness of the Attention Discriminator is depicted in Table 5.4. As shown in Table 5.4, by applying the attention discriminator, the performance of the model is improved with regard to AG, NIQE, and PI.

Table 5.4 : The effectiveness of the Attention Discriminator. Red color represents the metric value ranked as the 1st.

	PSNR↑	SSIM↑	AG↑	NIQE↓	PI↓
NLSA_MSHPRB_NAGC2	27.6428	0.9128	5.2840	5.3328	4.8255
NLSA_MSHPRB_NAGC2_AD	27.4159	0.9066	5.3954	5.1738	4.7367

The effects of the loss function. To explore the effects of the different loss functions, the perceptual loss of the VGG and LPIPS are trained and the results are listed in Table 5.5.

Table 5.5 : The effects of the loss function. Red color represents the metric value ranked as the 1st.

	loss function	numbers of Residual blocks	PSNR↑	SSIM↑	AG↑	NIQE↓	PI↓
NL	VGG	10	27.4735	0.9090	5.4272	4.7496	4.5365
		20	27.8003	0.9168	5.1369	5.0579	4.7774
		30	27.3459	0.9042	5.4325	5.0999	4.6708
		40	27.9899	0.9228	5.0689	5.4441	5.1996
	LPIPS	10	27.9962	0.9229	4.8150	6.2261	5.4576
		20	28.0166	0.9238	4.7716	5.3837	5.0297
		30	28.0134	0.9233	4.8026	6.0308	5.3823
		40	28.0352	0.9241	4.7726	5.8953	5.3983

As can be seen from Table 5.5, by applying the LPIPS loss function, the metric values increase relating to PSNR and SSIM. However, the AG, NIQE, and PI metric values do not show any improvements.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

In this study, a comprehensive review of the RSISR is presented that shows the RSISR has drawn widespread attention and been made excellent achievements with the speedy growth of the DL techniques. Furthermore, this study explores perceptual-driven methods both on single-frame and multi-frame RSISR.

For single-frame RSISR, firstly, the attention mechanism is integrated into CNN models to better learn the relationship among the channels. Instead of directly integrating the attention mechanism into the CNN model, a LAM module that tries to learn the relationships among the different residual groups is also proposed and is proven to be compelling in the performance improvement of the model.

Secondly, as the GAN-based method can generate more realistic images in SR fields, a GAN-based method - an improved ESRGAN which integrates the mixed high order attention module is proposed for the single-frame RSISR. Generally speaking, the presented attention GAN-based model provides high perceptual quality results both quantitatively and qualitatively.

Moreover, multi-frame RSISR which tries to generate the HR output from a couple of LR inputs is also explored. For the multi-frame RSISR, an attention GAN-based method that tries to learn the deep temporal relationships among the input frames by working on a single frame is proposed. A novel channel-spatial attention module and an attention-based discriminator are also proposed. These mechanisms are proven to be effective experimentally.

With the rapid development of the RS techniques and the wide applications of the RS images, RS images with higher resolution can further improve the performance of road and building extraction, object detection (such as plane or ship detections), land

cover/land use classification, and segmentation. Furthermore, major natural disaster monitoring and agricultural remote sensing fine monitoring can also benefit from the super-resolved RS images.

6.2 Recommendations

Despite the extensive advancements in RSISR with the involvement of DL base methods, there are still some challenges that need to be considered to improve the model results and increase the generalizability and transferability of the proposed approaches. We present a brief outlook of the future work in the following.

Real-world satellite image SR. In RSISR, generally, a reference HR image is not available and the degradation process of the satellite images is needed to generate the HR-LR image pairs to be used in model training. Creating HR-LR image pairs by using the Bicubic interpolation method or Gaussian blurring kernels and noise can not able to capture the realistic spectral and spatial detail that could be identical to the original LR image the corresponding remote sensor would observe.

Large-scale Satellite Image SR. The majority of the existing RSISR approaches are working on the scale factor of $\times 2$, $\times 3$, $\times 4$, and few are working on the larger factor of $\times 8$. The larger the scale factor is, the less information is provided in LR images. Thus, super-resolution with a large-scale (e.g. $\times 8$, $\times 16$) is a challenging task and one of the problems to be settled.

Assessment metrics. For the assessment of image quality, PSNR and SSIM are most widely used by researchers. However, they are not in line with the perceptual quality. For instance, Deep CNN-based methods can provide high performance with PSNR and SSIM but with blurring images. While the GAN-based methods can generate perceptually high-quality images in terms of LPIPS and PI but with low PSNR and SSIM. So, no universal assessment metric is available for both quantitative and visually proper results. Therefore, accurate and effective evaluation metrics for RSISR deserve to be explored in the future.

Applications with other tasks. High-resolution satellite imageries have been widely applied in classification, and object detection. Some researchers have made a positive contribution to improving the performance of object detection (EESRGAN (Rabbi, Ray, Schubert, Chowdhury, & Chao, 2020)) and feature extraction (Qin et al., 2020) by applying RSISR. These studies provide us a perspective for future work and need more exploration in the future.

Design of Network. Despite the outstanding achievements in the RSISR field in the past several years, DL techniques grows with rapid development. Even more recently, new technologies are proposed, such as Transformer-based image super-resolution (Khan et al., 2021)(Yang, Yang, Fu, Lu, & Guo, 2020), Graph Neural Network-based image SR (S. Zhou, Zhang, Zuo, & Loy, 2020), and Involution (Duo Li et al., 2021). We believe these new techniques can promote the improvement of the RSISR image quality.



REFERENCES

- Ahn, N., Kang, B., & Sohn, K.-A.** (2018a). Image Super-resolution via Progressive Cascading Residual Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 791–799). Salt Lake City, UT, USA, 18-22 June.
- Ahn, N., Kang, B., & Sohn, K.** (2018b). Fast , Accurate , and Lightweight with Cascading Residual Network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 252–268). Munich, Germany, 8-14 September.
- Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., & Selva, M.** (2008). Multispectral and Panchromatic Data Fusion Assessment Without Reference. *Photogrammetric Engineering & Remote Sensing*, 74(2), 193–200. Retrieved from <https://doi.org/10.14358/PERS.74.2.193>
- Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., & Zelnik-Manor, L.** (2018). The 2018 PIRM challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. Munich, Germany, 8-14 September. Retrieved from https://doi.org/10.1007/978-3-030-11021-5_21
- Bordone Molini, A., Valsesia, D., Fracastoro, G., & Magli, E.** (2020). DeepSUM: Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5), 3644–3656. Retrieved from <https://doi.org/10.1109/TGRS.2019.2959248>
- Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., & Shi, W.** (2017). Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (pp. 2848–2857). Honolulu, Hawaii, USA, 21-26 July (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2017.304>
- Chang, Y., & Luo, B.** (2019). Bidirectional convolutional LSTM neural network for remote sensing image super-resolution. *Remote Sensing*, 11(20). Retrieved from <https://doi.org/10.3390/rs11202333>
- Chen, A. A., Chai, X., Chen, B., Bian, R., & Chen, Q.** (2018). A Novel Stochastic Stratified Average Gradient Method: Convergence Rate and Its Complexity. In *2018 International Joint Conference on Neural*

Networks (pp. 1–8). Rio de Janeiro, Brazil, 8–13 July (IEEE). Retrieved from <https://doi.org/10.1109/IJCNN.2018.8489564>

- Chen, B., Deng, W., & Hu, J.** (2019). Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 371–381). Seoul, Korea (South), 27 October–2 November (IEEE).
- Chen, J., Zhang, Y., Hu, X., & Chen, C. Y. C.** (2021). Cascading residual–residual attention generative adversarial network for image super resolution. *Soft Computing*, 25(14), 9651–9662. Retrieved from <https://doi.org/10.1007/s00500-021-05730-4>
- Chen, L., Liu, H., Yang, M., Qian, Y., Xiao, Z., & Zhong, X.** (2021). Remote Sensing Image Super-Resolution via Residual Aggregation and Split Attentional Fusion Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 9546–9556. Retrieved from <https://doi.org/10.1109/jstars.2021.3113658>
- Cheng, X., Li, X., & Yang, J.** (2018). SESR : Single Image Super Resolution with Recursive Squeeze and Excitation Networks. In *2018 24th International conference on pattern recognition (ICPR)* (pp. 147–152). Beijing, China, 20–24 August (IEEE).
- Choi, Y., & Kim, Y.** (2020). A No-Reference Super Resolution for Satellite Image Quality Enhancement for KOMPSAT-3. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 220–223. Retrieved from <https://doi.org/10.1109/IGARSS39084.2020.9324422>
- Christian Ledig Ferenc Huszar, L. T.** (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). Honolulu, HI, USA, 21–26 July (IEEE).
- Dai, T., Cai, J., Zhang, Y., Xia, S. T., & Zhang, L.** (2019). Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 11057–11066). Long Beach, CA, USA, 16–17 June (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2019.01132>
- Data Fusion Contest.** (2019). Retrieved from <https://ieee-dataport.org/open-access/data-fusion-contest-2019-dfc2019>
- Deeba, F., Dharejo, F. A., Zhou, Y., Ghaffar, A., Memon, M. H., & Kun, S.** (2020). Single Image Super-Resolution with Application to Remote-Sensing Image. In *2020 Global Conference on Wireless and Optical Technologies, GCWOT 2020*. Malaga, Spain, 6–8 October. Retrieved from <https://doi.org/10.1109/GCWOT49901.2020.9391625>
- Deeba, F., Zhou, Y., Dharejo, F. A., Du, Y., Wang, X., & Kun, S.** (2021). Multi-scale Single Image Super-Resolution with Remote-Sensing Application

- Using Transferred Wide Residual Network. *Wireless Personal Communications*, 120(1), 323–342. Retrieved from <https://doi.org/10.1007/s11277-021-08460-w>
- Deudon, M., Kalaitzis, A., Goytom, I., Arefin, R., Lin, Z., & Sankaran, K.** (2020). HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery. *ArXiv*. Retrieved from <https://doi.org/arXiv:2002.06460>
- Dong, C., Loy, C. C., He, K., & Tang, X.** (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision* (pp. 184–199). Zurich, Switzerland, 6-12 September. Retrieved from https://doi.org/10.1007/978-3-319-10593-2_13
- Dong, C., Loy, C. C., He, K., & Tang, X.** (2016). Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2), 295–307. Retrieved from <https://doi.org/10.1109/TPAMI.2015.2439281>
- Dong, C., Loy, C. C., & Tang, X.** (2016). Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision* (pp. 391–407). Amsterdam, The Netherlands, 8-16 October (Springer, Cham). Retrieved from https://doi.org/10.1007/978-3-319-46475-6_25
- Dong, X., Sun, X., Jia, X., Member, S., Xi, Z., Gao, L., ... Zhang, B.** (2020). Remote Sensing Image Super-Resolution Using Novel Dense-Sampling Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(2), 1618–1633. Retrieved from <https://doi.org/10.1109/TGRS.2020.2994253>
- Dong, X., Wang, L., Sun, X., Jia, X., Gao, L., & Zhang, B.** (2020). Remote Sensing Image Super-Resolution Using Second-Order Multi-Scale Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4), 3473–3485. Retrieved from <https://doi.org/10.1109/TGRS.2020.3019660>
- Dong, X., Xi, Z., & Sun, X.** (2019). Transferred Multi-Perception Attention Networks for Remote Sensing Image Super-Resolution. *Remote Sensing*, 11(23), 2857. Retrieved from <https://doi.org/10.3390/rs11232857>
- Dong, X., Xi, Z., Sun, X., & Yang, L.** (2020). REMOTE SENSING IMAGE SUPER-RESOLUTION VIA ENHANCED BACK-PROJECTION NETWORKS. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium* (pp. 1480–1483).
- Dorr, F.** (2020). Satellite image multi-frame super resolution using 3d wide-activation neural networks. *Remote Sensing*, 12(22), 1–10. Retrieved from <https://doi.org/10.3390/rs12223812>
- Du, X., Jiang, S., Si, Y., Xu, L., & Liu, C.** (2021). Mixed high-order non-local attention network for single image super-resolution. *IEEE Access*, 9(1),

49514–49521. Retrieved from
<https://doi.org/10.1109/ACCESS.2021.3069777>

- Fu, J., Liu, Y., & Li, F.** (2018). Single frame super resolution with convolutional neural network for remote sensing imagery. In *International Geoscience and Remote Sensing Symposium (IGARSS)* (Vol. 2018-July, pp. 8014–8017). IEEE. Retrieved from <https://doi.org/10.1109/IGARSS.2018.8518584>
- Gao, L., Sun, H. H., Cui, Z., Du, Y., & Sun, H. H.** (2021). Super-resolution reconstruction of single remote sensing images based on residual channel attention, 15(1), 1–15. Retrieved from <https://doi.org/10.1117/1.JRS.15.016513>
- Gong, Y., Liao, P., Zhang, X., Zhang, L., Chen, G., Zhu, K., ... Lv, Z.** (2021). Enlighten-GAN for Super Resolution Reconstruction in Mid-Resolution Remote Sensing Images. *Remote Sensing*, 13(6), 1104. Retrieved from <https://doi.org/10.3390/rs13061104>
- Goodfellow, I. J., & Pouget-Abadie, J.** (2014). Generative Adversarial Nets. *Veterinary Immunology and Immunopathology*, 155(4), 270–275. Retrieved from <https://doi.org/10.1016/j.vetimm.2013.08.005>
- Gu, J., Sun, X., Zhang, Y., Fu, K., & Wang, L.** (2019). Deep residual squeeze and excitation network for remote sensing image super-resolution. *Remote Sensing*, 11(15), 1–16. Retrieved from <https://doi.org/10.3390/rs11151817>
- Guo, D., Xia, Y., Xu, L., Li, W., & Luo, X.** (2021). Remote sensing image super-resolution using cascade generative adversarial nets. *Neurocomputing*, 443, 117–130. Retrieved from <https://doi.org/10.1016/j.neucom.2021.02.026>
- Hara, M., & Tanaka, K.** (2016). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. [*Nihon Kōshū Eisei Zasshi*] *Japanese Journal of Public Health*, 60(8), 444–52. Retrieved from <https://doi.org/10.1109/TPAMI.2015.2439281>
- Haris, M., Shakhnarovich, G., & Ukita, N.** (2018). Deep Back-Projection Networks for Super-Resolution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1664–1673). Salt Lake City, UT, USA, 18-22 June. (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2018.00179>
- Haut, J. M., Fernandez-beltran, R., Paoletti, M. E., Plaza, J., & Plaza, A.** (2019). Remote Sensing Image Superresolution Using Deep Residual Channel Attention. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11), 9277–9289. Retrieved from <https://doi.org/10.1109/TGRS.2019.2924818>
- Haut, J. M., Fernandez-Beltran, R., Paoletti, M. E., Plaza, J., Plaza, A., & Pla, F.**

- (2018). A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11), 6792–6810. Retrieved from <https://doi.org/10.1109/TGRS.2018.2843525>
- He, Z., Li, J., Liu, L., He, D., & Xiao, M.** (2022). Multiframe Video Satellite Image Super-Resolution via Attention-Based Residual Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60. Retrieved from <https://doi.org/10.1109/TGRS.2021.3072381>
- Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E.** (2020). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), 2011–2023. Retrieved from <https://doi.org/10.1109/TPAMI.2019.2913372>
- Huan, H., Li, P., Zou, N., Wang, C., Xie, Y. Y., Xie, Y. Y., & Xu, D.** (2021). End-to-End Super-Resolution for Remote-Sensing Images Using an Improved Multi-Scale Residual Network. *Remote Sensing*, 13(666).
- Huang, B., He, B., Wu, L., & Guo, Z.** (2021). Deep Residual Dual-Attention Network for Super-Resolution Reconstruction of Remote Sensing Images. *Remote Sensing*, 13(14). Retrieved from <https://doi.org/10.3390/rs13142784>
- Huang, N., Yang, Y., Liu, J., Gu, X., & Cai, H.** (2017). Single-Image Super-Resolution For remote sensing data using Deep Residual-learning Neural Network. In *International Conference on Neural Information Processing* (pp. 118–125). Guangzhou, China, 14-18 November. Retrieved from <https://doi.org/10.1007/978-3-319-70096-0>
- Huang, Z. X., & Jing, C. W.** (2020). Super-Resolution Reconstruction Method of Remote Sensing Image Based on Multi-Feature Fusion. *IEEE Access*, 8, 18764–18771. Retrieved from <https://doi.org/10.1109/ACCESS.2020.2967804>
- Jetley, S., Lord, N. A., Lee, N., & Torr, P. H. S.** (2018). Learn to pay attention. In *6th International Conference on Learning Representations* (pp. 1–14). Vancouver, CANADA, 30 April- 3 May (Springer).
- Jia, S., Wang, Z., Jia, X., & Xu, M.** (2022). Multi-Attention Generative Adversarial Network for Remote Sensing Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15. Retrieved from <https://doi.org/10.1109/tgrs.2022.3180068>
- Jiang, K., Wang, Z., Yi, P., & Jiang, J.** (2018). A Progressively Enhanced Network for Video Satellite Imagery Superresolution. *IEEE Signal Processing Letters*, 25(11), 1630–1634. Retrieved from <https://doi.org/10.1109/LSP.2018.2870536>
- Jiang, K., Wang, Z., Yi, P., Jiang, J., Xiao, J., & Yao, Y.** (2018). Deep distillation recursive network for remote sensing imagery super-resolution. *Remote*

Sensing, 10(11), 1–23. Retrieved from <https://doi.org/10.3390/rs10111700>

Jiang, K., Wang, Z., Yi, P., Wang, G., Lu, T., & Jiang, J. (2019). Edge-Enhanced GAN for Remote Sensing Image Superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8), 5799–5812. Retrieved from <https://doi.org/10.1109/tgrs.2019.2902431>

Jilin-1 Smart Video Satellites. (2017). Retrieved from <https://www.satimagingcorp.com/satellite-sensors/jilin-1-satellite-sensor-1m/jilin-1-smart-video-satellite/>

Jo, Y., Oh, S. W., Kang, J., & Kim, S. J. (2018). Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA, 18-22 June (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2018.00340>

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711). Amsterdam, The Netherlands, 8-16 October (Springer, Cham). Retrieved from https://doi.org/10.1007/978-3-319-46475-6_43

Jolicoeur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard GAN. In *International Conference on Learning Representations*. Vancouver, CANADA, 30 April-3 May. Retrieved from <http://arxiv.org/abs/1807.00734>

Kappeler, A., Yoo, S., Dai, Q., & Katsaggelos, A. K. (2016). Video Super-Resolution With Convolutional Neural Networks. *IEEE Transactions on Computational Imaging*, 2(2), 109–122. Retrieved from <https://doi.org/10.1109/tci.2016.2532323>

Keshk, H. M., & Yin, X. C. (2020). Obtaining Super-Resolution Satellites Images Based on Enhancement Deep Convolutional Neural Network. *International Journal of Aeronautical and Space Sciences*, 22(1), 195–202. Retrieved from <https://doi.org/10.1007/s42405-020-00297-0>

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in Vision: A Survey. *ACM Computing Surveys (CSUR)*. Retrieved from <http://arxiv.org/abs/2101.01169>

Kim, J., Lee, J. K., & Lee, K. M. (2016). Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1637–1645). Las Vegas, NV, USA, 27-30 June (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2016.181>

Lai, W. S., Huang, J. Bin, Ahuja, N., & Yang, M. H. (2017). Deep laplacian pyramid

- networks for fast and accurate super-resolution. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (pp. 5835–5843). Honolulu, HI, USA, 21-26 July (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2017.618>
- Lan, R., Sun, L., Liu, Z., Lu, H., Su, Z., Pang, C., & Luo, X.** (2020). Cascading and Enhanced Residual Networks for Accurate Single-image Super-resolution. *IEEE Transactions on Cybernetics*, PP, 1–11. Retrieved from <https://doi.org/10.1109/TCYB.2019.2952710>
- Lanaras, C., Bioucas-dias, J., Galliani, S., & Baltsavias, E.** (2018). Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146(April), 305–319. Retrieved from <https://doi.org/10.1016/j.isprsjprs.2018.09.018>
- Lei, S., Shi, Z., & Zou, Z.** (2017). Super-resolution for remote sensing images via local-global combined network. *IEEE Geoscience and Remote Sensing Letters*, 14(8), 1243–1247. Retrieved from <https://doi.org/10.1109/LGRS.2017.2704122>
- Lei, S., Shi, Z., Zou, Z., Super-resolution, I., Lei, S., Shi, Z., & Zou, Z.** (2019). Coupled Adversarial Training for Remote Sensing Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5), 3633–3643. Retrieved from <https://doi.org/10.1109/TGRS.2019.2959020>
- Li, Dingyi, & Wang, Z.** (2017). Video Superresolution via Motion Compensation and Deep Residual Learning. *IEEE Transactions on Computational Imaging*, 3(4), 749–762. Retrieved from <https://doi.org/10.1109/tci.2017.2671360>
- Li, Duo, Hu, J., Wang, C., Li, X., She, Q., Zhu, L., ... Chen, Q.** (2021). Involution: Inverting the Inherence of Convolution for Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12321–12330). Nashville, TN, USA, 20-25 June. Retrieved from <http://arxiv.org/abs/2103.06255>
- Li, X., Zhang, D., Liang, Z., Ouyang, D., & Shao, J.** (2020). Fused recurrent network via channel attention for remote sensing satellite image super-resolution. In *Proceedings - IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). London, United Kingdom, 6-10 July (IEEE). Retrieved from <https://doi.org/10.1109/ICME46284.2020.9102948>
- Li, Yadong, Mavromatis, S., Zhang, F., Du, Z., Sequeira, J., Wang, Z., ... Liu, R.** (2021). Single-Image Super-Resolution for Remote Sensing Images Using a Deep Generative Adversarial Network With Local and Global Attention Mechanisms. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–24. Retrieved from

<https://doi.org/10.1109/TGRS.2021.3093043>

- Li, Yinhao, Iwamoto, Y., Lin, L., & Chen, Y. W.** (2020). Parallel-Connected Residual Channel Attention Network for Remote Sensing Image Super-Resolution. In *Proceedings of the Asian Conference on Computer Vision*. 30 November- 4 December. Retrieved from https://doi.org/10.1007/978-3-030-69756-3_2
- Liebel, L., & Körner, M.** (2016). Single-image super resolution for multispectral remote sensing data using convolutional neural networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41(July), 883–890. Retrieved from <https://doi.org/10.5194/isprsarchives-XLI-B3-883-2016>
- Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M.** (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1132–1140). Honolulu, Hawaii, USA, 21–26 July (IEEE). Retrieved from <https://doi.org/10.1109/CVPRW.2017.151>
- Liu, B., Li, H., Zhou, Y., Peng, Y., Elazab, A., & Wang, C.** (2020). A super resolution method for remote sensing images based on cascaded conditional wasserstein GANs. In *2020 3rd IEEE International Conference on Information Communication and Signal Processing, ICICSP 2020* (pp. 284–289). Shanghai, China, 12-15 September. Retrieved from <https://doi.org/10.1109/ICICSP50920.2020.9232066>
- Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., & Huang, T.** (2017). Robust Video Super-Resolution with Learned Temporal Dynamics. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2526–2534). Venice, Italy, 22-29 October (IEEE). Retrieved from <https://doi.org/10.1109/ICCV.2017.274>
- Liu, H., Ruan, Z., Zhao, P., Shang, F., Yang, L., & Liu, Y.** (2020). Video super resolution based on deep learning: A comprehensive survey. *Artificial Intelligence Review*, 1–55.
- Liu, J. G., Wald, L., & Ranchin, T.** (2000). Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 1161. Retrieved from <https://doi.org/10.1080/014311600750037499>
- 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), 1–19. Retrieved from <https://doi.org/10.3390/rs11131588>
- Luo, Y., Zhou, L., Wang, S., & Wang, Z.** (2017). Video Satellite Imagery Super Resolution via Convolutional Neural Networks. *IEEE GEOSCIENCE AND REMOTE SENSING LETTERS*, 14(12), 2398–2402.

- Ma, C., Yang, C. Y., Yang, X., & Yang, M. H.** (2017). Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158, 1–16. Retrieved from <https://doi.org/10.1016/j.cviu.2016.12.009>
- Ma, W., Pan, Z., Guo, J., & Lei, B.** (2018). Super-resolution of remote sensing images based on transferred generative adversarial network. In *International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 1148–1151). Valencia, Spain, 22-27 July (IEEE). Retrieved from <https://doi.org/10.1109/IGARSS.2018.8517442>
- Ma, W., Pan, Z., Guo, J., & Lei, B.** (2019). Achieving Super-Resolution Remote Sensing Images via the Wavelet Transform Combined with the Recursive Res-Net. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6), 3512–3527. Retrieved from <https://doi.org/10.1109/TGRS.2018.2885506>
- Ma, W., Pan, Z., Yuan, F., & Lei, B.** (2019). Super-Resolution of Remote Sensing Images via a Dense Residual Generative Adversarial Network. *Remote Sensing*, 11(21), 1578. Retrieved from <https://doi.org/10.3390/rs11212578>
- Ma, Y., Lv, P., Liu, H., Sun, X., & Zhong, Y.** (2021). Remote sensing image super-resolution based on dense channel attention network. *Remote Sensing*, 13(15), 1–20. Retrieved from <https://doi.org/10.3390/rs13152966>
- Mao, X.-J., Shen, C., & Yang, Y.-B.** (2016). Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections. *ArXiv*. Retrieved from <http://arxiv.org/abs/1606.08921>
- Mei, Y., Fan, Y., & Zhou, Y.** (2021). Image Super-Resolution with Non-Local Sparse Attention. In *Conference on Computer Vision and Pattern Recognition* (pp. 3517–3526). Nashville, TN, USA, 20-25 June.
- Mittal, A., Soundararajan, R., & Bovik, A. C.** (2013). Making a ‘completely blind’ image quality analyzer. *IEEE Signal Processing Letters*, 20(3), 209–212. Retrieved from <https://doi.org/10.1109/LSP.2012.2227726>
- Moustafa, M. S., & Sayed, S. A.** (2021). Satellite Imagery Super-Resolution Using Squeeze-and-Excitation-Based GAN. *International Journal of Aeronautical and Space Sciences*, 22(6), 1481–1492. Retrieved from <https://doi.org/10.1007/s42405-021-00396-6>
- Müller, M. U., Ekhtiari, N., Almeida, R. M., & Rieke, C.** (2020). SUPER-RESOLUTION OF MULTISPECTRAL SATELLITE IMAGES USING CONVOLUTIONAL NEURAL NETWORKS. *ArXiv*. Retrieved from <https://doi.org/arXiv:2002.00580>
- Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., ... Shen, H.** (2020). Single Image Super-Resolution via a Holistic Attention Network. In *European conference on computer vision* (pp. 191–207). Glasgow,

United Kingdom, 23-28 August. Retrieved from https://doi.org/10.1007/978-3-030-58610-2_12

- Pan, Z., Ma, W., Guo, J., & Lei, B.** (2019). Super-Resolution of Single Remote Sensing Image Based on Residual Dense Backprojection Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10), 1–16. Retrieved from <https://doi.org/10.1109/tgrs.2019.2917427>
- Peng, Y., Wang, X., Zhang, J., & Liu, S.** (2021). Pre-training of gated convolution neural network for remote sensing image super-resolution. *IET Image Processing*, 15(5), 1179–1188. Retrieved from <https://doi.org/10.1049/ipr2.12096>
- Pineda, F., Ayma, V., & Beltran, C.** (2020). A generative adversarial network approach for super-resolution of sentinel-2 satellite images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 43(B1), 9–14. Retrieved from <https://doi.org/10.5194/isprs-archives-XLIII-B1-2020-9-2020>
- pytorch.** (n.d.). Retrieved 7 January 2019, from <https://pytorch.org/>
- Qin, M., Hu, L., Du, Z., Gao, Y., Qin, L., Zhang, F., & Liu, R.** (2020). Achieving higher resolution lake area from remote sensing images through an unsupervised deep learning super-resolution method. *Remote Sensing*, 12(12). Retrieved from <https://doi.org/10.3390/rs12121937>
- R., S. H., & Bovik, A. C.** (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2), 430–444.
- Rabbi, J., Ray, N., Schubert, M., Chowdhury, S., & Chao, D.** (2020). Small-Object Detection in Remote Sensing Images with End-to-End Edge-Enhanced GAN and Object Detector Network. *Remote Sensing*, 12(9), 1432. Retrieved from <https://doi.org/https://doi.org/10.3390/rs12091432>
- Ren, C., He, X., Qing, L., Wu, Y., & Pu, Y.** (2021). Remote sensing image recovery via enhanced residual learning and dual-luminance scheme. *Knowledge-Based Systems*, 222, 107013. Retrieved from <https://doi.org/10.1016/j.knosys.2021.107013>
- Salvetti, F., Mazzia, V., Khaliq, A., & Chiaberge, M.** (2020). Multi-image super resolution of remotely sensed images using residual attention deep neural networks. *Remote Sensing*, 12(14), 2207. Retrieved from <https://doi.org/10.3390/rs12142207>
- Shao, Z., Wang, L., Wang, Z., & Deng, J.** (2019). Remote Sensing Image Super-Resolution Using Sparse Representation and Coupled Sparse Autoencoder. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8), 2663–2674. Retrieved from <https://doi.org/10.1109/JSTARS.2019.2925456>

- Sheikholeslami, M. M., Nadi, S., Naeini, A. A., & Ghamisi, P.** (2020). An efficient deep unsupervised superresolution model for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1937–1945. Retrieved from <https://doi.org/10.1109/JSTARS.2020.2984589>
- Shen, H., Qiu, Z., Yue, L., & Zhang, L.** (2021). Deep Learning Based Super-Resolution of Video Satellite Imagery by the Coupling of Multi-Frame and Single-Frame Models. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14. Retrieved from <https://doi.org/10.1109/TGRS.2021.3121303>
- 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). Las Vegas, NV, USA, 26 June-1 July (IEEE). Retrieved from <http://arxiv.org/abs/1609.05158>
- SpaceNet7.** (2020). Retrieved from <https://spacenet.ai/sn7-challenge/>
- Sustika, R., Suksmono, A. B., Danudirdjo, D., & Wikantika, K.** (2020). Generative Adversarial Network with Residual Dense Generator for Remote Sensing Image Super Resolution. In *Proceeding - 2020 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications, ICRAMET 2020* (pp. 34–39). Tangerang, Indonesia, 18-20 November. Retrieved from <https://doi.org/10.1109/ICRAMET51080.2020.9298648>
- Tai, Y., Yang, J., & Liu, X.** (2017). Image super-resolution via deep recursive residual network. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (pp. 2790–2798). Honolulu, Hawaii, USA, 21-26 July (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2017.298>
- Tao, X., Gao, H., Liao, R., Wang, J., & Jia, J.** (2017). Detail-Revealing Deep Video Super-Resolution. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4482–4490). Venice, Italy, 22-29 October (IEEE). Retrieved from <https://doi.org/10.1109/ICCV.2017.479>
- Tian, Y., Zhang, Y., Fu, Y., & Xu, C.** (2020). TDAN: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3357–3366). Seattle, Washington, USA, 14-19 June (IEEE). Retrieved from <https://doi.org/10.1109/CVPR42600.2020.00342>
- Tong, T., Li, G., Liu, X., & Gao, Q.** (2017). Image Super-Resolution Using Dense

Skip Connections. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4809–4817). Venice, Italy, 22-29 October (IEEE). Retrieved from <https://doi.org/10.1109/ICCV.2017.514>

Tuna, C., Unal, G., & Sertel, E. (2018). Single-frame super resolution of remote-sensing images by convolutional neural networks. *International Journal of Remote Sensing*, 39(8), 2463–2479. Retrieved from <https://doi.org/10.1080/01431161.2018.1425561>

Vaswani, A. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Wagner, L., Liebel, L., & Körner, M. (2019). DEEP RESIDUAL LEARNING for SINGLE-IMAGE SUPER-RESOLUTION of MULTI-SPECTRAL SATELLITE IMAGERY. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4(2/W7), 189–196. Retrieved from <https://doi.org/10.5194/isprs-annals-IV-2-W7-189-2019>

Wang, H., Hu, Q., Wu, C., Chi, J., & Yu, X. (2020). Non-Locally up-Down Convolutional Attention Network for Remote Sensing Image Super-Resolution. *IEEE Access*, 8, 166304–166319. Retrieved from <https://doi.org/10.1109/access.2020.3022882>

Wang, Jiaming, Shao, Z., Huang, X., Lu, T., Zhang, R., & Ma, J. (2021). Enhanced image prior for unsupervised remoting sensing super-resolution. *Neural Networks*, 143, 400–412. Retrieved from <https://doi.org/10.1016/j.neunet.2021.06.005>

Wang, Jiaming, Shao, Z., Lu, T., Huang, X., Zhang, R., & Wang, Y. (2021). Unsupervised Remoting Sensing Super-Resolution via Migration Image Prior. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). Taipei, Taiwan, 18-22 July (IEEE). Retrieved from <https://doi.org/10.1109/icme51207.2021.9428093>

Wang, Jin, Wu, Y., Wang, L., Wang, L., Alfarraj, O., & Tolba, A. (2021). Lightweight Feedback Convolution Neural Network for Remote Sensing Images Super-Resolution. *IEEE Access*, 9, 15992–16003. Retrieved from <https://doi.org/10.1109/ACCESS.2021.3052946>

Wang, Peijuan, Bayram, B., & Sertel, E. (2021). Super-resolution of remotely sensed data using channel attention based deep learning approach. *International Journal of Remote Sensing*, 42(16), 6050–6067. Retrieved from <https://doi.org/10.1080/01431161.2021.1934598>

Wang, Peijuan, & Sertel, E. (2021). Channel–spatial attention-based pan-sharpening of very high-resolution satellite images. *Knowledge-Based Systems*, 229, 107324. Retrieved from <https://doi.org/10.1016/j.knosys.2021.107324>

- Wang, Pengrui, Zhang, H., Zhou, F., & Jiang, Z.** (2019). Unsupervised Remote Sensing Image Super-Resolution Using Cycle CNN. In *International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 3117–3120). Yokohama, Japan, 28 July-2 August (IEEE). Retrieved from <https://doi.org/10.1109/IGARSS.2019.8898648>
- Wang, T., Sun, W., Qi, H., & Ren, P.** (2018). Aerial Image Super Resolution via Wavelet Multiscale Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 769–773. Retrieved from <https://doi.org/10.1109/LGRS.2018.2810893>
- Wang, Xiaolong, Girshick, R., Gupta, A., & He, K.** (2018). Non-local Neural Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803). Salt Lake City, UT, USA, 18-23 June (IEEE). Retrieved from <https://doi.org/10.1112/s0010437x21007144>
- Wang, Xintao, Chan, K. C. K., Yu, K., Dong, C., & Loy, C. C.** (2019). EDVR: Video restoration with enhanced deformable convolutional networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Long Beach, CA, USA, 16-17 June. Retrieved from <https://doi.org/10.1109/CVPRW.2019.00247>
- Wang, Xintao, Yu, K., Wu, S., Gu, J., & Liu, Y.** (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*. Munich, Germany, 8-14 September.
- Wang, Xinying, Wu, Y., Ming, Y., & Lv, H.** (2020). Remote Sensing Imagery Super Resolution Based on Adaptive Multi-Scale Feature Fusion Network. *Sensors*, 20(4), 1142. Retrieved from <https://doi.org/10.3390/s20041142>
- Wang, Z., Jiang, K., & Yi, P.** (2019). Ultra-Dense GAN for Satellite Imagery Super-Resolution. *Neurocomputing*, 398, 328–337. Retrieved from <https://doi.org/10.1016/j.neucom.2019.03.106>
- Wei, W., Feng, G., Zhang, Q., Cui, D., Zhang, M., & Chen, F.** (2019). Accurate single image super-resolution using cascading dense connections. *Electronics Letters*, 55(13), 739–742.
- 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 (ECCV)* (pp. 3–19). Munich, Germany, 8-14 September. Retrieved from https://doi.org/10.1007/978-3-030-01234-2_1
- Xiao, A., Wang, Z., Wang, L., & Ren, Y.** (2018). Super-resolution for “Jilin-1” satellite video imagery via a convolutional network. *Sensors (Switzerland)*, 18(4), 1–10. Retrieved from

<https://doi.org/10.3390/s18041194>

- Xiao, Y., Su, X., & Yuan, Q.** (2021). A Recurrent Refinement Network for Satellite Video Super-Resolution. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS* (pp. 3865–3868). Brussels, Belgium, 11-16 July (IEEE). Retrieved from <https://doi.org/10.1109/igarss47720.2021.9553281>
- Xiao, Y., Su, X., Yuan, Q., Liu, D., Shen, H., & Zhang, L.** (2021). Satellite Video Super-Resolution via Multiscale Deformable Convolution Alignment and Temporal Grouping Projection. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–19. Retrieved from <https://doi.org/10.1109/TGRS.2021.3107352>
- Xu, W., Xu, G., Wang, Y., Sun, X., Lin, D., & Wu, Y.** (2018). High quality remote sensing image super-resolution using deep memory connected network. In *International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 8889–8892). Retrieved from <https://doi.org/10.1109/IGARSS.2018.8518855>
- Yang, F., Yang, H., Fu, J., Lu, H., & Guo, B.** (2020). Learning Texture Transformer Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5791–5800). Seattle, WA, USA, 13-19 June (IEEE).
- Yi, P., Wang, Z., Jiang, K., Jiang, J., Lu, T., & Ma, J.** (2020). A Progressive Fusion Generative Adversarial Network for Realistic and Consistent Video Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828(c), 1–18. Retrieved from <https://doi.org/10.1109/TPAMI.2020.3042298>
- Yi, P., Wang, Z., Jiang, K., Jiang, J., & Ma, J.** (2019). Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3106–3115). Seoul, Korea (South), 27 October-2 November (IEEE). Retrieved from <https://doi.org/10.1109/ICCV.2019.00320>
- Ying, X., Wang, L., Wang, Y., Sheng, W., An, W., & Guo, Y.** (2020). Deformable 3D Convolution for Video Super-Resolution. *IEEE Signal Processing Letters*, 27, 1500–1504. Retrieved from <https://doi.org/10.1109/LSP.2020.3013518>
- Yu, J., Fan, Y., Yang, J., Xu, N., Wang, Z., Wang, X., & Huang, T.** (2018). Wide Activation for Efficient and Accurate Image Super-Resolution. *ArXiv*. Retrieved from <http://arxiv.org/abs/1808.08718>
- Yuhas, R. H., Goetz, A. F. H., & Boardman, J. W.** (1992). Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne*

Geoscience Workshop. Washington, D. C., October 25-29.

- Zhang, D., Shao, J., Li, X., & Shen, H. T.** (2020). Remote Sensing Image Super-Resolution via Mixed High-Order Attention Network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6), 5183–5196. Retrieved from <https://doi.org/10.1109/tgrs.2020.3009918>
- Zhang, F., Cai, N., Cen, G., Li, F., Wang, H., & Chen, X.** (2018). Image super-resolution via a novel cascaded convolutional neural network framework. *Signal Processing: Image Communication*, 63, 9–18. Retrieved from <https://doi.org/10.1016/j.image.2018.01.009>
- Zhang, Han, Goodfellow, I., Metaxas, D., & Odena, A.** (2019). Self-attention generative adversarial networks. In *36th International Conference on Machine Learning, ICML 2019* (pp. 7354–7363). Los Angeles, USA, 9-15 June.
- Zhang, Haopeng, Wang, P., & Jiang, Z.** (2020). Nonpairwise-Trained Cycle Convolutional Neural Network for Single Remote Sensing Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5), 1–12. Retrieved from <https://doi.org/10.1109/tgrs.2020.3009224>
- Zhang, J., Liu, S., Peng, Y., & Li, J.** (2020). Satellite image super-resolution based on progressive residual deep neural network. *Journal of Applied Remote Sensing*, 14(03), 1. Retrieved from <https://doi.org/10.1117/1.jrs.14.032610>
- Zhang, L., Zhang, L., Mou, X., & Zhang, D.** (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8), 2378–2386. Retrieved from <https://doi.org/10.1109/TIP.2011.2109730>
- Zhang, N., Wang, Y., Zhang, X., Xu, D., & Wang, X.** (2020). An Unsupervised Remote Sensing Single-Image Super-Resolution Method Based on Generative Adversarial Network. *IEEE Access*, 8, 29027–29039. Retrieved from <https://doi.org/10.1109/ACCESS.2020.2972300>
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O.** (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 586–595). Salt Lake City, UT, USA, 18-23 June (IEEE). Retrieved from <https://doi.org/10.1109/CVPR.2018.00068>
- Zhang, S., Yuan, Q., Yuan, Q., Li, J., Sun, J., & Zhang, X.** (2020). Scene-adaptive remote sensing image super-resolution using a multiscale attention network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7), 4764–4779. Retrieved from <https://doi.org/10.1109/TGRS.2020.2966805>

- Zhang, Y., Li, K., Wang, L., & Fu, Y.** (2018). Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 286–301). Munich, Germany, 8-14 September.
- Zhao, H., Kong, X., He, J., Qiao, Y., & Dong, C.** (2020). Efficient Image Super-Resolution Using Pixel Attention. In *European Conference on Computer Vision* (pp. 56–72). Glasgow, United Kingdom, 23-28 August. Retrieved from <http://arxiv.org/abs/2010.01073>
- Zhou, S., Zhang, J., Zuo, W., & Loy, C. C.** (2020). Cross-Scale Internal Graph Neural Network for Image Super-resolution. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)* (pp. 1–13). Canada, 6-12 December.
- Zhou, W., Alan, C. B., Hamid, R. S., & Eero, P. S.** (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 13(4), 7–24.
- Zhou, W., & Bovik, A. C.** (2002). A universal image quality index. *Signal Processing Letters, IEEE*, 9(3), 81–84. Retrieved from <https://doi.org/10.1109/97.995823>
- Zhu, J., Park, T., Efros, A. A., Ai, B., & Berkeley, U. C.** (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *2017 IEEE international conference on computer vision* (pp. 2223–2232). Venice, Italy, 22-29 October (IEEE). Retrieved from <https://doi.org/10.1109/ICCV.2017.244>
- Zhu, X., Talebi, H., Shi, X., Yang, F., & Milanfar, P.** (2020). Super-Resolving Commercial Satellite Imagery Using Realistic Training Data. In *Proceedings - International Conference on Image Processing, ICIP* (pp. 498–502). Anchorage, Alaska, USA, 19-22 September (IEEE). Retrieved from <https://doi.org/10.1109/ICIP40778.2020.9190746>

CURRICULUM VITAE

Name Surname : Peijuan WANG

EDUCATION :

- **B.Sc.** : 2007.09-2011.06, Wuhan University of Technology, School of Information Engineering, Communication Engineering
- **M.Sc.** : 2011.09-2013.06, Wuhan University of Technology, School of Information Engineering, Electronic and Communication Engineering
- **Ph.D.** : 2017.09-2022.07, Istanbul Technical University, Department of Communication Systems, Satellite Communication and Remote Sensing Programme

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Wang, P.**, Bayram, B. and Sertel, E., 2022. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, p.104110.
- **Wang, P.** and Elif Sertel., 2021. Channel–spatial attention-based pan-sharpening of very high-resolution satellite images. *Knowledge-Based Systems*, 229, p.107324.
- **Wang, P.**, Bayram, B. and Sertel, E., 2021. Super-resolution of remotely sensed data using channel attention based deep learning approach. *International Journal of Remote Sensing*, 42(16), pp.6048-6065.
- **Wang, P.**, Alganci, U. and Sertel, E., 2021. Comparative Analysis on Deep Learning based Pan-Sharpener of Very High-Resolution Satellite Images. *International Journal of Environment and Geoinformatics*, 8(2), pp.150-165.