UNIVERSITY OF WEST ATTICA

&

UNIVERSITY OF LIMOGES

Master Thesis

Comparative analysis of modern ESRGAN models in the ill-posed problems of blind super resolution and motion-blurred /low-light image restoration.

Student: Georgios Rouselatos

AIVC22016

Supervisor: Anastasios L. Kesidis, Professor

Egaleo, Athens - February 2025

Μέλη εξεταστικής επιτροπής συμπεριλαμβανομένου και του εισηγητή

Η διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή

Α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
1	Anastasios Kesidis	Professor	ANASTASIOS KESIDIS Digitally signed by ANASTASIOS KESIDIS Date: 2025.02.26 13:42:59 +02'00'
2	Paris Mastorokostas	Professor	PARIS MASTOROKOSTAS Digitally signed by PARIS MASTOROKOSTAS Date: 2025.02.26 13:22:52 +02'00'
3	Nikolaos Vasilas	Professor	NIKOLAOS VASILAS Digitally signed by NIKOLAOS VASILAS Date: 2025.02.26 14:51:01 +02'00'

The work for this thesis was fulfilled with the invaluable assistance and guidance of Dr. Anastasios L. Kesidis, whose overall supervision was a positive catalyst throughout the whole process.

Table of Contents

Abstract

This work explores the task of super-resolution using advanced GAN-based architectures tailored specifically for this purpose. A primary focus is addressing the inherent challenges of the ill-posed nature of super-resolution by incorporating synthetic images in various ways, emphasizing the importance of model selection that integrates synthetic image approaches. Additionally, we investigate the relationship between super-resolution and related image reconstruction tasks such as image deblurring and low-light image enhancement through transfer learning. Our experiments reveal the capabilities of state-of-the-art super-resolution models while also highlighting their potential and limitations, particularly in handling cues associated with deblurring and low-light scenarios. Despite their high performance in super-resolution, these models struggle with the complexities of upscaling processes inherent in these tasks, underscoring the necessity for pretraining on specialized datasets to address these challenges effectively. This study contributes insights into the strengths and limitations of advanced super-resolution techniques and emphasizes the importance of tailored training for addressing specific image enhancement tasks.

1. Introduction

Super-resolution is a pivotal challenge in the popular field of image processing and computer vision, aiming to enhance the spatial resolution of an image beyond its original acquisition capabilities, therefore transiting from a low resolution (LR) representation to a high resolution (HR) one. In various real-world applications such as medical imaging, surveillance, and satellite imagery, the ability to reconstruct high-quality, detailed images from low-resolution inputs is crucial. This problem becomes particularly pronounced when faced with limitations in hardware or data acquisition processes. Researchers and engineers grapple with the task of developing sophisticated algorithms and models to bridge the gap between low and high-resolution representations, opening avenues for clearer, more detailed visual information in scenarios where precision is paramount. The pursuit of effective super-resolution techniques addresses the inherent limitations of imaging systems and holds significant implications for advancing related fields of application.

There is a vast spectrum of challenges akin to image super-resolution, with each of these challenges addressing specific facets of visual data enhancement. A closely related problem is image de-blurring, (Figure 1.1) which aims to counteract blurriness induced by motion, defocus, or other factors. In this case, a blurry image is provided to an appropriate model which is tasked with returning a reconstruction of high image quality, while preserving concepts and visual aspects of the original image. In a similar sense there are challenges such as image dehazing, which addresses atmospheric interference by enhancing visibility and contrast in outdoor scenes, as well as image de-noising that attempts to refine image quality by ensuring a clearer representation of visual information in the reconstructed image. In cases where the original image is highly distorted due to increased levels of blurriness or noise in general, an accurate reconstruction can be significantly challenging, because fine details may be totally missing in the distorted image. Another area of works involves enhancements of the original image, encompassing a broader goal of refining overall image quality by adjusting factors such as brightness, contrast, sharpness, and color balance. Specifically, low-light image enhancement (Figure 1.2) presents several interesting research aspects, focusing on improving the visibility and quality of images captured under low-light conditions. In scenarios where insufficient illumination leads to reduced image clarity and increased noise, low-light image enhancement aims to bring out essential details, enhance contrast, and mitigate visual artifacts. A main difficulty attached to this problem is that even ground truth images may be dark enough so that certain details are hard to be discerned even by the human eye, posing more reconstruction obstacles in comparison to other enhancement tasks.

Figure 1.1*: An example of the task of image de-blurring. The image on the right is a reconstruction of the original, blurry image on the left [8].*

Figure 1.2: *An example of the task of low-light image enhancement. A dark image, obtained under low-light image settings on the left can reveal its details and obtain a proper illumination and color balance when reconstructed (right) [9].*

In the quest for achieving higher resolution images, there have been proposed several methods, mainly revolving around Generative Adversarial Networks (GANs) [1]. Specifically, the idea behind GANs employs two neural networks called generator (G) and discriminator (D), which compete in a zero-sum game fashion: the generator G produces an image by receiving a random noise vector z in its input, while the discriminator D decides upon the realism of the generated image by classifying it in the appropriate category, based on real images it has been trained on. The feedback from D drives G towards generating more and more realistic images as the training process proceeds, as G attempts to fool D by generating images lying closer to the training distribution. This procedure is called adversarial training. At a later stage, D becomes unable to discern between a real and a fake instance, therefore the equilibrium is reached, and the training process terminates. The trained G is now able to generate new images which are indistinguishable in comparison to the real images present in the training dataset –at least if G has been trained appropriately-. A variance of the aforementioned process involves the addition of a conditioning vector c together with the random noise z in the input of G; in the case of super-resolution, this conditioning can refer to a low-resolution image.

At this point, it is essential to obtain some LR sample images from existing HR ones; the conversion from HR to LR is important in order to maintain ground-truth pairings, in cases that such ground truth pairings are not provided in a dataset. For this reason, the following degradation function is employed [2]:

$$Ix = D(Iy, \delta), \quad (1)$$

Where Ix is the LR image, Iy its HR counterpart, δ is a suitable degradation parameter (such as noise, or any other degrading factor) and finally D is an unknown degradation function that creates the

desired mapping from the HR image set to a LR space. Often, the D function is represented as a downsampling process that utilizes a scaling factor s to produce a LR image; however, other operations are also favorable in literature, such as blurring, adding Gaussian noise or a combination of degradation techniques [2].

Consequently, a given generative model M will be tasked to perform the inverse process, i.e. transform a given LR image to a HR one: a GAN receives a LR image Ix as conditioning c among with the random noise z as inputs to G, while D aims to discriminate between the output generated from G and given corresponding HR instances Iy from the dataset. Nevertheless, we can easily assume that the way the degradation process is performed influences the reconstructed LR image Ix, since the generative model M is trained on a different mapping between the LR and the HR image distributions. In other words, given a LR image dataset and a set of models which have been trained on varying LR-HR mappings due to varying degradations, the corresponding output HR images will be different, based on the model M that produced them. Even in the case of a single generative model M that is tasked to perform the super-resolution process, a possible variability in the training hyperparameters, as well as the probabilistic nature of the generative process itself may lead to various HR outcomes. Notably, the generated HR images can be plausible and of high visual realism; therefore, by just considering images possessing desired characteristics in terms of quality, the super-resolution problem remains inherently ambiguous.

Apart from the original super-resolution task, related challenges such as de-blurring or low-light image enhancement also face ambiguity in reconstruction. For example, a significantly blurry image may be impossible to accurately recover, and a GAN should hypothesize on specific concepts or visual details to produce a realistic HR image. Similarly, an image captured under very low light conditions is possible to permanently miss concepts and details in shades, therefore they may not be recovered at all.

The aforementioned challenges place image super-resolution under the category of *ill-posed problems*. Formally, an ill-posed problem contains at least one of the following three characteristics [3]:

1. *No Unique Solution*: An ill-posed problem may have multiple solutions, and it may not be possible to determine a single, correct answer.
2. *Lack of Stability*: Small changes or perturbations in the input data or conditions can lead to significant variations in the output, making the problem sensitive and unstable.
3. *Solution Existence*: There may be cases where the problem does not have a solution, or the solution does not exist in the context of the given problem formulation.

Based on these, super-resolution satisfies the first and the second criteria: different models or hyperparameter settings result in a *non-unique solution*, while inherent stochasticity of generative models (mainly attributed to the random noise receiving in their input, such as the z vector in GANs) is tied to the observed *lack of stability*, even if the same model under the same settings is employed. The *solution existence* criterion is satisfied, since a generative model M will produce an output in any case, while a ground truth mapping between LR and HR images exists and is well-defined (as occurring from the degradation function (1)).

An example of the ill-posed nature of image super-resolution is presented in Figure 1.3. Specifically, the leftmost image is the ground-truth LR one, while the rest are proposed HR reconstructions provided by various generative models. We can observe variability in texture, even though they are all deemed as plausible reconstructions, since important semantics (such as color and generic texture cues) are preserved during reconstruction.

Figure 1.3*: On the ill- posedness of image super-resolution: a LR image (left) may receive several valid reconstructions from different models [11].*

The ill-posedness of super-resolution and its related reconstruction tasks (such as de-blurring and low-light image enhancement) correlates with several real-world concerns, where trained models are leveraged in applied super-resolution challenges. For example, real-world degraded images suffer from various artifacts [2]; these can be often connected with the image acquisition technique, as in the case of medical images where the imaging machine produces different artifacts in comparison to another one. Other artifacts can be correlated with the compression method utilized, for example employing a certain compression framework (e.g. JPEG) over another. Additional real-world degradations involve blurring and adaptive noise [2], which may be hard to be simulated through a well-defined degrading function (1), due to their unconstrained occurrence in real images. Overall, due to the ill-posed nature of super-resolution, the discrepancy between a well-crafted and strictly constrained LR dataset and a LR dataset containing real degradations poses significant constrains and is ultimately responsible for highly varying HR outcomes. These challenges have been addressed in prior literature [4, 5, 6], but were not fully resolved. At the same time, the evaluation ambiguity tied with image generation problems [7] bring the additional challenge of automatically defining the best possible HR reconstruction for a given LR image dataset.

State-of-the-art literature explores novel trajectories to tackle challenges related to image super-resolution tasks. One of the most promising endeavors is centered around the concept of localization training with adaptive target generation, as analyzed in [17], where mismatches between a ground truth HR image and a valid HR reconstruction are explored. Specifically, the utilization of an adaptive target training technique in super-resolution is advantageous for addressing the inherently ill-posed nature of the problem, offering increased flexibility and dynamism in learning the mapping from LR to HR images. Rather than adhering to a fixed ground truth target during training, the method of adaptive target generation enables the model to dynamically adjust its target patch according to the characteristics of the input image. Consequently, the model can adapt to the specific features and details inherent in each image. Through the integration of this tactic, the super-resolution model becomes adept at emphasizing pertinent details and structures within the input

image. This targeted focus results in a more precise and context-aware reconstruction of high-resolution images. The adaptability provided by this strategy plays a crucial role in mitigating the ill-posed nature of super-resolution, guiding the model to make well-informed decisions based on the unique content of each individual image. The outcome is an enhancement in reconstruction quality and a more effective preservation of image features.
Another promising way of tackling the inherently ill-posed nature of super-resolution tasks in state-of-the-art systems is the incorporation of synthetic data into the training process. Synthetic data, generated through a variety of data augmentation techniques, plays a crucial role in enhancing the robustness and generalization of super-resolution models. The incorporation of synthetic data offers several advantages. Firstly, it provides a diverse range of training examples, covering various scenarios and image degradations that may not be adequately represented in real-world datasets. This diversity helps to improve the model's ability to handle different types of degradation and variations in image content. Additionally, synthetic data can be generated with ground-truth HR images, allowing for precise control over image degradation parameters and facilitating the learning of complex mappings from low-resolution to high-resolution spaces. Moreover, synthetic data can help address the limited availability of high-quality training data, especially for specific applications or domains where annotated data is scarce or expensive to acquire, e.g. medical imaging. By synthesizing additional training samples, researchers can augment existing datasets and mitigate issues related to data scarcity, leading to more robust and effective super-resolution models. Furthermore, incorporating synthetic data enables the exploration of novel regularization techniques and loss functions tailored to handle specific challenges in super-resolution, such as addressing artifacts or enforcing perceptual quality metrics. This flexibility in designing training objectives contributes to improving model performance and enhancing the visual quality of super-resolved images. However, the successful integration of synthetic data into super-resolution training pipelines requires careful consideration of data quality, realism, and relevance to the target domain. Balancing the use of synthetic and real-world data is crucial to ensure that models generalize well to unseen data and real-world applications [10].

The choice of an appropriate GAN-based super-resolution model is significant in conjunction to the incorporation of synthetic data techniques during training. The advancements in the field of GAN architectures for super-resolution demonstrate a variety of successfully approaches, with models in the ESRGAN family [10, 11] serving as ideal choices: the superiority of their visual results, together with their provided functional codebase [1] render ESRGAN and Real-ESRGAN/Real-ESRGAN++ as viable solutions of high-quality. To this end, connecting the ideas behind the benefits of synthetic data, tackling ill-posedness of generative reconstruction tasks, as well as the generation capabilities of ESRGAN models can lead to improvements in super-resolution tasks, addressing the ambiguity in generation, while preserving the visual quality of the HR image.

The two main strategies mentioned in the previous paragraph bring the main motivation and focus of the present work which is the exploration of the capabilities and limitations of three similar GAN-based architectures, specifically within the context of image super-resolution, to address challenges

1 https://github.com/xinntao/Real-ESRGAN

posed by the inherently ill-posed nature of the problem. By leveraging synthetic data and advanced ESRGAN-based models, the study aims to improve the mapping between low-resolution and high-resolution images while preserving visual quality and addressing the ambiguities of generative reconstruction. The scope of this work encompasses examining synthetic data's role in enhancing model performance, assessing transfer learning across related image enhancement tasks such as deblurring and low-light enhancement, and critically evaluating state-of-the-art methods to identify avenues for further advancements. Through this comprehensive approach, the thesis seeks to provide insights into the strengths and limitations of existing solutions while presenting targeted strategies for mitigating key challenges.

The following chapters are organized as follows:

- Chapter 2 reviews related work in image super-resolution, including network designs and the evolution of architectures, establishing a foundation for the study.

- Chapter 3 is where the methodology is introduced, including a discussion of background concepts, the specifics of the ESRGAN and Real-ESRGAN architectures, and the topic of integration of synthetic data to address ill-posedness.

- Chapter 4 covers the experiments and the results, showcasing the performance of the proposed approaches and offering an analysis of their effectiveness across various scenarios.

- Chapter 5 concludes the thesis with a summary of key findings, their implications, and potential directions for future research. Together, these chapters provide a structured and detailed exploration of image super-resolution challenges and solutions.

2. Related work

2.1. Image super-resolution

The field of image super-resolution has received an abundance of successful implementations, focusing on varying aspects of the problem throughout the general advancements of deep learning. As a first step, supervised learning techniques have been established, requiring a ground-truth LR-HR pairing.

There are several supervised super-resolution frameworks and as seen in [2], they can be grouped into the following notable categories:

1. ***Pre-upsampling super-resolution***: This method has its roots in the works of [18] and [41]. It utilizes traditional upsampling techniques to generate higher-resolution images, which are subsequently refined using deep neural networks. On the positive side, pre-upsampling super-resolution solves the harder upscaling operation on an early stage, and then allows Convolutional Neural Networks (CNNs) to perform further visual enhancements. However, there are drawbacks to consider. Most notably, this approach often incurs higher computational costs due to the majority of operations being performed in high-dimensional space after upsampling. Moreover, predefined upsampling algorithms may introduce side effects such as noise amplification and blurring, potentially affecting the quality of the final output images. Despite these challenges, the pre-upsampling super-resolution framework remains a viable approach in supervised image super-resolution, providing a direct and interpretable method for enhancing image resolution using deep learning techniques. The pre-upsampling super-resolution technique is depicted in Figure 2.1.

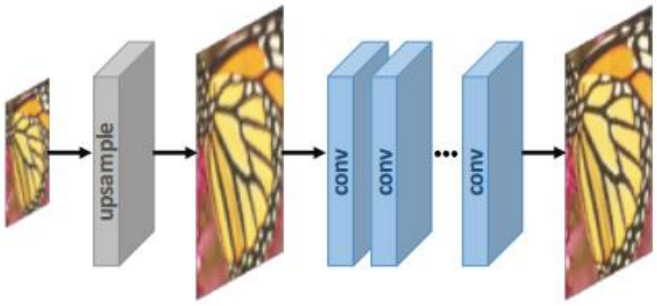

Figure 2.1. *Pre-upsampling super-resolution framework [2].*

2. ***Post-upsampling super-resolution***: In this technique, the majority of computations occur within a low-dimensional space, with end-to-end trainable upsampling layers added at the end to automatically enhance resolution. Pioneers of this approach [34,42] have made it obvious that it primarily aims to enhance computational efficiency, since operational costs are reduced when performed in lower dimensions, contrary to the pre-upsampling super-resolution technique. Overall, this technique has been favored for computational reasons and therefore adopted in later works. The post-upsampling super-resolution technique is depicted in Figure 2.2.

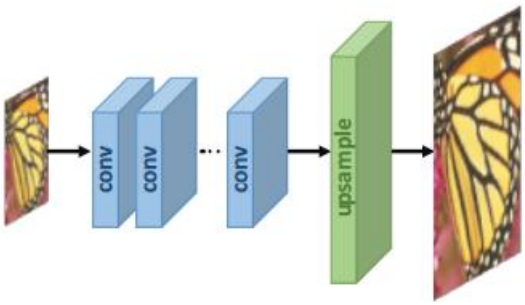

***Figure 2.2**. Post-upsampling super-resolution framework [2].*

3. ***Progressive Upsampling super-resolution***: this framework stipulates employing a cascade of CNNs to allow progressive reconstruction of HR images, aiming to break down the tough upscaling task into smaller steps to facilitate learning (particularly for large scaling factors), while keeping computational costs low enough. In works that have adopted this idea [43,44] one significant advantage is its prevalent ability to efficiently handle multi-scale super-resolution tasks, which is attributed to the progressive reconstruction. However, there are notable drawbacks to consider. Models within the progressive upsampling framework may necessitate training multiple stages or individual heads for intermediate-resolution image reconstruction, resulting in increased complexity in model training and design. Advanced training strategies need to be devised and incorporated to cope with training instabilities occurring, while the need for model guidance is also prevalent. Despite these obstacles, the progressive upsampling super-resolution framework demonstrates promise in addressing multi-scale super-resolution tasks and mitigating the learning difficulty associated with large scaling factors. The progressive upsampling super-resolution framework is illustrated in Figure 2.3.

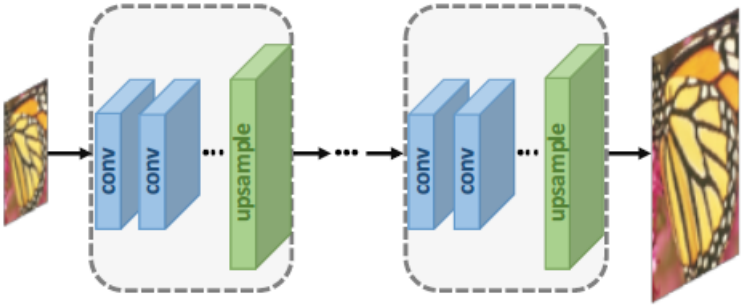

***Figure 2.3**. Progressive upsampling super-resolution framework. The dashed rectangles correspond to different intermediate upscaling stages [2].*

4. ***Iterative Up-and-down Sampling Super-resolution***: This framework aims to enhance the understanding of the mutual dependency between LR and HR image pairs by iteratively refining the estimation of the HR image. Key aspects of this framework include the utilization of back-projection refinement, where reconstruction errors are computed and fused back iteratively to adjust the intensity of the HR image, thus improving the quality of the HR image estimation by incorporating feedback from the reconstruction error. Additionally, by iteratively refining LR and HR image pairs, the framework efficiently models their mutual dependency, thereby enhancing overall super-resolution performance and generating more

accurate HR images. The proposals in [39] and [40] exhibit these characteristics. The iterative nature of the framework allows for multiple refinement steps, progressively improving HR image estimation and potentially leading to better image quality and detail preservation compared to single-stage super-resolution approaches. However, despite its effectiveness, the iterative up-and-down sampling super-resolution framework may encounter challenges related to model complexity, training stability, and the necessity for advanced modeling guidance and training strategies to optimize performance and convergence. Consequently, researchers continue to explore and refine iterative super-resolution approaches to elevate image quality and tackle the complexities associated with super-resolution tasks. Figure 2.4. illustrates the iterative up-and-down sampling super-resolution framework.

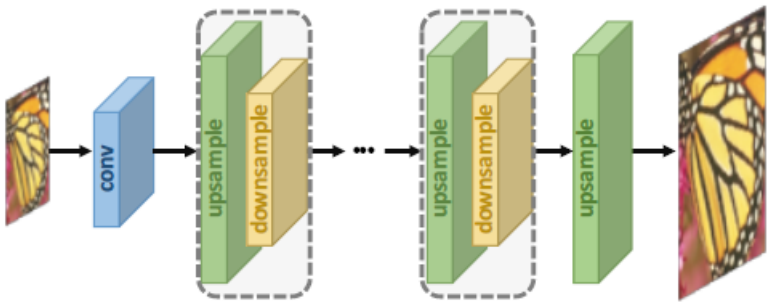

Figure 2.4. *Iterative Up-and-down Sampling Super-resolution framework [2].*

There are also techniques where supervision is not necessary and they come with certain advantages and some limitations. The list below is by no means exhaustive, but presents a few examples of the most common unsupervised techniques that can be found in the literature at present:

1. ***Generative Adversarial Networks (GANs)***: Generative Adversarial Networks (GANs) are a type of machine learning model that consist of two components: a generator, which creates high-resolution (HR) images from low-resolution (LR) inputs, and a discriminator, which distinguishes between real HR images and those generated by the model. There are of course many variations amongst different implementations, but the generator-discriminator pair architecture can be observed in all of them. One notable example of unsupervised variant is from the work of [35], called the CycleGAN (Figure 2.5). It employs a cycle-consistency loss to map LR images to HR images (and back) without the need for paired data. This enables the generator to produce HR images that are indistinguishable from real samples in the target domain, such as converting blurry images into high-quality outputs without requiring direct HR-LR pairs. GANs are particularly important for generating perceptually realistic HR images, often capturing fine details and textures with impressive fidelity. As already discussed in the introduction, GANs are also the area comprising the main focus and scope of the following chapters of the present work.

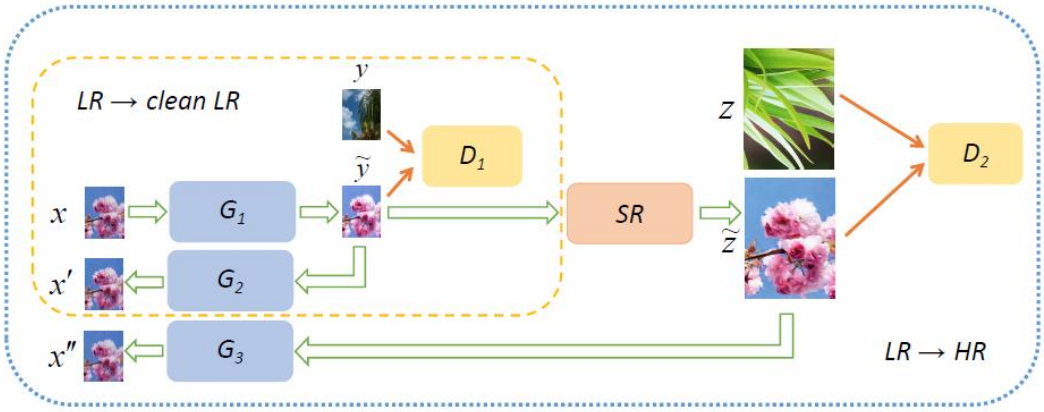

Figure 2.5. *Depiction of the framework of the proposed CinCGAN, the type of CycleGAN proposed in [35]. G1, G2 and G3 are generators, SR is a super-resolution network, D1 and D2 are discriminators.*

2. ***Self-Supervised Learning***: It is a method that generates pseudo-labels directly from the input data, enabling models to learn without requiring labeled datasets. In the context of super-resolution, this is typically achieved by artificially downscaling high-resolution (HR) images to create synthetic low-resolution (LR) versions, as shown in [34]. The model is then trained to reconstruct the original HR image from these synthetic LR inputs, allowing it to learn meaningful features during the process. Techniques such as data augmentation and contrastive loss can further improve performance. This approach is particularly significant as it leverages the natural hierarchy of image resolutions to develop powerful representations, eliminating the need for manually labeled data. An example can be seen in Figure 2.6.

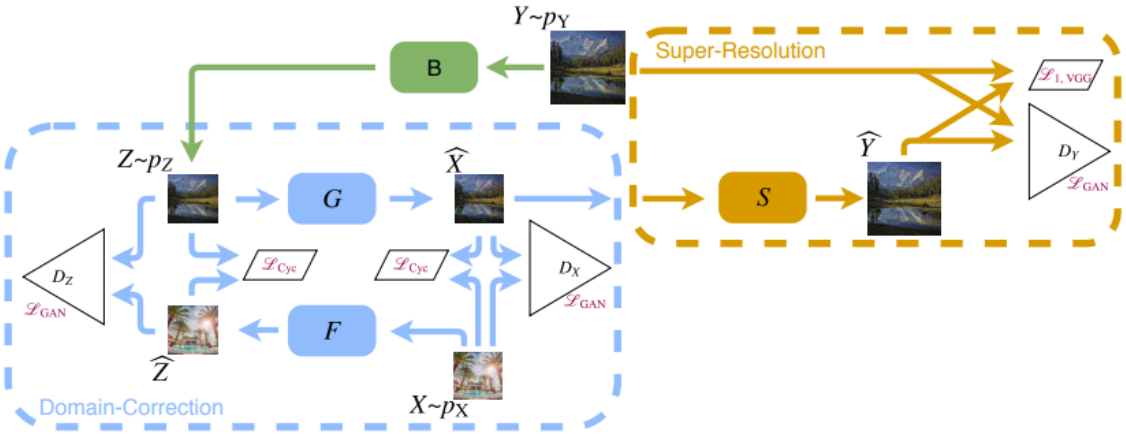

Figure 2.6. *Schematic overview the model in [34]. The model firstly learns the domain distribution network G, depicted in blue. Distributions of unpaired data from the input pX and output pY are used so the generator G is trained in a GAN framework by employing cycle consistency losses. The SR network S is trained in a second stage, depicted in orange, using pairs (X, Y ˆ) generated by their domain distribution, network G.*

3. ***Variational Autoencoders (VAEs)***: Variational Autoencoders (VAEs) are a type of model that learn a latent space representation of low-resolution (LR) images and use this representation to reconstruct high-resolution (HR) images by sampling from the latent space. It is demonstrated in [35] as an unsupervised approach, VAEs maximize a likelihood-based objective that combines reconstruction loss with a KL divergence term, eliminating the need for paired data. One of the key advantages of VAEs is their probabilistic framework, which enables the modeling of uncertainty in super-resolution tasks. This makes them particularly effective for generating multiple plausible HR outputs from the same LR input. Figure 2.7. shows a relevant example.

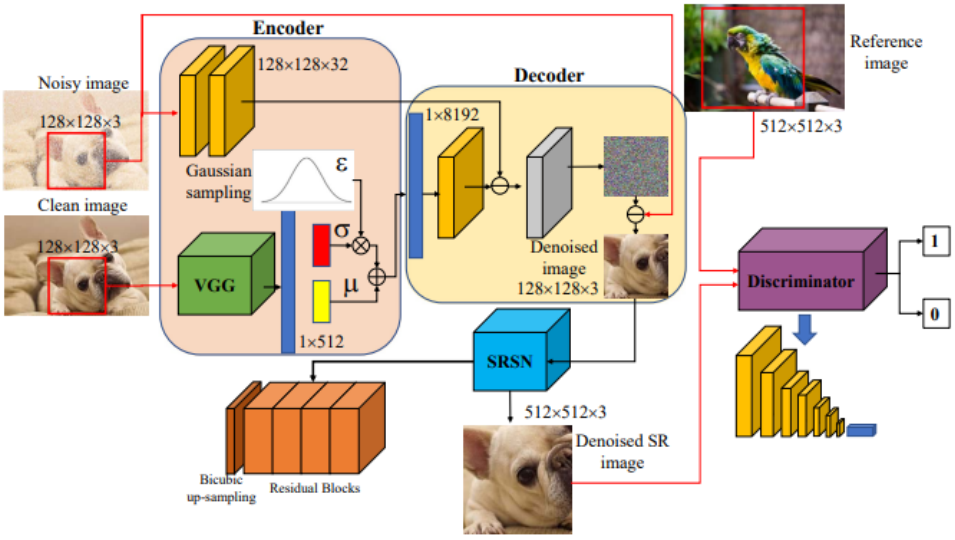

Figure 2.7. *Structural representation of the SRVAE model proposed in [35]. It includes a Denoising AutoEncoder (DAE) and Super-Resolution Sub-Network (SRSN).*

4. ***Patch-based Learning***: It is a technique that leverages local patches within an image or across multiple images to identify self-similar structures. The model learns correspondences between low-resolution (LR) and high-resolution (HR) patches to reconstruct detailed HR images. Two common techniques used in this approach are neighborhood embedding, which identifies similar patches and utilizes their HR counterparts for reconstruction, and non-local means, which aggregates information from similar patches to enhance upscaling. This method is particularly effective because it capitalizes on the inherent redundancy present in image data, allowing it to perform well even without explicit supervision [36]. The adaptive target ESRGAN method we discuss in this paper is also an example of patch-based learning utilization. In Figure 2.8. the proposed model from [36] is depicted.

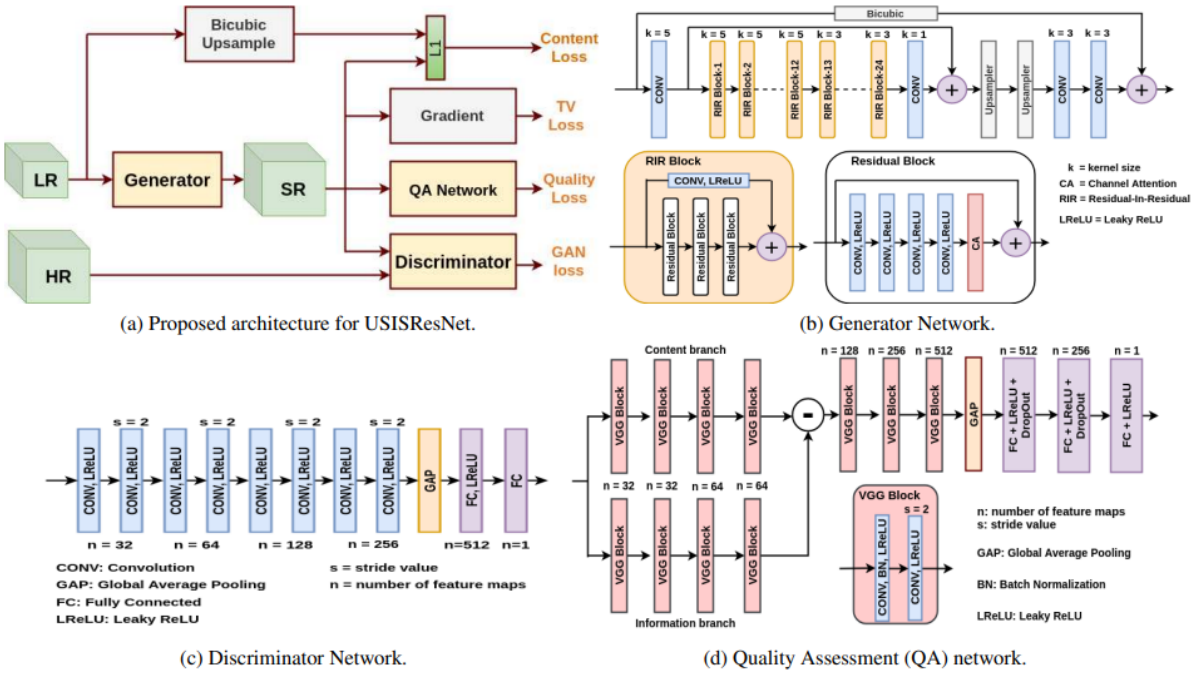

Figure 2.8. *Structural representation of the architecture of the Patch-based GAN model proposed in [36]. It includes an additional Quality Assessment (QA) Network..*

5. ***Perceptual Loss Functions***: Their focus is comparing high-level features of a generated high-resolution (HR) image to those of a real HR image, rather than relying on pixel-wise reconstruction loss. These features are typically extracted using pre-trained networks. In [37] we can observe that by minimizing perceptual differences, models can produce visually convincing HR images without requiring pixel-perfect HR-LR pairs. This approach is particularly significant because it aligns closely with human visual perception, allowing models to generate images that appear more realistic and visually appealing, even if they lack exact pixel accuracy. Figure 2.9. shows the perceptual loss function conceptualization by the authors of [37], as they are incorporated in their model's approach of awareness of the distance between two domains (The next unsupervised learning technique in paragraph 6).

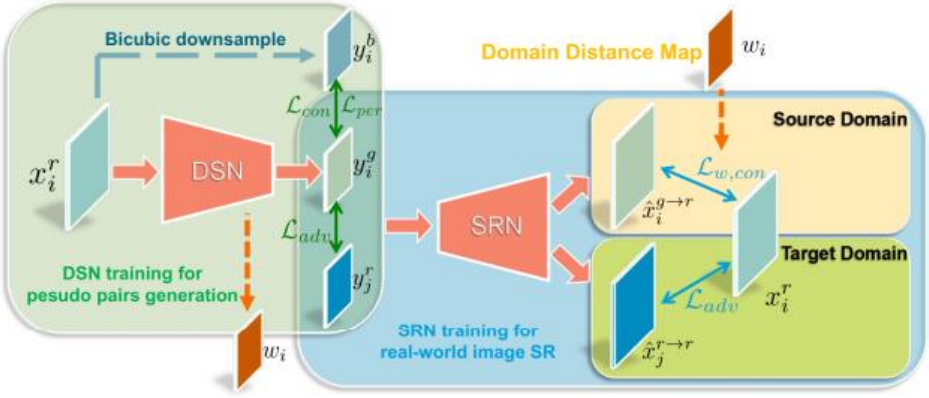

Figure 2.9. *Illustration of the proposed domain distance aware model of [37] , with the incorporation of perceptual loss functions.*

6. ***Cross-domain transfer learning***: It involves training a model on a source domain where paired high-resolution (HR) and low-resolution (LR) data are available, then adapting the model to a target domain using unpaired data through fine-tuning or domain adaptation techniques. In the target domain, unpaired images help bridge the gap between the source and target domains (seen in both [37] and [38]). For instance, a model trained on synthetic data can be adapted to work effectively in real-world scenarios. This approach is particularly valuable when labeled data is unavailable in the target domain but the two domains share similar underlying structures, enabling effective knowledge transfer. In Figure 2.10., the proposed cross-domain transfer learning from [38] is shown.

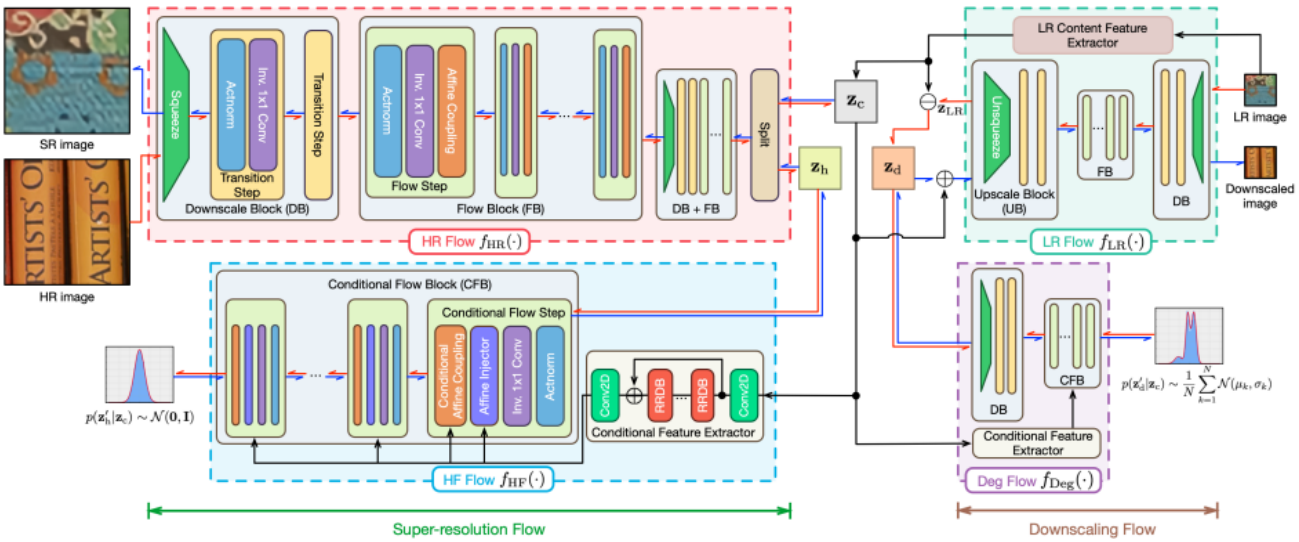

Figure 2.10. *Illustration of the cross-domain transfer learning model of [38] .*

2.2. Upsampling methods

In the scope of supervised super-resolution frameworks and in conjunction to their attributes, we discuss the various upsampling algorithms that have been employed in later models in this category. They refer to the image scaling operation from LR to HR images, and can be divided into interpolation-based and learning-based upsampling methods.

Interpolation-based algorithms incorporate the straightforward approach of nearest-neighbor interpolation, choosing the value of the closest pixel for each position during interpolation and yielding rapid but blocky outcomes. Bilinear and bicubic interpolation methods, on the other hand, utilize neighboring pixel values to calculate the interpolated pixel value, resulting in smoother outputs when compared to nearest-neighbor interpolation. Additionally, more sophisticated techniques such as Sinc and Lanczos resampling strive to maintain image details and minimize artifacts during upsampling, offering further enhancement in interpolation quality [2]. The interpolation-based upsampling logic is demonstrated in Figure 2.11.

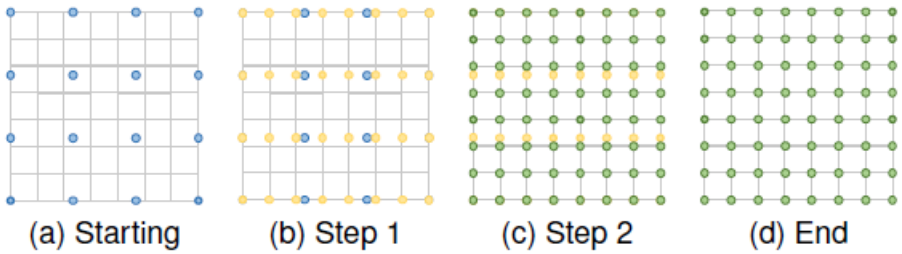

Figure 2.11. *As shown in [2] Commencing from a LR image (blue points in starting step), interpolation algorithms fill the gap with intermediate points (yellow dots in the first step), outputting a final HR image (green dots in the second step).*

Nevertheless, there are some drawbacks tied to the interpolation-based upsampling, including amplification of noise present in the original LR image and blurry outcomes, as well as increased computational complexity. To this end, learning-based upsampling algorithms are designed to overcome these limitations, by introducing neural layers that contain learnable parameters. Specifically, in the learning-based upsampling family, end-to-end learnable upsampling layers are incorporated within deep neural networks to acquire the upsampling process during training, facilitating a more adaptable and flexible approach to upsampling operations that cater to the unique requirements of the super-resolution task at hand. This integration allows for the seamless learning of complex upsampling patterns, leading to the generation of high-quality HR images. The utilization of Convolutional Neural Networks (CNNs) for end-to-end upsampling has emerged as a prevalent trend in image super-resolution, empowering networks to effectively learn intricate upsampling procedures and produce superior HR images. In recent literature, learning-based upsampling techniques are more favorable in accompanying super-resolution endeavors [2].

The ***transposed convolution layer***, often referred to as the deconvolution layer, constitutes a pivotal element within deep learning-driven upsampling techniques employed in image super-resolution. Firstly, the transposed convolution layer executes an upsampling operation by expanding the input

feature maps to a higher resolution. By predicting potential inputs based on feature maps akin to those generated by a convolutional layer, it effectively augments the spatial dimensions of the data, thereby enhancing resolution. Secondly, through the strategic insertion of zeros and the application of convolution operations, the transposed convolution layer magnifies the image size in an end-to-end fashion while preserving a connectivity pattern compatible with standard convolution operations. This iterative process contributes significantly to augmenting the resolution of the input data. Furthermore, within image super-resolution contexts, the transposed convolution layer commonly functions as an upsampling layer to produce higher-resolution feature maps from lower-resolution inputs. This feature allows the network to assimilate the upsampling process during training, enabling the model to reconstruct finer details within the high-resolution output image. However, it is crucial to acknowledge that transposed convolution layers may introduce artifacts, notably "uneven overlapping" and checkerboard-like patterns, into the output images. In response, researchers have endeavored to develop techniques aimed at mitigating these artifacts, thereby enhancing the overall performance of super-resolution models employing transposed convolution layers [2].
The transposed convolution layer is illustrated in Figure 2.12.

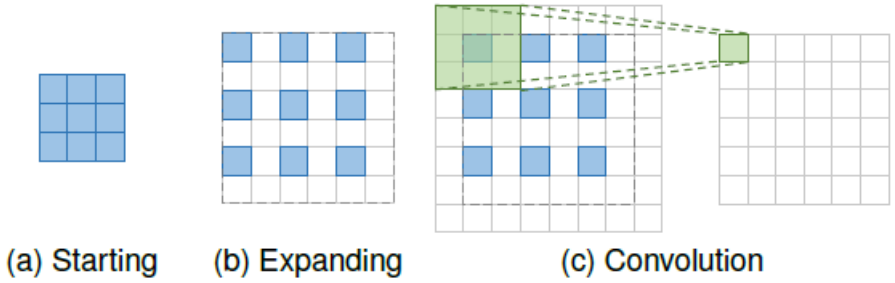

Figure 2.12. *Transposed convolution layer operation. The initial LR image (blue squares) is expanded and the gaps (white squares) are appropriately filled using convolutional operations, denoted in green. The convolutional kernel is projected to form an area of the final HR image, as denoted in the rightmost image [2].*

The Sub-pixel Layer is another approach within the deep learning-driven upsampling methodologies employed in image super-resolution undertakings. Primarily, the Sub-pixel Layer executes an upsampling operation by leveraging convolution operations to generate an increased number of channels. Subsequently, these channels are reshaped to amplify the spatial resolution of the data, thereby enhancing the overall resolution. Moreover, the initial convolution operation effectively augments the feature representation of the input data by generating outputs with a higher number of channels. This augmentation process significantly enriches the information available for generating higher-resolution outputs. Following the convolution operation, the reshaping operation, also known as shuffle, is performed to rearrange the channels and reshape the data into a format with heightened spatial resolution. This reshaping procedure plays a pivotal role in reconstructing the image at a finer scale, thereby enhancing its quality. Additionally, the Sub-pixel Layer offers a broader receptive field in comparison to alternative upsampling methods, affording more comprehensive contextual information. This augmented contextual understanding aids in generating realistic details and preserving crucial features during the upsampling process. Lastly, akin to other learnable upsampling layers, the Sub-pixel Layer enables end-to-end upsampling within deep neural networks. This feature empowers the model to learn the upsampling process autonomously, consequently facilitating the

generation of high-quality, high-resolution images during the training phase [32].
The sup-pixel layer operation is demonstrated in Figure 2.13.

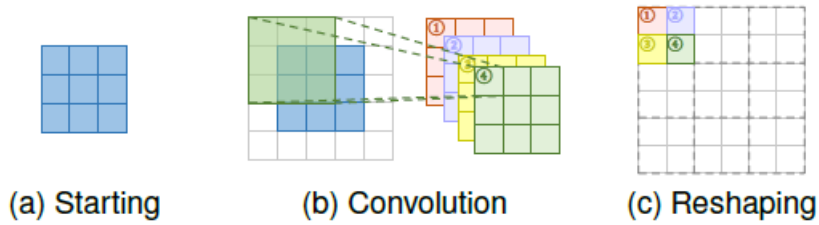

Figure 2.13*. Sub-pixel layer operation. Multiple channels are extracted in parallel during the convolution stage towards a HR counterpart, leveraging overlapping receptive fields. During the re-shaping stage, the channels are placed appropriately to constitute the final HR image [2].*

A limitation of the learning-based upscaling methods described above is the need for a predefined scaling factor which inhibits efficiency and applicability of related approaches in real-world scenarios. To this end, a meta-upscale module can dynamically accommodate arbitrary scaling factors, leveraging meta-learning techniques. This method is lightweight in practice, with the time taken for the upsampling module accounting for only a small percentage of the overall processing time, despite the need for predicting weights during the inference stage. As shown in [34], the performance of this module on varying upscaling factors remains competitive in comparison to fixed ones due to the large amounts of training data employed.

2.3. Network Design in image super-resolution methods

In terms of network design, an abundance of methods has been proposed throughout the years. Popular deep learning techniques leveraged in other computer vision tasks are adopted to enhance super-resolution approaches.

Residual and recursive learning are complementary techniques that enhance deep neural networks by improving feature refinement and hierarchical representation. Residual learning focuses on capturing differences between inputs and outputs through residual functions, addressing the degradation problem in deep networks by simplifying optimization and boosting training efficiency [49]. Shortcut connections enhance gradient flow and fine-tuning of input features, while global and local residual approaches target different scales for better performance in tasks like image super-resolution. Recursive learning, as highlighted in [45,46,47], iteratively processes modules to progressively learn hierarchical features, expand the receptive field, and capture broad contextual dependencies without increasing parameter counts. Despite challenges like gradient issues, recursive learning often benefits from residual strategies and multi-supervision, ensuring efficient and compact feature learning, as noted in [49].

Multi-path learning uses multiple pathways in a network to process input data differently and fuse extracted features for richer representations. In [35,48,49] we see how this enables the network to capture diverse global, local, and scale-specific features, improving modeling capabilities for complex tasks like image super-resolution. Parallel learning across paths also accelerates training and improves generalization.

Dense Connections are inspired by DenseNet. They link all layers in a block, promoting feature reuse, smooth gradient flow, and efficient information propagation. This structure reduces gradient vanishing, enables hierarchical representation learning, and compresses model size without sacrificing performance. Dense connections effectively integrate low- and high-level features, boosting super-resolution outcomes [50].

Attention mechanisms enhance a network's focus on relevant spatial areas and dependencies, an observation that can be made in [51]. Spatial attention emphasizes critical regions in the input, while channel attention dynamically reweights feature importance for better representation. These mechanisms improve contextual understanding, enabling precise reconstructions with intricate textures and structures in super-resolution tasks.

Dilated and group convolutions enhance super-resolution models by expanding receptive fields, improving feature reuse, and optimizing efficiency. Dilated convolutions capture broad contexts, while group convolutions reduce model complexity. These methods refine feature extraction, bolster modeling capabilities, and achieve high-quality image reconstruction with computational efficiency [52, 53, 54].

Region-recursive learning addresses pixel interdependence by modeling global context and serial dependencies during super-resolution. It generates coherent, high-resolution images by sequentially refining pixels, guided by long-range dependencies. Adaptive attention mechanisms enhance detail generation, but the approach increases computational costs and training complexity [55,56,57].

Pyramid pooling captures global and local context by dividing feature maps into bins, pooling information across spatial scales, and upsampling for resolution restoration. As demonstrated in [58] and [59], this process enriches feature representation by fusing multi-scale information, enabling holistic understanding and improved reconstruction quality in super-resolution models.

Wavelet transformation decomposes images into high-frequency details and low-frequency structures, enabling efficient multi-resolution analysis. Super-resolution models focus on enhancing high-frequency components for sharpness while ensuring coherence through low-frequency information. This approach improves computational efficiency, reduces model size, and minimizes artifacts [60].

Desubpixel accelerates inference by reducing feature space dimensionality, optimizing computational resources for faster processing. This method supports real-time super-resolution tasks but may require balancing speed and output quality to maintain visual fidelity [61].

xUnit integrates spatial feature processing with activation functions, enhancing representation learning by capturing complex spatial patterns. It combines spatial weighting and non-linear activations, improving efficiency and performance while reducing model size compared to traditional activation functions like ReLU [62].

2.3. Evolution of super-resolution architectures

Image super-resolution and other reconstruction tasks based on deep learning models received some of their first implementations around a decade ago. Based on approaches targeting image deblurring and denoising using neural networks [19], the seminal model of SRCNN set the basis for several consequent image super-resolution implementations stepping upon CNNs [18]. It offers an end-to-end solution, mitigating external operations (pre-processing or post-processing stages) outside the CNN pipeline. The advanced results of SRCNN over interpolation-based upsampling techniques or other non-neural approaches established the use of neural networks for image reconstruction tasks, including super-resolution, denoising and beblurring, while it presents efficiency in terms of computational speed compared to previous neural methods that had been proposed for other reconstruction tasks, such as denoising. Consequent implementations validated the successful usage of CNNs, contributing to frameworks and architectures, such as the ones analyzed in the previous section. To this end, a variety of implementations served the image restoration research community for quite some time.

Inspired by VGG-Nets, which demonstrated superior image classification performance, [20] achieves in surpassing SRCNN performance with a minor sacrifice in training time, leveraging very deep neural networks and residual connections. In a similar fashion, deeper networks are favored for performance, while a modification in residual connection structure is sufficient to boost image super-resolution results [21]. Moreover, altering the loss function can be beneficial. For example, stepping away from pixel-level losses and moving towards perceptual losses, which focus on high-level features or even combining pixel-level and perceptual losses, enhances reconstruction performance, while maintaining fast training time [22]. Targeting the reconstruction of fine details when higher reconstruction factors are imposed, SRGAN, based on GAN backbones and prior enhancements such as VGG-based architecture and perceptual loss functions achieves high-quality reconstructions, aiming one key aspect of image super-resolution ill-posedness [16]. Texture-related enhancements are targeted via feed forward CNNs that handle all RGB channels simultaneously and perceptual losses in conjunction to adversarial training, overcoming the over-smoothing resulting in previous works [23]. By focusing on intermediate layers of the neural network conditioned on semantic segmentation probability maps, which exemplify considering the semantic characteristics of the category of the image to be reconstructed, the resulting modifications lead to significantly more realistic HR textures in comparison to reconstructions that employ no image priors [24]. ESRGAN continues the GAN-based line of work, targeting advanced visual quality and realistic texture reconstruction through architectural enhancements, as well as adjustments on the loss functions (including adversarial loss and perceptual loss) [10].

Early approaches in image super-resolution are constructed using certain assumptions in order to be able to perform the reconstruction. Such assumptions are primarily centered around the degradation function that has been applied on the HR image to produce its LR counterpart. However, the

assumed bicubic degradations employed for downsampling do not correspond to most real cases, where more complex reconstruction mappings need to be learned in order to achieve successful HR images. The unknown nature of such degradations delineates the field of blind super-resolution, which either incorporates learning more complex mappings, involving an array of downsampling operations together (noise addition, compression, blurring etc.) or totally trespassing the need to explicitly assume the underlying degradation function and instead focus on learning the data distribution to implicitly infer the degradation function. To this end, Real ESRGAN employs a novel degradation scheme to achieve the optimal trade-off between efficiency and simplicity towards real world degradations, ultimately achieving a more complex degradation space. This fact necessitates some architectural alterations in Real-ESRGAN; specifically, the more capable U-Net architecture is employed for the discriminator in place of the VGG-style network used in the prior version of ESRGAN. This enhancement allows for more fine-grained understanding of details to enforce accurate discrimination between real and fake samples. Nevertheless, the more expressive U-Net architecture together with the complex degradation set impose several training instabilities, which can be however resolved when Spectral Normalization (SN) regularization is incorporated during training. Finally, another basic advancement brought by Real ESRGAN is the employment of purely synthetic data samples for training, which are able to drive real world restorations and are strongly connected with resolving the ill-posed nature of reconstructive tasks [11].

3. Methods

3.1. Background

A central idea employed in this work is tackling super-resolution by using synthetic data alongside the advanced ESRGAN architecture family, as introduced in [10, 11]. Synthetic data offers benefits like enabling diverse training sets that cover a wide range of image degradations not always present in real-world datasets, enhancing the model's ability to handle distortions. It also provides precise control over degradation parameters using paired high-resolution ground truth, facilitating the learning of complex mappings. By augmenting existing datasets, synthetic data addresses the scarcity of annotated samples, particularly in domains like medical imaging, improving model robustness. Additionally, it supports experimenting with novel loss functions and regularization techniques for better visual fidelity. However, effective integration requires balancing synthetic and real-world data to ensure generalization ability [10].

3.1.1. Synthetic data for super-resolution ill-posedness

There have been several different approaches for incorporating synthetic data in the training process of models dedicated to image reconstruction tasks. Two of the most notable ideas are the adaptive target generation[17] and data augmentation techniques [11] to construct synthetic degraded images. These two processes are based on different ideas and have independently benefited related models.

Adaptive Target Generation (ATG)

The idea behind the adaptive target technique[17] was a catalyst in generative reconstruction tasks, as it is proven to be a highly performing solution to the inherent ill-posedness of related problems. More specifically, adaptive target localization training offers a multifaceted set of advantages in the domain of super-resolution, especially in cases where most previous traditional models do not handle

the one-to-many mapping from LR images to HR ones as such, as well as struggling with the presence of unknown blur kernels. This innovation bears several key implications for mitigating the ill-posedness of super-resolution. Firstly, adaptive target patches introduce flexibility in output evaluation by allowing the consideration of multiple potential HR targets for a given LR input. This departure from the traditional one-to-one mapping constraint fosters exploration of a broader solution space, facilitating the generation of visually appealing outputs alongside mathematical validity. Secondly, adaptive targets play a crucial role in handling the inherent ambiguity of the super-resolution problem, arising from its one-to-many nature. By providing the algorithm with alternative targets derived from the original ground truth (GT) target through transformations, adaptive targets guide the algorithm to generate outputs capturing the essence of the GT target while accommodating variations aligned with the ambiguity intrinsic to the super-resolution task. Thirdly, adaptive targets serve as an effective training strategy, particularly in blind super-resolution scenarios featuring unknown blur kernels. Training the algorithm to minimize the difference between the generated output and the adaptive target enhances the algorithm's robustness to variations in blur kernels and input conditions, ultimately improving its capacity to generate high-quality HR images. Lastly, the utilization of adaptive targets significantly enhances the perceptual quality of generated HR images. By encouraging the algorithm to explore diverse solutions and align outputs with multiple valid targets, adaptive targets contribute to the production of visually pleasing results that retain the original content and perceptual impression of the images [17].

The procedure for generating an adaptive target within the super-resolution framework involves creating alternative targets derived from the original GT target, guiding the algorithm towards the production of high-quality synthetic HR images. The step-by-step process entails starting with the original GT HR image as a reference, applying diverse transformations such as scaling, rotation, and translation to introduce variations aligning with potential super-resolution outputs. The original GT HR image is divided into non-overlapping patches, and transformations are individually applied to each piece, thus achieving a more accurate transformation overall in comparison to applying the transformation on the whole image at once. A range of acceptable transforms is defined, maintaining content integrity while accommodating valid variations. The adaptive target is reconstructed by assembling the transformed pieces, capturing the essence of the GT target with variations. An example of a transformation leading to a desired adaptive target is provided in Figure 3.1, where f is the super-resolution model, x^i the LR image and y^i is the original target HR image.

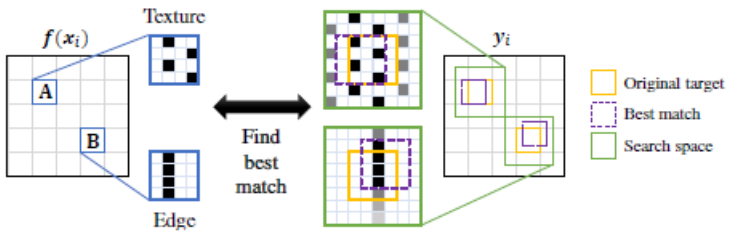

***Figure 3.1**. A sample demonstration of creating an adaptive target, in this case considering only translational transformation. Given pieces of the image f(x_i) denoted with blue boxes, the best match is obtained (purple boxes) within a contained search space denoted in green. The best matches are assembled in the final target y_i [17].*

The ideal adaptive target is the one being closer to the initial one, as occurring from the super-resolution model f. However, the approximation of the ideal target is a mathematically hard problem due to the large search space of possible acceptable solutions.

During training, the super-resolution algorithm minimizes the difference between the generated output and the adaptive target, optimizing for both mathematical validity and visual appeal. This comprehensive approach effectively addresses the ill-posed nature of the problem, handles ambiguity, and produces high-quality HR outputs that retain the original content and perceptual impression of the images. In the original paper, a two – stage model training strategy is explored. The first stage involves pre-training the super-resolution model f using the original target HR image from a selected dataset. This pre-trained model goes through further training incorporating the –ideal-adaptive target constructed via the aforementioned process.
An outline of the ATG process is depicted in Figure 3.2.

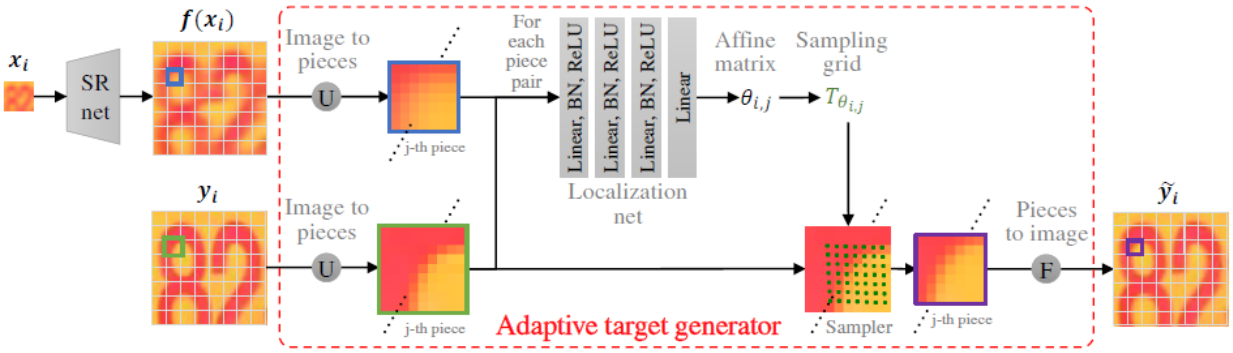

***Figure 3.2**. An outline of the adaptive target generation process [17].*

As seen in Figure 3.2 above, the ATG process involves several architectural components. These key elements include a super-resolution network, a localization network, and the application of affine transformations to produce alternative targets. The components are outlined as follows: [17]

At its core, the structure incorporates a super-resolution network, responsible for upscaling LR images to their HR counterparts. The output of this network, denoted as f(x_i), represents the generated HR image corresponding to a given LR input x_i. Complementing the super-resolution Network is a Localization Network, a critical element that estimates affine transformation matrices for transforming individual pieces of the original Ground Truth (GT) target, denoted as y_i, to align with the corresponding pieces of the generated HR image f(x_i). This network comprises 4 fully connected layers; excluding the first layer, the rest are followed by batch normalization and Rectified Linear Units (ReLU) as activation functions. Its functionality is crucial in order to determine the necessary transformations to achieve alignment. The process of Adaptive Target Generation is facilitated by the application of Affine Transformations, where matrices θ_{ij} are estimated to transform each piece of the original GT target y_i, incorporating operations such as translation and rotation. This introduces variations in the adaptive target generation process, allowing for diverse solutions. The adaptive target localization network technique in a sense augments the available training data due to the nature of the image transformation it utilizes to tackle the issue of ill-

posedness. This augmentation of the data can be considered as synthesizing extra data for the training, as a side effect of calculating the extra adaptive patches during the process [17].

The methodology adopts a Piecewise Processing approach, operating in a patch-wise manner by dividing the super-resolution network's output f(x_i) and the original target y_i into non-overlapping pieces of specific sizes. Each piece of the original target undergoes transformation using the estimated affine matrices to align with the corresponding HR image piece. To facilitate the transformation process, a Sampling Grid is generated from the estimated affine transformation matrices θ_{ij}. Bilinear Sampling is subsequently applied, ensuring a smooth and accurate mapping between the original target and the generated HR image. The Adaptive Target is then formed through the Combination of Transformed Pieces. Once all pieces of the original target have undergone transformation, they are unified to generate the adaptive target patch $\tilde{y}_i$. This adaptive target represents an alternative HR image, relaxing the strict one-to-one mapping constraint and allowing for variations in the output of the super-resolution algorithm [17].

The authors experiment on the influence of ATG as a synthetic data generation scheme over multi-stage training strategies, which aim at refining the performance of Super-resolution networks. The first step involves pretraining the Localization Network to ensure precise estimation of affine transformation matrices. This is achieved by pretraining the localization network using synthetic affine matrices, which are formed through random combinations of basic transformations, such as translation and rotation within defined ranges. The ATG incorporates the initial image and the randomly distorted image as occurring from the transformations, estimating the transformation matrices. Based on the matrices and the distorted images, the loss function for the ATG pre-training is formed. The pre-training stage provides direct supervision, enhancing the network's ability to accurately estimate transformations [17].

The next step involves training the already pre-trained super-resolution network with adaptive target patches generated through the ATG process on-the-fly, meaning that a new target patch is adapted to any consequent output image. The training loss is computed based on the disparity between the generated adaptive target and the super-resolution network output for each input LR image Operating in a patch-wise manner, the process divides images into non-overlapping pieces for transformation. This ensures local accuracy and spatial consistency, particularly in scenarios with spatially varying blur kernels. Training the super-resolution network with adaptive targets enables the super-resolution model to produce sharper and visually pleasing HR images, leading to improvements in performance metrics like PSNR and visual quality [17].
The impact of the two-stage training process is demonstrated in Figure 3.3.

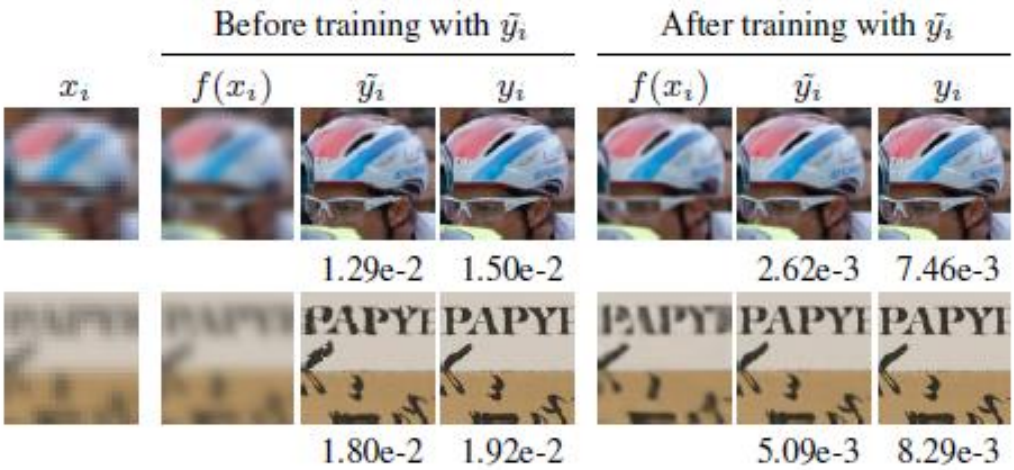

Figure 3.3. *Adaptive targets $\tilde{y}_i$based on the original y_i targets incorporated in a pre – trained super-resolution network, as well as in the model proposed in the paper. Corresponding outputs f(x_i) demonstrate the capabilities of the ATG process proposed in the paper. Mean squared error (MSE) is also reported as a measure of proximity between LR image and targets [17].*

The final results obtained from the adoption of ATG as proposed in [17] demonstrate the superiority of this approach. Such results as reported by the authors are provided in Figure 3.4, where a varying amount of noise from an isotropic Gaussian blur kernel is applied on the GT image (σ defines the amount of noise).

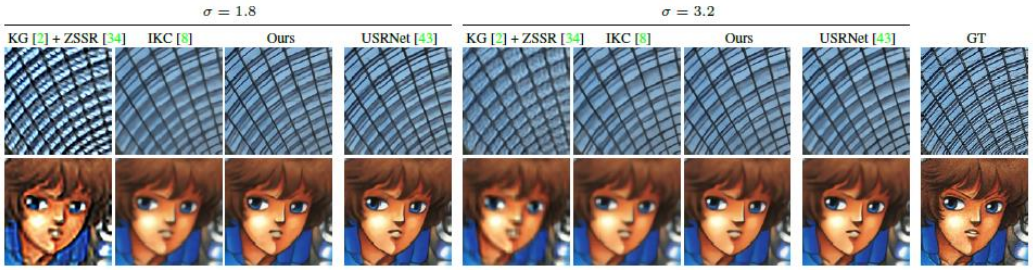

Figure 3.4. *Qualitative results of the adaptive targets vs other super-resolution implementations when isotropic Gaussian blur kernels are used to produce LR images. It becomes evident that adaptive kernels pose the ability to better reconstruct edges and texture, while preserving semantics of the GT image [17].*

Similarly, more qualitative results are presented in Figure 3.5, covering the capabilities of adaptive targets on handling random blur kernels (instead of isotropic ones, as in the previous case).

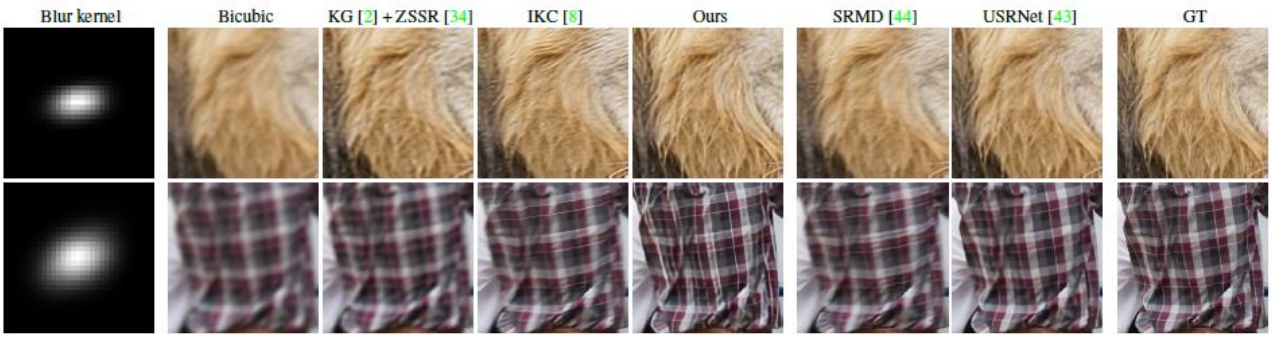

Figure 3.5. *Qualitative results of adaptive kernels vs other approaches on the same task when random blur kernels are used for LR images. Adaptive kernels better achieve texture reconstruction compared to the GT image on the right [17].*

By considering ATG as a synthetic data generation process, super-resolution reconstructions were enhanced, highlighting the merits of employing appropriate synthetic samples during training.

Data augmentation

While Adaptive Target Generation focuses on dynamically generating high-resolution target images during training to match the characteristics of input low-resolution images, allowing for adaptive learning and fine-grained optimization, the creation of synthetic data via appropriate degradations involves simulating real world image degradations on high-resolution images to generate paired low-resolution counterparts, enabling the training of super-resolution models under controlled conditions that mimic real-world challenges.

To this end, two key concepts promoted in the Real ESRGAN work [11] is the construction of more real-world degradation functions and the use of synthetic data. The synergy of these advancements concludes towards a novel data augmentation pipeline, which significantly enhances the quality of reconstructed images.

The degradation procedure in super-resolution offers numerous significant advantages for efficiently restoring and enhancing low-resolution images. Firstly, it allows for the accurate modeling of real-world distortions, including noise, blur, compression artifacts, and other common imperfections affecting images. Through this degradation simulation during training, the super-resolution model becomes adept at countering and eliminating these distortions in resulting high-resolution images. Additionally, the integration of a realistic degradation process into the training data enhances the model's ability to generalize to unseen or out-of-distribution images with similar degradation characteristics. This improvement contributes to the model's enhanced performance across a diverse range of real-world images, enabling it to handle various intricate degradation patterns. Furthermore, training the model with a degradation process augments its robustness against common imperfections in low-resolution images, such as noise and artifacts. The model learns to adapt to and mitigate these degradations, resulting in more precise and visually appealing high-resolution reconstructions. The degradation process also facilitates the synthesis of realistic training data that closely mirrors the characteristics of genuine low-resolution images. This authenticity in training data enables the model to learn the generation of high-quality reconstructions that faithfully represent the original content

while effectively addressing introduced degradations. Lastly, incorporating a degradation process into the training pipeline allows for performance evaluation under realistic conditions, emulating challenges found in real-world images. This comprehensive approach enables a more accurate assessment of the model's proficiency in handling diverse degradation types and producing high-quality super-resolved images.

The utilization of exclusive synthetic data generation in this research provides numerous advantages, significantly enhancing the effectiveness of the super-resolution model, especially in addressing real-world complexities. Firstly, the use of synthetic data establishes a controlled environment, allowing precise specifications of the characteristics of both input images and their corresponding high-resolution targets. This controlled setting facilitates targeted model training, concentrating on specific degradation types and image features for more efficient learning. Additionally, synthetic data generation introduces a diverse array of degradation types, noise levels, and artifacts into the training dataset, creating a broad spectrum of scenarios. This diversity is crucial for equipping the model to handle a wide range of real-world challenges and complex degradations that may not be readily available in actual data. Moreover, synthetic data generation proves to be more efficient and cost-effective than collecting and annotating extensive real-world datasets, streamlining the creation of comprehensive training datasets with diverse characteristics. Training on synthetic data enhances the model's generalization to unseen real-world images by exposing it to a broad spectrum of simulated degradations and challenges. This exposure contributes to refining the model's capacity to handle unknown and out-of-distribution degradations in practical applications. Finally, the use of synthetic data allows for controlled experiments and precise performance evaluation, leveraging the knowledge of ground truth high-resolution images. This accurate assessment aids in identifying areas for improvement and effectively refining the model's capabilities. An overview of the synthetic generation process involved in Real-ESRGAN is presented in Figure 3.6.

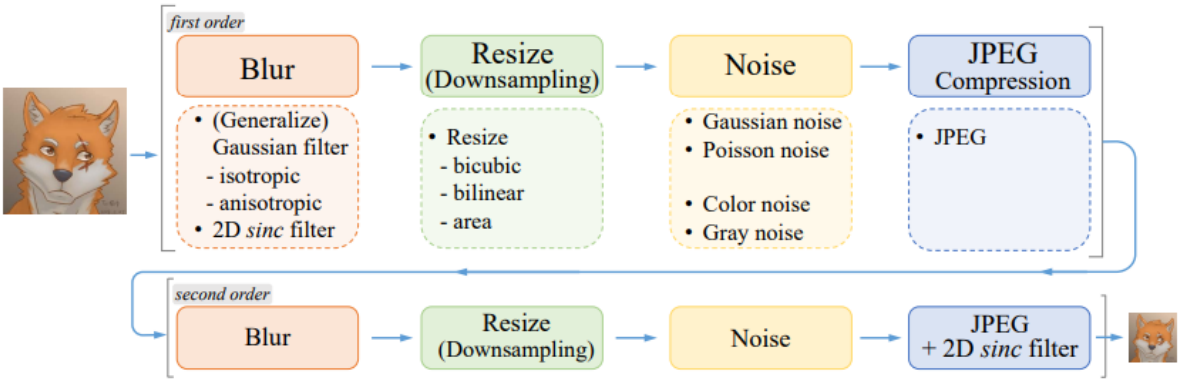

Figure 3.6. *The degradation process followed in Real-ESGAN to produce synthetic data for training. It is designed to incorporate common artifacts in images, with noise, blurring, compression and resize functions being among the most prevalent corruptions in practice [11].*

A sharpening technique employed during the training process alleviates the possibility of introducing artifacts, which may occur if sharpening is performed in a post-hoc manner.

3.1.2. Original ESRGAN architecture

ESRGAN (Enhanced Super-resolution Generative Adversarial Networks) distinguishes itself from previous state-of-the-art approaches at the time, such as SRGAN [16], through several key components. Initially, the Residual-in-Residual Dense Block (RRDB) is introduced as the fundamental building unit, excluding batch normalization, thereby enhancing the network's capacity to capture intricate features and improve image reconstruction (Figure 3.7). ESRGAN incorporates a relativistic GAN discriminator (Figure 3.7), predicting relative realness instead of absolute values, which aids in generating images with more realistic textures and reduced artifacts. The perceptual loss in ESRGAN is improved by utilizing features before activation, providing stronger supervision for brightness consistency and texture recovery. These modifications result in images with superior visual quality, featuring more realistic and natural textures. Additionally, ESRGAN refines its network architecture, adversarial loss, and perceptual loss components, contributing to the generation of images with detailed structures and fewer artifacts compared to SRGAN [16]. An outline of the architecture of ESRGAN is provided in Figure 3.7.

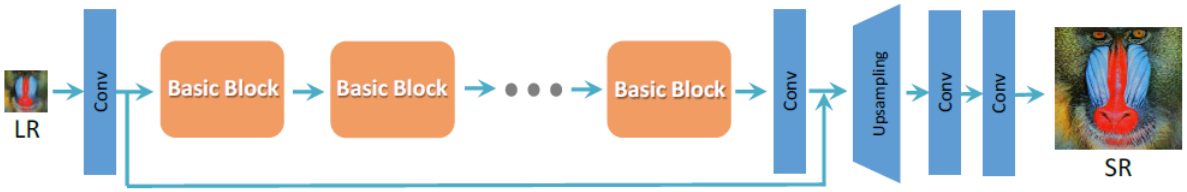

Figure 3.7. *ESRGAN architecture. The 'basic block' refers to any relevant structure, such as dense layers or residual connections block, even though authors propose the novel RRDB block [10].*

By zooming into the RRDB block of ESRGAN, as shown in Figure 3.8, we can easily observe that it comprises a combination of dense blocks and residual connections within. Deeper and more complex networks are proven to boost super-resolution results; thus, they are employed in the original ESRGAN architecture [10].

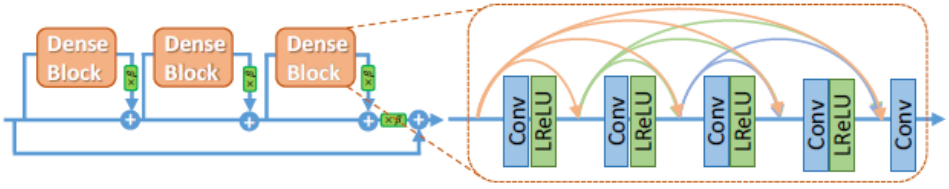

Figure 3.8*. RRDB structure as the basic block of ESRGAN [10].*

The relativistic discriminator, predicting the probability of a real image being more realistic than a fake one (instead of independently predicting whether an image is fake or real without a comparative measure), is showcased in Figure 3.9.

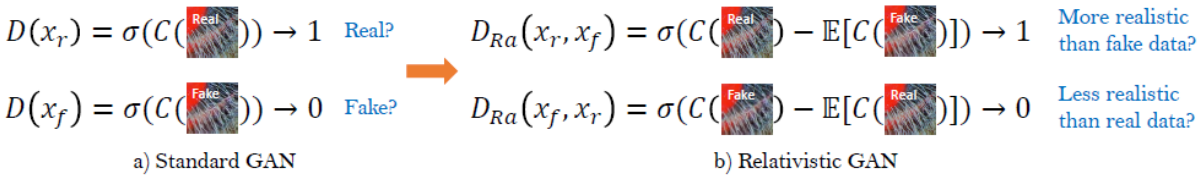

Figure 3.9. *Relativistic discriminator of ERSGAN, considering both real and fake instances at a time in order to assign a relevant label [10].*

The qualitative advancements of ESRGAN over SRGAN, as reported by the authors, are demonstrated in Figure 3.10.

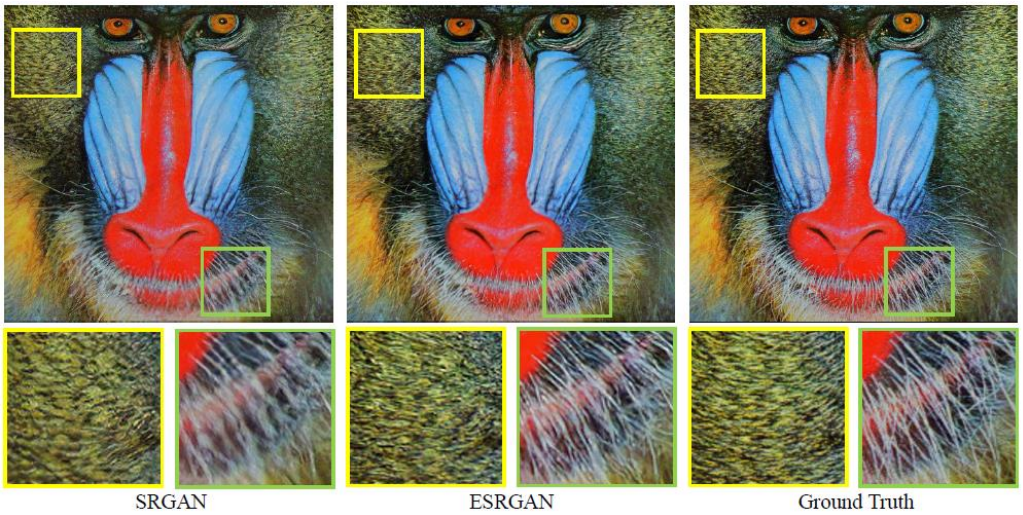

Figure 3.10. *Qualitative results of ESRGAN (x4 scale factor) in comparison to SRGAN and the ground-truth. There is an obvious ESRGAN improvement in terms of texture detail [10].*

The proposed enhancements of ESRGAN and its consequent variants analyzed later indirectly contribute to improving the overall quality and realism of super-resolved images, and thus the ill-posed nature of generative reconstructive tasks. The encouraging results of the original ESRGAN model led to the proposal of enhanced techniques, further advancing super-resolution and related tasks.

3.1.3. Real-ESRGAN architecture

Proceeding with the improved Real-ESRGAN variant [11], some enhancements further boost the quality of the recovered HR image. The main points proposed in [11] in comparison to prior art in image super-resolution are the following:

1. *Advanced Degradation Modeling*: The classical "first-order" degradation model is extended to a more sophisticated "high-order" degradation modeling approach. This extension is designed to capture the intricate complexities of real-world image degradations, providing a more accurate representation for the training process, ensuring that the model is well-equipped to handle the challenges prevalent in practical scenarios.
2. *Architectural Modifications*: Based on its predecessors [10], Real-ESRGAN focuses on improving discrimination capabilities of the underlying GAN structure. To this end, the

discriminator incorporates a U-Net architecture with spectral normalization, accompanied with a two-stage training process.

3. *Synthetic Data Generation Utilization*: The work utilizes pure synthetic data generation to create diverse training datasets with controlled degradations, noise levels, and artifacts, as described in the previous subsection. This approach enhances the model's adaptability to real-world scenarios and improves generalization by exposing it to a wide range of potential challenges.

Delving into the main contributions of [11], the employment of advanced degradation algorithms, as well as the utilization of synthetic data were analyzed in the previous section, serving as key components of improved super-resolution frameworks. We further analyze the architectural advancements introduced in Real-ESRGAN in the current section.

More specifically, Real-ESRGAN comprises of an ESRGAN [10] generator, a U-Net discriminator with Spectral Normalization and an enhanced training process integrating high-order degradation modeling, adaptive target mechanisms, and synthetic data generation. The ESRGAN generator employs an advanced deep network architecture featuring multiple Residual-in-Residual Dense Blocks (RRDB). Its primary function is to conduct super-resolution by improving the details and overall quality of low-resolution images. The architecture consists of numerous layers, allowing the model to comprehend intricate features and mappings necessary for high-quality image reconstruction. Moreover, the generator is specifically optimized to handle diverse scaling factors, such as ×2 and ×4, for various super-resolution tasks. The architecture of the Real-ESRGAN generator is depicted in Figure 3.11.

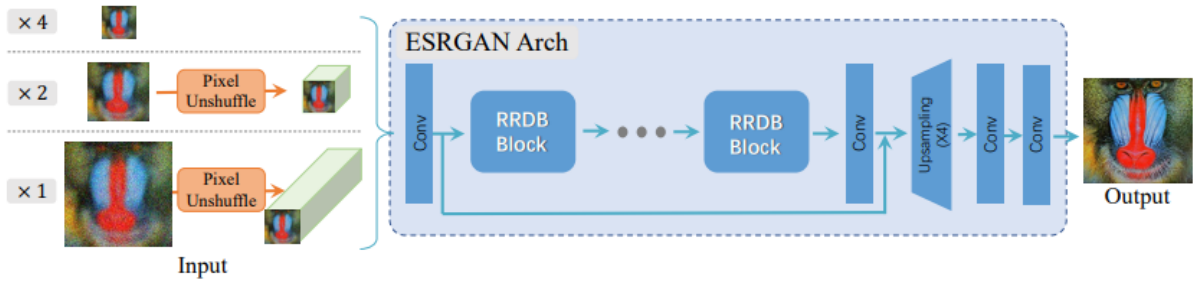

Figure 3.11. *Outline of Real-ESRGAN architecture, comprising the same generator module as in the original ESRGAN approach. The spatial size reduction is achieved by employing the pixel unshuffle operations for low resolutions (×1, ×2), contributing to the mitigation of resources consumption [11].*

The U-Net discriminator, an enhanced version of the VGG-style discriminator in ESRGAN, incorporates skip connections to boost overall performance. This discriminator provides realness values for each pixel, offering detailed per-pixel feedback to the generator. To ensure stable training dynamics and enhance the discriminator's ability to provide precise gradient feedback for local textures, spectral normalization regularization is applied. The U-Net structure is meticulously designed to manage complex training outputs and improve the model's capacity to differentiate between real and generated images. Figure 3.12 depicts the structure of the U-Net discriminator.

***Figure 3.12**. U-Net discriminator of Real-ESRGAN with spectral normalization for training dynamics stabilization. This component is crucial since the complicated degradation space imposes instabilities during training [11].*

The training process entails generating training pairs through a high-order degradation modeling process and utilizing exclusively synthetic data to train the Real-ESRGAN model. During training, Sinc filters are employed to model common ringing and overshoot artifacts, enhancing the model's proficiency in removing these artifacts from restored images. The primary focus of the training process is to enhance image details while eliminating disruptive artifacts in real-world images, with the ultimate goal of achieving superior visual performance compared to prior approaches. Additionally, the implementation includes efficient procedures for synthesizing training pairs on-the-fly, optimizing the training process and facilitating the generation of diverse training data with controlled degradations. Finally, training on sharpened images leads to an improved version named Real-ESRGAN+ [11].

Some outputs of the Real-ESRGAN and Real-ESRGAN+ models are presented in Figures 3.13 and 3.14. These qualitative results demonstrate superior capability in reconstructive tasks in comparison to previous state-of-the-art approaches, namely DAN [12], CDC [13], RealSR [14] and BSRGAN [15], as well as its predecessor ESRGAN [10]. The quality of the reported results sets the basis for our current work, which significantly relies on the Real-ESRGAN approach.

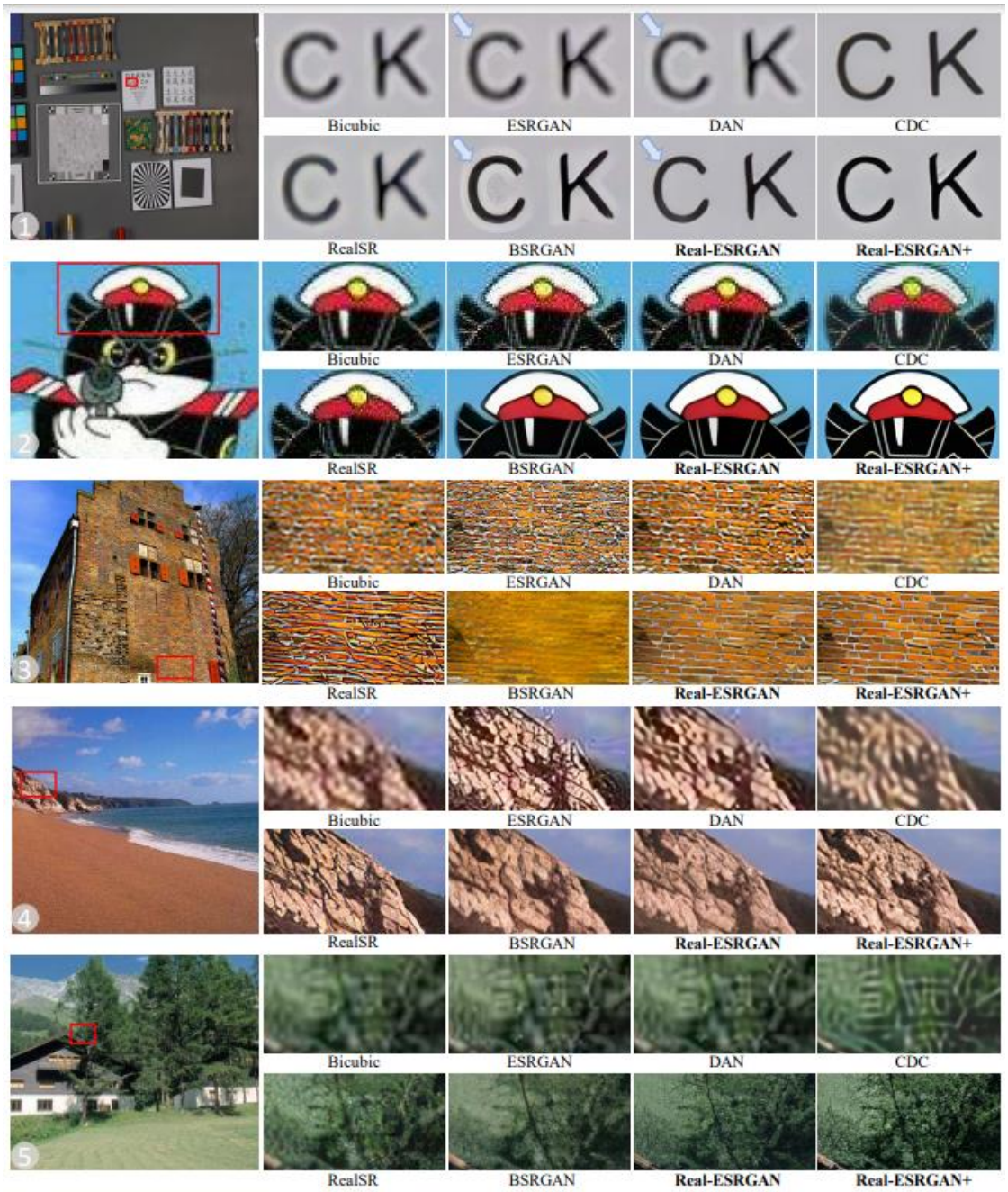

Figure 3.13. *Real-ESRGAN qualitative results as reported by the authors (scale factor of x4) in comparison to previous arts in super-resolution tasks. These outcomes verify the reconstructive power of Real-ESRGAN in terms of accuracy and quality in the details [11].*

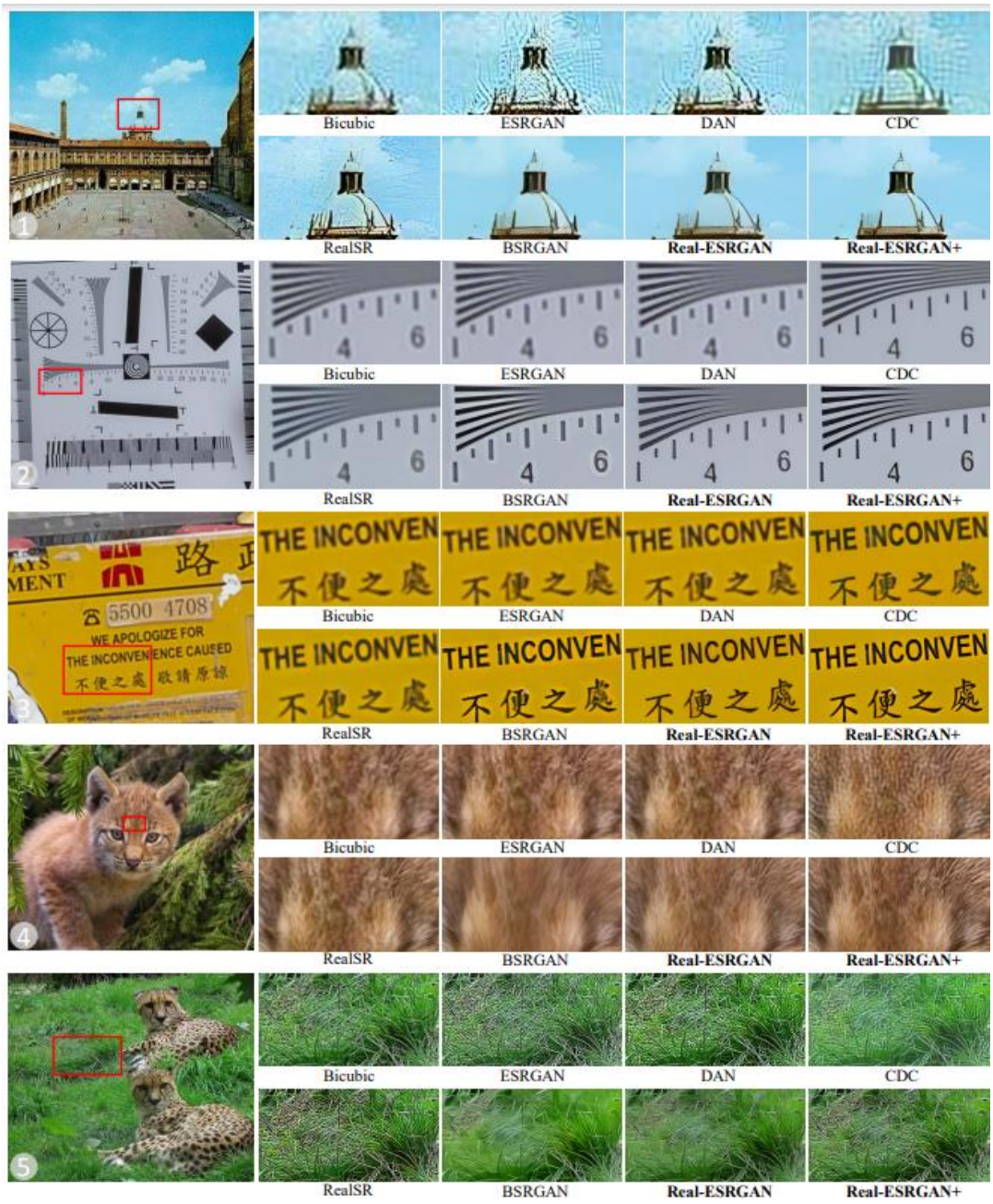

***Figure 3.14**. Continuation of previous Figure with more qualitative results from Real-ESGAN/Real-ESRGAN+ [11].*

A Real-ESRGAN variant called Real-ESRNet is a model focused on maximizing PSNR (Peak Signal-to-Noise Ratio), trained using the L1 loss function. Subsequently, the Real-ESRNet model serves as an initialization for the generator in the Real-ESRGAN framework. The Real-ESRGAN model is then trained using a combination of loss functions, including L1 loss, perceptual loss, and GAN (Generative Adversarial Network) loss. This sequential training approach leverages the strengths of Real-ESRNet as a PSNR-oriented model to initialize Real-ESRGAN, allowing for comprehensive optimization through multiple loss components to achieve enhanced super-resolution performance. Real-ESRNet is adapted from ESRGAN to facilitate quicker convergence during training.

In Figure 3.15 we present some results using Real-ESRNet.

Figure 3.15. *Qualitative results of Real-ESRNet for different degradation functions (top) and usage/no usage of sync filters (bottom).*

3.2. Methodology

In this section, we analyze our contributions in the field of super-resolution. Our work concentrates on three different image enhancement tasks, namely super resolution, image deblurring and low-light enhancement, employing the advancements of the ESRGAN architecture family.

For the task of super-resolution, the comparison is performed between the pretrained models of the Adaptive Target of [17] which uses the ESRGAN as its backbone, and the Real-ESRGAN [11]. This comparison is meaningful in terms of assessing different data augmentation processes, whether direct or indirect, i.e. the Adaptive Target Generation process [17] versus the usage of pure synthetic data during training based on the utilization of a novel degradation model, as in Real-ESRGAN [11].
Two Real-ESRGAN variants were tested, specifically the default Real-ESRGAN as described in [11] and the more optimized for convergence Real-ESRNet. To streamline the comparison of the models and in order to evaluate them at an equal basis, the scaling factor for all those experiments was selected to be x4. All models were tested in blind super resolution with random test kernels, using the specialized DIV2KRK dataset [25]. The Adaptive Target model was pretrained on the DIV2K dataset [26] and the Real-ESRGAN was pretrained on pure synthetic data extended from the DIV2K, Flickr2k [27] and OutdoorSceneTraining [24] datasets. The super-resolution experiments were evaluated using the PSNR, SSIM and LPIPS metrics. For LPIPS, the lower the value the better; it means the resulting image is closer to the ground-truth. The opposite holds for the other two metrics, for which higher values are better.

Since low-resolution can be similar to other types of image degradation, the experiments were extended to compare the performance between Adaptive Target model and those of the default Real-ESRGAN and its Real-ESRNet variant in the task of image deblurring. For this set of experiments the REDS dataset [28] was used. Specifically, we experiment with a REDS subset that exclusively contains motion-blurred images and another REDS subset that contains images which were motioned blurred and downscaled with a factor of x4. This was done because both GAN architectures are primarily designed as super-resolution models, so it was interesting to evaluate how they behave in

image deblurring both when super-resolution is and is not involved. By evaluating their performance on motion-blurred images and downscaled images (blurred and then downscaled by a factor of x4), we aim to understand how the models behave in scenarios where image quality is compromised by motion blur, which is a common issue in fast-moving scenes. The connection to super-resolution lies in the fact that both tasks involve enhancing image quality and details, albeit in different contexts—super-resolution aims to increase image resolution and clarity, while deblurring focuses on restoring sharpness and reducing blur. The experiments thus explore the pretrained models' ability to generalize beyond their original task and demonstrate their potential utility across multiple image enhancement domains, highlighting the versatility and robustness they possess in addressing diverse image degradation challenges.

The experiments were finally concluded with the task of low-light image enhancement in images taken from the RELLISUR dataset [29], which contains images that are both in low resolution and poor lighting conditions, where both degradations are achieved by their respective factor parameters; specifically, x4 for super-resolution and x2.5 & x3 for low light (negative) exposure. The dataset contains larger factors of negative exposure, however after initial testing it was deemed unnecessary to test such large values that are >x3 for this parameter, since the pretrained Adaptive Target ESRGAN and Real-ESRGAN (default and Real-ESRNet) are not specialized for this task, thus it would not produce any useful results in those scenarios. The task of low-light image enhancement, as explored with images from the RELLISUR dataset using these models including their variants, is inherently connected to the concept of super-resolution due to the overlapping goals of improving image quality and restoring visual fidelity. While they are not specifically tailored for low-light enhancement tasks, their capabilities in super-resolution make them relevant candidates for potentially extending their training to address challenges posed by low-resolution and poor lighting conditions. In the context of RELLISUR, where images suffer from both low resolution and negative exposure factors, the models' primary function of increasing image resolution (e.g., scaling images by a factor of x4) contributes to enhancing visual details and clarity, which is essential for mitigating the effects of poor lighting. By applying them to images with low resolution and inadequate lighting, the aim is to leverage their core strengths in super-resolution to improve image quality, brightness, and overall perceptual quality, thereby bridging the gap between low-resolution, poorly lit imagery and visually appealing, high-quality representations. This task underscores their potential for versatility beyond traditional super-resolution applications, showcasing its prospective utility in addressing broader image enhancement challenges encompassing both resolution and lighting considerations.

4. Experiments and Results

The codes used for the experiments were either borrowed entirely or otherwise largely based on the GitHub repositories for the publications of [17] and [11]; specifically, the test script for blind super resolution with random kernels from [17] (https://github.com/yhjo09/AdaTarget/blob/main/test_RK.py) and the inference script for Real-ESRGAN from [11] (https://github.com/xinntao/Real-ESRGAN/blob/master/inference_realesrgan.py)

4.1. Super-resolution

The quantitative results of Real-ESRGAN and Real-ESRNet are reported in Table 4.1. and are compared to the Adaptive Target technique [17] for a scale factor x4.

Dataset:	**DIV2KRK**		
Method:	PSNR ↑	SSIM ↑	LPIPS ↓
ATG-ESRGAN x4	28.43	**0.7868**	0.3394
Real-ESRGAN x4	31.12	0.6509	**0.2332**
Real-ESRNet x4	**31.89**	0.7246	0.3192

Table 4.1: *Comparison of our implemented Real-ESRGAN and Real-ESRNet models with Adaptive Target (ATG-ESRGAN).*

Based on the findings reported in Table 4.1, the adaptive target network has the best performance for SSIM, whereas default Real-ESRGAN has the best LPIPS score and Real-ESRNet slightly edges it on the PSNR score.

We further report some qualitative results on super-resolution to compare the three techniques. Figure 4.1. refers to super-resolution results acquired with Adaptive Target, whereas Figure 4.2 and Figure 4.3. contain results for Real-ESRGAN and Real-ESRNet respectively.

Low Resolution Images *Super-resolved with Adaptive Target Network*

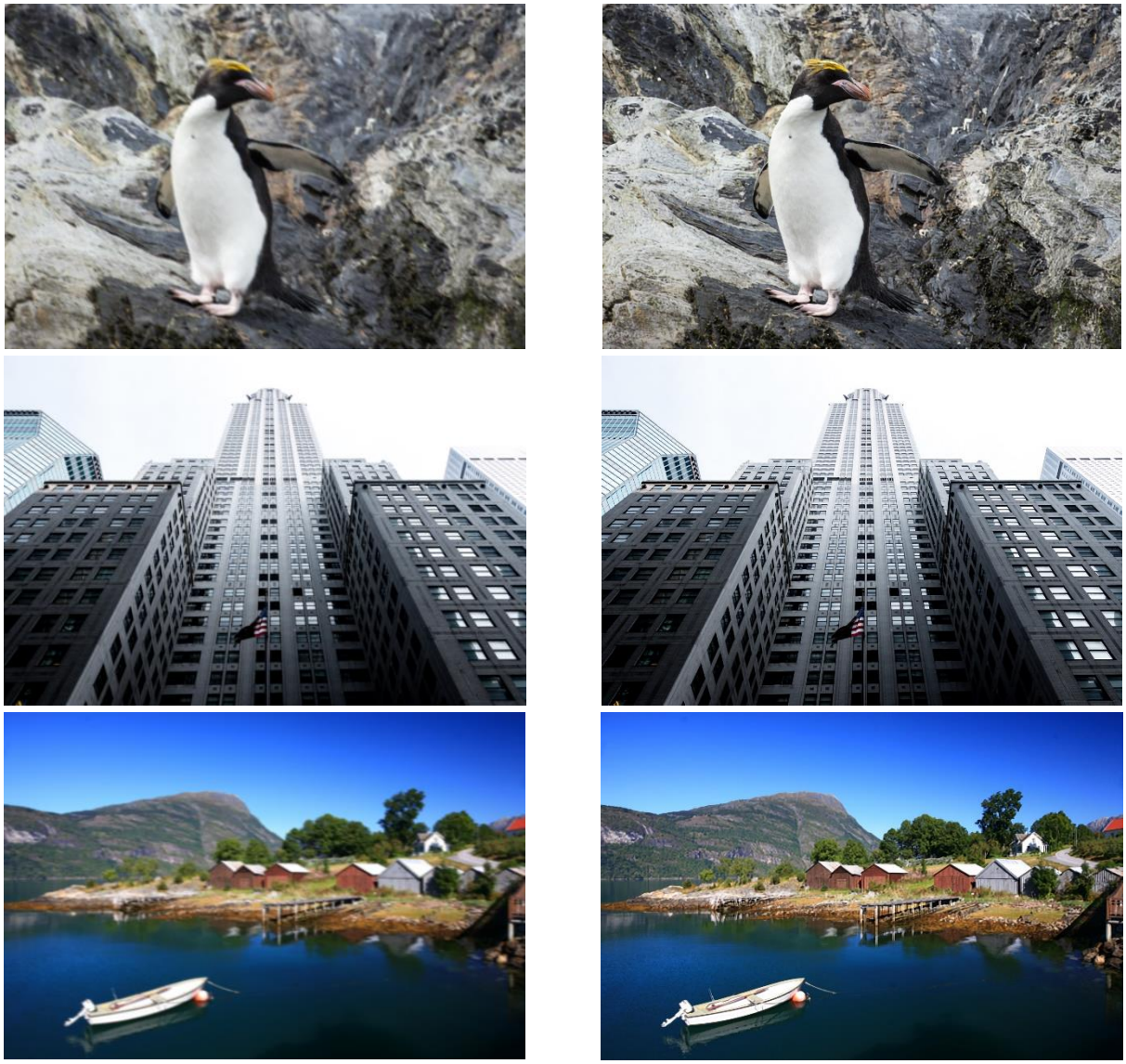

Figure 4.1: *ATG-ESRGAN Results*

Low Resolution Images *Super-resolved with Real-ESRGAN*

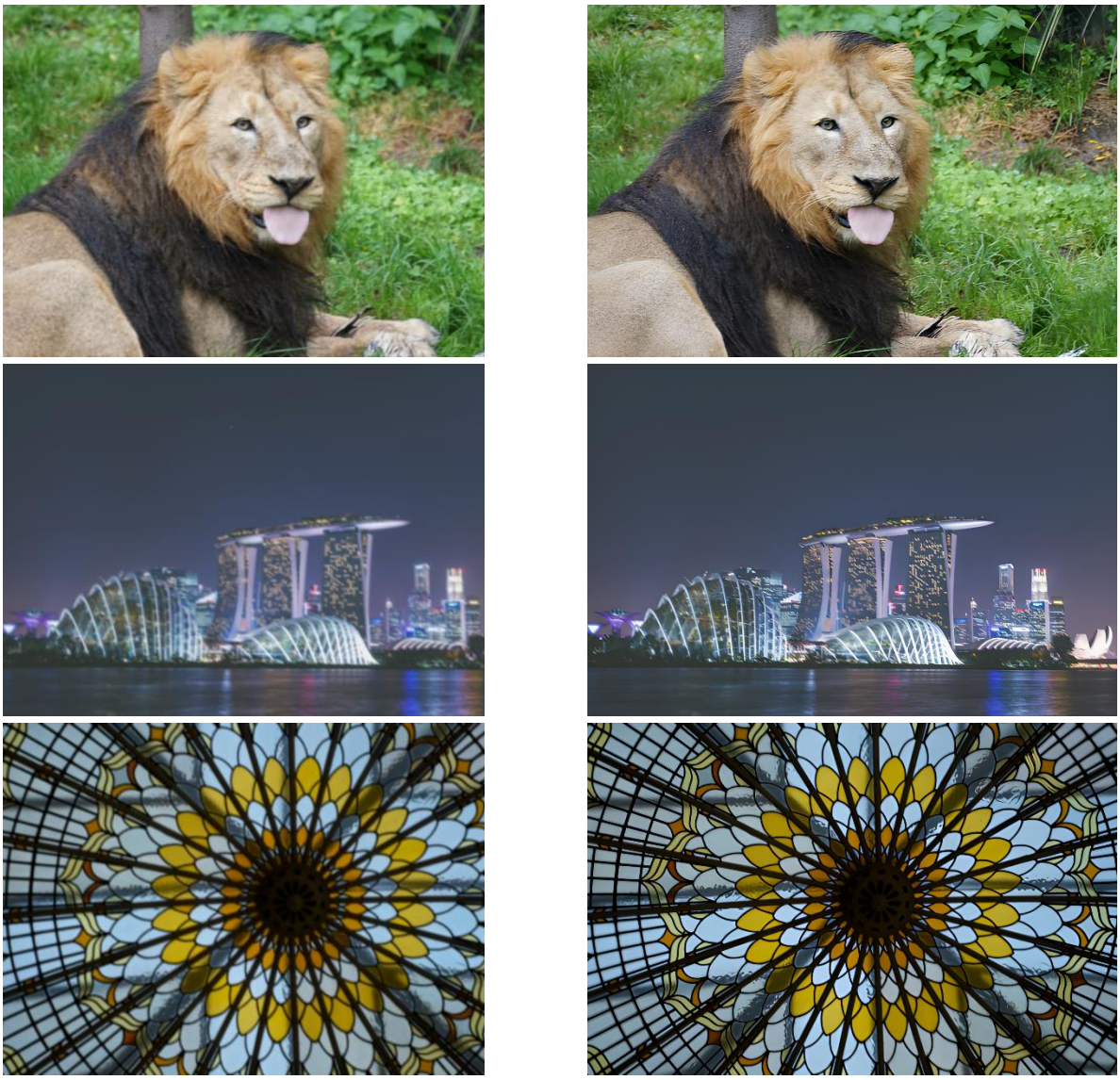

Figure 2: *Real-ESRGAN Results*

Low Resolution Images

Super-resolved with Real-ESRNet

Figure 3: *Real-ESRNet Results*

Below are some comparative qualitative results in comparison to the ground truth image for all three techniques presented previously. These results are illustrated in Figure 4.4.

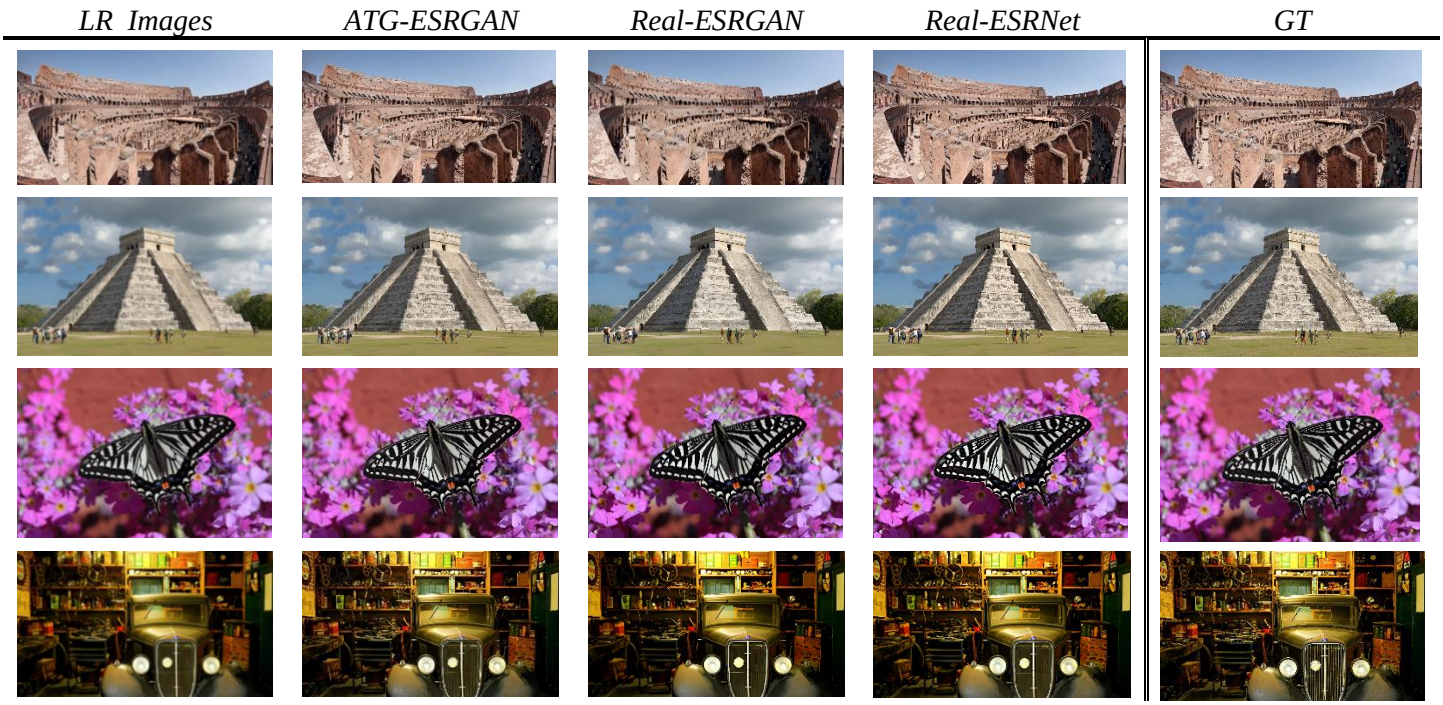

Figure 4.4: *Result Comparison to Ground Truth Images*

At a first glance, the experimental results detailed in Table 4.1 and visualized in Figures 4.1, 4.2, 4.3, and 4.4 demonstrate visually appealing outcomes across all models evaluated, indicating the effectiveness of these models in enhancing image resolution. When closely examining the images however, subtle differences in texture are observed. The Real-ESRGAN in particular seems to better in preserving overall image fidelity when the textures are a bit more complex (see figure 4.5).

Despite variations in quantitative metric scores, discernible patterns emerge in the visual quality of outputs among different models. More importantly, the results generated by the Real-ESRGAN exhibit a more natural appearance overall compared to the adaptive target and the Real-ESRNet variants, which produces clearer and less noisy results. An interesting observation on this is the relation between metric scores and visual perceptions; for instance, the default Real-ESRGAN achieves the best LPIPS score, so the resulting images do not exhibit an artificial appearance. Strong examples are for instance the *Chichén Itzá* pyramid and the car grille scenes of figure 4.5 where there are diverse textures. Notice how the comparisons between the models reveal differences in shape sharpness (crisper but more artificial door frames of pyramid car head lamp in Real-ESRNet) and in accurate representation of textures (pyramid stairs and car grille are better illustrated in

Real-ESRGAN and Adaptive Target respectively) These qualitative results suggest that the inherent characteristics of these models may contribute differently to visual outcomes depending on the subject in question.

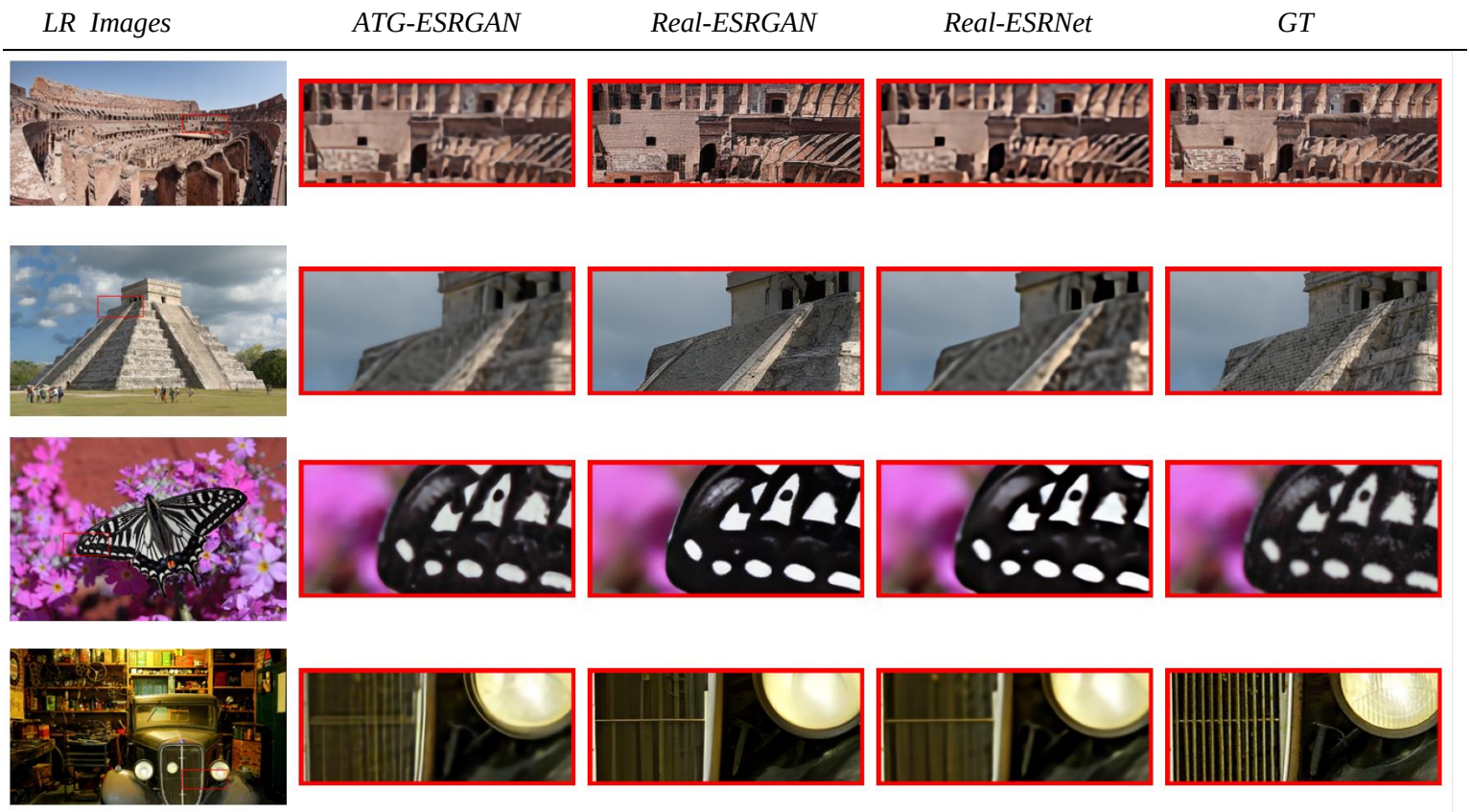

Figure 4.5: *Isolated area from images in figure 4.4 with a x10 zoom factor.*

Another thing to note is that all the methods struggled with facial features (especially eyes) in some way or another, resulting in distortion or other irregularities. This is more prominent when the face is part of a larger set in an image (a scene of people with focal depth variability) instead of a typical close-up portrait photo, which all models seemed to handle ok (figure 4.6). Consequently, we can hypothesize that models trained on non-specialized datasets for face reconstruction may struggle to properly learn and manage facial characteristics, hinting at a possible challenge towards that kind of direction when considering the aspect of generalization for various image restoration tasks.

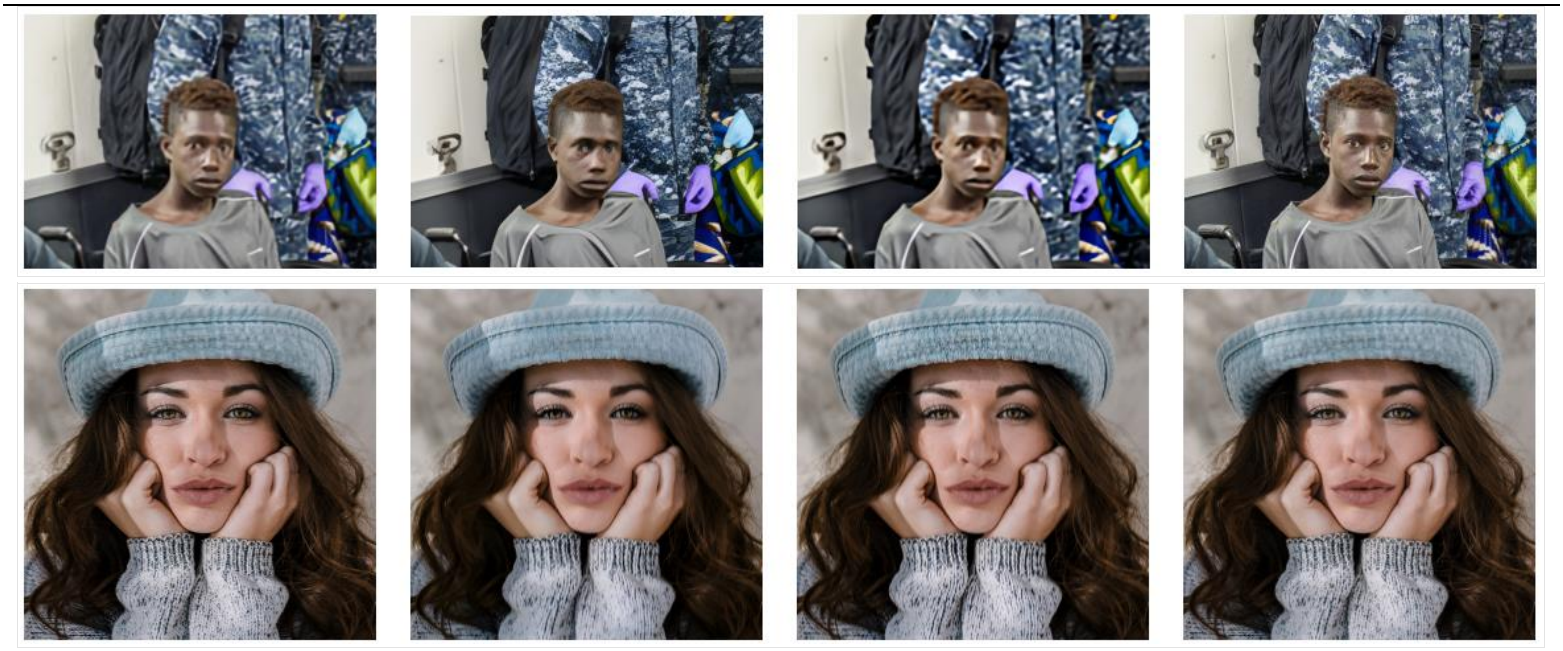

Figure 4.5: *Illustration of the performance of the models in regard to facial characteristics.*

In summary, while all models exhibit differences in quantitative scores, their actual super-resolution outputs display distinctions that are always relevant to the attribute under examination. This underscores the complexities of evaluating image enhancement models, where visual perception of textures, shapes and other details can often deviate from metric assessments, emphasizing the importance of holistic approaches that integrate both qualitative and quantitative analyses to comprehensively evaluate model performance. Based on all the aforementioned points, it is also reasonable to assume that adopting the adaptive target localization network technique during the training phase of Real-ESRGAN could potentially yield more powerful super-resolution images.

4.2. Image deblurring

In Table 4.2. we present results for the task of image deblurring applied to the REDS dataset, with emphasis given to super-resolution with scaling factor x4 and without super-resolution scaling (x1 scaling factor). For these two scaling scenarios, we evaluate Adaptive Target (ATG-ESRGAN), Real-ESRGAN and Real-ESRNet.

Dataset:		**REDS**		
Method:		PSNR ↑	SSIM ↑	LPIPS ↓
Deblurring from Same Resolution	**ATG-ESRGAN**	**32.51**	**0.8271**	**0.2214**
	Real-ESRGAN	27.91	0.1889	0.7337
	Real-ESRNet	27.92	0.1861	0.7218
Deblurring from Low Resolution x4	**ATG-ESRGAN x4**	31.59	0.7125	0.3156
	Real-ESRGAN x4	27.93	0.1926	0.7119
	Real-ESRNet x4	27.91	0.2047	0.7386

Table 4.2: *Deblurring results on REDS datasets without and with super-resolution using ATG-ESRGAN, Real-ESRGAN and Real-ESRNet models.*

The ESRGAN model combined with the adaptive target technique appear to have significantly outperformed the other two models in all metrics, although such dramatic differences in performance are not reflected in the qualitative results (we further discuss this below). The other two models performed slightly better when the super-resolution of scaling factor x4 was also involved in the image deblurring. Under this condition, the default Real-ESRGAN scored better in PSNR and LPIPS, while the Real-ESRNet variant had a better SSIM performance.

To expand upon the quantitative measurements of Table 4.2, especially the apparent significant difference in performance between the adaptive target and the other two models, we need to put these results under scrutiny. Due to the fact that the three networks rely on the same core model, the ESRGAN, we deduce this drastic difference cannot really be an accurate representation of the variance in performance among these three solutions, but rather a difference on how they handle the metric scores. This idea is backed up from the relatively similar qualitative results shown in figures 4.6, 4.7, 4.8 and 4.9. There must be other reasons therefore that are possible explanations.

Let us think this through for a moment. First of all, we know that all three networks are pretrained with the same, or at worst, very similar datasets, such as the DIV2KRK. We also observed their consistent qualitative and quantitative performance on a typical super-resolution task. Most likely then the large difference in metric scores for the deblurring task arises from differences in how the models handle the image deblur task, pointing to potential architectural or training approach differences that specifically influence the performance in metrics for this type of distortion. In other words, the reasons for this are most probably relevant to how the adaptive target model optimizes for

or interacts with these metrics, rather than outright differences in deblurring performance. In addition to this, other factors could stem from the metrics themselves and from the type of the image restoration task.

As we have seen in previous sections, the concept of the adaptive target is based on training an additional localization network which essentially boosts and improves the model's training process. This happens within the generator component of the model (see again figure 3.2), where it generates slightly different image patches from what was fed to it by deploying simple and limited pixelwise affine transformations. The resulting target patch is almost but not absolutely identical to the input. This of course has the benefit of make the most (if not more) of a given dataset, since this process can in some ways be viewed as "synthesizing" extra data images. But as a downside it can also introduce certain biases. Truly, adaptive target ESRGAN might be better at minimizing specific artifacts or distortions (e.g., ringing or oversmoothing) that disproportionately affect metrics that evaluate structural similarity like SSIM, LPIPS . Its localization model could by design give particular emphasis in this aspect when performing the previously mentioned affine transformation (i.e. optimized loss functions for this), making the model inherently score much better in these metrics. This of course is somewhat problematic, because it means that the model can be prone to overfitting towards certain types of metrics. However, in our study this is observed only with this type of image degradation (image blur) and not with the standard task the adaptive target model was designed for (super-resolution), meaning that the frequencies and types of tools used for a specific image distortion also play a role. This could mean that currently the model has a hindered ability of generalization unless its creators consider mitigating the issues described.

Given all the above then, the most plausible explanation as to why the adaptive target ESRGAN has drastically better scores for SSIM and LPIPS is that it ends up aligning more closely with the ground truth, either in terms of pixel-level accuracy or perceptual feature similarity, due to differences in a combination of architecture, loss functions, or how it processes high-frequency details. This alignment improves the scores, even if the perceptual quality appears comparable to the human eye, as shown below throughout figures 4.6-4.9.

Figure 4.6: *Result Comparison to Ground Truth for Image Deblurring without Super Resolution*

Figure 4.7: *Isolated area from images in figure 4.6 with a x10 zoom factor*

None of the models has produced a result that is drastically observable to the human eye. Perhaps a very close examination of the images might lead to the discovery of some minor differences in sharpness and overall quality. This of course underpins the notion that these models are primarily designed as SR solutions, therefore some level of modification needs to be done before they can be generalized to other tasks The testing however of their performance in the task of image deblurring has given some interesting observations.

Figure 4.8: *Result Comparison to Ground Truth for Image Deblurring with Super Resolution x4*

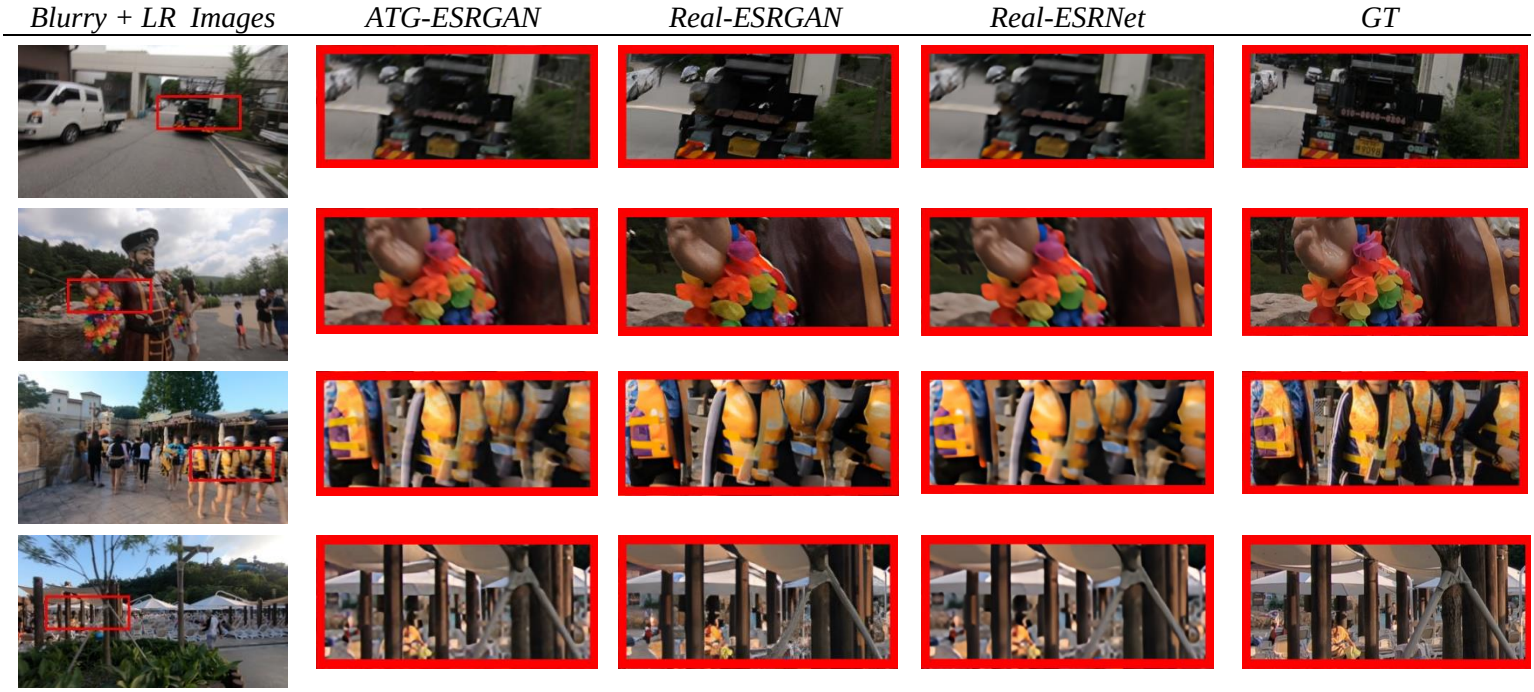

Figure 4.8: *Result Comparison to Ground Truth for Image Deblurring with Super Resolution x4*

Firstly, although none of the models perform exceptionally well, this effect is more apparent when the blur is intense (due to faster moving subjects). The discrepancy between pretraining being geared towards SR and the test data exhibiting primarily the issue of motion-blur with or without low resolution issues is a contributing factor in achieving proper deblurring, even though it is to be expected. In that case perhaps the choice of a more generalized model that has also been pretrained on numerous and more various datasets might be more appropriate, because it has incorporated some feature knowledge regarding motion blur in images. However, this option deviates from the scope of our current analysis and would requires larger datasets and more high-end computational resources.

It is also important to mention though that when the two models were tested on the subset of REDS which contains images both blurred and in low resolution, the results of the deblurring process were better. For some reason, the image upscaling using a factor x4 resulted in a perceptual side effect of improving some of the blur; however, this is unlikely to be the actual case since this is hardly observable and only in those stills where the subjects were moving slowly, thus the blur was not much in the first place. For faster moving subjects no such side effect was observed. Overall, the side effect was adequate towards distinguishing between the scenarios of merely deblurring the motion blur of the image (scale factor x1) and deblurring with super-resolution (scale factor x4). The reported finding can be attributed to several factors. In scenes where subjects exhibit mild and spatially coherent motion blur due to slow movement, the process of upscaling the image to a higher

resolution can inadvertently play a role in regard to this blur. The significant increase in image size during upscaling may introduce smoothing or averaging effects that help to partially reduce mild motion blur, enhancing the overall clarity of these scenes. This however was not consistently observed for faster moving subjects, where motion blur tends to be more severe and complex, making it impossible to describe it as successful image deblurring. This finding though hints to the natural connection between image super-resolution and motion deblurring, justifying the intention to select a SR model and test it to address image deblurring.

All the facts so far suggest that if the pretrained ESRGAN architectures are further trained with a dataset such as REDS, they could present promising potential to perform well in the image deblurring problem. This would be particularly true if those architectures could be enhanced with components aimed at removing the motion blur, examples of which can be found in the relevant challenge competition involving the dataset as shown in the respective paper [30]. Specifically, dynamic motion blurs and MPEG video compression artifacts incorporated in the REDS dataset are addressed in the NTIRE competition from many participating teams, concluding that employing temporal information, video deblurring techniques and ensembles which address varying directions of motion are essential to resolve persistent motion blur challenges. Thus, we could potentially propose a two-stream architecture, in which the one upsampling stream would be dedicated to motion blur artifacts, similar to the solutions proposed in [30] while the second one would target super-resolution similar to ESRGAN with adaptive target and Real-ESRGAN. The fusion of the two streams in a final module could address both challenges in the best possible way. Another simpler solution would focus on the data rather than the architecture; in this case, blurred and deblurred image pairs would be fed in a super-resolution architecture, such as the ESRGANs of this paper, and they would be trained to restore such relationships. However, it is unsafe to assume that any of them would be adequate for deblurring, since more specialized architectures yield more impressive results.

4.3. Low-light image enhancement

In Table 4.3 we present quantitative results for Adaptive Target (ATG-ESRGAN), Real-ESRGAN and Real-ESRNet variants for low-light image enhancement of negative exposure with factors x2.5 and x3. All models are once again used with a x4 super-resolution scaling factor in order to ascertain their performance on a more even and somewhat better suited basis, because of the inclusion of the super resolution task.

Dataset:		**RELLISUR**		
Method:		PSNR ↑	SSIM ↑	LPIPS ↓
Negative Exposure Factor x2.5	**ATG-ESRGAN x4**	**27.87**	0.2878	0.5706
	Real-ESRGAN x4	27.77	0.3111	**0.3898**
	Real-ESRNet x4	27.80	**0.3181**	0.4025
Negative Exposure Factor x3	**ATG-ESRGAN x4**	27.85	0.2440	0.5125
	Real-ESRGAN x4	27.84	0.2172	0.4438
	Real-ESRNet x4	27.85	0.2355	0.4589

Table 4.3: *Results on the low light image enhancement task (RELLISUR dataset) for ATG-ESRGAN, Real-ESRGAN and Real-ESRNet.*

Based on the findings of Table 4.3, there was not a great difference in numerical performance between the two negative exposure factors that were used in the tests. The models performed slightly better for the smaller 2.5 factor, a fact that is even more prevalent in the SSIM and LPIPS scores.
Below, we present some qualitative results in Figures 4.7 (negative factor x2.5) and 4.8 (negative factor x3) for all three models and the ground truth high-exposure image.

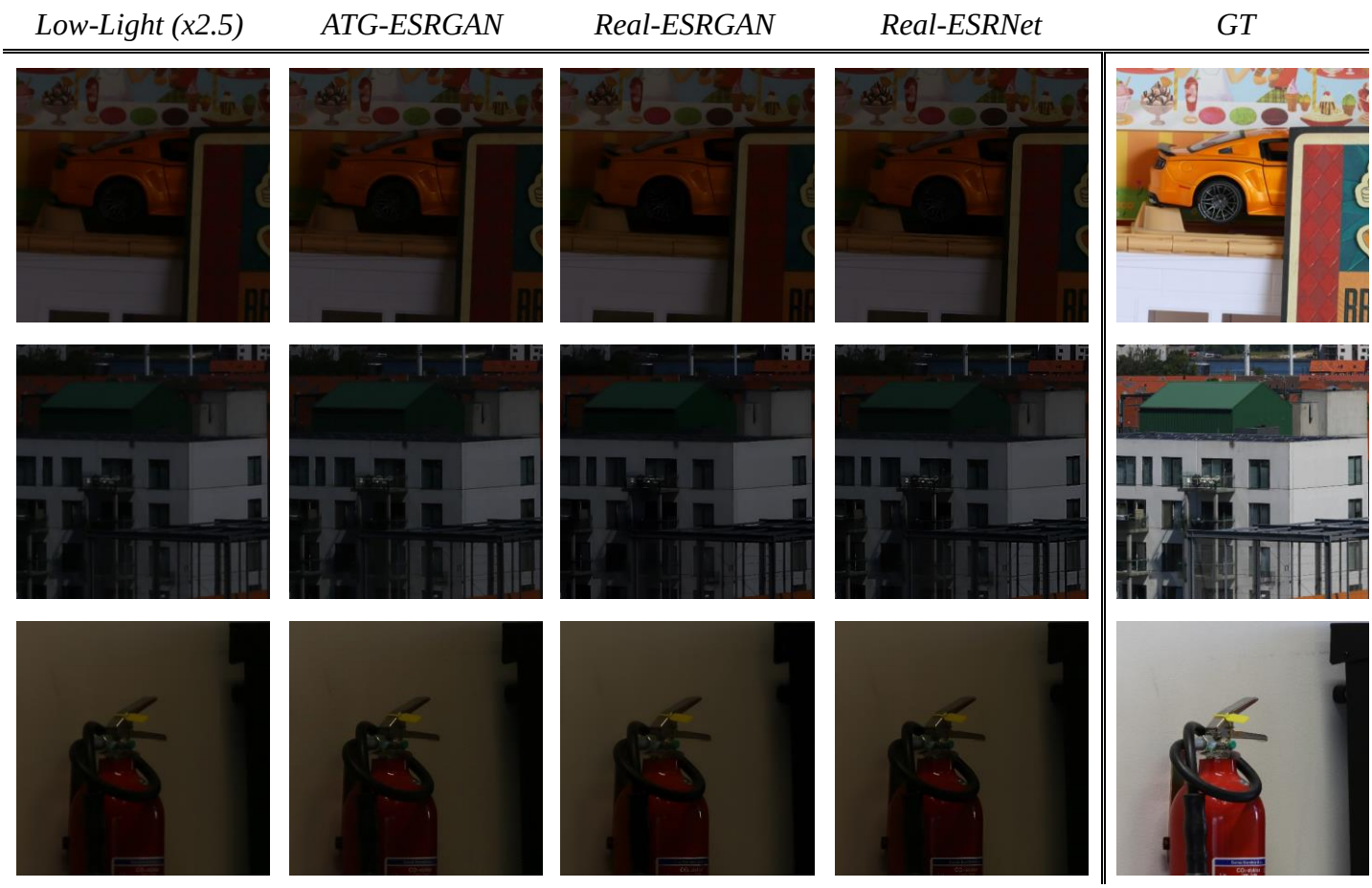

Figure 4.9. *Result Comparison to Ground Truth for Low-light Enhancement from Negative Factor x2.5*

It is apparent that there is not a great difference in terms of the lighting conditions. Although the models have no direct effect on the lighting, just by restoring the resolution of an image, some details become more visible, especially when it comes to aspects such as reflective surfaces etc. even when the exposure has largely remained the same. This is visible in figures 4.11 and 4.12.

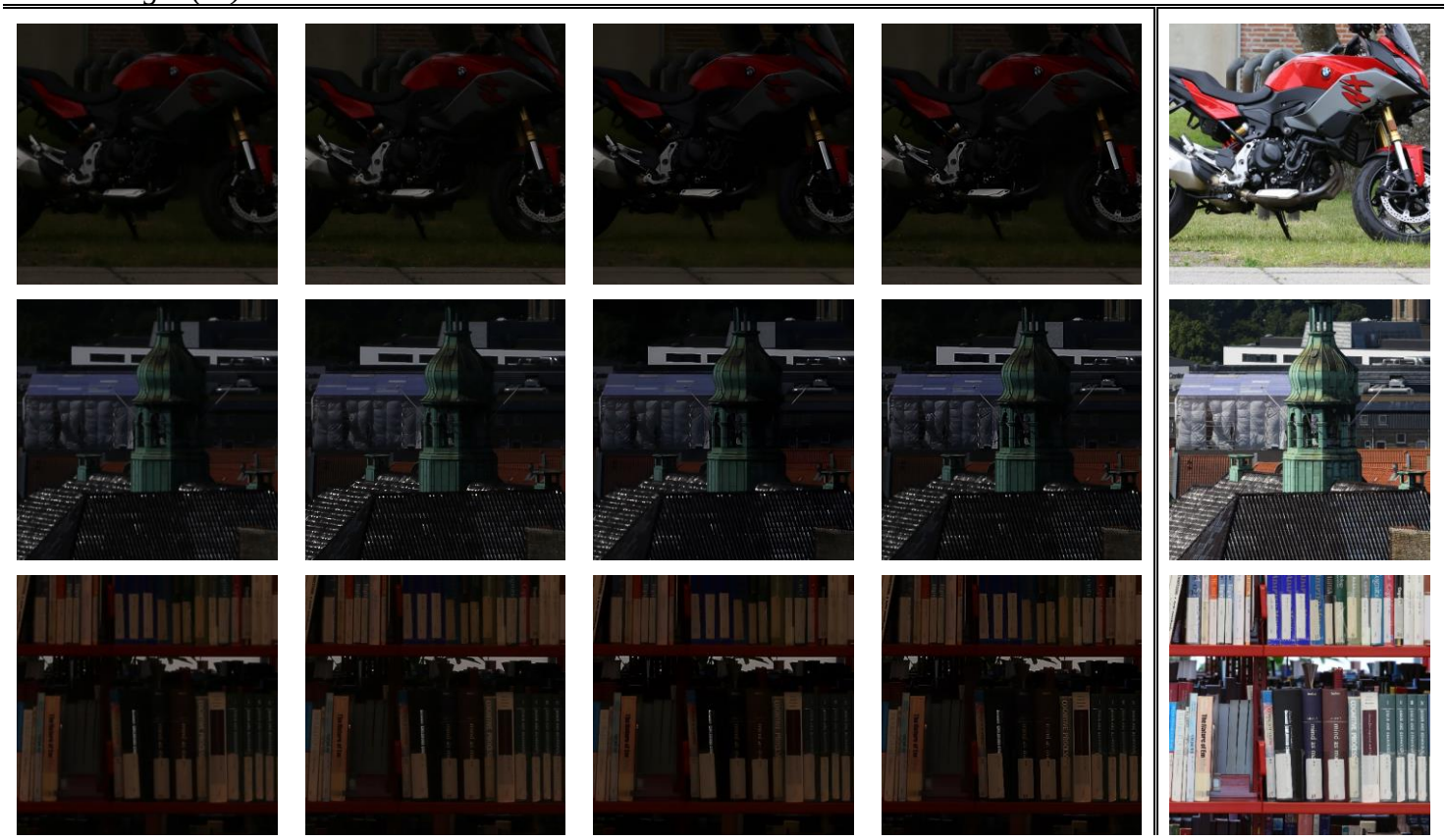

Figure 4.10: *Result Comparison to Ground Truth for Low-light Enhancement from Negative Factor x3*

The situation does not change among different negative exposure factors. The same thing can be observed with reflective surfaces, such as for example all the metal parts of the motorbike images in figure 4.10.

As mentioned above, we present here some additional results to illustrate the models' capabilities for text and reflective surfaces in Figures 4.11 (negative factor x2.5) and 4.12 (negative factor x3). Images featuring text and reflective surfaces present notable challenges for low-light enhancement algorithms, commonly encountered in scenarios like nighttime photography or dimly lit indoor environments. These elements contain intricate details crucial for assessing enhancement effectiveness, necessitating methods that can preserve fine details while minimizing noise and artifacts to ensure clear legibility. Despite the high contrast between bright and dark areas in such images which lead to issues like haloing or glare, the models' algorithms are capable of effectively clearing these contexts even if these contrast variations are not directly addressed in order to achieve natural and visually appealing results in low-light conditions.

Figure 4.11: *Enhancement Results for Text and Reflective Surfaces from Negative Factor x2.5*

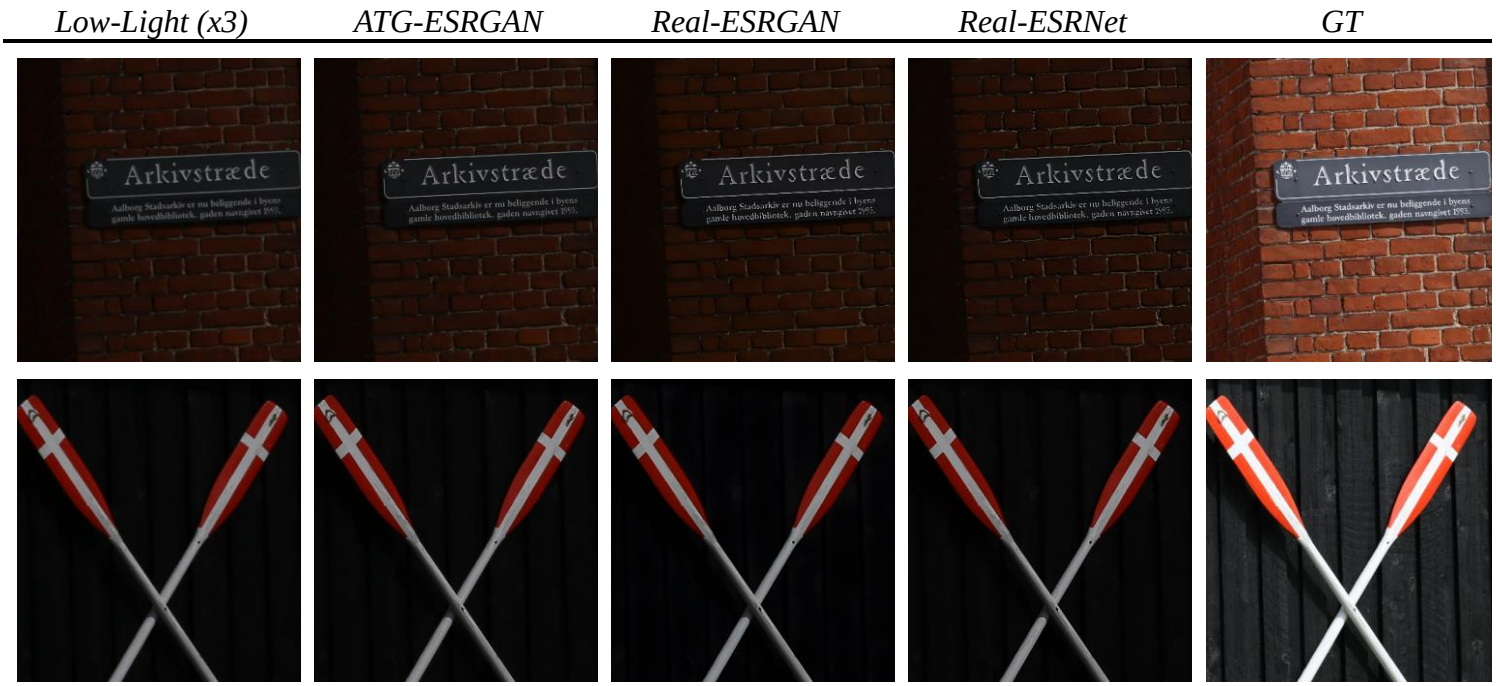

Figure 4.12: *Enhancement Results for Text and Reflective Surfaces from Negative Factor* x3

Conclusively, none of the models and their variants had a significant effect in restoring normal light condition for under-exposed images. There were however a few interesting observations.

The ESRGAN with the Adaptive Target implementation and Real-ESRGAN models are primarily developed and pretrained as general super-resolution frameworks, meaning they are designed to enhance the resolution and quality of images without specific modules or training dedicated to addressing light-related degradations. Unlike specialized models such as EnlightenGAN [31], which are specifically tailored for tasks involving light enhancement or low-light image restoration, the models used in this paper focus on improving image resolution and sharpness across a wide range of scenarios. As a result, they do not incorporate dedicated mechanisms to handle challenges related to low-light conditions, such as noise reduction, contrast enhancement, or artifact suppression caused by insufficient lighting. In contrast, models like EnlightenGAN are trained with a specific focus on addressing light-related degradations, leveraging techniques like conditional adversarial training to effectively enhance image quality in low-light environments by adjusting brightness, contrast, and overall illumination. Therefore, while Adaptive Target ESRGAN and Real-ESRGAN excel in general super-resolution tasks, specialized models like EnlightenGAN may offer superior performance and efficacy in scenarios where light-related degradations are a primary concern. This however does not mean that there is not any potential to generalize models of the ESRGAN family to tackle the low-light image enhancement problem.

This stems from the testing with the RELLISUR dataset, where notable observations were made regarding the performance of Adaptive Target ESRGAN and Real-ESRGAN variant models on images containing text and reflective surfaces, such as white writing on signs and white-painted window sills. Interestingly, certain images processed with the models showed a subtle yet noticeable

increase in perceived brightness specifically around these elements (as depicted in Figure 4.11). This effect can be attributed to the inherent characteristics of the models which excel in achieving highly effective super-resolution and may inadvertently enhance the appearance of other image components. The perceived increase in brightness around text and reflective surfaces is likely a by-product of the model's optimization for resolution enhancement, potentially leading to more artificial-looking results in some cases. This observation underscores the importance of critically evaluating the outputs of super-resolution models, particularly in scenarios involving text and reflective materials, to ensure that enhancements align with desired aesthetic and perceptual qualities without introducing unintended artifacts or distortions.

The findings discussed above suggest promising potential for further enhancing the performance of pretrained Adaptive Target ESRGAN and Real-ESRGAN models when trained with specialized datasets like RELLISUR. By fine-tuning them on datasets that specifically emphasize challenges related to low-light conditions, the models can adapt and learn to address nuances associated with brightness, contrast, and image fidelity in dimly lit environments. Training them with RELLISUR has the potential to enhance their ability to handle text and reflective surfaces, as observed in the dataset, by refining their capabilities to preserve and enhance critical image details while mitigating artifacts commonly associated with low-light image processing. This approach leverages the strengths of these models in super-resolution while tailoring their performance to excel in specific tasks related to light enhancement and low-light image restoration. Therefore, further training with datasets like RELLISUR holds promise for extending the models' applicability and effectiveness in addressing real-world challenges associated with low-light photography and imaging scenarios.

5. Conclusion

Within the scope of this study, we investigated the challenge of super-resolution by utilizing dynamic GAN-based structures that were specifically designed for this purpose. By complementary using synthetic pictures in a variety of various ways, one of the primary issues surrounding the ill-posed nature of the super-resolution problem may be handled. This, in turn, suggests the selection of model architectures that make use of methodologies that are also applicable to such synthetic images. When we take into account the fact that super-resolution is closely related to other picture reconstruction tasks, such as image deblurring and low-light image improvement, we conduct tests to evaluate transfer learning and determine the capacity of super-resolution models to perform on these related tasks. We obtain a variety of results throughout the course of our experimentation, which demonstrates the capabilities of high-end super-resolution models while also highlighting their shortcomings. For example, the disparity between the cues connected to deblurring and low-light image enhancement is not negligible from the perspective of the super-resolution model. As a result of this, even high-performing super-resolution models have difficulty modeling the upscaling techniques that are associated with these jobs, which demonstrates the necessity of pretraining on specific datasets.

As a way of conducting experiments on a variety of datasets that include synthetic and/or natural data pictures, it would be beneficial to examine the integration of an adaptive target network with the state-of-the-art architecture of the ESRGAN family of models in the course of future research activities. By using this approach, it is possible to improve the performance of the models in a manner that might potentially solve difficulties related to super-resolution, and even perhaps equivalent tasks such as picture deblurring and low-light image improvement. As a specific example, training this combination technique with specialized datasets that are targeted for both deblurring and low-light settings might further unlock the potential of these models. Although their pretrained versions have demonstrated initial promise in these areas (particularly the Real-ESRGAN), our findings indicate that additional training with specialized datasets is required in order to unlock their potential to significantly improve their effectiveness and generalization applicability for difficult image restoration tasks. This is because our findings suggest that extensive training with specialized datasets is required. The exploration of this combination with specialized training datasets is a tempting route for additional study, with the ultimate objective being to solve even more diversified picture enhancing difficulties to be addressed.

To further enhance the versatility of super-resolution models, future research should also explore the fusion of multi-stream architectures designed to target specific image degradation types. For instance, a two-stream approach could involve one stream dedicated to super-resolution and another optimized for complementary tasks such as deblurring or low-light enhancement. The integration of outputs from both streams could result in models that can better generalize across a broader range of image restoration scenarios. Additionally, leveraging advanced loss functions tailored to different perceptual and structural features could refine these models' abilities to handle variations in image textures, sharpness, and exposure. Such enhancements would address the inconsistencies observed in

our experiments, where models struggled to fully resolve complex visual challenges such as motion blur or the accurate restoration of reflective surfaces under low-light conditions.

Moreover, it is crucial to consider the role of diverse and robust training datasets when improving super-resolution models. Models that primarily rely on limited datasets for training and validation may encounter difficulties when exposed to real-world scenarios involving more severe and diverse image degradations. To mitigate this, expanding training datasets to include varied conditions such as different lighting environments, motion dynamics, and occlusions would provide models with a more comprehensive feature set to learn from. This, in turn, could lead to better handling of artifacts and distortions, such as those noted in scenes with intricate textures and other relevant attributes. The importance of dataset augmentation, including techniques like synthetic transformations, is another promising avenue to ensure models learn to adapt to unseen degradation patterns without overfitting to particular image types or restoration metrics.

Finally, we have seen that a usual evaluation framework for super-resolution models might need refinement to better reflect perceptual quality. Our study demonstrates that while traditional metrics like PSNR, SSIM, and LPIPS provide a numerical basis for comparison, they might not always align with human visual perception as was the case in the task of image deblurring. Incorporating perceptual studies, where human evaluators assess model outputs, can provide a more comprehensive understanding of model performance. Additionally, emerging perceptual metrics and neural network-based evaluators designed to simulate human vision could enhance the assessment of image restoration models, ensuring that improvements are both quantitatively measurable and visually meaningful. By combining better training data, improved architectures, and more reliable evaluation methods, the next generation of super-resolution models could achieve superior generalization across a wide range of image restoration/enhancement tasks.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems 27 (NeurIPS 2014) pp. 2672–2680. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[2] Wang, Z., Chen, J., & Hoi, S. C. H. (2019). Deep Learning for Image Super-resolution: A Survey. arXiv preprint arXiv:1902.06068. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9044873

[3] Tikhonov, A. N., & Arsenin, V. Y. (1977). "Solutions of Ill-Posed Problems." V. H. Winston & Sons. https://www.bibsonomy.org/bibtex/9aacc055724bc6d774982fca78c5d2d9

[4] K. Zhang, W. Zuo, and L. Zhang, “Learning a single convolutional super-resolution network for multiple degradations,” in CVPR, 2018 https://arxiv.org/abs/1712.06116

[5] Y. Bei, A. Damian, S. Hu, S. Menon, N. Ravi, and C. Rudin, “New techniques for preserving global structure and denoising with low information loss in single-image super-resolution,” in CVPRW, 2018 https://arxiv.org/abs/1805.03383

[6] A. Bulat, J. Yang, and G. Tzimiropoulos, “To learn image superresolution, use a gan to learn how to do image degradation first,” in ECCV, 2018 https://arxiv.org/abs/1807.11458

[7] Benny, Y., Galanti, T., Benaim, S., & Wolf, L. (2021). Evaluation Metrics for Conditional Image Generation. International Journal of Computer Vision, 129(5), 1712–1731. https://doi.org/10.1007/s11263-020-01424-w

[8] Li, Z. (2023). Image Deblurring using GAN. arXiv preprint arXiv:2312.09496. https://arxiv.org/abs/2312.09496

[9] Chen, C., Chen, Q., Xu, J., & Koltun, V. (2018). Learning to See in the Dark. arXiv preprint arXiv:1805.01934. https://arxiv.org/abs/1805.01934v1

[10] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., & Tang, X. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. arXiv preprint arXiv:1809.00219. https://arxiv.org/abs/1809.00219

[11] Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. arXiv preprint arXiv:2107.10833. https://arxiv.org/abs/2107.10833

[12] Luo, Z., Huang, Y., Li, S., Wang, L., & Tan, T. (2020). Unfolding the alternating optimization for blind super resolution. In NeurIPS. https://arxiv.org/abs/2010.02631

[13] Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L. (2020). Component divide-and-conquer for real-world image super-resolution. In ECCV. https://arxiv.org/abs/2008.01928

[14] Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., & Huang, F. (2020). Real-world super-resolution via kernel estimation and noise injection. In CVPRW. https://openaccess.thecvf.com/content_CVPRW_2020/papers/w31/Ji_Real-World_Super-Resolution_via_Kernel_Estimation_and_Noise_Injection_CVPRW_2020_paper.pdf

[15] Zhang, K., Liang, J., Van Gool, L., & Timofte, R. (2021). Designing a practical degradation model for deep blind image super-resolution. In ICCV. https://openaccess.thecvf.com/content/ICCV2021/papers/Zhang_Designing_a_Practical_Degradation_Model_for_Deep_Blind_Image_Super-Resolution_ICCV_2021_paper.pdf

[16] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super resolution using a generative adversarial network. In CVPR. https://arxiv.org/abs/1609.04802

[17] Jo, Y., Oh, S. W., Vajda, P., & Kim, S. J. (2021). Tackling the Ill-Posedness of Super-Resolution Through Adaptive Target Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 16236-16245). https://openaccess.thecvf.com/content/CVPR2021/html/Jo_Tackling_the_Ill-Posedness_of_Super-Resolution_Through_Adaptive_Target_Generation_CVPR_2021_paper.html

[18] Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In European Conference on Computer Vision (ECCV). https://link.springer.com/chapter/10.1007/978-3-319-10593-2_13

[19] Schuler, C. J., Burger, H. C., Harmeling, S., & Scholkopf, B. (2013). A machine learning approach for non-blind image deconvolution. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 1067–1074). https://ieeexplore.ieee.org/document/6618986

[20] Kim, J., Lee, J. K., & Lee, K. M. (2015). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. arXiv preprint arXiv:1511.04587. https://openaccess.thecvf.com/content_cvpr_2016/papers/Kim_Accurate_Image_Super-Resolution_CVPR_2016_paper.pdf

[21] Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. arXiv preprint arXiv:1707.02921. https://openaccess.thecvf.com/content_cvpr_2017_workshops/w12/papers/Lim_Enhanced_Deep_Residual_CVPR_2017_paper.pdf

[22] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. arXiv preprint arXiv:1603.08155. https://link.springer.com/chapter/10.1007/978-3-319-46475-6_43

[23] Sajjadi, M. S. M., Scholkopf, B., & Hirsch, M. (2017). Enhancenet: Single image super-resolution through automated texture synthesis. In ICCV. https://openaccess.thecvf.com/content_ICCV_2017/papers/Sajjadi_EnhanceNet_Single_Image_ICCV_2017_paper.pdf

[24] Wang, X., Yu, K., Dong, C., & Loy, C. C. (2018). Recovering realistic texture in image super-resolution by deep spatial feature transform. arXiv preprint arXiv:1804.02815. https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_Recovering_Realistic_Texture_CVPR_2018_paper.pdf

[25] Bell-Kligler S., Shocher A, Irani M. (2019). KernelGAN: Blind Super-Resolution Kernel Estimation using an Internal-GAN. NeurIPS 2019 (oral) (DIV2KRK dataset (https://www.wisdom.weizmann.ac.il/~vision/kernelgan/).)

[26] Agustsson, E., & Timofte, R. (2017). NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (pp. 1122-1131). July 2017. https://data.vision.ee.ethz.ch/cvl/DIV2K/

[27] Flickr2K, Huggingface https://huggingface.co/datasets/goodfellowliu/Flickr2K

[28] Son, S., Lee, S., Nah, S., Timofte, R., & Lee, K. M. (2021). NTIRE 2021 Challenge on Video Super-Resolution. In CVPR Workshops (pp. 166-181). June 2021. https://seungjunnah.github.io/Datasets/reds

[29] Aakerberg, A., Nasrollahi, K., & Moeslund, T. B. (2021). RELLISUR: A Real Low-Light Image Super-Resolution Dataset. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://vap.aau.dk/rellisur/

[30] S. Nah *et al.*, "NTIRE 2019 Challenge on Video Deblurring: Methods and Results," *2*019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019, pp. 1974-1984, doi: 10.1109/CVPRW.2019.00249. https://ieeexplore.ieee.org/document/9025548

[31] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., & Wang, Z. (2021). EnlightenGAN: Deep Light Enhancement without Paired Supervision. arXiv preprint arXiv:1906.06972. https://arxiv.org/abs/1906.06972

[32] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1874-1883). https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Shi_Real-Time_Single_Image_CVPR_2016_paper.pdf

[33] Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., & Lin, L. (2018). Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 701-710). https://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w13/Yuan_Unsupervised_Image_Super-Resolution_CVPR_2018_paper.pdf

[34] Lugmayr, A., Danelljan, M., & Timofte, R. (2019, October). Unsupervised learning for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 3408-3416). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9022038

[35] Liu, Z. S., Siu, W. C., Wang, L. W., Li, C. T., & Cani, M. P. (2020). Unsupervised real image super-resolution via generative variational autoencoder. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 442-443). https://openaccess.thecvf.com/content_CVPRW_2020/papers/w31/Liu_Unsupervised_Real_Image_Super-Resolution_via_Generative_Variational_AutoEncoder_CVPRW_2020_paper.pdf

[36] Prajapati, K., Chudasama, V., Patel, H., Upla, K., Ramachandra, R., Raja, K., & Busch, C. (2020). Unsupervised single image super-resolution network (USISResNet) for real-world data using generative adversarial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 464-465). https://openaccess.thecvf.com/content_CVPRW_2020/papers/w31/Prajapati_Unsupervised_Single_Image_Super-Resolution_Network_USISResNet_for_Real-World_Data_Using_CVPRW_2020_paper.pdf

[37] Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., & Song, H. (2021). Unsupervised real-world image super resolution via domain-distance aware training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13385-13394). https://openaccess.thecvf.com/content/CVPR2021/papers/Wei_Unsupervised_Real-World_Image_Super_Resolution_via_Domain-Distance_Aware_Training_CVPR_2021_paper.pdf

[38] Sun, W., & Chen, Z. (2024). Learning Many-to-Many Mapping for Unpaired Real-World Image Super-resolution and Downscaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://arxiv.org/pdf/2310.04964

[39] Haris, M., Shakhnarovich, G., & Ukita, N. (2018). Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1664-1673). https://openaccess.thecvf.com/content_cvpr_2018/papers/Haris_Deep_Back-Projection_Networks_CVPR_2018_paper.pdf

[40] Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., & Wu, W. (2019). Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3867-3876). https://openaccess.thecvf.com/content_CVPR_2019/papers/Li_Feedback_Network_for_Image_Super-Resolution_CVPR_2019_paper.pdf

[41] Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, *38*(2), 295-307. https://arxiv.org/pdf/1501.00092

[42] Dong, C., Loy, C. C., & Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (pp. 391-407). Springer International Publishing. https://www.researchgate.net/profile/Chen-Change-Loy/publication/305779418_Accelerating_the_Super-Resolution_Convolutional_Neural_Network/links/57c6331f08ae0a6b0dc8dff6/Accelerating-the-Super-Resolution-Convolutional-Neural-Network.pdf

[43] Lai, W. S., Huang, J. B., Ahuja, N., & Yang, M. H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 624-632). https://openaccess.thecvf.com/content_cvpr_2017/papers/Lai_Deep_Laplacian_Pyramid_CVPR_2017_paper.pdf

[44] Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O., & Schroers, C. (2018). A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 864-873). https://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w13/Wang_A_Fully_Progressive_CVPR_2018_paper.pdf

[45] Kim, J., Lee, J. K., & Lee, K. M. (2016). Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1637-1645). https://openaccess.thecvf.com/content_cvpr_2016/papers/Kim_Deeply-Recursive_Convolutional_Network_CVPR_2016_paper.pdf

[46] Tai, Y., Yang, J., & Liu, X. (2017). Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3147-3155). https://openaccess.thecvf.com/content_cvpr_2017/papers/Tai_Image_Super-Resolution_via_CVPR_2017_paper.pdf

[47] Tai, Y., Yang, J., Liu, X., & Xu, C. (2017). Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision* (pp. 4539-4547).
https://openaccess.thecvf.com/content_ICCV_2017/papers/Tai_MemNet_A_Persistent_ICCV_2017_paper.pdf

[48] Dahl, R., Norouzi, M., & Shlens, J. (2017). Pixel recursive super resolution. In *Proceedings of the IEEE international conference on computer vision* (pp. 5439-5448). https://openaccess.thecvf.com/content_ICCV_2017/papers/Dahl_Pixel_Recursive_Super_ICCV_2017_paper.pdf

[49] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Szegedy_Going_Deeper_With_2015_CVPR_paper.pdf

[50] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
https://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.pdf

[51] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141). https://openaccess.thecvf.com/content_cvpr_2018/papers/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.pdf

[52] Zhang, K., Zuo, W., Gu, S., & Zhang, L. (2017). Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3929-3938).
https://openaccess.thecvf.com/content_cvpr_2017/papers/Zhang_Learning_Deep_CNN_CVPR_2017_paper.pdf

[53] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).
https://openaccess.thecvf.com/content_cvpr_2017/papers/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.pdf

[54] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258). https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf

[55] Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., & Graves, A. (2016). Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/file/b1301141feffabac455e1f90a7de2054-Paper.pdf

[56] Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387-391. https://www.nature.com/articles/nature03390.pdf

[57] Cao, Q., Lin, L., Shi, Y., Liang, X., & Li, G. (2017). Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 690-698). https://openaccess.thecvf.com/content_cvpr_2017/papers/Cao_Attention-Aware_Face_Hallucination_CVPR_2017_paper.pdf

[58] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1904-1916. http://www.cs.utoronto.ca/~bonner/courses/2020s/csc2547/papers/discriminative/object-detection/spp-net,-he,-tpami-2015.pdf

[59] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890). https://openaccess.thecvf.com/content_cvpr_2017/papers/Zhao_Pyramid_Scene_Parsing_CVPR_2017_paper.pdf

[60] Mallat, S. (1999). A wavelet tour of signal processing. https://www.di.ens.fr/~mallat/papiers/WaveletTourChap1-6.pdf

[61] Vu, T., Van Nguyen, C., Pham, T. X., Luu, T. M., & Yoo, C. D. (2018). Fast and efficient image quality enhancement via desubpixel convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (pp. 0-0). https://openaccess.thecvf.com/content_ECCVW_2018/papers/11133/Vu_Fast_and_Efficient_Image_Quality_Enhancement_via_Desubpixel_Convolutional_Neural_ECCVW_2018_paper.pdf

[62] Kligvasser, I., Shaham, T. R., & Michaeli, T. (2018). xunit: Learning a spatial activation function for efficient image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2433-2442). https://openaccess.thecvf.com/content_cvpr_2018/papers/Kligvasser_xUnit_Learning_a_CVPR_2018_paper.pdf