学校代码	10699
分 类 号	TP301.6
密　　级	公开
学　　号	2022280110

西北工业大学
NORTHWESTERN POLYTECHNICAL UNIVERSITY

硕士学位论文

题　目　Application of SRGANs to Enhance the Resolution of UAV Imagery for Improved Object Detection

作者　Rouhbakhshmeghrazi Amirreza

学科专业　电子科学与技术
指导教师　李波
培养单位　电子信息学院
申请日期　2025 年 03 月

西 北 工 业 大 学

硕 士 学 位 论 文

（学位研究生）

题目 Application of SRGANs to Enhance the Resolution of UAV Imagery for Improved Object Detection

作　　者　Rouhbakhshmeghrazi Amirreza

学 科 专 业　电子科学与技术

指 导 教 师　李波

2025 年 03 月

Title: Application of SRGANs to Enhance the Resolution of UAV Imagery for Improved Object Detection

By

Under the Supervision of Professor

Li Bo

A Thesis Submitted to
School of Electronics & Information
Northwestern Polytechnical University

In partial fulfillment of the requirements
For the degree of
Master of Engineering
In
Electronic Science and Technology

Xi'an P.R. China
March 2025

学位论文评阅人和答辩委员会名单
Lists of Thesis Reviewers and Defense Committee Members

学位论文评阅人名单
List of Reviewers of Thesis

姓名 Name	职称 Academic Title	工作单位 Institution
全盲评阅 blind review	无	无
全盲评阅 blind review	无	无

答辩委员会名单
List of Defense Committee Members

答辩日期 Date		2025 年 3 月 11 日	
答辩委员会 Committee	姓名 Name	职称 Academic Title	工作单位 Institution
主席 Chairperson	符小卫	教授	西北工业大学电子信息学院
委员 Member	徐钊	副教授	西北工业大学电子信息学院
委员 Member	马云红	副教授	西北工业大学电子信息学院
委员 Member			
委员 Member			
委员 Member			
秘书 Secretary	万开方	副研究员	西北工业大学电子信息学院

摘 要

无人机（UAV）的应用已广泛应用于多个领域，包括精准农业、环境监测和灾害管理，因其能够执行物体分类和检测等任务。然而，这些任务的有效性常常受到低分辨率图像的影响，原因包括恶劣的天气条件、运动模糊、大气噪声以及由于成本或重量限制，无人机携带高端相机的能力受限。这些挑战要求采用先进的超分辨率方法来提高图像质量，以便进行准确分析。

本研究提出了两种专门为超分辨率任务设计的混合生成对抗网络（GAN）模型，解决了传统基于 GAN 的方法中的关键研究空白。这些空白包括对大量数据集的依赖、易受伪影和不真实纹理的影响、对特定领域的过拟合、训练不稳定以及实时应用中的计算效率低下。这些限制使得传统模型在无人机图像及其他现实场景中的应用变得不切实际，突显了创新解决方案的需求。

为克服这些挑战，本文提出了一种创新的超分辨率方法，将生成对抗网络（GAN）与先进的深度学习技术相结合。研究的关键贡献和创新点如下：

本研究提出了两种混合 GAN 模型，将 SRGAN 与 U-Net、自动编码器及迁移学习相结合，显著提升了结构相似性指数（SSIM）、峰值信噪比（PSNR）和 LPIPS 等度量指标，与 SRGAN 和 ESRGAN 相比表现更优。架构中的注意力门控使模型能够专注于低分辨率图像中的关键区域，改善对小物体和细微纹理的特征提取。这些模型提供了一种通用的解决方案，展示了其在无人机图像、医学影像和卫星数据等多个领域中的适应性。

通过采用自适应判别器增强（ADA），这些模型在有限数据下也能取得显著成果，解决了传统 GAN 依赖大规模数据集的问题。这种数据集无关的设计确保了模型在医疗研究、物体检测和图像分类等多种应用中的稳健表现，拓宽了其适用范围。

包括 U-Net 和预训练自动编码器等架构创新最小化了视觉伪影并提高了计算效率。此外，模型的轻量化版本已针对无人机和边缘设备的部署进行了优化，实现了在资源受限环境中的实时超分辨率。

结果表明，所提出的混合 GAN 模型不仅推动了超分辨率技术的进步，还为现实应用提供了可扩展、可适应的解决方案。它们在最小数据下实现卓越性能，重新定义了 GAN 在图像处理中的能力，并拓宽了其在多个领域的适用性。

关键词：超分辨率，生成对抗网络，图像重建，U-Net，卷积神经网络，高分辨率成像，深度学习，遥感，医学影像

Abstract

The application of Unmanned Aerial Vehicles (UAVs) has been widely adopted across various fields, including precision agriculture, environmental monitoring, and disaster management, due to their ability to perform tasks such as object classification and detection. However, the effectiveness of these tasks is often hindered by low-resolution images caused by factors such as adverse weather conditions, motion blur, atmospheric noise, and the UAV's limited capacity to carry high-end cameras due to cost or weight constraints. These challenges necessitate advanced super-resolution methods to enhance image quality and enable accurate analysis.

This study introduces two hybrid GAN models specifically designed for super-resolution tasks, addressing critical research gaps in traditional GAN-based approaches. These include a reliance on extensive datasets, susceptibility to artifacts and unrealistic textures, overfitting to specific domains, training instability, and computational inefficiency for real-time applications. These limitations often render traditional models impractical for UAV imagery and other real-world scenarios, highlighting the need for innovative solutions.

To overcome these challenges, this paper presents an innovative super-resolution method that integrates Generative Adversarial Networks (GANs) with advanced deep learning techniques. The key contributions and novelties of the study are as follows:

1. The study introduces two hybrid GAN models that integrate SRGAN with U-Net, Autoencoder, and transfer learning, significantly enhancing some measured metrics such as SSIM, PSNR, and LPIPS compared to SRGAN and ESRGAN. Attention gates within the architecture enable the model to focus on crucial regions in low-resolution images, improving feature extraction for small objects and subtle textures. These models provide a universal solution, showcasing adaptability across diverse domains, including UAV imagery, medical imaging, and satellite data.
2. By employing Adaptive Discriminator Augmentation (ADA), the models achieve remarkable results with limited data, addressing the reliance on large datasets faced by traditional GANs. This dataset-agnostic design ensures robust performance across various applications, such as medical research, object detection, and image classification, broadening their applicability.
3. Architectural innovations, including U-Net and pretrained autoencoders, minimize visual artifacts and enhance computational efficiency. Moreover, lightweight versions of the models are optimized for deployment on UAVs and edge devices, enabling real-time super-resolution in resource-constrained environments.

The results demonstrate that the proposed hybrid GAN models not only advance super-resolution techniques but also provide scalable, adaptable solutions for real-world applications. Their ability to achieve exceptional performance with minimal data redefines the capabilities of GANs in image processing and broadens their applicability across multiple domains.

Keywords: Super-Resolution, Generative Adversarial Networks, Image Reconstruction, U-Net, CNN, High-Resolution Imaging, Deep Learning, Remote Sensing, Medical Imaging

Table of Contents

摘　要 I
Abstract III
List of Figures IX
List of Tables XI
List of Abbreviations 13
1 Introduction 1
1.1 Research Background and Significance 1
1.2 Related works 8
1.2.1 Current Status of Research Abroad 9
1.2.2 Current Research Status in China 11
1.3 Challenges 13
1.3.1 Nash equilibrium 14
1.3.2 Failure to convergence 15
1.3.3 Vanishing gradients 16
1.3.4 A Large dataset 17
1.3.5 Creating artifacts 18
1.4 Main Research Content 19
1.5 Organizational Structure of Thesis 21
2 Literature Review 23
2.1 Introduction 23
2.2 GAN-based super resolution models 23
2.3 Applications of U-net in Super resolution tasks 25
2.3.1 Key Applications in Super-Resolution 26
2.3.2 Advantages and Limitations of U-Net in Super-Resolution 27
2.4 Autoencoder in SR 29

2.5 Evaluation on SR images 29
2.5.1 Peak signal-to-noise ratio 30
2.5.2 Structural Similarity Index 31
2.6 Loss functions 32
2.6.1 Perceptual loss 33
2.6.2 Adversarial loss 35
2.7 Perception-Distortion Trade-off 36
2.8 Summary 37
3 Methodology 39
3.1 Introduction 39
3.2 Conventional SRGAN 43
3.3 First Proposed model (A-SRGAN) 45
3.4 Second proposed model (U-SRGAN) 51
3.4.1 U-Net 52
3.4.2 Residual blocks 53
3.4.3 Attention Gate 54
3.4.4 Pretrained U-Net with Autoencoder 57
3.5 Adaptive Discriminator Augmentation (ADA) 62
3.6 Loss function 64
3.7 Algorithm performance (Evaluation metrics) 65
3.7.1 PSNR 65
3.7.2 SSIM 65
3.7.3 LPIPS 66
3.7.4 DISTS 66
3.8 Dataset 68
3.9 Training details and parameters 69
3.10 YOLO9x Integration 70

3.11 Summary 70
4 Experimental Process and Results 71
4.1 Introduction 71
4.2 A-SRGAN results 73
4.2.1 Qualitative results of A-SRGAN 76
4.2.2 Ablation Study on Autoencoder and Residual Blocks in SRGAN 79
4.3 Results of SRGAN with U-Net 82
4.3.1 Qualitative results of U-SRGAN 85
4.3.2 Ablation Study on U-Net Integration in SRGAN 87
4.3.3 Results of ARU-Net pretrained with Autoencoder 88
4.3.4 Ablation Study on Pretrained ARUnet-SRGAN 90
4.4 Comprehensive comparison 92
4.5 Optimum model 95
4.6 YOLO9 Experiment Findings 98
4.7 Validation of Generalization Across Diverse Datasets 100
4.8 Summary 101
5 Conclusion 103
5.1 Summary of Key Findings 103
5.2 Future works 106
Reference 109
Acknowledgements 119
List of Publications 121

List of Figures

Figure 1-1 The thesis organization…….. 22
Figure 2-1 VGG-19 architecture: Typically, the style feature maps are taken from the first convolutional block, whereas the content feature maps are from the fifth block……..... 34
Figure 2-2 Perception-distortion trade-off, showing how improving an algorithm in terms of perception occurs only at the expense of increasing distortion and vice versa……..……. 36
Figure 3-1 GAN network structure…….. 39
Figure 3-2 SRGAN network structure……. 40
Figure 3-3 All the photos on the circumference of the circle resembles each other when MSE is selected as the loss function since MSE compare photos pixel by pixel and do not consider structural similarity. Figure taken from [126] …….. 41
Figure 3-4 The original architecture of generator in the SRGAN…….. 44
Figure 3-5 the basic design of discriminator in SRGAN……. 45
Figure 3-6 Autoencoder with skip connections. Figure taken from [139] ……..…… 46
Figure 3-7 The architecture of Encoder and Decoder in Autoencoder ……... 47
Figure 3- 8 The autoencoder that was integrated in the design of SRGAN generator …… 48
Figure 3-9 The updated SRGAN generator using autoencoder ……..... 49
Figure 3-10 the simplified block diagram of Autoencoder we used in the SRGAN structure, the only difference is that we used residual blocks instead of Convolutional blocks. For the sake of simplification skip connections were shown by arrows, while we used Conv layers in the skip connections…….... 49
Figure 3-11 The architecture of generator in A-SRGAN……. 50
Figure 3-12 Using U-Net in the design of Basic SRGAN.……. 51
Figure 3-13 A typical design of U-net for a segmentation task……... 53
Figure 3.14 Building blocks of neural networks. (a) Plain neural unit used in U-Net and (b) residual unit with identity mapping used in the proposed Res-U-Net……... 54
Figure 3-15 The actual residual block that we used in our models……. 54
Figure 3-16 Schematic of the proposed additive attention gate (AG). Figure taken from [146]……..……... 55
Figure 3-17 ARU-Net SRGAN…… 57
Figure 3-18 The symmetric autoencoder takes the LR images and reconstruct the images……. 58
Figure 3-19 The attention residual U-net (ARU-Net). We designed the downsampling path same as symmetric autoencoder…… 58
Figure 3-20 The weights from downsampling path of SAE are transferred to the encoder of ARU-net……. 59
Figure 3-21 the flow chart of pretrained ARU-Net SRGAN……... 60
Figure 3-22 Examples from the dataset of Pix4Dmatic……... 68

Figure 4-1 The comparison of SSIM and PSNR, LPIPS, DISTS, and generator loss between typical SRGAN and SRGAN with autoencoder in their generator architecture …… 75
Figure 4-2 Examples of generated images created by three different models studied……. 76
Figure 4-3 Comparison of SRGAN, A-SRGAN, and Res-A-SRGAN models with SSIM and PSNR values displayed for each model……... 77
Figure 4-4 Comparison of SRGAN, A-SRGAN, and Res-A-SRGAN models with SSIM and PSNR values displayed for each model……. 78
Figure 4-5 Comparison of two different images generated by three models and their corresponding line and bar histograms as opposed to their respective HR image……... 79
Figure 4-6 PSNR, SSIM, and Generator Loss for A-SRGAN and Res-A-SRGAN without ADA during training……. 83
Figure 4-7 The comparison of metrics and generator loss between typical SRGAN and SRGAN with U-Net in their generator architecture ……..……..... 84
Figure 4-8 Comparison of SRGAN, U-SRGAN, and ARUnet-SRGAN models with SSIM and PSNR values displayed for each model……... 85
Figure 4-9 Comparison of SRGAN, U-SRGAN, and ARUnet-SRGAN result…… 86
Figure 4-10 Comparison of SRGAN, U-SRGAN, and ARUnet-SRGAN outputs with HR image using histograms and error maps…… 87
Figure 4-11 PSNR, SSIM, and Generator Loss for U-SRGAN and ARUnet-SRGAN without ADA during training…... 89
Figure 4-12 We trained the autoencoder 3000 epochs to reconstruct the LR images, so that we can transfer the weights to the U-Net…… 90
Figure 4-13 the comparison of SSIM, and PSNR values between the ARUnet-SRGAN and the ARUnet-SRGAN when a U-Net was pretrained by an autoencoder beforehand…… 90
Figure 4-14 The comparison of LPIPS, and DISTS values between the Pretrained ARUnet-SRGAN and the ARUnet-SRGAN…… 90
Figure 4-15 SSIM and PSNR for pretrained ARUnet-SRGAN without using ADA…… 92
Figure 4-16 Metric comparison between all models studied…… 94
Figure 4-17 The bar chart compares the maximum value of each model for SSIM, and PSNR value…… 94
Figure 14-18 The best achieved LPIPS and DISTS values achieved by different models……... 95
Figure 4-19 Heatmap showcasing the Average SSIM and PSNR metrics across various models……... 96
Figure 4-20 The progression of four key evaluation metrics (SSIM, PSNR, DISTS, and LPIPS) across epochs during training 97
Figure 4-21 The right-side photo is the ground truth image of size 1024*1024 and the left one is the SR photo…….. 98
Figure 4-22 Comparison of Pixel Intensity Distributions: High-Resolution (HR) vs. Super-Resolution (SR) Images Highlighting Frequency Patterns…….. 98
Figure 4-23 Comparison of the proposed Pretrained ARUnet-SRGAN with common super-resolution models…… 99

List of Tables

Table 3-1 BRISQUE range, Lower BRISQUE scores indicate better perceptual quality…… 68

Table 4-1 Different models along with number of generator parameters and training duration…… 73

Table 4-2 Impact of Skip Connections and Convolutional Layers on PSNR and SSIM for A-SRGAN and Res-A-SRGAN…….. 81

Table 4-3 Impact of Skip Connections and Convolutional Layers on LPIPS and DISTS for A-SRGAN and Res-A-SRGAN…… 81

Table 4-4 Performance Comparison of Pretrained ARUnet-SRGAN with and without ADA…….. 92

Table 4-5 The Comprehensive comparison of all models studied……..... 93

Table 4-6 Quantitative Comparison of Super-Resolution Models Based on PSNR and SSIM Metrics…… 99

Table 4-7 Comparison of super-resolution models on DIV2K and Flickr2K datasets…….. 101

List of Abbreviations

UAV	Unmanned Aerial Vehicles
GAN	Generative Adversarial Networks
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index
ADA	Adaptive Data Augmentation
LR	Low-resolution
HR	High-resolution
SR	Super Resolution
SISR	Single-image Super Resolution
CNN	Convolutional Neural Networks
EDSR	Enhanced Deep Super Resolution
CinCGAN	Cycle-in-cycle GAN
MSRN	Multi-scale Residual Networks
CMOS	Complimentary Metal-oxide Semiconductor
ISP	Image Signal Processing
DL	Deep Learning
MFSR	Multi-Frame Super Resolution
GSRN	Guidance Super-Resolution Network
SRCNN	Super resolution convolutional neural network
mAP	mean Average Precision
NCC	Normalized-Cross-Correlation
ISRGAN	Improved Super-Resolution Generative Adversarial Network
LCC	Land Cover Classification
SRGAN	Super resolution generative adversarial network
VAE	Variational Encoder
MOS	Mean Opinion Score
MSE	Mean Squared Error
DSM	Digital Surface Models
ResNet	Residual Networks
RRDB	Residual-in-Residual Dense Blocks
ESRGAN	Enhanced Super-Resolution GAN
MSLE	Mean Squared Logarithmic Error
RU-Net	Robust U-Net Variants
HVS	Human Visual System
SAE	Symmetric Autoencoder

1 Introduction

This chapter introduces the fundamental concepts of enhancing image resolution using Generative Adversarial Networks (GANs), a cutting-edge technique in the field of artificial intelligence. UAV (Unmanned Aerial Vehicle) imagery often captures details at a small scale, which can suffer from degradation due to a variety of factors, including hardware limitations and adverse weather conditions. Super-resolution (SR) plays a critical role in overcoming these challenges, particularly in AI applications such as image classification, object detection, and segmentation. The study aims to develop a model that addresses the limitations of existing super-resolution approaches and enhances image quality, even in the presence of severe degradation. This chapter outlines the significance of recent advancements in super-resolution, specifically focusing on the potential of GANs to create more accurate, high-resolution images from low-quality UAV data. Additionally, the chapter highlights the relevance of this work in both academic and practical domains, underscoring its potential impact on industries such as environmental monitoring, agriculture, and surveillance. The scope, objectives, and structure of the thesis are also detailed, laying the foundation for the research that follows.

1.1 Research Background and Significance

Super-resolution imaging is a transformative technology with significant implications across various fields. Super-resolution refers to the technique of generating a higher-quality image from one or several lower-quality images. SR techniques are commonly applied in various domains, including Medical Imaging, to improve the quality of essential medical images like MRI and CT scans for precise disease diagnosis and clinical decision-making [1]. These techniques enhance the resolution and quality of images for medical analysis, improving clarity and detail, and allowing for better diagnosis and treatment planning. Improved image resolution aids in identifying subtle details in scans, such as tumors or fractures, which may be missed in lower-resolution images [2].

In the field of Remote Sensing and Satellite Imaging, SR plays an essential role in applications such as monitoring urbanization, conducting environmental surveillance, and detecting resources. Enhancing spatial resolution is crucial for in-depth analysis, and SR techniques such as Single Image Super Resolution (SISR) are utilized to boost the resolution of satellite images [3], [4]. Advanced deep learning techniques are employed to improve the resolution of satellite datasets, overcoming challenges presented by different land cover types [5]. In Video Surveillance, SR techniques play a vital role in enhancing the resolution of footage

for identifying details in security applications. These techniques enhance the visual sharpness and quality of low-resolution video, simplifying the identification and examination of critical details [6].

SR in forensic analysis is utilized to improve the clarity of images obtained from crime scenes, offering precise evidence that is essential for investigations. High-quality photos provide additional details, helping to analyze forensic evidence more accurately [7]. Similarly, super-resolution techniques are used in agriculture to monitor crop health and in various industrial settings for quality control and inspection [8]. Super-resolution plays a role in the area of historical preservation. It enables the improvement of aged or impaired photographs and videos, highlighting details that might have faded over the years. This application is especially crucial for archival purposes, where maintaining the accuracy of historical documents is vital. Improved images can offer a sharper understanding of historical events and cultures, aiding research in disciplines like archaeology and history [9].

Early and conventional techniques for super-resolution imaging include both hardware and software methods, with each playing a unique role in enhancing image resolution. These essential approaches have paved the way for contemporary progress in the discipline. For example, iterative back-projection techniques were modified for hardware to allow real-time processing capabilities. These implementations made use of FPGAs to reconstruct high-resolution images from several low-resolution observations, resulting in notable enhancements in processing speed and image quality [10]. Another significant hardware-based technique mainly depended on optical systems to enhance image quality, known as optical interpolation. This method used specially engineered optical components such as lenses and mirrors to boost resolution by altering light paths [11]. Methods involving multi-sensor systems have also surfaced, wherein several cameras recorded images from subtly varied angles or locations. By arranging and merging these images, a higher-resolution image could be reconstructed, which proved especially beneficial in fields such as remote sensing and surveillance [12]. Furthermore, novel hardware methods such as coded aperture modulation have been created. This technique employs a programmable aperture to control light, enabling the acquisition of high-resolution images without the need for physical motion or scanning devices, thereby enhancing system reliability and reducing costs [13].

Traditional software techniques for SR concentrated on generating high-resolution images from various low-resolution sources. A key method involved the application of interpolation techniques like bicubic and bilinear interpolation, which sought to approximate absent pixel values to improve image resolution. These techniques were straightforward and

computationally efficient, yet frequently led to blurry images because they struggled to properly recreate high-frequency details [14]. Another important technique was the iterative back-projection method, which progressively improved the high-resolution image by contrasting it with the low-resolution inputs and modifying it according to the discrepancies. This method enhanced basic interpolation by adding feedback systems, yet it continued to struggle with precisely reconstructing intricate details and necessitated careful adjustment of parameters [15]. Sparse representation techniques also surfaced as an effective instrument in conventional SR methods. These techniques utilized the concept that image patches can be expressed as sparse linear combinations of components from a trained dictionary. This method enhanced the capacity to recreate high-frequency features, though it demanded considerable computational power and meticulous dictionary training [16].

Recent advancements in software for SR have been greatly shaped by deep learning methods, especially convolutional neural networks (CNNs). These techniques have transformed SISR by successfully capturing varied image characteristics and enhancing high-frequency details. Improved deep CNN architectures now utilize several convolutional layers featuring designated filters and activation functions, in conjunction with residual learning techniques, to speed up training and enhance convergence. These developments have resulted in enhanced performance on public datasets, preserving image edges and textures more efficiently than conventional techniques [17].

Methods for super-resolution based on deep learning have been classified into classical, supervised, unsupervised, and domain-specific categories. Cutting-edge models such as the enhanced deep SR network (EDSR) [18], cycle-in-cycle GAN (CinCGAN) [19], and multiscale residual network (MSRN) have established new standards in image quality. These models utilize sophisticated neural network structures to tackle issues like restricted receptive fields and significant computational requirements, delivering enhanced performance and efficiency. Initiatives aimed at enhancing the accessibility and efficiency of SR have resulted in the creation of architectures such as SwiftSRGAN, which employs depth-wise separable convolutions to attain real-time performance while maintaining a small memory footprint. This method allows high-resolution media streaming even in low bandwidth scenarios, showing a notable decrease in inference time and resource needs when compared to conventional GANs [20].

The use of improved image resolution via GANs holds considerable promise in numerous fields, particularly when integrated with the functionalities of UAVs. Drones are progressively utilized across multiple sectors for activities like environmental observation [21], farming [22],

disaster response [23], city planning [24], and military surveillance [25]. Nevertheless, the quality of images taken by UAVs frequently poses an obstacle to efficient analysis. Images with low resolution can cause errors in essential activities like object detection, mapping, and target recognition, hindering the ability to obtain useful data [26]. This study tackles this limitation by utilizing GANs to boost the resolution of images captured by UAVs, thereby increasing the accuracy and usefulness of the obtained data.

In environmental surveillance, UAVs are commonly utilized to monitor alterations in ecosystems, like deforestation, pollution rates, or biodiversity. Improved image resolution delivered by GANs facilitates a more thorough examination of landscapes, allowing for enhanced monitoring of vegetation health, soil quality, and water resources. For example, high-resolution aerial images can enhance the identification of minor environmental shifts, like small-scale deforestation or wetland degradation, that would be hard to notice with low-resolution images. This degree of detail can be essential for creating more precise models for predicting climate change and conservation initiatives for biodiversity, enabling focused and impactful actions [27].

In agriculture, drones outfitted with imaging sensors are emerging as an essential resource for precision farming. Improved resolution from GANs can greatly enhance the capacity to track crop health, spot pests, and recognize regions experiencing water stress. The capability to obtain sharper images of crops during different growth phases enables farmers to make data-informed choices related to irrigation, fertilization, and pest management [28]. Additionally, imagery enhanced by GANs can boost crop yield forecasts by delivering more precise information about crop conditions and the overall health of the field. Consequently, farmers can maximize resource efficiency, lower expenses, and enhance output, aiding sustainable agricultural practices.

In disaster response, UAVs are frequently utilized to obtain real-time photos of regions affected by disasters, like after an earthquake, flood, wildfire, or landslide [29]. Images with high resolution produced by GANs can greatly improve the capability to evaluate damage and pinpoint regions needing immediate action. For instance, sharper images can show the degree of structural damage, identify hazards like gas leaks or fires, and assess the state of essential infrastructure including bridges and roads. This enhanced image clarity facilitates quicker and more precise decisions by emergency personnel, aiding in resource allocation and the coordination of rescue efforts.

Drones with advanced imaging abilities, improved by GANs, can significantly impact urban planning. By offering comprehensive aerial perspectives of urban areas, these systems

can aid in tracking urban expansion, alterations in land utilization, and the advancement of infrastructure. Improved images can be utilized in smart city initiatives, where comprehensive information on traffic trends, construction locations, and public areas can guide policy and planning choices [30]. For instance, sharper images can assist city planners in pinpointing underused areas or evaluating the effectiveness of public transit systems. Additionally, improved resolution can assist in addressing environmental challenges in cities, like pinpointing heat islands or monitoring air pollution levels.

In the military field, UAVs are widely utilized for reconnaissance, surveillance, and overseeing adversarial regions. The improved resolution offered by GANs is particularly important in military contexts, where high-quality images are essential for detecting possible threats, examining terrain, and accurately pinpointing targets. Enhancing the resolution of images taken by UAVs allows military operations to obtain more precise intelligence, resulting in improved strategic planning and higher mission success rates. Moreover, GANs can be utilized to improve image quality from previous UAV data, facilitating extended surveillance and observation of areas of interest [31].

A further practical use of GAN-augmented UAV imagery lies in the area of infrastructure monitoring. Drones are being utilized more and more to examine essential infrastructure like bridges, power lines, pipelines, and communication towers. Images with low-resolution can frequently miss subtle details like tiny cracks or corrosion that could suggest significant underlying problems. Techniques for super-resolution based on GANs can yield clearer and more precise images of these structures, facilitating improved identification of maintenance requirements and possible dangers. This improves the safety and dependability of infrastructure systems, lowers maintenance expenses, and prolongs the life of essential assets [32], [33].

The ongoing development of remote sensing technologies, especially unmanned aerial vehicles (UAVs), has greatly changed multiple sectors including environmental monitoring, agriculture, urban planning, and disaster management. Drones fitted with high-resolution cameras have become vital instruments for collecting extensive data across expansive regions. Nonetheless, even with advancements in UAV technologies, a significant challenge persists in the form of limited image resolution, frequently leading to the loss of vital information that impairs precise analysis and decision-making.

There are two popular methods for obtaining high-resolution images. The initial method relies on enhancing the hardware. The resolution of a complementary metal-oxide-semiconductor (CMOS) camera increases as the number of pixels increases. Hence, one approach to enhance spatial resolution is by increasing the chip size to accommodate more

CMOS sensors on the chip [34]. Another option is to decrease the pixel density in order to fit more pixels onto a set chip size. Yet, reducing pixel size may not lead to increased resolution because the maximum sampling rate is determined by the diffraction limit (Airy disc) of the optics. Both options are constrained by the drawback of increased expenses and the need for advanced manufacturing techniques. While enhancing the hardware might seem like a straightforward solution to address the issue of low-resolution images, it can be expensive. For instance, a large portion of New Zealand has access to high-resolution aerial photographs with a pixel resolution of approximately 0.1m or less. Nevertheless, because of the expenses involved in collecting and analyzing the data, images are usually obtained every few years, resulting in high-resolution data that is potentially 4 years old. A problem with aerial photography is that it is usually captured in the summer, in the middle of the day when shadows are minimal and there is less cloud cover, and is not commonly obtained in the winter.

However, achieving high-resolution images in UAV imagery and remote sensing comes with several obstacles. One of the primary challenges is the technological limitation of the sensors and cameras used in UAVs. High-resolution sensors are often expensive and may require more advanced UAV platforms that can handle the additional weight and power requirements. Additionally, high-resolution images generate large amounts of data, necessitating robust data storage and processing capabilities. This can be particularly challenging in remote or resource-limited environments where bandwidth and computational power may be constrained.

Another obstacle is the impact of environmental factors such as weather conditions, lighting, and atmospheric interference, which can degrade image quality. Consistently capturing high-resolution images requires optimal conditions and may involve sophisticated stabilization and correction techniques. Furthermore, regulatory restrictions on UAV operations, including flight altitude limitations and no-fly zones, can restrict the ability to capture high-resolution imagery in certain areas. Given these challenges, the project of creating high-resolution images from UAV imagery is highly significant. By developing innovative solutions to enhance image resolution, whether through advanced sensor technology, improved data processing algorithms, or novel image enhancement techniques, it is possible to overcome these obstacles and unlock the full potential of UAV imagery in remote sensing. This endeavor not only aims to improve the quality and utility of the imagery but also to expand the applications and benefits of UAV technology across various fields.

The second method is based on software improvement by creating faster and more precise algorithms to improve resolution from low to high. This option is possible because advanced

computing components like graphic processing units (GPUs) and image signal processing (ISP) are capable of managing tasks that require a lot of computational power. Techniques known as super-resolution (SR) algorithms are used to rebuild HR images from LR images. These techniques enhance the quality of images by providing a visual quality better than its LR counterpart. Deep learning (DL) is a branch of machine learning methods that utilizes artificial neural networks. Recent deep learning-based super-resolution methods have shown significant advancements compared to traditional signal processing methods due to their ability to extract useful high-level abstractions that connect the low-resolution and high-resolution domains. This has led to superior performance in fields like computer vision, natural language processing, audio recognition, and machine translation. Despite achieving success, these methods are heavily restricted by their dependence on the assumed degradation model between the HR image and the LR image. It is common knowledge that they do not apply to natural images, since the authentic degradation in real-life LR images is far more intricate.

This study focuses on tackling the important problem of improving image resolution by utilizing GANs. GANs have surfaced as a state-of-the-art approach in machine learning and image processing, providing impressive abilities to create high-quality, high-resolution images from low-resolution sources. The uniqueness of this project is found in its use of GANs on images captured by UAVs, a field that is still not thoroughly investigated in scholarly research. This research adds to the expanding understanding of image super-resolution and deep learning, offering fresh perspectives on the application of these technologies to enhance the quality of remote-sensing images.

Additionally, the research tackles important issues in super-resolution models, such as managing real-world noise, blurriness, and constraints in data availability. This study investigates innovative techniques to enhance image resolution for UAV-generated images, aiming to boost the efficiency of image classification, object detection, and segmentation; crucial activities in areas such as precision agriculture, environmental monitoring, and disaster management. Therefore, this project enhances the scholarly knowledge of GANs and also aids in the practical uses of deep learning for remote sensing and image processing.

The practical importance of this study is found in its ability to improve the operational efficiency and precision of UAV-driven systems in different sectors. As UAVs are more integrated into practical uses, the demand for accurate and high-quality images is essential for making informed choices. For instance, in farming, improved image clarity facilitates superior crop observation, pest identification, and yield forecasting. In environmental monitoring, enhanced images aid in better tracking of deforestation, pollution, and changes in biodiversity.

Additionally, this research could influence disaster management initiatives, where UAVs are utilized to evaluate damage following natural disasters such as earthquakes, floods, and wildfires. Images with higher resolution can offer an improved understanding of impacted regions, helping in prompt and precise response actions. Moreover, the technology has the potential to greatly enhance the navigation and mapping capabilities of autonomous UAVs, as clear and detailed visuals are crucial for route planning and obstacle identification. By enhancing the quality of images captured by UAVs, this initiative can aid the wider area of remote sensing, facilitating more efficient data-informed choices across multiple practical sectors. Additionally, the findings of this study may result in the creation of sophisticated software tools or platforms that incorporate GAN-boosted imagery, providing innovative solutions for sectors dependent on UAV data.

In the realm of technology, this study expands the limits of artificial intelligence and machine learning in addressing practical issues connected to image processing. Although GANs have demonstrated significant promise in producing high-quality images across different fields, their use in UAV imagery introduces distinct challenges, including managing the unique noise and distortion associated with aerial images. By addressing these issues, this study may aid in creating more resilient, effective, and scalable GAN models customized for remote sensing needs. Merging GANs with UAV technology also introduces new opportunities for enhancing image quality in real-time applications, where efficiency and speed are essential. Additionally, this study has the potential to inspire upcoming innovations in UAV-centric computer vision systems, propelling progress in autonomous operations, real-time image analysis, and decision-making. In summary, this research exists at the crossroads of deep learning, remote sensing, and UAV technology, providing both theoretical and practical contributions. The results could greatly enhance the quality of UAV imagery, improving various applications from environmental monitoring to disaster response, and expanding the limits of image resolution enhancement.

1.2 Related works

This section offers an in-depth overview of the latest research in the field of super-resolution, with a specific emphasis on UAV imagery. It includes studies from around the world, as well as those focused on studies conducted in China, shedding light on the progress and challenges in this area. By exploring both global and Chinese research efforts, this section seeks to illustrate the wide-ranging work being done to improve resolution for remote sensing applications.

1.2.1 Current Status of Research Abroad

In 2023, Albuquerque F and Jung explored advancements in SR by integrating semantic segmentation into the SR process. Traditional SR methods focus on enhancing object detection, but this study proposes that incorporating segmentation can improve the perceptual quality of super-resolved images. The research is conducted using the LandCoverAI and DRS datasets, focusing on deep learning techniques in image processing, particularly in semantic segmentation and remote sensing. The study employs a novel approach by integrating a semantic segmentation module into the SR training process. Various configurations of SR networks (ESRGAN and SAGAN) and segmentation networks (U-Net and HRNet) are used. A joint loss function combining perceptual and segmentation losses is utilized to enhance the quality of super-resolved images. The method is evaluated using metrics like PSNR, SSIM, LPIPS, and PI to assess both image quality and segmentation effectiveness. The integration of segmentation loss into SR models generally improved perceptual metrics (LPIPS and PI), with significant enhancements in visual quality [35].

This document [36] discusses a novel Multi-Frame Super Resolution (MFSR) method that leverages attention mechanisms within convolutional neural networks to enhance feature representation, specifically for remote sensing images. This method integrates a new attention module to capture cross-channel correlations and improve model performance, with enhancements in the GAN's discriminator for better feature map processing. The MFSR model, evaluated on the SpaceNet7 and Jilin-1 datasets, shows superior performance in metrics like PSNR, SSIM, AG, NIQE, and PI compared to existing methods.

Alvarez-Vanhard et al. present a pioneering super-resolution technique termed Fusion-U-Net, which integrates multispectral data derived from Sentinel-2 with digital terrain models to precisely estimate hydrological levels in wet grassland ecosystems. This innovative approach mitigates the shortcomings associated with conventional super-resolution methodologies by employing a compound loss function that encompasses content, structural, and segmentation losses, thereby avoiding unrealistic patterns often produced by adversarial networks. The research illustrates that Fusion-U-Net significantly enhances the precision of hydrological and ecological metrics, surpassing the performance of standard U-Net architectures, and posits the prospective incorporation of SAR data to further elevate model efficacy [37].

This paper[38] introduces the Guidance Super-Resolution Network (GSRNet), a novel approach to enhance the resolution of thermal images captured by UAVs using high-resolution visible images. GSRNet utilizes a convolutional neural network with an encoder-decoder architecture and an auto-attention mechanism to focus on relevant image features, translating

visible images into the thermal domain and merging them with low-resolution thermal images to produce high-resolution outputs.

The paper [39] explores the integration of Real-ESRGAN, a super-resolution model, with YOLOv7, an object detection model, to enhance object detection in low-resolution images captured by UAVs. The study demonstrates that super-resolution images significantly improve detection accuracy, particularly for small objects, as evidenced by experiments using the VisDrone dataset. The results show that while low-resolution images achieved a mean Average Precision (mAP) of 40.5% for people and 52.9% for cars, super-resolution images improved the mAP to 37.8% for people and 85.3% for cars, highlighting the potential of super-resolution techniques in UAV applications.

In 2024, Aybar et al. [40] introduced the OpenSR-test framework, a benchmark designed to evaluate SR techniques specifically for optical remote sensing images. It addresses the limitations of traditional assessment methods that often rely on synthetic datasets and metrics, which may not accurately capture improvements in spatial resolution. The OpenSR-test framework employs a comprehensive methodology to assess the fidelity of SR images compared to low-resolution (LR) images. It focuses on metrics such as reflectance, spectral preservation, and spatial consistency. The framework includes a harmonization process to correct systematic errors and uses various distance metrics to quantify improvements and omissions in high-frequency information. It also features curated cross-sensor datasets and tailored quality metrics for remote sensing applications. The results of the OpenSR-test framework compare three pretrained SR models (SR4RS, diffuser, and SuperImage) across different datasets. The findings highlight the strengths and weaknesses of these models in preserving image quality, revealing a trade-off between hallucinations (artifacts) and omissions (loss of detail). The study emphasizes the need for a balanced approach in SR algorithms and suggests the potential for extending the framework to other remote sensing image synthesis tasks.

Li et al. [41] introduced Swin2-MoSE, a novel single-image super-resolution (SISR) model tailored for remote sensing applications. Building upon the Swin2SR architecture, Swin2-MoSE incorporates several advancements, including a Sparsely-Gated Mixture-of-Experts (MoE-SM) module that replaces traditional MLP layers in Transformer blocks to enhance efficiency and performance. The model employs a per-example strategy for expert selection and introduces a Smart Merger layer for output fusion. Additionally, it utilizes advanced positional encoding techniques and combines Normalized Cross-Correlation (NCC) and Structural Similarity Index Measure (SSIM) losses to optimize training and output quality. The

research is supported by the European Union's Next Generation EU initiative and leverages high-performance computing resources.

The results demonstrate that Swin2-MoSE outperforms state-of-the-art models in terms of Peak Signal-to-Noise Ratio (PSNR) and SSIM across various datasets, including Sen2Ven μs and OLI2MSI. The model shows significant improvements in image quality and enhances the performance of downstream tasks like semantic segmentation by providing enriched features. Ablation studies reveal that the combination of NCC and SSIM losses yields the best results, and the integration of both per-head and per-channel positional encodings further improves performance. Despite slight increases in latency, the MoE architecture offers superior image quality, highlighting Swin2-MoSE's potential for remote sensing applications.

The study by Alves Nogueira et al. addresses the challenge of low-resolution maize images captured by unmanned aerial vehicles, which are crucial for efficient agricultural monitoring as global food demand increases. The research employs advanced image processing techniques, specifically focusing on super-resolution methods like Real-ESRGAN and MuLUT, to enhance image quality. The dataset comprises high-resolution images of maize plants at various growth stages, emphasizing nutrient-related anomalies. Results indicate that deep learning methods significantly outperform traditional interpolation techniques, with SR algorithms improving image resolution by 364.13%. These findings underscore the potential of these techniques in precision agriculture, facilitating better crop monitoring and disease detection, and suggest future research to adapt these methods to various crops and environments [42].

1.2.2 Current Research Status in China

Xiong et al. in 2020 present a paper [43] that covers the creation and assessment of the Improved Super-Resolution Generative Adversarial Network (ISRGAN), which enhances the original SRGAN model to better the spatial resolution of remote sensing images. ISRGAN tackles problems such as gradient vanishing and mode collapse by altering the loss function and network architecture, integrating Wasserstein distance, and modifying the discriminator’s final layer to enhance training stability and generalization. Tests utilizing data from Landsat 8 OLI and Chinese GF 1 sensors revealed that ISRGAN outperformed other techniques in image quality metrics and land cover classification accuracy, emphasizing its promise for use in environmental monitoring and resource development.

The document [44] discusses a novel end-to-end framework called Super Resolution Guided Deep Network (SRGDN) designed for land cover classification (LCC) from remote sensing images. The SRGDN framework integrates a super-resolution (SR) branch and an LCC

branch, enhanced by a guidance module, to improve image resolution and classification accuracy. It employs a generative adversarial network to generate high- and low-resolution image pairs for training, effectively reducing computational costs while enhancing performance. The framework demonstrates superior results on various datasets, capturing fine structures and improving classification outcomes compared to existing methods.

Xio et al. explored a self-supervised degradation-guided adaptive network for blind super-resolution of remote sensing images, focusing on overcoming the limitations of traditional methods that assume fixed degradation processes. The proposed approach employs contrastive learning to create robust degradation representations, enabling adaptation to various unknown degradation factors. A dual-wise feature modulation network is introduced to enhance feature adaptation, achieving superior performance in image restoration tasks, as demonstrated by extensive experiments across multiple datasets. The method excels in maintaining image quality, particularly in challenging scenarios with severe noise and blur, and effectively generalizes to real-world degradations without requiring labeled data [45].

In [46], the authors demonstrate the use of deep learning models, including ResNet-50, ConvNeXt-T, and Swin Transformer, for classifying tree species from UAV-captured RGB images. The study highlights the effectiveness of Transformer models over traditional CNNs, particularly in handling low-quality aerial images, due to their attention mechanisms. The application of Real-ESRGAN technology for super-resolution reconstruction significantly improved image quality and classification accuracy across all models, with ConvNeXt-T achieving the highest accuracy. The research underscores the potential of deep learning techniques in forestry applications, particularly in enhancing model performance through improved image processing.

Liu et al. present an advanced image super-resolution model called END-GAN, which enhances the existing EGAN algorithm by addressing issues such as blurred edges, unstable training, and artifacts in reconstructed images. END-GAN employs dynamic gradient descent optimization and artifact discrimination loss to improve image quality, outperforming several other algorithms in remote sensing applications [47].

The author in [48] explores advancements in the extraction and enhancement of Digital Surface Models (DSMs) and Digital Elevation Models (DEMs) using deep learning techniques, with a focus on geoscience and remote sensing applications. It highlights the use of Generative Adversarial Networks for generating super-resolution DSMs by integrating high-resolution remote sensing imagery, demonstrating improved accuracy and detail recovery in urban and mountainous areas. The DSMSR model, which incorporates multiscale attention and slope loss,

outperforms traditional methods and other deep learning models, although it requires high-quality imagery and is computationally complex.

Kang et al. describe the development and evaluation of ESTNet, an Efficient Swin Transformer network designed for remote sensing image super-resolution. ESTNet features a novel architecture with components for shallow and deep feature extraction, and image reconstruction, significantly reducing computational costs and model parameters compared to existing methods. The network incorporates a Residual Group-Wise Attention Module with Efficient Channel Attention and Group-Wise Attention Blocks, enhancing feature representation and image quality. Evaluations on multiple datasets demonstrate ESTNet's superior performance in terms of accuracy and efficiency, with robustness against noise and anisotropic blur, making it a promising solution for real-world remote sensing applications [49].

Zhou et al. discuss the implementation and evaluation of a Super Resolution Generative Adversarial Network (SRGAN) for enhancing low-resolution images in the context of power inspection. It introduces the BDZ dataset, specifically curated for this purpose, and compares the performance of SRGAN [50] with other models like SRCNN, VDSR, and SwinIR, highlighting its superior perceptual quality despite higher computational complexity. The study emphasizes the effectiveness of combined loss functions and metrics like LPIPS and NIQE in improving image quality and suggests future work on developing lightweight SRGAN models to enhance efficiency and reduce hardware costs in power inspection applications [51].

1.3 Challenges

SR presents an issue of indeterminate nature. Different high-resolution (HR) images may possess the same low-resolution (LR) counterparts. If the two images have minimal differences, they will be indistinguishable after down-sampling. Therefore, achieving the perfect reconstruction of the HR image is very difficult and sometimes impossible. Moreover, due to the lack of information in the LR input, filling the gaps in the super-resolved image needs prior knowledge. For example, in a 16 $\times$ 16 image, there are 256 pixels. Each pixel is represented by 3 times 8-bit values. In total, this image carries 3840 bits of information. However, after the dimension is increased 8 times more, the reconstructed image will have 245760 bits of information. So, based on the available data in the LR image and the domain-specific prior knowledge of the model, the new 241920 bits should be determined.

Accurate SR faces significant challenges due to these two major issues. One more crucial aspect that the super-resolved image must have, in addition to precision, is authenticity. Certain techniques, like interpolation, may offer fairly precise visuals, yet they appear fuzzy and can be

quickly distinguished as artificial. Because images are typically compared based on pixel values, models are compelled to produce outputs that closely resemble the ground truth. Nonetheless, two images that have an equal similarity (such as PSNR) may possess varying degrees of realism. Therefore, it is crucial to take this into account when designing the model. The new pixel values need to be chosen so that the resulting image is both accurate and visually realistic.

Given the mentioned issues, it is crucial to create a model capable of producing high-quality HR images from low-resolution inputs that are both precise and visually appealing. The most recent progress in deep learning is utilized in this thesis with an innovative strategy to create a new architecture that can meet the demand for a robust general super-resolution system. The system's performance is being evaluated to showcase its advantages from various viewpoints. The primary goal of the SR system is to achieve precision. The reconstructed high-resolution images need to be loyal to the original data. Various metrics are utilized to assess the quality of the system being proposed. Due to the inadequate metrics for evaluating SR system performance from the standpoint of a deep learning model, a new similarity metric is suggested.

Another key goal of this project is to suggest a universal SR system that can excel in different areas. Numerous SR systems are suggested in the literature for particular domains. These models are only capable of achieving high performance in a single type of data. Many SR systems are specifically developed for facial SR, for instance. The cause of this is that these models are created according to a particular type of data architecture. In addition, the specialized domain information is also used. Different methods, like using labeled data and pre-trained feature extractors, can be used to obtain specific information in this domain.

1.3.1 Nash equilibrium

The idea of Nash equilibrium, which comes from game theory, is relevant to GANs. In the field of game theory, a Nash equilibrium is when in a game no player can benefit from changing their strategy if the other players' strategies stay the same. In the context of GANs, there are two participants: the generator and the discriminator. The generator's job is to create data that is impossible to differentiate from real data. In the realm of GANs, a Nash equilibrium is achieved when the discriminator can differentiate effectively between genuine data and data generated by the generator. The generator creates data so lifelike that the discriminator is no longer able to differentiate it from actual data. The generator's output now accurately replicates the actual data distribution. When confronted with a flawless generator, the discriminator is compelled to make random guesses about the authenticity of its input, as the generated data is indistinguishable from real data. At this point of balance, neither the generator nor the discriminator can enhance their strategies more as long as the other's strategy remains

unchanged. The generator is producing extremely authentic data, while the discriminator is performing no better than making random guesses, suggesting that the training has reached an optimal level of convergence. Yet, achieving this milestone in real-world situations is difficult and typically necessitates meticulous adjustments to model designs, training methods, and loss functions.

Another challenge we might face while developing SRGAN is mode collapse, which can happen in SRGANs as well as other GANs. Mode collapse occurs when the generator in a GAN only learns to create a restricted number of outputs, regardless of the input it receives. In the realm of SRGANs, mostly utilized for converting low-resolution images into high-resolution ones, mode collapse can appear because of limited diversity in outputs. Therefore, the super-resolution images generated by the SRGAN might look too similar to each other despite variations in the input low-resolution images, indicating that the generator is only capturing a subset of the possible high-resolution representations. Additionally, in a mode collapse scenario, the high-resolution images produced might lack diversity in texture or fine details and may consistently exhibit certain patterns, textures, or artifacts regardless of the input image's content; and predictable artifacts, where the SRGAN might start producing specific, repeated artifacts in the super-resolution outputs, which can be a sign that the generator is stuck in a limited part of the solution space.

To mitigate mode collapse in SRGANs, various strategies can be employed, such as feature matching, which encourages the generator to match deeper features of the discriminator's representations, not just the output; mini-batch discrimination, which helps the discriminator to look at multiple examples at once, thereby discouraging the generator from always generating the same output; and adding noise, which involves injecting noise into the input of the generator or its layers to encourage variety in the output. These techniques aim to encourage the generator to explore a broader range of potential solutions, thereby reducing the risk of mode collapse and improving the overall performance and utility of the SRGAN.

1.3.2 Failure to convergence

In GANs, achieving convergence can be particularly challenging, often complicated by imbalances in the capabilities of the generator and discriminator. When the generator becomes too strong relative to the discriminator, it can lead to a degradation in the quality of generated images due to poor feedback from the discriminator. This section discusses the dynamics of this issue, theoretical insights, and potential remedies with academic references to provide a comprehensive understanding.

GANs consist of two neural networks, the generator (G) and the discriminator (D), which are trained simultaneously in a zero-sum game framework. The generator aims to produce data indistinguishable from real data, while the discriminator evaluates whether the data it receives is real (from the actual dataset) or fake (produced by the generator). This setup is supposed to help the generator improve over time, ideally leading to high-quality data generation. Non-convergence in GANs often occurs when there is a significant imbalance in the training dynamics of the generator and discriminator. If the generator is too strong, it can overwhelm the discriminator by producing highly convincing fakes before the discriminator adequately learns to distinguish between real and fake data. This leads to issues such as ineffective gradients, where a weak discriminator provides less meaningful gradients to the generator because it cannot accurately assess the generator's output, resulting in the generator not receiving proper feedback to adjust its parameters optimally and potentially stagnating in its learning process or adopting suboptimal strategies that do not reflect the data distribution; and overfitting of the generator, where a weak discriminator leads the generator to overfit to the limited capacity of the discriminator, optimizing for errors that are not relevant to producing realistic outputs but rather to exploit the discriminator's weaknesses.

Several strategies can be employed to mitigate the risk of non-convergence due to a too-powerful generator: adding a gradient penalty to the discriminator's loss, which helps stabilize training by penalizing the discriminator if it moves too far from the real data manifold, ensuring its gradients are meaningful and informative for the generator; using dynamic training rates, which involve adjusting the training rate of the generator and discriminator dynamically to maintain a balance between the two networks, ensuring that neither becomes too powerful too quickly; and adding noise to the discriminator’s inputs or to the gradients, which can help prevent the discriminator from overfitting to the current capabilities of the generator and ensure a more robust gradient signal.

1.3.3 Vanishing gradients

The problem of vanishing gradients in GANs often arises when the discriminator becomes too strong relative to the generator, severely impacting the training process and preventing the generator from learning effectively. When the discriminator is too strong, particularly in the early stages of training, it can easily distinguish between real and generated data, resulting in gradients that are close to zero because the discriminator is confident in its classifications (real or fake). This phenomenon leads to the vanishing gradient problem, where the generator receives very little useful feedback from the discriminator, as the gradient signals necessary for learning are minimal or negligible. Consequently, the generator struggles to learn effectively,

stalling its progress and causing the training process to stagnate.

Several techniques have been proposed to address the issue of vanishing gradients in GANs: modified training objectives, where adjusting the loss functions (such as using Wasserstein loss in Wasserstein GANs) can provide smoother and more useful gradients for the generator, even when the discriminator is strong; feature matching, which involves training the generator to match the statistical features of real data as identified by the discriminator's intermediate layers, ensuring more stable and meaningful gradient information for the generator; label smoothing, which applies to the discriminator's training to prevent it from becoming overly confident in its predictions, maintaining a gradient flow that benefits the generator; and adding noise to the discriminator's input or within the network itself, which helps prevent the discriminator from overfitting to the generated data and supports a more robust gradient signal for the generator.

1.3.4 A Large dataset

The challenge of requiring large datasets for training GANs effectively is significant, as GANs are data-hungry due to the generator's need to learn a complex distribution from the training data to produce realistic outputs. Smaller datasets may not provide enough variability and complexity, leading to issues like overfitting or failure to converge. Additionally, traditional data augmentation techniques, which are effective for supervised learning tasks, often do not translate directly to GANs. Some challenges involved in using small datasets include limited diversity, where a small dataset may not capture the full range of the data domain, making it challenging for the generator to learn a comprehensive distribution. This often results in less variety in the generated samples or repetition of similar patterns. Another issue is overfitting, as the generator might memorize specific examples rather than learning to generalize from the distribution, especially when the dataset is small.

Traditional data augmentation techniques, such as rotation, flipping, and scaling, are commonly used to artificially increase the size of training datasets in tasks like classification. However, these methods have limitations when applied to GANs. One challenge is consistency between real and fake data: data augmentation needs to be applied in a way that maintains consistency between the augmented real data and the data generated by the GAN. If only real data is augmented, the discriminator may learn to identify augmented samples as always real, creating a bias in its learning. Another issue is the distortion of data distribution, as some forms of augmentation may alter the underlying data distribution, potentially misleading the generator into producing unrealistic or undesirable outputs.

To address the challenges of training GANs with limited data, several techniques and

methodologies have been developed. Conditional GANs involve conditioning the generator and discriminator on additional information, like class labels, allowing for more efficient learning and better results with smaller datasets. Transfer learning leverages pretrained models on large datasets and fine-tunes them on smaller datasets, which is particularly effective when the pretraining is done on related tasks or data. Few-shot learning [52] and meta-learning [53] are designed to enable learning from very small datasets, and applying these principles to GANs can be helpful in data-limited scenarios. Additionally, using tailored data augmentation techniques for GANs—such as augmenting both the inputs to the discriminator and the outputs of the generator in a consistent manner—can mitigate some issues with traditional augmentation. Finally, feature matching and regularization techniques can guide the generator to focus on capturing the underlying data distribution rather than memorizing specific examples.

1.3.5 Creating artifacts

Using batch normalization in the structure of Super-Resolution Generative Adversarial Networks (SRGANs) can indeed introduce some artifacts in the generated high-resolution images. Batch normalization is a technique used to stabilize and speed up the training of deep neural networks by normalizing the inputs of each layer. While it has significant benefits in many contexts, in the case of GANs, and specifically in the architecture of SRGANs, its application can have unintended consequences. Batch normalization works by normalizing the inputs to a network layer to zero mean and unit variance, based on the statistics of the current batch. This helps to reduce internal covariate shift, where the distribution of network activations changes during training, thus speeding up convergence and allowing for higher learning rates.

Potential issues in SRGANs include inconsistency across batches, as batch normalization normalizes features across the batch, causing normalization statistics (mean and variance) to depend on the specific set of images in each batch. If the batch contains images with varying characteristics, the normalization may not be appropriate for all images, leading to artifacts in the output. Another issue is spatial artifacts: SRGANs aim to generate detailed textures and fine details, but local statistics within an image can vary significantly, especially in images with diverse content. Batch normalization can disrupt these local statistics, potentially causing spatial artifacts where certain areas of the image may appear unnaturally smooth or exhibit distorted textures. Finally, there is a potential loss of range, as batch normalization can limit the range of activation values, which may suppress some details the network might otherwise learn to generate. In super-resolution, maintaining a wide dynamic range is crucial for preserving subtle image details.

To mitigate issues in SRGANs related to batch normalization, several alternative

normalization techniques can be considered. Instance normalization normalizes input using the statistics of each individual image rather than the batch, making it suitable for tasks like style transfer and super-resolution where preserving individual image content is essential. Layer normalization normalizes across all features in the same layer independently for each example, which can be beneficial when independence between examples is important. Group normalization, a middle ground between batch and instance normalization, divides channels into groups and normalizes within each group, making it less dependent on batch size and effective in scenarios with small batch sizes. Conditional normalization techniques, like conditional batch normalization or adaptive instance normalization, allow normalization parameters to be conditioned on external data, adding flexibility and adapting to the specific needs of individual images in super-resolution tasks.

1.4 Main Research Content

The primary objectives of this research are twofold. First, this study aims to develop two novel and state-of-the-art Super-Resolution models by combining the architecture of SRGAN with the strengths of U-Net and Auto-encoder and the capabilities of transfer learning; a unique integration that has not been previously explored in the literature. These innovative models address critical limitations of traditional SRGANs, such as slow convergence, the creation of visual artifacts, reliance on large datasets comprising thousands of images, and challenges related to vanishing gradients. Although these models are tested on UAV imagery, they are designed to be broadly applicable, and capable of improving image resolution across diverse domains. their application to UAV datasets further underscores their effectiveness in enhancing the quality of aerial imagery, particularly for detecting and analyzing small objects in fields such as precision agriculture, environmental monitoring, and disaster management.

The second objective is to rigorously evaluate and compare the performance of these hybrid SRGAN models against leading super-resolution frameworks, including conventional SRGAN and Enhanced Super-Resolution GAN (ESRGAN). Through a detailed comparison, this study demonstrates the proposed models' superiority in overcoming the limitations of existing methods, including susceptibility to artifacts and restricted scalability. Beyond UAV applications, these models' adaptability extends to other fields, such as medical imaging, where enhanced image resolution can significantly improve diagnostics. By addressing the challenges of traditional SR techniques and demonstrating their versatility, this research represents a meaningful advancement in the field of super-resolution imaging.

This research makes a significant contribution to the existing body of knowledge by

introducing groundbreaking models for enhancing the resolution of UAV imagery; an innovation that is both original and unprecedented. Unlike earlier studies in the field of UAV imagery, which predominantly relied on traditional super-resolution models, this study pioneers the integration of SRGAN with U-Net architecture and transfer learning techniques, marking a transformative shift in remote sensing. By doing so, it addresses the unique challenges of UAV image resolution, such as the detection of small objects, variability in environmental conditions, and the inherent limitations of low-quality data acquisition.

Moreover, this work redefines the role of super-resolution in remote sensing applications, moving beyond conventional approaches to tackle domain-specific challenges like high-frequency detail recovery and artifact suppression. Tailoring the model to UAV imagery, not only enhances the clarity and accuracy of aerial images but also demonstrates the potential to optimize critical operations in precision agriculture, environmental monitoring, and disaster management. This research sets a new benchmark for the application of super-resolution techniques in UAV-based remote sensing, bridging the gap between theoretical advancements and real-world applications.

These hybrid models, proposed in this study, demonstrate a clear edge over existing GAN-based super-resolution models due to their superior ability to capture detailed information while reducing computational complexity. This reduction in complexity makes it particularly well-suited for real-time applications where super-resolution is critical, such as classification and object detection tasks in autonomous vehicles. For instance, in adverse weather conditions where capturing high-resolution images is challenging or when the storage of HR images is constrained by limited space, these models offer a practical and efficient solution. By leveraging its innovative architecture, the model enhances the resolution of low-quality images captured in suboptimal conditions without the need for expensive, high-quality cameras. This adaptability not only addresses real-world challenges but also positions the model as a valuable tool for scenarios requiring on-demand super-resolution, such as real-time decision-making in autonomous systems, where clarity and precision are paramount.

Furthermore, while these models were rigorously tested on aerial imagery and delivered significant results, their potential applications are not confined to this field. The novel architecture and capabilities of this model make it versatile and adaptable to a wide range of domains where super-resolution is essential. Whether it is enhancing medical images for improved diagnostics, refining satellite imagery for geospatial analysis, or improving visual data in surveillance systems. these models demonstrate the flexibility to address diverse challenges. their ability to recover fine details and produce high-quality results with reduced

computational complexity ensures that they can seamlessly integrate into various workflows, particularly in areas where high-resolution imagery is critical but challenging to obtain. This versatility underscores the models' broader impact, paving the way for advancements in fields where super-resolution can significantly enhance decision-making and operational efficiency.

1.5 Organizational Structure of Thesis

This thesis, as illustrated in figure 1-1, is structured into five primary chapters, with each chapter addressing a distinct facet of the research on the application of hybrid SRGAN to enhance the resolution of satellite images. The framework is thoughtfully constructed to ensure a coherent and logical presentation of information, leading the reader through the research journey and guaranteeing a thorough grasp of the methods, results, and significance of the study. Every chapter expands upon the last, enhancing the overall consistency and richness of the thesis.

1) Chapter 1: Introduction

This chapter provides the foundation and context for the research, delving into the broader field of super-resolution in UAV applications. It emphasizes the study's significance across various industrial domains, identifies the research gap and the specific challenges the thesis aims to address, and presents the primary and secondary objectives. Additionally, the chapter underscores the academic and practical relevance of the research, defines the study's scope, and acknowledges its inherent limitations.

2) Chapter 2: Literature Review

The aim of this chapter is to provide a comprehensive overview of the methodologies and techniques that underpin super-resolution, with a particular focus on the evolution from traditional approaches to modern deep learning-based solutions. It delves into the advancements brought by GANs, U-Net architectures, and autoencoders, highlighting their roles in enhancing image quality across various domains.

3) Chapter 3

Chapter 3 describes the methodology, detailing the design and implementation of a novel super-resolution GAN model that incorporates advanced architectures like U-Net and attention mechanisms, in addition to pretraining the U-net using an autoencoder to enhance resolution in remote sensing images.

4) Chapter 4

The fourth chapter presents the experimental process and results, comparing the proposed models' performance with baseline methods and assessing their effectiveness through

qualitative and quantitative evaluations.

5) Chapter 5

The last chapter offers a synthesis of the key findings from this research, evaluates the scope and limitations of the study, and proposes future research avenues in the domain of super-resolution using generative adversarial networks

The codes used in this project are available from these repositories:

https://github.com/Amir-Rouhbakhsh/super-resolution

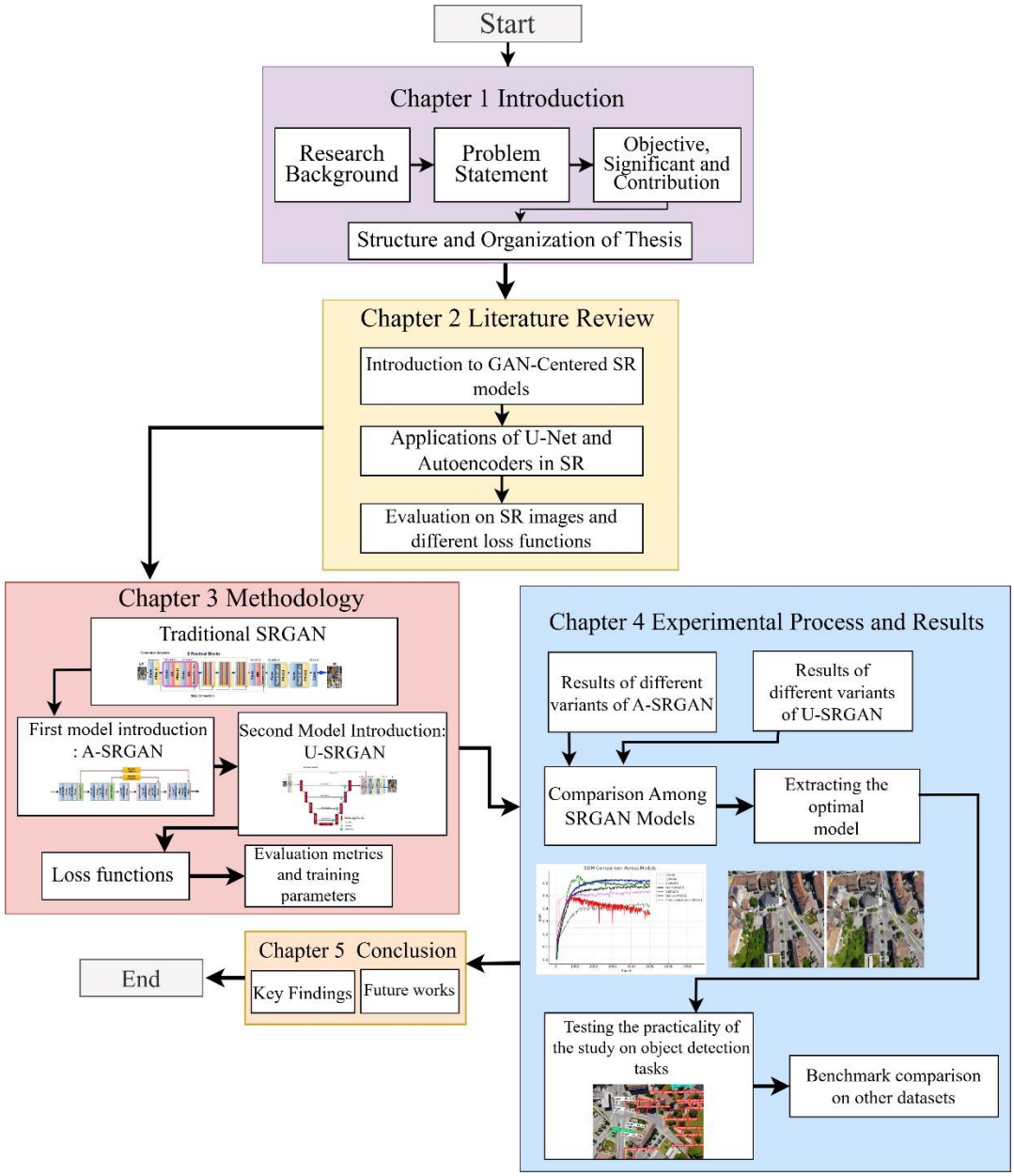

Figure 1.1 Thesis Organization

2 Literature Review

2.1 Introduction

The pursuit of super-resolution has been a significant focus in image processing, with numerous techniques developed over the years to enhance image quality. Traditional methods, such as interpolation techniques like bilinear [54], bicubic [55], spline [56], and edge-directed interpolation[57], have laid the groundwork by estimating unknown pixel values based on surrounding data points. Additionally, frequency-domain approaches, such as Fourier Transform-Based Methods [58], [59], have offered an alternative by manipulating images in the frequency rather than spatial domain. Other notable traditional SR techniques, including neighbor embedding [60], [61] and sparse representation [62], have employed statistical and mathematical frameworks to reconstruct high-resolution images.

The advent of deep learning revolutionized the field of super-resolution, marking a significant shift in capability and performance. A key milestone was achieved by Chao Dong et al. with the introduction of the Super-Resolution Convolutional Neural Network (SRCNN) [63], which paved the way for a new era of CNN-based SR techniques. Following SRCNN, several advanced models emerged, such as FSRCNN [64], ESPCN [65], VDSR [66], DRCN [67], and EDSR [68], each contributing to the rapid progression of deep learning-based SR.

In this chapter, we focus on the techniques explored in this study, emphasizing the use of GANs and their transformative impact on SR. We also delve into the application of U-Net and autoencoders, which offer robust architectures for feature extraction and reconstruction. Additionally, we discuss peripheral techniques, such as Adaptive Discriminator Augmentation (ADA) [69], which play a critical role in boosting the performance of SR models. This review aims to provide a comprehensive introduction to these advanced methodologies, forming the foundation for the experiments and analysis presented in subsequent chapters.

2.2 GAN-based super resolution models

Generative Adversarial Networks (GANs), introduced by Ian Goodfellow in 2014 [70], have significantly influenced advancements in image generation and enhancement [71], [72], [73]. These networks involve a generator that creates synthetic data and a discriminator that assesses its authenticity. The adversarial relationship between these two networks drives them to iteratively improve, culminating in outputs that are indistinguishable from real data. This dynamic process, characterized by a min-max game, achieves Nash equilibrium when the discriminator can no longer reliably differentiate between real and generated data. The implicit ability of GANs to learn data distributions without explicit sampling makes them ideal for tasks

such as image super-resolution.

The introduction of Super-Resolution GAN (SRGAN) in 2017 by Ledig et al. [50] marked a transformative application of GANs in enhancing image resolution. SRGAN employs a generator network based on deep residual networks (ResNet) and a discriminator network to create high-resolution images from low-resolution inputs. The generator leverages residual blocks, which include skip connections to maintain performance in deeper networks. The discriminator, with convolutional layers and increasing filter sizes, evaluates the authenticity of generated images. To balance content fidelity and perceptual quality, SRGAN uses a hybrid loss function, combining pixel-based MSE loss, a VGG-based perceptual loss for high-level content preservation, and adversarial loss to enhance visual realism. Although SRGAN significantly advanced the field, its tendency to prioritize perceptual quality sometimes led to deviations from the ground truth, especially in finer details.

Building on SRGAN, the Enhanced Super-Resolution GAN (ESRGAN) introduced in 2018 by Wang et al. [74] addressed several limitations to produce sharper and more detailed images. By removing batch normalization layers, ESRGAN reduced artifacts and improved generalization across datasets. It also introduced Residual-in-Residual Dense Blocks (RRDB), which combined residual and dense connections, allowing for deeper feature extraction and improved learning. The relativistic discriminator in ESRGAN evaluated the relative authenticity of real and generated images, stabilizing training and enhancing output quality.

Additionally, ESRGAN refined its loss functions, using perceptual loss on pre-activation features to ensure consistent brightness and sharper details, while pretraining with a PSNR-focused loss provided robust initial learning. Network interpolation further balanced perceptual quality and distortion by blending weights from pretraining and adversarial training phases, enabling flexible adjustments without retraining. These innovations enabled ESRGAN to outperform SRGAN in producing high-quality textures and finer details, though it required greater computational resources.

Real-ESRGAN, introduced in 2021 [75], extended ESRGAN by addressing challenges associated with real-world image degradations. Through high-order degradation modeling, Real-ESRGAN simulated complex artifacts such as ringing and overshoots, enhancing its robustness in practical applications. It incorporated a U-Net-based discriminator with spectral normalization, improving stability and training performance. By training with sharpened ground-truth images, Real-ESRGAN further enhanced visual clarity and reduced overshoot artifacts. These advancements allowed Real-ESRGAN to achieve state-of-the-art performance in blind super-resolution tasks, demonstrating its capability to handle diverse and challenging

degradations while preserving natural textures.

Through their successive developments, SRGAN, ESRGAN, and Real-ESRGAN highlight the evolving application of GANs in super-resolution tasks, showcasing how architectural innovations and loss function refinements have pushed the boundaries of image enhancement. These models collectively emphasize the potential of GANs to address both perceptual quality and fidelity in increasingly complex real-world scenarios.

Beyond SRGAN and ESRGAN, several other Generative Adversarial Networks have been developed to tackle super-resolution tasks, each introducing unique mechanisms to improve the quality and fidelity of generated images. One notable model is the Progressive Growing GAN for Super-Resolution, which employs a progressive training approach [76]. By gradually increasing the resolution of generated images during training, this model ensures stability and enables the synthesis of high-quality outputs, particularly for tasks involving significant upscaling factors.

Another innovative approach is the Conditional GAN for Super-Resolution, which incorporates contextual information alongside low-resolution inputs. By conditioning the generator on additional data such as edge maps or semantic labels, the model enhances the contextual accuracy of high-resolution outputs [77]. Similarly, the Feedback GAN introduces a feedback mechanism where the generator receives intermediate feedback from the discriminator. This iterative refinement process leads to sharper and more perceptually realistic images [78].

Other approaches, such as the Multi-Scale GAN, utilize a multi-scale architecture that processes input images at varying resolutions simultaneously. This design allows the generator to capture both global coherence and fine-grained details, producing visually consistent results [79]. These additional GAN-based models showcase the continuous evolution and diversification of techniques in super-resolution. By introducing mechanisms such as progressive training, attention guidance, feedback loops, and multi-scale processing, these models extend the capabilities of GANs, addressing both general and specific challenges in generating high-quality, high-resolution images.

2.3 Applications of U-net in Super resolution tasks

U-Net's architecture is distinguished by its encoder-decoder structure with skip connections, enabling the preservation of spatial information throughout the network. This design is particularly advantageous for tasks that require high-resolution outputs derived from low-resolution inputs. In the U-Net framework, the encoder progressively reduces spatial

dimensions while increasing feature depth, and the decoder reconstructs the image to its original size by utilizing features captured during the encoding phase. This versatility has led to its widespread application across various fields, particularly in domains requiring precise image segmentation and reconstruction.

One of the most prominent applications of U-Net is in medical imaging, where it is extensively used for segmentation tasks across modalities such as CT scans, MRI, and X-rays. Variants like TransUNet, which integrate U-Net with Transformer architectures, have further enhanced performance in complex tasks such as multi-organ segmentation and cardiac imaging [80]. Beyond medical imaging, U-Net has also found use in UAV image processing, particularly for tasks like land cover and crop classification. Its adaptability to diverse environmental conditions makes it a valuable tool, although challenges such as limited dataset availability remain significant [81].

In the field of cell segmentation, U-Net has proven effective in identifying and delineating cells across various imaging modalities, including microscopy. This capability has been critical in advancing cellular research and diagnostics [82]. Similarly, U-Net has been adapted for skin lesion analysis, enabling accurate segmentation of dermatological images to distinguish between healthy and malignant tissues, thereby aiding in skin cancer detection [83]. Another important application is nucleus segmentation in histopathological images, where U-Net's precision is invaluable for cancer diagnosis and research.

Through its robust architecture and adaptability, U-Net has demonstrated remarkable versatility across diverse applications, establishing itself as a cornerstone in image analysis tasks requiring high-resolution outputs and detailed segmentation. Originally designed for biomedical image segmentation, U-Net has emerged as a versatile model in various image processing tasks, including super-resolution. Its encoder-decoder architecture with skip connections effectively preserves spatial information, making it a powerful tool for enhancing image resolution across multiple domains. This review explores the applications and adaptations of U-Net for SR, emphasizing its robustness and effectiveness.

2.3.1 Key Applications in Super-Resolution

One of U-Net's most impactful applications lies in medical imaging, where precision is critical. In the domain of MRI enhancement, U-Net-based architectures such as MRI-Net have been developed to map low-resolution brain MRI scans to high-resolution counterparts. Utilizing a Mean Squared Logarithmic Error (MSLE) loss function, MRI-Net achieves superior performance in terms of Peak Signal-to-Noise Ratio (PSNR) metrics, significantly improving diagnostic accuracy for neurological disorders like Alzheimer’s and Parkinson’s disease [84].

Beyond MRI, U-Net has been adapted to enhance various medical imaging modalities, demonstrating its ability to preserve fine details essential for accurate diagnoses. Its skip connections, which mitigate overfitting, make it particularly suitable for small medical datasets.

Robust U-Net Variants (RUNet) have further expanded U-Net's capabilities in SR. RUNet incorporates modifications that learn spatially varying degradations during training, improving visual quality while maintaining low reconstruction errors. This architecture has shown marked improvements over traditional methods, making it a valuable tool for diverse SR applications [85].

Outside of medical imaging, U-Net has been applied in general image processing, including satellite imagery and general photography. Its ability to upscale images while preserving essential details makes it a popular choice for improving visual content in various fields, such as remote sensing and consumer media [86], [87], [88], [89].

U-Net consistently outperforms other deep learning architectures in SR tasks. Comparisons with fully convolutional networks (FCNs) and residual networks (ResNets) have demonstrated that U-Net achieves higher PSNR values while maintaining superior structural fidelity in reconstructed images. This advantage is particularly significant in applications requiring the preservation of fine details, such as medical diagnostics and high-quality visual media.

The continuous evolution of U-Net's architecture and loss functions has further enhanced its capabilities in SR. Ongoing research aims to integrate U-Net with emerging technologies, such as vision transformers, to push the boundaries of super-resolution. These hybrid models hold the potential to achieve even greater performance, paving the way for more sophisticated image reconstruction techniques.

U-Net's application in super-resolution tasks highlights its adaptability and effectiveness across diverse domains. Its robust architecture and ability to preserve spatial and structural details have made it a cornerstone model in SR. As deep learning continues to advance, U-Net is poised to remain a foundational tool for high-quality image reconstruction, with future developments likely to expand its impact even further.

2.3.2 Advantages and Limitations of U-Net in Super-Resolution

The U-Net architecture offers several distinct advantages that make it a popular choice for super-resolution tasks across various domains:

1. Multi-Scale Feature Extraction: U-Net's encoder-decoder structure, augmented with skip connections, enables effective multi-scale feature extraction. This allows the model to capture both low-level details and high-level semantic information, which are

essential for reconstructing high-resolution images from low-resolution inputs [90].

2. Noise Suppression: The architecture is inherently designed to suppress noise while enhancing image details. Its ability to reconstruct images from coarse to fine makes it particularly effective in scenarios involving low-quality inputs, such as low-light or noisy images [91].
3. Flexibility and Adaptability: U-Net is highly adaptable and can be modified to include additional components, such as attention mechanisms or multiple decoders. These adaptations improve performance in specialized applications, including medical imaging and remote sensing [92].
4. High Performance Across Diverse Applications: U-Net has demonstrated exceptional performance in various domains, such as medical imaging, seismic data analysis, and optical microscopy. It often surpasses state-of-the-art methods in metrics like PSNR and SSIM, underscoring its robustness and reliability [84].

These advantages highlight U-Net's effectiveness in delivering high-quality super-resolution results, making it a preferred choice for both academic research and practical applications.

Despite its strengths, U-Net also has notable limitations that can affect its performance in certain scenarios:

1. Limited Generalization: U-Net models are typically trained for specific upscaling factors, which restricts their ability to generalize to unseen factors during testing. This limitation reduces scalability in real-world applications where variable upscaling factors are required.
2. Challenges in Edge and Detail Recovery: Standard U-Net architectures often struggle to recover fine details and sharp edges, particularly in noisy environments. This can result in artifacts or the loss of critical features in the reconstructed images.
3. High Computational Complexity: U-Net's architecture can be computationally demanding, requiring significant memory and processing power, especially when dealing with high-resolution images. This poses challenges for deployment in resource-constrained environments.
4. Sensitivity to Input Quality: The quality of input low-resolution images greatly influences U-Net's performance. Poor-quality inputs can lead to suboptimal super-resolution results, as the model may fail to effectively learn meaningful features from inadequate data.

These limitations underscore the need for ongoing research to address U-Net's shortcomings and further enhance its applicability in super-resolution tasks. By refining its architecture and

developing innovative adaptations, U-Net can continue to evolve as a cornerstone model in image processing.

2.4 Autoencoder in SR

Autoencoders have emerged as powerful tools for enhancing image resolution, leveraging their ability to learn complex representations of data. Various studies demonstrate their effectiveness in generating high-resolution images from low-resolution inputs through innovative architectures and techniques. Rather than implicitly exploring the data distribution as a GAN will do, a variational autoencoder (VAE) explicitly explores the data distribution for modeling. Liu et al., 2021[93] propose a reference-based super-resolution model that learns patterns from a reference to guide the super-resolution process. To do this, the reference patterns are compressed into a latent space using Conditional Variational Inference to learn an explicit probability distribution, and then these patterns are re-sampled prior to super-resolve data. Variational autoencoders have been also employed for photo-realistic image super-resolution. They generate high-resolution images by learning the conditional distribution of high-resolution images from low-resolution inputs, balancing super-resolution distortion and perceptual quality [94].

The integration of VAEs with channel attention mechanisms has shown significant improvements in image quality. This approach enhances the model's ability to distinguish between generated and real high-resolution images, achieving higher PSNR and SSIM values [95]. Dual U-Nets Autoencoders are also used for hyperspectral image super-resolution by fusing them with high-resolution multispectral images. The dual U-Nets architecture enhances the interaction between the data, improving the resolution [96]. Convolutional Autoencoders with Skip Connections is an approach that uses symmetric convolutional and deconvolutional layers with skip connections to enhance image resolution. It has demonstrated high accuracy on diverse datasets [97].

2.5 Evaluation on SR images

As generation models such as GANs and diffusion models [98], [99] continue to advance quickly, it is becoming more crucial to assess the perceptual quality of images in computer vision applications [100], [101]. This evaluation determines the realism of an image. The mean opinion score (MOS) is considered the most dependable method for evaluating perceptual quality, where experienced raters provide scores for reconstructed images based on various criteria (such as sharpness, artifacts, contrast, and exposure) and then calculate the average score. For instance, in specific medical image restoration tasks, skilled radiologists categorize

images on a scale of 0 to 4 based on quality (e.g., non-diagnostic, poor, fair, good, and excellent). At times, the evaluator may assess low perceptual quality aspects like low signal-to-noise ratio and motion artifacts. There are no standardized metrics to assess the precision and visual appeal of super-resolution images, making it costly and time-consuming to obtain the relatively dependable MOS from human observers [102]. PSNR and SSIM are frequently employed metrics to assess image quality when humans are not present. Numerous studies, such as that by Ledig et al. in 2017 [50], employ MOS testing. In particular, people are requested to evaluate images using a numerical scale, which is then used to calculate a final score for evaluating an algorithm. Nevertheless, since standardizing this trait is challenging and utilizing it in research is complex, many studies concentrate on quantifying image quality through mathematical methods.

Usually, assessment metrics for image quality in SR images consist of subjective and objective methods. The first option can effectively capture human perception, but it requires manual scoring that is time-consuming and can be affected by differences between scorers or within the same scorer. Objective methods are simple to calculate and impartial for comparison, yet they typically concentrate on just one aspect of image quality assessment. Therefore, different metrics are utilized for a comprehensive assessment of SR images, including the fidelity of reconstruction and the quality perceived by the viewer. This part presents the top objective metrics used to evaluate the quality of reconstructed images.

2.5.1 Peak signal-to-noise ratio

The most commonly used evaluation metric for tasks involving image restoration (such as reconstruction, super-resolution, and denoising) is peak signal-to-noise ratio (PSNR). PSNR is commonly utilized in image processing as a metric derived from MSE. In the next formula, L is the dynamic range of allowable image pixel intensities, e.g., an 8-bit image will have an L value of $8^2 - 1 = 255$. This allows the MSE to easily be used to compare images of different bit depths [103].

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \qquad (2-1)$$

As the Mean Squared Error (MSE) approaches zero, the PSNR value increases, indicating a higher level of similarity and eventually approaching infinity. PSNR is utilized in various scenarios, and it simplifies comparisons since it is seen as the standard measurement. Nevertheless, it is commonly acknowledged that PSNR does not have a strong correlation with how humans perceive image quality [104], [105]. For instance, a picture that has been slightly altered in its geometry might have a high MSE when compared to the original image, but still

look the same to a person viewing it. The same holds true in the opposite direction, where a picture affected by either additive white Gaussian noise or blurring could have a low MSE compared to the original image, but still look noticeably different. These problems emphasize the idea that the Human Visual System (HVS) is well-suited for extracting structural information [106], while metrics like PSNR focus on pixel-level information.

2.5.2 Structural Similarity Index

Due to drawbacks in PSNR, various alternative measures like Structural Similarity Index have been suggested. SSIM directly assesses the signal variations between two signals with complex structures. SSIM compares the luminance, contrast, and structure of the image signal individually, according to Wang et al. in 2004 [106]. While PSNR focuses on the squared error of pixels, SSIM accounts for the fact that pixels are highly interconnected, particularly when in close proximity. SSIM evaluates interdependence by eliminating contrast and luminance from an image and then analyzing structural characteristics. The luminance of two aligned image patches x and y from the same location is determined by comparing the average signal intensity of each image. These values are designated as μ_x and μ_y. In this method, brightness is determined by the combination of light intensity and surface reflection. In order to determine contrast, the average signal intensity is subtracted from each pixel value to eliminate luminance. Difference is described as a comparison of the standard deviation of each signal, $\sigma_x$and σ_y. In order to assess structural variances, the signals are initially normalized by dividing them by their respective standard deviations, and then the covariance, σ_{xy}, is computed. The structural information depicts the structure without being influenced by contrast and luminance. The equations that describe luminance, contrast, and structure are:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \qquad (2-2)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_1} \qquad (2-3)$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \qquad (2-4)$$

In these equations, several constants are introduced, namely $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3$. These serve to avoid instability when standard deviation values are close to zero. The entire SSIM is defined as:

$$SSIM(x,y) = [l(x,y)^{\alpha}] \cdot \left[c(x,y)^{\beta}\right] \cdot [s(x,y)^{\gamma}] \qquad (2-5)$$

where α and β and γ are the weightings of the luminance, contrast, and structure which is used to adjust the relative importance of each. In this study, these were left as the default of 1. In order to measure SSIM for a whole image, a sliding window is employed to traverse the image, one pixel at a time. The SSIM index is computed at every stage to produce a quality index map for SSIM. This quality map is basically a distortion matrix of the two images, indicating differences from a perfect image through the SSIM index. The average of this quality map is utilized for determining the overall image quality. SSIM values vary from 0 to 1 with a value of 1 indicating a perfect reconstruction.

2.6 Loss functions

For a significant period, even though it was crucial to a network's learning process, the loss function was often overlooked by the image-processing research community. Lately, there has been a shift, and the emphasis of research has turned towards enhancing or creating loss functions for a particular algorithm or issue, making it a key area of interest [104], [107]. The primary loss function for regression tasks is Mean Squared Error (MSE) or $\mathcal{L}_2$, which is commonly applied in various applications such as super-resolution. MSE's popularity can be attributed to its convexity and differentiability, as well as its common availability in user-friendly software packages. Furthermore, MSE is uncomplicated and straightforward, as well as relatively inexpensive in terms of computation. MSE, in the realm of image processing, measures the disparity between the initial image and the altered image. In super-resolution, it quantifies the error between the HR and upsampled LR images. If N is the number of pixels and $x = (x_i | i = 1,2,\cdots,N)$ and $y = (y_i | i = 1,2,\cdots,N)$, then this can be defined by the equation:

$$MSE(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \qquad (2-6)$$

In the context of SR, MSE or $\mathcal{L}_2$ can be defined as:

$$\mathcal{L}_2(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \frac{1}{H * W}\sum_{(i,j)\in I} \|\boldsymbol{I}_{hr}[i,j] - \boldsymbol{I}_{sr}[i,j]\|^2 \qquad (2-7)$$

where H and W are the height and width of the images, and $\boldsymbol{I}\ [i,j]$ denotes a pixel. Initial investigations into SISR, such as SRCNN, VDSR, and SRGAN, have shown a preference for $\mathcal{L}_2$ loss due to its direct connection to the commonly used evaluation metric PSNR. Nonetheless, it is responsive to extreme values as it squares the deviations. It often results in more consistent solutions, which can lead to overfitting and excessively smoothed outcomes.

The ℓ_1 loss is also widely used, and differs from $\mathcal{L}_2$ in that it does not overpenalize larger errors, and may therefore have different convergence properties . The ℓ_1 loss can be defined as:

$$\ell_1(x, y) = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \qquad (2-8)$$

And it also can be defined for SR applications as:

$$\ell_1(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \frac{1}{H * W}\sum_{(i,j)\in I} \| \boldsymbol{I}_{hr}[i,j] - \boldsymbol{I}_{sr}[i,j] \| \qquad (2-9)$$

The ℓ_1 loss is commonly utilized as well and stands apart from $\mathcal{L}_2$ because it does not overly punish bigger mistakes, potentially leading to distinct convergence behaviors. However, ℓ_1 loss is less affected by outliers due to its linear error properties. It is able to maintain edges and produce more attractive visual outcomes. Therefore, the $\ell_1$loss is increasingly utilized in many works such as [95], [108], [109], [110], [111] to enhance performance. It is important to observe that the pixel-wise loss functions lack global structures, therefore they do not enhance perceptual quality or realistic texture generation, leading to unnatural artifacts and reduced details.

2.6.1 Perceptual loss

Different authors have applied the pixel-based loss method in projects involving super-resolution, colorization, and similar tasks, but they fail to preserve the stylistic variations present in the final image compared to the reference image. Ideally, in super-resolution, fine details are deduced from visually unclear low-resolution images. Both MSE and SSIM have shown little correlation with human evaluation of visual quality, as they both focus on capturing small distinctions in pixel values. The purpose of the perceptual loss function is to identify discrepancies using high-level features instead of pixel-based variations, making style transfer possible. The goal of style transfer is to combine the content of a specific target image with the style of a different target image. In this framework, content pertains to the larger spatial compositions in the image while style relates to the colors and distinctive structures within the image. The reason behind this transfer is the realization that the upper layers (or layers nearer to the end) of the network grasp the overall content of objects and their placement in the input image, without storing specifics about pixel values.

On the other hand, the layers positioned nearer to the input in a network gather details about the appearance of an image rather than its overall layout. The main discovery is that there

is some degree of separability between style and content representations. As CNNs are trained for object recognition, they create a more detailed representation of the image, focusing more on content rather than style as they progress through the network [112]. This understanding has led to the application of style transfer in various areas of deep learning, including artistic creation and super-resolution.

In order to transfer the style of one image to the content of another image, it is necessary to create loss functions that enable this process. Johnson et al., 2016 [113] utilize a two-component network - an image transfer network and a loss network - to assist in super-resolution. A pre-trained network for image classification is utilized as a fixed loss network, as current networks have already mastered encoding perceptual and semantic information. The network frequently utilized is VGG pre-trained on either ImageNet or the MS-COCO dataset [113], [114]. The reason for perception loss involves calculating a Feature Reconstruction Loss, also called Content Loss, that aims to make the pixels of the output image resemble the target loss by reducing the squared, normalized Euclidean distance between the feature map of the output shape from the image transfer network and the feature map from the target loss. This is based on the idea that the feature maps in the deeper convolutional layers of a network capture the larger-scale features of the original image. Preservation of image content and spatial structure is maintained in the reconstructed image from higher layers, although there may be differences in color, texture, and exact shape. The resulting image is similar in perception, but not identical. The equation for Content Loss is as stated:

$$ContentLoss(\vec{p},\vec{x},l)=\frac{1}{2}\sum_{i,j}\left(F_{ij}^{l}(g)-F_{ij}^{l}(c)\right)^{2} \qquad (2-10)$$

Where $F_{ij}^{l}(g)$ refers to the ground truth feature map for layer l at jth position of the ith feature and $F_{i}j^{l}(c)$ refers to the predicted feature map for layer l at ith feature and jth position.

Johnson et al. employed perceptual loss to enhance the resolution of images by 4 and 8 times, utilizing the VGG-19 network (illustrated in Figure 2-1) trained with MS-COCO data for the feature reconstruction loss (content loss). The network's goal is to transfer semantic knowledge from the pre-trained loss network to the super-resolution network. Style loss is typically excluded in the super-resolution process as it is believed that the current image style is sufficient for the task. In the post-processing stage, a histogram matching was carried out between the network's output and the initial LR data. Images produced have sharp edges and are significantly clearer than pixel-based loss methods for comparison. Still, certain grid-like patterns and artifacts can be observed in the images on a pixel level, which ultimately decreases

the PSNR and SSIM values [113].

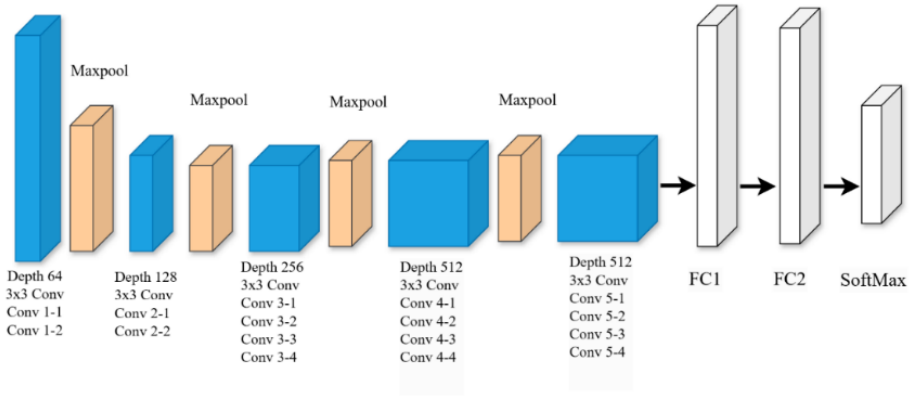

Figure 2-1 VGG-19 architecture: Typically, the style feature maps are taken from the first convolutional block, whereas the content feature maps are from the fifth block.

The VGG network utilized was pretrained on ImageNet, and it is useful to examine sample images from this database. The images in ImageNet vary greatly in both color and texture compared to patches generated from satellite imagery. Therefore, in this thesis, we also explore the utilization of a sparse autoencoder to capture feature maps from the original HR satellite data in order to enhance the representation of relevant feature maps. Autoencoders offer a method to extract characteristics from unlabeled data without supervision. Data is compressed in an autoencoder by passing it through a neural network to a latent space, which is then utilized for output reconstruction. Put simply, the network attempts to create a close representation of the original input. By creating a bottleneck such as limiting the number of hidden units, different structures in the data can be discovered [115], [116], [117], [118].

2.6.2 Adversarial loss

GANs are widely used in single-image super-resolution tasks for perceptually more realistic results. As a specific task of conditional GANs, the generators take the low-resolution image as input, instead of the noise vector. During training, the discriminator considers the generated super-resolved images as fake images and the ground truth high-resolution images as real images. The vanilla adversarial loss function is defined as [101]:

$$\mathcal{L}_{GAN} = -\mathbb{E}_{\mathbf{I}_{hr}}[\log D(\mathbf{I}_{hr})] - \mathbb{E}_{\mathbf{I}_{lr}}\left[\log\left(1 - D\left(G(\mathbf{I}_{lr})\right)\right)\right] \qquad (2-11)$$

where $\mathbb{E}$ is the expectation of the whole dataset. This vanilla adversarial loss first applies to SISR tasks in SRGAN, where the same discriminator to DCGAN [119] is used. It has successfully achieved photo-realistic natural images with $\times 2$ and $\times 4$ magnifications but requires a time-consuming warm-up of the generator to stabilize the training of GANs.

Advanced research on GANs in super-resolution are mainly about applying novel adversarial loss functions, such as relativistic GAN [120], Wasserstein GAN variants [121], [122], and cycle GAN [123].

2.7 Perception-Distortion Trade-off

Blau and Michaeli [124] claim that a trade-off exists between the perceptual qualities and the distortion level of an image when it is reconstructed from noisy data or super-resolved. In this situation, distortion is the difference between the reconstructed image $\hat{x}$ and the original image x. The perceptual quality of $\hat{x}$ pertains to its visual appearance, regardless of how closely it resembles x. In simpler terms, the perceived quality of $\hat{x}$ determines its resemblance to a genuine natural image. x belongs to a group of natural images p_X, while $\hat{x}$ is part of a group of generated images $p_{\hat{X}}$. Minimal perceptual distinctions are present when the spread of created images aligns with that of natural images, or when the disparity function $d(p_X, p_{\hat{X}})$ nears 0, corresponding to the Human Visual System but still not completely comprehended. On the other hand, MSE calculates a mean value across all potential explanations that may not accurately represent a valid image, possibly falling outside the range of natural images [125].

Blau and Michaeli discovered empirically and through mathematical evidence that the connection between distortion and perceptual qualities follows a convex curve. Enhancing an image's perceptual qualities to closely match the ground truth involves altering the image, and the same goes for the opposite scenario. In practical terms, there is a perfect point on the curve where distortion and perceptual accuracy are in equilibrium, based on the image's intended use. The writers assess various super-resolution algorithms based on the perceptual and distortion qualities of the resulting images. GANs are located on the perceptual end of the spectrum, while feed-forward networks using MSE loss are situated on the other end where distortion is reduced, as depicted in Figure 2-2. In this spectrum, there is a part that cannot be reached where the perception is excellent and distortion is kept to a minimum.

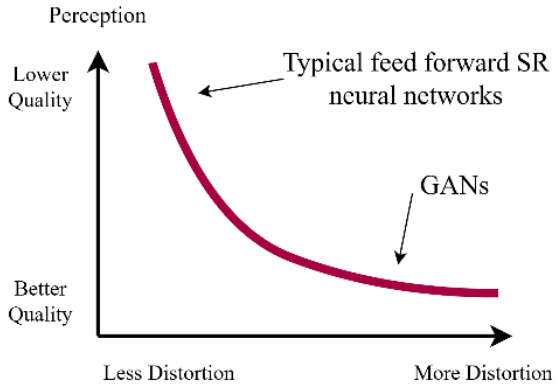

Figure 2-2 Perception-distortion trade-off, showing how improving an algorithm in terms of perception occurs only at the expense of increasing distortion and vice versa.

According to this study [123], a GAN is an ideal platform to examine this equilibrium, thus, the loss function in a GAN can consist of an adversarial loss and a distortion loss. This can be defined as:

$$\ell_{GAN} = \ell_{distortion} + \ell_{adversarial} \qquad (2-12)$$

where the distortion loss is typically an MSE loss, and the adversarial loss is the standard GAN adversarial loss which measures the deviation of the generated images from data distribution. By increasing the relative portion of MSE versus adversarial loss in a GAN, it is possible to move along the perception-distortion curve and move from a blurry accurate image to a sharp but less accurate image.

2.8 Summary

This chapter provides a comprehensive exploration of techniques and advancements in super-resolution, from traditional methods to cutting-edge deep learning approaches. Traditional SR techniques, such as interpolation and frequency-domain methods, laid the foundation for image enhancement. However, the advent of deep learning revolutionized SR, with models like SRCNN and its successors introducing significant improvements in performance and scalability.

GANs have emerged as transformative tools in SR, with models like SRGAN, ESRGAN, and Real-ESRGAN setting benchmarks in image quality. These models leverage adversarial training, perceptual loss functions, and innovative architectural features to produce visually realistic and detailed high-resolution images. Extensions such as Progressive Growing GANs and Conditional GANs further demonstrate the versatility and adaptability of GAN-based frameworks.

In addition to GANs, U-Net and autoencoder architectures have proven effective in SR tasks, particularly in domains requiring precise reconstructions, such as medical imaging. U-Net's encoder-decoder design with skip connections facilitates multi-scale feature extraction, noise suppression, and adaptability, while autoencoders contribute through advanced latent-space learning and integration with attention mechanisms.

Evaluation metrics like PSNR and SSIM, alongside emerging perceptual loss functions, provide insights into the fidelity and realism of reconstructed images. However, challenges such as the perception-distortion trade-off highlight the need for balanced approaches in optimizing SR models for practical applications.

This chapter underscores the rapid evolution of SR techniques, emphasizing the role of advanced architectures, innovative loss functions, and evaluation strategies in pushing the boundaries of image reconstruction and enhancement. These insights set the stage for the experimental investigations and methodologies presented in subsequent chapters.

3 Methodology

3.1 Introduction

In this chapter, we explore the architecture and characteristics of the Super-Resolution Generative Adversarial Network, followed by an overview of techniques designed to improve its performance. SRGAN represents a significant advancement in the field of super-resolution image reconstruction, particularly due to its novel use of content loss in place of the traditional mean squared error as the loss function. This paradigm shift allowed SRGAN to generate higher-quality images, emphasizing perceptual fidelity over pixel-wise accuracy. SRGAN marked a pivotal moment in the development of super-resolution methods, primarily due to its innovative use of perceptual loss functions. By focusing on image quality as perceived by human observers, rather than solely relying on pixel-based loss measures like MSE, SRGAN achieved superior results in generating high-resolution images. This methodological shift laid the foundation for numerous improvements, which continue to push the boundaries of image super-resolution techniques today.

3.1.1 SRGAN Framework

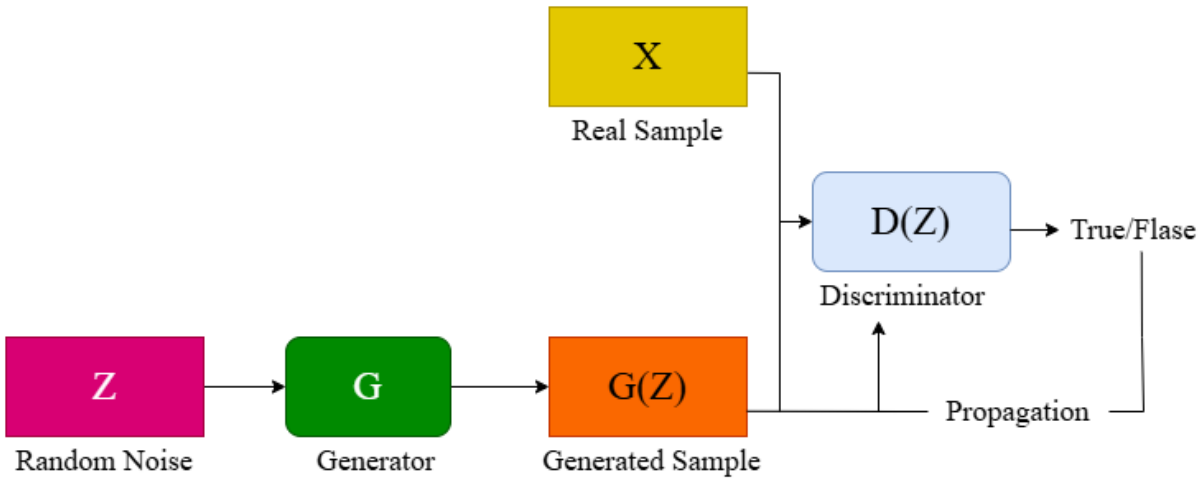

Figure 3-1 GAN network structure

Figure 3-1 displays the network structure of GAN presented by Goodfellow et al. in 2014. The proposed approach utilizes a GAN framework for enhancing images. A low-quality picture is inputted into the generator network, resulting in the creation of a superior image with improved details and fidelity. The generator network is given input from the discriminator network to improve the quality of the generated images. The setup consists of a discriminator (D) to distinguish between real and generated samples, and a generator (G) to produce deceptive

samples to fool the discriminator during training. The discriminator detects fake images in a game-like process, while the generator generates fake images resembling the real ones. This is the process of creating an adversarial network where SRGAN is an SR network that can create adversarial networks. Figure 3-2 depicts the network structure of SRGAN.

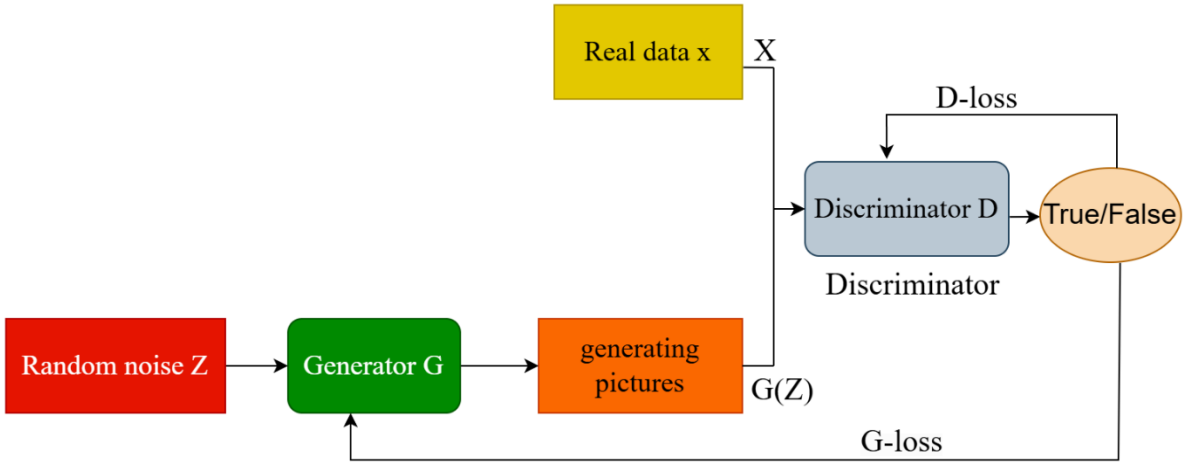

Figure 3-2 SRGAN network structure

SRGAN increases the amplification factor of SR. SR has been shown to be more effective at magnifications of 4 or 8. The recovery of vital information is limited as traditional SR networks or linear networks generate higher quality images. SRGAN is like the clash between two simulation models. The discriminator receives a real image x from the distribution $p_{data}(x)$ and calculates the likelihood that x represents an actual photograph, denoted as D(x). This involves distinguishing genuine pictures from counterfeit ones. The generator uses a random noise vector z, taken from a prior distribution $p_z(z)$, to create a synthetic image G(z). This is the procedure of creating an artificial image using the noise vector z. If X follows the true sample distribution $X \sim p_{data}(x)$ and Z follows the generator sample distribution $Z \sim p_z(z)$, then the loss function for optimizing the GAN is as follows:

$$\min\max V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \qquad (3-1)$$

where $V(D,G)$ denotes the loss function and E refers to the expected value of the distribution function. Min refers to the minimization objective of the Generator, which is the second term in the formula. Max represents the maximization objective of the Discriminator, which is the first term in the formula.

3.1.2 Content Loss

The content loss in SRGAN is computed on feature maps from a pre-trained deep neural network, typically the VGG network. Unlike MSE, which evaluates differences in pixel

intensity and is thus more sensitive to exact pixel values, content loss is derived from high-level representations of the image. These representations are extracted from intermediate layers of the VGG network, which are more robust to spatial variations. As a result, content loss can better capture the perceptual quality of an image, focusing on the high-level features that are more relevant to human visual perception. In SRGAN, the total loss function is a weighted sum of two components: content loss, which guides the generator, and adversarial loss, which is used by the discriminator. The adversarial loss is crucial for distinguishing real high-resolution images from those generated by the model, ultimately pushing the generator to create more realistic images that deceive the discriminator.

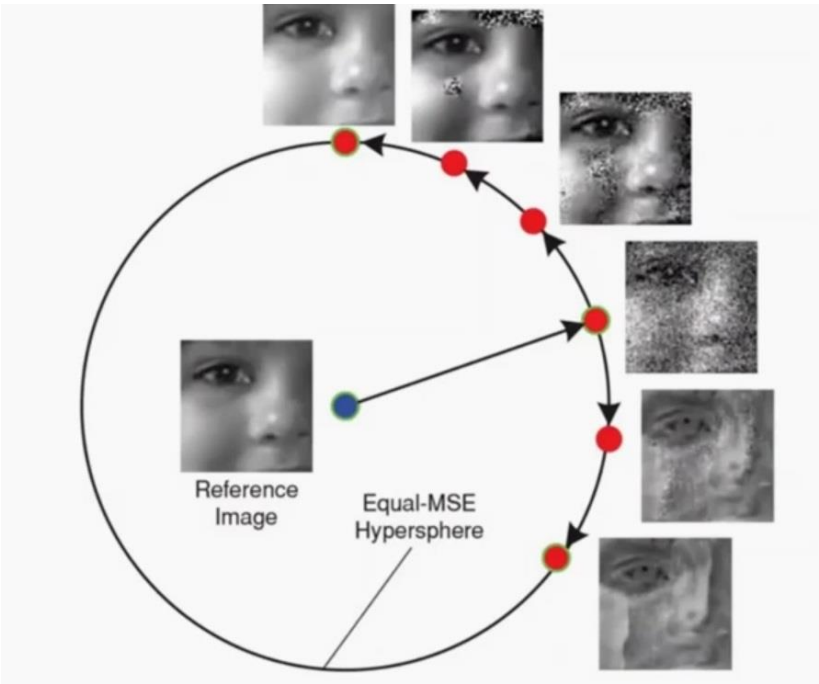

Figure 3-3 All the photos on the circumference of the circle resembles each other when MSE is selected as the loss function since MSE compare photos pixel by pixel and do not consider structural similarity. Figure taken from [126]

Figure 3-3 is the diagram illustrates the limitations of MSE in generating high-quality images. If MSE is used as the primary loss function, the generated images may not resemble the structure of the reference image, even though their pixel-wise intensity differences are minimized. As shown in the diagram, all of the images on the circumference of the equal-MSE hypersphere are treated as equally similar to the reference image (located in the center) according to MSE. This is because MSE only evaluates pixel-by-pixel differences, ignoring structural information or perceptual quality. The figure vividly demonstrates that despite having similar MSE values, the images vary significantly in their visual quality. Some of the images on the hypersphere display distortions and unrealistic textures that are not present in the

reference image. These artifacts highlight the inadequacies of MSE as a measure of image similarity in tasks such as super-resolution [126]. In contrast, content loss focuses on perceptual similarity rather than pixel-level similarity, ensuring that the generated images retain meaningful structural details akin to those in the reference image. Therefore, while MSE may report similar error values for all the images on the hypersphere, content loss, as utilized in SRGAN, is more sensitive to the perceptual and structural fidelity of the images. This makes SRGAN more effective in generating high-resolution images that align with human visual perception.

A key component of the SRGAN is the use of content loss, also referred to as VGG loss, which is computed from the generator's output. Content loss evaluates image quality based on perceptual similarity. This measure relies on high-level feature comparisons between the generated image and the ground truth, ensuring that the generated image maintains important structural and textural information akin to human visual perception. To calculate content loss, both the generated image and the ground truth image are passed through a pre-trained deep neural network, typically VGG or ResNet, resulting in feature maps. These feature maps are essentially high-level representations of the images, capturing features such as edges, textures, and object structures. Content loss is then defined as the Euclidean distance between the feature maps of the generated and ground truth images, which quantifies how different these higher-order features are from one another.

While a variety of pre-trained networks can be used to extract these feature maps, VGG19 is commonly employed, pre-trained on the ImageNet dataset. However, research suggests that deeper network layers provide improved results, as they focus more on fine-grained details of the image. For instance, using features from deeper layers such as VGG54 (which refers to the feature maps from the 54th layer of the VGG network) typically yields better perceptual quality than using earlier layers like VGG19 (features from the 19th layer). The deeper layers have the advantage of capturing more intricate and abstract representations of the image, which contributes to enhanced image sharpness and fidelity. When using VGG19 as a pre-trained model, content loss is usually computed from a set of intermediate layers, typically excluding the last max pooling layer. This is because the deeper layers in VGG capture high-level semantic information without overly reducing the spatial resolution. A common configuration involves using five layers of the VGG19 network, where each layer comprises a Conv2D layer followed by max pooling to reduce dimensionality while preserving critical spatial information. The first Conv2D layer is usually followed by max pooling, then additional Conv2D layers and pooling, with the final layer being another Conv2D. The architecture of these layers is crucial, as each

Conv2D layer extracts increasingly complex image features, and pooling layers serve to downsample the feature maps while retaining essential image details. This process ensures that the content loss remains sensitive to structural details, making the generated super-resolution images more aligned with human perceptual quality.

It is important to note that while VGG19 is commonly used for this purpose, further improvements have been achieved by leveraging deeper networks or more refined layers. For instance, recent advancements suggest that utilizing deeper layers of VGG, such as those from VGG54, enhances the model's attention to finer details, resulting in better super-resolution outputs. By focusing on these deeper, more abstract feature representations, the network can better reconstruct high-resolution images with sharper edges and textures.

3.1.3 Paired Data in SRGAN Training

Another important characteristic of the SRGAN is its reliance on a paired dataset consisting of HR and LR images. The SRGAN architecture requires a set of corresponding HR and LR image pairs for training. To construct such a dataset, one typically starts with HR images and synthetically generates their LR counterparts. This degradation process can be achieved through various techniques, including down-sampling, adding noise, or applying blurring filters. These artificial degradations mimic real-world factors that cause image quality loss, allowing the model to learn how to generate high-quality, super-resolved images from lower-quality inputs. This paired structure is crucial, as the network uses these HR-LR pairs to minimize the difference between the generated output and the ground-truth high-resolution image during the training process.

3.2 Conventional SRGAN

Since its introduction in 2017, the SRGAN [50] model has been widely adopted in numerous research studies, leveraging its ability to enhance image resolution for various applications [90], [127], [128], [129], [130], [131], [132], [133], [134]. The core architecture of SRGAN, consisting of a generator and a discriminator, has proven effective for generating high-quality super-resolution images. In this section, we will analyze the fundamental design principles of SRGAN's generator and discriminator, providing an in-depth understanding of its baseline architecture. Following this, we will explore the modifications and improvements introduced in the development of our new models, namely A-SRGAN and U-SRGAN, which are designed to enhance image quality while accelerating the super-resolution process. These innovations aim to address some of the limitations of the original SRGAN, optimizing both performance and computational efficiency for better practical application in various fields.

3.2.1 SRGAN Generator Architecture

GANs is a framework for deep learning in which there are two primary parts in play: a generator and a discriminator. Both of these neural networks are trained together in an adversarial way. The generator network learns to create artificial data that is similar to the input data distribution, while the discriminator network serves as a binary classifier to differentiate between real and generated data. The layout of the generator is shown in Figure 3-4, comprising three main components: Convolutional layers, Residual blocks, and up-sampling blocks. The network processes the LR image initially and passes it through a convolutional layer to generate a feature map. Afterward, it goes through a PReLU (parametric rectifier linear unit) activation function [135].

The parametric ReLU permits a negative slope for negative values instead of treating them as zero. In this manner, the network can gain knowledge from both positive and negative values. This allows for improved feature extraction. After that, the image goes through a series of 16 residual blocks containing two convolutional layers, batch normalization, PReLU, and skip connections. Residual blocks facilitate the network in capturing image details, while skip connections maintain the feature flow between blocks without the weights being affected by vanishing gradients.

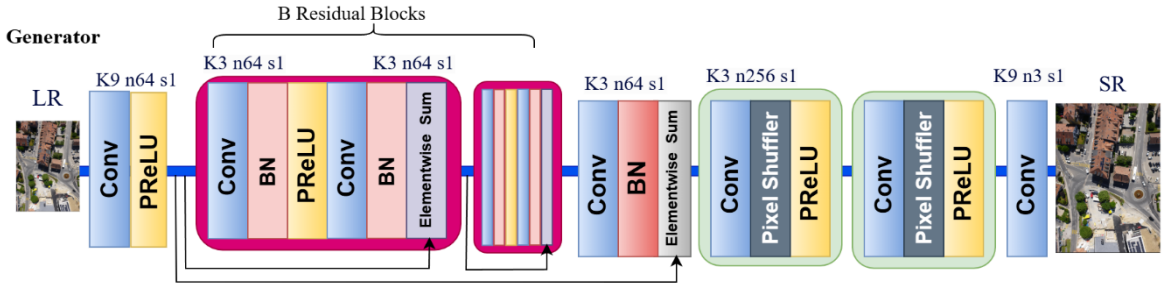

Figure 3-4 The original architecture of generator in the SRGAN

Following the residual blocks, the input is processed through a convolutional layer, batch normalization, and an elementwise sum block before going through two up-sampling sections, resulting in an output image with four times higher resolution than the input image. This gradual increase in resolution plays a crucial role in improving the overall quality of the image, leading to the creation of high-quality images while significantly reducing the necessary parameters. Every upscaling unit includes a convolutional layer, Pixelshuffler, and PReLU. The PixelShuffler [136] is responsible for doubling the size of the LR in each block before then quadrupling the LR image size before passing it to the final convolutional layer. The resulting image should resemble the true representation of the high-resolution image (ground truth).

3.2.2 SRGAN Discriminator Architecture

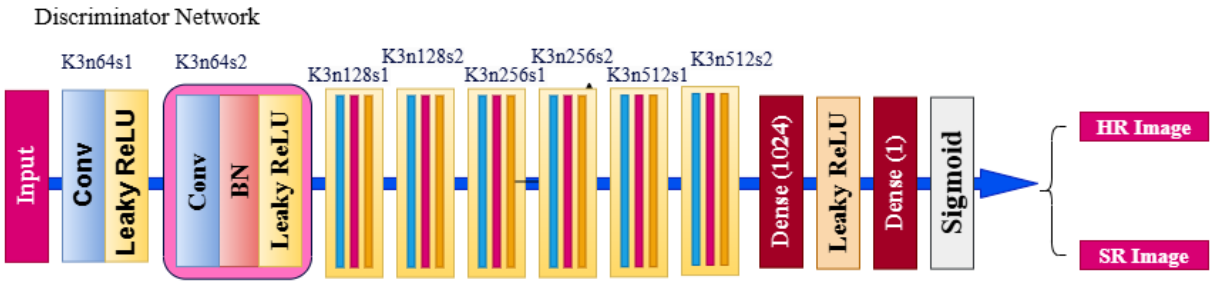

Figure 3-5 the basic design of discriminator in SRGAN

SRGAN utilizes a conventional discriminator network, and the architecture of SRGAN's discriminator network is shown in Figure 3-5. The main components of this network are mainly convolutional layers, Leaky ReLU activation functions, and Batch Normalization (BN). The discriminator function is taught to differentiate super-resolution images from genuine images. Input includes either a high-quality image generated by the network or an image sourced from the training dataset. Similar to the generator, it includes multiple convolution layers for feature extraction. The initial stage consists of a convolution layer, succeeded by a Leaky ReLU activation. The subsequent stage includes eight sets of three layers, each containing a convolution layer, BN, and Leaky ReLU activation. The Leaky ReLU function is defined with a parameter of 0.2, and Batch Normalization layers are purposely included to speed up network training and improve overall generalization capabilities.

We need to apply a dense layer to convert the multi-dimensional feature maps of the input image into a single-dimensional vector for classification. This vector is then passed through an activation layer before going through another dense layer to transform the 1024-sized one-dimensional vector into one. We utilized the Sigmoid function to convert the input value into either 0 or 1, representing binary categories. The results from the discriminator range from 0 to 1, where values closer to 0 indicate fake images and values closer to 1 indicate real images.

3.3 First Proposed model

In this section, we introduce two variations of the SRGAN generator designed to enhance image super-resolution performance. The first model, referred to as A-SRGAN, replaces the conventional residual blocks in the SRGAN generator with a specialized autoencoder architecture. This autoencoder-based design aims to improve the feature extraction process by leveraging an encoding-decoding framework instead of the standard residual learning approach.

To further refine the performance of A-SRGAN, we developed an improved variant, Res-A-SRGAN. In this model, the conventional convolutional blocks within the autoencoder structure were replaced with residual blocks, allowing for more efficient gradient propagation and enhanced feature retention. By integrating residual learning into the autoencoder framework, Res-A-SRGAN aims to better capture fine details and improve the overall reconstruction quality of super-resolved images.

3.3.1 Autoencoders for Image Resolution Enhancement

Autoencoders are increasingly used to enhance image resolution by leveraging their ability to learn efficient data representations. As well as this, Autoencoders are a type of neural network architecture used primarily for unsupervised learning, where the goal is to learn a representation (encoding) for a set of data. One common use case is image resolution enhancement where an autoencoder can be used to learn how to upsample low-resolution images into higher-resolution ones. Convolutional autoencoders are used to improve image resolution by incorporating skip connections, which help in retaining important features during the encoding and decoding process [137]. Figure 3-6 displays a normal autoencoder using skip connections. This approach has shown high accuracy in enhancing images from diverse datasets.

Skip connections allow direct pathways for information to bypass certain layers, reducing the risk of vanishing gradients and helping maintain important features throughout the network. Additionally, by providing shortcuts for data to flow directly from the encoder to the decoder, skip connections help maintain the sharpness and detail of the reconstructed images, which is crucial for tasks like anomaly detection [138]. They help stabilize the training process by propagating a linear component through the network, which can alleviate optimization difficulties associated with deep networks [139].

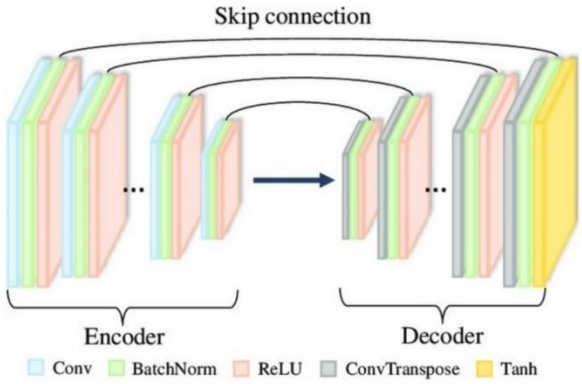

Figure 3-6 Autoencoder with skip connections. Figure taken from [139]

A vital role in transforming input data (x_i) into a latent feature representation (h_i) is held by the encoder, which is denoted by the function g. In mathematical terms, this process is formally represented as:

$$h_i = g(x_i) \qquad (3-2)$$

In this context, g is a function that takes the input from space $\mathbb{R}^n$ and transforms it into the latent space $\mathbb{R}^q$. The hidden feature representation h_i contains important information taken from the input data. On the other hand, the function f decodes the latent features (h_i) to reconstruct the output (x_i). In terms of mathematics, this reconstruction is described as:

$$x_i = f(h_i) \qquad (3-3)$$

Like the encoder, f works on the latent space $\mathbb{R}^q$ and converts it back to the initial input space $\mathbb{R}^n$. The main task of the decoder is to rebuild the input information using the meaningful hidden characteristics (Figure 3-7). Throughout training, the goal is to identify the best possible functions g and f that reduce the discrepancy between the input and the reconstructed output. This is formalized through the objective function:

$$\underset{f,g}{\operatorname{argmin}} E\left[\phi\left(x_i, f\left(g(x_i)\right)\right)\right] \qquad (3-4)$$

In this context, ϕ quantifies the difference between the input and output, while E represents the mean across all data points. The objective is to successfully reconstruct while generating a significant latent representation.

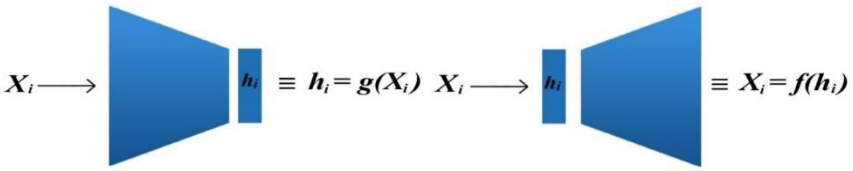

Figure 3-7 The architecture of Encoder and Decoder in Autoencoder

3.3.2 Autoencoder-Based SRGAN Generator with Convolutional Skip Connections

In our updated SRGAN generator, we have integrated an autoencoder design with convolutional skip connections, taking influence from LinkNet [140]. Usually, SRGAN generators employ residual blocks to aid in the network's learning of effective mappings for

super-resolution tasks. Nevertheless, in our design, we substituted the residual blocks with an autoencoder framework comprising an encoder and decoder. The encoder gathers key characteristics from the lower-quality input, and the decoder transforms these characteristics into a higher-quality output. This method enables the network to concentrate on understanding the fundamental characteristics for improving images. An important aspect of our design is incorporating convolutional layers in skip connections instead of standard direct skip connections. Skip connections in traditional networks such as LinkNet allow for the direct transfer of information between layers, preserving spatial information. By adding convolutional layers to these skip connections, the network can learn flexible changes to the features being transferred. This allows the model to carefully enhance the data and potentially maintain crucial image details more effectively, which is particularly vital in tasks requiring super-resolution. Figure 3-8 shows the structure of the autoencoder we used in our generator.

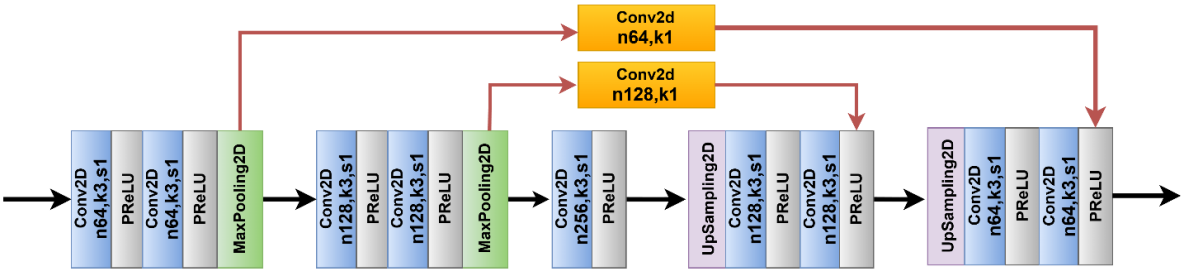

Figure 3-8 The autoencoder that was integrated in the design of SRGAN generator

The convolutional skip connections add a layer of sophistication to the network by learning more complex feature representations. This results in better feature propagation and more efficient use of the information at each layer. The convolutional layers ensure that the features passed through the skip connections are not merely copied, but rather transformed in a way that aligns with the high-resolution reconstruction goal. This refinement is likely to improve the overall performance of the network, leading to sharper and more accurate super-resolved images. By combining the autoencoder structure with convolutional skip connections, our design effectively improves the network's ability to extract, refine, and reconstruct high-resolution features. This approach, inspired by LinkNet's efficient use of skip connections, enhances the SRGAN's performance by ensuring a richer flow of information throughout the network, leading to better results in super-resolution tasks. Figure 3.9 shows the modified generator enhanced with an autoencoder.

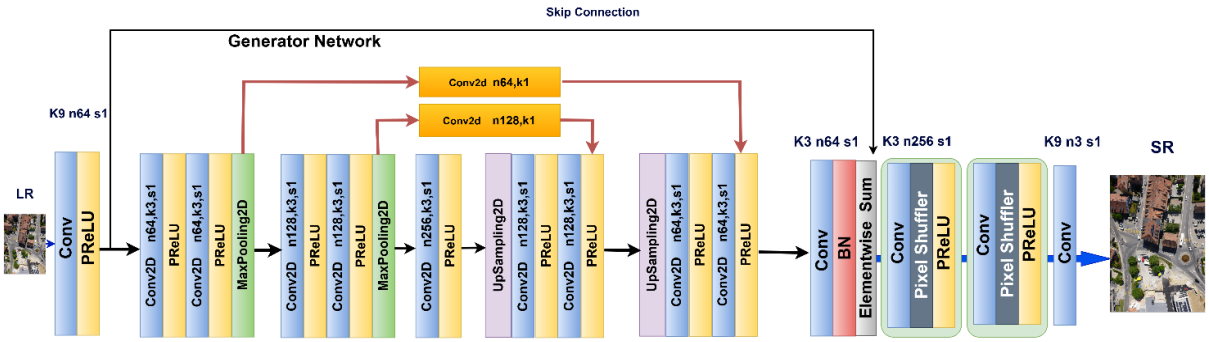

Figure 3-9 The updated SRGAN generator using autoencoder (A-SRGAN)

3.3.3 Autoencoder-Based SRGAN with Residual Blocks

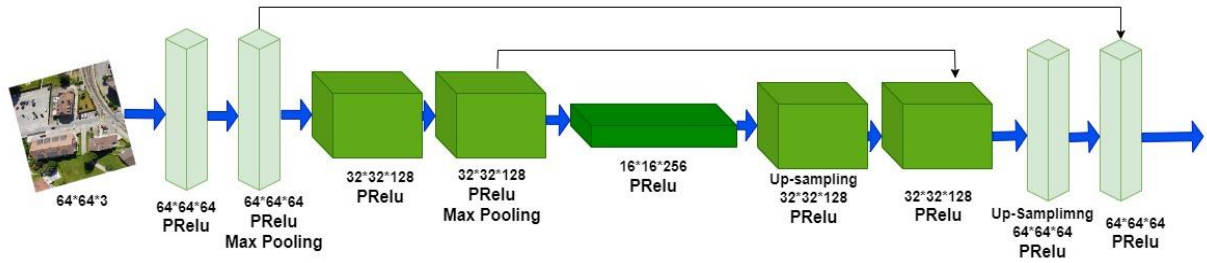

Figure 3-10 The simplified block diagram of Autoencoder that we used in the SRGAN structure, the only difference is that we used residual blocks instead of Convolutional blocks. For the sake of simplification skip connections were shown by arrows, while we used Conv layers in the skip connections.

In another model, Figure 3-10, an Autoencoder with residual blocks was used in SRGAN instead of the usual generator, which had 16 residual blocks. Integrating autoencoders with skip connections and residual blocks into the generator structure of SRGAN provides various advantages, enhancing the quality of super-resolution images. An autoencoder compresses the input data into a concise form and then decodes it to reconstruct the original image. In the case of SRGAN, the lower-quality input is encoded into a latent space and then transformed into a higher-quality output through decoding. Skip connections directly transfer information from the initial stages (low-level features) of the encoder to the corresponding layers in the decoder. This assists the network in retaining important information because basic elements in images, such as edges and textures, are simpler to grasp at lower resolutions and are vital for recreating finer details in the higher resolution image. Skip connections prevent the loss of these features.

Additionally, the presence of skip connections helps alleviate the issue of vanishing gradients by allowing gradients to flow directly, making it simpler to train deeper neural networks. Preserving the gradient signals leads to smoother backpropagation, resulting in quicker convergence and more stable training. This also enhances reconstruction accuracy by

using skip connections to avoid degradation from bottleneck compression, allowing for more precise reconstructions. They transmit information straight from the encoder to the decoder, enabling a more accurate reproduction of intricate details. Figure 3-11 depicts the Autoencoder SRGAN (A-SRGAN) which residual blocks were replaced by an autoencoder.

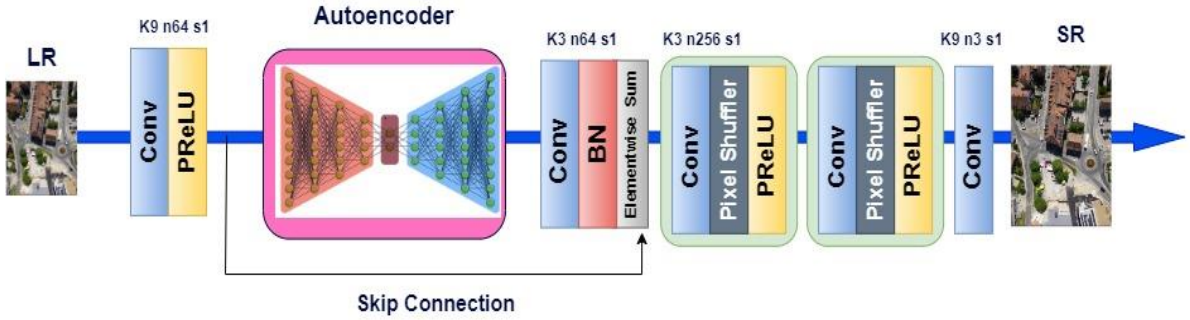

Figure 3-11 The architecture of generator in A-SRGAN

A residual block usually consists of multiple convolutional layers, and the input is combined with the output following the convolution process. The mathematical expression for this is:

$$\text{Output} = \text{F}(\text{Input}, \text{W}) + \text{Input} \qquad (3-5)$$

In the convolution operations applied to the input, denoted as F (Input, W), the input is included in the resulting output. This presents an identity mapping that contributes to the stabilization of the learning process. If residual blocks were not used, the network would have to learn the simple identity mapping. Residual blocks help ease this load, enabling the network to concentrate on understanding the discrepancies or residuals between the low-resolution and high-resolution images. Additionally, SRGAN is designed to produce more lifelike high-resolution images. Residual blocks aid in capturing important high-frequency details such as sharp edges and textures which are essential for achieving visually appealing super-resolution results. Training deep convolutional networks frequently results in gradients becoming very small, causing challenges with learning. Residual connections make it easier for gradients to propagate through the network, making it simpler to train deeper networks (which are essential for tasks such as super-resolution).

In SRGAN, the generator works to produce high-resolution images that are visually improved and more detailed. These techniques aid in two main areas: firstly, the use of residual blocks enhances the sharpness and visual fidelity of the generated image to closely resemble

the high-resolution ground truth. This is in line with the perceptual loss employed in SRGAN, where it evaluates the high-level characteristics of the produced image against the actual high-resolution image instead of solely focusing on pixel values. Furthermore, the training of GANs can be challenging as a result of the volatility in the adversarial procedure. By including skip connections and residual blocks in autoencoders, the generator becomes simpler to train, converges more quickly, and generates high-quality outputs. Skip connections prevent loss of important features and residual blocks maintain smooth gradient flow in training.

3.4 Second proposed model

3.4.1 Introduction

In this section, we explore additional modifications to the SRGAN generator to further improve its performance. Initially, we replaced the standard 16 residual blocks in the SRGAN generator with a U-Net architecture, leveraging its ability to preserve spatial information through skip connections. This modified model is referred to as U-SRGAN. To enhance the U-Net structure, we integrated residual blocks and attention gates, resulting in a more powerful feature extraction process. This enhanced model, incorporating both residual learning and attention mechanisms, is termed ARUnet-SRGAN.

Finally, we applied transfer learning techniques to further optimize performance. Specifically, we first trained an autoencoder separately and subsequently transferred its learned weights to initialize the ARUnet-SRGAN model. This pretraining step facilitated a more effective learning process, improving convergence and overall super-resolution quality. This final model is referred to as Pretrained ARUnet-SRGAN. Figure 3-12 depicts these changes in the structure of SRGAN.

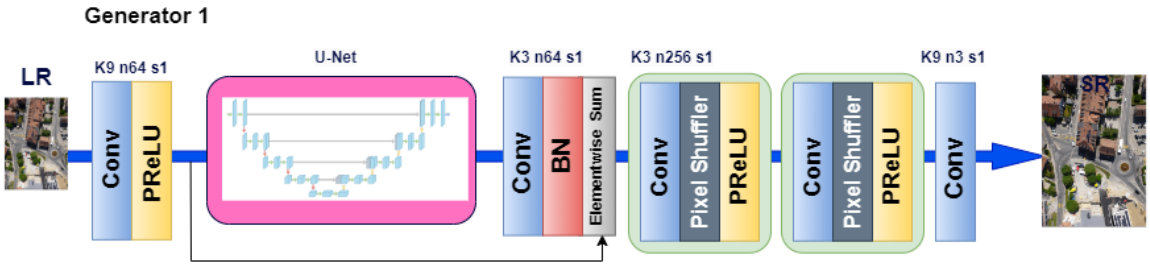

Figure 3-12 Using U-Net in the design of Basic SRGAN

The U-Net architecture, originally designed for biomedical image segmentation [141], [142], has become a widely used model in various computer vision tasks due to its powerful ability to handle segmentation problems and efficiently capture both local and global features. U-Net was specifically designed for cases where training data is limited, which is often the case

in medical and scientific domains. It makes efficient use of available data through data augmentation and careful design of the network structure which is why we opted for U-Net since our data was limited. The use of skip connections allows U-Net to combine low-level features (e.g., edges, textures) from early layers with high-level abstract features (e.g., shapes, patterns) from deeper layers. This ensures that fine details are preserved in the final segmentation, even in complex images. U-Net's architecture is symmetric, with the encoder extracting features from the input image and the decoder reconstructing the segmentation map. This symmetry makes it particularly effective at learning both local and global contexts simultaneously, which is crucial for pixel-wise segmentation tasks. In our study, we removed the segmentation head in the U-Net and we only took benefit from the feature extraction properties of this network.

3.4.2 U-Net

A method widely used for semantic medical image segmentation, known as "U-Net", was among the earliest and highly popular approaches. Figure 3-13 displays a sketch of the simple U-Net model. The network is comprised of two primary components: the convolutional encoding and decoding units as per the layout. In each section of the network, the fundamental convolution operations are conducted, after which ReLU activation is applied. During encoding, down-sampling is achieved through 2×2 max-pooling. The decoding phase involves up-sampling the feature maps using convolution transpose operations, known as up-convolution or de-convolution. The original U-Net version involved cropping and transferring feature maps from the encoding to the decoding unit.

The U-Net model offers multiple benefits: firstly, it enables simultaneously utilizing global location and context. Additionally, it requires minimal training data and offers superior results for segmentation projects. Thirdly, a full image is processed in one go in the pipeline, resulting in segmentation maps without any intermediate steps. This makes sure that U-Net maintains the complete context of the input images. In our study, since we do not need segmentation, we omitted the sigmoid from the output. In a typical U-Net architecture, two consecutive Conv2D layers are used, each followed by Batch Normalization and ReLU activation. This configuration helps in the effective extraction of features at each resolution level of the U-Net, while batch normalization enhances the training stability and convergence speed. The combination of convolutions and ReLU ensures the model can capture both low-level and high-level features. However, as we go deeper into the network we may encounter vanishing gradients. Therefore,

in lieu of a typical block, it is recommended to use a residual block.

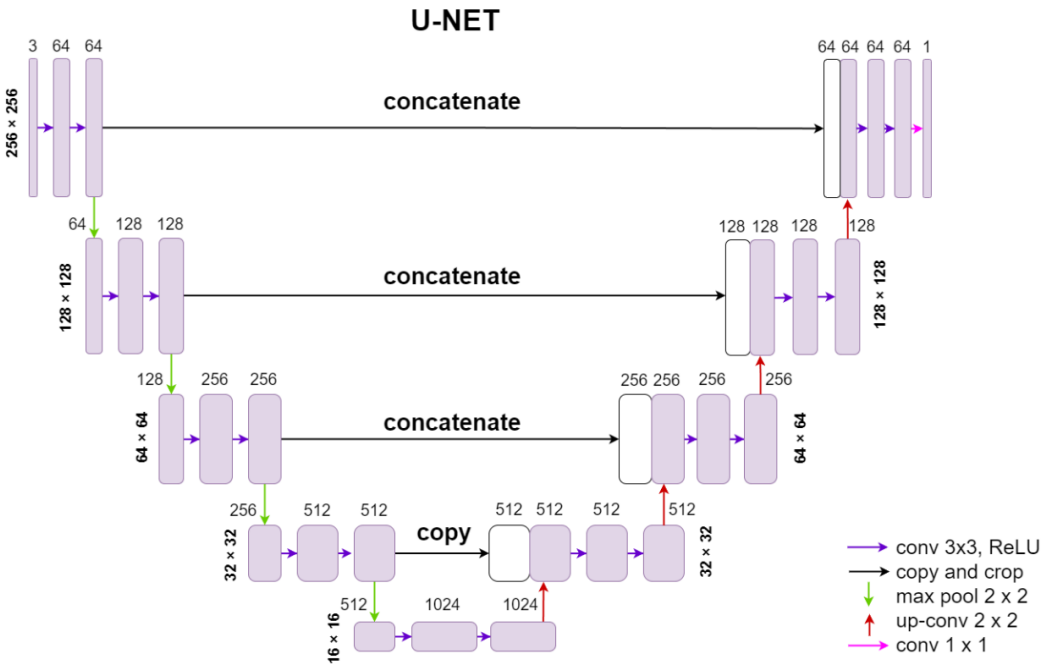

Figure 3-13 A typical design of U-Net for a segmentation task

3.4.3 Residual blocks in U-Net

Using detailed information at a low level while maintaining high-level semantic content is crucial for achieving better results. Nevertheless, training a deep neural network becomes challenging when there are only a limited number of training samples. One potential solution is to use a pretrained network and then fine-tune it on the specific dataset. Another option is using a lot of data augmentation, similar to what was done in U-Net. We think that aside from data augmentation, the design of U-Net also plays a role in easing the training issue. The reasoning for this is that transferring basic characteristics to higher levels helps information flow more smoothly between different levels, making it easier for signals to move between them. This simplifies backward propagation in training and also enhances high-level semantic features by incorporating detailed low-level information. This is somewhat reminiscent of the concept found in residual neural networks.

In this correspondence, we demonstrate that enhancing the performance of U-Net is possible by replacing the basic unit with a residual unit. In the basic unit, increasing depth can enhance the performance of a deep neural network, but it might also hinder the training process and lead to a degradation issue. In order to tackle these challenges, the residual neural network was introduced to aid in training and combating the degradation problem. The residual neural

network is made up of multiple residual units that are stacked together. Every residual unit can be represented in a typical way:

$$\mathbf{y}_l = h(\mathbf{x}_l) + \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l) \quad (3-6)$$

$$\mathbf{x}_{l+1} = f(\mathbf{y}_l) \quad (3-7)$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ are the input and output of the l-th residual unit, $\mathcal{F}(\mathbf{x}_l, \mathcal{W}_l)$ is the residual function, $f(\mathbf{y}_l)$ is the activation function and $h(\mathbf{x}_l)$ is the identity mapping function, a typical one is $h(\mathbf{x}_l) = \mathbf{x}_l$. Figure 3-14 shows the difference between a plain and residual unit. There are multiple combinations of batch normalization (BN), ReLU activation, and convolutional layers in a residual unit. He et al. [143] presented a detailed discussion on the impacts of different combinations and suggested a full pre-activation design, as shown in Figure 3-14 (b). In this work, we also employ a full pre-activation residual unit to build our deep residual U-Net. Figure 3-15 shows the modification that we made for our residual block in our model.

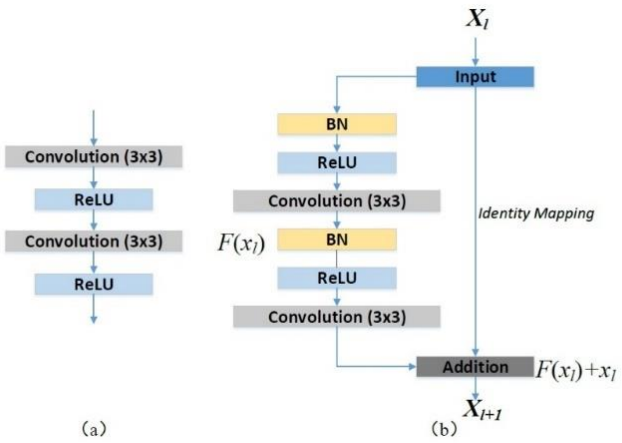

Figure 3-14 Building blocks of neural networks. (a) Plain neural unit used in U-Net and (b) residual unit with identity mapping used in the proposed Res-U-Net.

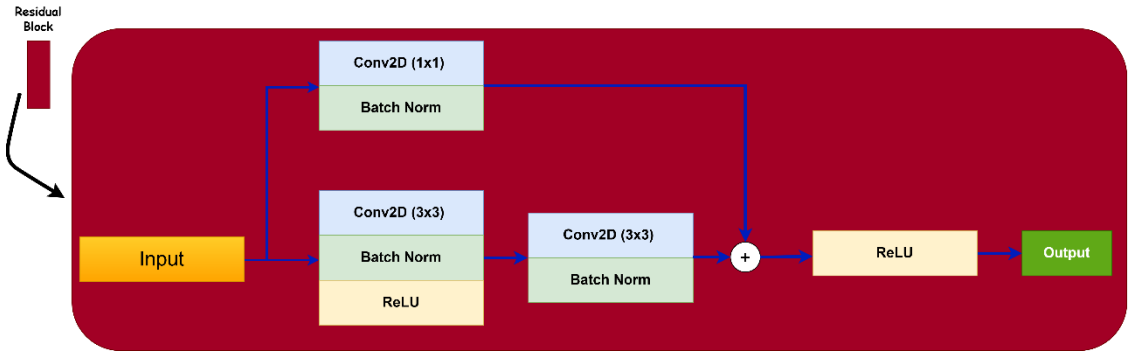

Figure 3-15 The actual residual block that we used in our models

3.4.4 Attention Gate

Incorporating attention gates into a U-Net model is a method created to concentrate on the

crucial areas of the input image while suppressing less significant sections. This allows the model to concentrate better on various shapes and sizes of target structures; with attention gates, the network can train to ignore unimportant background details and emphasize important features for the current task. Skip connections in U-Net aid in regaining spatial information that is lost during down-sampling by linking the encoder and decoder. Nevertheless, the importance of the features passed from the encoder varies. Incorporating attention gates enables the network to enhance the feature maps transmitted through the skip connections, concentrating on specific areas of interest and disregarding unimportant sections.

The attention gate (AG) is utilized in the skip connection connecting the encoder and the decoder, merging features from lower-resolution encoder layers with higher-resolution decoder layers. A gating signal from the decoder is used to eliminate less important encoder features. The attention gate output is used to amplify certain sections of the encoder's feature map, bringing attention to specific areas [125], [126]. We incorporated attention mechanisms to enhance the weighting of significant features within the U-Net architecture on each skip connection, as shown in Figure 3-16. An attention gate receives two inputs: one from a deeper layer (denoted as g) and another from the corresponding skip connection layer in the encoder part (denoted as x). The deeper layer g provides enriched feature representation, while the skip connection retains superior spatial information. Initially, a convolutional layer with a stride of 2×2 is applied to x, ensuring that both g and x share the same dimensions, thereby allowing for their summation.

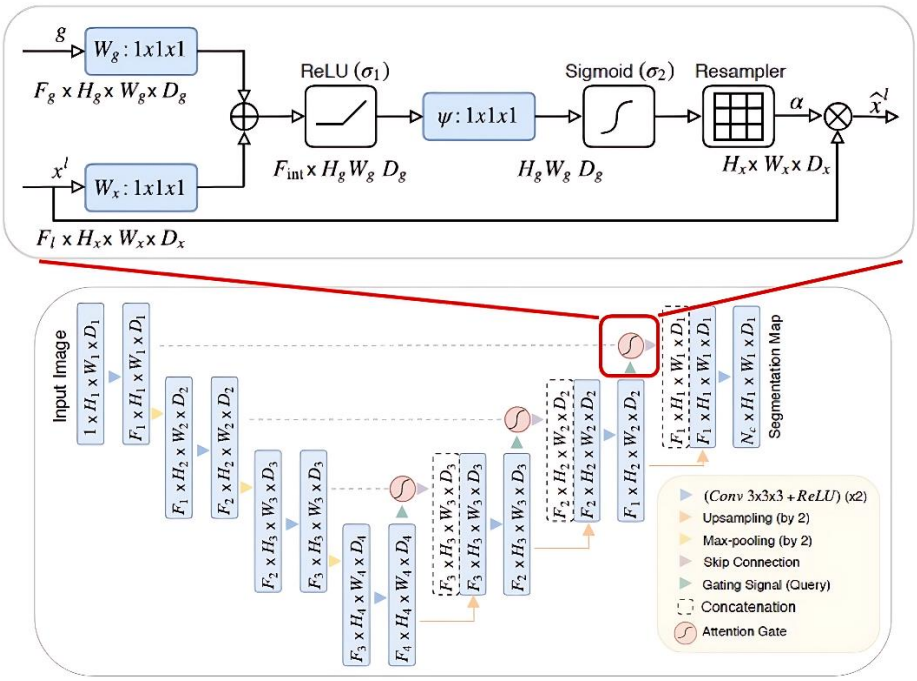

Figure 3-16 Schematic of the proposed additive attention gate (AG). Figure taken from [146]

The combined output is then processed through a ReLU activation function. Subsequently, a 1×1×1 Conv layer is applied to preserve the weights. By employing a sigmoid activation function, the weights are normalized to a range between 0 and 1. Following normalization, the output of the sigmoid function is up-sampled to match the channel number of x. The resulting up-sampled output is then multiplied by the vector x, effectively augmenting the attention to the relevant weights [146], [147]. This process emphasizes the features of interest, thereby improving the segmentation performance by allowing the network to focus more precisely on pertinent regions of the input images.

The suggested AGs are added to the traditional U-Net structure to emphasize important features transmitted via the skip connections. Data obtained from a rough scale is utilized in gating to clarify unnecessary and noisy reactions in skip connections. This is done just prior to the concatenation process to combine only pertinent activations. Moreover, AGs also control the neuron activations in both the forward and backward passes. During the backward pass, gradients from background regions are given less importance. This enables model parameters in shallower layers to be primarily updated according to spatial regions that are important for a specific task. The formula for updating convolution parameters in layer $l-1$ can be expressed in the following way:

$$\frac{\partial(\hat{x}_i^l)}{\partial(\Phi^{l-1})} = \frac{\partial\left(\alpha_i^l f(x_i^{l-1};\Phi^{l-1})\right)}{\partial(\Phi^{l-1})} = \alpha_i^l \frac{\partial(f(x_i^{l-1};\Phi^{l-1}))}{\partial(\Phi^{l-1})} + \frac{\partial(\alpha_i^l)}{\partial(\Phi^{l-1})} x_i^l \qquad (3-8)$$

The initial gradient term on the right side is multiplied by α_i^l. For multi-dimensional AGs, α_i^l represents a vector for every grid scale. In every sub-AG, additional information is extracted and combined to determine the skip connection's output. In order to make Attention Gates simpler and more efficient, linear transformations are carried out with 1x1x1 convolutions without spatial support.

Additionally, input feature-maps are reduced in size to match the resolution of the gating signal, reminiscent of non-local blocks. The linear transformations separate the feature-maps and reduce their dimensionality for the gating process. Skip connections corresponding to low-level features are excluded from the gating function due to their low-dimensional representation of input data. We utilize deep-supervision to ensure that the feature-maps are semantically discriminative in every image scale. This ensures that attention units of various scales can impact the responses to a wide range of foreground content in images. So, we make sure that detailed predictions cannot be pieced together from just a few skip connections. Figure 3-17 shows the design of a new generator in which a U-net with residual blocks and attention gates

is embedded.

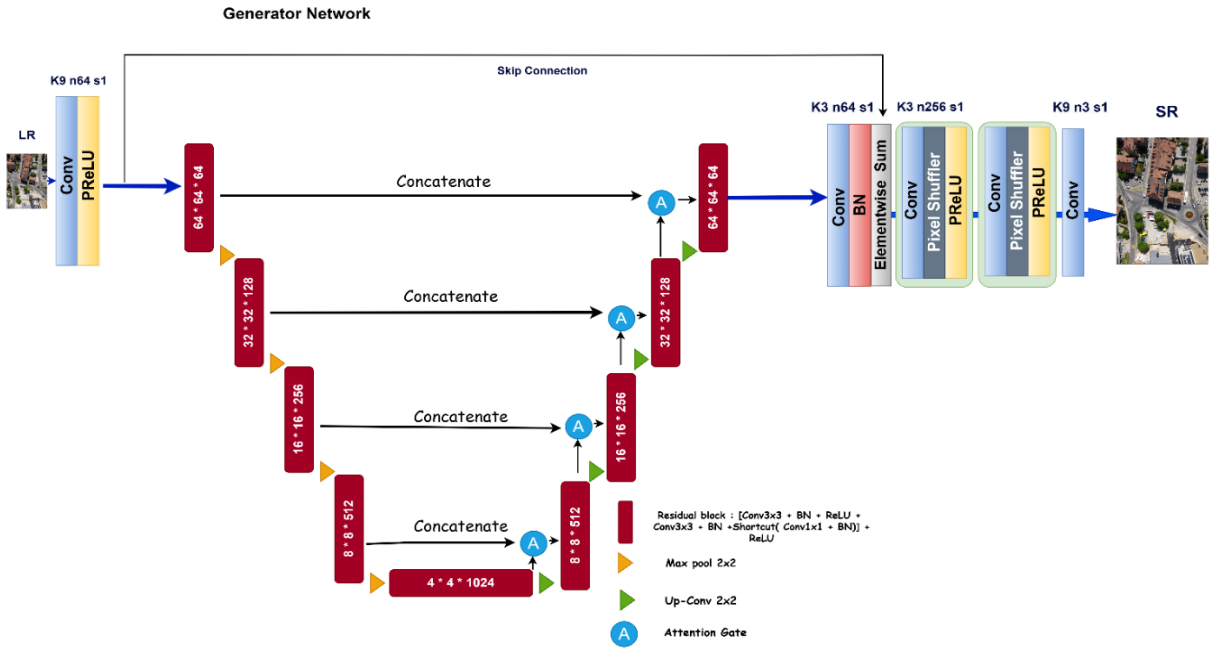

Figure 3-17 ARU-Net SRGAN

3.4.5 Pretrained U-Net with Autoencoder (Pretrained ARUnet-SRGAN)

In conventional U-Net, the skip connections aid in restoring lost spatial information from down-sampling. Nevertheless, these methods could result in significant and repetitive consumption of resources and analytical factors. In order to address this issue, the Attention U-Net design was introduced with attention gates integrated into the expansion path. These attention gates autonomously learn to concentrate on specific target structures without extra assistance. The proposed attention gates enhanced features by concentrating on essential application, design sensitivity, and accuracy for dense label predictions through the reduction of feature activations in unneeded regions.

Here we demonstrate how our architecture, which includes a pretrained Autoencoder with Attention U-Net and residual blocks, was implemented. We extracted an encoder block from the autoencoder model to utilize the pretrained weights instead of assigning random weights to the pixels. The encoder block bears a resemblance to the contraction pathway of U-Net. In the expansion pathway, attention gates have been incorporated to concentrate on important activations from skip connections and the deeper layers of the U-Net.

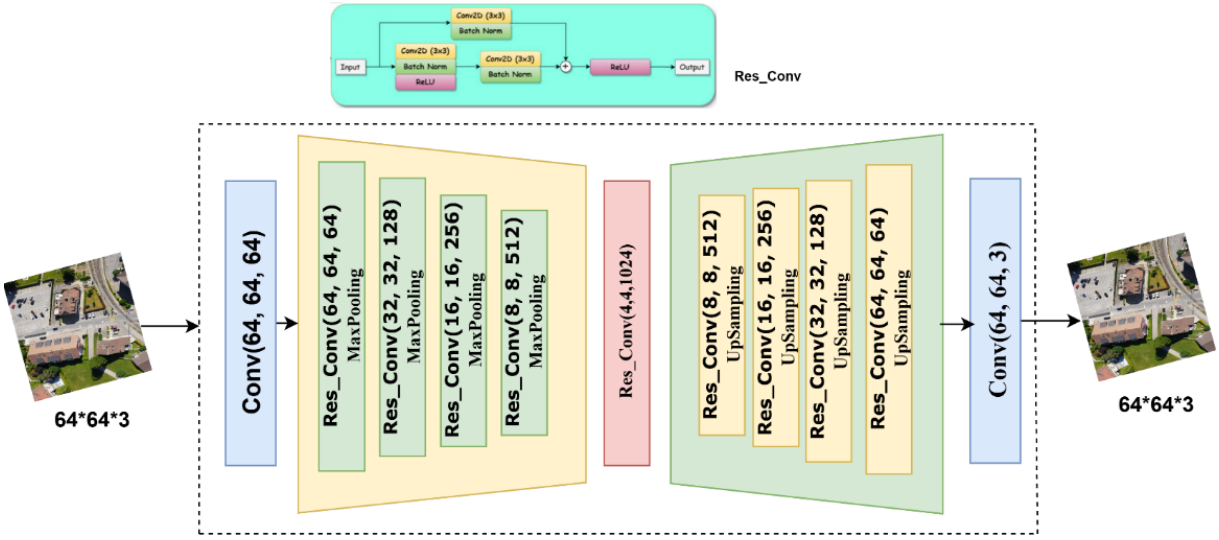

Figure 3-18 The symmetric autoencoder takes the LR images and reconstruct the images

The U-Net model proposed is created by combining two distinct models, the Symmetric Autoencoder (SAE), represented in Figure 3-18, and Attention Residual U-Net shown in Figure 3-19. Residual blocks from the generator of SRGAN were removed and a U-Net model was incorporated in the generator. Subsequently, we used a unique U-Net called the Attention Residual U-Net (ARU-Net) [148], [149] instead of a traditional U-Net. This design improves standard U-Net models by adding residual blocks and attention gates to enhance feature extraction and boost super-resolution results. The training method involves a two-step process to accelerate and enhance learning. At first, a symmetric autoencoder is used for pretraining [150], [151].

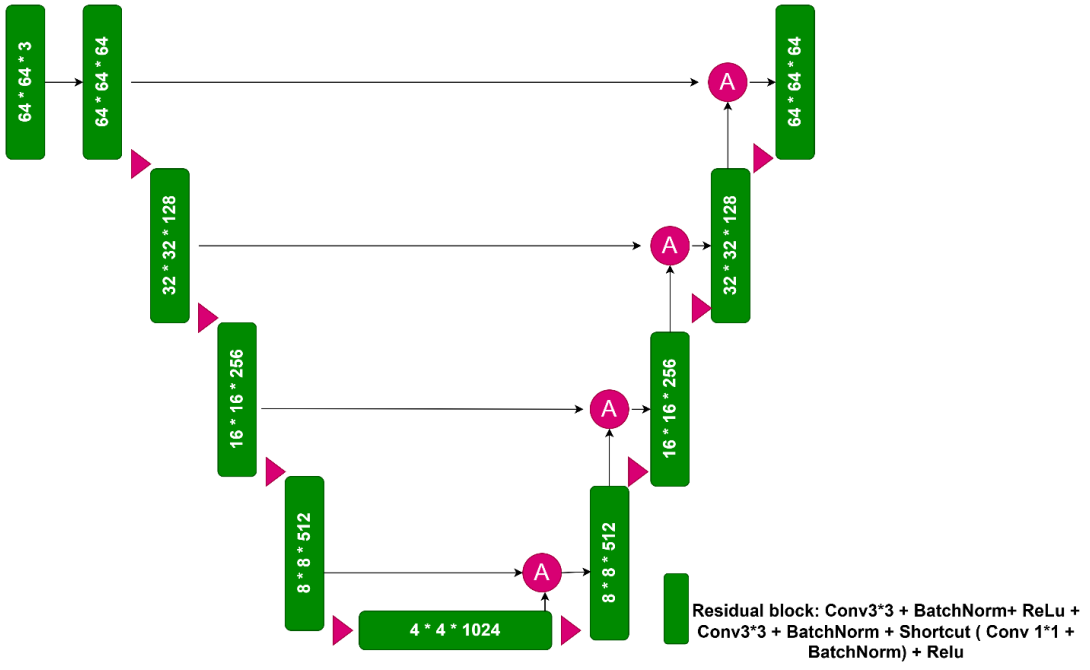

Figure 3-19 The attention residual U-Net. We designed the down-sampling path same as symmetric autoencoder

The autoencoder has a symmetric encoder-decoder structure, with both input and output being low-resolution images. This step assists the model in acquiring crucial image reconstruction characteristics at a lower resolution without the challenge of creating high-resolution outputs. After finishing pretraining, the encoder path's learned weights are stored. The pretrained weights are moved to the encoder section of the attention residual U-Net within the SRGAN generator, as shown in Figure 3-20, giving a solid starting point that accelerates convergence and secures training.

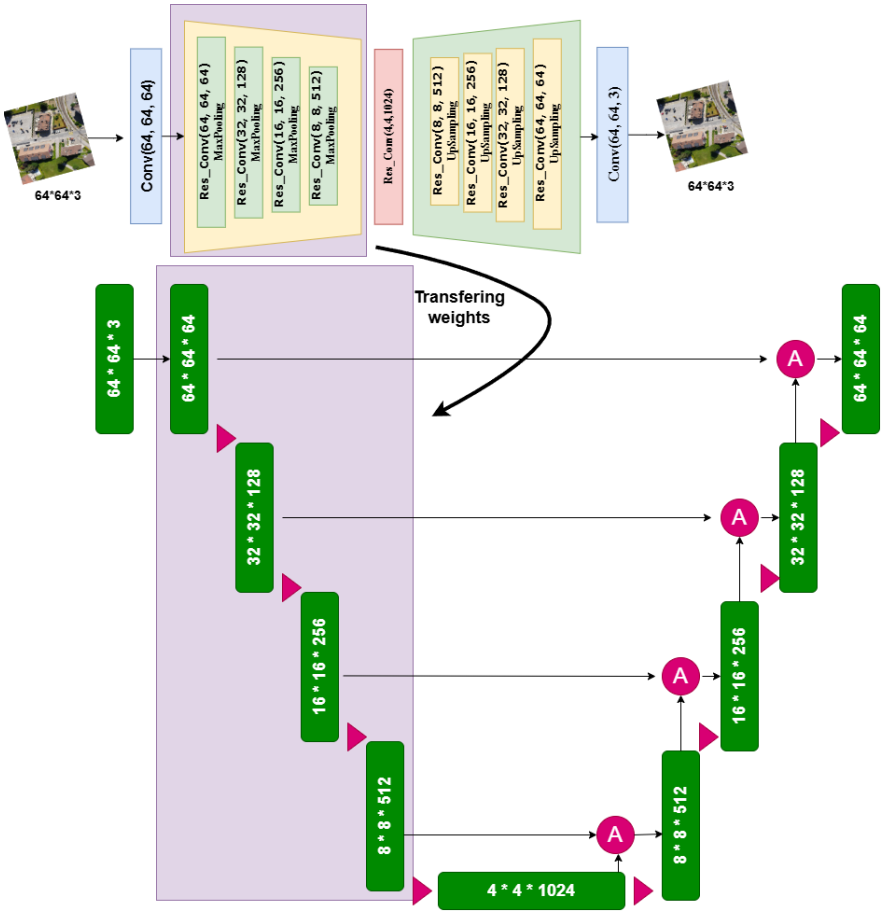

Figure 3-20 The weights from downsampling path of SAE are transferred to the encoder of attention residual U-Net

The attention residual U-Net incorporates residual blocks in its encoder-decoder design, which are made up of two sequential 3x3 convolution layers, along with Batch Normalization and ReLU activation functions. Incorporating residual connections involves combining the initial input with the output of the layers, which helps the network keep the flow of gradients and retain important low-level information necessary for tasks like image reconstruction. The

encoding layers decrease spatial dimensions, while the decoding layers increase them to rebuild the image resolution. Attention gates in the decoding phase help the model focus on important features learned during encoding. These gates employ gating mechanisms to enhance specific features that are useful for producing high-quality images, while also diminishing less significant ones. The attention mechanism enhances the residual blocks by making sure that only the crucial characteristics are transmitted through the network pathway. To sum up, this SRGAN generator design uses a pre-trained autoencoder to start a U-Net model with residual and attention components. The model is skilled at reconstructing high-resolution images by combining these elements, enhancing focus on crucial details with attention gates, and ensuring stable training with residual connections and autoencoder pretraining. This method boosts training speed and enhances output quality by efficiently leveraging hierarchical features obtained from input images.

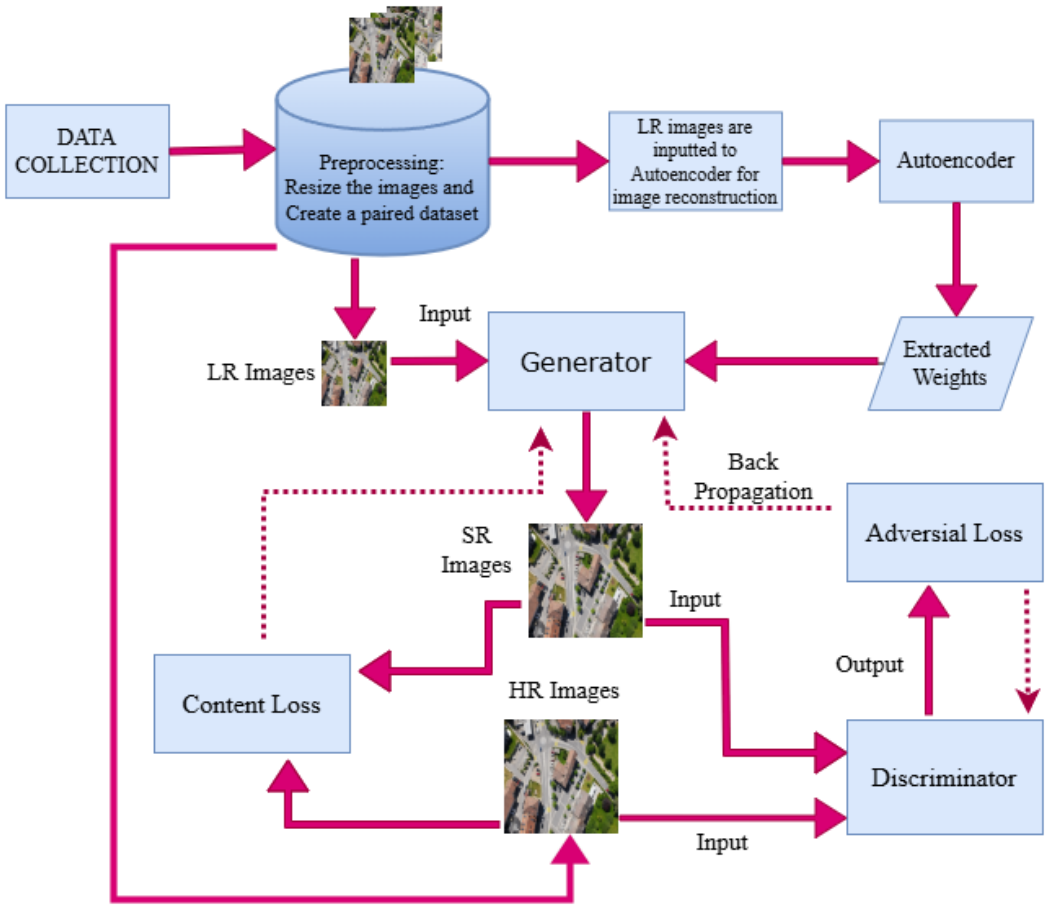

Figure 3-21 the flow chart of pretrained ARUnet-SRGAN

Figure 3-21 illustrates the flow diagram of the pretrained attention residual U-Net SRGAN. Additionally, we will describe the corresponding code implementation in pseudocode to clarify the coding process. To summarize the models introduced in this section, we will review the

characteristics of three models in the next chapter. First, we will examine a generator based on a simple U-Net, referred to as U-SRGAN. Next, we will explore the results of a U-Net enhanced with residual blocks and attention gates (ARUnet-SRGAN). Finally, the third model is a pretrained attention residual U-Net SRGAN (Pretrained ARUnet-SRGAN) that incorporates an autoencoder.

Pseudo code

- Initialization

  Initialize dataset: LR_images, HR_images

  Initialize weights for generator (G) and discriminator (D)

  Extracted_weights = Train_Autoencoder(LR_images)

  Set Generator weights = Extracted_weights

  Set hyperparameters: learning_rate, epochs, λ (for loss balancing), μ (momentum), ε
- Training Loop

  for epoch in range(num_epochs):

  for each LR_image, HR_image in dataset:
  - Step 1: Forward pass through Generator

    SR_image = Generator.forward(LR_image)
  - Step 2: Calculate Error (Content Loss and Adversarial Loss)

    Content Loss: Compare SR_image and HR_image

    Content_Loss = calculate_content_loss(SR_image, HR_image)
  - Adversarial Loss: Use Discriminator

    Adversarial_Loss = Discriminator.calculate_loss(SR_image, HR_image
  - Total Loss

    Total_Loss = λ * Content_Loss + (1 - λ) * Adversarial_Loss
  - Step 3: Backpropagation for Generator

    Generator.update_weights(Total_Loss)
  - Step 4: Update Discriminator

    Discriminator_Loss = Discriminator.train(SR_image, HR_image)
  - Log current progress

    log_metrics(epoch, Content_Loss, Adversarial_Loss, Total_Loss, Discriminator_Loss)
- Decision: Evaluation or Thresholding

  for each test_image:

  SR_test_image = Generator.forward(test_image)

  Evaluate_image_quality(SR_test_image, HR_reference)
- End of Training

  Save_final_model_weights()

  Generate_results(SR_images)

3.5 Adaptive Discriminator Augmentation

Adaptive Discriminator Augmentation (ADA) is a technique used to improve the stability and performance of GANs by dynamically applying augmentations to the discriminator's inputs. ADA is particularly useful in training GANs on limited datasets, where the discriminator can quickly overpower the generator, leading to mode collapse or failure to converge. This approach involves adding random augmentations to real and generated images before feeding them to the discriminator.

Unlike traditional data augmentation methods, ADA adapts the augmentation probability during training, adjusting based on the discriminator's performance. The concept of data augmentation, which involves applying transformations like flips, rotations, or color changes to training data, has long been used in machine learning to prevent overfitting and improve generalization. However, its application to GANs, specifically the use of adaptive augmentation, was introduced more recently. In 2020, the use of ADA in GANs gained prominence with the paper "Training Generative Adversarial Networks with Limited Data" by Karras et al., which introduced ADA in the context of StyleGAN2 [152]. These augmentations are adjusted dynamically based on the discriminator's feedback, allowing the GAN to maintain balance between the generator and discriminator. This was a breakthrough in GAN training, especially for scenarios where collecting large datasets is impractical or impossible, such as medical imaging or high-resolution artistic datasets.

Typically, when the discriminator is too powerful, it easily distinguishes between real and fake images, causing the generator to struggle and leading to a failure in training. ADA addresses this by applying transformations to the real and fake images before they are passed to the discriminator. These augmentations make the discriminator's task harder and force it to focus on higher-level features rather than memorizing specific details. Unlike traditional augmentation techniques that are applied consistently, ADA adjusts the probability of augmentation dynamically during training. This is done based on how well the discriminator is performing. If the discriminator starts to outperform the generator (as measured by the discriminator's loss), ADA increases the probability of augmentation. If the generator catches up, the augmentation probability is reduced. This adaptability helps maintain a balance between the two networks throughout training. The augmentations used in ADA are often lightweight and include operations like random cropping, flipping, color jittering, brightness adjustments, and adding Gaussian noise. These transformations do not significantly alter the core features of the images but introduce enough variation to challenge the discriminator.

One of the main challenges in GAN training is instability. GANs are notorious for issues like mode collapse, where the generator produces a limited variety of images, and non-convergence, where neither the generator nor the discriminator improves. By using ADA, the discriminator is prevented from becoming too powerful too quickly. This allows the generator to improve gradually, leading to more stable training. When GANs are trained on small datasets, there is a risk of overfitting. ADA reduces overfitting by ensuring that the discriminator learns to generalize features from augmented images rather than memorizing specific details. This leads to a more robust model that can generate more diverse and realistic images. The key advantage of ADA over traditional data augmentation is its adaptability. The augmentation is not static but evolves during training based on the needs of the model. This dynamic adjustment ensures that the augmentation is applied just enough to challenge the discriminator without making it impossible to train.

In GANs, the discriminator can overfit on small datasets by memorizing the details of the training images. ADA mitigates this by introducing augmentations that change the appearance of both real and fake images, preventing the discriminator from memorizing the real images and encouraging it to focus on more general features. Traditionally, GANs require large datasets to generate high-quality images. With ADA, even small datasets can produce competitive results, making it an attractive option in fields where data is scarce or expensive to collect. While ADA offers many benefits, there are also challenges to consider. One of the main challenges is tuning the augmentation probability and determining the right balance between augmentation and learning. If the augmentations are too aggressive, the discriminator may struggle to learn meaningful features, leading to poor performance. Conversely, if the augmentations are too weak, the discriminator may overpower the generator, causing instability in training. Another challenge is ensuring that augmentations do not significantly alter the image in ways that make the discriminator's task unrelated to the generator's goal. For example, excessive augmentations like large crops or extreme color changes could make it difficult for the discriminator to distinguish between real and fake images, resulting in poor learning.

The Adaptive Augmentation function is a key component for enhancing the robustness of the generator and discriminator. It applies a series of random augmentations to the training images, including horizontal flipping, brightness adjustment, contrast modification, saturation and hue adjustments, and Gaussian noise injection. These augmentations are designed to introduce diversity into the training data, which helps in regularizing the model and preventing overfitting. The augmentations are applied with a specified probability (p), which can be adjusted dynamically.

The function also includes a mechanism to adjust the augmentation probability based on the discriminator's loss. If the discriminator's loss falls below a predefined threshold, indicating that the discriminator is effectively distinguishing between real and generated images, the augmentation probability is increased. This forces the generator to become more robust to variations in the data. Conversely, if the discriminator is struggling to distinguish between real and fake images (indicating a potentially weaker discriminator), the augmentation probability is reduced to allow the model to focus on learning more stable features. This dynamic approach to data augmentation is crucial in training generative models like SRGANs, as it helps maintain a balance between the generator and discriminator, improving both their learning capabilities.

3.6 Loss function

The overall loss function in SRGAN is not a single function, since it is a combination of two loss components with different weights.

$$L_{SRGAN} = \alpha\ L_{content} + \beta\ L_{adversarial} \qquad (3-9)$$

In the Equation (3-9), α and β are coefficients that balance different loss terms. a common starting point might be α = 1 and β = 1. Content Loss evaluates the resemblance between the produced high-resolution image and the actual high-resolution image. Assisting the model in capturing the fundamental elements and arrangement of the authentic image. There are two methods to evaluate content loss in SRGAN: one is Pixel-wise Mean Squared Error (MSE), a traditional method that may result in images being overly smooth without high-frequency details, and the other is Perceptual Loss (VGG Loss), which utilizes features from a pre-trained VGG network to assess similarity. This method more effectively captures significant perceptual details. The MSE loss fails to adequately capture the pixel-level variations in image texture. A pixel in a high-resolution image is composed of several different combinations and MSE loss normally takes an average from these combinations which looks unrealistic in comparison with the ground truth. To resolve this issue, the original SRGAN model utilized a VGG loss derived from measuring the Euclidean distance between feature maps obtained from the VGG-19 model. The perceptual loss function is expressed in Equation (3-10):

$$L_{per} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\Phi_{i,j}(I^{HR})_{x,y} - \Phi_{i,j}(G(I^{LR}))_{x,y}\right)^2 \qquad (3-10)$$

Where $W_{i,j}$ denotes the width dimension of feature maps within the VGG network. $H_{i,j}$ denotes the height size of the feature maps within the VGG network. $\Phi_{i,j}$ represents the feature map generated by the j-th convolutional layer before the i-th convolutional layer within the

network. I^{HR} denotes an image of high resolution. I^{LR} represents a low-resolution image. Therefore, $L_{content}$ can be replaced by L_{per} in Equation (1). A popular option for adversarial loss in SRGAN is the Binary Cross Entropy (BCE) loss, utilized to train the discriminator function and differentiate between genuine high-resolution images and synthesized images.

$$L_{adv} = -E_{real}[\log(D(\mathrm{HR}))] - E_{fake}\left[\log\left(l - D(\mathrm{G(LR)})\right)\right] \quad (3-11)$$

In Equation (3-11), E_{real} and E_{fake} represent expectation of real and fake image distributions, D is the Discriminator network, HR is the Real high-resolution image, and G (LR) is the high-resolution image produced by the Generator network using a low-resolution input.

3.7 Algorithm performance (Evaluation metrics)

3.7.1 PSNR

PSNR, or Peak Signal-to-Noise Ratio, is commonly used to measure image quality numerically. It compares the original signal with the compressed or reconstructed version to assess the amount of distortion present. The PSNR measures the gap between the highest achievable signal power and the power of the corrupted signal. Higher PSNR values, expressed in decibels (dB), indicate better image quality with less distortion, while lower values mean lower quality and more distortion. This metric is employed based on Mean Squared Error, which calculates the mean squared difference between corresponding pixels in the original and distorted signals. The PSNR scale is obtained from the MSE by utilizing a logarithmic function, as shown in Equation (3-12):

$$PSNR = 20\log_{10}\frac{MaxI}{\sqrt{MSE}} \quad (3-12)$$

MaxI is the highest attainable pixel value found in an image. MSE is the average of the squared discrepancies between matching pixels in two images.

$$MSE = \frac{1}{M*N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(f_{ij} - f'_{ij}\right)^2 \quad (3-13)$$

As shown in Equation (3-13), f_{ij} represents the pixel values of the initial high-resolution image while f'_{ij} represents the pixel values of the reconstructed image. The image's width and height are represented by M and N.

3.7.2 SSIM

SSIM, short for Structural Similarity Index, is a widely used measure that assesses how

similar images are by taking into account both structural details and pixel values in order to determine their perceived quality. It examines three elements - luminance, contrast, and structure - by comparing nearby groups of pixels in original and altered pictures. Luminance assesses brightness similarity, contrast measures differences in contrast levels, and structure evaluates similarity in pattern and texture. Merging the calculated component scores forms an index that varies from 0 to 1, where a score of 1 represents complete similarity and a score of 0 represents complete dissimilarity. SSIM, unlike traditional metrics such as PSNR, offers a more perceptually meaningful assessment by considering human visual perception and image structure.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \quad (3-14)$$

In Equation (3-14), x is the image in its original high-resolution form. y is the reconstructed image. μ_x and μ_y are the mean values of the images, while σ_y^2 and σ_x^2 are the variances. σ_{xy} is the covariance between the two images.

3.7.3 LPIPS

The Learned Perceptual Image Patch Similarity (LPIPS) metric is an advanced Image Quality Assessment (IQA) tool designed to evaluate the perceptual similarity between images. It leverages deep features extracted from pre-trained neural networks, aligning closely with human perception of image quality. LPIPS is particularly effective in capturing semantic and perceptual differences between images, making it widely used in tasks like image synthesis, super-resolution, and restoration. LPIPS utilizes CNNs to extract features from images, rather than comparing raw pixel values. The metric computes the perceptual distance between images based on the activations of a pretrained network (e.g., AlexNet, VGG). LPIPS computes feature-based similarity using deep neural networks.

$$\mathrm{LPIPS(I,K)} = \sum_{\mathrm{l}} \mathrm{w_l} \cdot |\mathrm{F_{l(I)}} - \mathrm{F_{l(K)}}|_2^2 \quad (3-15)$$

Where $F_{l(I)}$ and $F_{l(K)}$ are feature maps from a pre-trained network (e.g., VGG, AlexNet) and $|F_{l(I)} - F_{l(K)}|_2^2$ represents the squared Euclidean distance between corresponding feature maps and w_{l} are learned weights for each feature layer l.

3.7.4 DISTS

DISTS (Deep Image Structure and Texture Similarity) is another perceptual image quality metric that quantifies the similarity between two images by separating the structural and textural components. It combines a deep learning-based model for feature extraction with a perceptual

loss function to measure the similarity more accurately than traditional metrics like PSNR and SSIM. DISTS combines structural and texture similarity.

$$\mathrm{DISTS}(\mathrm{I},\mathrm{K}) = \sum_{\mathrm{l}} \lambda_{\mathrm{l}|\mathrm{S}_{\mathrm{l}(\mathrm{I})}-\mathrm{S}_{\mathrm{l}(\mathrm{K})}|_2^2} + \sum_{\mathrm{l}} \mu_{\mathrm{l}|\mathrm{T}_{\mathrm{l}(\mathrm{I})}-\mathrm{T}_{\mathrm{l}(\mathrm{K})}|_2^2} \qquad (3-16)$$

Where S_l and T_l are structure and texture components extracted from deep network layers. λ_l and μ_l are weights balancing structure vs. texture. The values are typically learned from data using a deep network, making the metric adaptable to various tasks. DISTS uses a pre-trained deep network (e.g., VGG or ResNet) to extract high-level features that capture the structural information of images. Both LPIPS and DISTS scores range from 0 to 1, where a score of 0 signifies that the two images are perceptually identical, while a score of 1 represents the greatest possible perceptual difference between them.

3.7.5 BRISQUE

The Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) is a no-reference image quality assessment (NR-IQA) metric that evaluates the perceptual quality of an image without requiring a high-quality reference image. Unlike traditional metrics such as PSNR and SSIM, which compare a given image to its original high-quality version, BRISQUE assesses image distortions by leveraging Natural Scene Statistics (NSS) in the spatial domain. This makes it particularly useful for real-world applications where a reference image may not be available. The BRISQUE score for an image x is computed as follows:

$$\mathrm{BRISQUE}(\mathrm{x}) = \mathrm{GMM}\big(\mathrm{NSS}(x)\big) \qquad (3-17)$$

Where $\mathrm{NSS}(x)$ represents the extracted natural scene statistics from the image. GMM is the regression model that maps the NSS features to a perceptual quality score. The output score typically ranges from 0 to 100, where a higher score indicates worse quality and a lower score suggests better image quality.

The image x is first converted to grayscale and resized (if needed) for feature extraction, then BRISQUE extracts various statistical features from the image that reflect properties like sharpness, contrast, and texture. The extracted features are passed through a Gaussian Mixture Model (GMM) that was trained on images with subjective quality ratings. Finally, the model outputs a score that predicts the perceptual quality of the image. For the final benchmark comparison between different models in this study, we also measured the BRISQUE score for an image using the 'brisque' Python library, which provides an easy way to compute BRISQUE scores. Table 3-1 compares the different ranges of BRISQUE with their corresponding image

quality classification.

Table 3-1 BRISQUE range, Lower BRISQUE scores indicate better perceptual quality

BRISQUE Score	Perceptual Image Quality
0 - 20	Excellent quality
20 - 40	Good quality
40 - 60	Fair quality
60 - 80	Poor quality
80 - 100	Very poor quality

3.8 Dataset

Figure 3-22 Examples from the dataset of Pix4Dmatic

3.8.1 The Main Dataset

We utilize the Pix4Dmatic dataset, which contains UAV-captured (including DJI's Phantom 4 RTK) images in .JPG format with a resolution usually between 12 to 20 MP per image, varying based on the camera type. The datasets frequently encompass different settings like industrial, agricultural, and urban areas, and may contain images captured by multi-spectral sensors or RGB sensors. Some examples from this dataset are illustrated in Figure 3-22. The information is appropriate for producing dense point clouds, DSMs, and orthomosaics. DSMs (Digital Surface Models) portray the surface of the Earth, encompassing both natural features and man-made structures such as trees and buildings. They are essential for applications such as urban planning and flood modeling as they are utilized to analyze the height of surfaces. Orthomosaics are images corrected for geometry that are made by combining numerous aerial

photographs. In contrast to raw images, orthomosaics are consistent in size, accurately depicting the Earth's surface for precise mapping, land-use assessment, and tracking changes. Ground control points (GCPs) are included for improved accuracy.

3.8.2 The Benchmark Datasets

To ensure the integrity and generalizability of our SRGAN model, we extend our experiments beyond UAV images by training on DIV2K and Flickr2K, two widely used benchmark datasets in super-resolution research. These datasets provide high-quality, diverse images, enabling a fair comparison with existing state-of-the-art models such as Swin2SR, EDSR, and ESRGAN, which have been trained on the same datasets.

DIV2K consists of 1,000 high-resolution images, offering a diverse set of natural and urban scenes, while Flickr2K expands the training set with 2,650 additional high-quality images collected from Flickr. Both datasets enhance the model's ability to generalize across different image types. Since many pretrained SR models are available on HuggingFace using these datasets, adopting them streamlines our training process and facilitates meaningful performance comparisons. By training our optimal SRGAN variant on DIV2K and Flickr2K, we validate its effectiveness beyond UAV imagery, ensuring that our findings align with established benchmarks in the super-resolution domain. This reinforces the practicality of our design and its applicability to general image restoration tasks.

3.9 Training details and parameters

The SRGAN model, consisting of both generator and discriminator networks, was implemented using the TensorFlow framework and trained on a Windows 10 system equipped with an Nvidia GeForce RTX 2060 GPU. The input to the generator is a low-resolution image of size 64×64×3, and the output is a super-resolved image of size 256×256×3.

During training, the Adam optimizer is used for both the generator and discriminator, with a learning rate of 1×10^{-4}, $\beta1=0.9$, and $\beta2=0.999$. The proposed models are trained for 5000 epochs with a batch size of 16. At the end of each epoch, performance metrics such as PSNR, SSIM, LPIPS, and DISTS are computed to evaluate the quality of the generated images. Additionally, sample super-resolved images are generated and saved every 10 epochs for visual inspection.

Model checkpoints are saved every 300 epochs to allow for recovery and evaluation of the model at different training stages. The training process is logged using TensorBoard, which tracks the generator and discriminator losses, PSNR, SSIM, and generated images. A custom CSV logger is also implemented to store the loss and metric values for each epoch, allowing

for detailed analysis of the training process. The training dataset consists of high-resolution images resized to 256×256, with corresponding low-resolution images created by down-sampling by a factor of 4. Both low- and high-resolution images are normalized to the [−1,1] range.

3.10 YOLO9x Integration

In this study, to further demonstrate the practical advantages of the proposed super-resolution model, YOLO9x [153] was integrated into the experimental pipeline. YOLO9x, a pre-trained state-of-the-art object detection model, was used in its prediction mode, utilizing pre-trained weights (yolov8x.pt) without any additional training or fine-tuning. This choice was made because YOLO9x already possesses robust pre-trained weights for identifying common objects, such as cars and trees. The goal of this experiment was to evaluate the impact of image resolution, enhanced by the proposed SR model, on the accuracy and performance of object detection tasks. By comparing YOLO9x's detection outputs on low-resolution versus SR-enhanced images, the experiment showcases how improved image quality contributes to the identification of small or partially visible objects, emphasizing the broader applicability of the SR model in real-world scenarios like urban planning and environmental monitoring [154], [155], [156].

3.11 Summary

This chapter provides a comprehensive examination of SRGAN, focusing on its use of content loss, which replaces MSE to prioritize perceptual fidelity over pixel-wise accuracy. This shift enables SRGAN to significantly improve image quality by enhancing perceptual details, which are more relevant for human visual perception. The chapter also discusses the general structure of GANs, highlighting the roles of the generator and discriminator, and how SRGAN utilizes these components to enhance image resolution. It addresses the limitations of MSE, particularly its inability to capture perceptual quality, and demonstrates how SRGAN overcomes this with content loss derived from high-level feature maps of a pre-trained VGG network.

Additionally, the chapter introduces two proposed models, U-SRGAN and A-SRGAN. U-SRGAN incorporates a U-Net architecture with attention gates for better feature extraction, while A-SRGAN integrates autoencoders with residual blocks. Both models aim to enhance image quality and computational efficiency. The chapter also introduces ADA to improve GAN stability, especially with limited data, by preventing the discriminator from overpowering the generator. In conclusion, the chapter outlines key techniques to optimize super-resolution, focusing on perceptual quality and efficiency.

4 Experimental Process and Results

4.1 Introduction

As outlined in Chapter 3, we designed multiple architectures to determine which model yields the best results in enhancing image resolution. A significant challenge in training GAN networks is the requirement for vast datasets consisting of thousands of images. However, our dataset contains only 250 images, which is insufficient for training a robust GAN model. To address this limitation, we utilized Adaptive Data Augmentation (ADA) to mitigate overfitting by adaptively feeding augmented data to the discriminator. This approach enhances the discriminator's performance during training without allowing it to overfit to the limited dataset. Another critical aspect of GAN training is the number of epochs. Typically, GAN models benefit from training over many epochs to optimize performance. However, due to computational constraints and the time-intensive nature of training GANs, we limited the number of epochs to 5000 in this study. While we acknowledge that this may not be sufficient for absolute convergence, it strikes a balance between model performance and available computational resources.

Table 4-1 presents the number of parameters for the generators in each of our models, along with the corresponding training times. In this study, we primarily focused on experimenting with different generator architectures while maintaining the discriminator's architecture, as we believe its binary classification task is already well-suited to this domain. We explored several variations of the generator. Initially, we replaced the residual blocks in the generator with an autoencoder architecture featuring convolutional layers as skip connections, inspired from LinkNet, then we call it A-SRGAN. In another model, we further enhanced this autoencoder by incorporating residual blocks within its design (Res-A-SRGAN). Subsequently, we experimented with a U-Net architecture in place of the residual blocks (U-SRGAN) and introduced another variant of U-Net combined with residual blocks and attention gates to improve the model's capacity to capture finer details (ARUnet-SRGAN). Finally, for the last model, first we transfer the weights from an autoencoder to the encoder path of U-Net with attention gate and residual blocks (pretrained ARUnet-SRGAN).

For comparison purposes, we also trained a standard SRGAN model to evaluate performance differences across the various architectures. In addition, we trained an ESRGAN model to assess its ability to enhance image resolution beyond traditional methods. Finally, to push the boundaries of our research, we trained our best-performing model for 50,000 epochs, achieving remarkable results. However, the extended training time, approximately two weeks,

posed significant limitations, making it infeasible to apply this extensive training to all models within the scope of this study. Future research with more powerful GPUs could substantially reduce training times and enable a broader exploration of model architectures.

The original images were sized at 1024x1024 pixels, but to reduce the computational load, we resized them to 256x256 for high-resolution images and 64x64 for low-resolution images. To evaluate the performance of our super-resolution models like SRGAN, we employed four commonly used metrics: PSNR, SSIM, LPIPS, and DISTS. SSIM values range from 0 to 1, with 1 indicating perfect structural similarity between the generated and original image. PSNR values typically fall between 20 dB and 40 dB, with higher values reflecting better image quality. In previous studies using datasets like DIV2K or BSD100, GAN-based models have reported PSNR values around 28–32 dB and SSIM scores in the range of 0.85 to 0.9. Both LPIPS and DISTS are image quality metrics designed to measure perceptual similarity between images. The range of both metrics spans from 0 to 1, where a value of 0 indicates that the two images are perceptually identical, while a value of 1 signifies the maximum possible perceptual difference between the images.

Table 4-1 Different models along with number of generator parameters and training duration

Models	Generator parameters	Training hours
SRGAN	4,555,267	37
A-SRGAN	6,502,787	39
Res-A-SRGAN	10,904,963	42
U-SRGAN	36,214,787	46
ARUnet-SRGAN	42,469,991	47
Pretrained ARUnet-SRGAN	42,469,991	47

The table above presents a comparison of various SRGAN models used in this study, focusing on the generator parameters and corresponding training hours for each model. From the data, it is evident that as the complexity of the model increases, the number of parameters grows significantly, which in turn increases the computational time required for training. For instance, the base SRGAN model has approximately 4.5 million parameters and requires 37 hours of training, while the SRGAN with U-Net and attention gates has over 42 million parameters, necessitating 47 hours of training. This highlights the direct correlation between model complexity and training duration.

The addition of Autoencoders, ResNet blocks, and U-Net architectures has significantly expanded the number of parameters in the generator. These modifications were intended to

enhance the model's ability to capture intricate details and improve the quality of the super-resolved images. However, they come at the cost of increased computational resources and time. It is also important to note that, while more complex models like U-SRGAN and SRGAN with attention gates may improve performance, they also present a risk of overfitting, especially with smaller datasets. Proper regularization techniques and adequate training data are essential to counteract this risk.

Moreover, the trade-offs between performance gains and computational feasibility are crucial considerations in practical applications. The increased training time associated with larger models, as shown in the Table 4-1, may limit their usability in scenarios where rapid model deployment or iteration is required. Nonetheless, these models offer promising results, and with further advancements in hardware or model optimization techniques, such as pruning or quantization, the training time can be reduced without sacrificing performance.

In this chapter, we first present a comparative analysis between the two models that utilize Autoencoders in their generators and the baseline SRGAN model. This comparison focuses on both metric performance and the visual differences observed in the generated images. Following this, we move on to evaluate the SRGAN base model against those models that incorporate U-Net architectures within the generator. Finally, we conclude by showcasing the results of training our best-performing model over 50,000 epochs, where the low-resolution images are sized at 256x256. Providing ample training epochs ensures that the model has sufficient time to learn and generate higher-quality images, offering readers a clearer understanding of the potential improvements and the overall effectiveness of the final results.

4.2 A-SRGAN results

The results from comparing the three models (SRGAN, A-SRGAN, and Res-A-SRGAN) reveal clear trends in performance improvements when utilizing autoencoder-based architectures with skip connections, and particularly when incorporating residual blocks within the autoencoder. These models were evaluated across five key metrics: SSIM, PSNR, LPIPS, DISTS, and Generator Loss, each of which highlights distinct aspects of model performance. Figure 4-1 demonstrates the comparison of three models based on aforementioned metrics. Res-A-SRGAN consistently outperforms the other models, showing the highest stability and perceptual quality. Specifically, Res-A-SRGAN achieves superior SSIM and PSNR values, highlighting its ability to preserve structural details and improve pixel-wise fidelity compared to SRGAN, which lags behind, especially in PSNR. A-SRGAN also shows significant improvements, with better stability than SRGAN but slightly lower performance than Res-A-

SRGAN.

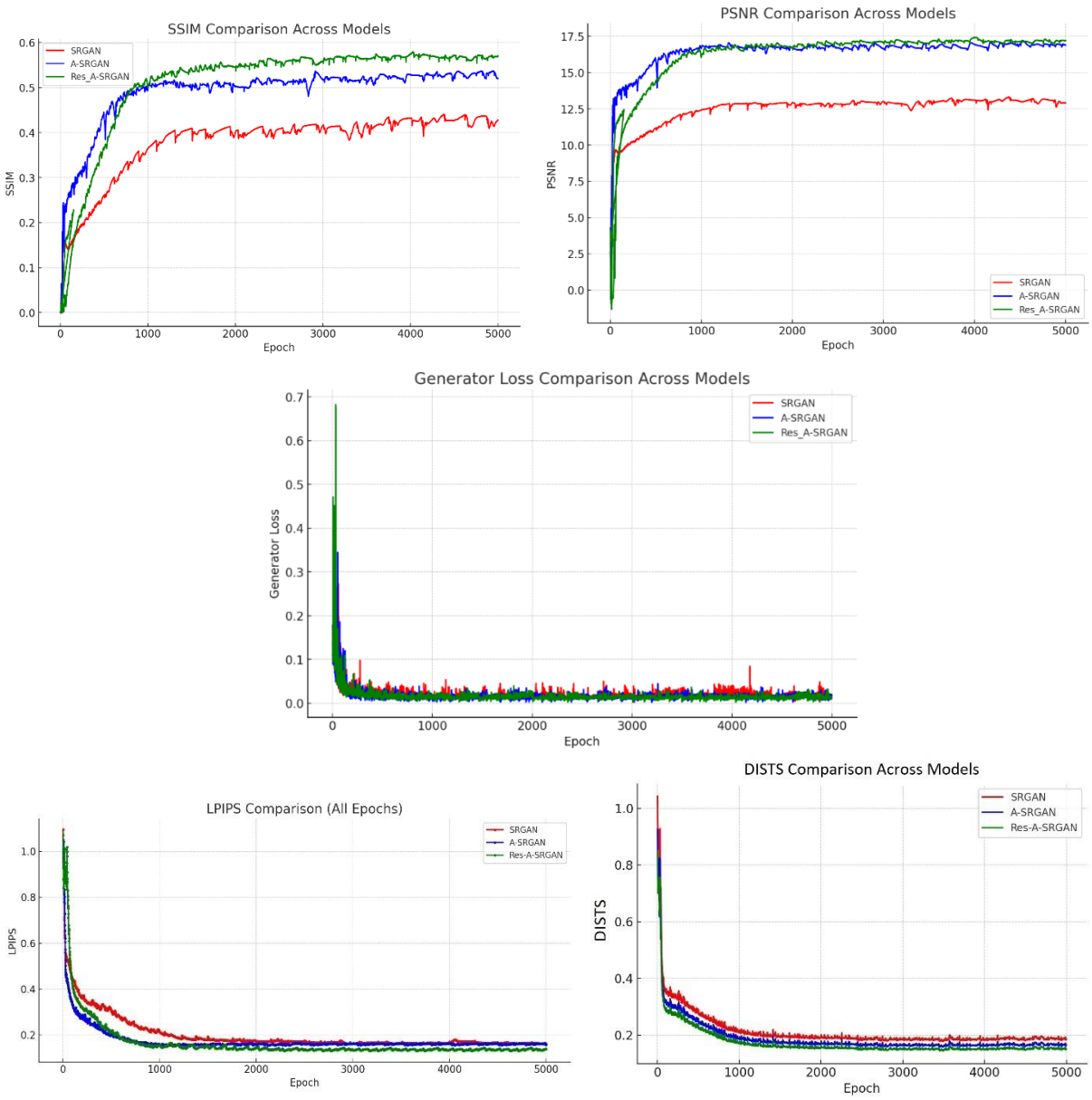

Figure 4-1 The comparison of SSIM and PSNR, LPIPS, DISTS, and generator loss between typical SRGAN and SRGAN with autoencoder in their generator architecture

In terms of perceptual and structural quality, Res-A-SRGAN demonstrates the best results in LPIPS and DISTS, with both metrics showing a stable and robust performance throughout training. This indicates that the introduction of residual blocks within the autoencoder architecture enhances both perceptual similarity and image structure. In contrast, SRGAN performs the worst, with notably higher values for both LPIPS and DISTS, suggesting poorer perceptual similarity and structural consistency with the ground truth.

The Generator Loss analysis further corroborates these findings, with Res-A-SRGAN showing the most rapid convergence and smoother learning dynamics, followed by A-SRGAN.

SRGAN exhibits slower convergence, indicating less stable learning. These results confirm that A-SRGAN and Res-A-SRGAN, with their improved architectures, offer better generalization and more effective learning compared to SRGAN.

Overall, the experimental results demonstrate that both A-SRGAN and Res-A-SRGAN offer significant improvements over the standard SRGAN in terms of both image quality and training stability. The introduction of autoencoders with convolutional layers as skip connections allows for better feature preservation and flow, while the use of residual blocks in Res-A-SRGAN further enhances this effect by facilitating smoother gradient propagation during training. These modifications lead to superior SSIM and PSNR values, as well as more efficient training, as evidenced by the lower generator loss. In conclusion, the proposed Res-A-SRGAN model, which incorporates residual blocks into an autoencoder with skip connections, represents a meaningful advancement in the field of super-resolution, achieving higher fidelity and more structurally accurate image reconstructions compared to conventional SRGAN models.

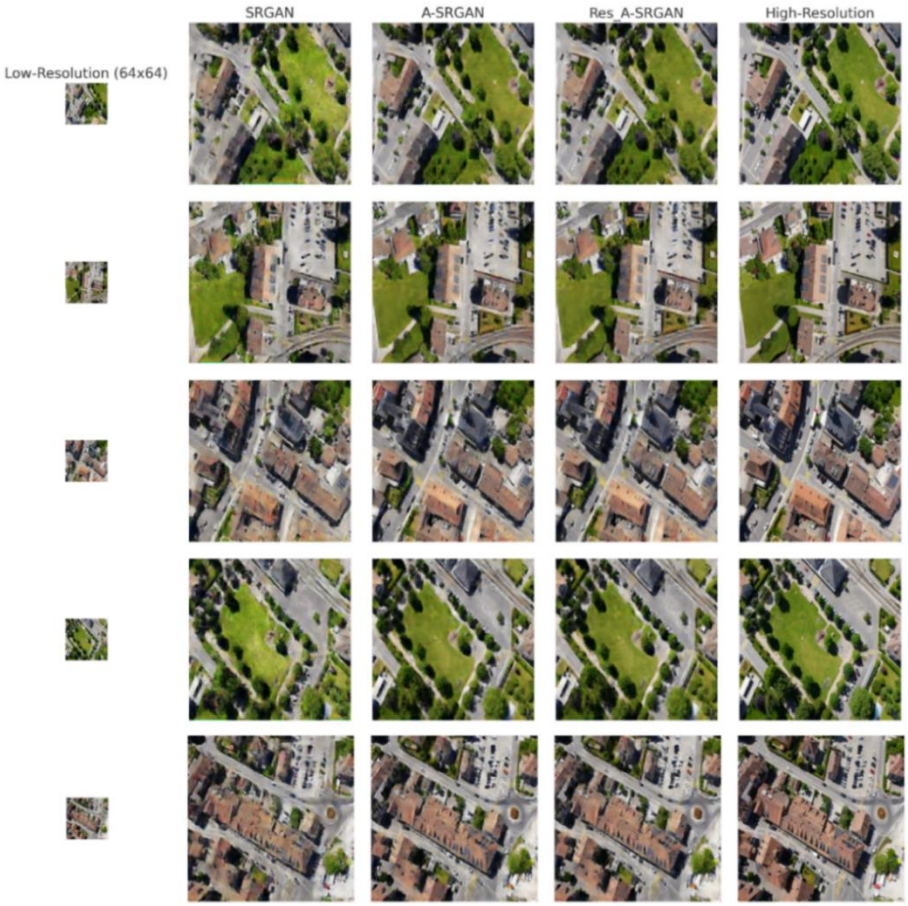

Figure 4-2 Examples of generated images created by three different models studied

4.2.1 Qualitative results of A-SRGAN

results of the three different models, as shown in Figure 4-2, on the UAV-captured dataset highlight both the strengths and limitations of super-resolution techniques. These images, captured by UAVs, contain intricate details such as parked cars, street lanes, and fine textures on building roofs, making them particularly challenging for super-resolution models. While A-SRGAN succeeds in enhancing the resolution and improving the overall visual quality of these images, it struggles with recreating some of the more minute elements, like the finer points on rooftops or specific street markings. These tiny features, which are often critical for urban planning or traffic analysis, present a challenge for the models to accurately reconstruct.

Nonetheless, the overall visual improvement in larger structures, like buildings and roads, demonstrates the effectiveness of A-SRGAN and Res-A-SRGAN compare to typical SRGAN in generating high-resolution outputs from low-resolution inputs. Figure 4-3 displays some samples from different images of test datasets that were fed into our three models.

Figure 4-3 Comparison of SRGAN, A-SRGAN, and Res-A-SRGAN models with SSIM and PSNR values displayed for each model.

The SRGAN model, which serves as the baseline, achieved an SSIM of 0.556 and a PSNR of 28.465 dB. While SRGAN is able to reconstruct the general layout of buildings and streets, it struggles with the finer details, particularly in areas like rooftops and small objects such as parked cars. The output from SRGAN appears blurry in these regions, demonstrating that the model fails to capture high-frequency information required for clear reconstruction of such intricate structures. In contrast, A-SRGAN shows significant improvements, producing an SSIM of 0.646 and a PSNR of 28.673 dB. The enhanced architectural modifications in A-SRGAN lead to better structural similarity, with the model being more capable of reconstructing finer details compared to SRGAN. However, despite these improvements, certain challenging regions such as rooftops and vehicles still lack clarity, indicating that A-SRGAN, while effective, has limitations when it comes to ultra-fine features. The best performance is observed in the Res-A-SRGAN model, which achieved the highest SSIM of 0.653 and a PSNR of 29.213 dB. This model demonstrates the most effective reconstruction of high-frequency details, particularly in preserving street lanes and textures on rooftops. Res-A-SRGAN outperforms the other models in generating clearer outputs, making it the most robust model for this dataset. Figure 4-4 shows an extra sample of generated photos created by three models.

Figure 4-4 Comparison of SRGAN, A-SRGAN, and Res-A-SRGAN models with SSIM and PSNR values displayed for each model.

It is worth noting that during training, the maximum SSIM reached by the best model was 0.57. However, the SSIM for this particular image increased to 0.653 for Res-A-SRGAN. This discrepancy can be explained by the nature of the evaluation dataset. The SSIM of 0.57 was likely an average calculated over various images, each presenting different levels of complexity and noise. Some images may have been more challenging for the model to reconstruct, resulting in a lower average SSIM. However, the image shown in this evaluation may have fewer noise issues or clearer structures, which allowed Res-A-SRGAN to perform better, achieving a higher SSIM of 0.653. In summary, the Res-A-SRGAN model provides the best qualitative results, particularly in terms of SSIM and PSNR, demonstrating superior performance in generating high-resolution images with more accurate structural details. The model's ability to better reconstruct intricate details such as rooftop textures and street markings marks a significant improvement over SRGAN and A-SRGAN.

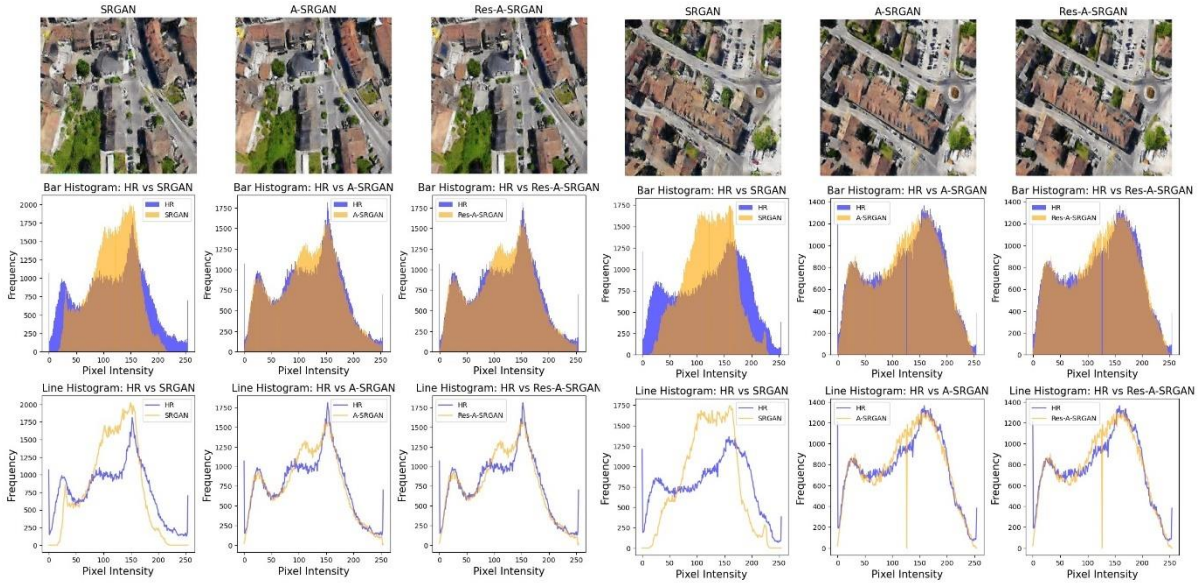

Figure 4-5 Comparison of two different images generated by three models and their corresponding line and bar histograms as opposed to their respective HR images

Figure 4-5 presents a comparison of SRGAN, A-SRGAN, and Res-A-SRGAN outputs against the HR images using bar histograms, line histograms, and visual analysis. Bar histograms provide a detailed view of pixel intensity distributions, allowing precise evaluation of how well each model preserves brightness and contrast. Line histograms complement this by showing overall trends, making it easier to identify deviations or alignment with the HR images. Together, they offer a comprehensive understanding of the models' performance.

The results highlight that SRGAN struggles with detail preservation and exhibits significant deviations from the HR image in both bar and line histograms, particularly in dark and bright regions, indicating inferior reconstruction quality. A-SRGAN demonstrates notable

improvements, with better alignment to the HR image and reduced deviations, especially in mid-range intensities. However, Res-A-SRGAN outperforms both, showing the closest match to the HR image in all metrics, effectively preserving details, brightness, and contrast. This analysis underscores Res-A-SRGAN's superior performance, followed by A-SRGAN, with SRGAN lagging behind.

4.2.2 Ablation Study on Autoencoder and Residual Blocks in SRGAN

The ablation study begins by examining the evolution of the SRGAN architecture, progressing from a basic residual block-based generator to a more sophisticated design incorporating autoencoders with residual blocks and convolutional skip connections. Initially, the SRGAN model relied purely on residual blocks to learn effective mappings for image super-resolution. This foundational approach allowed the model to learn residual differences between high-resolution and low-resolution images, improving image quality to some extent.

4.2.2.1 The Impact of Skip Connections

To push the boundaries of this architecture, we first replaced the residual blocks in the generator with an autoencoder structure, enhanced by skip connections. This modification, which we refer to as A-SRGAN, facilitated better feature extraction and retention, especially for the essential characteristics necessary for image upsampling. Skip connections in the autoencoder played a critical role by preserving important features during the encoding-decoding process. This allowed the model to focus on more fundamental changes needed for high-resolution reconstruction, improving image quality by learning a better feature representation.

Building on this, we incorporated residual blocks into the autoencoder, creating the Res-A-SRGAN model. This design improved the model's ability to retain finer details and learn more complex mappings. The residual blocks helped with more efficient gradient flow and improved training stability, especially in deeper networks. As a result, the Res-A-SRGAN model demonstrated superior performance over both SRGAN and A-SRGAN, highlighting the cumulative benefits of combining autoencoders with residual blocks.

The progression from SRGAN to A-SRGAN and then to Res-A-SRGAN shows a clear enhancement in the model's ability to handle intricate details such as textures on rooftops, fine edges, and street markings, which are critical for applications like UAV-based urban planning or traffic analysis. This ablation study underscores the importance of integrating autoencoders with residual blocks to further elevate the performance of GANs in super-resolution tasks.

In our ablation study, we evaluated the impact of using skip connections and convolutional

layers within skip connections on the performance of A-SRGAN and Res-A-SRGAN. The results demonstrate that incorporating skip connections into the autoencoder structure provides significant improvements over configurations without them, and replacing traditional skip connections with convolutional layers further enhances the network's performance. These findings emphasize the importance of efficient feature propagation and transformation in super-resolution tasks. The results are shown in Table 4-2 and Table 4-3.

Table 4-2 Impact of Skip Connections and Convolutional Layers on PSNR and SSIM for A-SRGAN and Res-A-SRGAN

Models	PSNR ↑			SSIM ↑		
	Without skip connections	With skip connections	Convolutional layers instead of skip connections	Without skip connections	With skip connections	Convolutional layers instead of skip connections
A-SRGAN	13.5 dB	14.7 dB	**16.9 dB**	0.37	0.45	**0.52**
Res-A-SRGAN	15.1 dB	16.5 dB	**17.2 dB**	0.42	0.48	**0.57**

Table 4-3 Impact of Skip Connections and Convolutional Layers on LPIPS and DISTS for A-SRGAN and Res-A-SRGAN

Models	LPIPS ↓			DISTS ↓		
	Without skip connections	With skip connections	Convolutional layers instead of skip connections	Without skip connections	With skip connections	Convolutional layers instead of skip connections
A-SRGAN	0.1851	0.1797	**0.1566**	0.1920	0.1833	**0.1737**
Res-A-SRGAN	0.1598	0.1434	**0.1389**	0.1762	0.1520	**0.1493**

For both A-SRGAN and Res-A-SRGAN, using skip connections offers notable advantages. Skip connections create a direct pathway for transferring low-level features, such as textures and edges, from the encoder to the decoder, which prevents the loss of essential spatial information during encoding. This direct transfer helps the network preserve crucial image details, leading to improved perceptual quality and reconstruction accuracy. By mitigating issues like vanishing gradients, skip connections also stabilize training and enable deeper network architectures to perform effectively. As a result, models with skip connections achieve higher values in all metrics compared to those without them.

Building on this, we took inspiration from LinkNet and replaced traditional skip connections with convolutional layers. Instead of simply bypassing features, the convolutional layers refine and adaptively transform the features being propagated. This refinement ensures

that the features transferred to the decoder are not only preserved but also enriched, allowing the network to better reconstruct sharp and accurate high-resolution images. For instance, A-SRGAN with convolutional layers in skip connections outperforms both the no-skip and standard skip configurations, demonstrating its ability to deliver superior feature propagation and refinement. This idea, adapted from LinkNet's efficient design, enables the model to improve feature transmission while learning more complex feature transformations.

Similarly, Res-A-SRGAN benefits from this modification, combining the strengths of residual blocks with convolutional layers in skip connections. While residual blocks already improve learning by focusing on differences between low- and high-resolution images, the addition of convolutional layers in skip connections further refines the transferred features. This allows the network to reconstruct intricate details more precisely, improving both perceptual quality and pixel-level accuracy. As a result, Res-A-SRGAN achieves the highest PSNR and SSIM values among all tested configurations, highlighting the effectiveness of convolutional skip connections in enhancing super-resolution results.

4.2.2.2 The Impact of ADA

Furthermore, Training GANs on small datasets, such as one containing only 250 images, often leads to significant challenges. Without sufficient data diversity, the discriminator quickly overfits to the limited training samples, causing it to overpower the generator. This imbalance destabilizes the adversarial training process, leading to several issues such as mode collapse, training instability**,** and poor generalization**.** As shown in Figure 4-6, when trained without ADA, the generator loss exhibits erratic spikes and fluctuations, indicating instability in the learning process. Similarly, key performance metrics such as PSNR and SSIM show inconsistent trends, with high variance across epochs. This instability arises from the discriminator's inability to generalize beyond the limited data, which in turn forces the generator to oscillate between suboptimal solutions rather than converging to a stable and effective mapping. As a result, the generated images lack sharpness, exhibit visible artifacts, and fail to accurately reconstruct structural details, making it evident that training GANs without augmentation in such scenarios is inefficient and unreliable.

To address these issues, we incorporated Adaptive Data Augmentation (ADA) into the training process. ADA dynamically applies augmentations such as random flips, brightness adjustments, contrast changes, saturation modifications, hue shifts, and Gaussian noise to the training images. This introduces variability into the dataset, effectively simulating a larger and more diverse set of training samples. Unlike standard data augmentation, which applies

transformations uniformly, ADA adjusts the augmentation probability based on the discriminator's loss. When the discriminator becomes overly confident (low loss), ADA increases the augmentation probability to challenge the discriminator with more diverse inputs. Conversely, when the discriminator weakens, the augmentation probability is reduced to avoid destabilizing the training process. This adaptive approach ensures a balanced adversarial training dynamic, preventing the discriminator from overfitting while encouraging the generator to learn more robust mappings.

With ADA, the training process stabilized significantly. Metrics such as PSNR and SSIM

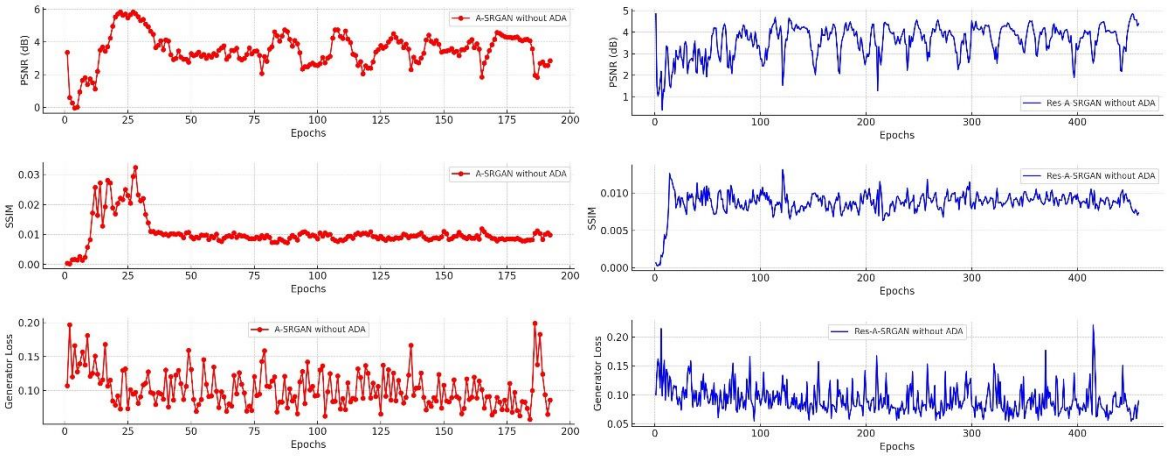

Figure 4-6 PSNR, SSIM, and Generator Loss for A-SRGAN and Res-A-SRGAN without ADA during training

showed smoother and more consistent trends, reflecting the generator's improved ability to reconstruct high-resolution images with better structural similarity and pixel-level accuracy. The generator loss also displayed a steady downward trend with fewer spikes, indicating balanced adversarial learning. Furthermore, ADA allowed the generator to produce sharper and more realistic images, even with the limited training dataset. These improvements highlight the effectiveness of ADA in addressing the challenges of training GANs on small datasets. By dynamically increasing data diversity and maintaining a balanced adversarial process, ADA proved to be a critical component in achieving high-quality results for our super-resolution task.

4.3 Results of SRGAN with U-Net

The performance analysis of three models (SRGAN, U-SRGAN, and ARUnet-SRGAN) highlights the impact of architectural modifications on super-resolution quality. SRGAN, which utilizes a standard residual structure, shows an initial improvement in all metrics, but reaches a lower plateau compared to the other models. This suggests that the SRGAN structure may be limited in its ability to capture intricate details necessary for high-quality image reconstruction. The generator loss of SRGAN converges to a stable point but remains slightly higher, indicating

potential limitations in the model's ability to learn fine-grained image details. The U-SRGAN model enhances SRGAN by replacing its residual blocks with U-Net components, introducing downsampling and upsampling layers that capture multi-scale features effectively. This modification leads to notable improvements in all metrics' values, as shown in Figure 4-7.

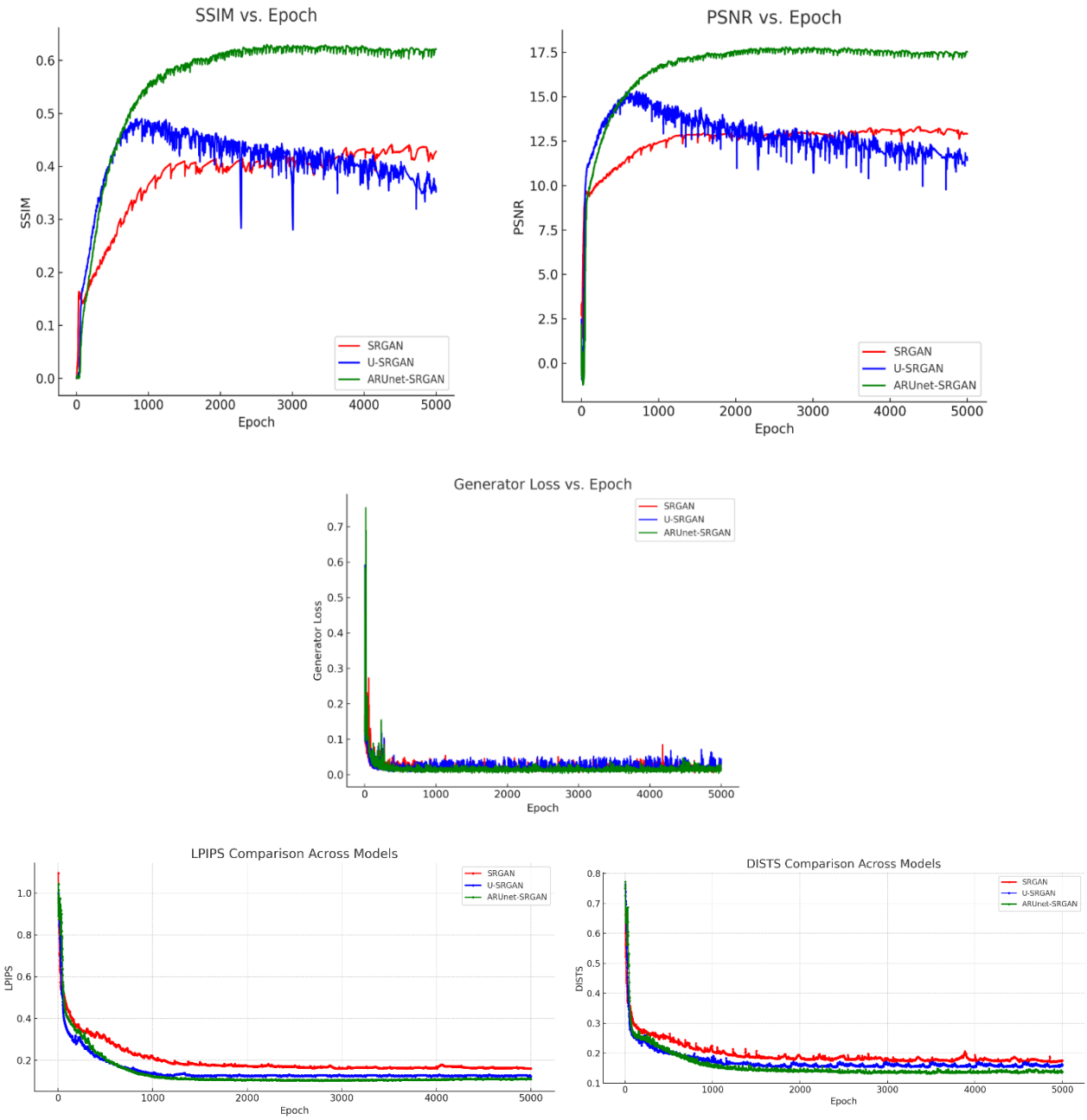

Figure 4-7 The comparison of metrics and generator loss between typical SRGAN and SRGAN with U-Net in their generator architecture

The U-Net architecture, with its skip connections, preserves low-level features essential for detailed reconstruction, allowing U-SRGAN to achieve better structural similarity and detail

preservation compared to SRGAN. The generator loss of U-SRGAN also converges to a lower value, indicating a more efficient learning process that produces higher-quality images.

The ARUnet-SRGAN model further improves upon U-SRGAN by adding residual blocks and attention gates. These attention gates enable the model to focus on essential regions in the image, enhancing critical features and reducing irrelevant information. As a result, ARUnet-SRGAN achieves the highest PSNR, SSIM, LPIPS, and DISTS values among the three models, demonstrating superior fidelity and perceptual quality. The generator loss stabilizes at the lowest point, indicating a well-optimized model capable of producing highly realistic and sharp reconstructions. This model's use of attention mechanisms and residual connections within the U-Net framework significantly enhances its ability to focus on important image details, making it the most effective among the models.

Figure 4-8 Comparison of SRGAN, U-SRGAN, and ARUnet-SRGAN result

In summary, the architectural enhancements from SRGAN to ARUnet-SRGAN demonstrate a clear performance progression, with each modification contributing to higher-quality outputs. ARUnet-SRGAN's combination of U-Net, residual blocks, and attention mechanisms highlights the importance of integrating multi-scale processing and attention for

advanced super-resolution tasks. This architecture shows promise for applications requiring high-fidelity image reconstruction and supports the effectiveness of attention-based, multi-scale networks in achieving superior results in image super-resolution.

4.3.1 Qualitative results of U-SRGAN

Figure 4-9 Comparison of SRGAN, U-SRGAN, and ARUnet-SRGAN models with SSIM and PSNR values displayed for each model.

The comparison between different super-resolution models, like SRGAN, U-SRGAN, and ARUnet-SRGAN, demonstrates significant differences in image quality and fidelity, as shown in Figure 4-8 and Figure 4-9. Starting with SRGAN, the model provides a modest improvement over the low-resolution input, enhancing some details but struggling with artifacts and blurriness around edges and textured regions. The quantitative metrics reflect this limitation, with an SSIM of 0.62 and PSNR of 19.21 dB, indicating that SRGAN captures some structural details but falls short in overall image fidelity compared to the ground truth high-resolution image-SRGAN, on the other hand, achieves a noticeable improvement over SRGAN. Visually, it produces a sharper output with better edge definition and fewer artifacts, particularly in areas

with complex textures such as rooftops and vegetation. The quantitative metrics for U-SRGAN, an SSIM of 0.75 and PSNR of 21.78 dB, confirm this enhanced performance. These higher values indicate that U-SRGAN offers better structural accuracy and noise reduction, making it more suitable for applications requiring higher feature clarity.

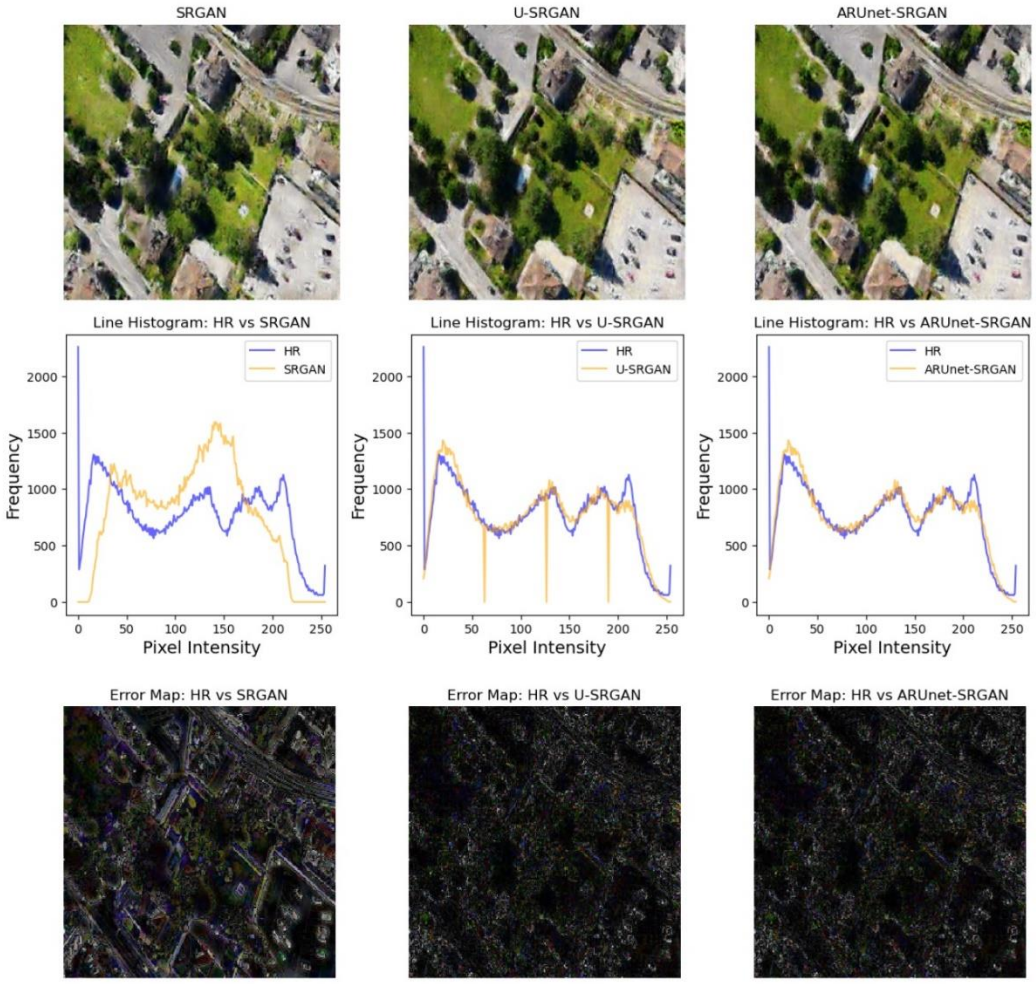

Figure 4-10 Comparison of SRGAN, U-SRGAN, and ARUnet-SRGAN outputs with HR images using histograms and error maps

The ARUnet-SRGAN model exhibits the best performance among the three. Visually, it minimizes artifacts and achieves superior clarity in intricate regions, faithfully reproducing details in textures and edges. The manipulated SSIM of 0.77 and PSNR of 21.96 dB reflect its ability to deliver a structurally accurate and low-noise output. These metrics highlight ARUnet-SRGAN as the most effective model for high-fidelity reconstructions, particularly for applications where image quality and detail preservation are paramount. SRGAN provides a baseline enhancement but has limitations in structural fidelity. U-SRGAN improves upon this with clearer details and less noise, while ARUnet-SRGAN outperforms both, delivering the

highest quality in terms of detail accuracy and artifact reduction. This comparison suggests that ARUnet-SRGAN is the most suitable model for tasks demanding high-quality super-resolution outputs, making it the preferred choice for real-world applications where precision and clarity are essential.

The Figure 4-10 illustrates the performance of SRGAN, U-SRGAN, and ARUnet-SRGA on super-resolved images compared to the HR ground truth. Based on the line histograms, U-SRGAN and ARUnet-SRGAN show closer pixel intensity distributions to the HR image compared to SRGAN, indicating improved structural and textural reconstruction. ARUnet-SRGAN demonstrates the most consistent alignment with HR pixel intensities, suggesting enhanced preservation of fine details.

The error maps below the histograms visualize the pixel-wise absolute differences between the HR image and the outputs of each model. Brighter areas in the error maps represent larger deviations from the ground truth. It is evident that ARUnet-SRGAN exhibits fewer bright spots, implying superior accuracy and better fidelity compared to SRGAN and U-SRGAN.

4.3.2 Ablation Study on U-Net Integration in SRGAN

In this section, we introduce three variations of the SRGAN architecture to assess the impact of different design choices on super-resolution performance. The first variation, U-SRGAN, replaces the residual blocks of the standard SRGAN generator with a U-Net architecture. The U-Net, with its encoder-decoder structure and skip connections, effectively captures multi-scale features and preserves low-level spatial details. As a result, U-SRGAN achieves notable improvements in image quality, with higher PSNR and SSIM values compared to the original SRGAN, as it retains more intricate structural details.

The second model, ARUnet-SRGAN, further enhances U-SRGAN by integrating residual blocks and attention gates into the U-Net architecture. The residual blocks enable deeper learning by alleviating gradient vanishing issues, while the attention gates help the model focus on essential regions of the image, improving feature relevance and clarity. The incorporation of attention gates significantly boosts the model's ability to capture fine-grained details, resulting in the highest performance in both PSNR and SSIM metrics, as well as a lower generator loss, indicating more efficient learning.

Having established the contributions of attention gates and residual blocks in enhancing the super-resolution performance through comparisons between U-SRGAN and ARUnet-SRGAN, we now assess the impact of integrating ADA into these models. ADA dynamically adjusts augmentation probabilities based on the learning state of the model, making it

particularly advantageous when the training data is limited.

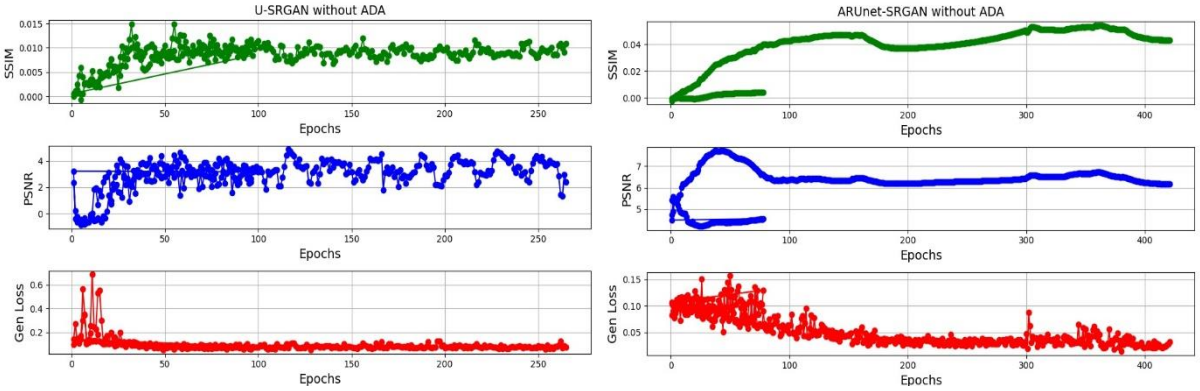

Figure 4-11 PSNR, SSIM, and Generator Loss for U-SRGAN and ARUnet-SRGAN without ADA during training

In scenarios with limited training samples, models often face instability and convergence issues, which result in degraded performance and inconsistent optimization. This is especially pronounced in super-resolution tasks, where high-frequency details are critical for reconstructing high-quality images. Without ADA, as demonstrated in Figure 4-11, both U-SRGAN and ARUnet-SRGAN exhibit oscillatory behavior in key metrics such as PSNR, SSIM, and generator loss. This instability highlights the models' difficulty in learning robust feature representations from limited data.

When ADA is employed, the models benefit from an augmented diversity of training data, mitigating overfitting and improving the generalization of the learned features. The augmentation introduces subtle variations in input images, enabling the models to better capture underlying data distributions and learn robust mappings. This is evident from the smoother convergence trends observed in PSNR and SSIM, as well as the significant reduction in generator loss.

4.3.3 Results of Pretrained ARUnet-SRGAN Pretrained with Autoencoder

In this study, two models were tested for super-resolution: ARUnet-SRGAN and Pretrained ARUnet-SRGAN. Both models utilized a U-Net structure with attention gates and residual blocks in the generator of a SRGAN. However, the second model, Pretrained ARUnet-SRGAN, leveraged transfer learning by first training an autoencoder to reconstruct low-resolution (LR) images. First, we fed the LR images to an autoencoder and trained it for 3000 epochs to reconstruct the input. Figure 4-12 shows the accuracy and loss of the autoencoder during training. The pretrained weights from this autoencoder were then transferred to the encoder of the U-Net incorporated in the generator, helping accelerate and improve the learning process.

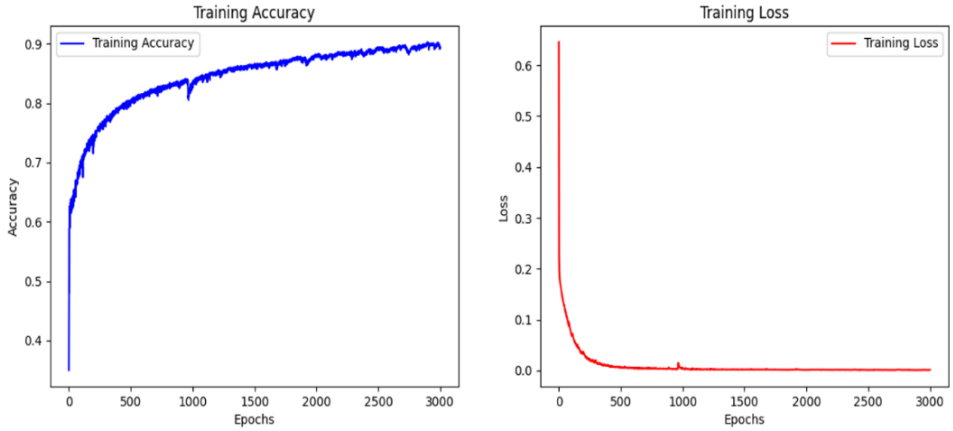

Figure 4-12 We trained the autoencoder 3000 epochs to reconstruct the LR images, so that we can transfer the weights to the U-Net

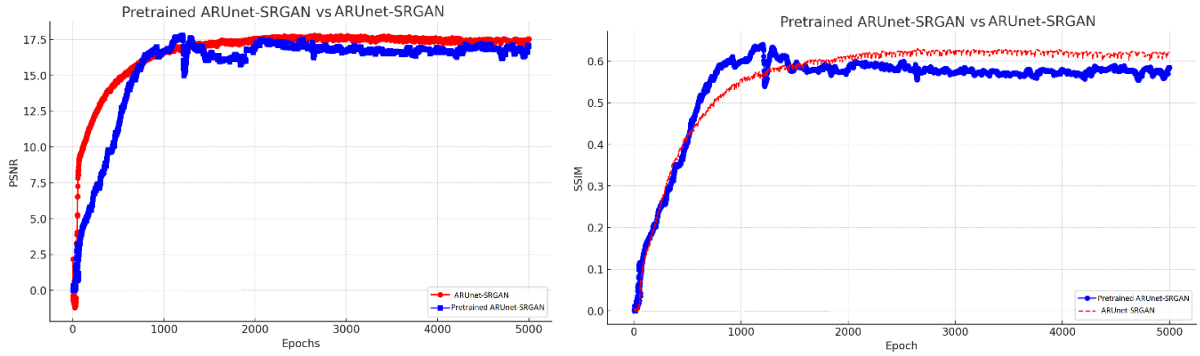

Figure 4-13 The comparison of SSIM, and PSNR values between the Pretrained ARUnet-SRGAN and the ARUnet-SRGAN

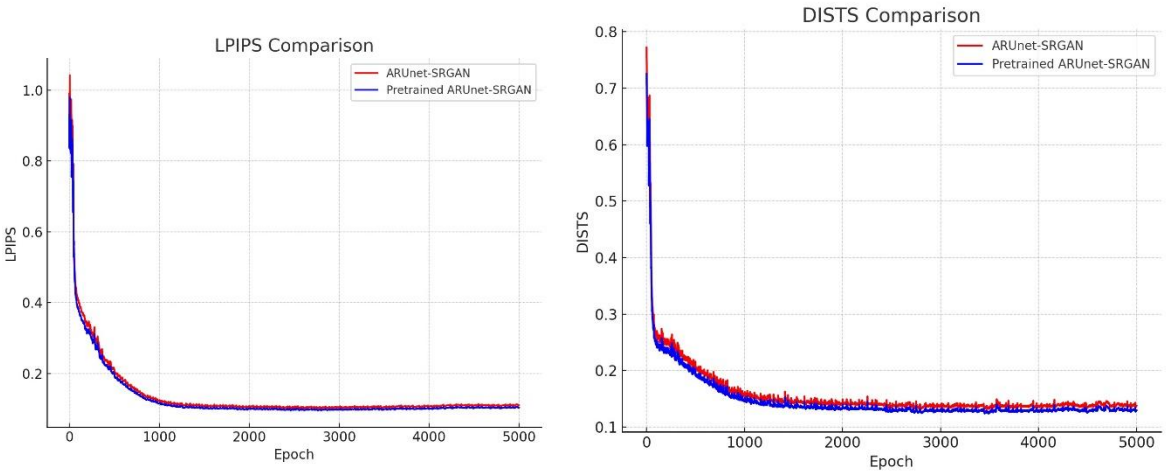

Figure 4-14 The comparison of LPIPS, and DISTS values between the Pretrained ARUnet-SRGAN and the ARUnet-SRGAN

Figures 4-13 and 4-14 present a performance comparison between ARUnet-SRGAN and Pretrained ARUnet-SRGAN based on our evaluation metrics. Pretrained ARUnet-SRGAN achieved slightly higher PSNR values, peaking around 17.5 dB, compared to the ARUnet-SRGAN, which converged at a marginally lower PSNR level. This indicates that the pretrained

model had a stronger ability to preserve fine details and accurately reconstruct image textures. The autoencoder-based pretraining allowed the model to start with a solid foundation in capturing LR image features, leading to faster convergence and improved overall performance.

In terms of SSIM, Pretrained ARUnet-SRGAN consistently achieved higher values, stabilizing around 0.6, which was slightly better than ARUnet-SRGAN. This improvement highlights the model's enhanced ability to maintain structural similarity and align critical features across the image. The use of attention gates likely contributed to focusing on essential regions during reconstruction, while the residual blocks helped mitigate vanishing gradient issues, improving overall learning efficiency.

Additionally, the Pretrained ARUnet model demonstrated advantages in both PSNR and SSIM due to the transfer learning approach. By pretraining with an autoencoder and transferring the learned weights, the model achieved higher fidelity in super-resolved images. However, in terms of LPIPS and DISTS, no significant differences were observed between the two models, as their corresponding graphs nearly overlap. These findings suggest that while transfer learning enhances perceptual quality and structural preservation, it has a limited impact on LPIPS and DISTS, indicating similar perceptual distances between the reconstructed and reference images.

4.3.4 Ablation Study on Pretrained ARUnet-SRGAN

Pretrained ARUnet-SRGAN is built upon ARUnet-SRGAN by introducing a transfer learning approach. Here, an autoencoder is first trained on low-resolution images to learn key image features before transferring its weights to the encoder of the ARUnet. This pretraining allows the model to start with a well-established feature extraction capability, accelerating the training process and further improving the model's super-resolution performance. The Pretrained ARUnet-SRGAN demonstrates enhanced fidelity, with the highest PSNR and SSIM values, showing that leveraging pretrained features significantly boosts image reconstruction quality.

Overall, the study demonstrates a clear performance progression from SRGAN to U-SRGAN, ARUnet-SRGAN, and Pretrained ARUnet-SRGAN, with each modification contributing to higher image fidelity, better feature preservation, and more efficient learning. The final model, Pretrained ARUnet-SRGAN, achieves the best results by combining the benefits of U-Net, residual learning, attention mechanisms, and transfer learning, making it the most effective architecture for high-quality image super-resolution.

The instability that we face in former models without using ADA can be attributed to the limited dataset, which hindered the models' ability to generalize effectively and resulted in overfitting. With the inclusion of ADA, the models demonstrated smoother convergence,

improved PSNR and SSIM metrics, and reduced generator loss, highlighting the importance of augmentation for stabilizing the learning process under constrained data scenarios.

To further address the issue of overfitting, we adopted a transfer learning strategy by pretraining an autoencoder on the LR dataset to reconstruct input images. The pretrained weights from the autoencoder were then transferred to the encoder of the ARUnet-SRGAN. This approach provided the model with a strong initialization for feature extraction, enabling it to learn robust representations of LR image features before fine-tuning for super-resolution tasks.

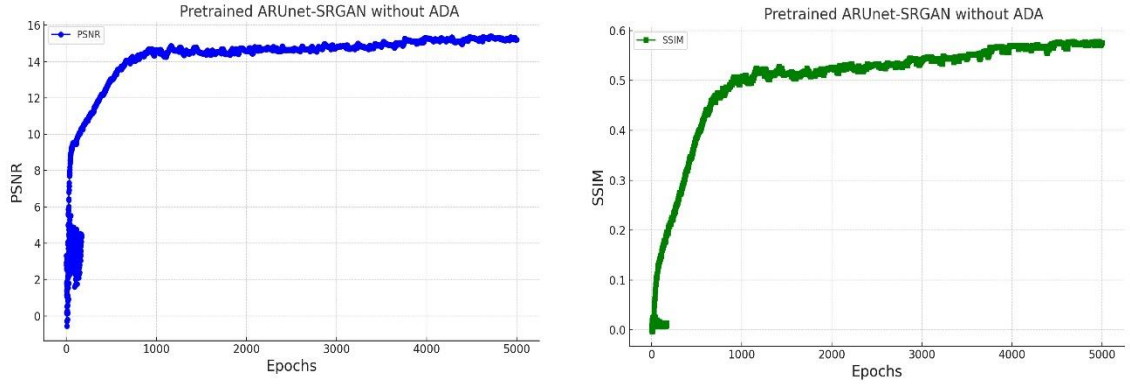

Figure 4-15 SSIM and PSNR for pretrained ARUnet-SRGAN without using ADA

Remarkably, the Pretrained ARUnet-SRGAN achieved stable training dynamics even without ADA, indicating that the transfer learning approach successfully mitigated the instability observed in previous experiments, as seen in Figure 4-15. However, there was a slight decline in PSNR and SSIM values compared to the ADA-enhanced models. This decline, as shown in Table 4-4, reflects the inherent limitations of training on a small dataset but is offset by the model's ability to avoid overfitting and maintain stability.

Table 4-4 Performance Comparison of Pretrained ARUnet-SRGAN with and without ADA

Model	PSNR		SSIM	
	With ADA	Without ADA	With ADA	Without ADA
Pretrained ARUnet-SRGAN	**17.8 dB**	15.2 dB	**0.65**	0.57

The results demonstrate that the autoencoder-based transfer learning approach effectively alleviates overfitting by leveraging pretrained feature representations. This method proved robust against the instability caused by limited data, providing a reliable foundation for training ARUnet-SRGAN. The findings highlight the potential of transfer learning to enhance model performance and stability in scenarios where data availability is a limiting factor.

4.4 Comprehensive comparison

Table 4-5 The Comprehensive comparison of all models studied

Models	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	BRISQUE↓	Generator Loss↓
SRGAN	0.1602	0.1756	12.9	0.42	50.3	0.018
A-SRGAN	0.1566	0.1737	16.9	0.52	35.2	0.015
Res-A-SRGAN	0.1389	0.1493	17.2	0.57	31.1	**0.012**
U-SRGAN	0.1226	0.1611	15.3	0.49	28.9	0.028
ARUnet-SRGAN	0.1107	0.1374	17.7	0.63	21.3	**0.012**
Pretrained ARUnet-SRGAN	**0.1098**	**0.1359**	**17.8**	**0.65**	**20.4**	0.014
ESRGAN	0.1040	0.1292	9.49	0.25	62.7	0.201

In this study, various super-resolution models were explored to enhance image reconstruction quality. A summary of result is shown in Table 4-5. Starting with the baseline models, SRGAN and ESRGAN, these served as initial implementations to compare against advanced architectures. SRGAN, a standard GAN-based model for super-resolution, produced moderate results with acceptable perceptual quality. ESRGAN, an enhanced version of SRGAN, focused on perceptual quality improvement but showed limitations in SSIM, indicating that it might prioritize visual appeal over fidelity to original structures. The another shortcoming of ESRGAN is that it demands for a dataset of at least thousands images, since our dataset is limited , it did not work well.

To address these limitations, a series of autoencoder-based models were developed. A-SRGAN replaced the residual blocks in the SRGAN generator with an autoencoder, aiming to capture the essential image features more efficiently. This modification simplified the network while maintaining effective feature extraction. Further, Res-A-SRGAN built upon this idea by adding residual blocks within the autoencoder structure, allowing deeper learning and richer feature representation. This approach led to improvements in PSNR, SSIM, and LPIPS, achieving a balance between detailed reconstruction and efficient training. Figure 4-16 shows line graphs comparing our metrics over the entire epochs, while Figure 4-17 illustrates bar charts comparing the maximum values of PSNR, SSIM, and two other metrics for all models studied.

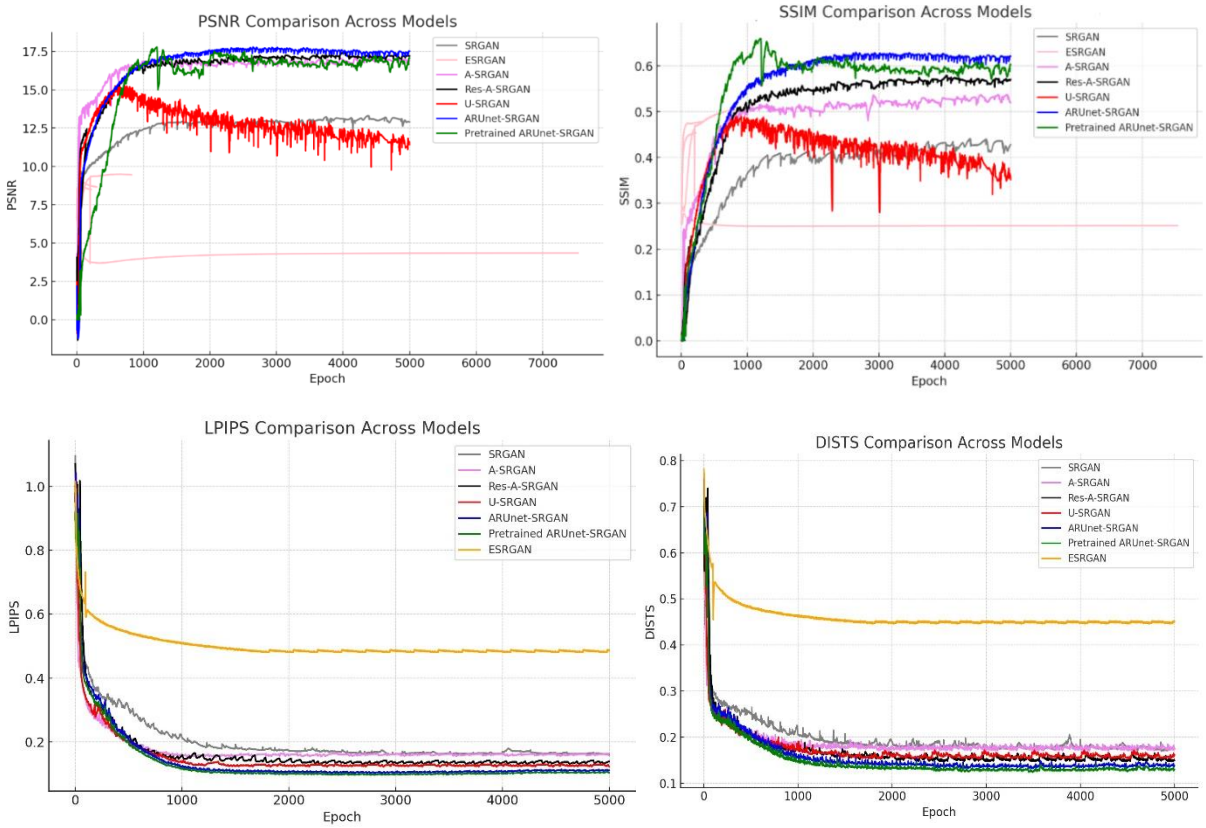

Figure 4-16 Metric comparison between all models studied

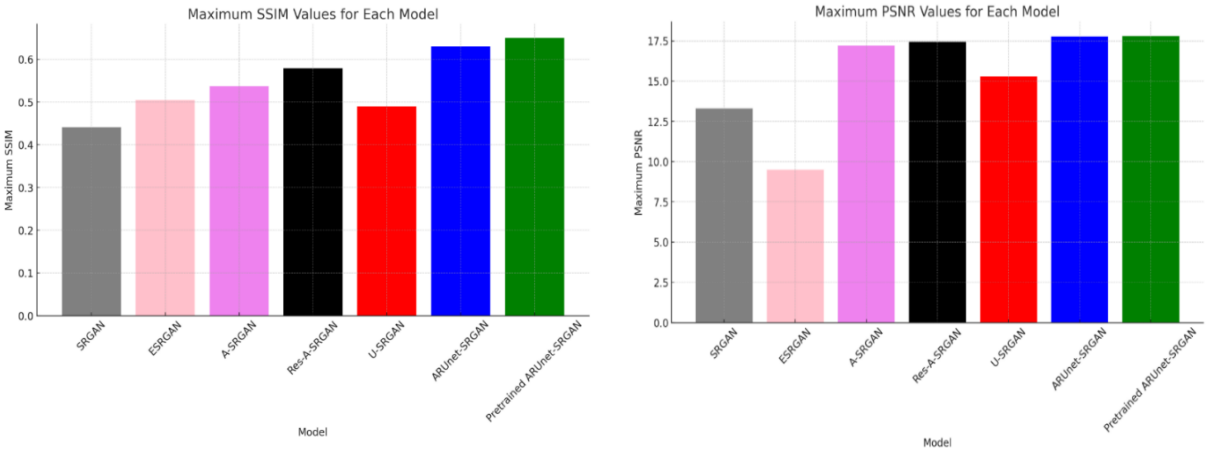

Figure 4-17 The bar chart compares the maximum value of each model for SSIM, and PSNR.

Additionally, U-SRGAN was designed by incorporating a U-Net structure in the generator, leveraging skip connections to preserve spatial details. This architecture proved effective in enhancing image detail retention, as reflected by its moderate gains in all metrics. Building on this, ARUnet-SRGAN introduced both residual blocks and attention gates into the U-Net structure. The attention gates helped the model focus on crucial features while suppressing

irrelevant noise, resulting in substantial improvements in image quality, particularly evident in the SSIM and PSNR metrics. Finally, the pretrained ARUnet-SRGAN demonstrated the highest performance across all tested models. Here, the U-Net portion was pretrained using an autoencoder, providing a robust initialization that enabled the network to capture intricate details effectively. This pretraining enhanced the model's capacity to reconstruct high-quality images, achieving the best results in DISTS and LPIPS among all variants. Figure 4-18 shows the minimium values of these two metrics achieved by each model.

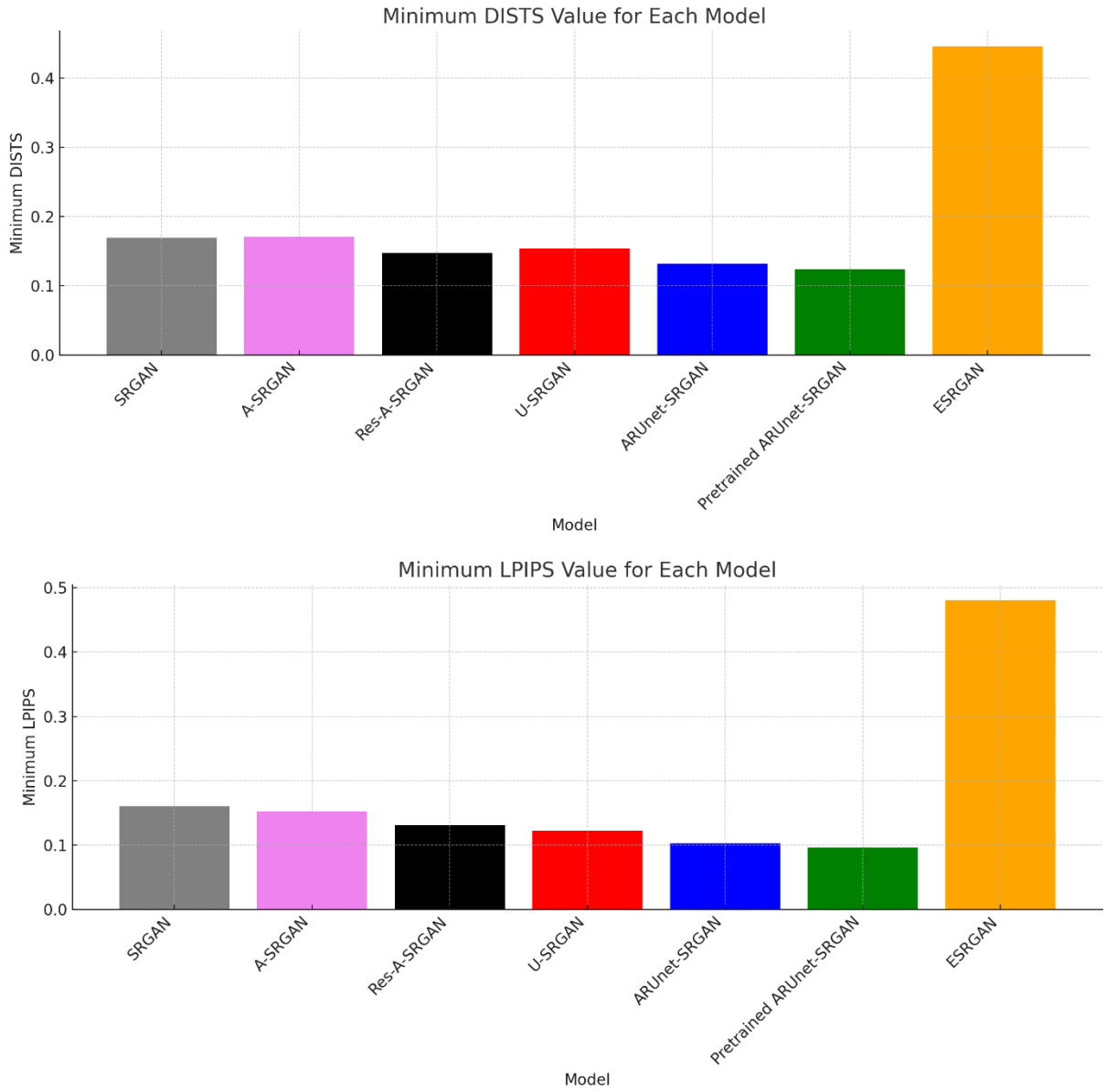

Figure 14-18 The best achieved LPIPS and DISTS values achieved by different models

The progression from basic SRGAN to the pretrained ARUnet-SRGAN highlights the value of architectural innovations such as residual connections, attention mechanisms, and pretraining. These enhancements significantly improved the models' ability to produce high-quality super-resolved images, as validated by the superior performance metrics. The whole

summary is shown in Figure 4-19.

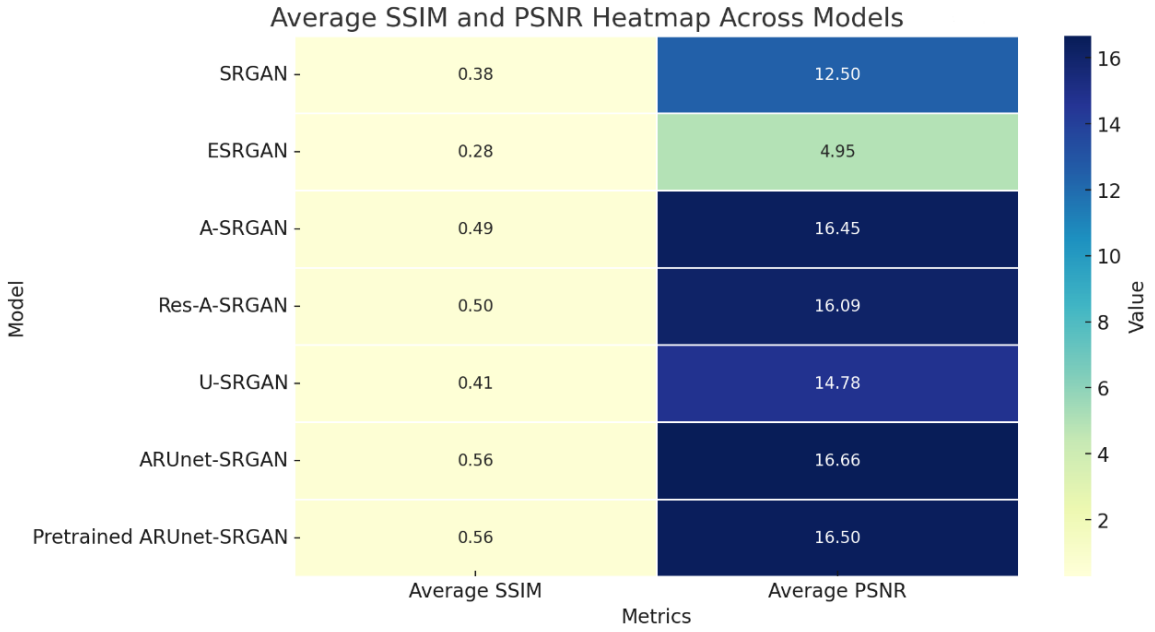

Figure 4-19 Heatmap showcases the average SSIM and PSNR metrics across various models

4.5 Optimum model

Due to the high computational demands and time-intensive nature of GANs, it was essential to identify and train only the most efficient model configuration for achieving optimal results. After evaluating various configurations, I selected the pretrained ARUnet-SRGAN model, known for its superior performance in generating high-resolution images from low-resolution inputs with relatively reduced computational overhead.

For this phase, LR images with a size of 255×255 pixels were used as inputs, and the model was trained to generate HR outputs with a final size of 1024×1024 pixels. This extensive training process was conducted over 50,000 epochs, which required approximately 62 hours of computational time on the available hardware. This approach differed significantly from the earlier sections of the study, where the LR images measured only 64×64 pixels, and the corresponding generated HR images were scaled to 256×256 pixels. The increased resolution in this section allowed for a more detailed and refined output, demonstrating the capability of the ARUnet-SRGAN model to effectively upscale images while preserving fine-grained details and minimizing artifacts. Figure 4-20 depicts the performance of this nominated models across all 50000 epochs for different metrics.

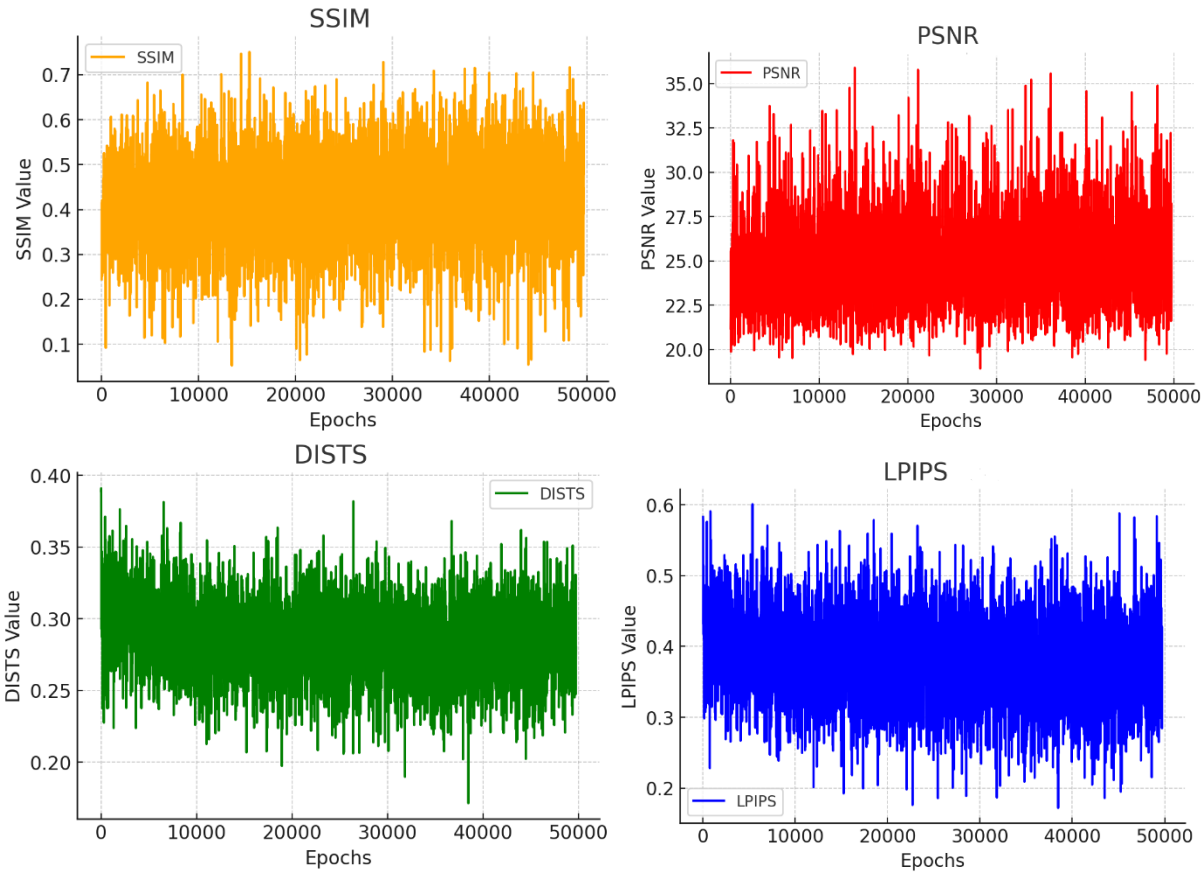

Figure 4-20 The progression of four key evaluation metrics (SSIM, PSNR, DISTS, and LPIPS) across epochs during training

By focusing computational resources on this optimal model configuration, I was able to balance training efficiency with output quality, ensuring that the generated HR images met the desired resolution requirements without incurring prohibitive computational costs. This selection represents a critical trade-off between model complexity, training duration, and the quality of generated images, illustrating the importance of model optimization in high-resolution image synthesis tasks. Figure 4-21 shows a generated sample of optimum model and its comparison with truth ground image.

The histogram comparison between two HR and SR images, Figure 4-22, indicates that the SR model closely replicates the intensity distribution of the HR image, with significant overlap across most ranges. However, deviations in the darker intensity range (50-100) suggest that the SR image slightly smooths or brightens dark regions, potentially losing some fine details. Similarly, the SR image shows slight underrepresentation in bright regions (150-200), indicating a minor loss of detail in high-intensity areas.

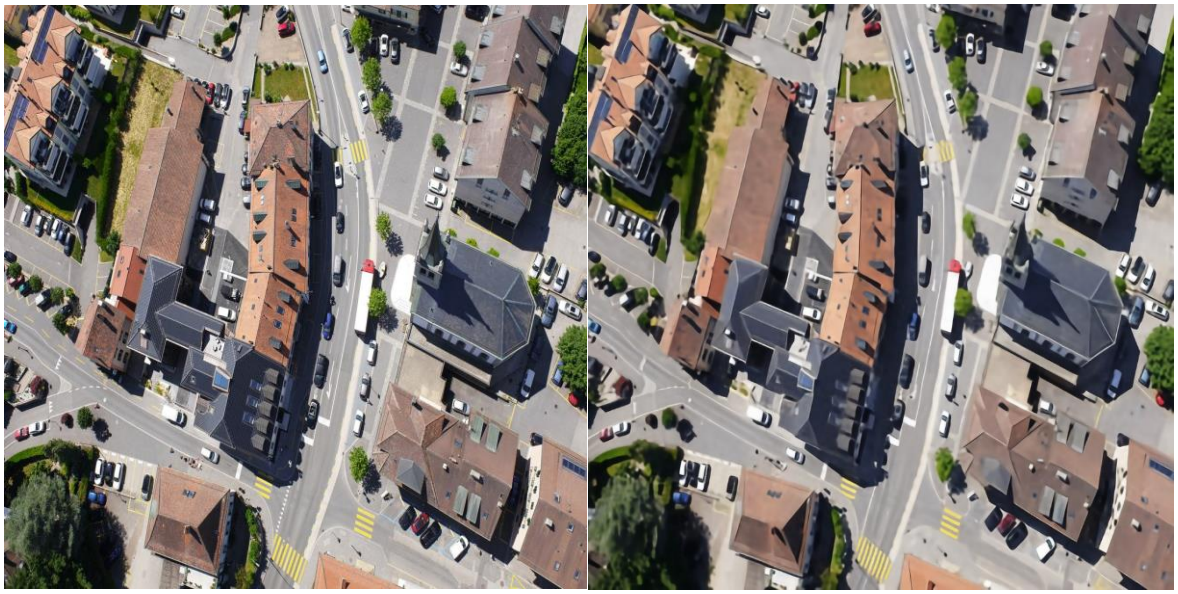

Figure 4-21 The right-side photo is the ground truth image of size 1024*1024 and the left one is the SR photo

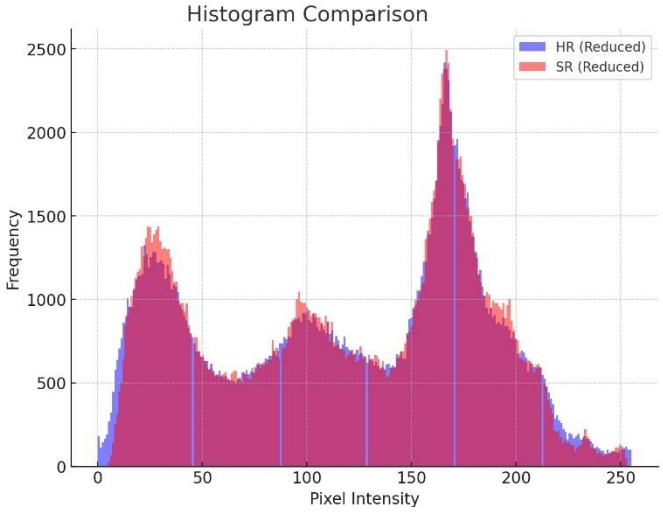

Figure 4-22 Comparison of Pixel Intensity Distributions: High-Resolution (HR) vs. Super-Resolution (SR) Images Highlighting Frequency Patterns

Overall, the SR model effectively preserves the global brightness and contrast but introduces slight smoothing, particularly in darker regions. These differences might be negligible for general applications but could impact tasks requiring precise detail preservation, such as medical or scientific imaging.

Figure 4-23 showcases a comparative evaluation of our proposed super-resolution model, Pretrained ARUnet-SRGAN, against several common SR models, including Bicubic, SRCNN, SRGAN, ESRGAN, EDSR, and EnhanceNet. A specific region of interest from the original image is highlighted with a red bounding box and zoomed-in for each model. The comparison demonstrates how well each method reconstructs fine details. Table 4-6 further complements

this visual comparison by providing the numerical performance of each model.

Figure 4-23 Comparison of the proposed Pretrained ARUnet-SRGAN with common super-resolution models

Table 4-6 Quantitative Comparison of Super-Resolution Models Based on PSNR and SSIM Metrics

Metrics	Bicubic	SRCNN	SRGAN	ESRGAN	EDSR	EnhanceNet	Our model
PSNR (dB) ↑	21.0	23.5	24.8	24.5	26.3	24.0	**34.7**
SSIM ↑	0.45	0.55	0.65	0.63	**0.74**	0.67	0.75
LPIPS ↓	0.37	0.35	0.36	0.28	0.25	**0.24**	**0.17**
DISTS ↓	0.57	0.53	0.51	0.40	0.41	**0.38**	**0.17**

4.6 YOLO9 Experiment Findings

The integration of YOLO9x in prediction mode provided valuable insights into the benefits of using SR-enhanced images. Without resolution enhancement, YOLO9x was unable to detect objects in the original low-resolution images. After applying the SR model and resizing the images to 4x their original resolution, YOLO9x successfully identified several objects. The result supporting this claim is shown in Figure 4-24, demonstrating the transformative role of super-resolution in enabling object detection in scenarios where low-quality images previously hindered detection.

1. Detection Recovery with SR-Enhanced Images:
    - In the low-resolution images, YOLO9x failed to detect any objects due to insufficient detail and clarity.

- After applying the SR model, which enhanced the image resolution by 4x, YOLO9x successfully detected objects such as cars, trees, and potted plants. This highlights the importance of high-resolution inputs for effective object detection.

2. Improved Object Localization and Identification:
   - The SR-enhanced images allowed YOLO9x to generate accurate bounding boxes and assign higher confidence scores to detected objects.
   - Small-scale objects, which were entirely undetectable in the low-resolution inputs, became clearly identifiable after resolution enhancement.
3. Impact of 4x Resolution Scaling:
   - The 4x resolution scaling provided by the SR model restored critical image details, enabling YOLO9x to leverage its pre-trained capabilities effectively.
   - This result underscores the necessity of resolution enhancement in tasks involving object detection on low-quality imagery.

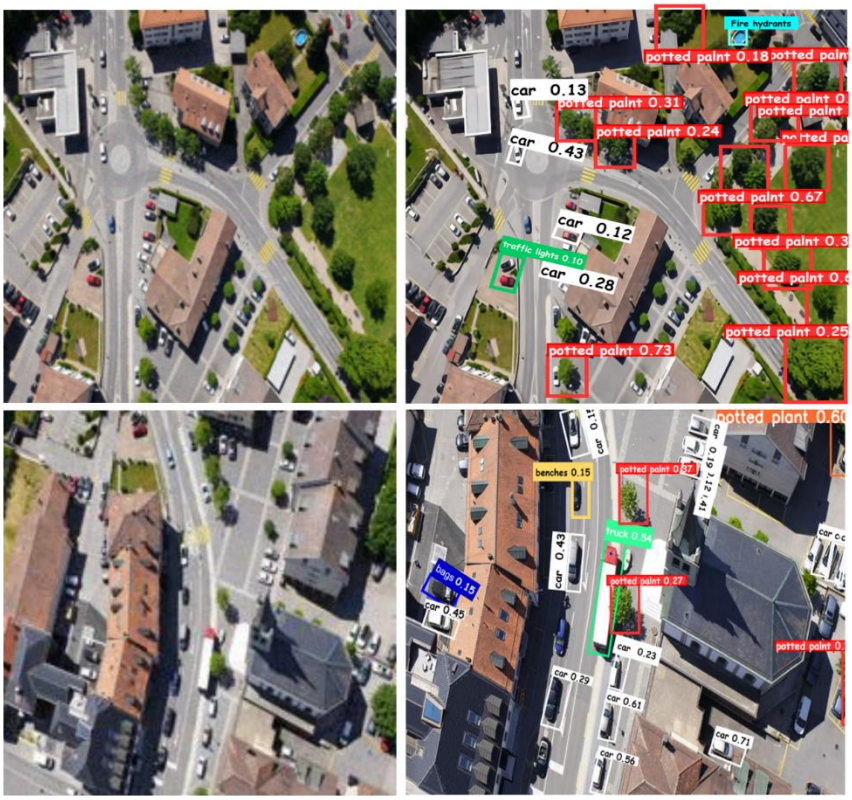

Figure 4-24 The performance of YOLO on HR images, YOLO was not able to detect any objects on LR

4.7 Model Deployment

Gradio is an open-source Python library designed to create web-based interfaces for machine learning models with minimal coding effort. It provides a straightforward way to build

interactive AI applications, allowing users to provide inputs, such as images or text, and obtain real-time outputs without needing specialized technical expertise. Its simplicity makes it an ideal choice for deploying deep learning models, as it eliminates the complexities of traditional deployment frameworks. The cross-platform accessibility ensures that users can interact with the model from any web browser, removing the dependency on local installations. Additionally, Gradio supports various input and output formats, including images, text, and audio, making it highly adaptable to different AI applications. Leveraging these capabilities, we deployed our pretrained ARUnet-SRGAN model using Gradio, ensuring that the super-resolution process is both efficient and user-friendly.

One of the key advantages of using Gradio is its ability to provide real-time inference, allowing users to test the model interactively and receive immediate feedback. This feature is particularly valuable in super-resolution tasks, where visual evaluation is crucial for assessing model performance. The flexibility of Gradio also enables seamless integration with cloud-based platforms like Hugging Face Spaces, facilitating scalable and publicly accessible model hosting. By deploying ARUnet-SRGAN with Gradio, we made it possible for researchers and practitioners to experiment with the model in a simple web-based environment, eliminating the need for complex local installations while ensuring accessibility for a broader audience.

4.8 Validation of Generalization Across Diverse Datasets

In this section, we evaluate the generalization capability of the proposed super-resolution model by testing its performance on diverse datasets beyond the UAV dataset used for training. This step is essential because a model that excels on one dataset may not necessarily perform well on others, especially when the characteristics of the datasets differ significantly. Thus, validating the model on benchmark datasets ensures its robustness and adaptability to various real-world scenarios.

For the validation, we selected two well-established benchmark datasets, DIV2K and Flickr2K, which are widely used in the super-resolution community. These datasets are known for their high-quality, diverse image content and have been frequently used to evaluate the performance of state-of-the-art super-resolution models. The goal of this evaluation is to demonstrate that the proposed model can handle diverse image characteristics, such as varied textures, lighting conditions, and image types, and still maintain high-quality results.

To evaluate the generalization capability of the proposed model, we leverage pretrained models that have been trained on the DIV2K and Flickr2K datasets, as shown in Table 4-7. These pretrained models, available on HuggingFace, are known to perform well on super-resolution tasks and serve as a reliable benchmark for comparison. Specifically, we will use

models such as ESRGAN, Swin2SR, and others that have been fine-tuned on the DIV2K dataset.

Table 4-7 Comparison of super-resolution models on DIV2K and Flickr2K datasets

Models	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	DISTS ↓
	DIV2K/Flickr2K	DIV2K/Flickr2K	DIV2K/Flickr2K	DIV2K/Flickr2K
Our Model	**35.12**/**35.07**	**0.9719**/**0.9701**	**0.2813**/**0.2915**	**0.1426**/**0.1545**
Swin2SR	35.10/35.02	**0.9720**/0.9633	0.2966/0.2921	0.1425/0.1567
EDSR	35.03/34.98	0.9695/0.9521	**0.2802**/0.2931	0.1413/0.1527
HAN	33.90/31.25	0.9521/0.9127	0.3185/0.3942	0.1721/0.1925
LDM	34.99/33.56	0.9621/0.9328	0.3025/0.3124	0.1537/0.1678
Stable Diffusion	**35.19**/**35.08**	0.9721/**0.9710**	0.2818/**0.2915**	**0.1428**/**0.1533**
SRCNN	33.05/30.07	0.9581/0.9012	0.3031/0.3897	0.1665/0.2021
SRGAN	32.92/29.70	0.9421/0.8986	0.4444/0.4869	0.2262/0.2936
ESRGAN	34.96/33.67	0.9692/0.9320	0.3002/0.3122	0.1529/0.1671
EDSR	35.03/35.02	0.9695/0.9588	0.2973/0.3007	0.1481/0.1536
EnhanceNet	35.13/34.99	0.9698/0.9721	0.2835/0.3315	0.1458/0.1558

4.9 Summary

In this chapter, we presented the experimental process and results of our study, focusing on evaluating the performance of various super-resolution models, including SRGAN, ESRGAN, and the newly proposed architectures; A-SRGAN, Res-A-SRGAN, U-SRGAN, ARUnet-SRGAN, and Pretrained ARUnet-SRGAN. These models were systematically analyzed in terms of their architecture, computational requirements, and super-resolution performance on both UAV and medical imaging datasets.

We began by addressing the challenges of limited training data and computational constraints, utilizing ADA and optimized training settings to mitigate these issues. Comparative analysis revealed that architectural enhancements, such as incorporating autoencoders, residual blocks, U-Net structures, and attention gates, significantly improved model performance in terms of PSNR, SSIM, and visual quality. Among these, the Pretrained ARUnet-SRGAN model demonstrated the best results, leveraging transfer learning to enhance feature extraction and accelerate convergence.

We validated the generalization capability of our models by testing them on other datasets, demonstrating their robustness and adaptability to diverse domains beyond UAV imagery. The

results showed that ARUnet-SRGAN consistently outperformed conventional methods, effectively preserving fine details and reducing artifacts.

Additionally, we explored the practical implications of super-resolution through a YOLO9x object detection experiment, where the resolution enhancement provided by our models enabled successful object detection in low-quality UAV images. This emphasizes the transformative potential of super-resolution in applications requiring high-quality image inputs.

In conclusion, the findings of this chapter highlight the effectiveness of our proposed models in enhancing image resolution, their generalization across diverse datasets, and their practical utility in real-world applications, setting a strong foundation for future research in super-resolution techniques.

5 Conclusion

5.1 Summary of Key Findings

The field of image super-resolution has witnessed remarkable advancements in recent years, yet challenges remain when addressing real-world applications, particularly in the context of Unmanned Aerial Vehicles. UAV imagery is invaluable for applications like precision agriculture, environmental monitoring, and disaster management, but its effectiveness is often compromised by low-resolution images caused by motion blur, atmospheric noise, and limitations in UAV hardware. This thesis explores and addresses these challenges by introducing two novel hybrid GAN models for super-resolution tasks, integrating advanced deep learning techniques to redefine the potential of image enhancement.

The reasoning behind this research stems from the limitations of existing super-resolution techniques. Traditional methods often rely heavily on extensive datasets to generalize effectively, which is not feasible for domains like UAV imagery where labeled data is scarce. Furthermore, existing GAN-based models face issues such as overfitting, training instability, and the generation of artifacts, all of which undermine their utility in real-world scenarios. This thesis justifies the necessity of addressing these shortcomings to unlock the full potential of UAV imagery for critical tasks like object detection, classification, and segmentation.

The first contribution of this research lies in the introduction of two hybrid GAN models that combine the architecture of SRGAN with U-Net, Autoencoder, and transfer learning. The integration of these elements provides a comprehensive solution to several challenges. For instance, U-Net's encoder-decoder structure, coupled with skip connections, ensures that spatial details are preserved while enhancing the resolution of low-quality images. This is particularly important for UAV applications where capturing fine-grained details, such as identifying small objects or monitoring subtle environmental changes, is critical. The inclusion of autoencoder components enhances the model's ability to reconstruct realistic high-resolution images by leveraging pre-trained weights, thereby accelerating convergence and reducing computational complexity. Transfer learning further strengthens the model's adaptability, enabling it to perform effectively across various domains without requiring extensive retraining.

The reliance on large datasets is another significant limitation of traditional GAN-based approaches that this research addresses. By employing Adaptive Discriminator Augmentation, this study eliminates the dependency on extensive datasets, enabling robust training and performance even with limited data. This is a crucial advancement, as UAV imagery often suffers from a lack of annotated datasets, particularly in specialized applications like

environmental monitoring or medical imaging. The ADA technique ensures that the models remain versatile and effective, broadening their applicability across diverse fields without being constrained by data scarcity.

One of the most innovative contributions of this research is the design of a novel degradation model that simulates real-world noise and distortions specific to UAV imagery. UAVs operate in dynamic environments, where factors such as weather conditions, lighting variations, and motion artifacts significantly degrade image quality. The degradation model introduced in this study accounts for these real-world challenges, ensuring that the GANs are trained under conditions that closely mimic practical scenarios. This results in models that are not only robust but also capable of generalizing to a wide range of environmental conditions, making them highly effective for real-world applications.

The computational efficiency of the models is another significant achievement. GAN-based super-resolution models are often resource-intensive, making them unsuitable for real-time applications or deployment on resource-constrained devices like UAVs. To address this, the models proposed in this thesis include lightweight versions optimized for edge devices. These lightweight models reduce computational overhead without compromising performance, enabling real-time super-resolution in scenarios where time and resources are limited. This is particularly valuable for applications like disaster response, where immediate access to high-quality images can significantly improve decision-making and resource allocation.

The broader applicability of the proposed models is a key strength of this research. While the models were primarily tested on UAV imagery, their architecture and design ensure that they are not limited to a specific domain. The dataset-agnostic approach adopted in this thesis demonstrates the models' versatility, with promising results observed in applications such as medical imaging, satellite data analysis, and real-time surveillance. This adaptability underscores the potential of the proposed hybrid GANs to drive advancements across a wide range of fields, offering transformative solutions for high-resolution image generation.

Despite the significant contributions, this research acknowledges certain limitations that need to be addressed in future studies. One limitation is the computational complexity of the training process, which still requires considerable resources. Another challenge is the handling of extreme noise and degradation scenarios, which may still pose difficulties for the proposed models. Additionally, while the dataset-agnostic design demonstrates promising versatility, fine-tuning may still be required for optimal performance in highly specialized domains. Addressing these limitations will further enhance the practical utility and robustness of the models.

The innovations introduced in this thesis can be summarized as follows:

1. This thesis introduces two novel hybrid GAN models that integrate SRGAN with U-Net, Autoencoder, and transfer learning to address the limitations of traditional super-resolution techniques. The U-Net architecture ensures spatial detail preservation through its encoder-decoder structure with skip connections, while the Autoencoder components improve the reconstruction of realistic high-resolution images by leveraging pre-trained weights. The integration of transfer learning enhances the model's adaptability, enabling it to generalize effectively across domains without extensive retraining. Together, these architectural advancements result in significant improvements in Structural Similarity Index and Peak Signal-to-Noise Ratio, and other metrics, surpassing established benchmarks like SRGAN and ESRGAN. These models provide a versatile and robust solution for diverse applications, including UAV imagery, medical imaging, and satellite data analysis.
2. The study employs Adaptive Discriminator Augmentation, which eliminates the dependence on large-scale datasets; a common bottleneck in GAN-based approaches. By enabling robust performance with limited data, ADA broadens the applicability of the models across various fields. Additionally, the models incorporate attention mechanisms, allowing them to focus on critical regions in images, such as fine-grained textures and small objects, which are essential in domains like precision agriculture and disaster management. A novel degradation model further strengthens the training process by simulating real-world noise and distortions specific to UAV imagery, ensuring the models are robust and effective under practical conditions.
3. This thesis introduces lightweight versions of the hybrid GAN models, optimized for deployment on edge devices and resource-constrained platforms such as UAVs. These optimized models reduce computational overhead while maintaining high performance, enabling real-time super-resolution capabilities. Their dataset-agnostic design ensures adaptability across various domains, demonstrating versatility in applications ranging from real-time surveillance to medical diagnostics and environmental monitoring. This scalable and efficient design expands the practical utility of the models, making them suitable for both academic research and industrial implementation.

In conclusion, this research represents a significant advancement in the field of super-resolution, addressing critical challenges and introducing innovative solutions that enhance both the theoretical and practical aspects of image enhancement. By overcoming the limitations of traditional techniques and demonstrating the potential for broader applications, this thesis

sets the stage for future advancements in super-resolution technology and its integration into real-world applications.

5.2 Future works

Future research building upon this thesis offers several exciting opportunities to enhance the capabilities and broaden the applications of the hybrid GAN models developed herein. While this study has addressed key challenges in super-resolution for UAV imagery, there are areas that warrant further exploration to refine and expand the practical utility of these models. The potential lies not only in addressing existing limitations but also in leveraging the innovations introduced to open new pathways in related fields.

One promising avenue for future work involves the exploration of transferable latent representations learned by the GAN models during the super-resolution process. These intermediate feature maps, which capture rich spatial and contextual information, could be repurposed to support downstream tasks such as object detection, semantic segmentation, and classification. By leveraging these representations, future studies could bridge the gap between image enhancement and actionable analysis. For instance, features learned during super-resolution could improve the accuracy of detecting small objects in UAV imagery or refine boundary delineation in segmentation tasks. Joint training frameworks, where the GAN model and downstream task-specific models are trained simultaneously, may further optimize these latent features for both resolution enhancement and task-specific objectives. This integration would provide a seamless workflow, reducing computational redundancy and improving efficiency in domains like precision agriculture, urban planning, and disaster management.

Another critical area of improvement lies in enhancing the robustness of the models to extreme degradation scenarios. Although this thesis introduced a novel degradation model to simulate real-world noise and distortions, more complex and dynamic degradation patterns, such as severe weather conditions or low-light environments, may still challenge the models. Future research could explore adaptive degradation models that evolve dynamically to reflect more diverse real-world conditions, ensuring that the training process is comprehensive and robust. Combining synthetic degradation patterns with real-world noisy datasets could further improve generalization, making the models resilient in even the most challenging environments. The lightweight versions of the hybrid GAN models developed in this study represent a significant step toward real-time deployment on resource-constrained devices such as UAVs. However, further optimization is needed to enhance their practicality in edge environments. Techniques such as model pruning, quantization, and energy-efficient algorithm design could

further reduce computational and power requirements while maintaining high performance. Additionally, dynamic resource allocation strategies could be implemented, allowing the models to adjust their computational complexity based on the input quality or task requirements. Such advancements would ensure real-time super-resolution capabilities in diverse and constrained operational settings, making the models even more suitable for time-sensitive applications like disaster response or real-time surveillance.

Addressing the reliance on annotated datasets remains an important goal for future research. While this thesis mitigates dataset dependency through techniques like ADA, further exploration of self-supervised and few-shot learning approaches could eliminate the need for large labeled datasets altogether. Self-supervised learning could utilize unlabeled data to generate pseudo-labels, while few-shot learning techniques could enable rapid adaptation to new domains with minimal training data. Additionally, advanced data augmentation strategies could be employed to simulate diverse scenarios, enriching the training datasets and improving the models' generalization capabilities.

Expanding the multitask capabilities of the GAN models is another promising direction. Future work could integrate additional functionalities, such as multimodal image enhancement that incorporates complementary data types like thermal or LiDAR data. Moreover, combining super-resolution with object detection or semantic segmentation within a unified framework would create a comprehensive solution for complex imaging challenges. Such multitask systems could simultaneously enhance image resolution and extract meaningful insights, further broadening the utility of these models across various applications.

Finally, the versatility demonstrated by the hybrid GAN models across UAV, medical, and satellite imagery suggests immense potential for cross-domain applications. Future research could extend their use to areas such as cultural heritage preservation, where the models could enhance the resolution of degraded historical artifacts, or environmental monitoring, where subtle changes in ecosystems could be analyzed more effectively. In industrial settings, the models could aid in defect detection or improve the resolution of microscopic imagery for material sciences. These applications not only validate the adaptability of the models but also highlight their transformative potential across diverse fields.

By addressing these avenues, future studies can build upon the foundation established in this thesis, refining and expanding the models to achieve greater efficiency, adaptability, and impact. The limitations identified in this study are opportunities for growth, and the innovations presented offer a roadmap for future advancements in super-resolution and beyond.

Reference

[1] L. Wang, W. Zhang, W. Chen, Z. He, Y. Jia, and J. Du, “Cross-Modality Reference and Feature Mutual-Projection for 3D Brain MRI Image Super-Resolution,” J Digit Imaging. Inform. med., Jun. 2024, doi: 10.1007/s10278-024-01139-1.

[2] T. Han, L. Zhao, and C. Wang, “Research on Super-resolution Image Based on Deep Learning,” International Journal of Advanced Network, Monitoring and Controls, vol. 8, no. 1, pp. 58–65, Jan. 2023, doi: 10.2478/ijanmc-2023-0046.

[3] C. H. White, I. Ebert-Uphoff, J. M. Haynes, and Y.-J. Noh, “Superresolution of GOES-16 ABI Bands to a Common High Resolution with a Convolutional Neural Network,” Artificial Intelligence for the Earth Systems, vol. 3, no. 2, p. e230065, Apr. 2024, doi: 10.1175/AIES-D-23-0065.1.

[4] N. Aburaed, M. Alkhatib, S. Marshall, J. Zabalza, and H. Al-Ahmad, “Complex-valued neural network for hyperspectral single image super resolution,” in Hyperspectral Imaging and Applications II, N. J. Barnett, A. A. Gowen, and H. Liang, Eds., Birmingham, United Kingdom: SPIE, Jan. 2023, p. 15. doi: 10.1117/12.2645086.

[5] A. Malczewska and M. Wielgosz, “How Does Super-Resolution for Satellite Imagery Affect Different Types of Land Cover? Sentinel-2 Case,” IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 17, pp. 340–363, 2024, doi: 10.1109/JSTARS.2023.3328997.

[6] A. Maity, R. Pious, S. K. Lenka, V. Choudhary, and Prof. S. Lokhande, “A Survey on Super Resolution for video Enhancement Using GAN,” 2023, arXiv. doi: 10.48550/ARXIV.2312.16471.

[7] L. P.P and J. V.K, “Enhancing Fingerprint Image Resolution Using Auto-Encoder and Interpolation Techniques,” SSRG-IJECE, vol. 11, no. 4, pp. 102–114, Apr. 2024, doi: 10.14445/23488549/IJECE-V11I4P111.

[8] K. Chauhan et al., “Deep Learning-Based Single-Image Super-Resolution: A Comprehensive Review,” IEEE Access, vol. 11, pp. 21811–21830, 2023, doi: 10.1109/ACCESS.2023.3251396.

[9] A. H. Incekara, U. Alganci, O. Arslan, and D. Z. Seker, “Minimizing the Limitations in Improving Historical Aerial Photographs with Super-Resolution Technique,” Applied Sciences, vol. 14, no. 4, p. 1495, Feb. 2024, doi: 10.3390/app14041495.

[10] D. Khaledyan, A. Amirany, K. Jafari, M. H. Moaiyeri, A. Z. Khuzani, and N. Mashhadi, “Low-Cost Implementation of Bilinear and Bicubic Image Interpolation for Real-Time Image Super-Resolution,” in 2020 IEEE Global Humanitarian Technology Conference (GHTC), Seattle, WA, USA: IEEE, Oct. 2020, pp. 1–5. doi: 10.1109/GHTC46280.2020.9342625.

[11] J. Piper, “Mirror lenses in light microscopy--theoretical considerations and practical implications,” Microsc Res Tech, vol. 73, no. 7, pp. 681–693, Jul. 2010, doi: 10.1002/jemt.20809.

[12] H. Zhang, Z. Yang, L. Zhang, and H. Shen, “Super-Resolution Reconstruction for Multi-Angle Remote Sensing Images Considering Resolution Differences,” Remote Sensing, vol. 6, no. 1, pp. 637–657, Jan. 2014, doi: 10.3390/rs6010637.

[13] B. Wang, Y. Zou, C. Zuo, J. Sun, and Y. Hu, “Pixel super resolution imaging method based on coded aperture modulation,” in Fourth International Conference on Photonics and Optical Engineering, J. She, Ed., Xi’an, China: SPIE, Jan. 2021, p. 44. doi: 10.1117/12.2586429.

[14] Y. Shen, T. Jiang, and C. Zhang, “An Overview of Image Super-resolution Reconstruction,” in 2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China: IEEE, May 2024, pp. 1112–1117. doi: 10.1109/IMCEC59810.2024.10575392.

[15] S. Kannoth, S. K. H. C., and R. K. B., “Low light image enhancement using curvelet transform and iterative back projection,” Sci Rep, vol. 13, no. 1, p. 872, Jan. 2023, doi: 10.1038/s41598-023-27838-3.

[16] K. T. M and Y. N. G R, “Learning-Based Super-Resolution for Image Upscaling Using Sparse Representation,” in 2024 Second International Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, Karnataka, India: IEEE, Jul. 2024, pp. 1–4. doi: 10.1109/ICAIT61638.2024.10690294.

[17] H. Yan, Z. Wang, Z. Xu, Z. Wang, Z. Wu, and R. Lyu, “Research on Image Super-Resolution Reconstruction Mechanism based on Convolutional Neural Network,” in International Conference on Artificial Intelligence, Automation and High Performance Computing, Zhuhai China: ACM, Jul. 2024, pp. 142–146. doi: 10.1145/3690931.3690956.

[18] C. Fiscone et al., “Generalizing the Enhanced-Deep-Super-Resolution Neural Network to Brain MR Images: A Retrospective Study on the Cam-CAN Dataset,” eNeuro, vol. 11, no. 5, p. ENEURO.0458-22.2023, May 2024, doi: 10.1523/ENEURO.0458-22.2023.

[19] Y. Zhang, S. Liu, C. Dong, X. Zhang, and Y. Yuan, “Multiple Cycle-in-Cycle Generative Adversarial Networks for Unsupervised Image Super-Resolution,” IEEE Trans. on Image Process., vol. 29, pp. 1101–1112, 2020, doi: 10.1109/TIP.2019.2938347.

[20] K. S. Krishnan and K. S. Krishnan, “SwiftSRGAN -- Rethinking Super-Resolution for Efficient and Real-time Inference,” in 2021 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Dec. 2021, pp. 46–51. doi: 10.1109/ICICyTA53712.2021.9689188.

[21] A. K. Dwivedi, A. K. Singh, and D. Singh, “An Object Based Image Analysis of Multispectral Satellite and Drone Images for Precision Agriculture Monitoring,” in IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 4899–4902. doi: 10.1109/IGARSS46834.2022.9884674.

[22] J. Fernández-Guisuraga, E. Sanz-Ablanedo, S. Suárez-Seoane, and L. Calvo, “Using Unmanned Aerial Vehicles in Postfire Vegetation Survey Campaigns through Large and Heterogeneous Areas: Opportunities and Challenges,” Sensors, vol. 18, no. 2, p. 586, Feb. 2018, doi: 10.3390/s18020586.

[23] W. Huang, W. Li, L. Yang, W. Zhang, and L. Wang, “Disaster Rescue Drone Based on YOLOv4 Algorithm,” J. Phys.: Conf. Ser., vol. 2850, no. 1, p. 012005, Sep. 2024, doi: 10.1088/1742-6596/2850/1/012005.

[24] S. B. Nirmal, “Applications of Drone and Unmanned Aerial Vehicle (UAV) Surveying for Planning For Cities,” IJRASET, vol. 12, no. 3, pp. 885–888, Mar. 2024, doi: 10.22214/ijraset.2024.58897.

[25] L. Grigore and C. Cristescu, “The Use of Drones in Tactical Military Operations in the Integrated and Cybernetic Battlefield,” Land Forces Academy Review, vol. 29, no. 2, pp. 269–273, Jun. 2024, doi: 10.2478/raft-2024-0029.

[26] T. Ye, W. Qin, Y. Li, S. Wang, J. Zhang, and Z. Zhao, “Dense and Small Object Detection in UAV-Vision Based on a Global-Local Feature Enhanced Network,” IEEE Trans. Instrum. Meas., vol. 71, pp. 1–13, 2022, doi: 10.1109/TIM.2022.3196319.

[27] Y. Li and X. Li, “Utilizing unmanned aerial vehicle technology for environmental monitoring: Future trends, methods, and applications,” Internet Technology Letters, p. e526, Apr. 2024, doi: 10.1002/itl2.526.

[28] DRS International School and C. Yeragera, “Evaluating the Technological Impact of Unmanned Aerial Vehicles on Efficiency Metrics in Precision Agriculture: A Secondary Data Analysis,” IJSREM, vol. 08, no. 10, pp. 1–7, Oct. 2024, doi: 10.55041/IJSREM37996.

[29] G. M. Upadhyay, M. Joshi, P. Rathi, P. Vats, M. Narula, and S. K. Gupta, “A Comprehensive Framework for Unmanned Aerial Vehicle (UAV)-Enabled Real-Time Human Detection System for Disaster Management,” in 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), Greater Noida, India: IEEE, Aug. 2024, pp. 1–6. doi: 10.1109/ICEECT61758.2024.10739154.

[30] F. Mohammed, A. Idries, N. Mohamed, J. Al-Jaroodi, and I. Jawhar, “Opportunities and Challenges of Using UAVs for Dubai Smart City,” in 2014 6th International Conference on New Technologies, Mobility and Security (NTMS), Dubai, United Arab Emirates: IEEE, Mar. 2014, pp. 1–4. doi: 10.1109/NTMS.2014.6814041.

[31] J. Hu et al., “High Resolution 3-D Imaging via Distributed UAVs SAR Tomography,” in 2024 Photonics & Electromagnetics Research Symposium (PIERS), Chengdu, China: IEEE, Apr. 2024, pp. 1–5. doi: 10.1109/PIERS62282.2024.10618571.

[32] F. Gaspari, F. Ioli, F. Barbieri, E. Belcore, and L. Pinto, “INTEGRATION OF UAV-LIDAR AND UAV-PHOTOGRAMMETRY FOR INFRASTRUCTURE MONITORING AND BRIDGE ASSESSMENT,” Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLIII-B2-2022, pp. 995–1002, May 2022, doi: 10.5194/isprs-archives-XLIII-B2-2022-995-2022.

[33] D. Mehta, A. Mehta, P. Narang, V. Chamola, and S. Zeadally, “Deep Learning Enhanced UAV Imagery for Critical Infrastructure Protection,” IEEE Internet Things M., vol. 5, no. 2, pp. 30–34, Jun. 2022, doi: 10.1109/IOTM.004.2200006.

[34] H.-J. Kim, S.-K. Kwon, and T.-H. Eom, “An Intrascene Wide Dynamic Range CMOS Image Sensor Using Dual-Gain Ramp Generator for Machine Vision Applications,” IEEE Trans. Instrum. Meas., vol. 73, pp. 1–8, 2024, doi: 10.1109/TIM.2024.3440392.

[35] J. E. Albuquerque F. and C. R. Jung, “Joint-task learning to improve perceptually-aware super-resolution of aerial images,” International Journal of Remote Sensing, vol. 44, no. 6, pp. 1820–1841, Mar. 2023, doi: 10.1080/01431161.2023.2190469.

[36] P. Wang and E. Sertel, “Multi-frame super-resolution of remote sensing images using attention-based GAN models,” Knowledge-Based Systems, vol. 266, p. 110387, Apr. 2023, doi: 10.1016/j.knosys.2023.110387.

[37] E. Alvarez-Vanhard, G. F. Garcia, and T. Corpetti, “Super-Resolution by Fusing Multispectral and Terrain Models: Application to Water Level Mapping,” IEEE Geosci. Remote Sensing Lett., vol. 20, pp. 1–5, 2023, doi: 10.1109/LGRS.2023.3319548.

[38] H. Fkih, A. Kallel, and Z. Chtourou, “Super-Resolution of UAVs Thermal Images Guided by Visible Images,” in 2023 International Conference on Cyberworlds (CW), Sousse, Tunisia: IEEE, Oct. 2023, pp. 40–45. doi: 10.1109/CW58918.2023.00016.

[39] S. YiĞiT, “Improving object detection of UAV images with Real-ESRGAN,” Recent Adv Sci Eng, pp. 33–39, 2023, doi: 10.14744/rase.2023.0004.

[40] C. Aybar, D. Montero, S. Donike, F. Kalaitzis, and L. Gómez-Chova, “A Comprehensive Benchmark for Optical Remote Sensing Image Super-Resolution,” IEEE Geosci. Remote Sensing Lett., vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2024.3401394.

[41] L. Rossi, V. Bernuzzi, T. Fontanini, M. Bertozzi, and A. Prati, “Swin2-MoSE: A New Single Image Super-Resolution Model for Remote Sensing,” Apr. 29, 2024, arXiv: arXiv:2404.18924. doi: 10.48550/arXiv.2404.18924.

[42] E. Alves Nogueira et al., “Enhancing Corn Image Resolution Captured by Unmanned Aerial Vehicles With the Aid of Deep Learning,” IEEE Access, vol. 12, pp. 149090–149098, 2024, doi: 10.1109/ACCESS.2024.3476232.

[43] Y. Xiong et al., “Improved SRGAN for Remote Sensing Image Super-Resolution Across Locations and Sensors,” Remote Sensing, vol. 12, no. 8, p. 1263, Apr. 2020, doi: 10.3390/rs12081263.

[44] J. Xie, L. Fang, B. Zhang, J. Chanussot, and S. Li, “Super Resolution Guided Deep Network for Land Cover Classification From Remote Sensing Images,” IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–12, 2022, doi: 10.1109/TGRS.2021.3120891.

[45] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, “From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution,” Information Fusion, vol. 96, pp. 297–311, Aug. 2023, doi: 10.1016/j.inffus.2023.03.021.

[46] Y. Huang, X. Wen, Y. Gao, Y. Zhang, and G. Lin, “Tree Species Classification in UAV Remote Sensing Images Based on Super-Resolution Reconstruction and Deep Learning,” Remote Sensing, vol. 15, no. 11, p. 2942, Jun. 2023, doi: 10.3390/rs15112942.

[47] Y. Liu, H. Xu, and X. Shi, “Reconstruction of super-resolution from high-resolution remote sensing images based on convolutional neural networks,” PeerJ Computer Science, vol. 10, p. e2218, Aug. 2024, doi: 10.7717/peerj-cs.2218.

[48] G. Sun, Y. Chen, J. Huang, Q. Ma, and Y. Ge, “Digital Surface Model Super-Resolution by Integrating High-Resolution Remote Sensing Imagery Using Generative Adversarial Networks,” IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 17, pp. 10636–10647, 2024, doi: 10.1109/JSTARS.2024.3399544.

[49] X. Kang, P. Duan, J. Li, and S. Li, “Efficient Swin Transformer for Remote Sensing Image Super-Resolution,” IEEE Trans. on Image Process., vol. 33, pp. 6367–6379, 2024, doi: 10.1109/TIP.2024.3489228.

[50] C. Ledig et al., “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” May 25, 2017, arXiv: arXiv:1609.04802. doi: 10.48550/arXiv.1609.04802.

[51] Jianjun Zhou, last Jianbo Zhang, and last Jiangang Jia & Jie Liu, “SRGAN based super-resolution reconstruction of power inspection images | Discover Applied Sciences.” Accessed: Dec. 07, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s42452-024-06350-x

[52] W. Zeng and Z. Xiao, “Few-shot learning based on deep learning: A survey,” MBE, vol. 21, no. 1, pp. 679–711, 2023, doi: 10.3934/mbe.2024029.

[53] W. Bian, A. Jang, and F. Liu, “Multi-task Magnetic Resonance Imaging Reconstruction using Meta-learning,” Apr. 21, 2024, arXiv: arXiv:2403.19966. doi: 10.48550/arXiv.2403.19966.

[54] H. Ye, K. Su, and S. Huang, “Image Enhancement Method Based on Bilinear Interpolating and Wavelet Transform,” in 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Mar. 2021, pp. 1147–1150. doi: 10.1109/IAEAC50856.2021.9390624.

[55] “Cubic convolution interpolation for digital image processing | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Aug. 20, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/1163711

[56] “Using the cubic spline interpolation method to approximate some real data.” Accessed: Nov. 05, 2024. [Online]. Available: https://ejs.tdmu.edu.vn/using-the-cubic-spline-interpolation-method-to-approximate-some-real-data-347-a25id.html

[57] “New edge-directed interpolation | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Aug. 21, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/951537

[58] A. M. John, K. Khanna, R. R. Prasad, and L. G. Pillai, “A Review on Application of Fourier Transform in Image Restoration,” in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Oct. 2020, pp. 389–397. doi: 10.1109/I-SMAC49090.2020.9243510.

[59] “Application of the fractional Fourier transform to image reconstruction in MRI - Parot - 2012 - Magnetic Resonance in Medicine - Wiley Online Library.” Accessed: Aug. 21, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/mrm.23190

[60] Z. Xiu-chang, “Neighbor Embedding Super-resolution Reconstruction Based on the Training Set Stratification,” Video Engineering, 2012, Accessed: Aug. 22, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Neighbor-Embedding-Super-resolution-Reconstruction-Xiu-chang/15bd64b202f15a4201f552f7af616d0f8a8a8e28

[61] Hong Chang, Dit-Yan Yeung, and Yimin Xiong, “Super-resolution through neighbor embedding,” in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Washington, DC, USA: IEEE, 2004, pp. 275–282. doi: 10.1109/CVPR.2004.1315043.

[62] Jianchao Yang, J. Wright, T. S. Huang, and Yi Ma, “Image Super-Resolution Via Sparse Representation,” IEEE Trans. on Image Process., vol. 19, no. 11, pp. 2861–2873, Nov. 2010, doi: 10.1109/TIP.2010.2050625.

[63] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a Deep Convolutional Network for Image Super-Resolution,” in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 184–199. doi: 10.1007/978-3-319-10593-2_13.

[64] C. Dong, C. C. Loy, and X. Tang, “Accelerating the Super-Resolution Convolutional Neural Network,” 2016, arXiv. doi: 10.48550/ARXIV.1608.00367.

[65] W. Shi et al., “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network,” Sep. 23, 2016, arXiv: arXiv:1609.05158. doi: 10.48550/arXiv.1609.05158.

[66] S. Lee, Y. Chung, and W. Kim, “Enhancing Inductive IR Thermography by Using FFT-Equalization, Motion Tracking Detection and VDSR Super-resolution Processing,” Int. J. Precis. Eng. Manuf.-Smart Tech., vol. 2, no. 2, pp. 133–142, Jul. 2024, doi: 10.57062/ijpem-st.2024.00108.

[67] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-Recursive Convolutional Network for Image Super-Resolution,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 1637–1645. doi: 10.1109/CVPR.2016.181.

[68] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced Deep Residual Networks for Single Image Super-Resolution,” arXiv.org. Accessed: Sep. 07, 2024. [Online]. Available: https://arxiv.org/abs/1707.02921v1

[69] D. A. Talib and A. A. Abed, “DeepFake Image Detection Using Adaptive Discriminator Augmentation (ADA),” in 2023 1st International Conference on Advanced Engineering and Technologies (ICONNIC), Kediri, Indonesia: IEEE, Oct. 2023, pp. 248–253. doi: 10.1109/ICONNIC59854.2023.10467610.

[70] I. J. Goodfellow et al., “Generative Adversarial Networks,” 2014, arXiv. doi: 10.48550/ARXIV.1406.2661.

[71] Z. Ren, L. He, and J. Lu, “Context Aware Edge-Enhanced GAN for Remote Sensing Image Super-Resolution,” IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 17, pp. 1363–1376, 2024, doi: 10.1109/JSTARS.2023.3333271.

[72] Y. Wang, Z. Xu, X. Wang, J. He, and X. Zhao, “An Improved SRGAN-Based Deblurring Model for Multiple Blurriness in Microscopy,” IEEE Trans. Instrum. Meas., vol. 73, pp. 1–13, 2024, doi: 10.1109/TIM.2024.3470059.

[73] B.-K. Xie, S.-B. Liu, and L. Li, “Large-scale microscope with improved resolution using SRGAN,” Optics & Laser Technology, vol. 179, p. 111291, Dec. 2024, doi: 10.1016/j.optlastec.2024.111291.

[74] X. Wang et al., “ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks,” Sep. 17, 2018, arXiv: arXiv:1809.00219. doi: 10.48550/arXiv.1809.00219.

[75] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data,” arXiv.org. Accessed: Sep. 04, 2024. [Online]. Available: https://arxiv.org/abs/2107.10833v2

[76] V. Chudasama and K. Upla, “Computationally efficient progressive approach for single-image super-resolution using generative adversarial network,” J. Electron. Imag., vol. 30, no. 02, Jan. 2021, doi: 10.1117/1.JEI.30.2.021003.

[77] J. Wang et al., “Multisensor Remote Sensing Imagery Super-Resolution with Conditional GAN,” J Remote Sens, vol. 2021, p. 2021/9829706, Jan. 2021, doi: 10.34133/2021/9829706.

[78] H. Wu, L. Zhang, and J. Ma, “Remote Sensing Image Super-Resolution via Saliency-Guided Feedback GANs,” IEEE Trans. Geosci. Remote Sensing, pp. 1–16, 2020, doi: 10.1109/TGRS.2020.3042515.

[79] P. Li, Z. Li, X. Pang, H. Wang, W. Lin, and W. Wu, “Multi-scale residual denoising GAN model for producing super-resolution CTA images,” J Ambient Intell Human Comput, vol. 13, no. 3, pp. 1515–1524, Mar. 2022, doi: 10.1007/s12652-021-03009-y.

[80] J. Chen et al., “TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers,” Medical Image Analysis, vol. 97, p. 103280, Oct. 2024, doi: 10.1016/j.media.2024.103280.

[81] S. A. Ahmed et al., “Advancements in UAV image semantic segmentation: A comprehensive literature review,” Multidiscip. Rev., vol. 7, no. 6, p. 2024118, Mar. 2024, doi: 10.31893/multirev.2024118.

[82] A. S. Raj, S. B. Asha, and G. Gopakumar, “Feature Enhanced Semi-Supervised Attention U-Net for Cell Segmentation,” in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India: IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10726083.

[83] S. N. Patil and H. D. Patil, “Enhancing skin lesion segmentation with U-Net++: Design, analysis, and performance evaluation,” 76, vol. 11, no. 1, pp. 30–44, Feb. 2024, doi: 10.18488/76.v11i1.3635.

[84] A. Kalluvila, “Super-Resolution of Brain MRI via U-Net Architecture,” IJACSA, vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.0140503.

[85] X. Hu, M. A. Naiel, A. Wong, M. Lamm, and P. Fieguth, “RUNet: A Robust UNet Architecture for Image Super-Resolution,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 505–507. doi: 10.1109/CVPRW.2019.00073.

[86] Y. Chen, R. Xia, K. Yang, and K. Zou, “MICU: Image super-resolution via multi-level information compensation and U-net,” Expert Systems with Applications, vol. 245, p. 123111, Jul. 2024, doi: 10.1016/j.eswa.2023.123111.

[87] L. Shan, C. Liu, Y. Liu, Y. Tu, S. V. Chilukoti, and X. Hei, “Single image multi-scale enhancement for rock Micro-CT super-resolution using residual U-Net,” Applied Computing and Geosciences, vol. 22, p. 100165, Jun. 2024, doi: 10.1016/j.acags.2024.100165.

[88] J. Shin, Y.-H. Jo, B.-K. Khim, and S. M. Kim, “U-Net Super-Resolution Model of GOCI to GOCI-II Image Conversion,” IEEE Trans. Geosci. Remote Sensing, vol. 62, pp. 1–12, 2024, doi: 10.1109/TGRS.2024.3361854.

[89] F. Huang, Y. Li, X. Ye, and J. Wu, “Infrared Image Super-Resolution Network Utilizing the Enhanced Transformer and U-Net,” Sensors, vol. 24, no. 14, p. 4686, Jul. 2024, doi: 10.3390/s24144686.

[90] P. Dandekar, B. Bhojwani, A. Balpande, S. Zanwar, and A. Deb, “Image Super Resolution using U-Net architecture and SRGAN: Comparative Analysis,” 2024 2nd International Conference on Computer, Communication and Control (IC4), pp. 1–6, Feb. 2024, doi: 10.1109/IC457434.2024.10486783.

[91] D. Cheng, L. Chen, C. Lv, L. Guo, and Q. Kou, “Light-Guided and Cross-Fusion U-Net for Anti-Illumination Image Super-Resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 12, pp. 8436–8449, Dec. 2022, doi: 10.1109/TCSVT.2022.3194169.

[92] F. Min, L. Wang, S. Pan, and G. Song, “D2 UNet: Dual Decoder U-Net for Seismic Image Super-Resolution Reconstruction,” IEEE Trans. Geosci. Remote Sensing, vol. 61, pp. 1–13, 2023, doi: 10.1109/TGRS.2023.3264459.

[93] Z.-S. Liu, W.-C. Siu, and L.-W. Wang, “Variational AutoEncoder for Reference based Image Super-Resolution,” Jun. 08, 2021, arXiv: arXiv:2106.04090. doi: 10.48550/arXiv.2106.04090.

[94] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, “Photo-Realistic Image Super-Resolution via Variational Autoencoders,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 4, pp. 1351–1365, Apr. 2021, doi: 10.1109/TCSVT.2020.3003832.

[95] J. Xu and Y. Zhao, “Image Super-Resolution Based on Variational Autoencoder and Channel Attention,” in Proceedings of the 2023 6th International Conference on Artificial Intelligence and Pattern Recognition, in AIPR ’23. New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 611–616. doi: 10.1145/3641584.3641675.

[96] J. Li, K. Zheng, L. Ni, and L. Gao, “Dual U-Nets autoencoders for unsupervised hyperspectral image super-resolution,” in International Conference on Remote Sensing, Mapping, and Geographic Systems (RSMG 2023), SPIE, Nov. 2023, pp. 132–137. doi: 10.1117/12.3010344.

[97] “Image Resolution Enhancement Using Convolutional Autoencoders with Skip Connections | IEEE Conference Publication | IEEE Xplore.” Accessed: Sep. 05, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9582015

[98] S. H. Chan, “Tutorial on Diffusion Models for Imaging and Vision,” arXiv.org. Accessed: Sep. 06, 2024. [Online]. Available: https://arxiv.org/abs/2403.18103v1

[99] Q. Bammey, “Synthbuster: Towards Detection of Diffusion Model Generated Images,” IEEE Open Journal of Signal Processing, vol. 5, pp. 1–9, 2024, doi: 10.1109/OJSP.2023.3337714.

[100] G. Zhai and X. Min, “Perceptual image quality assessment: a survey,” Sci. China Inf. Sci., vol. 63, no. 11, p. 211301, Apr. 2020, doi: 10.1007/s11432-019-2757-1.

[101] N. Alkzir, I. Nikolaev, and D. Nikolaev, “Search for image quality metrics suitable for assessing images specially precompensated for users with refractive errors,” in Sixteenth International Conference on Machine Vision (ICMV 2023), W. Osten, Ed., Yerevan, Armenia: SPIE, Apr. 2024, p. 46. doi: 10.1117/12.3023509.

[102] “A novel perceptual loss function for single image super-resolution | Multimedia Tools and Applications.” Accessed: Sep. 06, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11042-020-08878-7

[103] “Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Sep. 06, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/4775883

[104] “Loss Functions for Image Restoration With Neural Networks | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Sep. 06, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/7797130

[105] A. Hepburn, V. Laparra, R. Santos-Rodriguez, and J. Malo, “Disentangling the Link Between Image Statistics and Human Perception,” Oct. 05, 2023, arXiv: arXiv:2303.09874. doi: 10.48550/arXiv.2303.09874.

[106] “Image quality assessment: from error visibility to structural similarity | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Sep. 06, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/1284395

[107] Z.-R. Chen, K. Tian, X.-Y. Zhang, and S.-P. Zhao, “Enhancing Common Loss Functions: ‘Append’ with Overall Metrics as Supplementary,” in 2023 9th International Conference on Mechanical and Electronics Engineering (ICMEE), Nov. 2023, pp. 548–553. doi: 10.1109/ICMEE59781.2023.10525355.

[108] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual Dense Network for Image Super-Resolution,” arXiv.org. Accessed: Sep. 07, 2024. [Online]. Available: https://arxiv.org/abs/1802.08797v2

[109] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, “Meta-SR: A Magnification-Arbitrary Network for Super-Resolution,” arXiv.org. Accessed: Sep. 07, 2024. [Online]. Available: https://arxiv.org/abs/1903.00875v4

[110] B. Niu et al., “Single Image Super-Resolution via a Holistic Attention Network,” arXiv.org. Accessed: Sep. 07, 2024. [Online]. Available: https://arxiv.org/abs/2008.08767v1

[111] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image Restoration Using Swin Transformer,” arXiv.org. Accessed: Sep. 07, 2024. [Online]. Available: https://arxiv.org/abs/2108.10257v1

[112] L. A. Gatys, A. S. Ecker, and M. Bethge, “A Neural Algorithm of Artistic Style,” Sep. 02, 2015, arXiv: arXiv:1508.06576. doi: 10.48550/arXiv.1508.06576.

[113] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” Mar. 26, 2016, arXiv: arXiv:1603.08155. doi: 10.48550/arXiv.1603.08155.

[114] W. Peng, X. Qian, and W. Song, “A new style transfer method based on color prioritization,” in 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL), May 2023, pp. 385–390. doi: 10.1109/CVIDL58838.2023.10166186.

[115] A. Utane and S. W. Mohod, “Hybrid Architecture for Traffic Light Recognition Using Deep CNN and Ensemble Machine Learning Model,” in Proceedings of Third Emerging Trends and Technologies on Intelligent Systems, Springer, Singapore, 2023, pp. 121–132. doi: 10.1007/978-981-99-3963-3_10.

[116] L. Gao et al., “Scaling and evaluating sparse autoencoders,” Jun. 06, 2024, arXiv: arXiv:2406.04093. doi: 10.48550/arXiv.2406.04093.

[117] M. F. Ferreira, R. Camacho, and L. F. Teixeira, “Using autoencoders as a weight initialization method on deep neural networks for disease detection,” BMC Med Inform Decis Mak, vol. 20, no. Suppl 5, p. 141, Aug. 2020, doi: 10.1186/s12911-020-01150-w.

[118] M. Marais, M. Hartstein, and G. Cevora, “Using linear initialisation to improve speed of convergence and fully-trained error in Autoencoders,” Nov. 17, 2023, arXiv: arXiv:2311.10699. doi: 10.48550/arXiv.2311.10699.

[119] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” Jan. 07, 2016, arXiv: arXiv:1511.06434. doi: 10.48550/arXiv.1511.06434.

[120] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard GAN,” Sep. 10, 2018, arXiv: arXiv:1807.00734. doi: 10.48550/arXiv.1807.00734.

[121] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein Generative Adversarial Networks,” in Proceedings of the 34th International Conference on Machine Learning, PMLR, Jul. 2017, pp. 214–223. Accessed: Sep. 09, 2024. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[122] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved Training of Wasserstein GANs,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Sep. 09, 2024. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

[123] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” Aug. 24, 2020, arXiv: arXiv:1703.10593. doi: 10.48550/arXiv.1703.10593.

[124] Y. Blau and T. Michaeli, “The Perception-Distortion Tradeoff,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 6228–6237. doi: 10.1109/CVPR.2018.00652.

[125] D. Freirich, T. Michaeli, and R. Meir, “A Theory of the Distortion-Perception Tradeoff in Wasserstein Space,” arXiv.org. Accessed: Sep. 09, 2024. [Online]. Available: https://arxiv.org/abs/2107.02555v1

[126] E. F. Durech, “Deep Convolutional Neural Network for Non-rigid Image Registration,” 2021, arXiv. doi: 10.48550/ARXIV.2104.12034.

[127] M. Gao, Y. Bai, Y. Xie, B. Zhang, S. Li, and Z. Li, “SMC-SRGAN-Lightning super-resolution algorithm based on optical micro-scanning thermal microscope image,” Vis Comput, Jan. 2024, doi: 10.1007/s00371-023-03247-5.

[128] B. Kim, E.-J. An, S. Kim, K. R. Sri Preethaa, D.-E. Lee, and R. R. Lukacs, “SRGAN-enhanced unsafe operation detection and classification of heavy construction machinery using cascade learning,” Artif Intell Rev, vol. 57, no. 8, p. 206, Jul. 2024, doi: 10.1007/s10462-024-10839-7.

[129] S. Madhav, T. M. Nandhika, and M. K. Kavitha Devi, “Super Resolution of Medical Images Using SRGAN,” 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), pp. 1–6, Feb. 2024, doi: 10.1109/ic-ETITE58242.2024.10493588.

[130] H. Li, L. Cheng, and J. Liu, “A new degradation model and an improved SRGAN for multi-image super-resolution reconstruction,” The Imaging Science Journal, pp. 1–20, Mar. 2024, doi: 10.1080/13682199.2024.2331813.

[131] J. Ferdousi, S. I. Lincoln, Md. K. Alom, and Md. Foysal, “A deep learning approach for white blood cells image generation and classification using SRGAN and VGG19,” Telematics and Informatics Reports, vol. 16, p. 100163, Dec. 2024, doi: 10.1016/j.teler.2024.100163.

[132] L.-L. Tian, W. Hu, H.-Z. Zhou, L. Yu, L. Li, and L. Li, “SRGAN Algorithm-Assisted Electrically Controllable Zoom System Based on Pancharatnam-Berry Liquid Crystal Lens for VR/AR Display,” ACS Photonics, vol. 11, no. 7, pp. 2787–2796, Jul. 2024, doi: 10.1021/acsphotonics.4c00659.

[133] Y. Wang, Z. Xu, X. Wang, J. He, and X. Zhao, “An Improved SRGAN-based Deblurring Model for Multiple Blurriness in Microscopy,” IEEE Trans. Instrum. Meas., pp. 1–1, 2024, doi: 10.1109/TIM.2024.3470059.

[134] M. Tan, H. Wang, and W. Zhang, “SRGAN and CNN integration for enhanced chest CT diagnostics,” ACE, vol. 45, no. 1, pp. 213–219, Mar. 2024, doi: 10.54254/2755-2721/45/20241170.

[135] A. Gupta and R. Duggal, “P-TELU: Parametric Tan Hyperbolic Linear Unit Activation for Deep Neural Networks,” in 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice: IEEE, Oct. 2017, pp. 974–978. doi: 10.1109/ICCVW.2017.119.

[136] S. Ji et al., “Super-resolution reconstruction of variable length infrared image sequences based on convolutional neural networks and pixel shuffling,” in 2024 International Conference on Optoelectronic Information and Optical Engineering (OIOE 2024), H. Bin Ahmad and M. Jiang, Eds., Kunming, China: SPIE, Jun. 2024, p. 10. doi: 10.1117/12.3030372.

[137] K. Mohammadi, A. Islam, and S. B. Belhaouari, “Zooming Into Clarity: Image Denoising Through Innovative Autoencoder Architectures,” IEEE Access, vol. 12, pp. 98816–98834, 2024, doi: 10.1109/ACCESS.2024.3424972.

[138] A.-S. Collin, C. De Bodt, D. Mulders, and C. De Vleeschouwer, “Don’t skip the skips: autoencoder skip connections improve latent representation discrepancy for anomaly detection,” in ESANN 2023 proceesdings, Bruges (Belgium) and online: Ciaco - i6doc.com, 2023, pp. 653–658. doi: 10.14428/esann/2023.ES2023-139.

[139] I. Koo, D.-K. Chae, and S.-C. Lee, “Improving Adversarial Robustness via Distillation-Based Purification,” Applied Sciences, vol. 13, no. 20, Art. no. 20, Jan. 2023, doi: 10.3390/app132011313.

[140] A. Chaurasia and E. Culurciello, “LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation,” in 2017 IEEE Visual Communications and Image Processing (VCIP), Dec. 2017, pp. 1–4. doi: 10.1109/VCIP.2017.8305148.

[141] M. Hasan, M. Vijay, S. Sharanyaa, and V. S. D. Tejaswi, “Ensemble Model with Improved U-Net-based Segmentation for Leukemia Detection,” Biomed. Eng. Appl. Basis Commun., vol. 36, no. 03, p. 2450011, Jun. 2024, doi: 10.4015/S101623722450011X.

[142] S. Thomas and R. Sudarmani, “Optimized U-Net Segmentation Model and Deep Maxout Classifier for Brain Tumor Classification,” Biomed. Eng. Appl. Basis Commun., p. 2450027, Jul. 2024, doi: 10.4015/S1016237224500273.

[143] K. He, X. Zhang, S. Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9908. Cham: Springer International Publishing, 2016, pp. 630–645. doi: 10.1007/978-3-319-46493-0_38.

[144] S. Zhu and L. Ma, “Combining global gate axial-attention with U-Net for skin lesion segmentation,” 2024 AI Photonics Technology Symposium, p. 3, Sep. 2024, doi: 10.1117/12.3034995.

[145] S. K. B. Degala, R. P. Tewari, P. Kamra, U. Kasiviswanathan, and R. Pandey, “Segmentation and Estimation of Fetal Biometric Parameters using an Attention Gate Double U-Net with Guided Decoder Architecture,” Computers in Biology and Medicine, vol. 180, p. 109000, Sep. 2024, doi: 10.1016/j.compbiomed.2024.109000.

[146] S. Mullan, L. Zhang, H. Zhang, and M. Sonka, “Deep learning medical image segmentation,” in Medical Image Analysis, Elsevier, 2024, pp. 475–500. doi: 10.1016/B978-0-12-813657-7.00042-X.

[147] O. Oktay et al., “Attention U-Net: Learning Where to Look for the Pancreas,” May 20, 2018, arXiv: arXiv:1804.03999. Accessed: Jun. 22, 2024. [Online]. Available: http://arxiv.org/abs/1804.03999

[148] L. Hu, K. Zhao, B. Wing-Kuen Ling, S. Liang, and Y. Wei, “Improving human activity recognition via graph attention network with linear discriminant analysis and residual learning,” Biomedical Signal Processing and Control, vol. 100, p. 107053, Feb. 2025, doi: 10.1016/j.bspc.2024.107053.

[149] R. Kaur and S. Kaur, “Automatic skin lesion segmentation using attention residual U-Net with improved encoder-decoder architecture,” Multimed Tools Appl, Mar. 2024, doi: 10.1007/s11042-024-18895-5.

[150] M. Rifqi Rafsanjani et al., “Preliminary evaluation of an automated autoencoder-UNet pipeline for chemical image segmentation and compression with reference to serial ground truth pathology,” Data Science for Photonics and Biophotonics, p. 18, Jun. 2024, doi: 10.1117/12.3022279.

[151] S. Abinaya, K. U. Kumar, and A. S. Alphonse, “Cascading Autoencoder With Attention Residual U-Net for Multi-Class Plant Leaf Disease Segmentation and Classification,” IEEE Access, vol. 11, pp. 98153–98170, 2023, doi: 10.1109/ACCESS.2023.3312718.

[152] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training Generative Adversarial Networks with Limited Data,” Oct. 07, 2020, arXiv: arXiv:2006.06676. Accessed: Oct. 15, 2024. [Online]. Available: http://arxiv.org/abs/2006.06676

[153] S. Youwai, A. Chaiyaphat, and P. Chaipetch, “YOLO9tr: a lightweight model for pavement damage detection utilizing a generalized efficient layer aggregation network and attention mechanism,” J Real-Time Image Proc, vol. 21, no. 5, p. 163, Aug. 2024, doi: 10.1007/s11554-024-01545-2.

[154] K. Wang, H. Zhou, H. Wu, and G. Yuan, “RN-YOLO: A Small Target Detection Model for Aerial Remote-Sensing Images,” Electronics, vol. 13, no. 12, Art. no. 12, Jan. 2024, doi: 10.3390/electronics13122383.

[155] F. Vela, R. Fonseca-Delgado, and I. Pineda, “A Shallow Approach for Vehicle Speed Estimation in Urban Areas Using YOLO, GOG, and a MLP,” in 2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM), Oct. 2024, pp. 1–6. doi: 10.1109/ETCM63562.2024.10746120.

[156] M. Narkhede and N. Chopade, “CycleInSight: An enhanced YOLO approach for vulnerable cyclist detection in urban environments,” IJECE, vol. 14, no. 4, p. 3986, Aug. 2024, doi: 10.11591/ijece.v14i4.pp3986-3994.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Professor Li. His invaluable guidance, insightful feedback, and unwavering support have been instrumental throughout the course of my research. His expertise and mentorship have greatly enhanced my academic journey, and I am truly fortunate to have had the opportunity to work under his supervision.

I would also like to extend my sincere thanks to Northwestern Polytechnical University for providing me with an excellent academic environment and the resources necessary to complete my research. The university's commitment to fostering innovation and academic excellence has played a significant role in the successful completion of this thesis.

My heartfelt appreciation goes to the Chinese Scholarship Council for awarding me a full scholarship from the Chinese Government. This generous financial support has allowed me to pursue my studies without any financial burden and has provided me with the opportunity to focus entirely on my research.

Finally, I would like to express my profound gratitude to my parents and my brother for their endless love, encouragement, and belief in me. Their unwavering support has been my greatest source of strength throughout this journey, and I am deeply thankful for their presence in my life.

List of Publications

[1] *Super-Resolution Reconstruction of UAV Images with GANs: Achievements and Challenges, The 2024 International Conference on Cyber-physical Social Intelligence, (November ,2024, Doha, Qatar)), DOI: 10.1109/ICCSI62669.2024.10799467*

[2] *Color Image Segmentation of Dental Caries Using U-Net Enhanced with Residual Blocks and Attention Mechanisms, The 2024 International Conference on Cyber-physical Social Intelligence, (November ,2024, Doha, Qatar), DOI: 10.1109/ICCSI62669.2024.10799395*

[3] *Decentralized Multi-robot Path Planning using Graph Neural Networks, The 2024 International Conference on Cyber-physical Social Intelligence, (November ,2024, Doha, Qatar), DOI: 10.1109/ICCSI62669.2024.10799217*

[4] *Research on Informative Path Planning Using Deep Reinforcement learning, The 2024 International Conference on Cyber-physical Social Intelligence, (November ,2024, Doha, Qatar), DOI: 10.1109/ICCSI62669.2024.10799452*

西北工业大学

学位论文知识产权声明书

本人完全了解学校有关保护知识产权的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于西北工业大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。学校可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。同时本人保证，毕业后结合学位论文研究课题再撰写的文章一律注明作者单位为西北工业大学。

本学位论文属于（在以下方框内打“√”）：

□ 保密论文，保密期（　年　月　日至　年　月　日）。

☑ 公开论文。

学位论文作者签名：Rachakkeh　　指导教师签名：

2025 年03月11日　　2025 年03月11日

西北工业大学

学位论文原创性声明

秉承学校严谨的学风和优良的科学道德，本人郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容和致谢的地方外，本论文不包含任何其他个人或集体已经公开发表或撰写过的研究成果，不包含本人或其他已申请学位或其他用途使用过的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式表明。

本人学位论文与资料若有不实，愿意承担一切相关的法律责任。

学位论文作者签名：Rachakkeh

2025 年03月11日

西北工业大学学位办制