Deep Learning Approaches for Single Image and Video Super-Resolution: Leveraging CNN, RNN and GAN Architectures for Upscaling

by

Moinul Hossain Bhuiyan
20301002
Istiaq Ahmad
20301056
Abdullah Al Mamun
20301062
Labib Sadman Azam
21301643

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

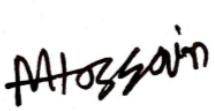
Department of Computer Science and Engineering
Brac University
September 2023

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Moinul Hossain Bhuiyan

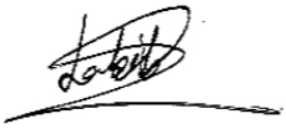
20301002

Istiaq Ahmad

20301056

Abdullah Al Mamun

20301062



Labib Sadman Azam

21301643

Approval

The thesis titled "Deep Learning Approaches for Single Image and Video Super-Resolution: Leveraging CNN, RNN and GAN Architectures for Upscaling" submitted by

1. Moinul Hossain Bhuiyan (20301002)
2. Istiaq Ahmad (20301056)
3. Abdullah Al Mamun (20301062)
4. Labib Sadman Azam (21301643)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September 20th, 2023.

Examining Committee:

Supervisor:
(Member)

Dr. Muhammad Iqbal Hossain

Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

In a world where visual content plays a crucial role in anything imaginable, the need for sharper, more detailed images and videos has never been more important. This research paper explores innovative approaches to improve the quality of both single images and videos through the application of Deep Learning techniques, specifically Convolutional Neural Networks (CNN) Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GAN). Upscaling is essential because many people out there own older devices that can not properly output high-definition content. These devices struggle to display high-quality videos or images that are stored locally. That is the reason why video upscaling methods are needed to help these devices to first render lower-quality content and then enhance it to the quality people desire to see. This research explores how smart computer systems, using advanced techniques of CNNs, RNNs, and GANs, may remarkably increase the quality of pictures and videos. Imagine converting grainy photos into clear, vivid ones and making visuals smoother and more detailed. The examination delves into how different technologies collaborate to enhance individual photographs and videos. The results not only improve entertainment but also have practical consequences in the medical sector, security, and beyond. By harnessing the power of deep learning, a new level of clarity and richness is added to the pictures the user sees every day. Also, the research paper aims to demonstrate the potential of deep learning techniques to improve the visual quality of any Low Resolution content.

Keywords: Super Resolution (SR), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Generative Adversarial Networks (GAN)

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

CNN Convolutional Neural Network

$EGAN$ Enhanced Generative Adversarial Network

$FFMPEG$ Fast Forward Moving Picture Experts Group

$FPGA$ Field Programmable Gate Array

GAN Generative Adversarial Network

$MMCNN$ Multi-Memory Convolutional Neural Network

RNN Recurrent Neural Network

$SISR$ Single Image Super Resolution

SR Super Resolution

VSR Video Super Resolution

Chapter 1

Introduction

1.1 Background

In a world filled with visual content, the desire for crisper, more colorful pictures and videos has become a fundamental component of an individual's digital experience. It is a common phenomenon that, when a user has looked at a photo or watched a video that appeared somewhat hazy or lacking the desired sharpness? The number of these people is a lot. Many of us have this difficulty, particularly on older devices that struggle to display high-definition material. Fortunately, this research looks into an interesting field of technology that may convert these less-than-perfect pictures or videos into something absolutely extraordinary.

High-definition displays have grown less expensive and more common since the mid-2000s, and can now be found in TVs, computer monitors, and even the devices a common person uses most, which is their smartphones. In the era where online streaming is at its peak, here only video is responsible for 88% of total internet traffic in 2023 [12].

Before the machine learning revolution of the 2010s, most upscaling tasks were accomplished using proprietary algorithms; in recent years, scholars have moved their emphasis towards employing deep neural networks instead, particularly making use of CNNs or GANs. This method has previously discovered widespread usage in industries such as picture editing [16] and video games [33].

While video production has Adapted to changes in display technology by simply recording at higher resolutions, material generated before these resolution shifts is trapped at its prior resolution, and does not look well on today's monitors or displays without the use of some form of scaling.

In this case, the video upscaling method plays a significant role. Using neural networking, there are some techniques that can accomplish the task. Imagine having the ability to take those older, lower-quality videos and pictures and suddenly make them appear as though they were recorded with the newest high-end equipment. That's exactly what this research paper emphasizes on. This paper emerges into the area of deep learning, a form of computer intelligence that's surprisingly effective at learning from examples and improving things. In particular, this paper will

investigate three strong technologies: CNNs, RNNs, and GANs.

Additionally, bicubic interpolation is used to upgrade low-resolution information, such as vintage broadcast recordings in PAL resolution (720 x 576) [2]. video filters in FFMPEG [31], with Lanczos [1] filtering being utilized in various tests.But, " In terms of a single image, the CNN model is highly efficient where the model can detect the object and recognize it, opening a new era where images are being processed automatically" [17].

To enhance the resolution of single images, it is seen utilizing the power of CNNs, which excel at recognizing patterns and features in visual data. After training these networks on vast datasets, it enables them to generate high-resolution (HR) images from low-resolution (LR) inputs, allowing a person to witness the final output which has more depth or definition into it.

As a result, when it comes to videos, maintaining consistent high-quality output is a challenging task. Here, RNNs come into play, as they excel at processing sequential data.

Lastly, a little bit of Artificial intelligence touch is needed to properly make an image or video look real. Here, GANs add an extra layer of magic. The abbreviation of GAN stands for generative adversarial networks. They're like artists who make the enhanced images look even more realistic. GANs help fill in the missing pieces and make the upscaled pictures and videos look like they were shot in high definition from the start.

So, eventually, if a user has ever desired sharper, more vivid images on their older devices, this study paper is a ticket to knowing how cutting-edge technology might turn a dream into reality. Welcome to the world of deep learning super-resolution, where photographs and videos are about to receive a stunning makeover.

1.2 Problem Statement

Video upscaling plays a critical role in preserving the quality and interoperability of visual material in today's digital age. When a video is not upscaled as required, it may give rise to a number of difficulties that influence both the watching experience and the overall efficacy of the information. If a video has a low quality and has not been correctly enhanced then the initial problem that can be detected is the loss of details. In a low-quality video, it lacks the essential pixel density. In conclusion, a viewer may find diminished clarity, resulting in it being harder to recognize objects, or facial expressions within the content. Additionally, a low-quality video that is not upscaled might seem hazy or pixelated, particularly when presented on bigger displays or displays that support higher resolutions.

Not only that, content inside the video, such as subtitles, on-screen graphics, or essential comments, might become unreadable when the video is not upscaled. This may hamper understanding, especially in instructional or informative topics. When

analyzing the role of a video, it typically acts as a medium for transmitting information, generating emotions, or describing tales. When details are lost owing to the lack of upscaling, the footage's capacity to effectively deliver its intended message is weakened, resulting in misconceptions or distracted viewers. All those aspects add up to a terrible user experience.

Besides, the screens that are used today, these all have precise aspect ratios. For instance, a high-definition (HD) video has to be played on a 16:9 ratio. If it has to scale up to 4K, the aspect ratio would be 16:9 or 21:9. Whenever a video is not upscaled to fit the right aspect ratio, it may seem stretched, deformed, or have black bars on the sides, which is aesthetically unpleasant and breaks the intended composition of the video. Users do not prefer having bezels in the videos, they prefer a much more immersive experience. As a high-resolution display is meant to deliver immersive and cinematic viewing experiences. While the video is not enhanced to make use of these capabilities, consumers might feel disenchanted with the content and lose out on the engaging characteristics that higher resolutions may bring. In terms of media consumption, a person can understand that video upscaling plays a crucial role in one's day-to-day life. Not just media consumption it also plays a significant role in the gaming industry.

Supersampling approaches like DLSS (Deep Learning Super Sampling) and FSR (Fidelity Super Resolution) have become vital tools in the area of gaming. Their value rests in their capacity to dramatically improve the game experience on many platforms.

One of the key advantages of these technologies is their potential to increase the visual quality of video games. DLSS and FSR do this by automatically upscaling lower-resolution visuals in real time. This leads to photos that are not only crisper but also more precise and aesthetically attractive. The upgraded visuals add to a heightened feeling of realism inside the game world, making virtual settings and people more alive and engaging. Gamers may immerse themselves in finely detailed settings and notice subtler subtleties in character design, all of which contribute to the overall attractiveness of the gaming experience.

However, it's not only about appearances. DLSS and FSR also play a crucial part in ensuring games operate smoothly on a broad variety of gaming devices. By decreasing the computing burden necessary for generating high-quality images, these technologies allow games to reach greater frame rates, resulting in smoother gameplay. The importance of this cannot be emphasized, since smoother gameplay not only increases the gaming experience but also gives a competitive advantage in online multiplayer games, where split-second reflexes may make all the difference.

More on that, upscaling is akin to waving a magic wand over older computers and gaming consoles. It breathes fresh life into aged gear by allowing it to run complex, expensive games that were formerly out of reach. Users no longer need to feel confined by their equipment's limits; they can now explore the newest games with increased performance and visuals. Not only in PC or mobile gaming it creates a new dimension in Virtual Reality (VR) gaming as well. In the immersive realm

of VR gaming, upscaling acts like a pair of high-tech glasses. VR headsets strive to transport users to compelling virtual worlds, and visual quality is important for a genuine experience. Upscaling boosts the clarity, detail, and realism inside VR settings. As a consequence, users may explore these virtual environments with more ease and interest. Reduced visual abnormalities and enhanced picture quality lead to a more comfortable and motion sickness-free VR experience, making the technology more accessible and attractive to a larger audience. This technology not only increases the game experience but also makes more effective use of hardware resources. By utilizing the potential of available processing power, upscaling eliminates bottlenecks, lowering system strain and overheating concerns. This optimal resource usage leads to longer device lifespans, cheaper maintenance costs, and a smoother gaming experience overall.

By applying deep learning architectures for upscaling, the gaming industry, and the individuals who love to consume content may present the user with more spectacular, detailed, and engaging experiences, finally pushing the boundaries of what's possible in the media and the gaming world.

1.3 Research Objectives

In this research paper, the purpose is to build a hybrid model that utilizes CNN, RNN, and GAN methods. The fundamental purpose of this paper is to examine and develop the use of deep learning methods, including CNNs, RNNs, and GANs, in the arena of single image and video super-resolution. The research objectives of this paper are:

- Investigate the effectiveness of CNNs in enhancing the resolution and quality of single images
- Evaluate the capabilities of RNNs to improve video super-resolution, ensuring temporal coherence between frames.
- Examine the role of GANs in refining and adding realism to the upscaled images and videos
- Explore various combinations and architectures of CNNs, RNNs, and GANs to determine optimal solutions for different scenarios.
- Assess the practical applications of deep learning super-resolution techniques in fields such as medicine, security, and entertainment.
- Provide insights and recommendations for leveraging these technologies to enhance visual content on older or lower-quality devices.

By addressing these goals, this paper will contribute to the evolution of technology that can improve the visual quality of photos and videos, making them sharper, more detailed, and acceptable for a broad variety of real-world applications.

Chapter 2

Related Work

The paper "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network" describes a novel strategy to improve the quality and speed of single-image super-resolution using deep neural networks [4]. Current techniques upscale low-resolution images before improving them, which is ineffective. The authors provide a revolutionary CNN architecture that works directly with low-resolution data and employs a smart upscaling method. This methodology exceeds conventional approaches, delivering better image quality and being significantly faster.

So, the main idea of this research paper is to highlight the importance of super-resolution (SR) in digital image processing, where the objective is to enhance the quality and resolution of low-resolution (LR) photos or videos. It highlights that SR has practical applications in numerous areas including HDTV, medical imaging, satellite imaging, face recognition, and surveillance. The research paper analyzes the problems of SR, including the loss of high-frequency information during the LR-to-HR transition and the inherent uncertainty in the mapping from LR to HR space. It additionally provides two main kinds of SR methods: multi-image SR, which depends on numerous LR photographs of the same scene, and single image super-resolution (SISR), which aims to retrieve HR information from a single LR instance exploiting implicit redundancy discovered in natural data. Both methods have to deal with the ill-posed nature of the issue and need restrictions or previous information to guide the reconstruction process.

After that, the paper "Real-time image upscaling with commonly available resources" discusses the rising demand for high-quality video streaming and the challenge of decreasing data transmission while preserving visual quality [20]. Traditional compression algorithms are getting closer to their limitations, increasing the need for alternative techniques. The idea provides a technique comprising reducing picture resolution on the sender side and upscaling on the receiver's side. However, the difficulty is that consumers have variable resources for upscaling, and conventional techniques may not give sufficient outcomes. The idea seeks to solve this using a real-time deep learning network built to operate on a 2 GHz CPU core, giving higher picture quality compared to standard interpolation approaches. This study illustrates the importance of bit rate in video streaming, highlighting the necessity for reducing data transmission while having acceptable picture and audio

quality. It illustrates that broadcasting raw 1080p video at 60 FPS needs a high bit rate, which may be reduced by lowering the image resolution and frame rate. Compression algorithms currently assist lower bit rates, but additional reductions while still maintaining acceptable quality might lead to tremendous cost savings for video streaming services. The paper also emphasizes that contrary to expectations, consumers value smooth video delivery over video quality, making choices like lower resolution appealing if they result in smoother video playback.

More on the topic, the paper known as "Deep upscaling for video streaming: a case evaluation at SVT" analyzes the study on how deep learning, specifically a convolutional neural network (CNN), may enhance video quality via super-resolution [19]. A large-scale A/B video quality test was conducted in order to compare CNN-based upscaling to the standard bicubic method. The research results show that viewers usually prefer CNN-scaled video, but not necessarily for content that is generally upscaled. The research reveals that deep upscaling technology offers potential but needs more optimization and flexibility to become suitable for mainstream use. So, the main goal of this paper is to highlight the growth of display technology, the rising prevalence of high-definition monitors, and the growing popularity of video streaming services. It also introduces the idea of super-resolution (SR), which includes upscaling low-resolution pictures or videos to better resolutions. The paper discusses the issues connected with SR and how deep learning, especially CNNs and generative adversarial GANs, have become essential to tackling these challenges. The promise of deep super-resolution is highlighted, including its possible applications to improve the quality of historical material and decreases distribution cost for streaming services. Finally, the paper discusses the particular case study that is being conducted to assess whether viewers prefer video scaled using deep learning over traditional bicubic interpolation, highlighting the significance of subjective evaluation in addition to measurements that are objective.

Moving on, the paper "Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications" provides the importance of GANs as a potent framework for image and video synthesis, both unconditionally and with input conditions [18]. GANs have changed the development of high-resolution, photorealistic visuals, a task previously believed tough or unachievable. Their emergence has prompted various unique applications in the creation of content. This paper presents an overview of GANs, highlighting their algorithms and applications in visual synthesis. It dives into critical strategies geared at stabilizing the famously hard GAN training process. Additionally, it covers GAN applications in image translation, image processing, video synthesis, and neural rendering. This paper addresses the fundamental concept of Generative Adversarial Networks (GANs) in the area of deep learning and its major impact on several aspects of visual content synthesis. GANs comprise a generator and discriminator network engaged in a competitive training process, resulting in the development of synthetic data that resembles actual data. GANs have successfully replaced hand-designed components in computer vision pipelines, especially for generation tasks, by deriving objective functions from training data. However, GANs are tough to train because of the changing nature of the generator's output distribution, which demands careful control of training dynamics. Various ways have been proposed to stabilize GAN training over time.

Moreover, the paper also differentiates between unconditional and conditional GAN frameworks, where conditional GANs utilize control signals for more precise generation tasks. This breakthrough has led to numerous intriguing applications in semantic picture synthesis, image-to-image translation, image processing, video synthesis, and neural rendering. The overall topic is that GANs have become an essential tool in the area of computer vision, allowing numerous creative visual content-generating applications.

After that, the paper "Algorithm and Architecture Design of High-Quality Video Upscaling Using Database Free Texture Synthesis" introduces a low-complexity super-resolution (SR) algorithm and hardware architecture that improved the quality of TV pictures in real time without costly resources [3]. The algorithm employs a texture synthesis method without having a database, producing detailed visuals by evaluating the input itself. It also preserves temporal consistency. The hardware design decreases computation by 76% utilizing a partial-sum reuse method and optimizes memory utilization using a tile-based processing approach. Experimental findings reveal that this technique produces high-quality output in real-time at a reasonable hardware cost, addressing difficulties observed in traditional scalers such as zigzag and blurring effects. The main objective of this paper is to emphasize the relevance of TV scalers in enhancing the viewing quality of low-resolution content on high-resolution screens, particularly with the increasing resolution gap between content sources and display devices. This paper's primary goal is to highlight the significance of TV scalers in improving the viewing experience of low-resolution information on high-resolution displays, especially given the increasing resolution gap between content sources and display devices. The zigzag effect, blur effect, and flickering are a few common abnormalities in TV scaling that are examined in this study. To address this problem, interpolation-based methods and super-resolution (SR) algorithms are presented. The goal of the project is to improve image quality while addressing hardware constraints and real-time demands by integrating SR methods into TV scalers. It underlines the need for an effective SR method and architecture that is hardware-friendly in order to close the resolution gap. A summary of the paper's organization and structure is also provided.

A new single-image upscaling method that improves picture quality and efficiency was studied in the work "Image and Video Upscaling from Local Self-examples" [32]. This approach emphasizes local self-similarity in the picture as opposed to other approaches that rely on external databases or a totally input image. It reduces search time while maintaining quality by extracting patches from small, localized portions of the input picture. Particularly effective for lesser scaling factors is the approach. It applies specialized filters for these small scalings, providing high-resolution outcomes compatible with the original picture. The algorithm is basic, efficient, and can be implemented in parallel on a GPU. It shows high-quality resolution enhancements, works perfectly with video sequences, and is capable of real-time enhancement of low-resolution videos into high-definition formats. The basic goal of this paper is to illustrate the importance and difficulties of image upscaling, a fundamental image-editing procedure. The paper highlights the limits of traditional upscaling approaches, which frequently result in artifacts and image abnormalities. It presents an innovative method that uses local scale invariance in real images, concentrating

on small, localized patches in order to improve the accuracy and efficiency of upscaling. The author underlines that this creative method works better for small scaling factors and includes multiple upscaling steps employing specific filter banks to accomplish high-quality resolution enhancement. Additionally, it demonstrates the advantages of the proposed technique for both video and image sequences, along with its efficient implementation on GPUs for real-time performance.

The paper "Learning Spatio-Temporal Downsampling for Effective Video Upscaling" highlights the challenges connected to downsampling in image processing, especially in the context of videos, where improper downsampling may lead to aliasing problems including moire patterns and the wagon-wheel effect [26].To solve this challenge, the researchers offer a framework for neural networks that concurrently learns spatiotemporal downsampling and upsampling. The idea is to keep necessary patterns in the source video while enhancing reconstruction during upsampling. To maintain compatibility with standard image and video storage formats, the downsampling results are encoded as uint8 using a differentiable quantization layer. Two new modules for explicit temporal propagation and space-time feature rearrangement are presented to make more use of spatio-temporal correspondences.

Eventually, results from experiments reveal that this method considerably enhances the quality of space-time reconstruction by maintaining spatial textures and motion patterns throughout both downsampling and upsampling. Additionally, the suggested framework enables many applications, including video resampling, blurry frame reconstruction, and efficient video storage. The fundamental goal of this paper is to solve the issues of scaling high-resolution, high-frame-rate movies in small devices, such as mobile phones and glasses, where constraints in memory and bandwidth for transfer demand a trade-off between spatial and temporal resolution. The research paper notes that traditional nearest-neighbor downsampling, which is commonly used in these kinds of situations, may lead to aliasing concerns owing to high-frequency information being folded over in the downsampled frequency domain.

So, to address aliasing, the research paper recommends using deliberately spatial and temporal anti-aliasing filters, such as optical blur and motion blur, to smudge high-frequency information, allowing the reconstruction of fine features during post-capture processing. These anti-aliasing filters may be pre-designed and deployed with the downsampler during capture. The fundamental uniqueness of this study is the notion of concurrently learning both downsampling and upsampling, which is typically handled separately in traditional methods. By simultaneously learning a downsampler and an upsampler, this paper intends to retain and recover high-frequency information in both space and time during image and video processing. This technique provides advantages for video restoration tasks by enabling the co-design of an upsampler to restore missing details. In summary, the core aim of the study is to offer a unified framework that jointly learns spatiotemporal downsampling and upsampling, highlighting the necessity of keeping high-frequency details in both spatial and temporal dimensions. The paper also covers possible applications of this framework, including video resampling, fuzzy frame restoration, and efficient video storage.

The paper "Super Resolution of Videos using E-GAN" provides a unique strategy called EGAN (Enhanced Generative Adversarial Network) for video super-resolution, which comprises enhancing the details and upscaling the resolution of videos1 [23]. EGAN is highlighted as a more accessible and reliable alternative to conventional neural networks, giving greater performance with fewer artifacts It specializes at maintaining high-frequency details and delivering visually stunning outcomes from substantially downsampled input videos. EGAN varies from standard Generative Adversarial Networks by eliminating Batch Normalization layers. The research paper outlines the benefits of this decision and presents significant improvements in human perceptual quality based on thorough Mean Opinion Score (MOS) and Video Multimethod Assessment Fusion (VMAF) testing compared to state-of-the-art methods. The fundamental concept of this study is to suggest an enhanced method of super-resolution, a technique utilized to improve the resolution and visual quality of videos and images. The research paper explores the use of neural networks, especially GANs, for super-resolution and illustrates the limits and benefits of diverse neural network structures.

Besides, the research offers an enhanced version of SRGAN (Super-Resolution Generative Adversarial Network) by including Residual-in-Residual Dense (RRDB) blocks and eliminating batch normalization layers. These modifications seek to enhance training efficiency and shorten the time necessary for super-resolution. The authors also underline the need to keep internal textures and improve visual quality in the super-resolved images and movies. Additionally, the study describes adjustments made to the discriminator architecture, proposing the idea of a Realistic Average GAN to improve the comparison of visual and realistic quality in images. The suggested method is claimed to provide better textures and improve the frame generation rate compared to SRGAN.

Moreover, the research paper evaluates the suggested method's performance using multiple quality evaluation measures such as VMAF, PSNR, SSIM, and MOS, while noting that training GANs consumes significant time and suggests a powerful graphics processor for quicker frame generation. In summary, the research paper focuses on enhancing super-resolution methods, notably via modifications in neural network architectures to generate higher-quality super-resolved images and videos.

More on this topic, the paper "Multi-Memory Convolutional Neural Network for Video Super Resolution" analyzed Multi-Memory CNN (MMCNN) for video super-resolution (SR), and emphasizes on improving the quality of high-resolution (HR) frames from low-resolution (LR) video sequences [9]. Unlike previous methods, which generally employ a direct connection and single-memory module inside convolutional neural networks (CNNs), MMCNN utilizes spatio-temporal complementary information across LR frames more efficiently.

The MMCNN design contains an optical flow network and an image-reconstruction network, connected in a way that cascades. It employs a sequence of residual blocks for feature extraction and reconstruction, emphasizing intra-frame spatial correlations. Notably, instead of a single-memory module, convolutional long short-term memory is integrated into the residual block, generating a multi-memory residual

block. This approach gradually isolates and keeps inter-frame temporal correlations between consecutive LR frames.

Furthermore, extensive trials on multiple testing datasets with varied scaling factors illustrate the superiority of MMCNN over state-of-the-art approaches. MMCNN delivers greater Peak Signal-to-Noise Ratio (PSNR) and improved visual quality, exceeding the best available approach by up to 1 dB.

The basic goal of this work is to emphasize the relevance of super-resolution methods in computer vision, especially in the context of the rising demand for high-definition videos in formats like 4K and 8K. The paper highlights the significance of video super-resolution (Video SR), which seeks to rebuild high-resolution video frames from low-resolution input frames. Besides, the study examines the evolution of super-resolution methods, from interpolation-based techniques to learning-based approaches, with an emphasis on Convolutional Neural Network (CNN)-based methods that have attracted interest owing to their capacity to learn features from various image samples.

Moreover, it also analyzes the difference between single-image super-resolution (SISR) and video super-resolution (Video SR), emphasizing the necessity to utilize temporal correlations between low-resolution frames in Video SR. Traditional Video SR methods are given, however, they often require computational costs that are higher and struggle with large scaling factors and motions. The research paper covers current CNN-based Video SR techniques and their limitations, notably in terms of computational complexity and the usage of inter-frame temporal information. Finally, the core concept of the paper's suggested method, the Multi-Memory Convolutional Neural Network (MMCNN), is described. The MMCNN contains an optical flow network and an image-reconstruction network, intending to use both spatial and temporal information for enhanced video super-resolution. The paper's contributions and the arrangement of the future parts are also detailed.

In summary, it sets the stage for the paper by emphasizing the importance of Video SR in the era of high-definition displays, highlighting the limitations of existing methods, and introducing the innovative MMCNN framework as a solution to enhance both the quality and efficiency of video super-resolution.

Chapter 3

Work Plan

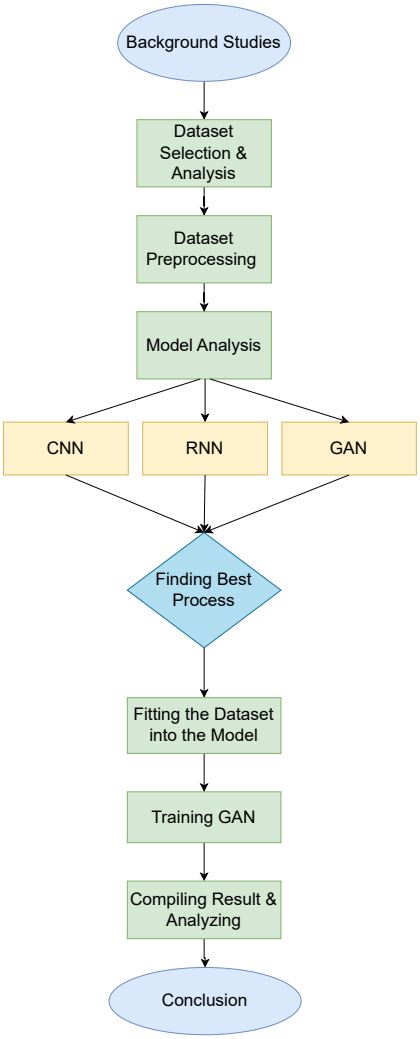


Figure 3.1: Workplan in a flowchart diagram

The research work plan 3.1 starts with a rigorous selection of an appropriate dataset including high-quality photos and videos. This basic step is critical to guarantee that our models get the greatest training data available. After we have obtained our dataset, we will continue to preprocess it and improve it for usage as inputs in our training process.

Our major emphasis focuses on building and improving models for the process of upscaling single photos and videos to attain super-resolution. This key step will include training our algorithms using the carefully produced datasets, allowing them to learn and increase their capacity to boost picture and video resolution.

After completing the training process, we will start the key step of upscaling testing. Here, we will methodically tweak our code and incorporate many layers to refine our models, seeking to obtain the greatest potential performance. This iterative refining process will enable us to examine alternative techniques and combinations to maximize our models.

Ultimately, our study will conclude with a complete examination of the outcomes gained from each model. We will examine how they perform based on multiple metrics and benchmarks to establish their usefulness in super-resolution jobs. Through this comprehensive review, we want to give significant insights and ideas for enhancing the upscaling of single photos and videos.

In summary, our study tries to contribute to the improvement of super-resolution approaches by employing carefully selected datasets, robust model training, and iterative optimization. By thoroughly studying and comparing the outputs of our models, we seek to give a superior strategy for upscaling single photos and videos, thereby boosting the quality of visual information.

Chapter 4

Methodology

Neural networks are computational models inspired by biological neurons in the human brain, consisting of interconnected nodes organized into layers: input, hidden, and output. These layers process input data, learn complex patterns through weighted connections, and produce the final output.

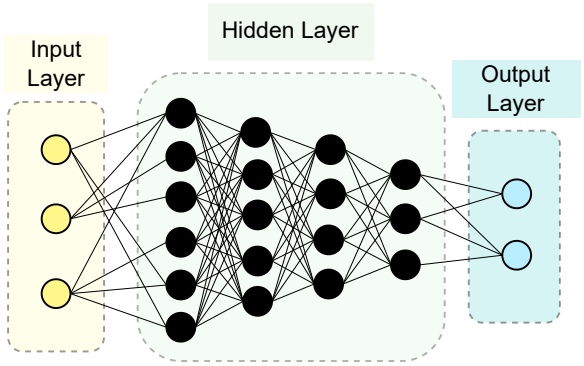


Figure 4.1: Working Mechanism of Neural Network.

Deep learning is a subset of machine learning and it uses neural networks with multiple layers to model and solve complex problems. Deep learning automatically extracts useful information from photos and performs end-to-end learning, allowing networks to learn from raw data and tasks. It scales with data, unlike shallow learning, which converges when additional instances and training data are added. Deep learning's "deep" layers allow the network to learn hierarchical features and their representation, making it successful in natural language processing, reinforcement learning, and computer vision [35].

However, it requires substantial labeled data and computational resources, and the interpretability of complex models can be a challenge. Despite these limitations, neural networks remain the basic building blocks for deep learning, enabling more powerful and automated learning from trained data.

4.1 Convolutional Neural Network (CNN)

The artificial intelligence technique known as a convolutional neural network (CNN) is made specifically for tasks involving images and visual input. It can be viewed as an intelligent system with the ability to "see" and comprehend images.
For making a computer to identify an object in a picture this CNN architecture is important. A CNN consists of layers, and each layer has a distinct function. It is possible for the first layer to identify basic objects like edges and colors. Deeper layers integrate these basic characteristics to comprehend increasingly intricate patterns, such as textures and forms. Each layer gives a unique value. These values of each segment add up and in the end, tell a computer what it is. The concept of applying filters or tiny grids that move over the image is where the word "convolutional" originates. These filters assist the network in learning relevant data and concentrating on particular details. When it comes to managing spatial hierarchies, CNNs excel. CNNs use a technique called pooling, which minimizes the amount of data they must evaluate while maintaining crucial features, to make this function effective. The multidimensional output of the following layer in Convolutional Neural Networks (CNNs) is transformed into a one-dimensional array using a flatten layer. Then using the fully connected layer the model gives the probability of an object.CNN can identify brand-new, unidentified photographs after it has been trained [15] [28] [14] [10].

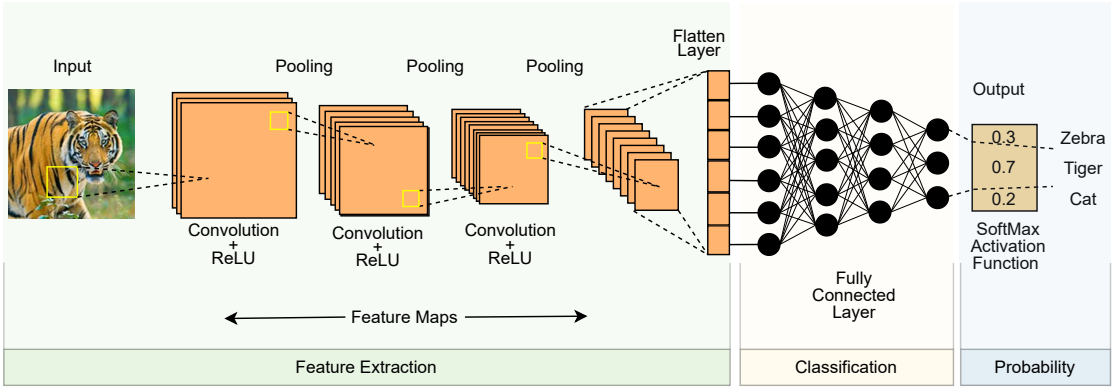


Figure 4.2: The working mechanism of CNN

To sum up, CNNs are similar to visual investigators. They deconstruct images into their more basic components, develop their ability to spot patterns, and apply these skills to comprehend and categorize new images. This makes them an essential tool for many computer vision tasks, including object and image recognition.

4.1.1 ResNet

A kind of convolutional neural network (CNN) used for tasks involving images is called a residual network, or ResNet. The usage of residual blocks, which are intended to aid in the training of extremely deep neural networks, is one of its distinguishing features. A quick connection is used in a residual block to merge the input and output of a layer. The network can concentrate on acquiring residual

information, that is, the variation between a layer's input and output thanks to this connection. The goal is to facilitate the network's ability to recognize and maximize an image's essential characteristics. These shortcut connections solve a common problem in deep networks that is referred to as vanishing gradients. Gradients in deep architectures may decrease during backpropagation as the network gets more complicated, which makes learning more difficult. By giving gradient flow a direct path via the residual connections, training very deep networks is possible without losing crucial data. In ResNet, every residual block picks up unique picture features [34] [13]. The network can recognize and comprehend progressively complex patterns in the data by stacking these blocks. As the network gets deeper, performance deterioration is avoided because of this architecture, which encourages information to go through the network smoothly. In image identification tasks, ResNet has proven remarkably successful, attaining state-of-the-art performance on multiple benchmarks.

4.1.2 VGGNet

The Visual Geometry Group Network, or VGGNet, is a type of Convolutional Neural Network (CNN) designed for picture categorization applications. VGGNet is an intelligent investigator that can study photographs and infer the items or scenes contained in a collection of pictures where a computer is required to identify the contents of each picture. VGGNet stands out for being both efficient and straightforward. Each of the multiple layers in the system is responsible for identifying different aspects of the image. The architecture is deep, with layers upon layers arranged hierarchically, beginning with basic elements such as edges and working up to more complex patterns and forms [30].

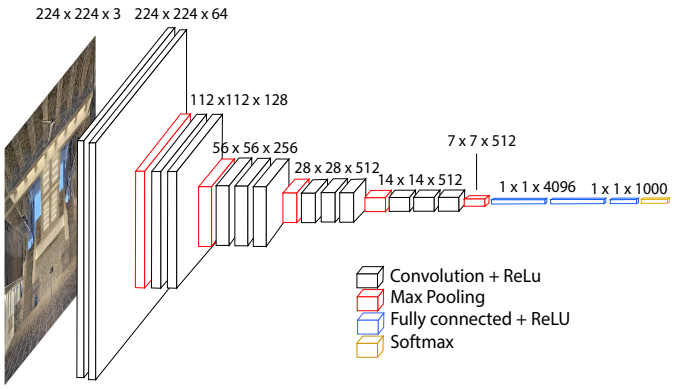


Figure 4.3: The working mechanism of VGG19

One notable feature of VGGNet is that it makes use of some filters, which are small convolutional filters. As these filters move across the input image, they meticulously

pick out patterns [30]. The network's robustness in image identification is enhanced by the acquisition of a wide variety of features and combinations made possible by the integration of numerous layers with these tiny filters. The VGGNet architecture is characterized by the recurrence of layer blocks, and the number of these blocks can be changed to modify the depth of the network. The model VGG-19 is the most commonly used. VGG-19 generates a deep architecture with 19 layers by alternating a series of 3x3 convolutional layers with max-pooling. The network can gradually extract complex information from input photos thanks to this approach. While max-pooling downsamples spatial dimensions and preserves important information, the convolutional layers function as efficient feature detectors. Both low-level and high-level image details are captured by the deep stack of layers, which enables hierarchical feature representation. VGG-19 develops a wide understanding of images through pre-training on datasets. It is effective as a feature extractor for a variety of computer vision tasks, including perceptual loss in tasks like super-resolution, thanks to its simple architecture and weight-sharing approach.

4.2 Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are neural networks used for sequence-based tasks like time series prediction, natural language processing, and speech recognition. However, they have limitations, such as difficulty in capturing long-term dependencies and computational costs. RNNs are not typically used for image or video up-sampling or upscaling tasks, as training them for image generation can be computationally expensive and time-consuming, especially when dealing with high-resolution images. The sequential nature of RNNs can also result in slow generation processes. Recurrent neural networks (RNNs) have feedback loops in their recurrent layer, allowing them to store information in memory over time. However, training RNNs for long-term temporal dependencies can be challenging due to the exponential decay of the loss function gradient [29].

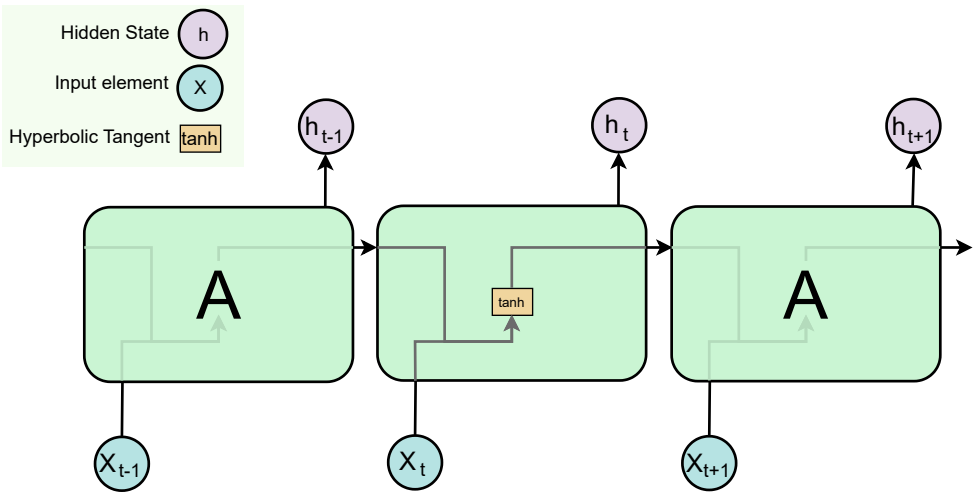


Figure 4.4: The working mechanism of RNN

In the above figure RNN model's repeating module is being explained. Here, a small network is replicated for each time step in the input sequence. It inputs the current input element (X) and the hidden state from the previous time step (ht), capturing the model's memory of previous inputs. The output is a new hidden state (ht+1) and an output value(A), used to predict the current input element. The RNN model consists of a stack of repeating modules, learning long-range dependencies in the data.

Video and photo-up sampling tasks often involve the use of Recurrent Neural Networks (RNNs), but some variants like Long Short-Term Memory (LSTM) networks can be useful for image generation and synthesis. LSTMs can be used to generate sequences of data. This includes generating sequences of images for video or photo up sampling. Each time step in the sequence corresponds to a frame or an image, and the LSTM is trained to generate coherent sequences by learning the temporal dependencies between frames. Video sampling is a process where LSTMs

are used to generate sequences of data, including images, by learning the temporal dependencies between frames. These sequences are then used to create coherent data. LSTMs are advanced machine learning techniques that can be applied to video frame interpolation, generating additional frames between existing frames in a video sequence. These LSTMs can predict intermediate frame content based on adjacent frames.LSTMs, when combined with other architectures like CNNs, can incorporate temporal dependencies into the up-sampling process, such as a convolutional network for spatial features and an LSTM for temporal dependencies in video up sampling. LSTM networks are a type of RNN that use special units, including a

memory cell, to maintain information in memory for extended periods. They address issues by introducing new gates like input and forget gates, which improve control over gradient flow and preserve long-range dependencies. These gates control when information enters, outputs, and is forgotten:

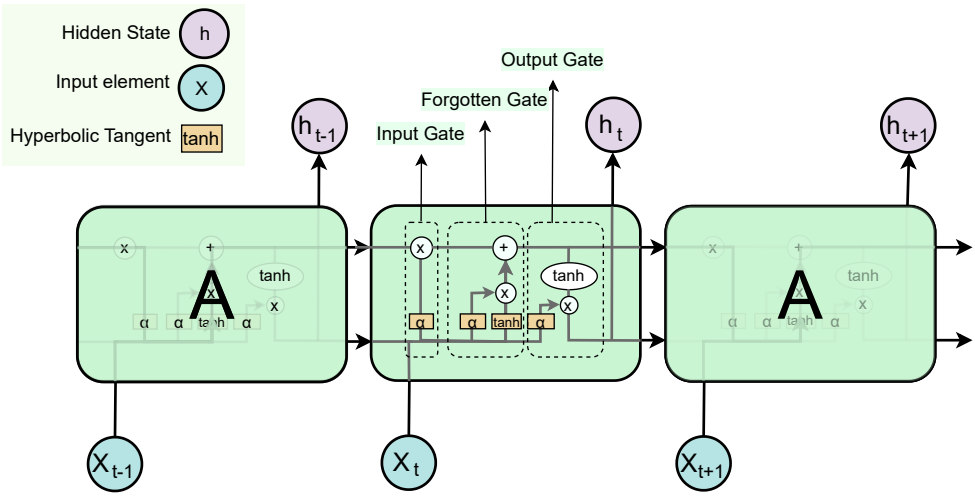


Figure 4.5: The Working mechanism of the LSTM model

So, there are 3 gates in the above picture. Those are the input gate, forget gate, and output gate. Firstly, the input gate decides which information from the current input (X) to add to the cell state. The forgot gate decides which information from the previous hidden state (ht) to forget. Lastly, the output gate decides which information from the cell state to output as the new hidden state (ht+1). The decisions are made based on the generated output (0 or 1) by each gate.

Though LSTM and RNN are not completely similar. There are some core differences between them. The main factor to be considered about when we should use one of those is the main aspect. RNNs are ideal for tasks involving sequences, such as machine translation, where the sequences are sentences or words. However, LSTMs are often used due to their computational efficiency, as they solve the vanishing gradient problem, a problem that standard RNNs struggle with. LSTMs are more computationally effective [21]. There are various applications of LSTMs and video analysis is one of them.

Despite these potential use cases, it's important to note that even LSTMs have their limitations. Especially when dealing with high-dimensional data like images. LSTMs are more commonly associated with sequential data processing and their application to image up-sampling may not be as straightforward as with other architectures. RNNs, particularly LSTMs, are considered for image generation tasks, they are not the primary choice for straightforward image up-sampling due to computational challenges and the availability of more effective architectures like GANs. Convolutional neural networks and generative models are more commonly used.

4.3 Generative Adversarial Network (GNN)

The discipline of generative modeling has seen a change because of the revolutionary class of machine learning models known as Generative Adversarial Networks, or GANs. A new approach for producing fresh data instances that closely mimic an existing dataset is provided by GANs. The interaction between the generator and discriminator neural networks, which are involved in a competitive learning process, is the fundamental concept of GANs.

In this model, there are two sectors. One is the "Generator" and the other one is the "Discriminator". In the GAN architecture, the "Generator" acts as a creative element. The job of the "Generator" is to take in random noise and convert it into data instances that, as closely as possible, resemble the patterns found in the training data [27]. The generator can be an artist trying to create original paintings. It begins with a blank canvas (random noise) and improves its capacity to produce increasingly convincing samples over time through iterative learning.

On the other hand, the "Discriminator" undertakes the position of a critic. Its job is to differentiate between instances of actual data from the training set and artificial data generated by the generator. In the analogy with art, the discriminator plays the role of an art critic attempting to distinguish between skillfully constructed forgeries and authentic masterpieces.

The generator and discriminator engage in continuous back-and-forth throughout the GAN training process. The "Discriminator" wants to improve their ability to discriminate between actual and fake cases, while the "Generator" wants to create data that is identical to real examples. Due to the competitive dynamic created by this adversarial training, both networks are constantly improving and challenging one another to achieve greater performance levels [22].

An objective function that has been carefully designed is the foundation of the GAN architecture. The generator aims to maximize the likelihood that its generated samples would be mistakenly classified as real by the discriminator. On the other hand, the discriminator aims to maximize the accuracy of its distinction between authentic and fraudulent data. This antagonistic goal creates a delicate balance, and careful parametric adjustment is frequently necessary for GAN training to be effective.

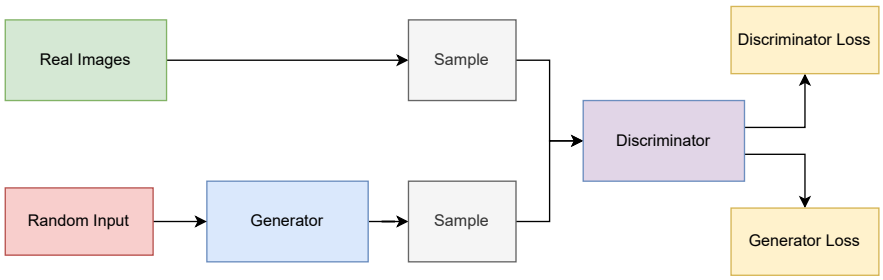


Figure 4.6: Working procedure of GAN model

If we see the equation for the GAN that is:

$$min_G max_D V(D,G) = E_{x \sim Pdata(x)}[logD(x)] + E_{z \sim Pz(z)}[log(1 - D(G(z)))]$$

G is a neural network that generates synthetic data, attempting to mimic the real data distribution. D is another neural network that acts as a binary classifier. It takes input data and tries to distinguish between real data (x) and fake/generated data (G(z)). P(Sub)Data(x) represents the distribution of real data. The generator aims to generate data that is similar to this distribution. Pz(z) represents the distribution of the noise that is fed into the generator. It's a source of randomness that allows the generator to produce diverse outputs. The GAN's objective is formulated as a minimax game between the generator and the discriminator. The goal is to minimize the generator's loss while simultaneously maximizing the discriminator's loss.

The expected value of the log probability that the discriminator assigns to the produced data is represented by the second term, $E_{x \sim Pdata(x)}[logD(x)]$. In order to increase the likelihood that created data will be seen as authentic by the discriminator, the generator aims to reduce this.

The second term $E_{z \sim Pz(z)}[log(1 - D(G(z)))]$ represents the expected value of the log probability that the discriminator assigns to the generated data. The generator

wants to minimize this, making generated data more likely to be classified as real by the discriminator.

The GAN framework has a lot of sub-categories. This research focuses on Super-Resolution Generative Adversarial Networks (SRGAN) and Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN). First and foremost we will be looking into SRGAN.

Consider a low-resolution picture that has a grainy or distorted snapshot. The goal of super-resolution is to incredibly transform a low-resolution image into a detailed, high-resolution counterpart. It's similar to sharpening and clarifying an image by enhancing its quality to display finer details.

Using conventional techniques for image upscaling often generates a little higher resolution, but they are unable to produce realistic, finely detailed images. Introducing SRGAN, a unique type of technology created to make super-resolution more than just adding pixels. Besides, it also aims to produce high-quality, realistic-looking images.

To calculate perceptual loss, SRGANs frequently use neural networks that have already been trained, like VGG networks [24] [5]. These networks are skilled at catching and identifying complicated features because they have been trained on vast datasets for image classification tasks. Making use of this prior knowledge improves SRGANs' capacity to produce more realistic images with a wealth of perceptual aspects. SRGANs are particularly good at producing images with fine features and increased realism since they use pre-trained networks and incorporate perceptual loss. Not only are the produced images clearer, but they also more closely resemble high-resolution material. In short, SRGANs are more specialized and efficient in tasks requiring the creation of detailed, high-quality images from low-resolution inputs because they are designed for super-resolution applications.

By combining novel approaches and improving essential elements, the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) is a significant improvement over the Super-Resolution Generative Adversarial Network (SRGAN). During super-resolution, SRGAN's generator used a deep neural network to enhance image details. It nevertheless lacked a specific structure for recording complex patterns.

In its generator architecture, ESRGAN introduced the Residual-in-Residual Dense Block (RRDB) [24] [8]. The model can capture tiny details and textures in a deeper manner thanks to the more complex and efficient structure of the RRDB. The quality of super-resolved photographs has improved dramatically as a result of this architectural modification.

The RRDB (Residual-in-Residual Dense Block) enhances image quality in deep learning models. It features nested residual blocks, each with dense connections. In simple terms, it preserves information and promotes feature reuse. Residual con-

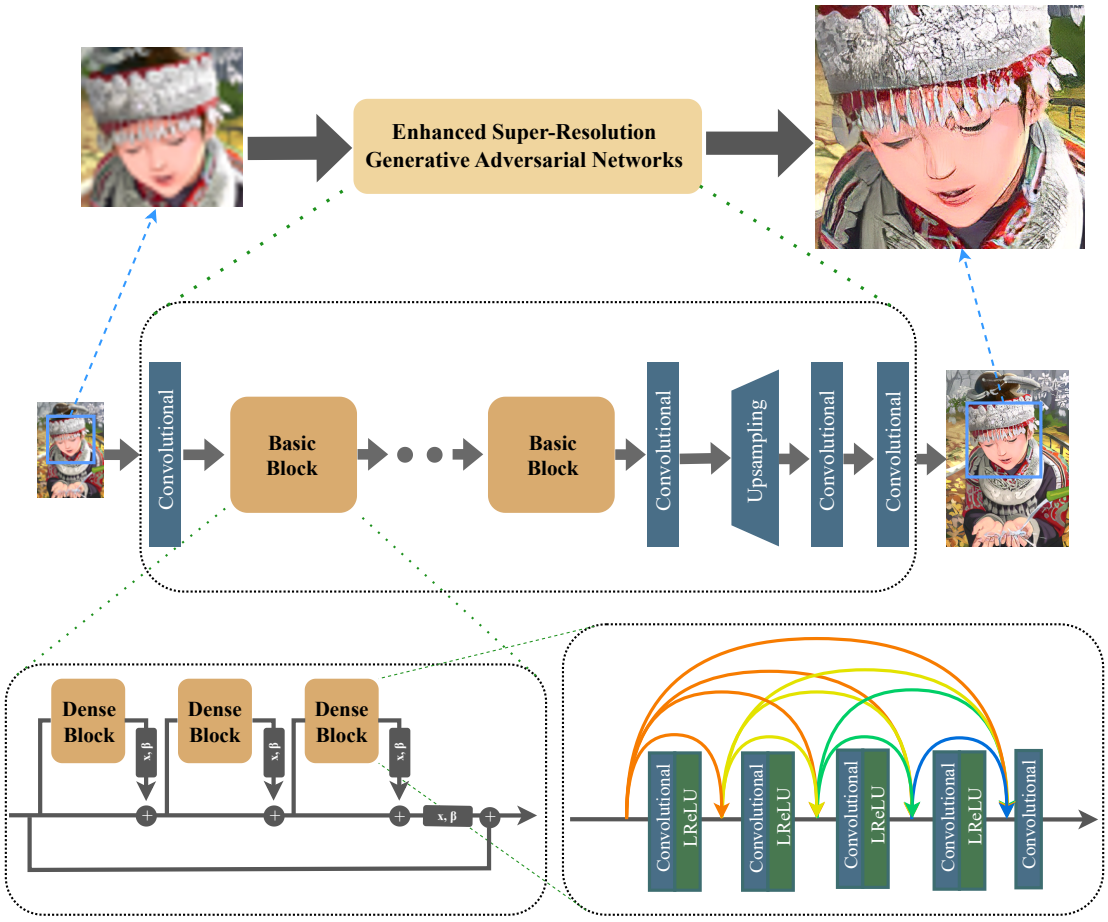


Figure 4.7: ESRGAN working process through the RRDB model.

nections enable direct access to earlier features, preventing information loss. Dense connections connect all layers within a block, aiding gradient flow. This architecture efficiently refines and consolidates features, allowing deep networks to generate high-quality images with improved details and realism.

ESRGAN goes beyond SRGAN's introduction of the idea of perceptual loss to improve realism. ESRGAN improves perceptual loss functions by prioritizing the preservation of critical visual characteristics more strongly [25]. For the perceptual loss, VGGNet-19 is used. The working mechanism of VGGNet-19 was explained earlier in the research. ESRGAN generates super-resolved images that closely mimic real high-resolution content while also appearing sharper thanks to its more efficient use of perceptual loss, which results in a more organic and beautiful final product. ESRGAN is made to take into account various upscaling factors. This increases its versatility in handling a range of scaling requirements by enabling it to adjust to different super-resolution levels. ESRGAN shows flexibility in its applications, whether it is upscaling small photos for greater visual clarity or increasing features in larger images.

In conclusion, ESRGAN outperforms SRGAN by implementing a more sophisticated RRDB architecture, improving perceptual loss functions, and showcasing adaptability while managing various upscaling factors. All of these enhancements add up to a better ability to produce realistic, detailed, and high-quality super-resolved photos.

Till now this paper discussed relative topics regarding image enhancement scenarios. Now the focus is going to be shifted to Videos. Videos are more complex than images as they have numerous numbers of frames. Videos are made up of frames, and it's critical to keep the timing consistent. Upscaling each frame separately may produce jerky or strange movements. To achieve smooth and coherent motion, video upscaling algorithms need to consider the temporal correlations between frames. Videos generally carry a lot more info than individual photos. Each frame of a video represents a picture, and there are sometimes numerous frames every second. Processing and scaling such a vast volume of data in real-time or near real-time can be computationally demanding.

4.3.1 Frame Recurrent Video Super Resolution (FRVSR)

To outcome these challenges, various types of models are already being used in various video upscaling techniques. such as FRVSR ,SSL and BSRNET. Amongst these, FRVSR uses frame recurrence and generative adversarial networks. The main idea behind FRVSR is to encourage temporally consistent results and reduce computational costs. Instead of processing and warping each input frame multiple times, FRVSR warps only one image in each step [7]. This approach not only reduces computational demands but also improves temporal consistency as it uses information from multiple similar adjacent frames (both future low-resolution frames and previous super-resolution estimates), in addition to the current frame. But one downside to FRVSR is that it needs lots of data to train.

4.3.2 STRUCTURED SPARSITY LEARNING (SSL)

Next is Learning with Structured Sparsity (SSL). In VSR models, the SSL approach is used to prune unnecessary filters. The application of VSR models on devices with limited resources, such as drones and smartphones, may be impeded by their high computational costs. Many redundant filters are present in existing VSR models, which can significantly reduce inference efficiency. In addition to SSL, pruning methods are developed for several crucial VSR model elements, including residual blocks, recurrent networks, and up-sampling networks. However, SSL in VSR requires extremely high computational expenses, which are frequently difficult or impossible to manage.

4.3.3 Blind Super-Resolution Network(BSRNET)

Then comes BSRNET. It stands for Blind Super-Resolution Network and it is a model used for image super-resolution tasks. In the context of video super-resolution, BSRNet might improve the quality of low-resolution videos. However, the actual functioning method of BSRNet in video super-resolution may vary depending on the specific implementation and nature of the video data. The primary concept underlying super-resolution models such as BSRNet is to develop a mapping from low-resolution inputs to high-resolution outputs. This is often accomplished using a deep learning framework that employs vast quantities of training data to learn this mapping. The trained model may then be applied to improve the resolution of previously unnoticed low-quality photos or videos. However, the main problem of implementing BSRNET in video upscaling tasks is, that BSRNET is best suited to image upscaling tasks. To use BSRNET in the context of video upscaling tasks, much computational data may be needed [6].

4.3.4 Temporally Coherent Generative Adversarial Network (TECOGAN)

To overcome these challenges this paper focuses on a new type of GAN architecture, and that's called TecoGAN. The phrase "Teco" in TecoGan refers to TEmporally COherentGAN. This whole TecoGAN concept is based on the self-supervision of GAN. Also, TecoGAN's are easy to use compared to VSRNET or other Video super-resolution methods. One crucial reason for this is TecoGAN needs a low amount of pre-trained data compared to other models. Also, this model employs adversarial learning across time. By doing so, it ensures that the generated frames are not only visually appealing but also consistent over time. But one thing to consider while implementing TecoGAN is the usage of the loss function [11]. TecoGAN introduces a novel loss function called the ping-pong loss. This loss encourages the model to create results that are free of temporal artifacts. Essentially, it prevents flickering or jitters in the generated videos.

Chapter 5

Data Analysis

5.1 Dataset Description

This thesis is working on a vast dataset of images (.jpeg, .img, .png, .jpg) and videos (.mp4, .gif). In the case of images for the training purposes the images are classified under various categories. The different categories of images and videos are used for the purpose of training the model.

In the dataset, we have taken around 4239 images and 250 videos. For the images, different categories have been used. Classification is important for CNN models as CNN works on features and gives the probability of identifying the image. But in the case of Super Resolution, synthetic data or pixel generation is more important rather than the classification of the preliminary data. Here the primary data are categorized for pre-processing and properly analyzing. These are Astronomy, CCTV Footage, Human Faces, Low Light, Nature, Pets, Text, Texture, and Vehicles collected from different datasets and sites like DIV2k, REDs and Kaggle. The table below consists of the categories and their respective image numbers.

Categories	Number of Images
Astronomy	39
CCTV Footage	559
Human	420
Low light	897
Nature	800
Pets	200
Text	516
Texture	232
Vehicles	576
Total	4239

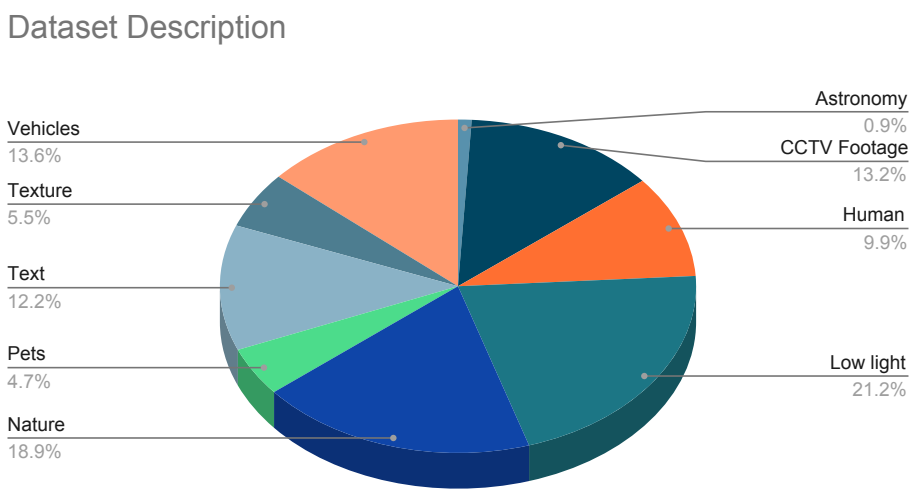


Figure 5.1: Category based data description.

5.2 Dataset Preprocessing

The ESRGAN model requires a set of pair images for training purposes. The initial dataset consists of high-resolution images which can be considered the ground truth (GT). To train the model a low-resolution (LR) set of the data is required. To achieve the low-resolution images OpenCv library was used to convert the main dataset into a four times low-resolution (4x LR) dataset. This low-resolution dataset will go through the ESRGAN model where the generated data will be compared in the discriminator part with the main high-resolution (HR) dataset as validation set. Based on the generator and discriminator loss the model will be trained.

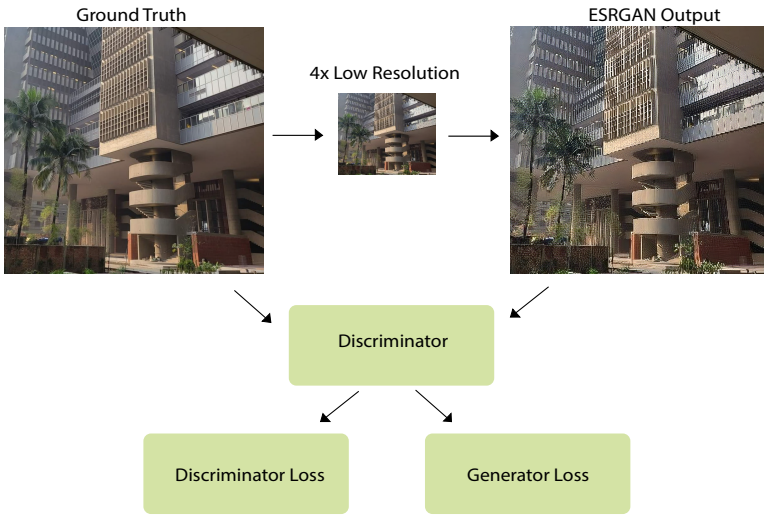


Figure 5.2: Data pair training process.

5.2.1 Primary Implementation and Results

Using ESRGAN pretrained model a low resolution 62×90 pixel image converted into super resolution form by two steps. In each step the resulotion upraised four times which gives a 990×1440 pixel output.

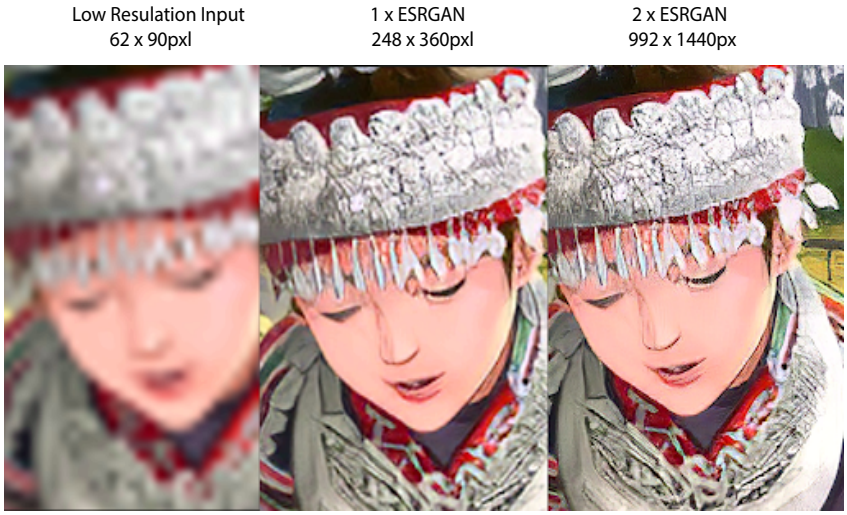


Figure 5.3: ESRGAN Model testing for 8X SR.

For initial testing of the model some data from DIV2k dataset has been used. Here some of results are given to prove the further possibilities of ESRGAN. In this process the RRDB model of ESRGAN requires huge amount of visual ram to process the data. So, according to the study and depending on the resolution the moderate up-scaling is 4x.

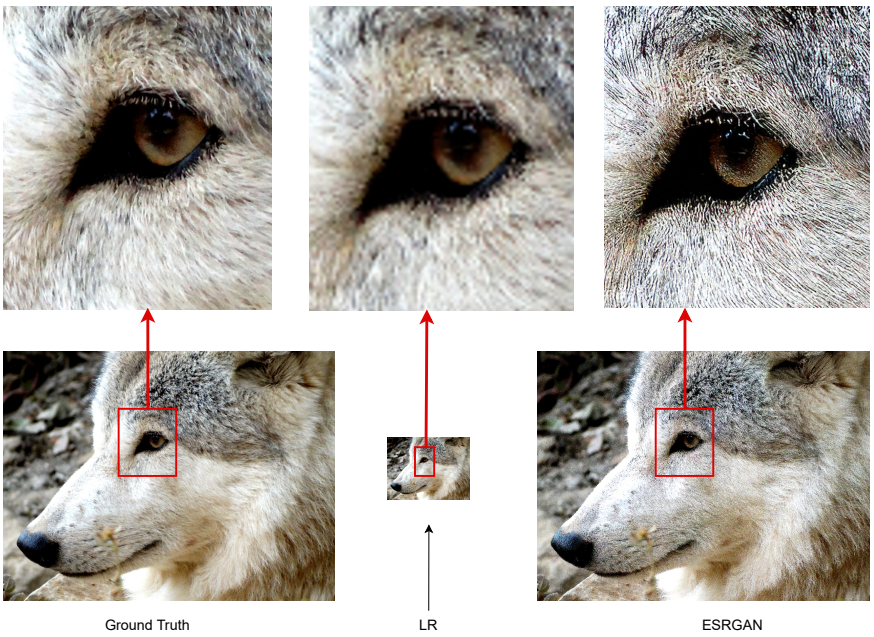


Figure 5.4: Model testing for single step 4x..

Chapter 6

Conclusion

In the world of photographs and Videos, where clarity and detail matter, our research on deep learning algorithms for super-resolution has uncovered a route to dramatic visual improvement. Through the synergy of CNNs, RNNs, and GANs, we've observed the potential of technology to breathe new life into the ordinary.

After surveying different papers on Super Resolution where CNN, RNN and GAN models were implemented, the GAN is leading for it's generative ability which is required for SR. Analyzing the models for Single Image Super Resolution ESRGAN and for Video Super-Resolution TecoGAN is dominating over other models. The existing models of ESRGAN has been tested and the results proven the possibilities towards this thesis. Based on the models the primary dataset has been developed and preprocessed to fulfill the pair base testing requirments. From improving single photos to enhancing video quality, this strategy has broad implications across numerous areas, which is visible in our everyday lives. Our venture into upscaling has proved that, with the correct tools or algorithms, we can expand our digital experiences and enable older devices to go beyond their boundaries.

As we come to the end of this paper, we embrace the combination of artificial intelligence and creativity, opening possibilities for a future where every pixel tells a clearer, more colorful story. With deep learning as our guide, we encourage eveyone to visualize a future where visual perfection knows no limitations.

Bibliography

[1] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology and Climatology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[2] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE transactions on acoustics, speech, and signal processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[3] Y.-N. Liu, Y.-C. Lin, Y.-L. Huang, and S.-Y. Chien, "Algorithm and architecture design of high-quality video upscaling using database-free texture synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 7, pp. 1221–1234, 2014.

[4] W. Shi, J. Caballero, F. Huszár, *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[5] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[6] S. Y. Kim, J. Lim, T. Na, and M. Kim, "3dsrnet: Video super-resolution using 3d convolutional neural networks," *arXiv preprint arXiv:1812.09079*, 2018.

[7] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6626–6634.

[8] X. Wang, K. Yu, S. Wu, *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[9] Z. Wang, P. Yi, K. Jiang, *et al.*, "Multi-memory convolutional neural network for video super-resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2530–2544, 2018.

[10] S. Bansari, *Introduction to how cnns work*, 2019.

[11] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, "Learning temporal coherence via self-supervision for gan-based video generation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 75–1, 2020.

[12] U. Cisco, "Cisco annual internet report (2018–2023) white paper," *Cisco: San Jose, CA, USA*, vol. 10, no. 1, pp. 1–35, 2020.

[13] J. Liang, *Image classification based on resnet*, 2020.

[14] J. C. Martinez, *Introduction to convolutional neural networks cnns*, 2020.

[15] L. Alzubaidi, A. J. H. Jinglan Zhang, O. A.-S. Ayad Al-Dujaili Ye Duan, M. A. F. J. Santamaría, and M. A.-A. L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of Bigdata*, vol. 8, no. 53, 2021.

[16] E. Chan, "Pack more megapixels into your photos with adobe super resolution," *Adobe Blog*, 2021.

[17] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.

[18] M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang, and A. Mallya, "Generative adversarial networks for image and video synthesis: Algorithms and applications," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 839–862, 2021.

[19] F. Lundkvist, *Deep upscaling for video streaming: A case evaluation at svt.* 2021.

[20] N. Tovar, *Real-Time Image Upscaling with Commonly Available Resources.* California State University, Long Beach, 2021.

[21] P. Gopalani, *Lstm vs rnn confusion cleared*, 2022.

[22] *How to use gan to generate images*, 2022.

[23] A. Rakvi, J. Shah, P. Singh, and S. Malik, "Super resolution of videos using e-gan," *Available at SSRN 4012374*, 2022.

[24] R. Sarode, S. Varpe, O. Kolte, and L. Ragha, "Image super resolution using enhanced super resolution generative adversarial network," in *ITM Web of Conferences*, EDP Sciences, vol. 44, 2022, p. 03 054.

[25] J. Song, H. Yi, W. Xu, X. Li, B. Li, and Y. Liu, "Dual perceptual loss for single image super-resolution using esrgan," *arXiv preprint arXiv:2201.06383*, 2022.

[26] X. Xiang, Y. Tian, V. Rengarajan, L. D. Young, B. Zhu, and R. Ranjan, "Learning spatio-temporal downsampling for effective video upscaling," in *European Conference on Computer Vision*, Springer, 2022, pp. 162–181.

[27] B. Zhang, S. Gu, B. Zhang, *et al.*, "Styleswin: Transformer-based gan for high-resolution image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 304–11 314.

[28] M. Mandal, *Introduction to convolutional neural networks (cnn)*, 2023.

[29] S. Poudel, *Recurrent neural network (rnn) architecture explained*, 2023.

[30] G. Boesch, *Vgg very deep convolutional networks (vggnet) - what you need to know.*

[31] *FFmpeg a complete, cross-platform solution to record, convert and stream audio and video.* https://ffmpeg.org/, Accessed: 2023-09-18.

[32] G. F. Freedman, "R. 2011. image and video upscaling from local self-examples," *ACM Transaction on graphics*,

[33] *NVDIA dlss 3*, https://www.nvidia.com/en-us/geforce/technologies/dlss/?nvid=nv-int-gfhm-55050, Accessed: 2023-09-18.

[34] C. Shorten, *Introduction to resnets*.

[35] *What Is Deep Learning? — How It Works, Techniques & Applications — mathworks.com*, https://www.mathworks.com/discovery/deep-learning.html, Accessed: 13-01-2024.