Super-Resolution and Deblurring of Hyper-Hemispherical Images using Deep Neural Networks

M. Barbierato*, W. Erb*, E. Simioni†, C. Pernechele†
*Department of Mathematics, Università degli Studi di Padova
†INAF - National Institute for Astophysics, Padua Observatory

***Abstract*—For the exploration of lunar caves, we study the application of deep neural networks to the problem of super-resolution (SR) of images captured by hyper-hemispherical lenses, in order to enhance both their spatial resolution and perceptual quality. Our focus is on panoramic images generated by a lens with an ultra-wide field of view, which introduces spatially variant blur due to its intrinsic point spread function. We investigate two distinct architectures - convolutional networks and transformers - alongside with specific network designs tailored specifically for SR. To address the complexities of the degradations in the image generation process, we refine a pipeline for generating supervised training data that enables our models to learn and adapt to the spatially varying distortions inherent in panoramic images. The performance of our approach is evaluated using a no-reference image quality metric, demonstrating its usefulness in this setting.**

***Index Terms*—Super Resolution, Deblurring, Image Enhancement, Panoramic Images, Deep Learning.**

I. Introduction

The field of image enhancement is composed of various subtasks, including super-resolution (SR), deblurring, denoising, and contrast enhancement. Each of these attempts to address a specific type of image degradation: SR focuses on spatial low-resolution (LR), deblurring mitigates blur, denoising removes noise, and contrast enhancement improves poor contrast. All of these tasks are hard to tackle because of the ill-posed nature of inverting these defects, e.g. in SR there are many high-resolution (HR) counterparts that are theoretically possible for a single LR image, but only one of them is understood to be the 'real' image. Similarly, deblurring and denoising are ill-posed tasks. Moreover, in real-world application it is usually necessary to correct for multiple of these degradations at the same time, rendering the task even more complicated.

For SR, two tasks are usually distinguished: *classical* SR, where the aim is to improve the spatial resolution of images, and *real-world* SR, where also other degradations as blur and noise need to be corrected. If the nature of these degradations is not known, the task is referred to as *blind* SR. Classical SR algorithms are usually iterative and assume that multiple LR images of the same object are available [1], which is not always the case. When only a single LR image is available, one talks about *single image* SR. The problem of single image SR consists in, given a LR image I_{LR} of size $H \times W$, recovering the corresponding HR image I_{HR} of size $sH \times sW$, where the positive integer $s \geq 1$ is called the scale of the SR. We suppose that the LR has been obtained by the application of a degradation operator $\mathscr{D}$, understood as a complex combination of downsampling, blur, noise and other degradations, such that $I_{LR} = \mathscr{D}(I_{HR})$, and the objective is to learn an approximate inverse of $\mathscr{D}$.

In this work, we focus on blind SR of single images captured by the PANCAM [2], a hyper-hemispherical camera that generates panoramic images with an ultra-wide field of view (FOV). Thanks to this wide FOV, the PANCAM plays a key role in various projects [3], ranging from studies of Earth's satellite environment to lunar explorations. In 2021, the PANCAM was selected as imaging payload for the ESA-funded DAEDALUS (Descent and Exploration in Deep Autonomy of Lava Underground Structures) mission. Currently, the Italian National Space Agency (ASI) is funding the enhancement of the DAEDALUS CAM, to explore and increase its immersive imaging capabilities, with a particular focus on the software for visualizing the immersive images of lunar caves.

These images pose a particularly complex challenge for SR due to the spatially varying nature of their degradations and the particular geometric properties of the camera. Most of the SR approaches in the literature address spatially *invariant* degradations [14] and can not be applied directly to PANCAM images. Moreover, as for the PANCAM only LR images are available, it is not possible to evaluate the performance of SR models with conventional SR metrics. In fact it is infeasible to replicate the images captured by the PANCAM with other lenses due to its unique geometric proprieties.

To address these challenges, we present a degradation pipeline that is able to generate training images with spatially variant blur degradations by means of combining constant blur degradations, and we encode the variant nature of these through a distortion channel that neural networks are able to take advantage of; we then investigate the results of training our networks in this manner and analyze their performance with a no-reference image quality assessment metric, comparing different modalities of adversarial training.

II. Related Work

Since their introduction, deep neural networks (DNN) have managed to achieve state-of-the-art performance in almost all tasks of computer vision, including image enhancement. For single image classical SR, the first breakthrough was SRCNN [5], a deep convolutional network (CNN) which managed to outperform previous machine learning methods by leveraging a large training dataset; convolutional networks were then extensively employed for SR and improved on [6] [7] [8] until the introduction of the ViT [9] which leverages the transformer architecture and improved on the performance

Fig. 1: An image taken by PANCAM, in the original stereographic projection (left), and in equirectangular projection (right). The left image has a central Zoom objective that is not covered by the hemispherical lens' FOV, so it is not included in the equirectangular projection.

of CNNs for SR, at the need of higher training resources. Since then, models based on the transformer architecture and its variants have been representing the current state-of-the-art [10] [4] [11]. Another important contribution was the implementation of Generative Adversarial Networks for high-scale SR [12], which has subsequently been employed to enhance performance both at high SR scale factors and in real-world SR [10].

In classical SR, the availability of training data for DNNs poses no problems. Given a dataset of HR images it is sufficient to down-sample their spatial resolution to obtain a LR image that can be used as a training input. However, for blind SR, where the nature of the added degradations is unknown, collecting image data that is appropriate for the task at hand becomes more challenging, as one needs at least to make some hypothesis on the nature of the degradations. Since the availability of paired image datasets appropriate for blind SR is rare, an easier approach is to synthetically generate the training pairs starting from HR images and by applying a series of specifically modeled degradations. This has classically been done by applying blur, noise, and down-sampling the HR images to generate training pairs [13]. This degradation pipeline has been successively improved by varying the model for the distribution of the blur and by successively applying more complex degradations [14].

Classical SR of panoramic images is more challenging because of the particular geometry of the images and the limited availability of suitable dataset compared to standard images. It is still possible to apply conventional SR methods to panoramic images, and initial work achieved some success using DNNs in this manner [15] [16]. Successive attempts [17] improved on this task by generating pseudo-panoramic images from ordinary ones to obtain larger training dataset, and by introducing tailored layers and geometric features to account for the structure of these images [18].

III. PANCAM

The aim of this work is to SR and visually enhance images captured by a special hyper-hemispherical lens system referred to as PANCAM [2]. Developed by the INAF institute, the PANCAM camera integrates features of fisheye and omni-directional lenses to achieve a FOV exceeding 180° while avoiding the characteristic "donut" shape of very wide-angle optics. Its optical design consists of a catadioptric lens for panoramic imaging, a frontal field lens, and an objective lens for image formation. Unlike traditional fisheye lenses, PANCAM enables a Z-field (zenithal) angle of up to 135° by optimizing the entrance pupil orientation. The result of this process is the possibility to capture panoramic images with a FOV of 360° in the azimuth plane, and more than 100° in the zenith plane, with more than 40° below the horizontal plane in which the lens is positioned.

Classical panoramic imaging can be accomplished using three primary approaches: (i) moving cameras that capture and merge multiple images as generally performed by Mars rovers, (ii) multi-camera systems with independent lenses, and (iii) stationary omnidirectional cameras (the PANCAM case). The first two methods can be resource-intensive, requiring time, cost, and complex calibration. In contrast, omnidirectional cameras simplify the process by utilizing a single detector without moving parts, offering advantages such as low power consumption, lightweight design, and increased reliability.

As shown in Figure 2, the PANCAM requires three optical components: a panoramic catadiopter, a nested frontal optic, and a shared three-lens imaging objective. The lenses use a shared optical layout, consisting of fore-optics (catadiopter and speed-down lens) before the aperture stop (AS) and an objective lens (OBJ) to project the image onto the focal plane.

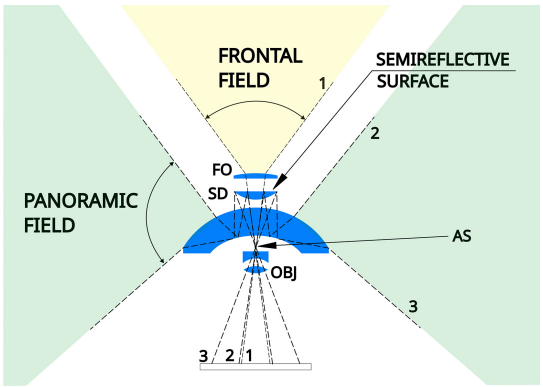

Fig. 2: Optical design of the PANCAM, in yellow the frontal field, in green the panoramic one. The resulting images are shown on the left hand side of Figure 1 while the right hand side shows the equirectangular projection limited to the panoramic field of view.

The panoramic field captures a 360°x270° FOV, using a catadiopter with a reflective concave surface and a semi-

reflective speed-down (SD) lens to decelerate light entering the OBJ. In the PANCAM, the frontal field (normally blind in common panoramic cameras) is covered by a higher focal length frontal lens (FO) to enhance the resolution in the central 20° circular view while retaining the panoramic FOV of 360°x100°.

Although panoramic cameras have lower spatial resolution, they are well-suited for applications in robotics, collision avoidance, and scene monitoring. As a result of the particular construction of the lens, PANCAM images result visually distorted especially in the areas at the limit of the FOV (below the horizon, the chief rays converge, reducing the lens's zenithal IFOV), and present blur degradations that are spatially variant as functions of the zenith angle. To facilitate the processing of these images, they are projected in an equirectangular plane according to the geometric model defined in [19], an example of which can be seen in Figure 1. The images are 1-channel monochrome.

The main scope of the PANCAM is to be used in a proposed robot named DAEDALUS [20] whose task is the exploration of lunar lava tubes. The PANCAM was, in fact, selected in 2021 for the ESA-funded DAEDALUS mission (Descent and Exploration in Deep Autonomy of Lava Underground Structures). Currently, the Italian National Space Agency (ASI) is supporting the advancement of the DAEDALUS CAM, aiming to enhance its immersive imaging capabilities, particularly by improving the software for visualizing lunar cave environments. These features are of great interest to the scientific community, as the Kaguya, LRO, and GRAIL missions have unequivocally confirmed the presence of deep voids beneath the lunar surface. These subsurface voids could serve as a foundation for long-term human presence in lunar exploration [21] because they: i) act as a natural shield against micrometeorite impacts and cosmic radiation, where radiation levels are only 30 of those on the surface; ii) provide a stable thermal environment; iii) may grant access to essential resources, including potential water sources. This discovery could revolutionize the approach to future lunar exploration and open new pathways for Martian exploration. Targets such as lunar caves have led to the selection of lenses like the PANCAM because of their ability to provide highly immersive images, thanks to their ultra-wide FOV. The main goal of the instrument will be the characterization of the pit of the lava tube from an imaging and photogram-metric point of view, by applying a consolidated stereo processing chain [22] to the PANCAM images. As a result, the main images that PANCAM should capture are of rock-like formations, which are composed of irregular and complex forms with little geometric structure.

IV. Problem Description - SR of PANCAM images

For the task of SR and deblurring, the degradations in the image acquisition process that are of interest to us are of three types, which we can describe as follows.

1) The imaging system has a point spread function (PSF) that, convolved with the object captured in the image, produces blur. Particularly when viewed in equirectangular projection, the blur spatially varies as a function of the zenith angle: as this angle increases to the limit of the lens FOV so does the 'spread' of the PSF and the resulting blur is therefore more pronounced. As a result, the regions captured at 'low' zenith angle are much more distorted than regions at 'higher' angles. An example of this behavior can be seen in Figure 3. Moreover, the spread of the PSF doesn't vary significantly as a function of the azimuth angle.
2) The process of projecting the images into an equirectangular plane is composed of both down-sampling and over-sampling operations, which we suppose to be unknown. Again, these operations vary, in particular, as a function of the zenith angle, though in a more complex manner compared to the PSF. Nevertheless, the images do not appear to exhibit a high level of aliasing, a required condition for effective SR [1].
3) Moreover, as with any imaging system, the images are subject to read-out noise at pixel level.

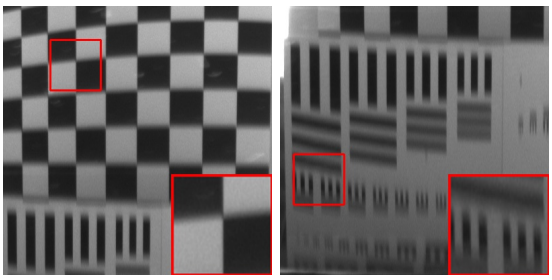

Fig. 3: Degradation difference as a function of zenith angle in images taken by PANCAM on a testboard. These image patches are taken with respect to the same azimuth angle, but the left patch is taken above the horizontal plane of the camera (high zenith angle), and the right patch is taken below it, approaching the FOV limit of the lens (higher angle). It can be seen that in the right patch the blur is significantly more pronounced.

Approaching the SR of PANCAM images presents two challenges: (1) the availability of an appropriate training dataset for images of rock formations, and (2) of a method of evaluating the performance of the SR of images without a HR counterpart. To address the first problem, we decided to train our neural network with a generalized dataset, composed of HR images representing a wide range of subjects. This has the immediate disadvantage of not providing the models with data tailored to the problem, and so performance is not expected to be optimal; nonetheless, models trained with generality can be used as a starting point to fine-tune models on more task-specific datasets when they become available, and doing so has been seen to result in increased performance compared to straightforward training on the specific datasets. For evaluating the performance of our models without a HR target, like previous works [14] we use a no-reference image quality assessment method, in our case the IL-NIQE metric [23], which attempts to quantify the perceptual quality of a single image by giving it a lower-is-better positive score. IL-NIQE consists in learning the parameters of certain statistical distributions from a dataset of pristine HR images, and then producing a score based on a distance defined between the

'pristine' distributions and the ones of other images.

Another specific challenge in PANCAM images is the spatially variant nature of the blur. Previous works generally deal with blur that is unknown in character but still uniform in spread, while our images have blur that is both unknown and variant. To treat this problem specifically, we modify the degradation pipeline in [14] to model a spatially variant blur PSF as a linear combination of two constant PSFs, see Section VI-A.

V. Model Description

In this section, we describe the architecture of SR networks we employed, one convolutional and the other transformer-based. As mentioned in the previous sections, the stereographic PANCAM images are projected into an equirectangular plane [19] to facilitate the processing. We indicate an equirectangular image I of size $H \times W$ as a matrix of values I_{ij}, where i and j represent the zenith and azimuth angle respectively.

A. Distortion Channel

As an attempt to encode the nature of the equirectangular images, the spatially variant nature of the degradations, and to better guide the models in their SR and deblurring, we follow the work of [17] and introduce a *distortion channel* to be appended to the PANCAM images as input to our models.

For our problem, we want this distortion channel to encode the variation of the zenith angle, as it is the main dependency of the blur degradations. To do this, given an equirectangular image I_{ij} of size $H \times W$, we define the distortion channel D_{ij} to have a value that changes linearly with the zenith angle:

$$D_{ij} = \frac{i}{H} \tag{1}$$

for $0 \leq i < H, 0 \leq j < W$, having the same size as I_{ij}. Note that the value of D_{ij} only varies as a function of i, which represents the zenith angle; this poses the problem that each row of D_{ij} contains the same value, and so there is much redundant information. However, having D_{ij} to be the same size of I_{ij} permits the use of known image data augmentation techniques not only on the training images themselves, but also on the distortion channels, to synthetically provide our models with more training pairs and prevent over-fitting. This channel is concatenated with the PANCAM images so that our networks input is 2-channels, and can be seen as an analogue of positional encoding techniques, as a way to preserve the spatial information of the zenith angle within the image.

B. Convolutional Model

The convolutional model we used is based on RRDBNet [24], which starts with an initial Pixel Unshuffle layer [25], and a narrow feature extraction module composed of a convolutional layer that produces a feature map with C channels. This feature map is passed through a deep feature extraction module composed of a number of 'Residual in Residual Dense' blocks (RRDB), which interleave convolutional layers and dense residual connections. We enhance this architecture by employing the parameter-free attention module introduced in [26] after each block: we verify that this attention module increases the performance of the architecture with basically no computational overhead. After the deep feature extraction, the SR image is reconstructed by upsampling the resulting feature map two times with bilinear interpolation and passing it respectively through two convolutional layers. Each convolutional layer has kernel size of $3x3$, and uses the leaky-RELU activation function. The residual connections are weighted by element-wise multiplication by a parameter α: in RRDBNet this parameter is set to the value 0.2, while we let it be a learnable parameter of our network.

C. Transformer Model

The transformer-based model we used is based on SwinIR [10], which has a Swin Transformer [27] backbone. A narrow feature extraction module composed of a single convolutional layer is first applied, similarly to the convolutional model. Then, we modify the original Swin Transformer Block architecture to introduce dense residual connection similar to the ones of RRDBNet; the feature maps goes through a number of 'Residual Swin Transformer Blocks', which are composed of a first group of Swin Transformer Blocks, after which we introduce a residual connection by concatenating feature maps on the channel dimension, paired with a single convolutional layer; the feature map goes through another group of Swin Transformer Blocks, after which there is another convolutional layer with a residual connection. After the deep feature extraction, the SR images are reconstructed via a convolutional and PixelShuffle [25] layers.

Introducing a residual connection by concatenating feature maps doubles the channel dimension, making the computations in the Transformer Blocks more expensive; to reduce this effect, the number of Swin Transformer Blocks that the feature maps passes through is halved compared to the original SwinIR architecture; overall, this does not increase computational costs and the introduction of more residual connections actually improves model performance.

D. Discriminator

Adversarial networks [28] are composed of two parallel networks, a *generator*, which corresponds in our case to the SR network whose task is to output images, and a *discriminator*, whose task is instead to learn to distinguish between 'genuine' images (i.e. the target HR images in our case) and those that are instead output of the generator. The latter learns to maximize its ability to correctly distinguish between these two classes, while the former learns, concurrently to its primary objective of SR, to also fool the discriminator into classifying its output as 'genuine'.

We design a simple discriminator architecture composed of convolutional layers with spectral normalization applied [29]. The input image goes through a first convolutional feature extraction layer, and then through four convolutional layer with kernel size $4x4$ and stride 2, which effectively halve each time the spatial size of the image. These layers are interleaved with residual connections, obtained by interpolating the previous feature maps by a scale of $1/2$. At the end of the convolutional

layers, we apply an average pool and linear layers to output a single number, representing the probability of the input image being genuine.

Previous works such as [14] use a U-Net based discriminator which outputs pixel-wise probabilities; these end up having no use because the Binary Cross Entropy adversarial loss is computed on their average, thus making the U-Net architecture needlessly complex as they are. Using an architecture that outputs pixel-wise probabilities would be useful if the training procedure takes them into account, for example by using data augmentation techniques on the discriminator inputs.

A visual representation of our architectures can be seen in Figure 4.

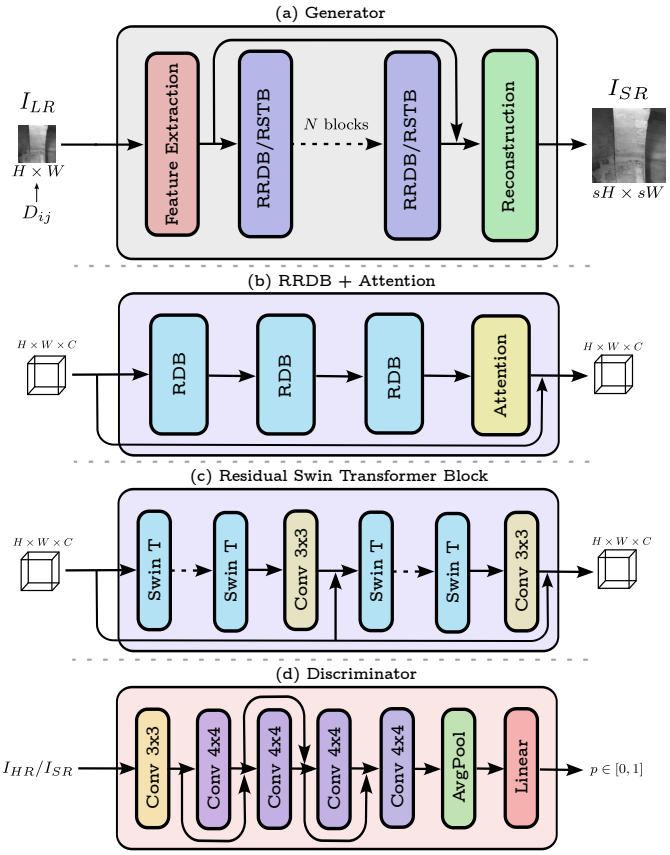

Fig. 4: (a) The general architecture of our generator, composed of an initial feature extraction module followed by N either convolutional (RRDB) or transformer (RSTB) blocks, and a final image reconstruction module. (b) The structure of a RRDBlock, based on [24], composed of three Residual Dense Blocks and a final parameter-free attention module [30] with a residual connection. (c) The structure of a RSTBlock, based on [10], composed of two subsequent series of Swin Transformers followed by a convolutional layer and a residual connection. (d) The structure we used for our discriminator, composed of 4x4 convolutional layers with stride 2 alternated by residual connections.

VI. Training Details

As explained in the introduction, we train our models starting from a dataset of HR images, and the LR images are generated through a *degradation pipeline* adapted from Real-ESRGAN [14] which we now describe.

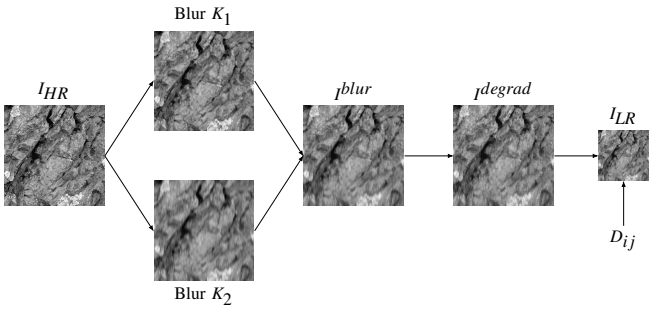

Fig. 5: Image degradation pipeline procedure. The HR image I_{HR} is blurred twice with increasing kernel variance and the two results are combined following (2) to obtain $I^{blurred}$. The image is then degraded with the addition of Gaussian noise to get I^{degrad} and then downsampled to obtain I_{LR}, to which the distortion channel (1) D_{ij} is appended. Zoom in for best view.

A. Degradation Pipeline

To generate training pairs composed of a LR image, input to the model, and a HR target image, we used the Real-ESRGAN degradation pipeline [14], which starts with a HR image and applies a series of degradations such as blur, noise, and downsampling to generate a LR degraded counterpart; we moreover modified this pipeline to adapt its capabilities to PANCAM images. In the following, we provide the details.

1) For a given HR image I_{HR} of size $H \times W$, first a number $\theta_1 \in [0,1]$ is chosen at random, which encodes the starting zenith angle $\Theta \in [0,\pi]$ that we want I_{LR} to represent in the following manner: $\Theta = \pi\theta_1$. Based on the size of I_{LR}, which is going to be equal to the size of I_{HR} divided by the scale, we compute $\theta_2 \in [0,1]$ which represents the final zenith angle of I_{LR}. More specifically, suppose that I_{LR} is going to be $n \times n$ pixels, which in PANCAM images of size 1800×3600 would represent a zenith angle of $\pi n/1800$: then we compute $\Theta_2 = \Theta_1 + \pi n/1800$ and $\theta_2 = \Theta_2/\pi$. Note that by doing this procedure in this manner has the possibility of resulting in Θ_2 being greater than π: to fix this, the initial θ_1 is actually chosen in the interval $[0, 1 - \frac{n}{1800}]$, for the above example.
2) Based on the angles $\theta_1, \theta_2 \in [0,1]$ above, we choose four variances σ_1^x, σ_2^x, σ_1^y, σ_2^y with the following rule:

$$\sigma_i^z = (1-\theta_i)\Sigma_1^z + \theta_i \Sigma_2^z$$

for $i = 1,2$, $z = x,y$, where $\Sigma_{1,2}^z$ are hyperparameters representing the variances of the PSF at a zenith angle of 0 (Σ_1^z) and π (Σ_2^z). Note that since $\theta_1 < \theta_2$, we have $\sigma_1^z < \sigma_2^z$.
3) Based on the variances above, the HR image is convolved with two bivariate Gaussian blur kernels K_1, K_2, respectively modeled as Gaussian distributions with mean zero and covariance matrices:

$$K_1 \sim \begin{pmatrix} \sigma_1^x & 0 \\ 0 & \sigma_1^y \end{pmatrix} \quad K_2 \sim \begin{pmatrix} \sigma_2^x & 0 \\ 0 & \sigma_2^y \end{pmatrix}$$

to produce two blurred images $I_1 = I_{HR} * K_1, I_2 = I_{HR} * K_2$, which are then joined together as a convex combination

dependent on the zenith angle:

$$(I^{blur})_{ij} = \left(1 - \frac{i}{H}\right)(I_1)_{ij} + \frac{i}{H}(I_2)_{ij} \tag{2}$$

where H indicates the height of the images I_1, I_2. This combination attempts to mimic a spatially variant blur that we have highlighted is specific to PANCAM images. We note that because of the linearity of the convolution operation, we can write (with a slight abuse of notation):

$$I^{blur} = I_{HR} * \left[\left(1 - \frac{i}{H}\right)K_1 + \frac{i}{H}K_2\right] = I_{HR} * K_i \tag{3}$$

meaning that the image I^{blur} is obtained by I_{HR} by convolving it with a kernel K_i which varies only as a linear function of the zenith angle i. We note that since K_1, K_2 are taken to be kernels modeled by a Gaussian PDF, a convex combination of them cannot be expressed as another Gaussian PDF and exhibits heavy-tails if the two variances of the K_1, K_2 are significantly different. However, by generating I^{blur} in this manner we are able to take advantage of the fast nature of convolutions by an invariant kernel while also managing to represent a spatially variant PSF.

4) To the resulting image I^{blur} Gaussian noise of constant variance is added; specifically, a matrix N of the same size of I^{blur} with independent entries $N_{ij} \sim N(0, \sigma^2)$ is generated and added:

$$I^{degrad} = I^{blur} + N$$

Finally, the resulting image is down-sampled by the chosen integer scale s of resolution with a method chosen between nearest, bilinear, and bicubic interpolation to obtain the LR image I_{LR}.

5) The Real-ESRGAN pipeline includes *higher-order* degradations, which consist in applying the above procedure of blur+noise a second time, to obtain more complex degradations that represent better the ones found in practical applications. However, we found that applying this process twice generates blur that are too complex for our purposes; we then decided to apply higher-order degradations only with a small probability $p = 0.05$ for each image batch.

6) Concurrently to the above pipeline, we generate the distortion channel described in equation (1) to be concatenated to the LR image. This needs to take into account the angle θ_1 chosen at the start and the final down-sampling of the image, and is thus:

$$D_{ij} = \frac{\theta_1 + i}{H/s} \tag{4}$$

for $0 \leq i < H/s, 0 \leq j < W/s$.

The image pair (I_{LR}, I_{HR}) is then used for training, where the LR image is the input to our model, which will output the super-resolved image I_{SR}, and I_{HR} is used as target. A graphical representation of the degradation pipeline is represented in Figure 5.

We use the Mean Absolute Error (also called L_1) loss to train the SR network:

$$L_1(I_{SR}, I_{HR}) = \sum_{i,j} |(I_{SR})_{ij} - (I_{HR})_{ij}| \tag{5}$$

which we find offers better reconstruction capabilities compared to the Mean Square Error. For adversarial training, both for the generator and the discriminator the Binary Cross Entropy loss is used.

The total loss $\mathscr{L}_{tot}$ of the generator is then

$$\mathscr{L}_{tot} = L_1 + \lambda_{GAN} \cdot \mathscr{L}_{GAN}$$

where λ_{GAN} is a hyperparameter controlling how much weight the adversarial loss has on the generator's learning. The choice of this hyperparameter is critical for the final performance of the generator: a higher adversarial loss weight makes it able to output images that are more realistic and higher quality, but also tend to contain more artifacts and amplification of noise. Since a greedy hyperparameter search is impossible given the size of our models and the absence of a consistent metric to evaluate their performance on PANCAM images, we opted to start by training a generator with $\lambda_{GAN} = 0$ and after a number of iterations we continuously increase this parameter until a certain value. This is advantageous as the generator is already pre-trained when the adversarial loss is introduced, so it is should be able to reach a better local optimum; moreover, if the discriminator is also pre-trained, the generator also doesn't have to deal with unimportant feedback from the discriminator at the start of its training.

We list here the hyperparameters chosen in the degradation pipeline. For the Kernel Variances, we chose $[\Sigma_1^x, \Sigma_1^x] = [0.05, 5.0]$ and $[\Sigma_1^x, \Sigma_1^x] = [0.5, 8]$. For Gaussian Noise, the variance is chosen uniformly in the interval $[.1, 3]$.

B. Training Parameters

We train our models using the Nomos-v2 dataset [31]. The images in this dataset are 3-channel RGB so for our purposes they are converted to 1-channel gray-scale, and the pixel values are normalized in the $[0, 1]$ interval range.

We train our models on a single Nvidia V100 GPU with a batch size of 12 using the Adam Optimizer algorithm with Nesterov momentum [32] for a total of 200k iterations, starting with a learning rate of 2e-4 and halving it at iterations $[100k, 150k, 175k]$. For the first 100k iterations, we only train the generator, with $\lambda_{GAN} = 0$; for the next 50k iterations, we also start training the discriminator, still with null loss weight. From 150k to 175k iterations, we linearly increase λ_{GAN} from 0 to 1e-3, and then to the last iteration we linearly increase it to 5e-3. The generator and the discriminator are trained alternatively for one mini-batch iteration each.

The training HR/LR pairs are augmented via rotation, mirroring, and the operations of Mix-Up [33], Cut-Mix [34], Resize-Mix [35], Cut-Blur [36]; for each batch iteration, one of these augmentations is chosen at via a weighted random choice and applied to it. These augmentations are employed in both the training LR image and the generated distortion channel (4) to provide the networks with a varied range of input images.

We report in Table I the complexity of our two architectures, together with the number of MAC operation on the forward pass to provide a measure of their complexity; note that our

Model	N blocks	Parameters	MACs
Convolutional	6	3,449k	16,95G
Transformer	4	2,191k	40,9G
Discriminator	-	64k	230M

TABLE I: Number of parameters and MACs operations for our two architectures. MACs are computed for an input image of pixel size 128×128.

transformer-based architecture has significantly less parameters than the convolutional one but is more computationally expensive.

Our implementation is based on NeoSR [31] and BasicSR [37].

VII. Results

In Figure 6, we report some image excerpts taken from Figure 1, comparing the outputs of the convolutional and Transformer-based architectures to the LR original images. In the SR images we see clear visual improvements, especially in correspondence to regular structures such as edges. The lower scores of the IL-NIQE metric compared to the LR images confirm the increase in perceptual quality. The convolutional architecture manages to achieve lower scores compared to the transformer one, possibly thanks to its higher number of parameters and thus ability to interpret complex structure and degradations. We observe that at higher zenith angles, our networks perform well particularly in debblurring and enhancing the geometric and regular structures captured by PANCAM images. However, at zenith angles close to the limit of the lens' FOV our networks struggle to SR effectively, as some examples show in Figure 7. Since degradations at these angles are much more pronounced this behavior is understandable, but it highlights the fact that our networks have not managed to learn the full extent of the degradations in PANCAM images, suggesting that better hyperparameter tuning for the degradation pipeline is necessary; moreover, it also puts into question the validity of the IL-NIQE metric (and in general, of no-reference quality metrics) for quality assessment of SR PANCAM images. We also note that when image features are particularly complex and irregular our networks struggle to SR as effectively as they do with regular image structures, and instead they tend to overproduce noisy artifacts, of which Figure 7 is an example. This has been pointed out previously in [12] and has been attributed to GAN training.

Adversarial Networks are in fact known to be unstable and hard to train due to the complex interaction and continuous feedback of the generator and the discriminator. In Figure 8 we compare the output of our convolutional model, and an equal one trained however without a discriminator. The adversarial architecture achieves visually sharper features and lower IL-NIQE scores, suggesting a better perceptual quality; however closer inspection reveals the presence of many artifacts especially in areas where features are vary complex, that are absent in the results from the non-adversarial network. In general there is a trade-off between perceptual clarity and presence of artifacts/noise amplification, which can be tuned by selecting appropriately the λ_{GAN} loss weight parameter. Moreover, while IL-NIQE is able to give an indication of quality improvement, it is not able to take into account these unpleasant artifacts.

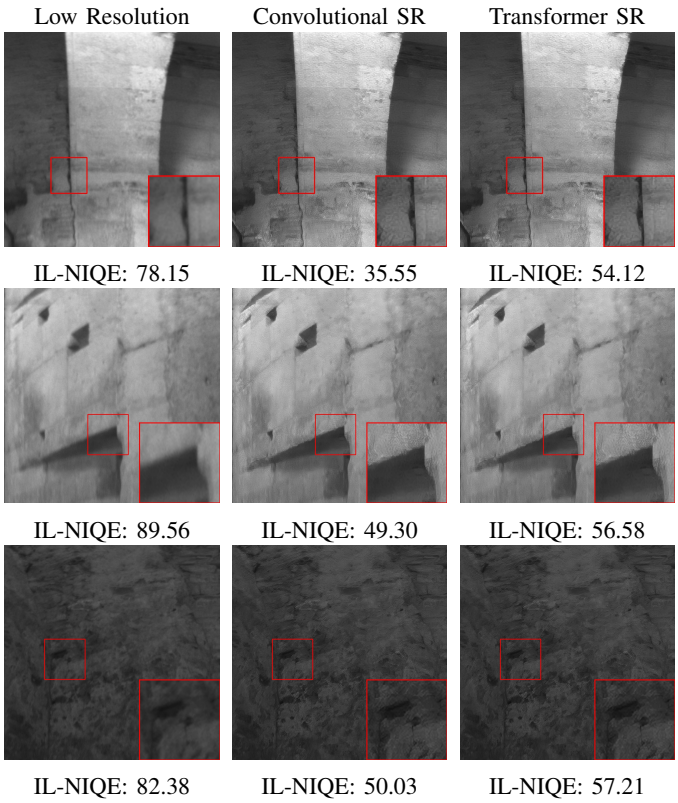

Fig. 6: Example PANCAM image details and the SR images with our two architectures. Below each image we report its IL-NIQE metric (lower is better). PANCAM LR Images have been upscaled with bicubic interpolation. In all cases we have a decrease in IL-NIQE scores, suggesting an improvement in image quality.

To gauge in a more interpretable manner the capabilities of our models we employed the Local Attribution Map (LAM) and the Diffusion Index (DI) score [38], which computes the Integrated Gradients [39] of the output image with respect to the blurred input baseline. The LAM is computed on a small patch of the output image, and can be interpreted as to give to different parts of the input image a measure of influence on the respective model outputs. The LAMs for two sample images for both of our models are shown in Figure 9 together with the relative DI scores; one can see that in both models the LAM follows the geometric structure of the image features, in particular in the first sample image, and so our models have learned to take basic shape and structures into account; when no particular features are present, the LAM is instead more spread out. The transformer model generally achieves higher DI scores, suggesting that it is capable of making use of more information on the input image than the convolutional model.

VIII. Conclusion

We have seen in this work an application of super-resolution with deep neural networks to a specific real-world imaging

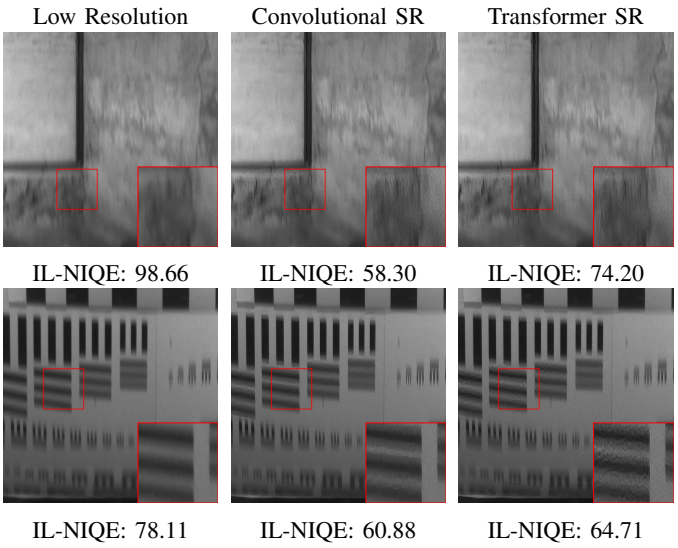

Fig. 7: Examples of SR failure of our models. The SR images have much lower IL-NIQE scores, suggesting a clear increase in perceptual clarity, but visually inspecting the images doesn't confirm this improvement, and instead we see that there is a tendency to amplify noise

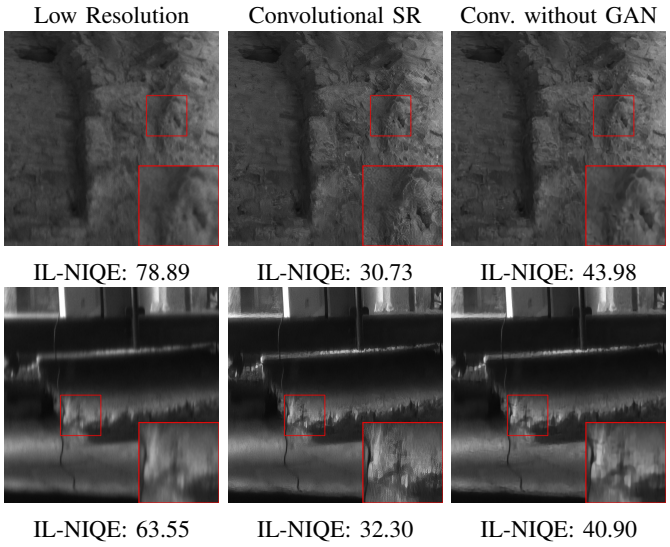

Fig. 8: Comparison of different SR architectures on a LR image that tends to produce artifacts. The convolutional architecture trained without a discriminator, while achieving higher IL-NIQE scores and less feature sharpness, does not present the same unpleasant artifacts of the architecture trained with GAN.

problem characterized by spatially variant and unknown degradations which presented a challenge for our approach.

We considered the two most common architectural backbones for SR networks, convolutional and transformer networks. We have seen that thanks to their efficiency, CNN manage to provide better performance for our task, but we have also confirmed thanks to the LAM visualization that transformers-based models should be able to take advantage of more information for the task of SR, as it was pointed out by the authors of LAM [38].

We have introduced a degradation pipeline that is able to

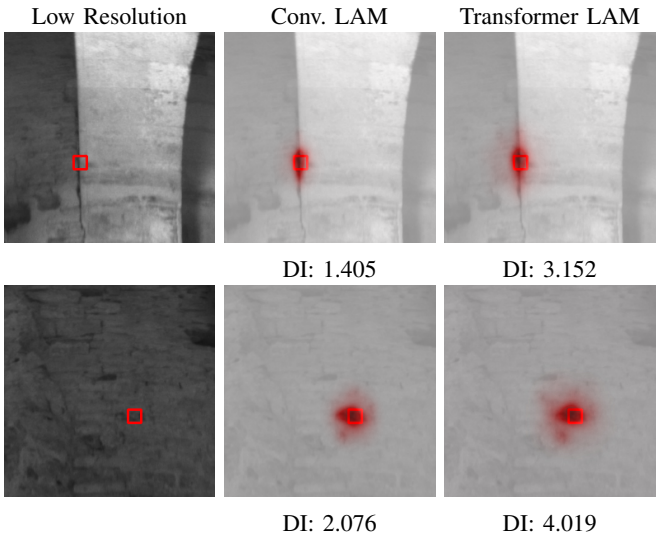

Fig. 9: LAM [38] comparison for our CNN and Transformer architectures, with relative DI scores (higher is better). The red square indicates the patch on which the LAM gradients are computed.

model a spatially variant PSF by a combination of two constant blur kernels, while also explaining the problems of such an approach, i.e. the non-standard behavior of the resulting blur. We note that this problem can be curbed by using a combination of a larger number of constant blur kernels, but this obviously comes with a higher computational cost.

We have briefly highlighted the shortcomings of training an adversarial SR network, namely the trade-off between perceptual quality and visual artifacts. For PANCAM images specifically, where the consistency of the SR might be non-negotiable for scientific use, adversarial training might be employed in a very conservative manner. Moreover, we have seen how a no reference image quality assessment metric such as IL-NIQE, while being correlated to visual perceptive quality of images, is not able to take into account artifacts, and thus is an unreliable metric for SR with neural networks. We note that also standard metrics such as the Peak-Signal-to-Noise-Ratio have been criticized for not being able to correlate well with human perception [40]. Moreover, given the presence of artifacts, SR with neural networks should be used with caution on images intended for scientific and research-critical uses, in particular with the employment of adversarial training.

Some further developments to build upon this work include: (1) Incorporating degradation pipeline kernels modeled after skewed distributions, such as the skew-normal distribution [41] (2) Employing other loss functions that focus on image fidelity and reconstruction quality, such as the focal frequency loss [42] (3) Making more effective use of the distortion channel (4) by including it in the pipeline of a tailored network to handle spatial variations, such as DeformConv [43] (4) eventually fine-tuning our models on task-specific datasets as the become available.

Acknowledgments

The second author is funded by the Università degli Studi di Padova - Dipartimento di Matematica under the project SID

BIRD 2023 entitled “ALISIA - ALgorithms for Immersive Stereoscopic Imaging with Appliciations to the Daedalus camera system”, and by the European Union – NextGenerationEU under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.1 - Call PRIN 2022 No. 104 of February 2, 2022 of Italian Ministry of University and Research; Project 2022FHCNY3 (subject area: PE - Physical Sciences and Engineering) “Computational mEthods for Medical Imaging (CEMI)”.

REFERENCES

[1] Jonathan D Simpkins and Robert L Stevenson. An introduction to super-resolution imaging. In *Mathematical Optics*, pages 555–580. CRC Press, 2018.

[2] Claudio Pernechele. Hyper hemispheric lens. *Optics express*, 24(5): 5014–5019, 2016.

[3] R. Opromolla, et al. Performances characterization of a non-conventional star tracker based on a hyper hemispherical panoramic camera *70th International Astronautical Federation Congress, Washington D.C*, 32 (20):IAC-18, #45987,2019

[4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015. URL https://arxiv.org/abs/1501.00092.

[6] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[7] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*, 2016.

[8] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 252–268, 2018.

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[11] Chih-Chung Hsu, Chia-Ming Lee, and Yi-Shiuan Chou. Drct: Saving image super-resolution away from information bottleneck. *arXiv preprint arXiv:2404.00722*, 2024.

[12] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, et al. Photo-realistic single image super-resolution using a generative adversarial network [c]. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 2017, pages 4681–4690, 2017.

[13] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997.

[14] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

[15] Yupeng Zhang, Hengzhi Zhang, Daojing Li, Liyan Liu, Hong Yi, Wei Wang, Hiroshi Suitoh, and Makoto Odamaki. Toward real-world panoramic image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 628–629, 2020.

[16] Qinglong Chang, Kwok-Wai Hung, and Jianmin Jiang. Deep learning based image super-resolution for nonlinear lens distortions. *Neurocomputing*, 275:969–982, 2018.

[17] Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong. Osrt: Omnidirectional image super-resolution with distortion-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13283–13292, 2023.

[18] Qi Jiang, Shaohua Gao, Yao Gao, Kailun Yang, Zhonghua Yi, Hao Shi, Lei Sun, and Kaiwei Wang. Minimalist and high-quality panoramic imaging with psf-aware transformers. *IEEE Transactions on Image Processing*, 2024.

[19] Emanuele Simioni, Claudio Pernechele, Wolfgang Erb, Andrea Marchini, Paolo Martini, Luigi Lessio, Luca Penasa, and Monica Beghini. A-central model for the geometric calibration of hyper-hemispherical lenses. *Opt. Express*, 32(20):34777–34795, Sep 2024. doi: 10.1364/OE.527318. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-32-20-34777.

[20] Dorit Borrmann, Andreas Nüchter, A Bredenbeck, Jasper Zevering, Fabian Arzberger, CA Reyes Mantilla, Angelo Pio Rossi, Francesco Maurelli, Vikram Unnithan, H Dreger, et al. Lunar caves exploration with the DAEDALUS spherical robot. In *52nd Lunar and Planetary Science Conference*, number 2548 in Lunar and Planetary Science Conference, page 2073, 2021.

[21] Haruyama, J., Morota, T., Kobayashi, S., Sawai, S., Lucey, P. G., Shirao, M., Nishino, M. N. (2012). Lunar holes and lava tubes as resources for lunar science and exploration. In Moon: prospective energy and material resources (pp. 139-163). Berlin, Heidelberg: Springer Berlin Heidelberg.

[22] Simioni, E., Re, C., Mudric, T., Cremonese, G., Tulyakov, S., Petrella, A., ... Thomas, N. (2021) 3DPD: A photogrammetric pipeline for a PUSH frame stereo cameras. Planetary and space science, 198, 105165.

[23] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[24] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[25] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 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*, pages 1874–1883, 2016.

[26] Cheng Wan, Hongyuan Yu, Zhiqi Li, Yihang Chen, Yajun Zou, Yuqing Liu, Xuanwu Yin, and Kunlong Zuo. Swift parameter-free attention network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6246–6256, 2024.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

[29] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[30] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*, pages 11863–11874. PMLR, 2021.

[31] NeoSR. NeoSR: Open-source framework for training super-resolution models. https://github.com/neosr-project/neosr, 2023.

[32] Timothy Dozat. Incorporating nesterov momentum into adam. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–4, 2016.

[33] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. URL https://arxiv.org/abs/1710.09412.

[34] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019. URL https://arxiv.org/abs/1905.04899.

[35] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels, 2020. URL https://arxiv.org/abs/2012.11101.

[36] Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy, 2020. URL https://arxiv.org/abs/2004.00448.

[37] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. BasicSR: Open source image and video restoration toolbox. https://github.com/XPixelGroup/BasicSR, 2022.

[38] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021.

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[40] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[41] Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

[42] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13919–13929, 2021.

[43] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.