SUPER-RESOLUTION ASSESSMENT AND DETECTION

ENRIQUE LÓPEZ CUENA

Thesis supervisor: DARIO GARCÍA GASULLA (Department of Computer Science)

Thesis co-supervisor: ADRÍÁN TORMOS

Degree: Master Degree in Artificial Intelligence

Thesis report

School of Engineering
Universitat Rovira i Virgili (URV)

Faculty of Mathematics
Universitat de Barcelona (UB)

Barcelona School of Informatics (FIB)
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

28/06/2023

Abstract

Super Resolution (SR) techniques are powerful digital manipulation tools that have significantly impacted various industries due to their ability to enhance the resolution of lower quality images and videos. Yet, the real-world adaptation of SR models poses numerous challenges, which blind SR models aim to overcome by emulating complex real-world degradations. In this thesis, we investigate these SR techniques, with a particular focus on comparing the performance of blind models to their non-blind counterparts under various conditions. Despite recent progress, the proliferation of SR techniques raises concerns about their potential misuse. These methods can easily manipulate real digital content and create misrepresentations, which highlights the need for robust SR detection mechanisms. In our study, we analyze the limitations of current SR detection techniques and propose a new detection system that exhibits higher performance in discerning real and upscaled videos. Moreover, we conduct several experiments to gain insights into the strengths and weaknesses of the detection models, providing a better understanding of their behavior and limitations. Particularly, we target 4K videos, which are rapidly becoming the standard resolution in various fields such as streaming services, gaming, and content creation. As part of our research, we have created and utilized a unique dataset in 4K resolution, specifically designed to facilitate the investigation of SR techniques and their detection.

Contents

List of Figures

List of Tables

1
Introduction

Digital content manipulation techniques, such as deepfakes, automatic colorization, or generative models, have garnered substantial attention in recent years. They have notably improved in quality and found numerous practical applications across diverse industries. Among these techniques is SR, which essentially aims to increase the resolution of lower quality images or videos and enhance the fine details that are missing in the Low-Resolution (LR) source. Image enhancing applications are successfully applied in medical imaging [4] [118] [19], security camera image footage [2] [71], remote sensing tasks [123] [30], gaming [75], and the entertainment industry [98] [86].

Manipulation techniques can be categorized into fully-synthesized and partial manipulation, according to the extent of modifications made to the original data. Full-synthetic techniques involve the generation of entirely synthetic visual content, which have significantly gained popularity due to the advances in Generative Adversarial Network (GAN) [107] and Latent Diffusion Model (LDM) [22]. Conversely, partial manipulation techniques involve altering specific regions of existing images or videos. The most notable examples in this area are deepfakes [6], which manipulate visual content in existing images, videos, or audio. Super-Resolution (SR) methods can fall into either of these categories, depending on the upscaling process.

Super-resolution finds its ultimate and most relevant application in real-life scenarios, where the input data could have numerous sources. The premise of SR lies in enhancing the quality and resolution of visuals - a valuable resource when the sources are as diverse as images captured by a variety of phone devices or surveillance cameras. In this context, SR can be utilized to enhance the visual experience, as well as to reduce the cost of data storage and transfer. This can be achieved by storing the content at low-resolution and then upscaling it to high-resolution on each display device. For their use in such general-purpose applications, SR models should be able to achieve good performance on a wide range of real-world degradations, such as motion blur, noise, artifacts produced by mobile phones, etc.

Training a supervised SR model requires of paired -HR (High-Resolution) samples that belong to the same image. The goal of the model is then to learn how to create the glshr version given only the input. To obtain these -HR pairs, most contributions downscale a set of the original glshr images synthetically (i.e. using a specific algorithm like bicubic interpolation). However, this introduces a bias, as it does not fully replicate the complex and varied degradation processes that occur in real-world low-resolution image acquisition. As a consequence, models trained under these conditions may exhibit limitations when dealing with real-world images due to the discrepancy between synthetic and real-world degradations.

The operation to obtain images is a key factor that heavily impacts the ability of the model to perform in real-world scenarios. If the same predefined downsampling operation is applied to generate all training samples, the model will specialize in that specific downscaling process.

As an alternative, a more robust approach would be to train the model with a more complex downsampling (or *degradation*) method. This way, the model could potentially adapt to various types of inputs, enhancing its performance in diverse scenarios.

Given that premise, *real*, also known as *blind* SR models propose a more complex approach towards generating the training image pairs by emulating real-world degradations. The field of blind-SR has drawn considerable attention in recent years, resulting in extensive research and significant advancements. As a result, the majority of today's models employ a more sophisticated method to obtain image pairs that are closer to real images. Alternatively, there are blind models which utilize unique *Real* datasets that already include original glshr and glslr images, that is, images where both the glshr and glslr versions are directly obtained by a recording device. This is possible by setting up a special camera setup that allows to capture different resolution images of the same content. Using two cameras or applying zoom represent two instances of this methodology.

Modern SR models, particularly those focused on Video Super-Resolution (VSR), generally require high computational resources. This requirement is further amplified by the increasing tendency towards high-resolution 4K content. According to the Visual Networking Index by Cisco [21], it is estimated that, by 2023, two-thirds of the installed flat-panel TV sets will be UHD. Video devices, in particular, can have a multiplier effect on traffic. A high-definition television with Internet capabilities, streaming approximately two to three hours of content daily, would generate the same amount of Internet traffic as an average household does in the present day. As a result, many on-demand and streaming platforms have turned to SR techniques to upscale their content to 4K. SR algorithms have allowed content that was originally in a smaller resolution to meet the viewer's expectations for detail and clarity, considering the limitations of SR networks.

This tendency has led to a surge of interest in the field of SR detection and also created a new set of challenges that threaten the authenticity of visual media, especially after the recent and socially impactful development of generative models. LDMs for image and video, image-to-image, text-to-image, and advanced SR methods based on Deep Learning (DL) are an example of the recent IA innovation in the digital content field. Digital forgeries, ranging from elementary manipulations like object cloning or removal to complex alterations involving deepfakes and SR pose substantial issues across different sectors, including digital forensics, cybersecurity, the legal system, media veracity, and privacy. Therefore, developing effective and reliable *forgery* detection mechanisms has become paramount. The urgency of this research line is clearly motivated by the widespread availability of these techniques through popular applications like Adobe's Photoshop[1], Deepfacelab [2], and TopazLab[3].

In a historical context, forgery detection methods have primarily targeted common manipulations such as copy-move(copying parts of an image and then pasting them into the same image), splicing (portions of an image come from another image), and deletion. However, with the rapid development of Artificial Intelligence (AI), the amount of digital forgery techniques has expanded. One of those manipulations is SR, which presents significant benefits, such as preserving and restoring old footage or enhancing the quality of video in resource-limited contexts like medicine, but can also be misused. Ethical concerns include the use

[1] `https://www.adobe.com/creativecloud/photography/discover/photo-manipulation.html`

[2] `https://github.com/iperov/DeepFaceLab`

[3] `https://www.topazlabs.com/`

of synthetic glshr images for decision-making in sensitive domains like radiology or law enforcement. Legal concerns include the incoming AI Act [81], which enforces the disclosure of synthetic images.

In conclusion, SR methods alter the original digital content and can potentially distort the truth by misrepresenting it, highlighting the need for robust SR detection mechanisms. Moving forward, this thesis aims to explore several research challenges in SR. First, to validate the applicability and impact of SR models, it investigates the effect of the downsampling method on the output of non-blind and blind SR methods. Secondly, it explores the current landscape for high-resolution upscaled content detection and proposes an extension over the existing methods to evaluate the discriminability of artificial glshr videos.

2
Related work

The following section is dedicated to providing an overview of indispensable concepts found in the literature about SR, that help understand the problem at hand and give the necessary context for the rest of the thesis. The content of the section is inspired by recent surveys about various SR or SR-related topics [49] [41] [7] [18] [101] [122] [96] [51].

In order to present a more comprehensive view of the various topics or areas of SR addressed in this thesis, we have constructed a summary graph (see Figure 2.1). Our objective is not to display a complete breakdown of every field within SR but to include a high-level overview that describes the content discussed throughout the document.

Super-resolution is a fundamental field in the low-level vision area that aims to recover a High-Resolution (HR) image or video from its Low-Resolution (LR) counterpart. This process generates an glshr output with increased resolution and improved visual quality by enhancing the structure and finer details present in a glslr input. SR can be considered an image or video manipulation technique since it essentially involves modifying the original data. Interpolation methods can introduce new pixels and alter existing ones, while deep neural networks generate the glshr output from a deep representation of the input image through hierarchically structured layers.

There are two main categories of Image Super-Resolution (ISR) techniques, Single-Image Super-Resolution (SISR) and Multi-Image Super-Resolution (MISR). The former methods focus on upscaling a single glslr image, while the latter combines information from multiple glslr images, a process that has naturally evolved into video super-resolution (VSR). VSR methods exploit information from neighbor frames to reconstruct the target frame, thereby utilizing extra information that is not present in SISR. For ISR, SISR is by far the most popular method, due to its higher efficiency and the difficulty to obtain multiple images for the same target. However, both of them have their own unique challenges, such as the single reference limitation in SISR and motion estimation, motion compensation, and handling temporal consistency in VSR. In this context, understanding the process behind SISR is key to developing and exploring VSR models, as videos can be understood as sequences of frames that can be processed individually or as a group.

2.1. Problem Formulation

SR is an ill-posed problem, as there exist multiple possible glshr images that could correspond to a given original glslr image. During the process of scaling an image, some details are lost and cannot be recovered. For example, if a glslr image contains a human face with poor detail, it might not be clear whether the person was frowning or smiling. When trying to upscale that image, either option could be plausible, leading to non-uniqueness in solutions.

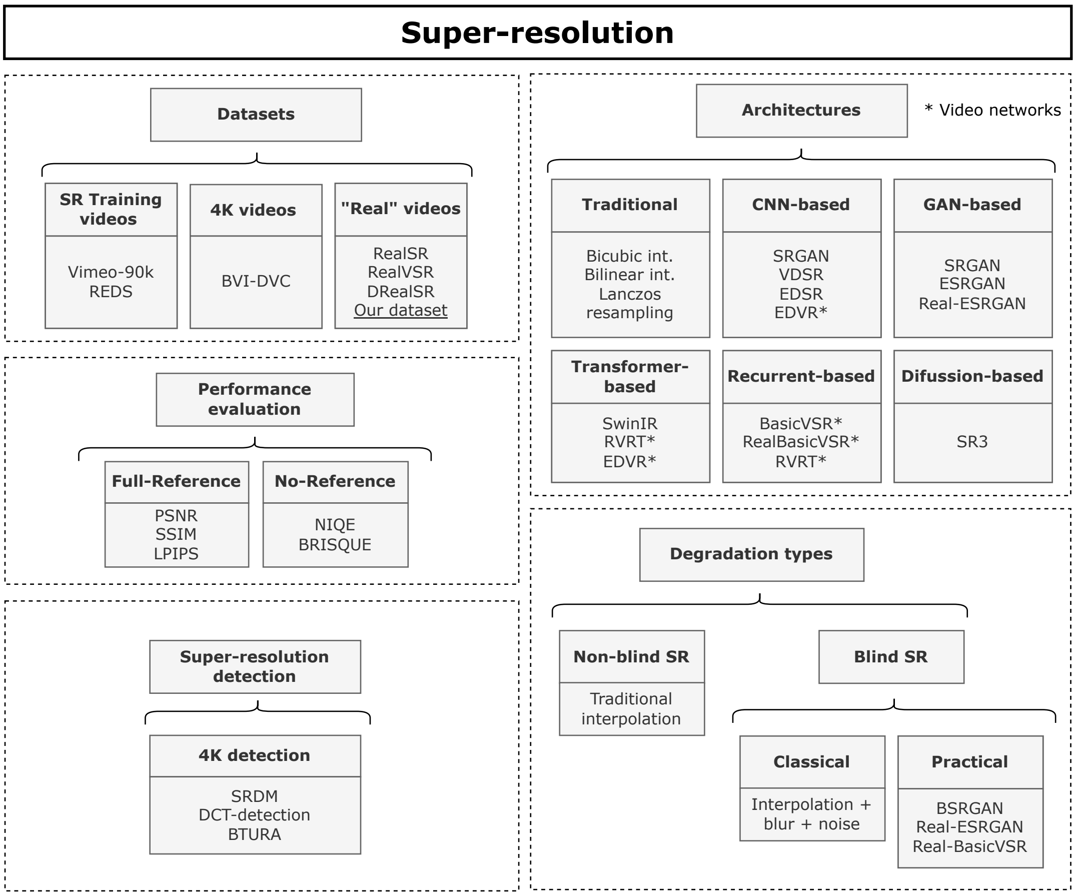


Figure 2.1: Organized diagram of relevant SR-related topics that are mentioned or explored through the thesis

Thus, the upscaling process is not the retrieval of an exact image, but rather an example of one of the many plausible images from the high-resolution and ”realistic” domain.

Building and training a SR model usually involves a supervised learning approach, where a dataset containing pairs of high-resolution (HR) and low-resolution (LR) images is needed. These pairs serve as ground truth and input images to the network respectively, as depicted in Figure 2.2. Original glshr images have a corresponding glslr version at a predefined scale (e.g. x2, x4), which are used as input glslr images. The SR model will then upscale the glslr input images to their original size and update the network's parameters to minimize a loss function. The loss function measures the discrepancy between the output generated by the network and the original glshr image. A common loss function used for SR is Mean Squared Error (MSE), which compares the sum of pixel-wise Euclidean distance between image pairs to obtain a final score. The MSE is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_{\text{HR}}(i) - I_{\text{SR}}(i))^2 \tag{2.1}$$

Where $I_{\text{HR}}(i)$ refers to the pixel intensity at the i-th position in the original high-resolution image and $I_{\text{SR}}(i)$ is the corresponding pixel intensity in the super-resolved image. Finally, N is the total number of pixels in the image.

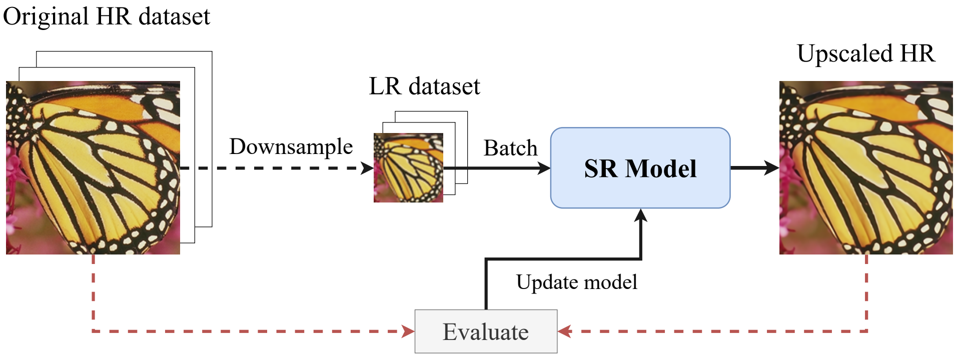


Figure 2.2: Supervised training scheme for SISR

2.2. Degradation Process in Super-resolution

The process of generating glslr images from glshr ones involves a series of modifications known as the degradation method. The degradation of the glshr image could include blurring or adding noise before the downsampling operation. It can be expressed as $y = f(x; s)$ where y is the resulting glshr image, x is the input glslr image, and f is the degradation function with a scale factor s.

Most existing methods assume a pre-defined degradation process (e.g., bicubic downsampling [36]) from an glshr image to an glslr one. In bicubic downsampling or interpolation, a cubic function is applied to a grid of pixels in the original image to calculate the pixel values in the downscaled image, resulting in a smooth glslr representation of the original. Assuming a specific traditional downsampling method to generate the glslr images is convenient and easy to calculate, but can hardly hold true for real-world images with complex degradation types. In the literature, the domain where degradations are assumed is referred to as *non-blind* SR, where the degradation model is known and accurately characterized, so the SR problem turns into the modeling of a function that reverses that known process to recover the glshr image.

2.2.1. Non-blind SR

The most simple function in non-blind SR assumes f to be bicubic downsampling or interpolation, although various other interpolation methods exist to fulfill the same task:

$$\boldsymbol{y} = \boldsymbol{x} \downarrow_s^{bic}, \tag{2.2}$$

or a more complex approach defined by:

$$\boldsymbol{y} = (\boldsymbol{x} \otimes k_g) \downarrow_s, \tag{2.3}$$

Where $\otimes$ indicates a fixed convolutional operation with a Gaussian blur with kernel k_g. Constructing a large LR-HR pair dataset to train a SR model is challenging, so researchers often opt to use simple degradation patterns to create said datasets, both in image and video SR [11] [33].

However, there is a substantial problem with non-blind degradations. A SR model that is trained with a fixed degradation will only be able to handle inputs with the same characteristics and will struggle to produce satisfactory results for other glslr images with different degradations. [49]

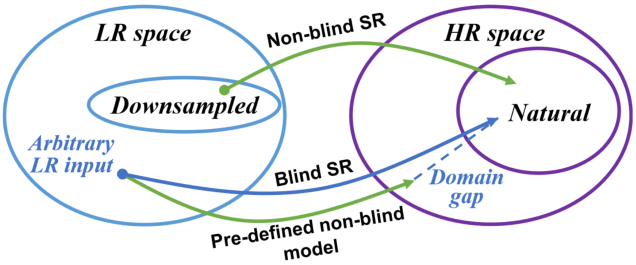


Figure 2.3: Domain interpretation of differences between non-blind and blind SR. Source: 'Blind Image Super-Resolution: A Survey and Beyond'[49]

As illustrated in Figure 2.3, a pre-defined degradation function confines the domain of glslr training samples to a reduced space. Using a model trained on that downsampled space on an arbitrary glslr input from the real world causes the non-blind model to generate unrealistic images. This mismatch produces a domain gap between the actual SR output and the desired output, which should ideally be closer to the glshr natural image domain and contributes to the suboptimal quality of the super-resolved outputs.

The limitations of non-blind models are worrisome given the diversity of imaging devices. Smartphones, DSLR (digital single-lens reflex) cameras, and surveillance devices greatly differ in the characteristics of taken images, producing a high range of quality content. Most video content today comes from smartphone cameras, that process image and video through a digital signal processor on the chip, involving several steps, such as pixel correction, white balance correction, denoising, and sharpening, which introduce unknown and complex degradations to the content. Another relevant factor is the transmission of digital content through the internet, where it can be subject to several compression types. As an illustration, a low quality-image downloaded from the internet may have undergone a complex and mixed set of degradations. The original image could have been taken with any mobile device a long time ago, introducing blur, sensor noise, and compression artifacts. Furthermore, post-processing activities such as sharpening and resizing create additional artifacts, which are amplified by the digital transmission of the image.

2.2.2. Blind SR

To address the domain gap in non-blind SR and to model such a complicated deterioration sequence, blind SR degradation proposes an approach based on unknown degradations. This way, the model learns the degradation process in addition to learning how to reconstruct the glshr image, making it more suitable for real-world SR scenarios, as training glslr images are more faithful to the real world. The degradation model is designed to simulate complex degradations caused in real images, as a consequence of various factors such as sensor noise, compression artifacts, motion blur, aliasing...

Blind SR methods can be divided into classical and practical modeling, and often revolve around the application of three common operations: blur, noise, and downsampling:

2.2.2.1. Classical Models

Classical degradation, widely adopted in explicit modeling, refers to approaches that incorporate a predefined mathematical model to generate a glslr image from the glshr one. On the contrary, implicit models utilize data distribution learning to simulate real degradations, generally employing adversarial deep learning architectures to learn the degradation models. Two great examples are CinCGAN [119] and DASR [111]. Noticeably, these blind SR methods do not generalize well to out-of-distribution images, as they are limited to degradations within the training dataset.

The classical degradation model is based on an extension of the non-blind proposal, consisting of a more complex degradation defined by:

$$\boldsymbol{y} = ((\boldsymbol{x} \otimes \boldsymbol{k}) \downarrow_s + \boldsymbol{n})_{JPEG_q} \tag{2.4}$$

Where the blur kernel k and the added noise n will have unknown parameters to generate the glslr images used to train the SR model. Some models also apply a JPEG compression operation with an unknown quality factor q. Other families of methods [125] [132] make use of external datasets to approximate the kernel and noise values, or leverage the internal statistics within a single image to model the degradation process [89].

Classical models based on an explicit combination of degradations are more versatile than non-blind methods and generalize better to different imaging conditions. Nonetheless, they are still not able to properly simulate complex real-world degradations.

2.2.2.2. Practical Models

Contrastingly, and as a step forward towards handling more realistic degradations, practical models simulate a complex combination of various degradation factors observed in real-world imaging. There has been a substantial increase in the utilization of practical models over recent years, leading to state-of-the-art results in the blind SR domain. An example of a widely used practical model is BSRGAN (Blind Super-Resolution Generative Adversarial Network) [126].

BSRGAN

BSRGAN was presented as the first applicable model for general blind SR, employing a practical degradation model to train deep blind SISR methods for real applications. The common classical blind SR models tackle blur, noise, and downsampling in a sequential manner. These operations provide a simple way of obtaining degraded glslr images. Applying these transformations with random parameters adds more diversity in terms of degradations, increasing the robustness of the model and its adaptability to more types of degradations.

BSRGAN proposed a cascade framework to better emulate real degradations, by increasing the degradation space of the three key factors and adopting a random shuffle strategy. Instead of using the commonly-used blur/downsampling/noise-addition pipeline, randomly shuffled degradations are performed to synthesize glslr images. The degradation space is increased through the use of random parameters, which allow the generation of more diverse content. The random shuffle strategy eliminates the fixed sequence in the typical degradation process, introducing more variation and unpredictability in image synthesis.

Each degradation type is characterized by random parameters. Particularly, the blur operation can use two convolutions types from both the glshr space and glslr space (before and after downsampling the ground truth glshr image). All blur kernel settings are randomly sampled from a set list of options to maximize the degradation space.

To generate the glslr image from the glshr version, a downsampling operation is uniformly sampled from nearest-neighbor, bicubic or bilinear interpolation, and lastly, a down-up combination. It first downsamples an image with a scale factor sa and then upscales it with scale factor a sampled from [1/2,s].

Concerning noise, BSRGAN considers JPEG compression noise and camera sensor noise, on top of the common Gaussian noise. In essence, the random sequence of degradations with different parameters allows BSRGAN to produce multiple glslr versions from the same glshr image.

Following the described configuration, BSRGAN can be employed to train deep blind SR networks using a set of paired LR-HR images, by producing unlimited degraded and aligned glslr samples, bypassing the data limitation that is prominent in blind SR. Nonetheless, it can produce glslr images with degradations that very rarely happen in the real world [126]. Regardless, BSRGAN proved to be an effective way of training existing networks in a blind SR setting, by producing the necessary training data. At the time of its publication, this model was the first work to adopt a new hand-designed degradation model for general blind image SR, and it has been successfully adopted in state-of-the-art blind SR models [44].

Real-ESRGAN

The second example of a popular practical model is an extension of the ESRGAN [104] network, to restore real-world images by covering a broad range of real-world glslr images. In Real-ESRGAN [105], the classical degradation model (blur/downsampling/noise/compression) referred to as ”first-order” is enhanced to a ”high-order” degradation model for real-world scenarios. A *second-order degradation process* is proposed for simplicity and effectiveness, meaning that the classical degradation model is applied twice. Notably, the two passes use different hyper-parameters, and the downsampling operation is designed to keep the image resolution at a reasonable range.

Similar to BSRGAN, the blur operation is performed by isotropic and anisotropic Gaussian filters, with random kernel parameters. Two types of noise functions are employed, additive Gaussian noise, with varying intensity, and Poisson noise, in order to account for the sensor noise caused by the variation in the number of photons sensed at a given exposure level. The downsampling operation is randomly chosen between bicubic interpolation, bilinear interpolation, and area resize operation shown in the publication. Lastly, images go through a JPEG compression with a random quality factor.

Gated degradation system More recent degradation models have been introduced, such as the one referenced in [129], which highlights the inability of BSRGAN and Real-ESRGAN to deal with easy degradation cases while they achieve great results in complex situations. [129] proposes a unified degradation model by introducing a gate mechanism to randomly select the base degradation types to be included in the degradation process, covering important corner cases that are prevalent in the real world. The presented model is suitable for non-blind SR, classical blind SR, and practical blind SR by introducing a gate controller to generate various combinations of base degradation types, that are the same as in BSRGAN and Real-ESRGAN (Gaussian blur, additive Gaussian noise, JPEG compression, and interpolation based downsampling operations).

There is an additional fundamental approach for obtaining training pairs for blind SR. The concept involves using a special camera setup that allows the production of original glshr and glslr images or videos. There exist various image and video datasets with paired glslr and glshr images or frames, captured directly with no downsampling involved, which allow training real SR networks with original degradations. Existing *real* datasets, techniques and limitations are discussed in chapter 4.

In conclusion, the process to generate the LR-HR pairs to train a SR is a key that determines the performance of SR models in the real world and their ability to generalize. To illustrate this point and provide a general overview of the differences between non-blind and blind SR methods, Figure 5.2 shows an example scenario where the same images have been upscaled with different techniques.

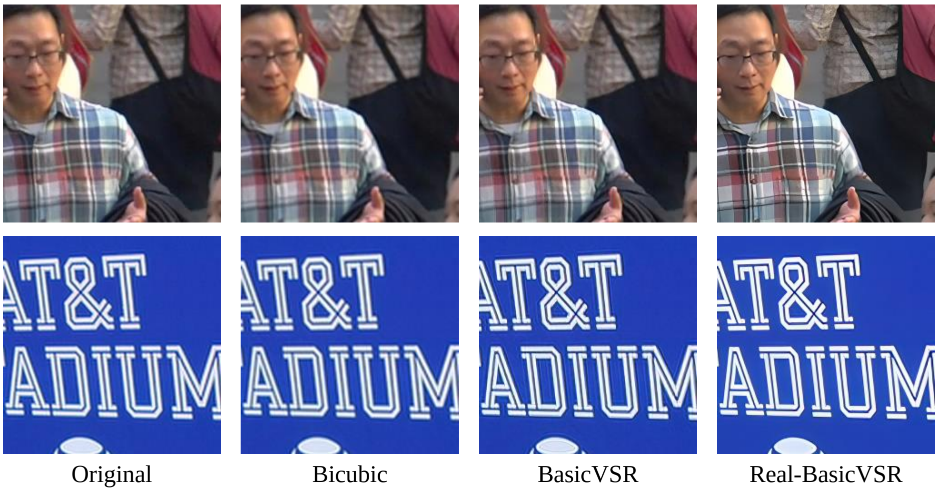


Figure 2.4: Demonstration images for traditional (bicubic) blind (Real-BasicVSR) and non-blind (BasicVSR) methods. Zoom in for better view

As can be observed in Figure 2.4 there is a significant difference between displayed SR methods. Especially, the modern VSR models are able to successfully enhance the details, generating results that can be more visually appealing than original images. Different training degradation types can lead to variations in performance, even under the same SR method. Understanding how degradations may affect the performance of SR techniques makes it necessary to investigate the various architectural strategies in SR.

In the following section, we will examine a variety SR techniques that can be roughly categorized into distinct categories depending on their design. Thus, we move from the impact of downsampling methods in SR to reviewing the types of upscaling techniques.

2.3. Categories of SR methods

SR methods can be broadly classified into interpolation-based methods and learning-based methods, depending on their approach to enhancing image quality. Exploring the categories of SR methods underscore the variety of existing strategies,

2.3.1. Traditional methods

To reconstruct the upscaled image, interpolation-based methods estimate the value of new pixels by considering the value of existing neighboring pixels, thus utilizing the spatial information and local context within the images. These methods are fast to compute and have been successfully used over recent years to provide satisfactory results in a variety of applications, such as zooming in digital photography and computer graphics. However, they are limited by their inability to reconstruct high-frequency details, such as fine textures and sharp edges. In addition, traditional interpolation techniques struggle to deal with more complex degradations such as blur, noise, or compression artifacts. As a result, machine learning and deep learning-based techniques have emerged, evolved, and gained popularity, outperforming traditional methods. Some popular interpolation methods are nearest-neighbor interpolation, bilinear interpolation, bicubic interpolation, and Lanczos interpolation.

2.3.2. Deep-learning-based Methods

Deep learning (DL) based methods employ deep neural networks to learn intricate features for the upscaling process, allowing networks to produce a high-resolution image that contains more detail and accuracy. Over time, DL-based methods have managed to outperform the classical interpolation methods, offering substantial improvements in the quality of the upscaled images.

2.3.2.1. CNN-based

More recent DL approaches used Convolutional Neural Network (CNN)s, such as SRCNN [27], a pioneering DL SISR method that uses a three-layer CNN to learn the mapping between glslr and glshr images. The initial layers are designed as feature extractors, which convert the initial image data into an internal representation that comprises meaningful information. Then that representation is transformed into a higher dimension feature vector, which is processed by the final convolutional layer to obtain the resulting glshr image. SRCNN demonstrated considerable improvements over traditional methods and cleared the path for future more sophisticated techniques. The training data consists of small patch pairs that represent glslr and glshr images, generated by applying bicubic interpolation on the original glshr images.

A deeper CNN-based network was presented in [37], based on the popular VGG-net [90], typically consisting of 20 convolutional layers. This architecture allows the model to learn more complex image representations, resulting in a better SR performance at a higher computational cost. The training process is similar to SRCNN, the low-resolution images (not patches) are first downscaled and then upscaled with an interpolation method, to create the LR-HR pair dataset. During training, VDSR learns to predict the residual image, that is the difference between the ground truth glshr image and the generated artificial image. The residual is added to the upscaled low-resolution input to produce the final output, and the network is optimized based on the distance between the output and the ground truth glshr image. More modern examples that are based on residual learning are EDSR [47] and CARN [5], which adds a recursive component.

2.3.2.2. GAN-based

Generally, described methods seek to reconstruct the glshr images by maximizing the MSE value, a reconstruction accuracy metric that does not ensure improvement in visual quality. As an attempt to produce more perceptually accurate results, Generative Adversarial Networks (GANs) gained popularity and showed great capability in the SR task. Perceptual accuracy refers to how closely the upscaled image resembles the original high-resolution image from a human observer's point of view. The general GAN architecture consists of generator and discriminator networks, which compete against each other to produce more realistic and visually appealing images, compared to previous CNN models.

Considered one the most influential works in GANs for SR, SRGAN was presented in [39]. The success of SRGAN is attributed to the effective combination of the generator-discriminator architecture plus the incorporation of the perceptual loss function, which helped produce more photo-realistic glshr images compared with the common MSE loss function that produced overly smooth details. In the architecture, both the generator and the discriminator are deep convolutional networks. The generator is responsible for producing the glshr images from glslr inputs. The discriminator acts as a binary classifier, determining whether a given image is an artificial product of the generator or a ground truth glshr image. During the training phase, SRGAN utilizes a combination of two loss functions, Content Loss and Adversarial Loss. Content loss is obtained by measuring the difference between the generated image and the ground truth image, to ensure that the generated images are similar to the ground truth. The content loss consists of two parts, the pixel-wise error, that is, the MSE between both images and the perceptual loss. Perceptual loss is calculated by measuring the similarity between the feature maps extracted from a pre-trained neural network. It encourages the network to create images that possess similar structures and high-level features, resulting in more visually appealing results.

The adversarial loss, on the other hand, measures the ability of the generator to produce images that can deceive the discriminator. It encourages the generator to create more natural and realistic images that resemble the ground truth. GAN-based SR networks have evolved and been used in different training settings, like CinCGAN [119] that introduces the cycle consistency loss to learn the implicit data distribution in an unsupervised manner, Real-ESRGAN [105] (based on ESRGAN [104], which supposed a great advance for blind SR.

2.3.2.3. Transformer-based

Much like the advancement sparked by GANs in SR tasks, the integration of Transformer architectures in computer vision has similarly opened new paths in this field. The recent success of the Transformer [97] in NLP tasks has inspired researchers to extend its capabilities to the domain of computer vision. Transformers, by their nature, are capable of learning global contextual relationships within data, which have shown to be particularly effective in a variety of computer vision tasks, including SR. Earlier implementations such as [17] demonstrated the ability of transformers to deal with low-level-vision task, but SwinIR [44] soon became the reference for SR transformer architectures. In SwinIR, the deep feature extraction module is composed of several residual Swin Transformer [55] blocks. It is widely used for different vision tasks besides SR, like image denoising (including grayscale and color image denoising), and JPEG compression artifact reduction. The version trained with BSRGAN degradations achieves state-of-the-art results in x2 and x4 camera shots according to the MSU Video Upscalers Benchmark [99]. The most prominent application of transformers for video is RVRT [45], which improves the performance and computational cost of its predecessor [46] by applying a recurrent video restoration transformer architecture. RVRT processes local neighboring frames in parallel within a globally recurrent framework, it aggregates features from different video clips and aligns them to reconstruct the output.

2.3.2.4. Recurrent Models for VSR

Recurrent mechanisms are especially interesting for VSR [85], as they allow propagating features from neighboring frames before performing the upscaling operation. The added temporal dimension introduces new challenges, mainly *Propagation*, *Alignment, Aggregation*, and *Upscaling.* They are considered and studied in [11], to create BasicVSR, a simple framework baseline for VSR. BasicVSR opts for a bidirectional propagation and a simple flow-based alignment at feature level. For aggregation and upsampling, it relies on popular methods such as feature concatenation and pixelshuffle [88].

BasicVSR was further improved in BasicVSR++ [12], obtaining state-of-the-art results for video, and also later adapted to a blind SR setting, in [13]. Real-BasicVSR implements an image pre-cleaning stage, indispensable to reduce noises and artifacts prior to propagation. Their objective is to "clean" the input sequence so that the degradations in the inputs have a weaker effect on the subsequent VSR network. Trained with Real-ESRGAN degradations achieves state-of-the-art results in VSR [99].

2.3.2.5. Diffusion Models

Rounding the examination of SR architectures, it is necessary to introduce the role of LDMs in SR. Diffusion models introduced a novel perspective into generational AI, causing a fundamental shift in the field by enabling the production of remarkably high-quality and customizable synthetic data. Overall, LDMs work by adding noise to the existing training data and then reversing the process. Over time, the model learns to eliminate the added noise, enabling the creation of high-quality synthetic images from random noise. LDMs for SR are still in the early stages, but there have been significant advancements and applications. Most notably, [84], [28] adapted denoising diffusion to SR, through a model that exhibits strong performance in comparison with other SISR methods.

2.4. Image/Video Quality Assessment

Due to the increasing use of SR techniques, evaluating the quality of the resulting upscaled images has become increasingly important. Evaluation techniques allow the comparison and identification of best-performing methods and provide insights into the strengths and shortcomings of different SR methods. Finally, developing more robust quality assessment methods assures the results and comparisons are reliable and trustworthy, a challenging task for artificial intelligence in computer vision.

Image or video quality is a term that represents the visual qualities of images and focuses on how observers perceive them. Image or video quality assessment incorporates techniques to quantify and analyze the factors affecting the visual experience of observers and includes two branches, subjective and objective evaluation. Subjective methods rely on human judgment to determine the quality of reconstructed images, by rating each image based on the visual perception. The most used subjective method Mean Opinion Score (MOS) is obtained by asking a group of human observers to manually rate the quality of a set of images on a predefined scale, e.g., from 1 (poor) to 5 (excellent). This method is considered to be the most reliable but requires human involvement, which translates into increased time and effort, making it unfeasible for instances where there are tens of thousands of images.

Objective measures use mathematical models to provide a numerical representation of the image quality, a more practical approach that relies on the implementation of such algorithms and keeps humans out of the loop. Objective metrics are designed to fit the human evaluation of the input, which will ultimately be the recipients of the image transmission system. Objective evaluation methods are more widely used in quality assessment, often combined with subjective methods, to save human and material resources.

Objective Image Quality Assessment (IQA) methods can be classified into three categories: Full Reference, Reduced Reference, and No Reference IQA.

2.4.1. Full-Reference IQA

Full Reference IQA (FR IQA) makes use of the original and distorted images to obtain a quality score based on the difference between the two. This process is only possible in settings where the LQ-HQ image pairs are available, which is often the case, because LQ images are obtained by downscaling the HQ ones. However, it limits its applicability in real-world scenarios where reference images are not accessible.

The most popular FR metrics in SR are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

2.4.1.1. PSNR

PSNR considers the pixel-wise changes between the reference and upscaled image by measuring the MSE, or the average square difference between pixel values at the same location. Considering a ground truth image $I_y \in \mathbb{R}^{H \times W}$ and the reconstructed image $I_{SR} \in \mathbb{R}^{H \times W}$, PSNR is defined as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{2.5}$$

Where MAX is the maximum possible pixel value in the image. PSNR is measured in decibels (dB), where a higher value indicated higher quality. While a higher PSNR typically suggests superior image reconstruction quality, it primarily accounts for per-pixel MSE, causing it to fall short in recognizing perceptual discrepancies. Despite having several limitations, PSNR is the most popular metric in SR, mainly caused by its low computational cost and convenience for optimization purposes.

2.4.1.2. SSIM

SSIM is a perception-based method that considers the luminance, contrast, and structure features to measure the similarity between two images. Unlike PSNR, which computes absolute errors at the pixel level, SSIM proposes that there are significant interdependencies among spatially adjacent pixels. These dependencies hold crucial information concerning perceptual structures. As such, SSIM can be articulated as a weighted combination of three comparative metrics [41]:

$$
\begin{aligned}
\operatorname{SSIM}\left(I_{SR}, I_y\right) &= \left(l\left(I_{SR}, i_y\right)^{\alpha} \cdot c\left(I_{SR}, I_y\right)^{\beta} \cdot s\left(I_{SR}, I_y\right)^{\gamma}\right) \\
&= \frac{\left(2\mu_{I_{SR}}\mu_{I_y} + c_1\right)\left(2\sigma_{I_{SR}I_y} + c_2\right)}{\left(\mu_{I_{SR}}^2 + \mu_{I_y}^2 + c_1\right)\left(\sigma_{I_{SR}}^2 + \sigma_{I_y}^2 + c_2\right)}
\end{aligned} \tag{2.6}
$$

Where l, c, and s denote the luminance, contrast, and structure between I_{SR} and I_y, respectively. $\mu_{I_{SR}}, \mu_{I_y}, \sigma_{I_{SR}}^2, \sigma_{I_y}^2$ and $\sigma_{I_{SR}I_y}$ correspond to the are the average (μ)/ variance $\left(\sigma^2\right)$ / covariance(σ) of their respective elements. SSIM ranges from -1 to 1, with 1 representing perfect similarity. Additionally, there exist several variants of SSIM, for instance, Multi-Scale SSIM (MS-SSIM)[109], which compares image structures at multiple scales, making it more versatile for images of different sizes and resolutions.

There exist other learning-based quality assessment methods, that instead of directly comparing the images as three-channel (RGB) matrixes, employ feature extraction and integration techniques.

2.4.1.3. LPIPS

Learned Perceptual Image Patch Similarity (LPIPS) [127] is another perceptual IQA metric that has gained attention in recent years and aims to provide a more accurate representation of human perception by leveraging deep features extracted from CNNs, which have been trained on large scale classification tasks. The metric computes the distance between feature representation of the reference and original images, assuming that they capture the perceptually meaningful information.

Consider an image X and its distorted version Y. The LPIPS score S between them can be calculated using:

$$
S(X, Y) = \sum_{i=1}^{n} d(F^i(X), F^i(Y)) \tag{2.7}
$$

Where F^i refers to the feature maps at layer i of the pre-trained network (for example, VGG or AlexNet), d is a distance function (usually cosine distance or L2 distance), and n is the total number of layers considered in the network.

The distance function d for the L2 distance can be represented as:

$$
d(F^i(X), F^i(Y)) = \sqrt{\sum_j (F_j^i(X) - F_j^i(Y))^2} \tag{2.8}
$$

where j iterates over the elements in the feature maps.

2.4.2. No-Reference IQA

No Reference IQA, also known as blind IQA (BIQA) methods are evaluated by different performance metrics, that are defined by the similarity between the results of the models and the subjective scoring by the human eye. Among existing BIQA metrics Natural Image Quality Evaluator (NIQE) [69]and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [68] are the most used ones.

2.4.2.1. NIQE

NIQE requires no prior knowledge about expected distortions as training image pairs, unlike FR metrics. This method uses a Natural Scene Statistics (NSS) model to extract a collection of localized, quality-aware features from images, which are then adapted to fit a Multivariate Gaussian (MVG) model. The final image quality is based on the distance between its MVG model and the MVG model derived from a natural image.

2.4.2.2. BRISQUE

BRISQUE analyzes the spatial nature scene statistics, focusing on locally normalized luminance coefficients. The main difference with NIQE is that it requires training on human-rated images, but it is usually reported to outperform NIQE in terms of correlation with human quality assessments. BRISQUE first computes the local Mean Substracted Contrast Normalized (MSCN) coefficients and their pairwise products. Then, a variety of statistics are calculated from the coefficients, including means, standard deviations, and higher-order moments, forming a feature vector that represents the spatial NNS of the image. The quality of the image is predicted by a Support Vector Machine (SVM) model, trained on a set of features from human-rated images.

2.4.3. Video Quality Assessment

Same as IQA, Video Quality Assessment (VQA) seeks to build models for evaluating the quality of videos, often applied in streaming and compression algorithm research industries. Similarly, it can be divided into FR, RR, and NR VQA. Although FR research has matured and several models are widely used, recent focus has shifted towards creating better NR metrics that can evaluate the quality of distorted videos in real-world scenarios, where references are not available. In this context, it is essential to highlight that SR detection and quality assessment techniques share some similarities and limitations. The tasks of classification and regression are different, but the underlying problem, data, and common architectures are notably analogous. SR and QA are heavily affected by any form of degradations on the input data, as both attempt to provide a method that can generalize outside the training dataset. For both the process of QA and detecting artificially upscaled content, deep neural networks are often employed for feature extraction. Equally, these two fields encounter shared challenges when dealing with high-resolution videos and added temporal dimension, primarily due to the associated computational complexity.

2.4.4. Limitations of current QA methods

Image and video quality assessment metrics play a pivotal role in diverse computer vision disciplines. Offering a quantitative measure of the 'perceived quality' is a relevant concept that remains subjective, but provides processing algorithms with a metric that can be used for optimization and evaluation. However, the existing and most widely used QA metrics have known limitations that prevent them from accurately capturing the nuances of human perception. As a result, the field of QA is expanding, to find metrics that can better represent human perception. PSNR and SSIM specifically, have been recently criticized for not being applicable for SR benchmarks [83], [82], despite being the most popular metrics for estimating the quality of SR, according to a post made by the Video processing, compression and quality research group [1]. The publication [110] effectively demonstrates a limitation of the MSE, the base of PSNR. In the study, different forms of distortions are applied to the original image, resulting in different quality levels among the resulting images. Despite these visually apparent disparities in quality, the MSE values for each distorted image are the same. This claim is backed by further research [108] [38] [29], where popular IQA metrics such as PSNR and SSIM have been proven to display low correlation with subjective scores.

No-Reference metrics face similar problems, namely, the struggle to accurately emulate human perception. More complex NR metrics like BRISQUE, NIQE, and PIQA are trained on specific distortion types, so they may not adapt to unseen distortions or real-world images that are more diverse. Moreover, there are a number of additional limitations with regard to more specific aspects in the literature. [117], for instance, claims that most of the current NR-VQA methods are still aimed at low-resolution videos, and they do not perform well when applied to UHD videos. To solve the limitations of existing NR-VQA methods, which primarily use CNNs on image patches, researchers have proposed an architecture based on a combination of SR and deep reinforcement learning. Further, [40] focus on the effects of compression on NR-IQA solutions, and highlight some underlying assumptions of other methods, such as the Gaussian assumption in NIQE. This serves as an example of the current research in QA methods, illustrating the variety of possible strategies towards video quality evaluation.

2.5. Synthetic Content Detection

The rapid advancements in AI and in Deep Learning in particular, have led to significant progress in the field of synthetic content generation. When Generative Adversarial Networks started gaining popularity, we experienced considerable growth in the quality of synthetic images. Since then, image generation and modification techniques such as inpainting, style-transfer, and SR have seen remarkable advancements, achieving an ability to generate incredibly realistic content. The latest synthetic content generation algorithms can outperform GANs [26] [78], and pose a challenge for humans to distinguish real from fake or modified content. This has led to the development of various detection techniques, which have been mainly focused on Deepfakes, due to their severe social and political implications.

Developing detection methods may seem like a straightforward task, where a classifier is trained on real and synthetic images. However, there are several unique limitations that must be overcome. The dataset will most likely be tied to the model's performance, affecting its generalization ability [130], [102]. Content type, degradations, and fake image generation models suppose a challenge in capturing the full range of potential manipulations. Furthermore, image generation or modification algorithms are in constant development, causing the features learned by these models to become ineffective and outdated for new techniques.

Earlier image generation methods were mostly built upon CNNs, which are the backbone of GAN-based networks. Even if results could be convincing, [103] found that CNN-generated images were notably easy to detect, by cause of particular CNN *fingerprints* with a great ability of generalization. Several CNN-based image generation methods were considered, in-

[1] `https://videoprocessing.ai/metrics/ways-of-cheating-on-popular-objective-metrics.html`

cluding GAN-based methods and one SR method. In [103], a classifier model is trained on one specific model, to resemble real-world detection problems. Moreover, they highlight the critical importance of data augmentation for generalization, as well as using diverse training images. By using 224 pixel crops and augmentations based on blur and JPEG compression, they achieve close to perfect Average Accuracy (AP) on unseen content generated by GANs, which belong to the same generative model family. Curiously, except for SR and Deepfake detection, which achieve 93.6 and 98.2 AP respectively without augmentations, the implementation augmentation helps performance in all other cases. The reason behind this phenomenon is that in SR models, only high-frequency components can differentiate between real and fake images, hence, applying blur or other similar degradations at training reduces performance [103].

The generalization ability described in [103] was further explored in [76], who tested the model on a different family of generative models such as LDM (Latent Diffusion Model). The classification accuracy drops to near chance, a fact that is supported by a study of the internal feature representations, which are able to distinguish content that has been generated by any GAN from other types of content, either real or fake. This is supported by a comparison of the frequency spectra visualizations between images generated by GANs and LDMs. There is a common discernible artifact pattern in GANs that does not exist for LDMs or real images, suggesting that the classifier's decision is based on the recognition of said artifacts. As a universal fake image detector, [76] choose a variant of the vision transformer, ViT-L/14, trained for the task of image-language alignment, CLIP. The feature representations from the network are then used to decide if a given image is real or not. Results show a highly better generalization performance in detecting real/fake images. Interestingly, this does not hold true for the only studied SR method, SAN [23], where the method based on Linear Probing achieves 79.02 AP score and 57.50 classification accuracy.

Moving to SR detection, is a relatively unexplored topic, especially for 4K content. The main problem of existing detection techniques is that they do not adapt to new SR proposals, showing deterioration in performance. This is a well-known topic in the field of synthetic content detection and poses the main challenge in SR detection.

2.6. Review of existing super-resolution detection methods

Artificial upscaling techniques have existed for several years, and have significantly evolved from earlier signal processing-based approaches. Prior works mainly focus on the detection of simpler interpolation algorithms, such as bicubic interpolation, bilinear interpolation, and nearest-neighbor interpolation. More recent research explores the detection of upscaled content generated by neural networks. However, it is an area that is still in development, and publications that tackle 4K images or videos are limited.

2.6.1. Super-Resolution Detection Model (SRDM)

[67] presents an approach to detect compressed and uncompressed upscaled videos, by incorporating a supervised and contrastive learning architecture. The classifier is trained with small crops taken from numerous upscaled videos (obtained by applying six SR architectures). It is evaluated on portions from the REDS and Vimeo-90 datasets, which consist of

720p videos (1280 x 720). In addition, it was tested on the MSU Video SR[58] (not available) and RealSR [9] benchmarks, achieving competent accuracy, detecting 30 out of 32 upscaling methods.

The architecture is based on a contrastive framework that takes three inputs: an anchor, a real image from the dataset; the negative case, that is, the same image but upscaled with a SR technique; the positive case, another original image from the dataset. To take advantage of the video format, two consequent frames are concatenated and used as input. From the input images, a ResNet-50 feature encoder and a projection head, a typical component in contrastive networks, composed of an MLP with three hidden layers are used to generate a feature representation for one of each image. Then, a composed contrastive loss is used, by combining cross-entropy with the result of the classification head, a regularization loss and, a contrastive loss.

In this system, a video is considered fake if at least 5% of all frames are detected as fake. Authors report a 100% accuracy for real videos in the MSU Video SR Benchmark [2], but a lower accuracy for upscaled videos. Using the RealSR benchmark, results are considerably better than previous methods, surpassing 95% accuracy for all upscaled methods.

The adoption of contrastive networks is not new in SR [106], but their adaptation to high-resolution data would require major modifications to account for the impractical computation and memory consumption. The authors of the described architecture do not mention how the evaluation process is carried out, considering that the model is trained with 224 x 224 patches and tested with higher resolution videos or images.

While deep learning methods have gained relevance due to their potency in handling complex tasks, detection methods based on frequency analysis have always been a popular approach for digital content scenarios. Frequency analysis, based on the field of signal processing, provides a computationally faster alternative and complementary approach to deep learning in SR. It exploits the fundamental nature of images as signals with different frequencies, making it an effective tool for discerning alterations in images and videos that have undergone SR.

2.6.2. DCT-based Detection

[116] propose a native resolution detection method for 4K-UHD videos, based on extracting features from the Discrete Cosine Transform (DCT) frequency domain. Authors support that for many upscaling methods, a sharp decline will be detected at the edge between high-resolution and low-frequency areas. To study the video frequency components, the DCT coefficients are summed along rows or columns, producing the accumulative log spectra of the frames. Then, the aggregated accumulative log spectra for a video is obtained by averaging the spectra of each frame. To detect sharp declines, a median filter is applied and subtracted from the original log spectra, to detect outliers, videos that are possibly upscaled. [116] sets an arbitrary threshold of -8, obtained by their experiments, to conclude if a video has been upscaled or not.

The method is evaluated on 10 videos from the MCML 4K UHD video quality dataset and considers four traditional interpolation methods and three DL-based SR methods. Reported

[2] https://videoprocessing.ai/

accuracy is perfect except for the RBPN method, which achieves an 80% True negative rate and a 100% True positive rate.

The main problem with SR detection on 4K videos is the resolution itself. Processing multiple frames consisting of more than eight million pixels at once is computationally impractical at best and unfeasible at worst, so various methods adopt a patch-based approach to the problem. Each image or frame is divided into small patches that are processed individually, and results are aggregated to obtain a final prediction.

2.6.3. Frequency domain + Natural Scene Statistics

Utilizing a patch-based approach combined with frequency features [132] proposed a no-reference image quality assessment metric to distinguish real and fake 4K contents. Natural scene statistics (NSS) and features from the frequency domain are extracted and processed by a support vector regressor (SVR), which aggregates them to obtain the quality score of the input image. The model is trained with a private database that true and pseudo 4K content, composed of 21 sequences from public 4K databases and 36 video sequences from channel of China Central Television (CCTV). To establish the pseudo or fake 4K content, 4K images are first downscaled to 1080p and combined with additional 1080p content from the Internet and from the video database of CCTV. Then, all 1080p images are upscaled by 14 different interpolation methods, including traditional and deep learning-based SR algorithms as well as video editing software. It bears noting that only SISR methods are considered in this analysis, and most of them are outdated. The final dataset includes 2,802 pseudo 4K, in which 1962 images are interpolated from 1080p and 840 images are upsampled from 720p, and it is not available for the general public as of May 2023.

2.6.4. Blind Texture-Aware UHD Content Recognition and Assessment (BTURA)

Drawing upon the research by [132], [57] was later published, a blind texture-aware UHD content recognition and assessment (BTURA) metric, a dual system that aims to recognize real and fake 4K images and measure their quality at the same time. The proposed model consists of three parts: a textured patch selection module, a quality-aware feature extraction module, and a quality evaluation module. The patch selection module picks three representative patches with the highest texture complexity, measured by the Grey-level Cooccurrence Matrix (GLCM) algorithm. Then, the feature extraction module, a pre-trained ResNet-18, is used to extract and aggregate features from all intermediate layers. This allows the system to capture high-frequency detail from the first layers and keep the more complex representations from the deeper stages. Extracted quality-aware features are concatenated and processed by the Quality Evaluation Module. A classification sub-network (MLP) and a quality prediction sub-network (MLP) are trained to predict a class label and a quality score respectively. Finally, a multi-task loss is applied to optimize the network. The same idea of a CNN-based feature extractor is common in other Video Quality Assessment methods, such as [92], [94] [128], good examples of how related SR detection and QA tasks are in terms o methodology and algorithmic design. This method is validated on four datasets: 4K IQA database established in [132], BVI-SR video quality database, MCML 4K UHD, and Waterloo IVC 4K video quality database. All databases are divided into training and test sets, that are evaluated independently. Reported results show an almost perfect (0.999% accuracy) score on 4K IQA and a 0.94% accuracy on BVI-SR. This type of multi-task learning is only possible because the training dataset contains collected quality scores for each sample, additional information that is expensive to get, and uncommon in large 4K video datasets. About this publication, neither the model nor the dataset are available.

2.6.5. Two-Stage Authentic Resolution Assessment (TSARA)

In a similar approach, [87] develop a two-stage system that classifies a video frame to have real or fake 4K resolution. The first stage classifies local patches using a CNN, and the second stage aggregates local assessments into a global image-level decision using logistical regression. According to the publication, the model is trained on a dataset consisting of ”Fake” and ”True” 4K images obtained by extracting frames from videos recorded at UHD resolution, which was released together with the model at [3]. They obtain the Fake images by upscaling the images extracted from 1080p video and a wide variety of native resolution images of 102 classes of flowers [74]. Moreover, only three traditional upscaling methods are tested, bicubic, faster-bilinear, and Lanczos.

[3] `https://github.com/rr8shah/TSARA`

3 Hypothesis and objectives

This section introduces and expands on the two central hypotheses and objectives that drive this thesis. It explores the complex relationship between degradation processes in SR methodologies, including blind and non-blind models, and their impact on the performance and generalization of these approaches. Further, it studies the limitations that currently exist in the detection of artificially upscaled content (specifically 4K videos), and offers a proposal over existing methods.

3.1. Hypothesis 1

The degradation process in SR techniques is critical in determining their performance and generalization ability. Blind models play a crucial role as they have the capacity to adapt to unknown degradations and produce more visually appealing results. Given the fact that the application of SR in the real world is performed over digital content with a great diversity of artifacts and degradations, blind models become the most interesting research line. Existing datasets and non-blind methods rely on a limited set of predefined degradations, therefore SR models trained with such data fail to fully represent the variety of real-world degradations, leading to reduced performance in such cases [49] [126].

We aim to explore the key differences between non-blind and blind models in their application to processing real and synthetic degradations. Our intention is to demonstrate that blind SR models exhibit a higher level of robustness against different degradations, while non-blind methods tend to learn features directly related to the training degradations. We believe that deep feature representations can uncover distinct *semantics* in video SR networks, which relate to image degradation rather than image content. We aim to substantiate these findings by conducting a quantitative performance analysis, derived from both original and synthetic glslr input data, and by comparing traditional, non-blind and blind models. Furthermore, we plan to utilize a degradation generalization metric based on the deep feature components to strengthen our conclusion.

All this analysis is facilitated by the glshr video dataset that we have manually collected in this thesis. It contains original and synthetic video pairs that serve as a comprehensive set for studying a broad range of degradation types in real scenarios. We want to take the opportunity to analyze the key differences between original and synthetic content in the same resolution.

This novel approach offers the unique advantage of enabling the generation of 4K and 1080p original video pairs, and serves as a benchmark for our study.

3.2. Hypothesis 2

Current existing detection methods often fall short of accurately identifying upscaled 4K content. This thesis proposes an improvement over existing proposals by extending an existing architecture and training it on more diverse upscaled content.

We intend to design, train, and evaluate a system that can accurately distinguish SR methods present within the training dataset. We are interested in understanding and explaining what the detection model is learning, and propose to analyze the feature representation of the model. By implementing different training strategies and architectures, we aim to optimize the detection model's performance. Our goal is to ensure it can consistently recognize a variety of upscaled content, focusing on methods that are outside of the training dataset.

The internal feature analysis can shed light on the similarities between different SR types (blind, non-blind) and architectures. The visualization of feature representations from various upscaling methods could provide a visual mechanism for understanding the model's behavior. It could explain which methods are treated similarly or help elucidate the underlying reasons for any erroneous predictions made by the model.

We will build the training dataset upon the public BVI-DVC, which offers a broad range of content. Our intention is to extend the available SR data resources by generating new upscaled videos from BVI-DVC, by employing one traditional upscaling method, and three DL modern video SR models. We believe that acquiring 200 original videos and 800 4K upscaled videos can facilitate the task of current and future SR research. We believe it is a suitable data source to train and evaluate the SR detection model studied in this thesis.

4 Data

As in any other computer vision task, the quality and diversity of the datasets, including image and video datasets, are integral to the success of SR, especially as we transition towards high-definition formats such as 4K. Most existing SR methods require LR-HR image pairs for training and evaluation, regardless of whether they are individual images or frames from a video. Due to the difficulty of obtaining real paired data, popular training datasets are synthetic, that is, glslr images are generated by downscaling their glshr counterparts. There are several paired real-world datasets, but there is a lack of accepted guidelines within academia for training and evaluation of higher-resolution (4K) content. This section will will serve as a guideline detailing the specific datasets that will be employed throughout the thesis. In addition, it will detail the various datasets utilized in SR tasks: common training video sources, available 4K video datasets for more general research purposes and, finally, *Real* Datasets used in blind SR. In

4.1. Public Datasets

The most popular glshr datasets for non-blind networks are also used in blind SR. DIV2K [3], Vimeo90K [114], REDS [72] and Flickr2K[48] are commonly used for training, while BDS100 [65], Urban1000 [31], Vid4 [50], and Set5 and Set14 [121] are used for testing. These datasets have resolution of 1280×720 or lower, to they are not suitable for the study of 4K content.

Regarding 4K video datasets, Table 4.1 provides a more comprehensive overview of data sources and specifications. It is notable that the majority of video sources do not offer a large quantity of videos. The purposes of datasets in Table 4.1 varies, from high-frame ratio studies to compression and codec performance assessment.

Name	No. of videos
BVI-DVC [59]	200
BVI HFR [61] [62]	22
LIVE HDR Video Quality Assessment Database	31
4KMedia [1]	165
Netflix ”Chimera” video sequence [73]	52
UVG Dataset [66]	16
SJTU Medialab [91]	15
MCML 4K UHD video quality database [20]	10
DAREFUL Browse Free Stock Clips Footage [24]	99
A Collection of 100 4K Video Sequences [34] [35]	100
Xiph.org Video Test Media (derf’s collection)[2]	18
CableLabs [3]	9
AVT-VQDB-UHD-1 video quality database [79]	16
BVI-SR Database [124]	24
LIVE YouTube High Frame Rate (LIVE-YT-HFR) Database [64] [63]	5
Waterloo IVC 4K Video Quality Database [42]	20

Table 4.1: List of available 4K datasets and number of videos

4.1.1. Real Datasets

The second approach to obtain LR-HR image pairs is to capture them directly, which often requires more complex acquisition techniques and digital devices. They offer a strong advantage over synthetic datasets, as captured glslr videos contain authentic degradations. In other words, no downsampling process is involved, since the recorded noise, motion blur and compression artifacts are inherent to real-world imaging processes.

Available datasets are generated using diverse methodologies, recording formats and resolutions. Let us start with static image datasets, where one of the most common approach is to achieve the LR/HR pairing through camera zoom. DRealSR [77] uses five DSLR cameras to cover indoor and outdoor scenes. glslr and glshr image pairs are collected by zooming the cameras and cropping the images. For each scaling factor, a SIFT (Scale-Invariant Feature Transform) [56] methods is used to crop an glslr image to match the content of its glshr counterpart. The final dataset contains approximately 800 images for different scales, each cropped to corresponding sizes suitable for each scale (from 192x192 to 380x380).

Utilizing same zoom-based approach as [77], SR-RAW [131] contains raw sensor data and ground-truth high-resolution images taken with a zoom lens at various zoom levels. It encompasses 500 sequences in indoor and outdoor scenes.

The main problem with collecting LR-HR pairs though optical zooming is the impracticability of data acquisition. Recording samples with the exact same motion and pixel alignment is quite difficult in practice, so additional processing techniques are often required. This limits the variety and volume of videos that can be recorded and published using this approach.

A similar approach is followed in CameraSR[14] and RealSR [8, 9], where a DSLR camera with different focal lengths is used to capture the image pairs. CameraSR proposes an additional method based on mounting a phone on a translation stage for data acquisition. Images taken from a closer distance are regarded as the glshr ground truth, while those captured from a long distance constitute the glslr versions. Smartphones cannot match the flexibility in focal lengths provided by DSLR cameras, but they are capable of recording at 4K resolution and capture degradations that are inherent to their specific hardware and software configurations (e.g. sensor noise, lens distortions, and compression artifacts).

As in [77], the dataset generation process is costly, due to the spatial misalignment, intensity variation, and colour mismatching caused by employing different focal lengths.

Leaving static image datasets behind and moving to video datasets, the most relevant ones are RealVSR [115], and Real-RawVSR [120]. RealVSR [115] builds a collection of 500 sequence pairs captured using the multi-camera system of a modern mobile device. To eliminate the alignment artifacts around the boundary, they are aligned and cropped at the center region of size 1024x512.

Taking a different approach, Real-RawVSR [120] builds a two-camera system with a beam-splitter to make sure that there is no parallax between the two cameras.While the approach seems sound, its hardly scalable which shows in the limited size of the dataset. They perform alignment on the capture content to generate 150 video pairs of about 50 frames each. The maximum resolution of the dataset is 1440x640, for 2x scale.

The limitations of the datasets reviewed above are discussed in the work [16] identifies two main limitations in current SR datasets. The first one is that the perceptual quality of glshr images is not high enough, which limits the performance of SR models. Models trained on original glshr images can yield blurry details and irregular patterns, caused by the low quality of ground-truth glshr images. The second limitation is the lack of human involvement in ground truth generation. As such, ground-truth images of poor quality may be inaccurately considered as high quality, due to the absence of human review and quality control. As a consequence, SR models trained suboptimal ground truths tend to produce over-smoothed results [16].

To deal with those limitations, they propose a human guided strategy towards ground-truth image generation. First, several image enhancement models are used to improve the perceptual quality of HR, while keeping the resolution. Then, human subjects annotate the high quality regions and label the regions with unpleasant artifact as negative examples. That information is finally used to create a dataset with multiple positive and negative samples, that can be used to train Real-ISR models.

While this approach seems promising, it also introduces several challenges. One potential issue is the human bias during the annotation process. The perceptual quality can significantly differ between individuals, leading to inconsistencies in annotations. Another relevant challenge lies in the scalability. Manual annotations of image quality require a consistent setting for evaluation (e.g. same illumination and display device). It requires extensive human involvement, which is time consuming and may not be feasible for large datasets.

In conclusion, *Real* video datasets today cover a wide array of methodologies, but there are still distinct limitations to be addressed. Primarily, the complexity involved in data acquisition, related with maintaining identical pixel and motion alignment. Artifacts and inconsistencies introduced by zooming or employing different focal lengths requires additional processing steps, which is technically challenging and resource intensive. To mitigate those limitations, researchers tend to manually crop the glshr images into lower resolution patches, which are used to train blind SR models.

In this thesis, we propose a new *Real* dataset in 4K resolution, which will be used in the following chapters as a tool to study non-blind and blind SR. Our dataset will also be employed to create the testing data for the SR detection proposal.

4.2. Overview of Datasets Under Study

We are interested in the study of SR at 4K to address the current trend of digital media towards higher resolution. Common SR datasets as REDS4, and existing *Real* datasets do not meet such requirements, so we decide to build our own to compliment the public 4K videos that we will study.

There are mainly two data sources employed for the development of this thesis.

4.2.1. BVI-DVC-SR

The first one is based on BVI-DVC [59], published to train CNN-based video compression systems. It contains 800 video sequences at various spatial resolutions from 270p to 2160p. Among the 800 sequences, 200 correspond to original 4K videos from different sources, and the remaining 600 correspond to the same videos in lower resolutions (artificially downscaled). As we perform x2 upscaling, we take all original 4K videos and their corresponding glslr (1080p) counterparts.

We extend the base database by upscaling the 200 1080p videos with different methods to create the BVI-DVC-SR dataset. The selected upscaling methods include one traditional technique and three DL-based video SR models:

- Bicubic interpolation (traditional): We use the implementation from OpenCV to downscale 4K videos to half their original size.
- BasicVSR (non-blind): The official implementation at MMediting [70] (replaced by MMagic as in 26/05/2023) is used to perform the upscaling. We use the pre-trained model on REDS dataset, with the recurrent framework on a maximum of six frames.
- RealBasicVSR (blind): Similarly, the pre-trained model on REDS is used with the same configuration.
- RVRT (non-blind): The official RVRT Pytorch implementation[4] is adopted, with default parameters except for the number of frames. We consider 10 consecutive frames instead of testing the videos as a whole.

Therefore, we build a 1,000 4K video dataset out of which 200 videos are original, and 800 artificially upscaled. Figure 4.1

[4] https://github.com/JingyunLiang/RVRT

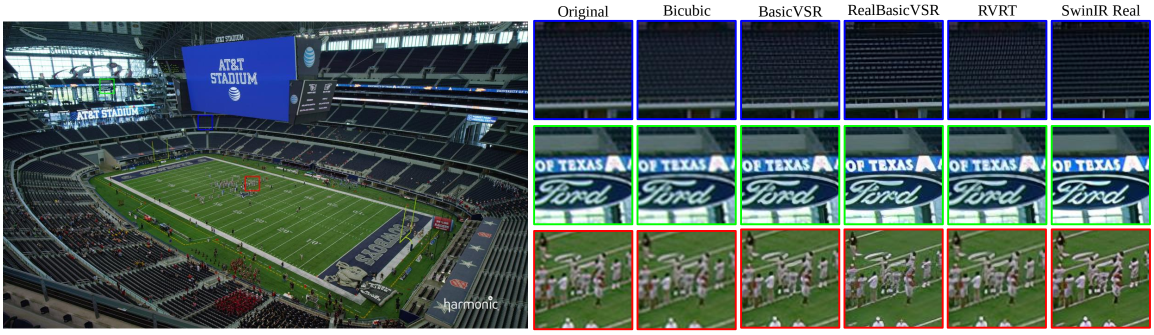


Figure 4.1: Frame extracted from a BVI-DVC-SR dataset. Shows the comparison between studied SR methods. Zoom in for a better view

4.3. Proposed Real Video Super-Resolution Dataset: BSC4K

In an effort to expand the current data on 4K SR, we present a dataset with paired video sequences at 1080p and 4K resolution recorded simultaneously. The dataset provides a valuable tool for analyzing the degradation nuances in the SR process by utilizing a unique camera setup to record the videos.

The motivation of the dataset is to overcome the challenges introduced by artificially downscaling glshr content to obtain the glslr counterparts. Generating original and synthetic glslr pairs allow us to study the domain gap that exists in non-blind SR methods and compare them to blind SR methods. Furthermore, the generated 4K videos will serve as a testing dataset for the SR detection proposal.

The first version of the dataset contains 33 4K and 33 1080p videos, cut to 64 frames each, recorded indoor and outdoor with a single DSLR camera. We restrict the SR problem to x2 scale, as the camera records at 4K and 1080p. The main advantages of our dataset are that the acquisition process is straightforward, while the post-processing primarily entails the separation of videos into individual frames.

While our methodology includes glslr and glshr recorded originally in that resolution, we add synthetically up and downsampled videos for evaluation purposes. Figure 4.2 shows an illustration of the data acquisition and naming conventions:

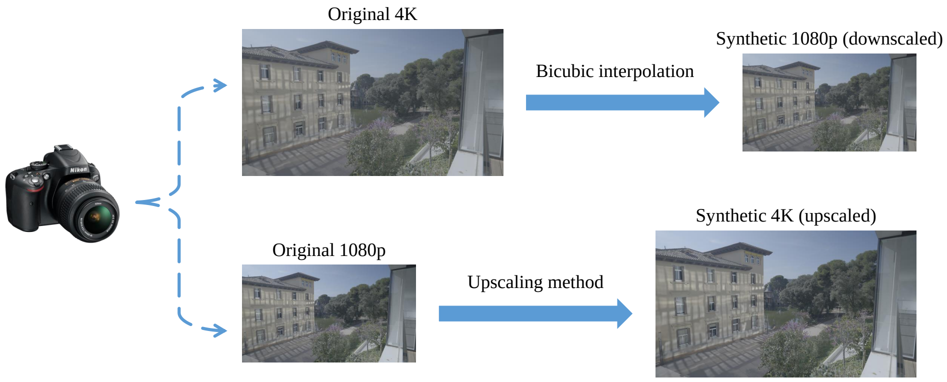


Figure 4.2: Data acquisition and processing diagram. We build pairs of LR-LR and HR-HR videos from real and synthetic content.

In following sections, we will reference glslr and glshr content with different names, depending on their origin. The camera records original videos at 4K and 1080p. Videos at each resolution are modified to obtain a pair of the opposite resolution. This way, we obtain ”original” glshr - ”synthetic” (or upscaled) glshr and original glslr - ”synthetic” (or downscaled) glslr video pairs (as showed in Figure 4.2.

In such manner, we generate several versions of glslr and glshr content. For LR, we consider original and degraded videos. The degraded versions include bicubically downsampling, BSRGAN degradations, and blur. The first two are obtained from the glshr videos and the latter by applying blur to the glslr samples. The glshr videos include the original and upscaled versions. For a better understanding of the dataset’s structure see Figure 4.3

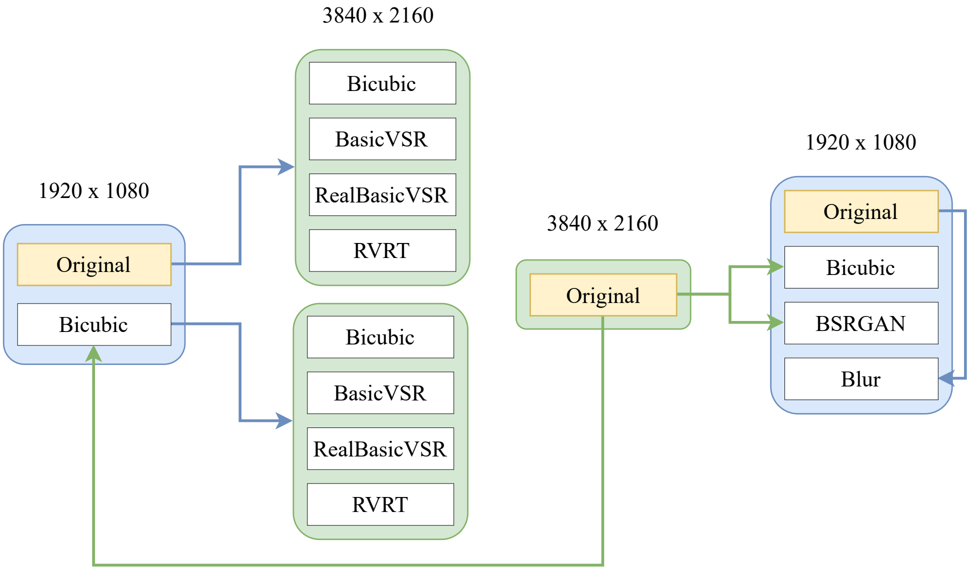


Figure 4.3: Illustration of the available data sorted by resolution. Arrow indicate where the data originates from. For instance, the 3840 × 2160 bicubic data comes from upsampling the original 1920 × 1080 videos

Following the same upscaling methodologies described in subsection 4.2.1, we increase the resolution of all videos by four upscaling techniques. Moreover, we also employ additional upscaling methods for testing purposes throughout the experiments in the thesis, most notably:

- SwinIR: Official implementation[5] for both the *Classical* and *Real* variants. The network weights are different but the transformer-based backbone is kept the same.
- Real-ESRGAN: We use the pre-trained model and default parameters from the official repository [6].
- Nearest-neighbor interpolation: The implementation from OpenCV is used.

4.3.1. Data Aquisition and Post-processing

This dataset has been generated by taking videos with a Panasonic Lumix DC-S5 in MOV containers with a H.264 codec at a resolution of 5.9k pixels. The sensor of the S5 camera is a full-frame 35.6x23.8mm CMOS sensor. The signal captured by the sensor, before being converted into H.264, was passed through to an HDMI interface towards a Blackmagic Video Assist 5'' 3G, recording in a MOV container with a ProRes HQ codec at a resolution of 1080p. The lens used has been a Panasonic Lumix S 20-60mm zoom lens at varying settings of focal length, ISO (the two native ISOs: 640 and 4000), shutter speed (from 30 to 80), aperture (from 3.5 to 22) and white balance (from 3200 to 5600k).

[5] https://github.com/jingyunliang/swinir
[6] https://github.com/xinntao/Real-ESRGAN

According to the AVC-Intra documentation [7], it , AVC-Intra 100 records the full 1920x1080 raster, representative of master-quality recording

After generating the .MOV files, we employ FFMPEG [95] to split the videos into frames in .png format. Due to a small difference between timecodes of 4K and 1080p videos, we methodically align the frames pairs of each video. This alignment procedure does not alter the content of the images. Instead, it ensures that the frame numbers from both videos align with the same timestamp and are consistent across both resolutions.

4.3.2. Content Description

Understanding the inherent characteristics of video sequences is fundamental for multiple aspects in the field of video processing and quality evaluation. These characteristics can provide an accurate and quantitative representation of a video's content, which is important when dealing and comparing different video data sources.

Following the method proposed by Winkler [112], we characterise the video sequences by using three descriptors: Spatial Information (SI), Temporal Information (TI) and Colourfulness (CF), and show the results in Figure 4.4. We adopt the SI and TI indicators defined in ITU-T Rec. P.910 (11/21) [80] and implemented in [100]. Spatial Information is an estimator of the amount of edge energy in the video sequence, and can be used to quantify the spatial complexity of a scene. First, each video frame is filtered with the Sobel filter. Then, the standard deviation of the pixels is calculated for each frame within the video sequence, resulting in a time series of spatial information. The highest value in the time series represents the spatial information content of the frame sequence:

$$SI = \max_{\text{time}} \sigma_{\text{space}}[\text{Sobel}(F_n)] \tag{4.1}$$

TI predicts the magnitude of motion based on the difference between the pixel values at the same location but at successive times or frames. The motion difference feature $M_n(i,j)$ as a function of time is defined as:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$$

Where $F_n(i,j)$ is the pixel at the i-th row, j-th column and nth frame. TI is computed as the maximum over time ($\max_{\text{time}}$)) of the SD over space (σ_{space}):

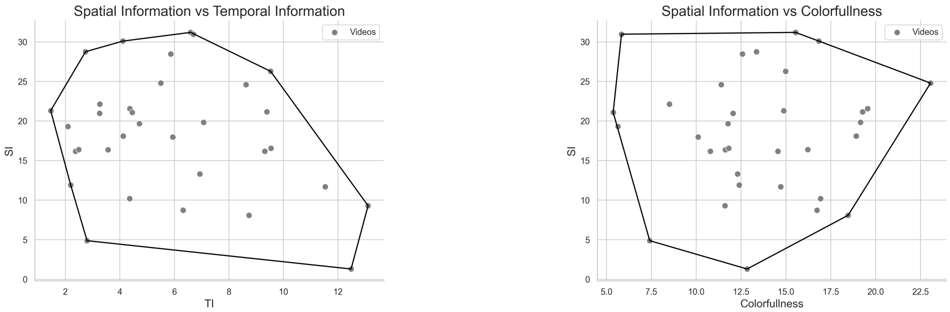
$$\text{TI} = \max_{\text{time}} \{\sigma_{\text{space}} \ [M_n(i,j)]\}$$

Finally, CF quantifies the variety and the intensity of colours within a scene. Using $rg = R - G$ and $yb = 0.5(R+G) - B$ as a simple opponent colour space, it is defined as [62]

$$\text{CF} = \sqrt{\sigma_{rg}^2 + \sigma_{by}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{by}^2} \tag{4.2}$$

The following figure shows the coverage of the three metrics for all videos in the dataset. This can serve as a reference for future iterations or other datasets, and can help evaluate the relation between spatial and temporal information with performance.

[7] https://resources.avid.com/SupportFiles/attach/FAQ_AVC-Intra.pdf



(a) Spatial Information (TI) vs Temporal Information (TI)

(b) Spatial Information (SI) vs Colourfullness (CF)

Figure 4.4: Illustration of the coverage of three video descriptors: Spatial Information (SI), Temporal Information (TI), and Colourfulness (CF). Each point in the scatter plot represents a single video in the dataset, mapped based on its SI, TI and CF values. A broad coverage across these metrics indicates a diverse set of videos in terms of visual complexity and colorimetry. The greater the spread of points, the more varied and representative the dataset, making it better suited for robust video processing and quality evaluation tasks.

This analysis serves as a foundation for future research, and can be used to compare existing databases, although the main purpose of our videos is to provide paired high-resolution sequences for SR research. Calculated metrics alone cannot ensure that methods trained or benchmarked with this dataset will adequate to real-world applications, so to draw meaningful conclusions it is crucial to analyze the problem at hand and the content of the videos independently.

Next, we show the value distribution for the individual low-level descriptors:

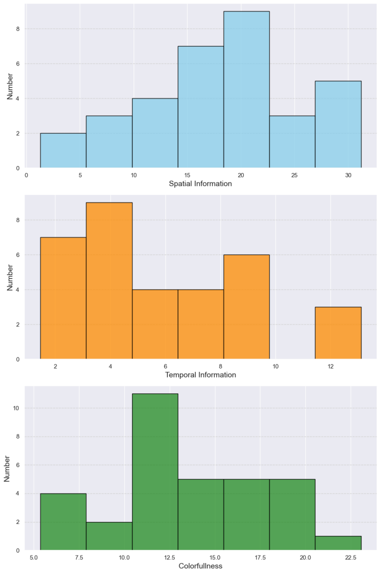


Figure 4.5: Distribution of the three low-level descriptors.

Presented figures and calculations establish a standard procedure for future versions and comparative datasets. The plots indicate that the dataset appears to be diverse in terms of content and motion components, without any unusual anomalies. The ultimate goal is to refine and increase the dataset and keep studying the unique properties that are present in the videos by comparing them to other data sources.

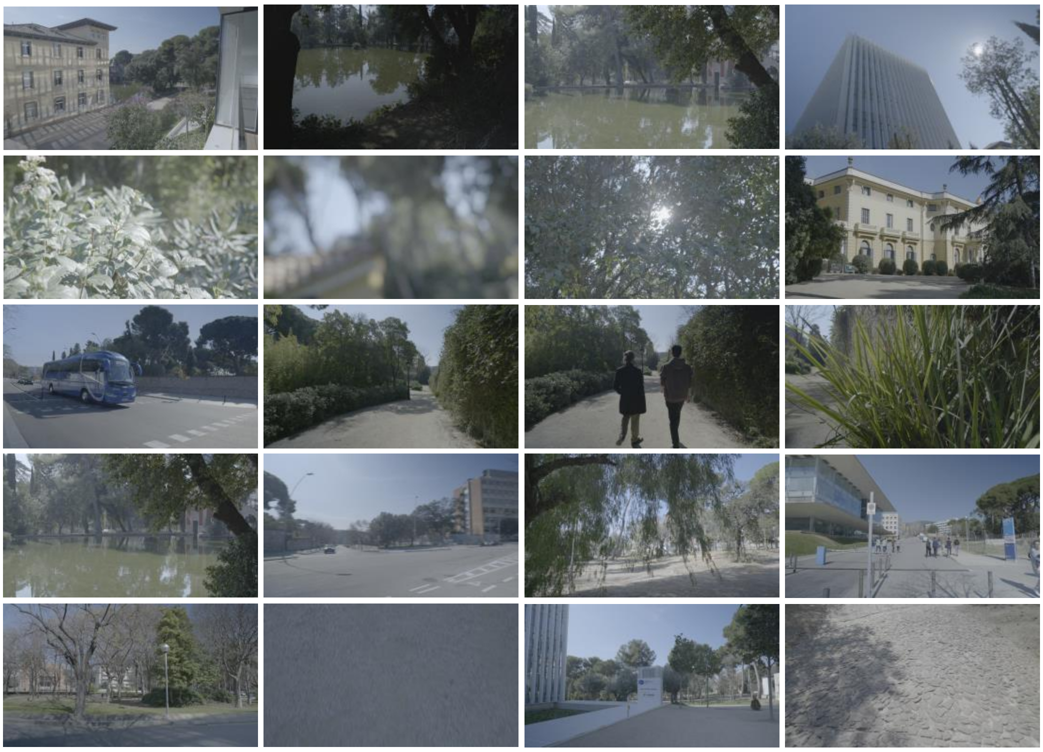


Figure 4.6: Example frames from the dataset.

The dataset is expected to be expanded to include more diverse content and environmental settings. Nevertheless, we try to cover several relevant categories in this first iteration, Even if the environmental and climactic aspects of the dataset are more similar. We include various types of motion speeds, zoom-out and zoom-in, blur and a variety of content types, as shown in Table 4.2.

Category	People	Vegetation	Text	Vehicles	Animals	Buildings	Textures	Close-up	Blurry	Focus change	Zoom
No. of videos	4	20	1	5	1	10	5	4	1	2	5

Table 4.2: Coverage of different content texture categories for BSC4K

We include scenes with minimal and high movement, as well as sequences involving panning shots and sequences that incorporate handheld camera movement.

4.3.3. Dataset licensing and GDPR Compilance

We intend to make the dataset open-access to promote accessible scientific research. The open license will allow researchers to download, use and modify the data freely as long as it is cited correctly.

In the process of recording and saving the data, we have taken care to respect each person's privacy. We have followed the guidelines in the General Data Protection Regulation (GDPR) to ensure that our practices correspond to their standards.

4.3.4. Limitations

The proposed dataset is created to study the effect of artificial degradation on blind and non-blind SR networks. It overcomes some limitations on current real datasets, like the costly alignment processes or the complex camera setups. However, our recording methodology is tied to certain restraining factors. Most notably, all videos are recorded with the same camera, which significantly reduces the range of available degradations, an important fact if the purpose of the data is to train a SR model. Furthermore, all videos are produced by the Image Signal Procesor (ISP) from raw data through multiple operations, which are non-invertible and tend to degrade the information content of the original raw videos. For that reason, other datasets [52] [131] propose to directly exploit camera sensor data, a recognized method in several areas of low-level vision [15] [32] [113]. Super resolution with raw videos is fairly unexplored domain, mainly due to the technical challenges associated with handling raw data and computational costs involved in processing such data.

On the other hand, the reduced number of videos and the similar recording settings present a limitation of the dataset to represent real-world diverse scenarios. The homogeneity of lightning conditions and environmental factors may hinder the applicability of the dataset and cause overfitting on some scenarios. Lastly, another limitation of our dataset is the slight difference in colour between glslr and glshr video, a common factor in many real SR datasets. In our case, the 1080p videos display a brighter color that the 4K versions. We keep the original videos in this dataset version, but a color matching step and its influence will be studied in the future.

5 Downsampling analysis

In SR methodologies, LR-HR pairs are an essential component for model training and evaluation. Therefore, the manner in which the training pairs are generated, that is, the downsampling operation that is applied to the original glshr image to obtain the glslr variant, can greatly influence the performance of the SR model. Biases or artifacts introduced during the downscaling process can reduce a model's ability to generalize to real-world situations, as the model becomes specialized to a specific type of downscaling that does not correspond to real degradations.

LR-HR image pairs are often generated through these types of predefined degradations, such as bicubic interpolation, deteriorating the model's performance for inputs that are outside of that training distribution. Blind or ”Real” models attempt to create a more complex degradation process that can emulate degradations in the real world.

The aim of this section is to conduct an exploration of the impact of different degradations on the performance of non-blind and blind SR models. We will consider the custom dataset BVI-DVC-SR and our BSC4K dataset to compare the quantitative metrics, the internal deep feature representations, and their frequency domain components. To do this, we will treat video frames as separate images that can be studied individually.

5.1. Quantitative performance comparison

We evaluate the effectiveness of various upscaling methods by measuring Full-Reference (FR) and No-Reference (NR) metrics across two separate datasets. Firstly, we use a subset comprising 50 videos from BVI-DVI, where the glslr videos have been obtained by applying a Lanczos filter to their glshr counterparts. In the second instance, we compute the performance on our dataset, where upscaled videos are obtained from real glslr inputs. To assess the performance, we employ three FR metrics - SSIM, PSNR, and LIPIS and two NR metrics - NIQE and BRISQUE.

PSNR and SSIM are often computed after taking the luminance (Y) component in the YCbCr color space and cropping the border [27] [44] [104]. Following the literature, we convert each RGB image into YCbCr, select the luminance channel, and crop 8 pixels from the border (implementation from the BasicSR repository [1]. To calculate LPIPS, we adapt the implementation in [2] (version 0.1), and use the default AlexNet variant. Similarly, we employ the adaptation of NIQE from BasicSR [3], which obtains very similar results as the

[1] `https://github.com/XPixelGroup/BasicSR`

[2] `https://github.com/richzhang/PerceptualSimilarity`

[3] `https://github.com/XPixelGroup/BasicSR`

original Matlab code. Lastly, we calculate the BRISQUE using the Python package with the same name[4]

In terms of SR models, we test one traditional method (bicubic interpolation) along three DL-based networks, two non-blind (BasicVSR and RVRT) and one blind (Real-BasicVSR) approach.

5.1.1. BVI-DVC-SR dataset

We average the results for each method and display them in Table 5.1

Method	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	BRISQUE↓
RVRT	**47.76 ± 4.93**	**0.992 ± 0.01**	**0.024 ± 0.02**	5.94 ± 1.26	49.23 ± 9.13
BasicVSR	47.52 ± 4.96	0.992 ± 0.01	0.03 ± 0.02	5.91 ± 1.32	48.88 ± 9.47
Bicubic	47.24 ± 4.64	0.991 ± 0.01	0.04 ± 0.03	6.68 ± 1.06	54.62 ± 8.76
RealBasicVSR	30.49 ± 3.82	0.89 ± 0.05	0.33 ± 0.12	**4.14 ± 1.20**	**25.00 ± 14.04**

Table 5.1: Quantitative metrics for BVI-DVC-SR dataset

Table 5.1 shows the average values of each upscaling method, along with the standard deviation. The first noticeable observation is related to the high PSNR and SSIM values overall. In contrast, the second remarkable fact is the low performance of Real-BasicVSR according in all FR metrics.

According to FR metrics, RVRT generates better quality results that are more similar to the original image, although the difference among non-blind methods (Bicubic, BasicVSR, RVRT) is quite small. It has the highest PSNR value, which suggests its effectiveness to retain information from the original image. The SSIM value indicates that the structural changes between the original and upscaled image are minimal, and the small LPIPS metric signifies that perceptual differences are also low. BasicVSR displays very similar performance in regards to FR metrics, as well as bicubic interpolation, despite the latter obtaining a lower perceptual similarity (and worst among all methods). Lastly, FR metrics for the non-blind method (Real-BasicVSR) are notably lower in comparison, mainly due to the tendency of non-blind methods to smooth the texture of resulting images.

Interestingly, NR metrics contradict the quality assessment made by FR metrics, indicating that videos generated by Real-BasicVSR are more visually pleasing. Once again, this occurs because the surfaces and textures from the videos are more smooth in contrast with other non-blind methods, and NR metrics that do not have access to the original reference content. In particular, BRISQUE, which has a higher correlation with human judgment, indicates a significantly higher quality in comparison. In addition, NR metrics are the highest for bicubic upsampling, indicating that it may provide the least natural-looking or lower-quality images.

Images with higher resolution contain a greater amount of pixels compared with lower resolution images. Consequently, even after the reconstruction process, they can still retain a significant amount of detail from the original image. This results in higher PSNR and SSIM values, although it does not necessarily mean that glshr images will obtain higher FR scores. Moreover, higher resolution images have higher pixel density, which means that even if some

[4] `https://github.com/rehanguha/brisque`

information is lost during the upscaling process, the high pixel density could help mitigate the visual impact of such distortions.

For a better visual understanding, Figure 5.1 shows some examples of frames generated by non-blind and blind networks, along with their PSNR and SSIM scores.

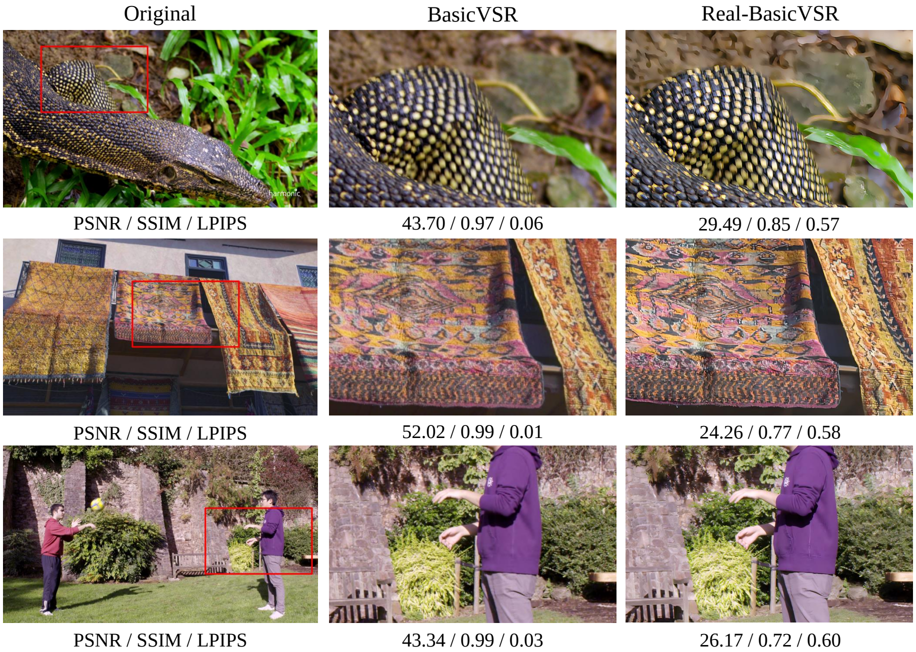


Figure 5.1: Original and upscaled frames extracteed from BVI-DVC dataset. We show the difference in visual quality and FR metrics for two models: BasicVSR (non-blind) and Real-BasicVSR (blind)

Ultimately, the perception of an image's quality is subjective, but Figure 5.1 helps to perceive the unique differences between both SR model generations. In general, Real-BasicVSR produces more sharp images, that may be more visually appalling for some people. However, FR metrics give the understanding that those images generated by the blind network have much less quality, which illustrates the possible limitations of PSNR, SSIM, and LPIPS to effectively evaluate images.

5.1.2. BSC4K (original LR)

Next, we compute the same metrics for all videos in our dataset. Having pairs of original and synthetic degradations for the same videos and resolutions poses an advantage in this scenario. The discrepancies in quantitative quality scores will not be caused by the video content but by the degradation type. As with any other computer vision model, SR networks may perform better in certain scenarios like indoors, spaces with people, nature, or vehicles, so our method ensures that the metrics are not skewed towards particular types of content.

Looking at Table 5.2, we notice a lower average value for all FR metrics, which is consistent for RVRT, BasicVSR, and Bicubic. Interestingly, the PSNR, SSIM, and LPIPS value for Real-BasicVSR is very similar in this dataset and in BVI-DVC.

Method	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	BRISQUE↓
RVRT	33.14 ± 2.66	0.96 ± 0.03	0.07 ± 0.04	5.76 ± 2.17	45.77 ± 10.07
BasicVSR	**33.46 ± 2.82**	**0.96 ± 0.03**	**0.06 ± 0.03**	5.93 ± 2.46	46.43 ± 11.21
Bicubic	33.11 ± 2.82	0.96 ± 0.03	0.11 ± 0.05	6.33 ± 1.65	52.26 ± 8.40
RealBasicVSR	29.74 ± 3.07	0.86 ± 0.06	0.29 ± 0.08	**4.27 ± 2.07**	**13.11 ± 10.95**

Table 5.2: Performance metrics for our dataset. The upscaled results come from the original glslr videos captured by the camera

The first takeaway from this fact is that non-blind methods could suffer an important drop in performance according to FR metrics because the input images do not follow the same distribution as the predefined degradation process. This could be a result of degradations specific to the image content, camera, or signal processing method.

The second takeaway is that Real-BasicVSR, which performs similarly across datasets, shows the robustness of its approach in handling diverse and complex degradation processes. This demonstrates a significant advantage over conventional, non-blind methods that often struggle to maintain performance consistency across different input distributions.

When examining NR metrics, we detect the same pattern and comparable values, where images generated by Real-BasicVSR are ranked the highest, while those produced by bicubic interpolation rank the lowest. Once more, BRISQUE indicates that Real-BasicVSR generates images that are substantially more aesthetically pleasing.

5.1.3. BSC4K (synthetic LR)

Lastly, we replicate the preceding step, but instead of employing the original images, we downscale the 4K versions to glslr (1080p) by applying bicubic interpolation. This allows us to obtain pairs of original-synthetic glslr images, and compare the performance depending on the input glslr degradation (original vs. bicubically downsampled). This time, the input glslr images are inside the training distribution data of non-blind methods (See Table 5.3 for the results).

Method	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	BRISQUE↓
RVRT	**40.84 ± 6.44**	**0.92 ± 0.04**	0.14 ± 0.08	4.02 ± 1.81	29.28 ± 12.61
BasicVSR	40.66 ± 7.82	0.95 ± 0.05	**0.13 ± 0.09**	4.25 ± 1.75	31.04 ± 12.48
Bicubic	24.20 ± 3.61	0.76 ± 0.12	0.34 ± 0.12	5.91 ± 1.55	49.49 ± 7.91
Real-BasicVSR	33.54 ± 4.10	0.90 ± 0.05	0.28 ± 0.08	**3.60 ± 0.81**	**10.69 ± 7.45**

Table 5.3: Performance metrics for BSC4K. The upscaled results come from the synthetic glslr videos downscaled from the original glshr videos with bicubic interpolation

Reviewing Table 5.3, we observe a considerable improvement according to PSNR, NIQE and BRISQUE for BasicVSR and RVRT. It confirms that these non-blind methods can better approximate ground truth and generate more visually appealing results (according to PSNR, the metric that is used to optimize the models) when the input images follow the same degradation process as the one used in the training phase.

The bicubic method shows a decline in performance metrics, particularly PSNR and SSIM, but a small improvement over the previous test according to NR metrics. Even in the presence of a familiar degradation, it fails to provide satisfactory results, which puts it at the bottom in terms of quantitative metrics.

Overall, Real-BasicVSR also sees an improvement, substantiating the method's robustness and its ability to generate visually attractive results. It is important to note that part of the difference in terms of PSNR and SSIM might be a consequence of the color difference between original glshr and glslr videos. In this last experiment, the glslr videos that are upscaled come from the original 4K variants (after bicubic interpolation downscaling). The FR metrics use the same original 4K frames as reference to make the score computations, thus the color components are the same. The original 1080p videos have a slightly brighter color, which is kept after the upscaling process. Nevertheless, the NR metrics indicate that even in those conditions, and without reference images, all methods produce better results.

5.1.4. Visual Effect of LR Degradations

Finally, we display the visual disparities depending on the input degradation. Serving as demonstrative models, we consider two SISR networks with the same backbone, SwinIR classical (non-blind), trained with assumed degradations, and SwinIR Real (blind), trained with BSRGAN degradations. To make this possible, We use two versions of each image at the same resolution as input, taken from our BSC4K dataset. Then the sample at 4K is downscaled by bicubic interpolation to obtain a 1080p pair that is comparable to the real sample (More details about the data acquisition process in chapter 4).

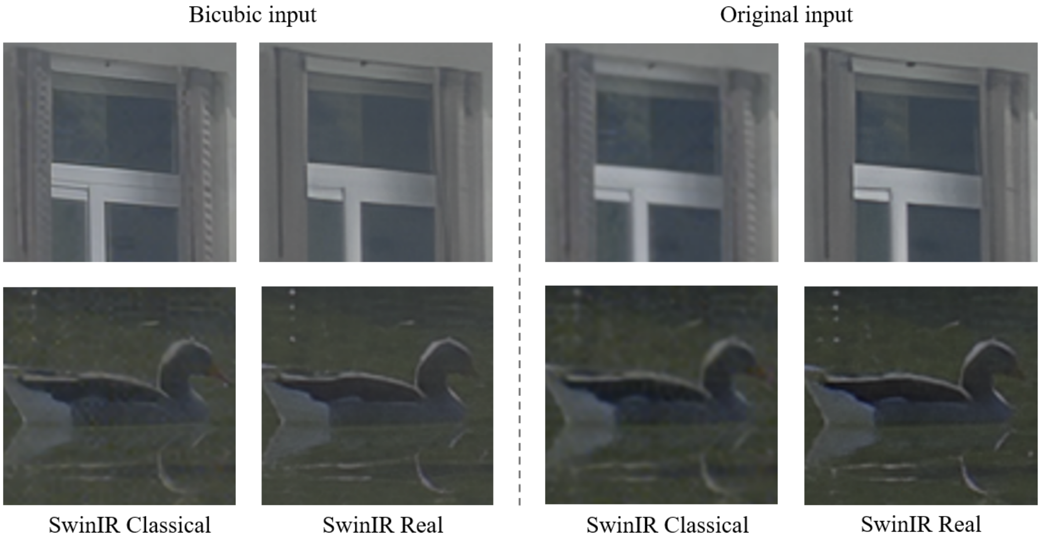


Figure 5.2: SwinIR output comparison. The classical and real SwinIR variants have been trained with simple and complex degradations respectively. The input bicubic image is a downscaled version of the 4k image. The scene has been recorded at 4k and 1080p simultaneously, so there is an original image in 4k and in 1080p. Zoom in for a better view

Figure 5.2 shows two 100 × 100 crops extracted from a glslr image in the dataset. We observe a tendency of the blind method to produce smoother results and a subtle but noticeable difference between each input type. While the blind method generates similar outputs regardless of the input type, there is a significant difference between the quality of images from the non-blind method. The non-blind network trained on inputs under an assumed degradation achieves superior quality when the input image belongs to the subspace of downsampled images. This helps to prove the point that using complex degradations in the training phase helps with generalization.

5.2. Deep features analysis

Deep features, which constitute the internal representations of a model learned during the training process, are critical to the reconstruction of high-resolution (HR) images. These deep features serve as pivotal factors in the realm of computer vision, facilitating a more comprehensive understanding of the model's outcomes and allowing explainability methods to give humans some guidance. Prior research on ISR [53] indicated that SR networks appear to discern the specific degradation types inherent in their training data. It further suggests that differences in data distribution might deactivate this discernment ability. In their exploration, the authors focus on what they reference as deep degradation representations or DDR. To achieve those representations, deep features are extracted from a SRCNN, and their dimensionality is reduced using Principal Component Analysis (PCA). The reduced feature maps are subsequently clustered through t-Distributed Stochastic Neighbor Embedding (t-SNE) [60]. A deeper study is conducted about the differences between shallow and deep layers in CNNs and GANs, which shows that different networks will learn different semantic representations.

Dimensionality reduction techniques are widely used in machine learning, but they come with their own set of advantages and disadvantages. In fact, the authors suggest t-SNE as a choice for CNN features, but do not claim it is the better option.

In our study, we adopt the same methodology to corroborate that modern VSR networks follow the same pattern. Additionally, we attempt to demonstrate the difference at feature level between synthetic degradations and original degradations present in our proposed dataset. Finally, we use the deep features to measure the generalization ability of analyzed SR networks by comparing their distributions, an idea proposed in [54].

5.2.1. Clustering VSR deep features by degradation

To investigate the effect of degradations in a video SR setting, we select two SR networks, BasicVSR and SwinIR, which are a recurrent network for VSR and a transformer-based model for ISR respectively. We consider two variants for each network, one trained on glslr inputs downscaled by bicubic interpolation and another one trained on ”real” glslr inputs. Particularly, we first compare the deep features from BasicVSR and RealBasicVSR, which adopt the same backbone, but the latter implements a pre-cleaning module for blind SR. In the case of SWinIR, we compare the classical version, trained with bicubic interpolation with Real SwinIR, trained with BSRGAN degradations.

To obtain the deep features, we exploit the nature of SR networks by directly extracting features from the intermediate and deepest layer before the upscaling operation. Generally, for an input of size $W \times H$ (width and height) and scaling factor s, the network will first increase the number of features maps, moving from an RGB input of $3 \times W \times H$ to a representation of size $64 \times W \times H$. This larger representation is then used to scale the image, by reducing the number of feature maps and increasing the resolution, until reaching the final result of $3 \times w \cdot s \times h \cdot s$.

Intuitively, using high-resolution images of size 4096x2160 will heavily increase the number of features. Following the previous example, each image from the deepest layer would contain $4096 \cdot 2160 \cdot 64 \simeq 566M$ features, a remarkably high number that further complicates the analysis process. To overcome this challenge, we create a lower resolution video dataset based on our glslr (1080p) videos. We select six points for each video, equally spaced, and crop patches of size 240×240 centered on each one. This method allows us to obtain more than 100 low-resolution videos, that we process by different degradations. In such a manner, we collect the original videos that were natively recorded at a resolution of 1080p, along with synthetic versions of those same clips. We study the following synthetic degradations: bicubic interpolation, from the 4K counterpart, blur, and BSRGAN degradations (see Figure 5.3). Original 1080p videos are used to generate the latter degradations.

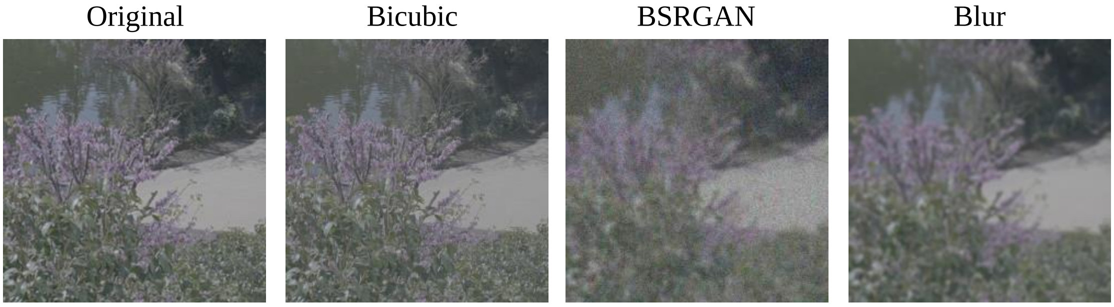


Figure 5.3: Cropped patch examples for each studied degradation (Zoom-in for a better view)

Instead of following the methodology in [53], we apply a mean pooling layer to condense the spatial information into a single value per channel, producing a feature vector of length 64 for each image. By computing the mean for each filter, we deal with an inevitable loss of information. However, we are more interested in the activation values of the neurons and can ignore their spacial information or localization. Furthermore, we avoid having to use PCA or other techniques on higher dimensional data. We reduce dimensionality to two dimensions with UMAP, so data points can be visualized in a 2D plane. Finally, we calculate the mean among all frames to get a single representation of each video.

To better illustrate and measure the discrimination ability, we adopt the Calinski-Harabaz Index (CHI) [10], a ratio of the mean between-cluster dispersion to the mean within-cluster dispersion. A high CHI score indicates that the clusters are well separated, and the data points within a cluster are close to each other. Given K number of clusters and N total number of data points, CHI score is formulated by:

$$CHI = \frac{B(K)}{W(K)} * \frac{(N-K)}{(K-1)} \tag{5.1}$$

Where $B(K)$ corresponds to the between-cluster dispersion, considering n_k the number of points in cluster k, c_k centroid (mean) of cluster k and c grand centroid (mean of all data points):

$$B(K) = \sum_{k=1}^{K} n_k \|c_k - c\|^2 \tag{5.2}$$

Lastly, $W(K)$, or within-cluster dispersion, is the sum of squared distances of all data points to their respective cluster center:

$$W(K) = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - c_k\|^2 \tag{5.3}$$

Where C_k is the set of points in cluster k, x is a data point in k, and c_k is the centroid of k.

The CHI score serves as an efficient way of complementing the visual results by assessing the quality of the cluster algorithm.

Results in Figure 5.4 and Figure 5.5 show that the studied complex video and image SR networks are good descriptors of degradation information. In these comparisons, the frame content of the studied videos is exactly the same for all degradations.

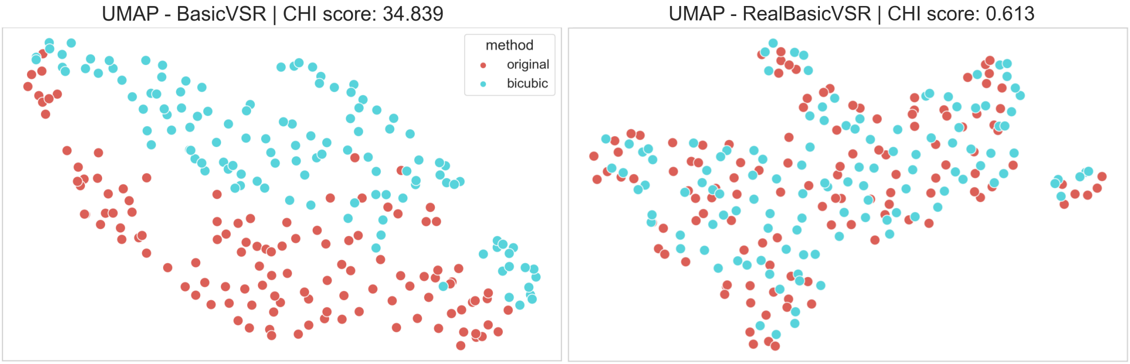


Figure 5.4: Deep feature representation differences between original and synthetic videos obtained by bicubic interpolation (Left: BasicVSR, Right: Real-BasicVSR)

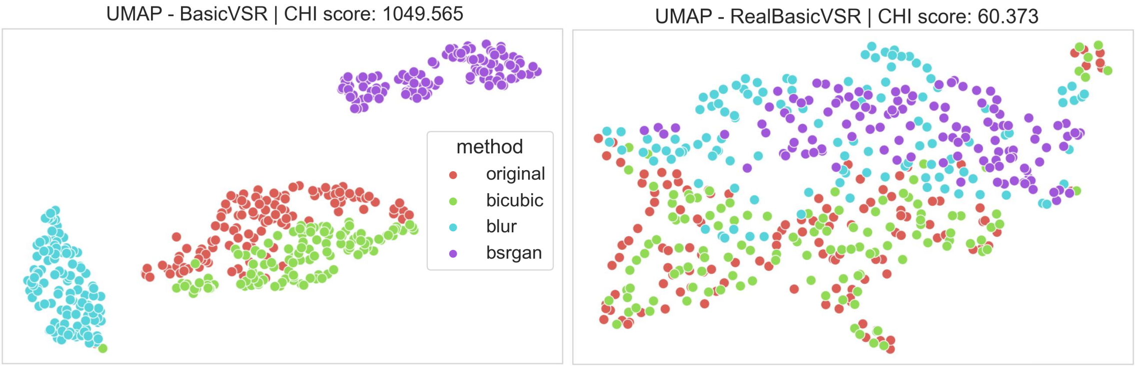


Figure 5.5: Deep feature representation differences between original and synthetic videos obtained by all degradation types (Left: BasicVSR, Right: Real-BasicVSR)

As shown in Figure 5.4, there exists a feature discriminability between original and bicubic samples in the case of BasicVSR, while RealBasicVSR does not distinguish between degradation types as clearly. BasicVSR is trained with glslr inputs that have been obtained by bicubic interpolation, so it recognizes images that follow the same distribution and images that do not. All frames of training videos suffer from the same degradations, so in the process of feature propagation among the temporal dimension, those assumptions are preserved. RealBasicVSR's image pre-cleaning module is crucial to remove degradations prior to propagation and suppressing artifacts in the outputs. As a consequence, deep features do not capture the information about degradations, as shown in Figure 5.5.

In theory, a model with a good generalization ability should be resilient to any type of input degradation. Looking at the feature representations, Real-BasicVSR shows more adaptability in its generalization capacity, as it shows less ”semantic” discriminability. A lower CHI score denotes better generalization (Table B.1 shows CHI scores for each method pair and for all of them together).

To offer an additional perspective and further explore the effect of degradations in SRs, we perform the same process on two variants SwinIR: SWinrIR Classical, trained on DIV2K, and SwinIR Real, trained with BSRGAN degradations. Features are extracted from the deepest layer, also consisting of 64 filters, and then the same dimensionality reduction method is applied.

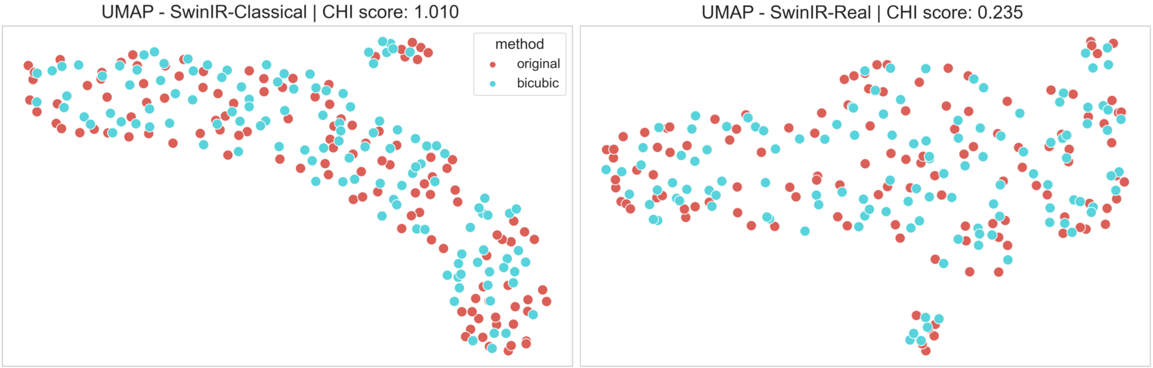


Figure 5.6: Deep feature representation differences between original and synthetic videos obtained by bicubic interpolation (Left: SwinIR Classical, Right: SwinIR Real

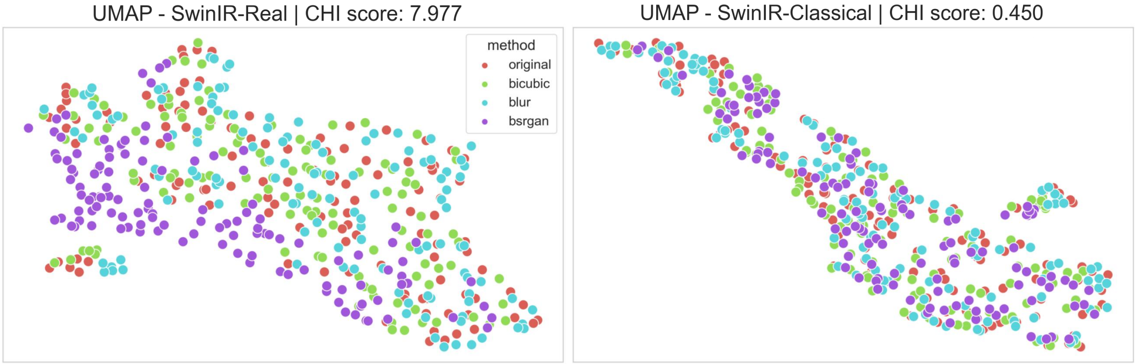
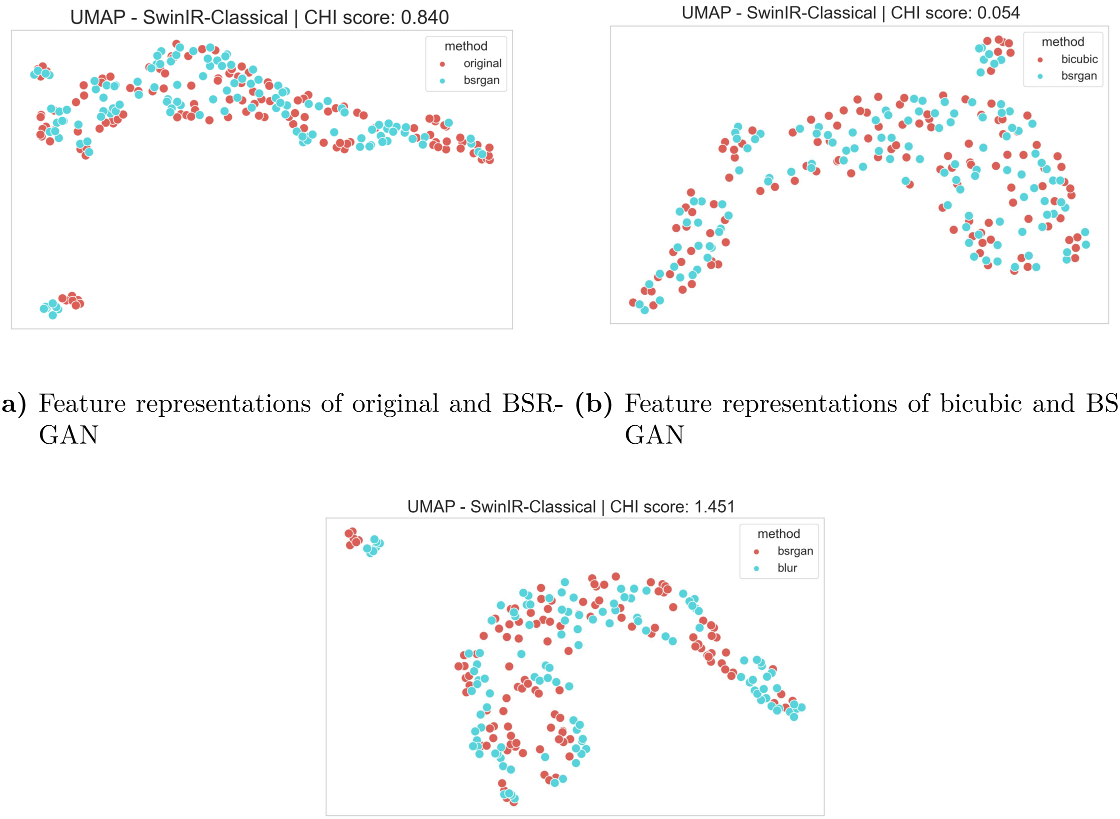


Figure 5.7: Deep feature representation differences between original and synthetic videos obtained by all degradation types (Left: SwinIR Classical, Right: SwinIR Real

Both visual and quantitative results in Figure 5.6 and Figure 5.7 reveal a surprising outcome, that is very different from the previous scenario. The classical model does not exhibit signs of degradation discrimination, while the real model can just barely differentiate images obtained by applying BSRGAN degradation from the rest. To affirm this statement we compare more degradation combinations in Figure 5.8:


UMAP - SwinIR-Classical | CHI score: 0.840

UMAP - SwinIR-Classical | CHI score: 0.054

UMAP - SwinIR-Classical | CHI score: 1.451

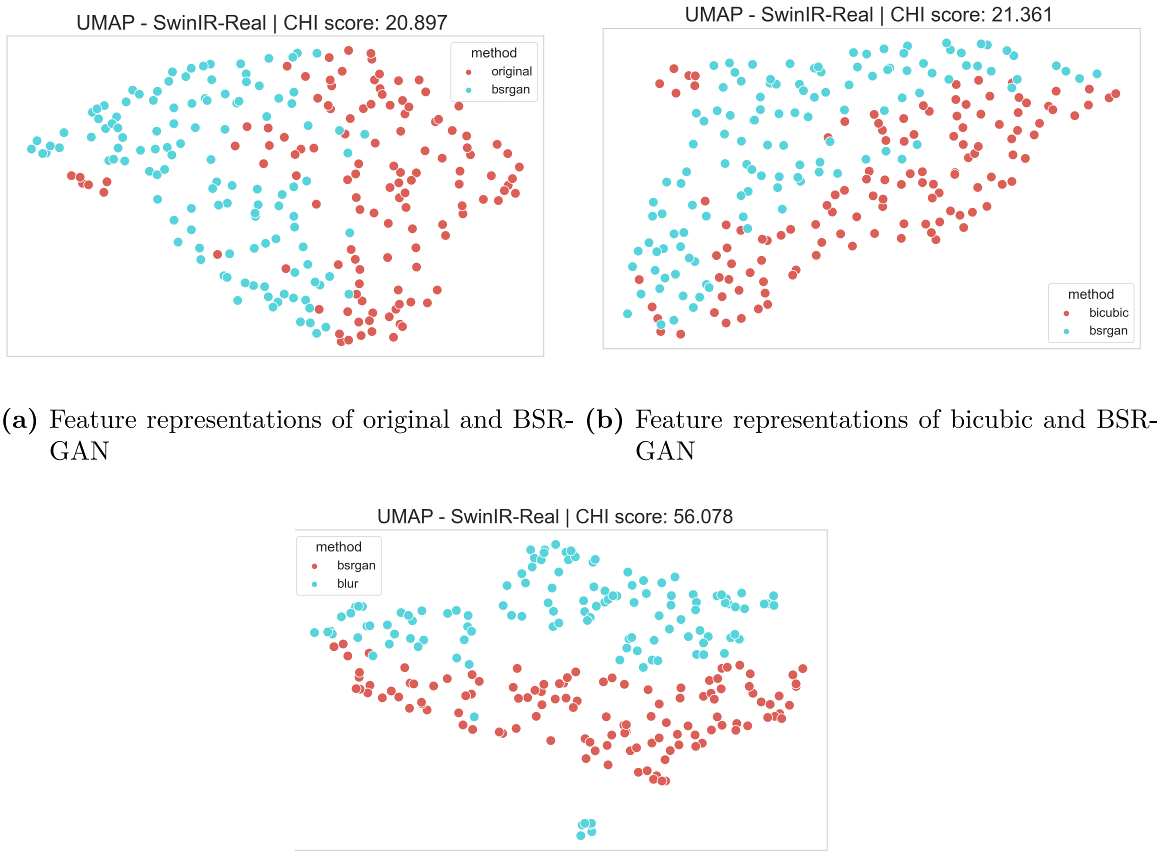
(a) Feature representations of original and BSRGAN

(b) Feature representations of bicubic and BSRGAN

(c) Feature representations of blur and BSRGAN

Figure 5.8: Deep feature representation differences for SwinIR Classical

It appears that the non-blind model does not possess the ability to discriminate between different forms of degradations, a theoretically strange behavior which would require more extensive research to fully understand. The same combinations with BSRGAN are tested for the Real version of the model in Figure 5.9



(a) Feature representations of original and BSRGAN

(b) Feature representations of bicubic and BSRGAN

(c) Feature representations of blur and BSRGAN

Figure 5.9: Deep feature representation differences for SwinIR Real

The real version, contrarily, seems to be capable of identifying BSRGAN degradations, a curious case considering that its design seeks better generalization by adapting to all kinds of inputs.

There are various things to consider in this situation. First, the parameters employed for BSRGAN are established by default settings. These may impose somewhat harsh conditions, leading to the generation of visually unrealistic images. However, on the flip side, such a rigorous approach can potentially enhance the model's performance in realistic scenarios. Secondly, the authors of SwinIR specify that the classical version has been trained on DIV2K [3]. According to the supplementary data [43] the training portion of the dataset is composed of images generated by the MATLAB bicubic kernel downsampling operation, which suggests that the discrepancy may be caused by other external factors.

SwinIR is based on a visual transformer architecture, a design that inherently leverages the principle of self-attention. The ability to establish relationships and dependencies between various regions of the image might be a contributing reason towards the obtained result.

The reason behind selecting SwinIR to complement the deep feature analysis is that both classical and real models share the same backbone, and the publicly available model can be used to directly extract the features. Even if SwinIR was not subjected to a quantitative evaluation, it is a good fit for the problem since the classical version is trained on bicubic data and the real version on BSRGAN data, two of the studied degradations. The unexpected behavior highlights the complexity of the problem at hand and the intricate relation between model architecture, training data, and model performance. Nevertheless, we believe the obtained results can help to shed light on the underlying mechanisms behind non-blind and blind model behavior and performance.

5.2.2. Evaluating the degradation generalization ability of VSR networks based on their deep features

Achieving better generalization is the main objective of blind SR networks. Until now, we have argued that models that learn similar feature representations for any input type exhibit a higher adaptation ability to real degradations. How to measure this generalization ability is complex yet very important, and there is no specific assessment to measure it. Moreover, the word 'generalization' can be ambiguous in this context. A SR that achieves great performance in any physical environment regardless of the lightning (indoor, outdoor, day, night...), content (people, nature, buildings...) would be labeled with a good generalization ability. However, this study does not consider the video or frame content, it focuses on degradation generalization, or the ability of a model to adapt and be resilient to any kind of input degradation.

As we have seen, IQA metrics are not ideal for this task, as they have several limitations, but most importantly the quality score does not represent generalization ability. A model's generalization ability should characterize the consistency of the model's processing effects across different types of input data, rather than absolute performance values like IQA metrics.

As an attempt to provide a quantitative evaluation of the generalization ability of SR networks, we base the evaluation protocol on the Generalization Assessment Index (GA Index) described in [54]. It is a non-parametric metric that is based on the statistical characteristics of internal features of the model, and it is applied to any test dataset.

The GA Index is computed using extracted feature values for each frame, which are transformed into probability distributions to compute the Kullback-Leibler divergence (KLD). In the publication, a generalized Gaussian distribution is used to model the feature sets, but we will use the Kolmogorov-Smirnov metric instead to avoid estimating the GGD parameters with our limited data.

Thus, given two feature sets that represent different degradations for the same model, $\mathrm{X}^{\mathcal{D}_1}$ and $\mathrm{X}^{\mathcal{D}_2}$, the Kolmogorov-Smirnov metric is computed as follows:

$$KS_{X_{\mathcal{D}1},X_{\mathcal{D}2}} = \sup x \left| F_{X_{\mathcal{D}1}}(x) - F_{X_{\mathcal{D}_2}}(x) \right| \tag{5.4}$$

where $F_{X_{\mathcal{D}1}}(x)$ and $F_{X_{\mathcal{D}2}}(x)$ are the empirical cumulative distribution functions (ECDF) of the feature sets $X_{\mathcal{D}1}$ and $X_{\mathcal{D}2}$, respectively. The supremum $\sup x$ denotes the maximum over all values of x.

Given a video sequence v_k, where k is the video number, we partition it into non-overlapping blocks of 8 frames. Each block is flattened to form a 1-D feature vector of $8*64 = 512$ features $\mathrm{X}_{j,k}^{\mathcal{D}_i}$ where j corresponds to the block within a video, k is the video number and $\mathcal{D}_i$ denotes the i^{th} degradation type. Features are normalized to account for a possible difference in

Method	BasicVSR	Real-BasicVSR	SwinIR Classical	SwinIR Real
Original	0.150	-	0.217	-
Bicubic	-	0.072	-	0.227
Blur	0.352	0.133	0.286	0.190
BSRGAN	0.305	0.150	0.232	0.375
Mean	0.269	0.118	0.245	0.264

Table 5.4: KS scores for each SR method

magnitude between methods. A model with strong generalization performance is expected to perceive the features extracted from degraded inputs as closely aligned as possible with the features of the training data distribution [54]. Thus, BasicVSR is expected to generate the best results in instances where the input has been subjected to bicubic interpolation, while Real-BasicVSR is designed to adapt to real-world scenarios.

For that reason, we compare the features from bicubic inputs with the rest of the degradations in the case of BasicVSR and we use original inputs as reference for RealBasicVSR. Then, we calculate the KS metric for each pair of reference-degradation, per block and degradation method. Table 5.4 shows average results by method, obtained by aggregating scores from all blocks.

We repeat the same process for SwinIR and observe the differences. According to our KS metric Real-BasicVSR achieves a better generalization ability, while the difference between blind and non-blind SWinIR models is very small. This confirms our previous suggestions, highlighting the resilience of RealBasicVSR and the abnormal behavior of SwinIR.

To get a better understanding of this idea, we plot the kernel density estimate (KDE) to visualize the distribution of observations, where KDE represents the data using a continuous probability density curve in two dimensions. In the same manner, we compare both network variants with each other. Figure 5.10 shows two representative examples from the dataset.

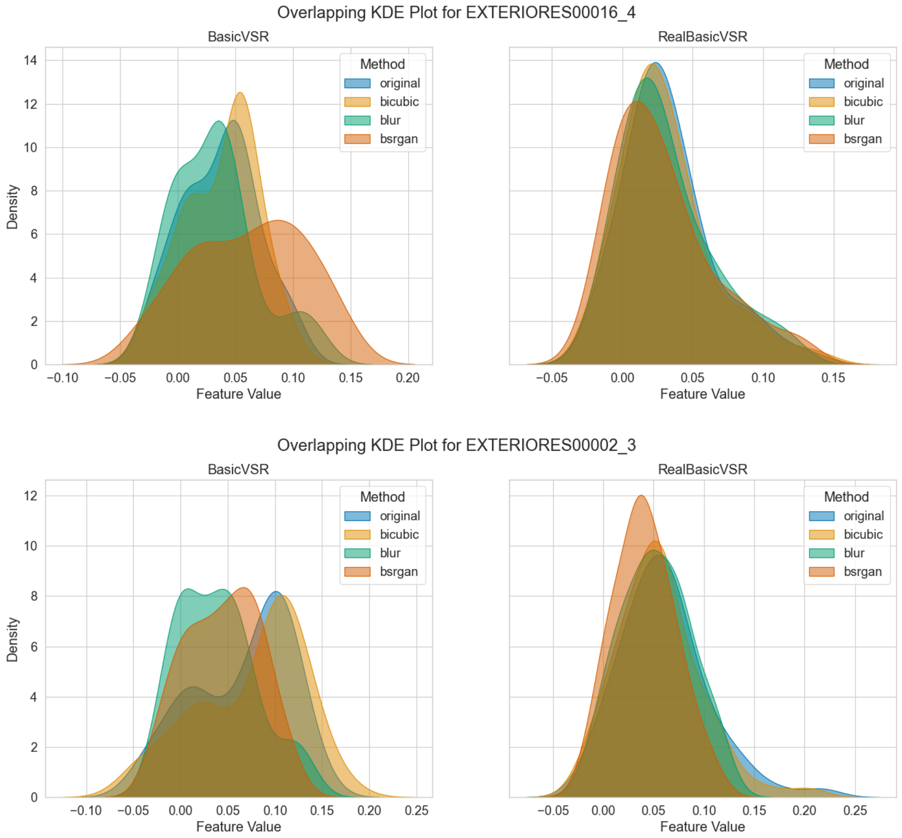


Figure 5.10: KDE plots of the features for example videos from the BSC4K dataset

5.3. Image frequency analysis

Image frequency analysis can be a fundamental tool in the study of images or videos, as it allows the examination of the image data in the frequency domain. This technique is widely used in digital forgery detection methods [130] [103], and it becomes particularly insightful to analyze the image frequency when studying the degradations in blind super resolution and detecting images that have been artificially upscaled. Various degradations, such as blurring, aliasing, or noise interference, can impact the frequency components of an image and thus alter its spectral characteristics. Understanding these transformations can provide valuable insights into the degradation process and lead to more effective super resolution strategies.

An important advantage of frequency analysis is that is is computationally more efficient, so it can be applied at a global level, considering the full resolution frame. However its application in the study of degradations is unexplored. We plot the video spectra by calculating the Discrete Cosine Transform (DCT) coefficients and following the procedure described in subsection 6.2.1. DCT coefficients correspond to frequency descriptors of videos, that provide information about the changes in low and high frequency, with the objective of finding significative similarities or differences between original and downscaled content. Figure 5.11 shows two representative examples of the obtained plots.

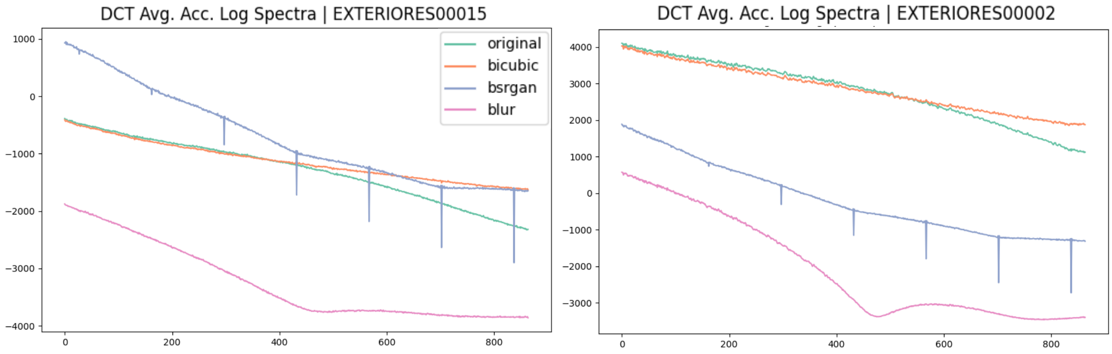


Figure 5.11: Accumulative Log Spectra of native 4K and upscaled methods

As represented in the visual data, bicubic and original method display very similar frequency characteristics, as their changes in high-frequency components are comparable. Given the original 4K images already are already detail-rich, the application of bicubic downscaling preserves those high-frequency details. Conversely, blurring the original image tends to result in a lower curve, a reasonable output considering the blur operation. Lastly, there are some noticeable sudden drops in the case of BSRGAN, predominantly caused by the more harsh degradations made of noise and blur operations, generating a distinctive pattern evident in the plot.

This data implies that the output generated by the internal processor of the DSLR camera contains similar high-frequency components to those obtained by using bicubic interpolation to downscale the 4K glshr video. Even so, the quantitative performance metrics still show that the input glslr image degradations have an impact on the results. This is not to suggest that training a SR model with either degradation is superior, but rather it attempts to better understand the effect of input degradations. Published models trained on ”Real” datasets claim to improve performance on real-world degradations, but it is complicated to make assumptions about their generalization ability. Ultimately, while this data provides insights into degradation impacts and SR model performance, it does not serve as conclusive evidence and serves as a process that could be beneficial in further research.

6 SR Detection

This section will be focused on high-resolution video content upscaling detection, a relatively underexplored area in digital forgery detection. First, we will review our proposed architecture, and then we will compare our results with other existing SR detection methods. In addition we present the results and observations from the conducted experiments to complement the research process.

6.1. Proposed system architecture

To try to improve the current state-of-the-art, we propose a network inspired by Lu et al.'s work [57] (BTURA). Our network's architecture is based on the feature extractor proposed in [57], which processes the small patches in the training dataset. It consists of a ResNet-18 (pre-trained on ImageNet [25]), where intermediate features are extracted from each block, grouped by a Global Average Pooling operation, and concatenated, as seen in Figure 6.1. In [57], a two-layer Multilayer Perceptron (MLP) takes the features from the feature extractor and outputs a probability for each category. We add a Dropout layer in the MLP to avoid overfitting and keep the softmax function to the output, which transforms the values into probabilities for each class. We refer to this architecture of a feature extractor and MLP as the baseline.

We name our architecture Synthetic Upscaling Detector with DCT features and Staircase module or SUDDS.

We propose to incorporate additional modules into the baseline architecture to evaluate their impact on performance and feature representations. The modules represent techniques or modifications that we believe should improve the performance. Furthermore, they can be optionally removed to keep the original baseline as it is.

Our goal in this context is generalization, as ideally, the model should be able to identify fake content upscaled with a method that is not in the training data. Our reproduction of the state-of-the-art methods (section 6.2) displayed in Table 6.1 show this task is far from solved, seeing that tested models seem to excessively adapt to the training methods.

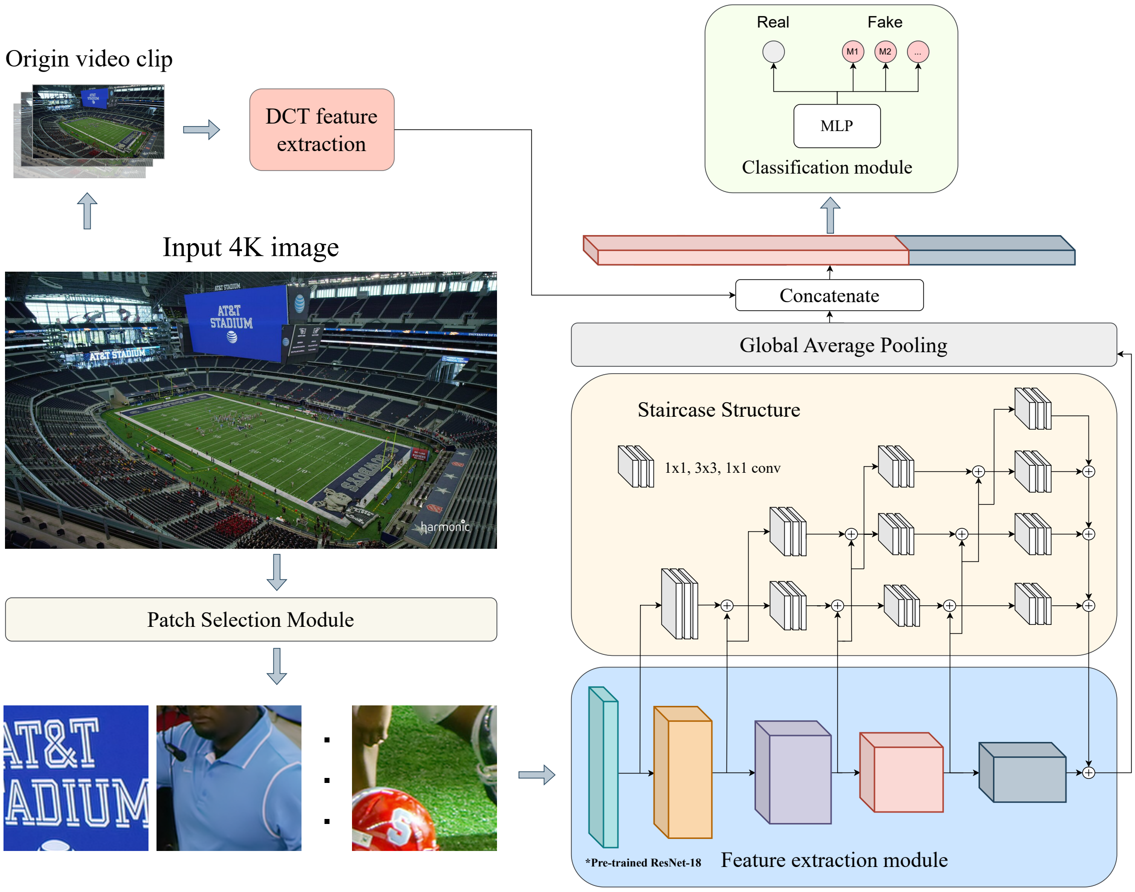


Figure 6.1: Network architecture of SUDDS. Original feature extraction and patch selection module from from Lu et al [57] and the DCT feature extraction and Staircase Structure are new additions

In our experiments, we attempt to improve performance by incorporating two new modules. First, the staircase structure, proposed in [93], attempts to fully utilize the visual information from low-level to high-level and learn the better feature representations for quality evaluation. The assumption is that the bottom convolutional layers from the ResNet capture the low-level information, such as edges and corners, while the more advanced layers capture the semantic information. The staircase architecture hierarchically integrates low and high-level features into a final feature map that is the input to the classifier.

The second module integrates a technique to combine local features from the patches and global features from the videos. By using the same methodology described in subsection 6.2.1, we save the Discrete Cosine Transforms (DCT) features for each video in the dataset. The DCT essentially decomposes an image to its spatial frequency spectrum. At training time, those features are concatenated with the local features from the feature extractor or staircase architecture. This means that patches from the same image will contain the same global DCT-based features, but different local features. Both modules are optional and evaluated independently, allowing for a broader understanding of the learning process at hand.

Finally, the concatenated feature map is fed into a classifier, constituted by a two-layer MLP (where the intermediate layer has a size of 256), which outputs a probability value for each class. We study the setting of binary classification, where all synthetic upscaling methods are grouped together, and multiclass classification, where each method is represented by an individual label.

6.2. Implementation of baselines

We evaluate the DCT-based Detector, TSARA and SRDM on our datasets. TSARA and SRDM are both publicly available. To provide additional perspective into the problem we implement the model presented in [116] and adapt the BTURA system from [57] (created from scratch), without the additional quality prediction.

6.2.1. Implementation of DCT-based Detecion

DCT is a widely used technique in image and video processing, especially in compression applications. In fact, it plays a crucial role in the JPEG (Joint Photographic Experts Group) compression algorithm, a lossy process that generally accomplishes perceptually great results.

DCT transforms a signal from the spatial domain into a representation in the frequency domain. It expresses a sequence of data points as a combination of cosine waves, which oscillate at different frequencies. In the context of SR detection, it is used to extract features from the frequency domain of the video frames to determine whether a video has been upscaled or not. DCT coefficients are summed along rows or columns to calculate the accumulative log spectra of a frame, to create a compact representation of the frequency content of each frame. Upscaling methods tend to introduce specific patterns in the frequency content of videos, so the followed methodology can analyze the features to detect the presence of those patterns. In particular, a sharp decline in the frequency information is used to make the final prediction. More specifically, the original paper highlights the presence of a decline at a normalized frequency of $1/s$, where s is the scaling factor. The reason behind this declining pattern is that upscaling methods can effectively increase the low-resolution component frequency, but often struggle to accurately enhance high-frequency details. As a result, the frequency content of an upscaled image tends to drop off sharply at a certain point

The implementation for [116] follows the process described in the paper. The 2D-DCT coefficients are obtained from the luminance component of video frames, so they are converted from RGB to YUV420 format. For easier analysis, the logarithm of the DCT spectrum is calculated, since most DCT coefficients are close to zero. Coefficients are then summed along rows or columns to calculate the accumulative log spectra of a frame:

$$\begin{aligned} \text{AccumLogSpecRow }_k[u] &= \sum_v \log\left(\left|YF_k[u,v]\right| + \text{ bias }\right) \\ \text{AccumLogSpecCol }_k[v] &= \sum_u \log\left(\left|YF_k[u,v]\right| + \text{ bias }\right) \end{aligned}$$

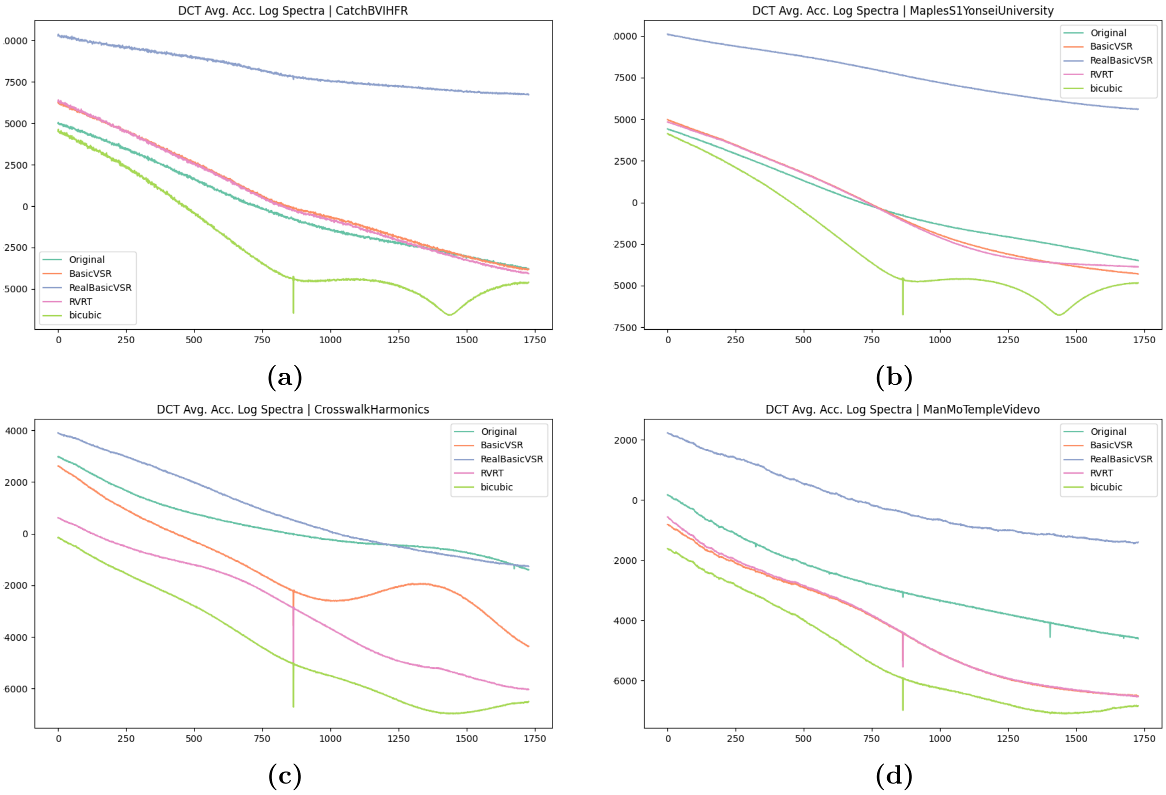
Where for N frames, luminance components of a video are $Y_k[\mathrm{i},\mathrm{j}], \mathrm{k} \in \{0,\ldots,N-1\}$, DCT coefficients of luminance components are $YF_k[u,v]$, and bias is set to 10^{-4} to prevent taking logarithm of zero. [116]. Then, the accumulative log spectra for a video is obtained by averaging the spectra of each frame:

$$\begin{aligned} \text{AveAccumLogSpecRow }[u] &= \frac{\sum_{k=0}^{N-1} \text{AccumLogSpecRow}_k[u]}{N} \\ \text{AveAccumLogSpecCol }[u] &= \frac{\sum_{k=0}^{N-1} \text{AccumLogSpecCol}_k[v]}{N} \end{aligned}$$

Next, a median filter is applied and used to calculate the difference between the spectra before and after filtering. This process allows the detection of sharp declines. Finally, the distribution of the log spectra difference is studied, where outlier values represent sharp declines in averaged accumulative log spectra, and indicate that a video is possibly upscaled. Finally, a video is considered fake if the ratio of the standard histogram is less than than an arbitrary threshold, set to -8 in the paper.

First, we use this method to visualize the average spectras of real and fake 4K videos and then evaluate it by calculating the accuracy on our data.

We put the method to the test with the BVI-DVC dataset, considering the real 4K videos combined with their upscaled versions, by four different upscaling techniques: bicubic interpolation, BasicVSR, Real-BasicVSR, and RVRT.


(a)

(b)

(c)

(d)

Figure 6.2: Several examples from calculated DCT average log spectras for videos in BVI-DVC-SR

Figure 6.2 shows four indicative examples of average log spectras. The first noticeable factor is that videos upscaled with bicubic interpolation show a similar pattern, where a clear sharp decline occurs at 1/2 of the frequency range. This phenomenon is consistent across all 200 videos, which displays a clear indication of the limiting effect of bicubic interpolation on the frequency content of the upscaled videos. This effect appears to be independent of the original video's content, suggesting a fundamental characteristic of bicubic interpolation that is used to detect fake videos. In regards to the DL-based methods, there is more irregularity. We observe that BasicVSR and RVRT behave in similar ways, probably due to their recurrent architecture and non-blind training setting. In some instances, there is no sign of a sharp decline, as in Figure 6.2a and Figure 6.2b, but in Figure 6.2c and Figure 6.2d, it is very clear. Lastly, we can see that RealBasicVSR shows a distinctive tendency to maintain a more consistent spectral distribution throughout the entire frequency range. This may indicate that blind-SR methods are more resistant to this particular detection method.

6.2.2. SRDM

To test SRDM, we use the provided pre-trained model based on ResNet. We follow the provided criterion for fake-resolution video detection. Considering all frames, if quantile of probabilities is bigger than 0.5 the video will be considered as fake-resolution video.

6.2.3. TSARA

We follow the same methodology to test TSARA. As it is a method for Image SR detection, we test each frame individually and then average the predicted probabilities. The final prediction corresponds to the closest value (0 or 1) from the average prediction among all video frames. This makes the method more robust to noise and other intra-frame variations.

6.3. Preliminary Performance comparison

As a preliminary evaluation step, we test the methods with 50 videos from BVI-DVC-SR and show the results in Table 6.1.

Method	DCT	SRDM	SRDM-Patches	TSARA
Bicubic	1.00	0.62	0.60	0.98
RVRT	0.72	0.46	0.42	0.44
BasicVSR	0.70	0.46	0.42	0.42
Real-BasicVSR	0.10	0.56	0.48	0.00
TOTAL REAL	0.30	**0.72**	0.70	0.68
TOTAL FAKE	**0.63**	0.52	0.48	0.46

Table 6.1: Consolidated performance metrics for all studied methods

The performance metrics displayed in Table 6.1 display considerable variation in accuracy for DCT. As expected, the bicubic method achieves the highest accuracy, as it was detected as fake 100% of the time. RVRT and BasicVSR present similar accuracy rates and reasonable results that show that there is still room for improvement in terms of VSR methods. On the other hand, the detection of original videos and videos upscaled by Real-BasicVSR seem to present a bigger challenge, as reflected by their low accuracy rates. The 0.3 accuracy for original videos indicates that the current system cannot properly identify real samples that have not been upscaled, as most of them were classified as fake. The misclassification could be due to a misinterpretation of the inherent noise or natural variations in the original videos as synthetic manipulations or distortions. Lastly, we observe a particularly low accuracy of Real-BasicVSR over the other methods, displaying the difference between non-blind and blind methods' outputs. As seen in Figure 6.2, it can consistently create smooth curves that fool the detection method, leading to a high rate of misclassification.

As an attempt to find a threshold that allows the model to achieve higher performance, we plot the distributions of calculated threshold distribution for each method.

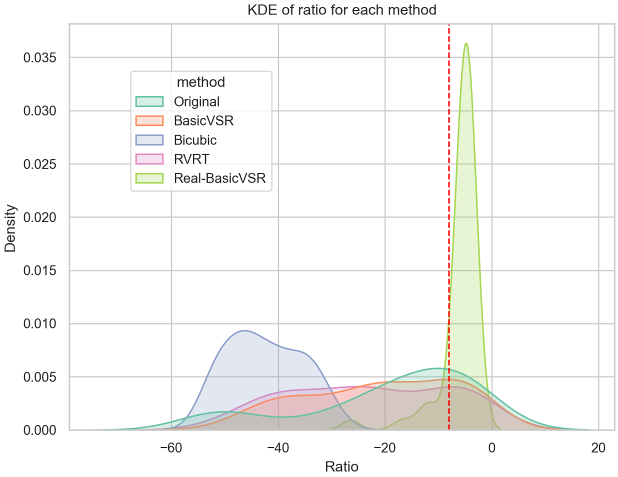


Figure 6.3: Ratio distribution per method. The vertical red line indicates the tested threshold (-8) to distinguish real and fake videos

Observing the results in Figure 6.3, it is evident that setting a constant threshold value will not yield optimal results. While this approach may be convenient and straightforward, the data distribution in Figure 6.3 shows significant variations among modern VSR methods, which cannot be accounted for by a constant threshold value. Therefore, we conclude that the current implementation is not an effective way to directly detect real and fake videos, although the extracted DCT features could be useful to train another classification model.

For SRDM, as results in Table 6.1 show, the model cannot distinguish real and fake videos. Results are somewhat expected, as the model has been trained on 224 × 224 images, which forces 4K videos to be downscaled to match the required input size.

To solve that problem, we attempt to crop each frame into smaller patches and use them as input, but results prove to be similar.

In the case of TSARA, videos upscaled with bicubic interpolation seem easier to identify. In fact, TSARA correctly classified 98% of bicubic videos but showed significantly lower accuracy for RVRT and BasicVSR, both below 50%. This suggests that modern DL-based VSR methods introduce characteristics that make the task of identifying the upscaling significantly more challenging for the current network, trained on other data. Finally, TSARA did not correctly predict any of the videos upscaled with Real-BasicVSR.

6.4. Results

This section will cover the detailed findings derived from the conducted experiments. First, we will describe the training data and evaluation methodology. Then, we will compare our results with the existing detection methods. Finally, we will show the experimentation and ablation process to get to the final detection system.

In our experiments, we use the ResNet-based feature extractor, we fix the optimizer AdamW with a learning rate of 0.002 and the intermediate layer of the MLP to 256. All metrics during the training phase are assessed at the level of individual patches. However, when it comes to testing videos, the evaluation is conducted at video-level as a whole.

6.4.1. Training Data and Evaluation Methodology

6.4.1.1. Training Data

The dataset used to train and validate the networks is BVI-DVC-SR, which consists of 200 videos of 64 frames each, split into original 4K videos and their corresponding upscaled counterparts, for a total of 1,000 videos. The 200 videos are divided into 175 videos for training and 25 videos for validation.

Following, [57], the model is trained on smaller non-overlapping image patches of 240 × 240, which serve as input for the feature extractor. For each 4K frame, there exist 144 non-overlapping patches. Yet, using all 144 patches from all video frames would significantly increase the dataset size, leading to longer training times and increased computational requirements. Moreover, it could lead to overfitting, due to the redundancy in the data, as consequent frames for a video are usually very similar. Lastly, including all patches would mean processing a multitude of contentless pictures like skies or other plain background textures.

To create the final training dataset, which we call BVI-DVC-SR-Patches we select 120 patches from the first frame of each video, thus avoiding temporal redundancy. We select the patches to create the training set by employing the Grey-level Co-occurrence Matrix (GLCM). GLCM quantifies the texture complexity of an image by evaluating the frequency in which pairs of pixel brightness values occur in a given space. For each frame, we can select the top k patches with the highest texture complexity, which will contain the high-frequency components used to identify real and upscaled videos. This approach not only cuts down the computational load but also ensures that the model's attention is focused on the most informative parts of each frame. This GLCM-based approach is also convenient for testing purposes. Incoming sequences of frames can be reduced to a lower number of patches that condense the most complex information, facilitating the evaluation process. Moreover, modifying the number of selected patches can assist in adapting videos from a range of resolutions. GLCM is defined in section A.1

As a last processing step. we remove all crops that contain a plain texture, as there exist videos where a big portion of each frame contains a plain color with no variation.

If the temporal dimension was necessary, we create a smaller subset (which is not used in this thesis), BVI-DVC-SR-Video-Patches. It comprises eight equally spaced patches extracted from the initial eight frames of each video. The total number of images for this version is $8 \cdot 8 \cdot 200 = 12,800$. A certain amount of visual redundancy is to be expected, due to the similarity among frames of the same video. However, the temporal dimension is kept, as there are sequences of 8 frames for each crop in a video.

Therefore, we create BVI-DVC-SR-Patches, which encompasses around 24,000 patches per method. We consider the original patches, and the corresponding upscaled patches (obtained by upscaling the glslr original images with bicubic, RVRT, BasicVSR, and Real-BasicVSR).

6.4.1.2. Evaluation Methodology

The ultimate purpose of this method is to discern real and fake videos. The network deals with image crops, so an extra data pipeline is needed to process videos directly (videos are equivalent to frame sequences) and match the model's input size.

Evaluation videos are processed individually, where each frame is divided into k patches selected by the GLCM method. Each frame is first classified by a majority voting mechanism based on the k individual predictions. A patch is considered fake if it belongs to any of the synthetic method classes in the training set (if it is not classified as original). Finally, if at least half of frames are identified as fake, the video is considered unauthentic or upscaled.

We test our SUDDS with videos from BSC4K, upscaled with methods that are not included in the training set. Specifically, we employ one traditional (nearest neighbor interpolation), two transformer-based (SwinIR-Classical, SwinIR-Real), and one GAN-based (Real-ESRGAN) method to evaluate the generalization ability.

6.4.1.3. Frequency Analysis for SR Detection

As a first step to understanding the data and the effect of upscaling methods, we contribute to the study of frequency component visualization by showing our obtained frequency spectra graphs for four upscaling methods: bicubic interpolation, BasicVSR, Real-BasicVSR, and RVRT and also incorporate a GAN-based SR method, Real-ESRGAN. Following prior work [103] [76], we perform a high-pass filtering by subtracting the real image from its median blurred version. Then, we calculate the Fourier transform for a more informative visualization, as in [103]. We first try capturing the frequency spectra for each frame in a video and averaging the values. Then, we select a set of random images from each method and do the same thing. After a careful look at the results, we do not find any common pattern in any of the methods, even if there are some distinguishable shapes. Figure 6.4 shows examples of the obtained visualizations.

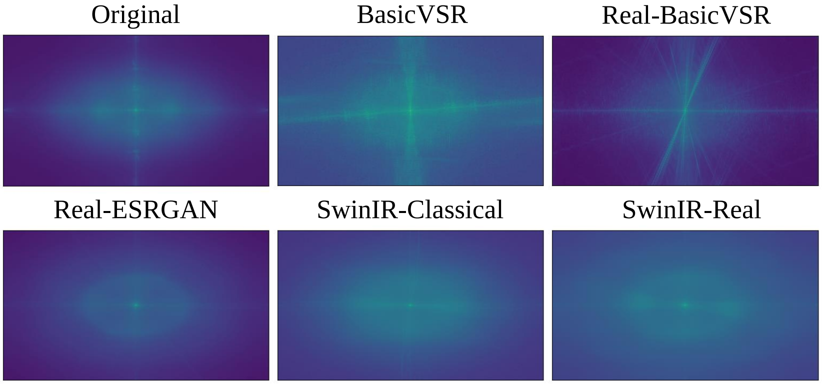


Figure 6.4: Average spectra of each high-pass filtered image, for several blind and non-blind SR methods.

This frequency analysis method was originally applied in the detection of CNN-generated content, which is related by distinct from SR. Authors of [103] showed consistent patterns of artifacts for GAN generated images, but it was later proved that content generated by LDMs does not exhibit such characteristics. The high-pass filter seems to be ineffective at describing any upscaling method, although other frequency analysis tools can be employed.

6.4.2. Comparison with Existing Detection Methods

In Table 6.2, we compare the existing detection methods with our own, on a varied list of upscaling methods.

Model	DCT	TSARA	SRDM-Patches	SUDDS (ours)
Original	0	0.72	0.88	**0.94**
SwinIR-Real	0.4	0	0.2	**0.9**
SwinIR-Classical	1	0.3	0.65	**1**
Real-ESRGAN	0	0	0.68	**0.94**
Nearest Neighbor.	1	0.05	0.4	**0.9**
BasicVSR	1	0.44	0.4	0.9*
Real-BasicVSR	0.06	0	0.59	1*
RVRT	1	0.36	0.42	0.8*
Bicubic	1	1	0.85	1*

Table 6.2: Accuracy Metrics for all studied SR and detection methods. * denotes SR methods that are in the training set

As we can see, our proposal outperforms the rest on unseen modern upscaling methods. This test is performed on our BSC4K dataset considering approximately 30 of the total videos.

6.4.3. Performance Across Upscaling Methods

In this set of experiments we evaluate the model's generalization ability training it with a single upscaling method from the training set and testing it with the rest.

As Table 6.3 shows, the model has a strong tendency to overfit to the training upscaling method, suggesting that it is adapting to the specific characteristics and patterns of the training data. This supports the use of blind methods, and suggests the use of a large variety of upscaling methods to increase model performance across domains.

	Validation Method				
Training Method	Original	Bicubic	BasicVSR	Real-BasicVSR	RVRT
Bicubic	0.963	0.999	0.156	0.001	0.172
BasicVSR	0.981	0.751	0.982	0.015	0.977
RealBasicVSR	0.999	0.001	0.001	0.990	0.001
RVRT	0.940	0.754	0.981	0.014	0.980

Table 6.3: Accuracy scores of each training-validation pair. The network is trained on four individual methods and evaluated with all of them. Each cell represents the performance score of the network when trained with one specific training method and validated with the method from the column.

Results in Table 6.3 indicate there is a clear similarity between the artifacts produced by BasicVSR and RVRT, as a model trained on one is capable of detecting the other with high accuracy. This may suggest that the two methods share similar characteristics and generate the same patterns in the outscaled images. They already showed a similar behaviour in terms of quantitative results in section 5.1, and both share a recurrent video architecture. Hence, the result is consistent with the previous observations.

Generally, the model trained with one specific method can accurately identify original videos, but it shows a propensity to misclassify fake videos from outside the training distribution.

In the following experiments, we consider all four methods as part of the training dataset. On one hand, we seek for a model that can effectively discern real and fake videos from any upscaling method. On the other hand, we are interested in understanding what the model is learning, that is, how differently it is processing all kinds of inputs.

6.4.4. Overall Performance

In this section we compare our best model with the baseline on the test data, in both binary and multiclass training.

We consider two ways of approaching the classification problem. First, as a binary classification task where one class corresponds to original videos and the other to upscaled videos, regardless of their upscaling method. We compare both ideas in Table 6.4

	Train acc.	Val. acc.	Original	Bicubic	BasicVSR	Real-BasicVSR	RVRT
Binary	0.944	0.98	0.920	0.991	0.995	0.99	0.98
Multiclass	0.93	0.89	0.95	0.99	0.75	0.99	0.75

Table 6.4: Baseline architecture comparison for binary and multiclass training

At first glance, validation shows that training the network with only two classes achieves better performance. To get a more informative view, we display the features extracted from our dataset in Figure 6.5, only considering training methods.

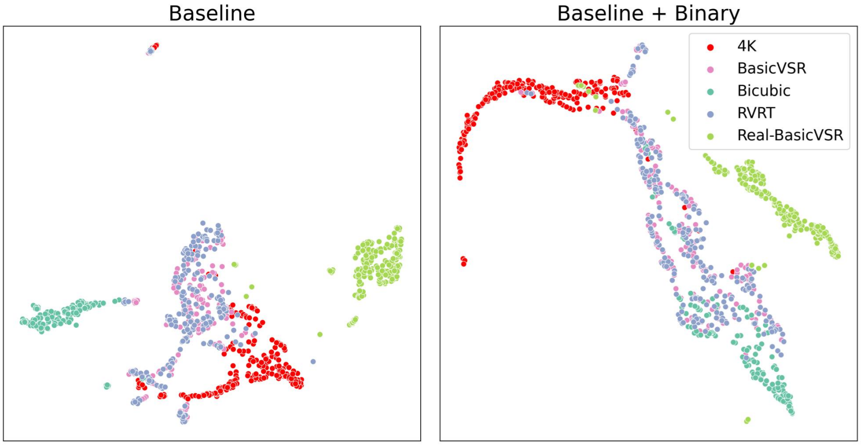


Figure 6.5: Feature representations for the baseline architecture with multiclass (left) and binary (right) training.

Both approaches yield comparable results in terms of feature visualizations. In binary classification, all upscaling methods share the same label, but the network has learned to distinguish them to some degree. Table 6.4 reinforces that BasicVSR and RVRT behave similarly, and are closer to real images that the rest. The network trained with a multiabel approach seems to add more nuance in the case of bicubic interpolation. However, the original cluster is closer to BasicVSR and RVRT, which explains the lower accuracy, as it is easier to distinguish them.

In an attempt to obtain more explainable results, we continue our experiments with the multiclass approach, and incorporate the Staircase and DCT feature modules to explore the networks behaviour. Table 6.5 shows the impact in accuracy of the different techniques.

Staircase	DCT	Train	Val.	Original	Bicubic	BasicVSR	Real-BasicVSR	RVRT
No	No	0.93	0.89	0.95	0.99	0.75	0.999	0.75
Yes	No	0.93	0.88	0.92	0.99	0.68	0.99	0.8
Yes	Yes	0.78	0.78	0.97	0.99	0.4	0.99	0.64

Table 6.5: Performance comparison with module combinations in multiclass classification.

Observing Table 6.5, we notice a low impact of the Staircase module. It slightly increases accuracy for RVRT but drops performance for BasicVSR. The full (all three modules) network achieves the lowest performance, but the most balanced in terms of training and accuracy metrics.

To evaluate their generalization ability, we test them on our dataset with unseen upscaling methods.

Staircase	DCT	Original	SwinIR-C	SwinIR-R	NN	Real-ESRGAN
No	No	0.9	1	0.5	0.06	0.5
Yes	No	0.63	1	0.7	0.1	0.79
Yes	Yes	0.94	1	0.9	0.9	0.94

Table 6.6: Test performance metrics for multiclass classification. NN represent Nearest Neighbor

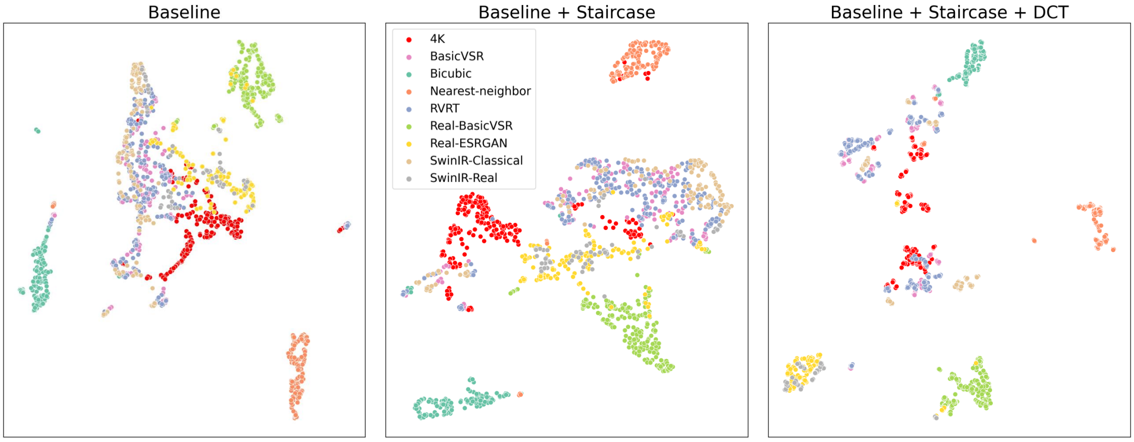


Figure 6.6: Feature representations for all training and testing data

Looking at Figure 6.6, we can see how the upscaling are grouped according to the detection network. In general, it seems like BasicVSR and RVRT show similar features, a pattern that we discovered in earlier sections. Traditional methods (bicubic and nearest neighbor) seem to be the most particular, as they form separated clusters evey time. In addition, it seems like the distinction between original and synthetic content is not clear, and there are instances where that difference is not perceptible at feature level. Finally, blind methods (SwinIR Real, BasicVSR, Real-ESRGAN) seem to be grouped in the same space, suggesting that they possess similar artifacts, which help the model identify them.

To complete our research, we repeat the same process for binary classification:

Staircase	DCT	Train	Val.	Original	Bicubic	BasicVSR	Real-BasicVSR	RVRT
No	No	0.994	0.98	0.92	0.99	0.995	0.999	0.98
Yes	No	0.996	0.98	0.91	0.99	0.998	1	0.99
Yes	Yes	0.99	0.99	0.98	1	0.993	0.993	0.991

Table 6.7: Performance comparison with module combinations in binary classification.

Table 6.7 shows comparable results in all cases, with a minor improvement with DCT in the original sample detection. To better investigate this behavior, we turn into the feature visualization (see Figure 6.7) and test metrics (Table 6.8).

Staircase	DCT	Original	SwinIR-C	SwinIR-R	NN	Real-ESRGAN
No	No	0.61	1	0.09	0.95	0.5
Yes	No	0.62	1	0.1	0.94	0.5
Yes	Yes	1	0.8	0.1	1	0.21

Table 6.8: Test performance metrics for binary classification

Looking at the test metrics, we can see that accuracy is very high for some methods but low for others, which suggests that its generalization ability is poor.

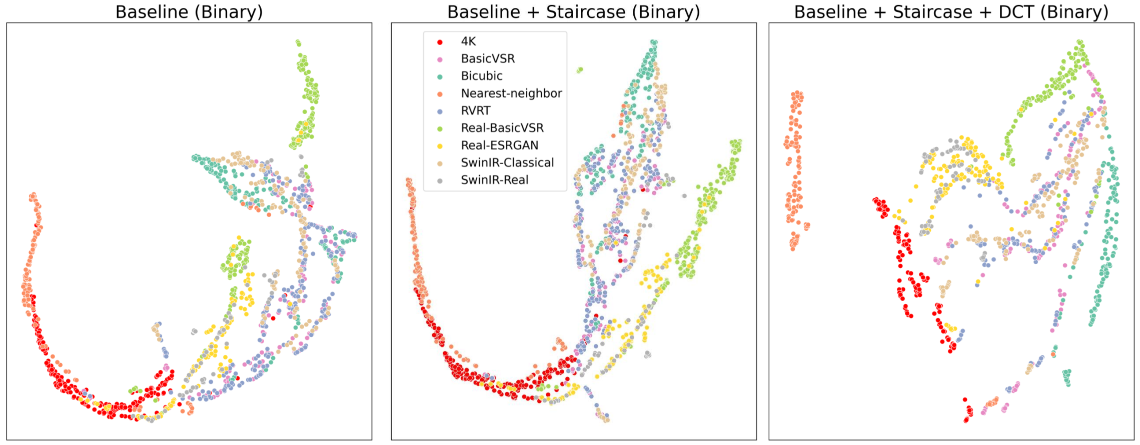


Figure 6.7: Feature visualization from binary classification and all tested models

Figure 6.7 displays the feature representations for the same components in a binary classification setting. The observed clusters form more distinct shapes compared with the ones in multiclass. However, the SR methods are grouped together in a similar manner.

6.4.4.1. Ablation study

We perform an ablation study on two components: data augmentation and freezing/training the feature extractor.

In order to achieve a more robust model and reduce overfiting, we turn to the practice of data augmentation. Table 6.9 represents the impact of the selected data augmentation and the effect of freezing the feature extractor's layers during the training phase. The performance metrics provided are for the training and validation sets.

	Aug.	Freeze	Train	Val.	Original	Bicubic	BasicVSR	Real-BasicVSR	RVRT
Baseline	No	Yes	0.93	0.806	0.88	0.94	0.61	0.98	0.6
Baseline	Yes	No	0.64	0.62	0.5	0.86	0.4	0.97	0.32
Baseline	Yes	Yes	0.93	0.89	0.95	0.99	0.75	0.999	0.75

Table 6.9: Performance comparison with and without data augmentation

The data augmentation pipeline includes includes a set of modifications made with random probabilities to increase the data variety, all from the Albumentations library [1]. First, we use a Horizontal Flip transformation with a 0.35 probability. To modify the images but keep the texture details we adjust the brightness condition with a 0.5 application rate. The brigthess is adjusted up or down by up to 20%. Finally, Coarse Dropout is applied to encourage the network to find more patterns that identify fake images.

As Table 6.9 shows, the combination of data augmentation and retraining the feature extractor yield the most positive rewards. Thus, we have fixed these parameters in the previous experiments (subsection 6.4.4).

6.5. Advantages and Limitations

There are two main advantages of the proposed system. Firstly, it can be easily adapted for image or video SR detection, as the DCT module is optional and the other modules are based on image crops. Secondly, the individual crop predictions can be aggregated in diverse ways, which allows the search for more suitable thresholds. The network outputs predictions for individual patches, which can be first aggregated to predict the authenticity of a single frame. Similarly, a video comprises numerous frames, each contributing towards the final overall video-level prediction.

By investigating the network's feature representations, we attempt to enhance the system's explainability. Observing the clusters that are formed during training can provide a helpful insight into the underlying patterns and relationships learned by the model. Moreover, the detection of anomalies or outliers could aid in the post-training analysis process.

Despite the advantages, there exist some limitations that need to be considered. One limitation of the proposed method is its evaluation process. The lack of a standard dataset for testing purposes makes it difficult to benchmark the performance and reliability of SR detection methods. What is more, it is impossible to anticipate if and when new upscaling methods will be proposed in this very active field, and if the developed method will be able to detect those. That being said, it is easy to imagine how to adapt the proposed model to new upscaling models released: Simply by adding samples of those to the model's training set.

The dataset needs to be expanded to include compression artifacts in its LR samples. For doing so, it would be ideal to use a variety of compressors, encoders and any other video codec popularly used in streaming services, smartphones or digital media companies.

7 Discussion

Through the exploration of SR techniques, their effectiveness, and generalization ability, this thesis offers valuable insight into non-blind and blind SR models. Our study of blind models addresses their resilience against different degradations, in comparison with non-blind models. We show that non-blind models can learn *semantics* about the training degradation process, which makes them less suitable for real-world scenarios where the input degradations are more diverse. To compare these techniques under equal conditions, we propose a dataset created through a manual collection of high-resolution videos (4K and 1080p), providing a unique benchmark for high-definition video SR.

This dataset allows us to calculate quantitative performance metrics and compare the impact of different input types. We demonstrate the domain-gap that exists for non-blind models, making them less robust in comparison to blind models. Our proposal includes the analysis of deep features from SR networks to measure their degradation generalization ability.

Further, we investigate the relatively under-studied area of 4K video SR detection, highlighting the need for more sophisticated and publicly available detection methods. We compare the performance of existing methods and propose an improvement by extending a pre-existing architecture. We train it on a more diverse range of upscaled content to analyze the learning process of synthetic upscaling identification. We find that in practice, the task of generalizing to unseen upscaling methods is not trivial, as the model tends to overfit to training SR techniques. This explains the decrease in performance for existing methods on more modern SR and VSR methods. We place emphasis on the training process, including data and learning strategy.

In this respect, studying feature representations proved to be a helpful tool for understanding the learning process and the decisions made by the prediction network. Results suggest that a more simple CNN-based feature extraction module can learn to accurately discern between upscaling methods that are in the training dataset. However, the generalization of these results to other upscaling methods poses a bigger challenge, as it is common in any sort of detection problem. We propose to add global features from the video level to create a system that exhibits superior performance and improved generalization ability.

There are not many public datasets in 4K resolution, and fewer that contain content specifically upscaled using a range of modern SR methods. This scenario constrains the training and validation processes, potentially introducing complications as a consequence of inferior quality data. This may introduce unwanted biases, negatively affecting performance. Moreover, the comparison with existing or future methods also poses a challenge, as there are no standardized SR detection benchmarks.

We hope that our research and datasets can help pave the way for further exploration in this domain. We will continue exploring and expanding the ideas in this thesis in hopes of uncovering new insights and better understanding the complex world of SR.

7.1. Challenges and Future Directions

Despite the meaningful progress accomplished in this thesis, there are several challenges to be addressed. The field of SR is an area of research that covers a wide amount of methodologies and applications. There are several key distinctions, such as image-video, blind-non-blind, HR-LR, and supervised-unsupervised models. In the future, we would like to expand the current framework to consider more types of architectures and data sources. This would consolidate our findings and provide a more insightful point of view on the studied areas.

In such a manner, including a set of quantitative metrics that better align with human judgment would significantly enhance the evaluation process. It would encourage further research to adopt more varied metrics and address the limitations of popular methods like PSNR and SSIM.

In the thesis, we propose a new dataset used to evaluate the degradation generalization ability and performance of blind models. While there are comparable datasets available, we create a strategy that allows the recording of paired 4K and 1080p videos without further post-processing techniques. Nonetheless, there are some aspects that should be addressed in the future. First, a more in-depth study of the internal process of the device to capture the video pairs would help clarify how they differ from other degradation methodologies, such as bicubic interpolation. In connection with the previous point is the fact that our dataset has been obtained with a single camera, which only captures degradations exclusive to the device. It serves as a great tool for blind SR method analysis, but real-world images and videos contain an extensive amount of degradation types. Moreover, a more complete version of the dataset could be used to train or fine-tune a blind SR network, enhancing its capability to handle real degradations. In this regard, a color correction mechanism could make the HR-LR frame pairs more similar in terms of pixel value, as there is a difference caused by the camera could that impact the training process of a network.

In future work, we plan to expand our dataset to include a more diverse set of glslr and glshr samples. Our first goal is to incorporate a wider variety of downsampled content. While our dataset currently contains original glslr videos and degraded glslr videos through bicubic downsampling, blur, and BSRGAN, there is potential to include additional degradation methodologies. For instance, we could downsample our glshr videos with Real-ESRGAN or similar degradation models.

Moreover, we intend to include samples downscaled via third-party software. A possible approach for this is uploading the videos to platforms like YouTube and downloading them through browser plugins. This would make the dataset account for common compression artifacts that are generated through internet transmission.

Lastly, our work has contributed to improving the detection of synthetic 4K content by highlighting the main challenges when training a SR detection model. The rapidly advancing field of DL continues to develop better and more powerful models, rendering previous detection models ineffective in these new instances. In our study, we try to cover a reasonable range of SR network types. However, it is hard to extrapolate our conclusions to the entire spectrum of existing and forthcoming models. On that account, increasing the number of training and test models and datasets would provide a richer understanding of the differences between them when it comes to the detection task. Furthermore, the generation of synthetic 4K videos is a quite demanding process computationally, due to the increased resolution of the input compared with the testing data. This is amplified in the case of videos, where several frames are usually processed concurrently. For that reason, we will investigate more efficient SR models, which may yield worst quality results but provide the advantage of computational efficiency.

In conclusion, this thesis has offered a comprehensive exploration of several SR techniques, particularly in the domains of degradation effect in blind SR models, detection of upscaled content, and model generalization. Our contributions have highlighted the current limitations and also offered a methodology to better explain the behavior of SR models. However, the advancements in media consumption requirements (higher resolution) and the landscape of DL continue to introduce new challenges and considerations. Therefore, it is necessary to expand on the current research, to achieve more accurate SR models, evaluation metrics, and detection mechanisms. The insights and methods proposed in this thesis provide an understanding of several areas of SR upon which future research can build and expand, to further comprehend their capabilities in such a challenging field.

Bibliography

[1] A. Buslaev, A. P. "Albumentations: fast and flexible image augmentations." In: *ArXiv e-prints* (2018). eprint: `1809.06839`.

[2] Aakerberg, A., Nasrollahi, K., and Moeslund, T. B. "Real-world super-resolution of face-images from surveillance cameras." In: *IET Image Processing* 16.2 (2022), pp. 442–452.

[3] Agustsson, E. and Timofte, R. "Ntire 2017 challenge on single image super-resolution: Dataset and study." In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops.* 2017, pp. 126–135.

[4] Ahmad, W. et al. "A new generative adversarial network for medical images super resolution." In: *Scientific Reports* 12.1 (2022), p. 9533.

[5] Ahn, N., Kang, B., and Sohn, K.-A. "Fast, accurate, and lightweight super-resolution with cascading residual network." In: *Proceedings of the European conference on computer vision (ECCV).* 2018, pp. 252–268.

[6] Akhtar, Z. "Deepfakes Generation and Detection: A Short Survey." In: *Journal of Imaging* 9.1 (2023), p. 18.

[7] Anwar, S., Khan, S., and Barnes, N. "A deep journey into super-resolution: A survey." In: *ACM Computing Surveys (CSUR)* 53.3 (2020), pp. 1–34.

[8] Cai, J. et al. "Ntire 2019 challenge on real image super-resolution: Methods and results." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 2019.

[9] Cai, J. et al. "Toward real-world single image super-resolution: A new benchmark and a new model." In: *Proceedings of the IEEE International Conference on Computer Vision.* 2019.

[10] Caliński, T. and Harabasz, J. "A dendrite method for cluster analysis." In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[11] Chan, K. C. et al. "Basicvsr: The search for essential components in video super-resolution and beyond." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021, pp. 4947–4956.

[12] Chan, K. C. et al. "BasicVSR++: Improving video super-resolution with enhanced propagation and alignment." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2022, pp. 5972–5981.

[13] Chan, K. C. et al. "Investigating tradeoffs in real-world video super-resolution." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022, pp. 5962–5971.

[14] Chen, C. et al. "Camera lens super-resolution." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2019, pp. 1652–1660.

[15] Chen, C. et al. “Learning to see in the dark.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018, pp. 3291–3300.

[16] Chen, D. et al. “Human Guided Ground-truth Generation for Realistic Image Super-resolution.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2023, pp. 14082–14091.

[17] Chen, H. et al. “Pre-trained image processing transformer.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021, pp. 12299–12310.

[18] Chen, H. et al. “Real-world single image super-resolution: A brief review.” In: *Information Fusion* 79 (2022), pp. 124–145.

[19] Chen, Y. et al. “Brain MRI super resolution using 3D deep densely connected neural networks.” In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018).* IEEE. 2018, pp. 739–742.

[20] Cheon, M. and Lee, J.-S. “Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience.” In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.7 (2017), pp. 1467–1480.

[21] Cisco, U. “Cisco annual internet report (2018–2023) white paper.” In: *Cisco: San Jose, CA, USA* 10.1 (2020), pp. 1–35.

[22] Croitoru, F.-A. et al. “Diffusion models in vision: A survey.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[23] Dai, T. et al. “Second-order attention network for single image super-resolution.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2019, pp. 11065–11074.

[24] *Dareful - Royalty Free 4k & HD Stock Video Footage Clips — dareful.com.* `https://dareful.com/`.

[25] Deng, J. et al. “Imagenet: A large-scale hierarchical image database.” In: *2009 IEEE conference on computer vision and pattern recognition.* Ieee. 2009, pp. 248–255.

[26] Dhariwal, P. and Nichol, A. “Diffusion models beat gans on image synthesis.” In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794.

[27] Dong, C. et al. “Image super-resolution using deep convolutional networks.” In: *IEEE transactions on pattern analysis and machine intelligence* 38.2 (2015), pp. 295–307.

[28] Gao, S. et al. “Implicit diffusion models for continuous super-resolution.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2023, pp. 10021–10030.

[29] Gu, J. et al. “NTIRE 2022 challenge on perceptual image quality assessment.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2022, pp. 951–967.

[30] Gu, J. et al. “Deep residual squeeze and excitation network for remote sensing image super-resolution.” In: *Remote Sensing* 11.15 (2019), p. 1817.

[31] Huang, J.-B., Singh, A., and Ahuja, N. "Single Image Super-Resolution From Transformed Self-Exemplars." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 5197–5206.

[32] Ignatov, A., Van Gool, L., and Timofte, R. "Replacing mobile camera isp with a single deep learning model." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 536–537.

[33] Isobe, T. et al. "Video super-resolution with recurrent structure-detail network." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer. 2020, pp. 645–660.

[34] Katsenou, A. V., Sole, J., and Bull, D. R. "Content-gnostic bitrate ladder prediction for adaptive video streaming." In: *2019 Picture Coding Symposium (PCS)*. IEEE. 2019, pp. 1–5.

[35] Katsenou, A. V., Sole, J., and Bull, D. R. "Efficient bitrate ladder construction for content-optimized adaptive video streaming." In: *IEEE Open Journal of Signal Processing* 2 (2021), pp. 496–511.

[36] Keys, R. "Cubic convolution interpolation for digital image processing." In: *IEEE transactions on acoustics, speech, and signal processing* 29.6 (1981), pp. 1153–1160.

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

[38] Kotevski, Z. and Mitrevski, P. "Experimental comparison of psnr and ssim metrics for video quality estimation." In: *ICT Innovations 2009*. Springer. 2010, pp. 357–366.

[39] Ledig, C. 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.

[40] Li, B. et al. "Learning to predict the quality of distorted-then-compressed images via a deep neural network." In: *Journal of Visual Communication and Image Representation* 76 (2021), p. 103004.

[41] Li, J., Pei, Z., and Zeng, T. "From beginner to master: A survey for deep learning-based single-image super-resolution." In: *arXiv preprint arXiv:2109.14335* (2021).

[42] Li, Z. et al. "AVC, HEVC, VP9, AVS2 OR AV1?—A comparative study of state-of-the-art video encoders on 4K videos." In: *Image Analysis and Recognition: 16th International Conference, ICIAR 2019, Waterloo, ON, Canada, August 27–29, 2019, Proceedings, Part I 16*. Springer. 2019, pp. 162–173.

[43] Liang, J. et al. "Supplementary for SwinIR: Image Restoration Using Swin Transformer." In: ().

[44] Liang, J. et al. "Swinir: Image restoration using swin transformer." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 1833–1844.

[45] Liang, J. et al. "Recurrent video restoration transformer with guided deformable attention." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 378–393.

[46] Liang, J. et al. “Vrt: A video restoration transformer.” In: *arXiv preprint arXiv:2201.12288* (2022).

[47] Lim, B. et al. “Enhanced deep residual networks for single image super-resolution.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 136–144.

[48] Lim, B. et al. “Enhanced Deep Residual Networks for Single Image Super-Resolution.” In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017.

[49] Liu, A. et al. “Blind image super-resolution: A survey and beyond.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[50] Liu, C. and Sun, D. “On Bayesian adaptive video super resolution.” In: *IEEE transactions on pattern analysis and machine intelligence* 36.2 (2013), pp. 346–360.

[51] Liu, H et al. “Video super resolution based on deep learning: A comprehensive survey. arXiv 2020.” In: *arXiv preprint arXiv:2007.12928* (2020).

[52] Liu, X. et al. “Exploit camera raw data for video super-resolution via hidden Markov model inference.” In: *IEEE Transactions on Image Processing* 30 (2021), pp. 2127–2140.

[53] Liu, Y. et al. “Discovering Distinctive” Semantics” in Super-Resolution Networks.” In: *arXiv preprint arXiv:2108.00406* (2021).

[54] Liu, Y. et al. “Evaluating the generalization ability of super-resolution networks.” In: *arXiv preprint arXiv:2205.07019* (2022).

[55] Liu, Z. et al. “Swin transformer: Hierarchical vision transformer using shifted windows.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[56] Lowe, D. G. “Distinctive image features from scale-invariant keypoints.” In: *International journal of computer vision* 60 (2004), pp. 91–110.

[57] Lu, W. et al. “Deep Neural Network for Blind Visual Quality Assessment of 4K Content.” In: *IEEE Transactions on Broadcasting* (2022).

[58] Lyapustin, E. et al. “Towards true detail restoration for super-resolution: A benchmark and a quality metric.” In: *arXiv preprint arXiv:2203.08923* (2022).

[59] Ma, D., Zhang, F., and Bull, D. “BVI-DVC: a training database for deep video compression.” In: *IEEE Transactions on Multimedia* (2021).

[60] Maaten, L. Van der and Hinton, G. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[61] Mackin, A., Zhang, F., and Bull, D. R. “A study of subjective video quality at various frame rates.” In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2015, pp. 3407–3411.

[62] Mackin, A., Zhang, F., and Bull, D. R. “A study of high frame rate video formats.” In: *IEEE Transactions on Multimedia* 21.6 (2018), pp. 1499–1512.

[63] Madhusudana, P. C. et al. “Capturing video frame rate variations through entropic differencing.” In: *arXiv e-prints* (2020), arXiv–2006.

[64] Madhusudana, P. C. et al. “Subjective and objective quality assessment of high frame rate videos.” In: *IEEE Access* 9 (2021), pp. 108069–108082.

[65] Martin, D. et al. “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics.” In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001.* Vol. 2. IEEE. 2001, pp. 416–423.

[66] Mercat, A., Viitanen, M., and Vanne, J. “UVG dataset: 50/120fps 4K sequences for video codec analysis and development.” In: *Proceedings of the 11th ACM Multimedia Systems Conference.* 2020, pp. 297–302.

[67] Meshchaninov, V., Molodetskikh, I., and Vatolin, D. “Combining contrastive and supervised learning for video super-resolution detection.” In: *arXiv preprint arXiv:2205.10406* (2022).

[68] Mittal, A., Moorthy, A. K., and Bovik, A. C. “No-reference image quality assessment in the spatial domain.” In: *IEEE Transactions on image processing* 21.12 (2012), pp. 4695–4708.

[69] Mittal, A., Soundararajan, R., and Bovik, A. C. “Making a “completely blind” image quality analyzer.” In: *IEEE Signal processing letters* 20.3 (2012), pp. 209–212.

[70] MMEditing Contributors. *MMEditing: OpenMMLab Image and Video Editing Toolbox.* `https://github.com/open-mmlab/mmediting`. 2022.

[71] Mudunuri, S. P. and Biswas, S. “Low resolution face recognition across variations in pose and illumination.” In: *IEEE transactions on pattern analysis and machine intelligence* 38.5 (2015), pp. 1034–1040.

[72] Nah, S. et al. “Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.* 2019, pp. 0–0.

[73] *NETFLIX OPEN CONTENT — opencontent.netflix.com.* `https://opencontent.netflix.com/`.

[74] Nilsback, M.-E. and Zisserman, A. “Automated flower classification over a large number of classes.” In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing.* IEEE. 2008, pp. 722–729.

[75] NVIDIA. *Pixel Perfect: RTX Video Super Resolution Now Available | NVIDIA Blog.* `https://blogs.nvidia.com/blog/2023/02/28/rtx-video-super-resolution/`. 2023.

[76] Ojha, U., Li, Y., and Lee, Y. J. “Towards universal fake image detectors that generalize across generative models.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2023, pp. 24480–24489.

[77] Pengxu Wei, Z. X. “Component Divide-and-Conquer for Real-World Image Super-Resolution.” In: *Proceedings of the European Conference on Computer Vision.* 2020.

[78] Ramesh, A. et al. “Hierarchical text-conditional image generation with clip latents.” In: *arXiv preprint arXiv:2204.06125* (2022).

[79] Rao, R. R. R. et al. “AVT-VQDB-UHD-1: A large scale video quality database for UHD-1.” In: *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE. 2019, pp. 17–177.

[80] Rec, I. “P. 910: Subjective video quality assessment methods for multimedia applications.” In: *International Telecommunication Union, Geneva* 2 (2008).

[81] *REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206`. 2021.

[82] Reibman, A. R., Bell, R. M., and Gray, S. “Quality assessment for super-resolution image enhancement.” In: *2006 International Conference on Image Processing*. IEEE. 2006, pp. 2017–2020.

[83] Reibman, A. R. and Schaper, T. “Subjective performance evaluation of super-resolution image enhancement.” In: *Second Int. Wkshp on Video Proc. and Qual. Metrics (VPQM’06)* (2006).

[84] Saharia, C. et al. “Image super-resolution via iterative refinement.” In: *arXiv:2104.07636* (2021).

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

[86] SAMSUNG. *How to use the Intelligent Mode of Samsung QLED TV | Samsung CA*. `https://www.samsung.com/ca/support/tv-audio-video/how-to-use-the-intelligent-mode-of-samsung-qled-tvs/`.

[87] Shah, R. R., Akundy, V. A., and Wang, Z. “Real versus fake 4K-authentic resolution assessment.” In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 2185–2189.

[88] Shi, W. 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.

[89] Shocher, A., Cohen, N., and Irani, M. ““zero-shot” super-resolution using deep internal learning.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3118–3126.

[90] Simonyan, K. and Zisserman, A. “Very deep convolutional networks for large-scale image recognition.” In: *arXiv preprint arXiv:1409.1556* (2014).

[91] Song, L. et al. “The SJTU 4K video sequence dataset.” In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE. 2013, pp. 34–35.

[92] Sun, W. et al. “A deep learning based no-reference quality assessment model for ugc videos.” In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 856–865.

[93] Sun, W. et al. “Blind Quality Assessment for in-the-Wild Images via Hierarchical Feature Fusion Strategy.” In: *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE. 2022, pp. 01–06.

[94] Sun, W. et al. “Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training.” In: *IEEE Journal of Selected Topics in Signal Processing* (2023).

[95] Team, F. *FFmpeg*. 2000.

[96] Tian, C. et al. “Generative adversarial networks for image super-resolution: A survey.” In: *arXiv preprint arXiv:2204.13620* (2022).

[97] Vaswani, A. et al. “Attention is all you need.” In: *Advances in neural information processing systems* 30 (2017).

[98] Vavilala, V. and Meyer, M. “Deep Learned Super Resolution for Feature Film Production.” In: *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks*. 2020, pp. 1–2.

[99] Video Processing, c. and group, quality research. *Video Upscalers Benchmark*. `https://videoprocessing.ai/benchmarks/video-upscalers.html`.

[100] *VQEG/siti-tools: SI TI calculation tools*. `https://github.com/VQEG/siti-tools`.

[101] Wang, L. “A survey on IQA.” In: *arXiv preprint arXiv:2109.00347* (2021).

[102] Wang, R. et al. “Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces.” In: *arXiv preprint arXiv:1909.06122* (2019).

[103] Wang, S.-Y. et al. “CNN-generated images are surprisingly easy to spot... for now.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8695–8704.

[104] Wang, X. et al. “ESRGAN: Enhanced super-resolution generative adversarial networks.” In: *The European Conference on Computer Vision Workshops (ECCVW)*. 2018.

[105] Wang, X. et al. “Real-esrgan: Training real-world blind super-resolution with pure synthetic data.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1905–1914.

[106] Wang, X., Ma, J., and Jiang, J. “Contrastive learning for blind super-resolution via a distortion-specific network.” In: *IEEE/CAA Journal of Automatica Sinica* 10.1 (2022), pp. 78–89.

[107] Wang, Z., She, Q., and Ward, T. E. “Generative adversarial networks in computer vision: A survey and taxonomy.” In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–38.

[108] Wang, Z. and Bovik, A. C. “Mean squared error: Love it or leave it? A new look at signal fidelity measures.” In: *IEEE signal processing magazine* 26.1 (2009), pp. 98–117.

[109] Wang, Z., Simoncelli, E. P., and Bovik, A. C. “Multiscale structural similarity for image quality assessment.” In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[110] Wang, Z. et al. “Image quality assessment: from error visibility to structural similarity.” In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[111] Wei, Y. et al. “Unsupervised real-world image super resolution via domain-distance aware training.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 13385–13394.

[112] Winkler, S. “Analysis of public image and video databases for quality assessment.” In: *IEEE Journal of Selected Topics in Signal Processing* 6.6 (2012), pp. 616–625.

[113] Xu, X., Ma, Y., and Sun, W. “Towards real scene super-resolution with raw images.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1723–1731.

[114] Xue, T. et al. “Video enhancement with task-oriented flow.” In: *International Journal of Computer Vision* 127 (2019), pp. 1106–1125.

[115] YANG, X. et al. “Real-world Video Super-resolution: A Benchmark Dataset and A Decomposition based Learning Scheme.” In: (2021).

[116] Yang, Z. et al. “Native resolution detection for 4k-uhd videos.” In: *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE. 2020, pp. 1–5.

[117] Ying, Z., Pan, D., and Shi, P. “Blind Video Quality Assessment for Ultra-High-Definition Video Based on Super-Resolution and Deep Reinforcement Learning.” In: *Sensors* 23.3 (2023), p. 1511.

[118] You, C. et al. “CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE).” In: *IEEE transactions on medical imaging* 39.1 (2019), pp. 188–203.

[119] Yuan, Y. et al. “Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 701–710.

[120] Yue, H., Zhang, Z., and Yang, J. “Real-RawVSR: Real-World Raw Video Super-Resolution with a Benchmark Dataset.” In: *European Conference on Computer Vision*. Springer. 2022, pp. 608–624.

[121] Zeyde, R., Elad, M., and Protter, M. “On single image scale-up using sparse-representations.” In: *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*. Springer. 2012, pp. 711–730.

[122] Zhai, G. and Min, X. “Perceptual image quality assessment: a survey.” In: *Science China Information Sciences* 63 (2020), pp. 1–52.

[123] Zhang, D. et al. “Remote sensing image super-resolution via mixed high-order attention network.” In: *IEEE Transactions on Geoscience and Remote Sensing* 59.6 (2020), pp. 5183–5196.

[124] Zhang, F. *BVI-SR Database. University of Bristol*. 2020.

[125] Zhang, K., Zuo, W., and Zhang, L. “Learning a single convolutional super-resolution network for multiple degradations.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3262–3271.

[126] Zhang, K. et al. “Designing a practical degradation model for deep blind image super-resolution.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 4791–4800.

[127] Zhang, R. et al. “The unreasonable effectiveness of deep features as a perceptual metric.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[128] Zhang, W. et al. “Blind image quality assessment using a deep bilinear convolutional neural network.” In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.1 (2018), pp. 36–47.

[129] Zhang, W. et al. “A closer look at blind super-resolution: Degradation models, baselines, and performance upper bounds.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 527–536.

[130] Zhang, X., Karaman, S., and Chang, S.-F. “Detecting and simulating artifacts in gan fake images.” In: *2019 IEEE international workshop on information forensics and security (WIFS)*. IEEE. 2019, pp. 1–6.

[131] Zhang, X. et al. “Zoom to Learn, Learn to Zoom.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[132] Zhu, W. et al. “Perceptual quality assessment for recognizing true and pseudo 4K content.” In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 2190–2194.

List of acronyms

AI Artificial Intelligence.

BRISQUE Blind/Referenceless Image Spatial Quality Evaluator.

CNN Convolutional Neural Network.

DCT Discrete Cosine Transform.

DL Deep Learning.

DNN Deep Neural Network.

GAN Generative Adversarial Network.

GLCM Grey-level Cooccurrence Matrix.

HR High-Resolution.

ISR Image Super-Resolution.

LDM Latent Diffusion Model.

LPIPS Learned Perceptual Image Patch Similarity.

LR Low-Resolution.

MISR Multi-Image Super-Resolution.

ML Machine Learning.

MLP Multilayer Perceptron.

MSE Mean Squared Error.

NIQE Natural Image Quality Evaluator.

PCA Principal Component Analysis.

PSNR Peak Signal-to-Noise Ratio.

RNN Recurrent Neural Network.

SISR Single-Image Super-Resolution.

SR Super-Resolution.

SSIM Structural Similarity Index.

SVM Support Vector Machine.

t-SNE t-Distributed Stochastic Neighbor Embedding.

VSR Video Super-Resolution.

A
Anexo I

A.1. GLCM calculation

$$P(i,j|d,\theta) = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\begin{cases} 1 & \text{if } I(x,y) = i \text{ and } I(x+\Delta_x, y+\Delta_y) = j \\ 0 & \text{otherwise} \end{cases} \tag{A.1}$$

Where $P(i,j)$ is the GLCM matrix and $p_{i,j}(d,\theta)$ is the pixel at location (i,j), for a specific displacement d and direction θ.

B
Additional Tables

Table B.1: CHI scores for all degradation combinations and all them united

CHI score	Blur - BSR-GAN	Bicubic - BSR-GAN	Original - Bicubic	Original - Blur	Original - BSR-GAN	Blur - Bicubic	All degradations
SwinIR Classical	1.11 ± 0.21	0.06 ± 0.02	0.64 ± 0.23	0.01 ± 0.01	1.28 ± 0.30	0.65 ± 0.30	0.61 ± 0.12
SwinIR Real	43.22 ± 6.03	22.38 ± 2.28	0.17 ± 0.07	0.03 ± 0.02	24.62 ± 2.79	0.50 ± 0.20	6.19 ± 1.77
BasicVSR	3418.60 ± 1515.84	3652.98 ± 1515.66	35.76 ± 4.96	770.52 ± 220.83	1873.02 ± 848.58	1753.16 ± 588.46	1401.87 ± 521.56
RealBasicVSR	20.59 ± 2.99	193.93 ± 25.38	0.48 ± 0.07	24.97 ± 3.74	168.28 ± 16.86	58.18 ± 3.78	61.47 ± 2.97