Learning Images Across Scales Using Adversarial Training

KRZYSZTOF WOLSKI, Max-Planck-Institut für Informatik, Germany
ADARSH DJEACOUMAR, Max-Planck-Institut für Informatik, Germany
ALIREZA JAVANMARDI, Max-Planck-Institut für Informatik, Germany
HANS-PETER SEIDEL, Max-Planck-Institut für Informatik, Germany
CHRISTIAN THEOBALT, Max-Planck-Institut für Informatik, Germany
GUILLAUME CORDONNIER, Inria, Université Côte d'Azur, France
KAROL MYSZKOWSKI, Max-Planck-Institut für Informatik, Germany
GEORGE DRETTAKIS, Inria, Université Côte d'Azur, France
XINGANG PAN, Nanyang Technological University, Singapore
THOMAS LEIMKÜHLER, Max-Planck-Institut für Informatik, Germany

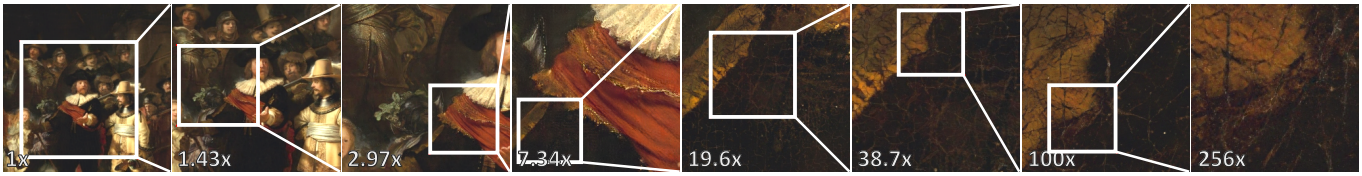

Fig. 1. Given an unregistered collection of image patches depicting an environment at vastly different scales, our approach uses adversarial training to obtain continuous and coherent scale spaces. Here, we showcase the reconstructed scale space of a painting, captured in its entirety, from the overall structure (1x) to the cracks in the oil paint (256x). Users can freely explore the scale space at interactive rates.

The real world exhibits rich structure and detail across many scales of observation. It is difficult, however, to capture and represent a broad spectrum of scales using ordinary images. We devise a novel paradigm for learning a representation that captures an orders-of-magnitude variety of scales from an unstructured collection of ordinary images. We treat this collection as a distribution of scale-space slices to be learned using adversarial training, and additionally enforce coherency across slices. Our approach relies on a multiscale generator with carefully injected procedural frequency content, which allows to interactively explore the emerging continuous scale space. Training across vastly different scales poses challenges regarding stability, which we tackle using a supervision scheme that involves careful sampling of scales. We show that our generator can be used as a multiscale generative model, and for reconstructions of scale spaces from unstructured patches. Significantly outperforming the state of the art, we demonstrate zoom-in factors of up to 256x at high quality and scale consistency.

CCS Concepts: • **Computing methodologies** → *Neural networks*; *Image processing*; *Image representations*.

Authors' addresses: Krzysztof Wolski, Max-Planck-Institut für Informatik, Saarbrücken, Germany, kwolski@mpi-inf.mpg.de; Adarsh Djeacoumar, Max-Planck-Institut für Informatik, Saarbrücken, Germany, adjeacou@mpi-inf.mpg.de; Alireza Javanmardi, Max-Planck-Institut für Informatik, Saarbrücken, Germany, alireza.javanmardi@dfki.de; Hans-Peter Seidel, Max-Planck-Institut für Informatik, Saarbrücken, Germany, hpseidel@mpi-sb.mpg.de; Christian Theobalt, Max-Planck-Institut für Informatik, Saarbrücken, Germany, theobalt@mpi-inf.mpg.de; Guillaume Cordonnier, Inria, Université Côte d'Azur, Sophia-Antipolis, France, guillaume.cordonnier@inria.fr; Karol Myszkowski, Max-Planck-Institut für Informatik, Saarbrücken, Germany, karol@mpi-inf.mpg.de; George Drettakis, Inria, Université Côte d'Azur, Sophia-Antipolis, France, george.drettakis@inria.fr; Xingang Pan, Nanyang Technological University, Singapore, Singapore, xingang.pan@ntu.edu.sg; Thomas Leimkühler, Max-Planck-Institut für Informatik, Saarbrücken, Germany, thomas.leimkuehler@mpi-inf.mpg.de.

Additional Key Words and Phrases: Image Synthesis, Reconstruction, Scale Space, Compression

ACM Reference Format:
Krzysztof Wolski, Adarsh Djeacoumar, Alireza Javanmardi, Hans-Peter Seidel, Christian Theobalt, Guillaume Cordonnier, Karol Myszkowski, George Drettakis, Xingang Pan, and Thomas Leimkühler. 2024. Learning Images Across Scales Using Adversarial Training. *ACM Trans. Graph.* 43, 4, Article 131 (July 2024), 13 pages. https://doi.org/10.1145/3658190

1 INTRODUCTION

The physical world exhibits a vast variety of scales, ranging from subatomic particles to galaxy clusters [Eames and Eames 1968]. A significant subset of these scales is within the reach of the human observer. A rich visual representation of the world, therefore, needs to account for as many scales as possible [Lindeberg 2013], while ordinary images can only capture a small slice of the scale spectrum due to two fundamental limitations: They have *finite extent* and *finite resolution* [Koenderink 1984].

Solutions to overcome these limitations can roughly be divided into three categories (Fig. 2a-c). *Level-of-detail* methods [Mallat 1989; Witkin 1987] construct a set of coarser-scale versions from a given full high-resolution image (Fig. 2a), but obtaining the entire original image becomes infeasible for large scale spans. In contrast, *super-resolution* infers finer scales from a coarse-scale image (Fig. 2b), hallucinating plausible higher-frequency content [Moser et al. 2023; Wang et al. 2020], but seems to reach an upper limit of upsampling in the order of 10x. Finally, methods that perform *structured aggregation* [Halladjian et al. 2019; Xiangli et al. 2022] combine multiple images into a multiscale representation (Fig. 2c), but require dense capture and careful registration of images across all scales.

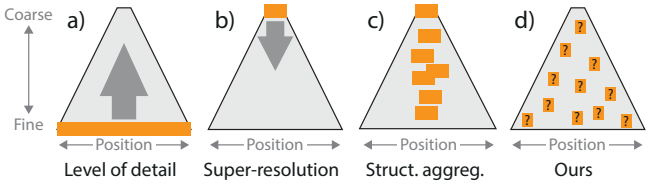

Fig. 2. Different paradigms to obtain a multiscale image representation. Orange blocks indicate the location of the input data in scale space (trapezoid). *a)* Level-of-detail methods require a full image at the finest scale and construct the scale space using low-pass filtering. *b)* Super-resolution infers slightly finer scales from a coarse-scale image. *c)* Approaches relying on structured aggregation assume registered images. *d)* Our approach relies on an *unstructured* collection of low-resolution input images: The locations of the images in scale space are unknown (question marks in the orange blocks) and do not even necessarily depict the same scene. We nevertheless produce full coherent scale spaces.

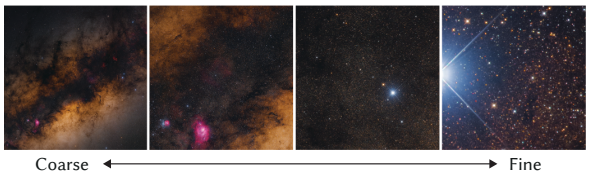

Fig. 3. Typical samples from a multiscale dataset. The images have a fairly low resolution (256 × 256 for us) and are unstructured, i.e., we do not have information about relative 2D location, allowing uncomplicated capture or collection without the need for registration. Images courtesy of Bartosz Wojczyński [2021].

We introduce a novel approach for the construction of a multiscale image representation from a set of *low-resolution, unstructured* images of a 2D environment (Fig. 2d). In particular, we consider images that have the same resolution but observe the environment at different scales, i.e., finer-scale observations cover smaller patches (Fig. 3). To alleviate the need for costly registration, *we do not require information on the location of the patches*, but only an approximate indication of scale. Such data corresponds, for example, to remote sensing applications whose objective is to capture a geographical point of interest at scales as various as the different acquisition tools e.g., satellites, airplanes, or drones flying at different altitudes. Note that the previous work discussed above cannot process this data. The absence of a single high-resolution image does not allow level-of-detail, the scale spans we consider are orders of magnitude above the capabilities of super-resolution, and missing positional cues prevent structured aggregation.

We observe that an unstructured collection of image patches across scales constitutes a data distribution that can be learned using adversarial training [Goodfellow et al. 2014]. Our key contribution is a neural architecture and training paradigm that treats multiscale image patches as slices of an underlying continuous scale space and enforces coherency across space and scale. This leads to two complementary training goals: *(i)* 2D slices of the generated scale-space(s) should match the distribution of the training data, and *(ii)* the generated scale space(s) should be coherent across all dimensions.

We build our representation upon an alias-free StyleGAN generator network [Karras et al. 2021] augmented with a set of progressive Fourier features distributed across multiple layers that generate tailored latent frequency content across the scale spectrum. Our training process is stable despite the wide range of scales and is specifically designed to enforce cross-scale consistency.

Once trained, we can interactively query our network with a position and scale level to obtain a corresponding generated scale-space slice. As shown in the companion video, it is possible to continuously zoom into *any* location of the sample or pan over the image at a chosen scale. Our generator synthesizes a single fixed-resolution image at a time but multiple adjacent samples can be seamlessly stitched together to yield coherent composites of up to several gigapixels at arbitrary scales. We demonstrate magnifications of up to 256x, i.e., one pixel in a 256x256 image can be enlarged to yield a full-resolution image with plausible high-frequency structure and details while being coherent across the continuous scale spectrum (Fig. 1).

We demonstrate two applications of our method: First, we show that we are able to aggregate the highly unstructured input into a coherent scale space, i.e., our approach produces a *pseudo-reconstruction* of the underlying scale space by learning a regularized distribution of the input patches. In doing so, our approach does not only implicitly register the patches, but also creates a compact representation, requiring 885x less parameters than the number of pixel values than an equivalent gigapixel images. Second, we show that we can learn a *generative model of scale spaces*, i.e., provided with image patches from different environments, our model allows to draw multiple, independent yet consistent scale-space samples.

We evaluate our method on several satellite datasets, a multiscale dataset consisting of an unstructured collection of images from the internet, as well as synthetic datasets created by extracting multiscale patches from gigapixel images. The latter provides us with ground-truth data for quantitative analysis. Our datasets, code, and trained models can be found at https://scalespacegan.mpi-inf.mpg.de.

In summary, our contributions are 1) a novel paradigm based on adversarial training to obtain a compact continuous multiscale image representation from unstructured ordinary images, 2) a generator architecture and training methodology for stable and scale-consistent scale space generation, 3) the application of our approach for multiscale unstructured image aggregation and as a multiscale generative model, 4) an interactive rendering application (approx. 20 fps) demonstrating the inference speed and data compression performance of our method.

2 RELATED WORK

2.1 Multiscale Representations

The representation of signals at multiple scales has a long history in mathematics, signal processing, as well as computer graphics and computer vision. Arguably the most concise description of phenomena at multiple scales is delivered by fractals [Mandelbrot 1982]. It provides a framework for modeling self-similarities and complexity across the (possibly infinite) scale spectrum, but is typically readily

applicable only to a narrow class of signals. Linear scale space theory [Iijima 1959; Lindeberg 2013; Witkin 1987] aims at representing images at multiple scales by embedding them into a one-parameter family of progressively smoothed versions. Scale-space methods are omnipresent in computer vision from the earliest methods [Marr and Hildreth 1980] till today [Lindeberg 2022]. Pyramids [Burt 1981; Williams 1983] are more compact multiscale representations, which perform scale discretization and spatial subsampling in addition to smoothing. Wavelets [Daubechies 1988; Mallat 1989] represent signals across scales by constructing a multiscale basis.

Within the torrent of deep learning in the last decade, many neural continuous multiscale representations have been developed. A broad range of visual-computing primitives has been considered, including images [Belhe et al. 2023; Chen et al. 2021a; Paz et al. 2022; Xu et al. 2021], geometry [Takikawa et al. 2021], materials [Kuznetsov et al. 2021], radiance fields [Barron et al. 2021; Xiangli et al. 2022], as well as general-purpose neural architectures [Fathony et al. 2020; Lindell et al. 2022; Saragadam et al. 2022; Shekarforoush et al. 2022].

All of the works listed above are designed to represent an original signal alongside its coarser-scale equivalents. Consequently, they require explicit *access to the entire signal at the finest scale*. This poses a severe problem when considering an orders-of-magnitude variety of scales, as, in this case, the finest scale contains details that are hard to synthesize or capture. In contrast, in this work, we develop a multiscale generative image model, which only requires an *unstructured* collection of fairly *low-resolution* images that capture *patches* of an environment at different scales. Similar to previous works [Barron et al. 2021], we employ progressive Fourier embeddings to steer our generator.

In the context of multiscale data visualization, exploration systems commonly blend between different representations best suited to convey information at a specific scale [Halladjian et al. 2019; Klashed et al. 2010; Mohammed et al. 2017; Tao et al. 2019]. Similar ideas have been utilized in stitching-based variable-resolution image creation [Licorish et al. 2021]. These approaches require registered images to be able to combine different sources of information, while our method relies on unstructured image collections. We, too, support interactive exploration of the scale space of our samples, due to a lightweight generator design.

2.2 Super-resolution

With origins in image restauration and deconvolution [Parker 2010; Wiener et al. 1949], single-image super-resolution methods aim at increasing the resolution of an image while synthesizing plausible higher-frequency content. Over the last years, feed-forward CNN-based approaches have established strong baselines for fixed-scale upsampling [Chen et al. 2021b; Liang et al. 2021; Lim et al. 2017; Lu et al. 2021; Shi et al. 2016; Wang et al. 2020; Yang et al. 2020; Zhang et al. 2018b,c]. Arbitrary-scale super-resolution methods take the upsampling factor as an additional input [Hu et al. 2019; Son and Lee 2021; Song et al. 2023; Vasconcelos et al. 2023; Wang et al. 2021a; Wei and Zhang 2023], allowing them to synthesize a range of scales. Generative models have been used as strong priors for super-resolution [Moser et al. 2023], with a particular focus on GANs [Chan et al. 2021; Menon et al. 2020; Wang et al. 2021b, 2018], and, recently, diffusion models [Gao et al. 2023; Kawar et al. 2022; Lin et al. 2023a; Wang et al. 2023c,b]. Typical upsampling factors for super-resolution methods are in the order of 10x – more than an order of magnitude less than what our approach can handle.

2.3 Scale-aware Generative Models and Infinite Images

The research on synthesizing infinite images and multi-scale images starts with textures. Early works on texture synthesis employed non-parametric methods to generate infinite [Efros and Freeman 2001; Efros and Leung 1999] and multi-scale [Han et al. 2008] textures. Subsequent advancements revealed that matching statistics of a pretrained CNN increases quality [Snelgrove 2017].

Since the invention of GANs [Goodfellow et al. 2014], there has been a remarkable surge in the quality of natural image synthesis, with the StyleGAN family being a representative example [Karras et al. 2020a, 2021, 2019, 2020b; Sauer et al. 2022]. The success of GANs has motivated researchers to explore the synthesis of very high-resolution or infinite images. A number of works have modified the GAN pipeline to produce a high-resolution image [Frühstück et al. 2019; Lin et al. 2022, 2023b; Rodriguez-Pardo and Garces 2022; Zhu and Kelly 2021]. Apart from GANs, high-resolution synthesis based on transformers [Esser et al. 2021; Liang et al. 2022] and diffusion models [Bond-Taylor and Willcocks 2024; Lee et al. 2023; Zhang et al. 2023] have been studied. However, all these approaches only consider images at a single scale but do not study how to learn and synthesize multi-scale images.

Even within a *single* natural image, content re-appears at multiple scales [Glasner et al. 2009; Zontak and Irani 2011]. This property has been exploited to perform image deblurring and super-resolution [Bell-Kligler et al. 2019; Michaeli and Irani 2014; Shocher et al. 2018], and to synthesize images with new layouts, structures, and sizes [Shaham et al. 2019; Shocher et al. 2019; Zhou et al. 2018]. This class of methods can only operate on a narrow range of scales, as self-similarities typically do not persist across orders-of-magnitude scale ranges. For example, when viewing a large painting from a distance, overall patterns and statistics are quite different from the individual paint strokes and cracks seen up close (Fig. 1). Our approach can also be used to reconstruct a single scale space, but assumes unstructured patches at multiple scales as input and therefore does not have to rely only on self-similarity.

The works most related to ours are AnyresGAN [Chai et al. 2022] and ScaleParty [Ntavelis et al. 2022], which can handle multi-scale images in both training and inference. However, these methods consider images of the same semantic level (e.g., images of human faces or animals) with a relatively narrow scale range, with the finest scale being only a 4-8x zoom of the coarsest scale. In this work, we consider a much broader scale range supporting zoom levels up to 256x, which involves the emergence of semantically new content (e.g., from an entire galaxy to individual stars) and thus introduces significant challenges. A concurrent work also studies such drastic multi-scale image synthesis based on diffusion models [Wang et al. 2023a]. However, it requires carefully crafted text prompts for each scale, which is prone to imprecise descriptions. In addition, this

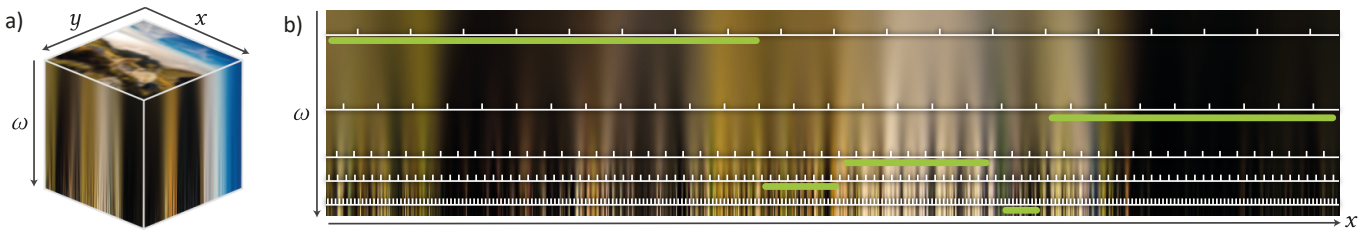

Fig. 4. (*a*) A scale space is a multiscale representation of an image. It is a continuous function of spatial coordinates $\mathbf{x} = (x, y)^T$ and bandwidth ω. Increasing ω introduces higher and higher frequencies. (*b*) An x-ω-slice through the volume in *a*). The resolution of spatial discretizations (white grids) needs to be adapted to a given ω to capture all frequency content. Input to our method is an unstructured collection of 2D image patches that sample the scale space (green bars). Each patch has a continuous location and scale. All patches have the same resolution ($N_p = 8$ in this visualization), which leads to different coverage of the spatial domain depending on their scale. Our method generates orders-of-magnitude scale spaces from this unstructured information. Notice that it is difficult to depict the actual resolution levels in a figure this size. We consider scale spaces, where the image at the top already requires a resolution of 256×256, leading to tens of thousands of pixels at the bottom.

method can only zoom into the center of the image, while our method creates full scale spaces.

3 MULTISCALE IMAGES

The central object of interest in this work is the continuous scale space $\bar{I}$ of an image [Iijima 1959; Witkin 1987], which captures versions of that image with different upper bounds on frequency content. We write

$$\bar{I}(\mathbf{x}, \omega) = \mathbf{c}, \tag{1}$$

where $\mathbf{c} \in \mathbb{R}^3$ is an RGB color, $\mathbf{x} = (x, y)^T \in [-0.5, 0.5]^2$ is a continuous location in the image plane, and $\omega \in \mathbb{R}_+$ is an upper frequency limit, also referred to as bandwidth. A low value of ω corresponds to an image with only little detail, while increasing ω progressively reveals structures of higher frequencies (Fig. 4a).

We are concerned with scale spaces in which ω spans a substantial interval $[\omega_{\min}, \omega_{\max}]$. Here, $\omega_{\min}$ is already high enough to represent an ordinary image (top face of the cube in Fig. 4a), and $\omega_{\max}$ is orders of magnitude larger than $\omega_{\min}$. To conveniently handle this large dynamic range, we introduce the notion of *scale* s, which we define in the logarithmic domain:

$$s = s(\omega) = \log_2 \left(\frac{\omega}{\omega_{\min}} \right) \in \mathbb{R}_{\geq 0}. \tag{2}$$

We further define $s_{\max} = s(\omega_{\max})$ to denote the full dynamic range of scales for a particular $\bar{I}$. In this work, a typical dynamic range is $s_{\max} = 8$, i.e., the most detailed image contains frequencies that are 256x higher than those of the coarsest image.

We consider both location $\mathbf{x}$ and scale s continuous parameters. However, in order to learn a scale-space representation from ordinary images, i.e., pixel arrays, as we set out to do in this work, we need to be able to handle discretizations in $\mathbf{x}$. We denote coordinate samples on a regular grid as $\mathbf{x}_i$. According to the Nyquist-Shannon sampling theorem [Antoniou 2006], a signal with bandwidth ω needs to be sampled at a rate of at least 2ω to capture all available detail and to avoid aliasing, referred to as the Nyquist limit. In our setting, consequently, the required spatial resolution increases with scale (white grids in Fig. 4b). Specifically, we need images with $N \times N$ pixels, where

$$N = N(s) = \lceil \sqrt{2} \omega_{\min} 2^{s+1} \rceil. \tag{3}$$

The factor $\sqrt{2}$ accounts for frequency content along the diagonal of the image plane, where the effective sampling rate is lower. Given the very high dynamic ranges $s_{\max}$ we consider, the required resolution at the finest scale $N_{\max} = N(s_{\max})$ quickly leads to gigapixel images. While, in theory, these ultra-high-resolution images constitute perfect data to create a scale-space representation $\bar{I}$, they are extremely difficult to produce. In practice, a large number of ordinary images is captured using a sophisticated camera setup and stitched together in a post-process [Brady et al. 2012; Cossairt et al. 2011; Kopf et al. 2007]. In the next section, we describe a novel, fundamentally different approach for learning a scale space based on adversarial training.

4 METHOD

We propose an algorithm to obtain an orders-of-magnitude scale space $\bar{I}$, which relies on an *unstructured* collection of ordinary, low-resolution images that constitute patch samples of $\bar{I}$. Each patch has a fixed resolution of $N_p \times N_p$ pixels, slices the scale-space volume at a continuous scale s_p, and is centered at a continuous location $\mathbf{x}_p$. The combination of fixed resolution N_p and varying scale s_p across patches leads to varying effective patch sizes in the spatial domain (green bars in Fig. 4b): A patch at a low s_p occupies a significant portion of the spatial domain, while a patch at a high s_p only covers a tiny fraction of it. Crucially, *we do not assume any knowledge about the 2D location* $\mathbf{x}_p$ *of each patch.* This significantly lifts the capture burden, as neither a specialized device nor a sophisticated acquisition protocol is required. In fact, we will demonstrate that our approach generates high-quality scale spaces even when applied to unstructured image collections from the internet that do not depict the same scene. We require, however, a coarse estimate of scale s_p per patch. We argue this does not impose unreasonable restrictions, as, for many application domains, s_p can be estimated from image metadata, e.g., focal length.

Instead of relying on the common approach of patch alignment and stitching, we train a deep generative model for obtaining $\bar{I}$ from the input data (Fig. 5). Specifically, we design a generator G that is able to learn a continuous orders-of-magnitude scale space $\bar{I}_G$. G takes as input a random vector $\mathbf{z}$, as well as a continuous patch location $\mathbf{x}_p$ and scale s_p, and renders one $N_p \times N_p$ image at a time.

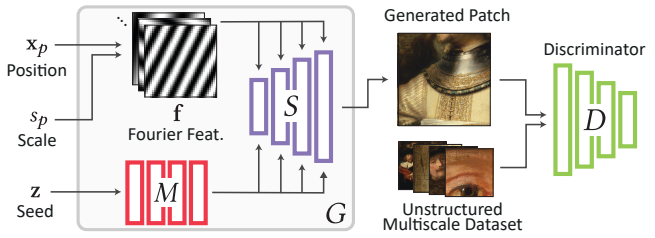

Fig. 5. Overview of our approach. Our multiscale generator G takes a patch location $\mathbf{x}_p$ and scale s_p, as well as a random seed $\mathbf{z}$ as input and synthesizes a corresponding image. A discriminator D compares the distributions of synthesized and data patches. Our generator architecture augments an alias-free StyleGAN with carefully designed Fourier features that are distributed across network layers, which allows to synthesize image patches from continuous orders-of-magnitude scale spaces. Dataset patches courtesy of Rijksmuseum [2019].

Continuously varying $\mathbf{x}_p$ and s_p produces corresponding slices of the learned $\bar{I}_G$ [Bora et al. 2018]. This allows interactive exploration of $\bar{I}_G$, while adjacent synthesized patches can be stitched together to yield seamless, arbitrary-resolution composites. We train G in an adversarial fashion [Goodfellow et al. 2014], i.e., we jointly train a discriminator D that compares the distribution of generated patches from $\bar{I}_G$ to the distribution of data patches from $\bar{I}$. During training, $\mathbf{z}$, $\mathbf{x}_p$ and s_p are randomly sampled. An additional consistency loss [Chai et al. 2022; Ntavelis et al. 2022] encourages coherency across scales.

Our statistical approach is surprisingly versatile and supports two modes of operation. First, given a collection of unstructured input patches from a *single* scale space $\bar{I}$, i.e., multiscale observations of the *same scene*, $\bar{I}_G$ converges to a plausible, coherent approximation of $\bar{I}$. While we find providing random inputs $\mathbf{z}$ to G is necessary for stable training, a converged generator G disregards $\mathbf{z}$, producing negligible variations of the output. We refer to this solution as *pseudo-reconstruction*, as $\bar{I}_G$ might differ from $\bar{I}$ in the arrangement of details, but captures the overall multiscale structure well. Note that in addition to missing knowledge about the patch locations $\mathbf{x}_p$, the patches do not exhaustively cover $\bar{I}$ at all scales. However, due to the generative capabilities of our framework, missing content will be seamlessly and consistently hallucinated across all scales. Along with implicitly performing an alignment of the input patches, G is a very compact representation: It has up to 885x less parameters than the corresponding gigapixel image at $s_{\max}$ requires RGB values stored in its pixel grid.

In the second mode of operation, we train G with a collection of input patches from *different* environments. The training patches depict different scenes from a (narrow) class at different scales. In this case, a converged G produces a *distribution of scale spaces* $p(\bar{I}_G)$, where different $\mathbf{z}$ result in different samples from that distribution.

To achieve our goal of learning images across scales using adversarial training, we require two essential ingredients: First, we design a generator G that can synthesize scale spaces $\bar{I}_G$ with large dynamic range $s_{\max}$. Second, we develop a training procedure that allows G to *robustly* learn *coherent* scale spaces. We give details on these ingredients in Sec. 4.1 and Sec. 4.2, respectively.

4.1 Multiscale Generator

We require a generator G that is able to encode orders-of-magnitude scale spaces $\bar{I}_G$. Importantly, while the output of G is a pixel grid, the model has to be intrinsically continuous to be a faithful representation of $\bar{I}$ and to allow for arbitrary translation and zooming. Fortunately, a continuous generator design is available in the form of alias-free StyleGAN (StyleGAN3) [Karras et al. 2021], which allows continuous translation of the generated content and has been shown to support some zooming [Chai et al. 2022]. However, we find that a vanilla StyleGAN3 generator architecture is not able to synthesize scale spaces for the high $s_{\max}$ we require. Therefore, we extend it to the multiscale setting.

The StyleGAN3 generator relies on Fourier features $\mathbf{f} \in \mathbb{R}^{d_\mathbf{f}}$, i.e., directional 2D sinusoids evaluated on a regular grid $\mathbf{x}_i$, i.e., they can be represented in the spatial domain as a pixel grid with $d_\mathbf{f}$ channels (Fig. 6a):

$$\mathbf{f}_j(\mathbf{x}_i) = \sin\left(2\pi\boldsymbol{\omega}_j^T\mathbf{x}_i\right), \tag{4}$$

where $\boldsymbol{\omega}_j \in \mathbb{R}^2$ are $d_\mathbf{f}$ different frequency vectors. The features $\mathbf{f}$ are fed into a sequence of layers of a synthesis network S, each of which performs non-linear operations. Occasionally, intermediate neural features are up-sampled to a higher spatial resolution. The non-linear operations are modulated by "style" vectors, which arise from feeding the latent code $\mathbf{z}$ through a mapping network M (Fig. 5). Both, non-linear operations and up-sampling, are carefully designed such that the resulting neural features only contain spatial frequencies below the Nyquist limit dictated by the resolution of the respective layer. A direct consequence of this approach is that the entire generator can be treated as a continuous function, despite relying on regular grids for the actual computations. We choose a variant of the generator, where the neural operations are applied point-wise (R-configuration) [Karras et al. 2021]. In addition to obtaining rotational equivariance, this configuration is well suited for multiscale generation, as the alternative – spatial convolutions – typically operates on fixed-size neighborhoods, whose meaning varies with scale. We synthesize images of resolution $N_p = 256$.

By design, procedurally shifting Fourier features $\mathbf{f}$ at the beginning of the processing sequence results in alias-free translation of the output image (Fig. 6b). To incorporate scaling, in the first step, we transform all grid coordinates $\mathbf{x}_i$ via

$$g(\mathbf{x}_i; \mathbf{x}_p, s_p) = 2^{-s_p}\left(\mathbf{x}_i - \mathbf{x}_p\right), \tag{5}$$

and feed the corresponding shifted and scaled Fourier features $\mathbf{f}_j\left(g(\mathbf{x}_i; \mathbf{x}_p, s_p)\right)$ into S. Unfortunately, choosing a high s_p stretches $\mathbf{f}$ to such an extent that it degrades to an almost constant function (Fig. 6c). Unsurprisingly, we find that S cannot learn to synthesize meaningful images given such an input. Simply increasing frequencies $\|\boldsymbol{\omega}_j\|$ is not a solution to the problem, as we need to stay below the Nyquist limit of the first layer. Therefore, in a second design iteration, we could indeed sample higher frequencies $\boldsymbol{\omega}_j$, but progressively blend them in only after g has stretched out the corresponding $\mathbf{f}_j$ far enough to stay below the Nyquist limit, using some blending function w (Fig. 6d). While this strategy provides S with meaningful frequency content across all scales, we observe drifting and tearing in the output images during zooming. This is because we are interfering with the positional encoding provided

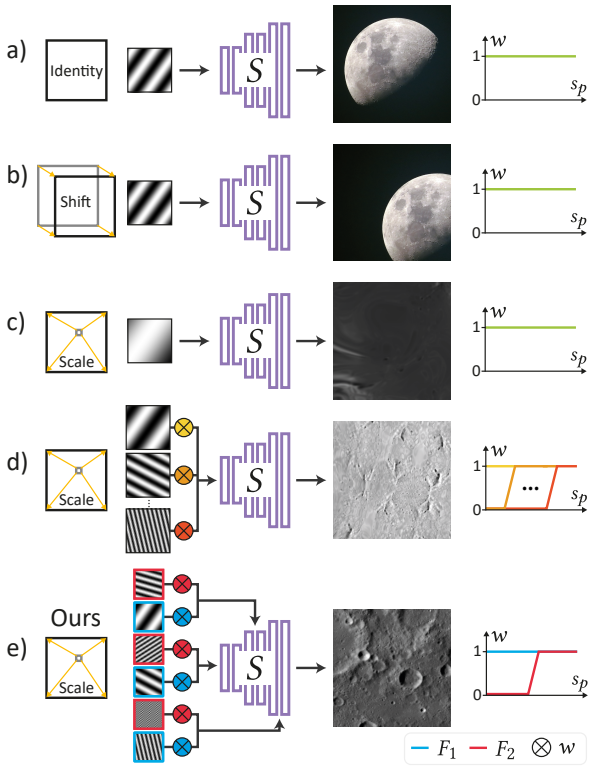

Fig. 6. *(a)* The StyleGAN3 generator rasterizes Fourier features $\mathbf{f}$ (one is shown) and feeds them through a synthesis network S to obtain an output image. *(b)* A spatial offset of $\mathbf{f}$ results in a shifted image. *(c)* Scaling up $\mathbf{f}$ leads to flat feature maps, which S cannot translate into a meaningful image. In the setups *a)-c)*, the features $\mathbf{f}$ are not modulated (constant weighting w). *(d)* Progressively blending in different $\mathbf{f}$ using the weighting function $w(s_p)$ results in a permanent re-scaling of individual features $\mathbf{f}$ (three out of many are shown), leading to unstable results. *(e)* We create Fourier features in bins (two bins – pink and blue – are shown) and blend in all features per bin at the same time. This leads to a significant reduction of blending (here, only the pink bin needs blending). Additionally, we inject features into different layers of S, significantly enhancing coherency across scales.

by the Fourier features in Eq. 4: The effect of scaling $\mathbf{f}_j$ is not distinguishable from the effect of shifting the input position $\mathbf{x}$. This ambiguity is exacerbated by the severe non-linearity of S.

We address this problem in our final design (Fig. 6e), which is based on two crucial observations. First, careful binning of Fourier features and simultaneous blending per bin significantly reduces the amount of re-weighting necessary. Second, the less non-linear layers are operating in between Fourier features and the final image, the less positional distortions they can cause. Consequently, we employ a re-assignment of binned Fourier features to different layers of S.

For frequency binning, we consider non-overlapping scale intervals of size Δs, and create $N = \lceil s_{\max}/\Delta s \rceil$ corresponding frequency bins F_k, where $k \in \{0 \dots N-1\}$. For each F_k, we randomly sample 512 frequencies with a maximum magnitude of $2^{s_{\text{base}}+\Delta s \cdot k}$, where s_{base} is a hyper-parameter. Fig. 7 shows a frequency distribution for two bins. We set $\Delta s = 3$ and $s_{\text{base}} = 6$ in all our experiments. We

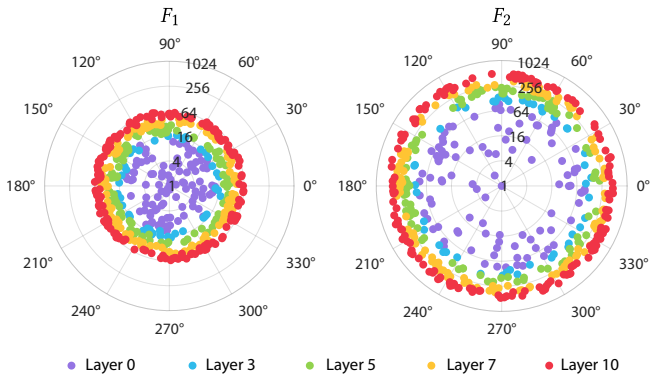

Fig. 7. Distribution of Fourier features $\mathbf{f}_j$ across bins and generator layers. Each point represents a 2D frequency $\boldsymbol{\omega}_j$ (magnitude is visualized in log space). Frequencies for different bins F_1 and F_2 are displayed separately. Colors signify the layer of the synthesis network into which a frequency is injected – the higher the frequency, the later the injection.

now define a weighting function that blends in all Fourier features per bin F_k as a function of scale s_p:

$$w_k(s_p) = \min\left(1, \max\left(0, s_p - \Delta s \cdot k + 1\right)\right). \tag{6}$$

This weighting scheme blends in entire bins of Fourier features simultaneously across regularly spaced, narrow intervals, which significantly reduces the amount of blending happening during zooming.

We inject the so-obtained weighted Fourier features into different layers of S [Diolatzis et al. 2023]. Specifically, in addition to injection into the first layer, we concatenate Fourier features $\mathbf{f}_j$ to neural features after each up-sampling layer. The assignment of $\mathbf{f}_j$ to the individual layers is based on per-layer Nyquist limits. We iterate over all injection layers in order and assign to the current layer those $\mathbf{f}_j$ that have not yet been assigned, if the following condition is met: At the scale s_p where the feature is blended in completely via Eq. 6, the scaling in Eq. 5 lets $\mathbf{f}_j$ fall below the layer's Nyquist limit (color coding in Fig. 7).

Our generator is now able to synthesize *detailed* and *coherent* content across orders-of-magnitude scale ranges. Let's train it!

4.2 Training

Our multiscale generator is trained in an adversarial fashion [Goodfellow et al. 2014] using an image discriminator D that compares the distribution of generated patches to the distribution of training patches (Fig. 5). In addition to sampling the random vector $\mathbf{z}$, we also randomly sample the patch location $\mathbf{x}_p$ and scale s_p. Regarding training setup and hyper-parameters, we largely follow the official StyleGAN3 [Karras et al. 2021] implementation and corresponding recommendations of the authors, involving R1 regularization [Mescheder et al. 2018] and adaptive discriminator augmentation [Karras et al. 2020a]. However, we find that our multiscale setting poses two significant challenges: *(i)* training stability, and *(ii)* consistency of the learned $\bar{I}_G$ across scales. We address these items using a progressive patch sampling scheme and a scale consistency loss, respectively.

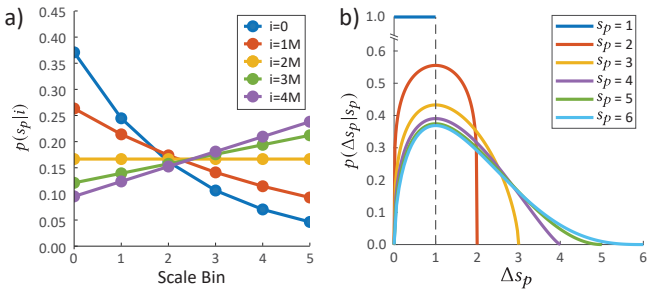

Fig. 8. (*a*) Sampling distribution of scale bins during training, where i denotes the number of images used so far. (*b*) Sampling distributions of scale offset Δs_p for different patch scales s_p.

4.2.1 Progressive Patch Sampling. We observe that simple uniform random sampling of patches across all scales does not converge to satisfactory results [Chai et al. 2022]. We therefore use a patch sampling scheme that prioritizes coarse scales at the beginning of training and progressively shifts attention towards the finer scales (Fig. 8a). This happens both for generated and for data patches; recall that we assume access to a coarse estimate of scale s_p per data patch. We operate on scale bins of width one to avoid relying on exact scale labels in the data. Specifically, we transition from a negative exponential distribution (blue curve in Fig. 8a), over a uniform distribution (yellow curve), to a linearly increasing distribution (pink curve). We find that this progressive strategy leads to significantly improved training stability.

4.2.2 Scale Consistency Loss. To encourage scale-coherent $\bar{I}_G$, we follow ideas from Chai et al. [2022] and add an additional loss term that compares two generated patches, separated by a scale offset $\Delta s_p \in \mathbb{R}_+$ [Irani and Peleg 1991]:

$$\mathcal{L}_s = \mathbb{E}_{\mathbf{z},\mathbf{x}_p,s_p,\Delta s_p}\left[d\left(R_{\Delta s_p}\left(G(\mathbf{z},\mathbf{x}_p,s_p)\right), G(\mathbf{z},\mathbf{x}_p,s_p-\Delta s_p)\right)\right]. \tag{7}$$

Here, $R_{\Delta s_p}$ is a function that downsamples an image by a factor $2^{\Delta s_p}$. d is an image distance metric that we apply to the two generated patches, only considering pixels that appear in *both* patches. We implement d using a linear combination of ℓ_1-norm and LPIPS [Zhang et al. 2018a] distance. While we could backpropagate gradients through both generator instances in Eq. 7, we find that training stability improves when we randomly choose only one of the generator instances to receive gradients in each training iteration.

While we uniformly sample $\mathbf{x}_p$ in Eq. 7, we find that the choice of sampling distribution for Δs_p has a strong influence on the scale consistency of $\bar{I}_G$ for our large $s_{\max}$. To make sure $\bar{I}_G$ is globally consistent, we want Δs_p to be sampled in the full range $[0, s_p]$, i.e., every patch is compared against arbitrarily zoomed-out counterparts up to $s_p = 0$. However, a uniform distribution over $[0, s_p]$ is not an optimal choice. On the one hand, sampling a very low value for Δs_p results in almost identical patches, wasting training resources. On the other hand, sampling a very high value results in images of significantly different relative resolutions, i.e., $R_{\Delta s_p}$ produces very low-resolution patches to be compared against, providing not much of a supervision signal either.

Our solution relies on the beta distribution B, which accounts for all the above considerations:

$$p(\Delta s_p|s_p) = \frac{B(\frac{\Delta s_p}{s_p};\alpha,\beta)}{s_p}, \tag{8}$$

where $\alpha = \sqrt[4]{\max(1,s_p)}$, and $\beta = (\alpha-1)\max(1,s_p) - \alpha + 2$. The parameters are chosen such that $p(\Delta s_p|s_p)$ has its mode at $\Delta s_p = 1$ and gradually falls off in both directions, while still covering the entire available scale interval (Fig. 8b).

5 EVALUATION

We evaluate our approach on the tasks of multiscale pseudo-reconstruction (Sec. 5.1) and generation (Sec. 5.2). We further analyze the components and properties of our method (Sec. 5.3). We urge the reader to watch our supplemental video, in which we demonstrate continuous zooming and panning through our obtained scale spaces. Our generator runs at 20 fps, which allows highly interactive exploration of our scale spaces.

Datasets and Training Details. We use a total of seven datasets for our evaluation, all containing unstructured patches at a resolution of $N_p = 256$. For Himalayas and Spain, we consider multiscale satellite data [Copernicus 2024]. Both datasets cover a square geographic region with $s_{\max} = 8$. Scale labels are obtained from satellite metadata. To enable a broader range of quantitative evaluations, we additionally employ three gigapixel images – Milkyway [Wojczynski 2021] ($s_{\max} = 6$), Moon [Speyerer et al. 2011] ($s_{\max} = 6$) and Rembrandt [Rijksmuseum 2019] ($s_{\max} = 8$) –, from which we extract random patches at multiple scales to simulate our input setting.

To evaluate generative capabilities, we only consider patches sampled from the finest four scales of each dataset source, denoted HimalayasGen, SpainGen, MilkywayGen, MoonGen, and RembrandtGen, respectively. The so-obtained data forces our models to learn scale-space distributions. Further, the Sunflowers and Bricks datasets are composed of a collection of images from Flickr. For Sunflowers, images were queried using the text strings "sunflower field" and "sunflower", while for Bricks the search strings were "brick wall" and "bricks and cracks". Obtained images are randomly cropped and down-sampled to our target resolution to achieve $s_{\max} = 4$. Coarse scale labels are assigned semi-automatically, taking into account the specific query, image resolution, and crop window size. Naturally, these datasets contains a variety of different scenes.

Datasets for pseudo-reconstruction contain 96k (milkyway and moon) or 156k patches (himalayas, spain, rembrandt), while those for generation contain 120k images, except for Sunflowers (185k patches) and Bricks (234k patches). The supplemental document lists more detailed dataset statistics.

We obtain converged models after 52-75 hours of training using eight A100 GPUs. Our models occupy 38-62MB of disk space. During inference, they require 2.8-3.1GB of VRAM.

Evaluating Scale Consistency. One crucial property of a scale space is its consistency across scales. We employ two procedures to quantify this. First, we create sequences of images, gradually zooming in. Using off-the-shelf optical flow estimation [Teed and Deng 2020],

Table 1. Quantitative evaluation of multiscale pseudo-reconstruction.

Dataset	s_{max}	FID ↓	Scale Consistency				$PSNR_{GT}$↑
			Bias ↓	Angle ↓	EMD ↓	$PSNR_{inter}$ ↑	
Himalayas	8	19.1	0.06	1.2	1.08	19.5	-
Spain	8	8.4	0.07	1.0	0.98	23.3	-
Milkyway	6	8.6	0.21	1.5	1.83	25.9	24.0
Moon	6	9.0	0.38	1.1	1.28	28.9	25.8
Rembrandt	8	14.6	0.20	1.2	1.91	28.9	-

Table 2. Quantitative evaluation of multiscale generation.

Dataset	Method	FID ↓	Scale Consistency		
			Bias ↓	Angle ↓	EMD ↓
MilkywayGen	AnyresGAN	30.6	0.81	11.3	12.22
	PULSE	-	0.43	6.9	8.69
	Ours	28.3	0.11	1.2	1.55
MoonGen	AnyresGAN	18.4	0.95	11.8	12.8
	PULSE	-	0.41	19.7	17.75
	Ours	6.3	0.23	2.2	2.73
HimalayasGen	**Ours**	19.4	0.11	1.1	1.28
SpainGen	**Ours**	9.7	0.05	1.1	1.17
RembrandtGen	**Ours**	18.9	0.36	3.7	7.37
Sunflowers	**Ours**	9.8	0.09	1.1	1.24
Bricks	**Ours**	6.8	0.08	1.03	1.16

we compute per-pixel motion trajectories (Fig. 12). As a first metric (inspired by tOF in [Chu et al. 2020]), we average all flow vectors to obtain an estimate of overall bias; a perfect solution has radial trajectories only (Fig. 12, right) and, thus, zero bias. For our second metric, we fit a line to each trajectory and compute the angular difference to the ground-truth trajectory [Çoğalan et al. 2023]. As a third metric, we compute the earth mover's distance (EMD) [Rubner et al. 1998] between each motion trajectory and the ground truth.

In a second procedure, we generate full-resolution scale-space slices at all integer scales. Notice that this involves spatial resolutions from 256×256 for $s = 0$ up to 65k×65k for $s = 8$. We now compute the PSNR between all slice pairs ($PSNR_{inter}$), where higher-resolution images are downsampled to match the lower-resolution ones. We also compute the PSNR with respect to ground truth when it is available ($PSNR_{GT}$). As reconstructed scale spaces do not exactly align with the references, we perform a global alignment using translation and isotropic scaling.

5.1 Multiscale Pseudo-Reconstruction

We show scale-space pseudo-reconstructions in Fig. 10, top and a corresponding quantitative evaluation in Tab. 1. More results can be found in the supplemental video. We observe that we are able to successfully learn orders-of-magnitude scale spaces, which allow coherent zooming into any location. The last column in Tab. 1 reveals that our reconstructions are quite close to the ground truth on average. We investigate this further in Fig. 9a, where we break down

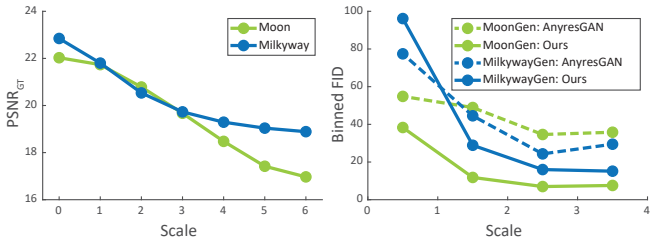

Fig. 9. (*a*) Reconstruction accuracy of our approach as a function of scale. (*b*) FID scores as a function of scale in the generative setting.

reconstruction accuracy per scale. We see that accuracy decreases for higher scales. This is because our model has the freedom to hallucinate high-frequency content as long as the overall structure is coherent i.e., our approach is a form of generative compression [Santurkar et al. 2018]. We study this behavior in an additional experiment, where we intentionally filter out training patches that overlap a certain spatial region. In Fig. 13, we show that this exclusion leads to inaccurate yet highly plausible content. In the supplemental document, we show a best-effort result of stitching a subset of our training patches using Adobe Photoshop.

5.2 Multiscale Generation

Results on multiscale generation are shown in Fig. 10, bottom and Tab. 2. We compare against two baselines, AnyresGAN [Chai et al. 2022] and PULSE [Menon et al. 2020] on two datasets. Details on how we modify these baselines to be able to handle our setting are provided in the supplemental document. We observe that our scale spaces are of significantly higher quality than those of the baselines, both in terms of patch distributions measured using FID [Heusel et al. 2017] and scale consistency. In Fig. 9b, we break down FID scores into scale bins, revealing that slices of our scale-space samples are well-behaved across scales. Notice that FID scores for coarse scales are less reliable due to less available data. In Fig. 11 we demonstrate qualitative results, while Fig. 12 shows flow trajectories of representative samples across methods, confirming that our approach delivers highly scale-consistent results.

5.3 Analysis

Compression. Our model requires 14M parameters, which is more compact than the models used in AnyresGAN (32M parameters) and PULSE (18M parameters). In contrast, an RGB gigapixel image at a corresponding resolution of 65k × 65k pixels requires 13B scalars to be stored. Thus, in terms of raw parameter reduction, our approach achieves a compression of 885x. Lossless or lossy compression can be applied on top of both approaches, e.g., JPEG for images and model weight compression for StyleGAN [Belousov 2021]. To shed some light on practical compression capabilities, we JPEG-compress the raw Milkyway gigapixel image to obtain a file size equal to our *uncompressed* model (JPEG quality: 32) and measure image quality at the finest scale. As our model only performs a pseudo-reconstruction in which details do not align with the reference (Fig. 13), pixel-wise PSNR is not an expressive metric for this task. We instead opt for patch-based FID, which yields a score of 44 for our model and 114 for

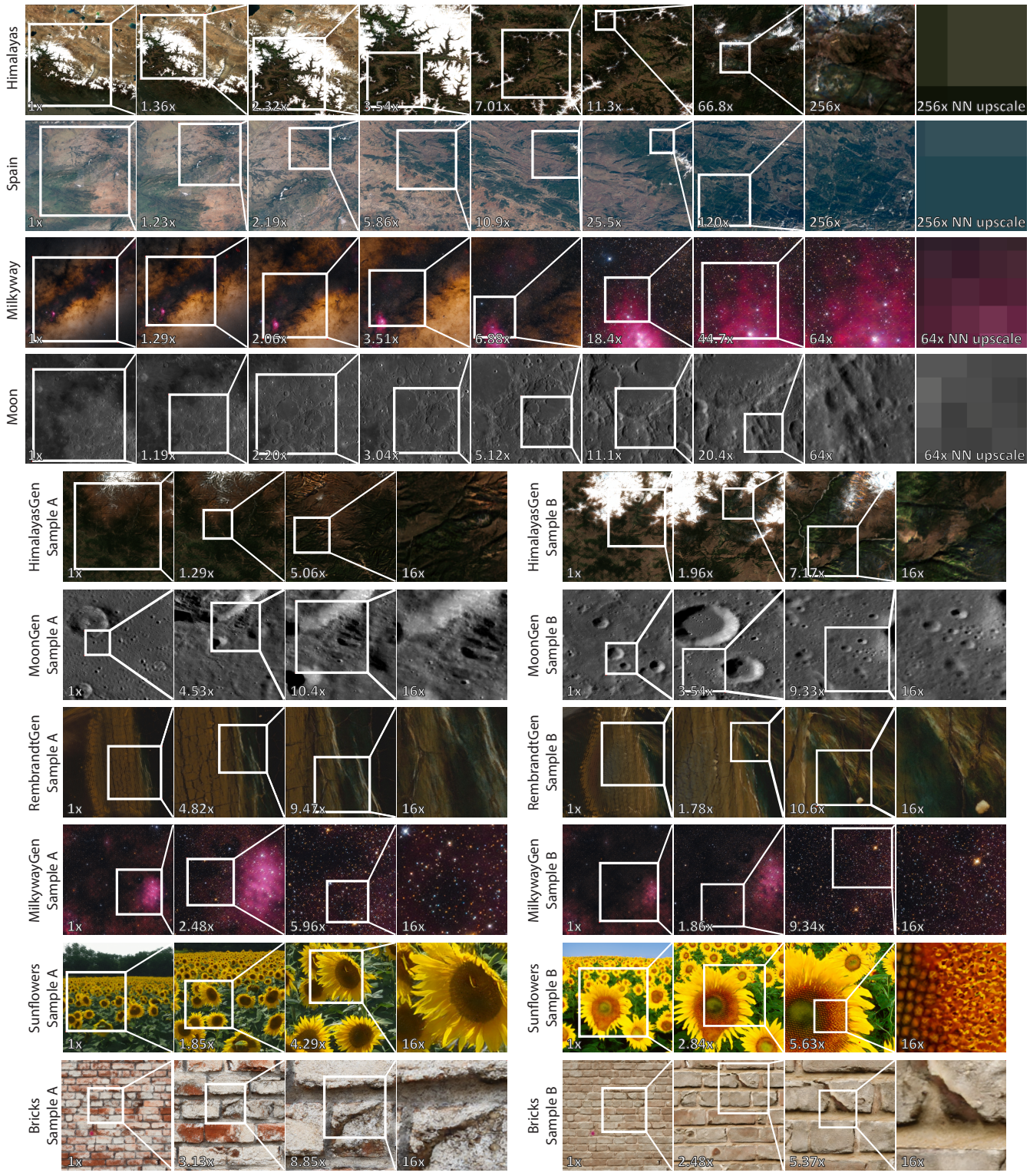

Fig. 10. Traversal of our scale spaces. The top four rows show pseudo-reconstructions, while the bottom six rows demonstrate generative scale spaces. The last column in the top four rows shows the upsampled version of the image in the first column that corresponds to the area that the image in the second-to-last column covers. Please refer to our supplemental video for demonstrations of continuous zooming and panning.

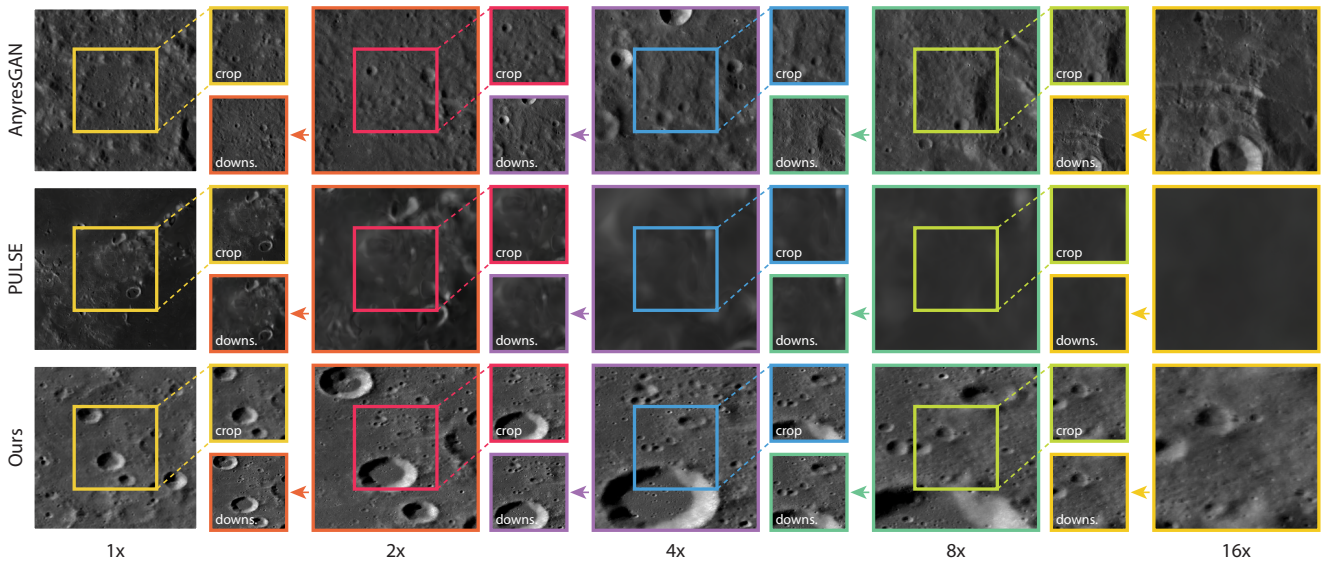

Fig. 11. Qualitative comparison of different methods (rows). Every column (large images) depicts a 2x zoom into the center of the image in the previous column. The small insets compare the same image region across two adjacent scales (the coarser scale is cropped, the finer scale is downsampled). Please note the lack of scale consistency for AnyresGAN and the loss of details for PULSE. Only our method is consistent and produces rich details across scales.

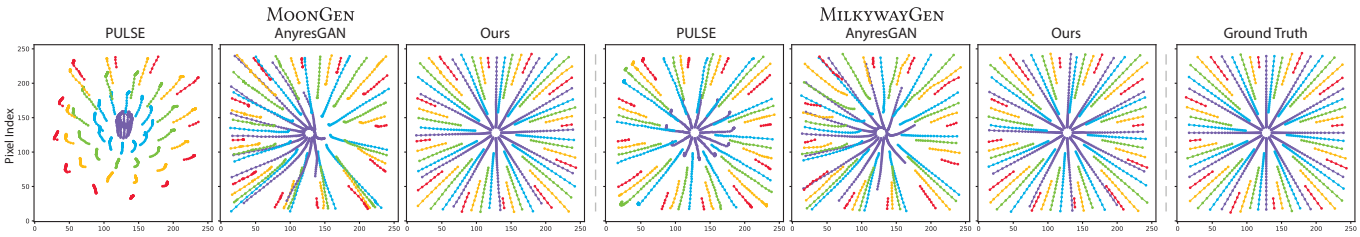

Fig. 12. Pixel trajectories for representative samples of different methods. To obtain the trajectories, we compute an image sequence zooming into the image center and concatenate optical flow vectors [Teed and Deng 2020] of adjacent images in the sequence per pixel. Here, we visualize the trajectories for 60 pixels only to avoid clutter. A perfect flow field is radial, i.e., linearly expanding from the center (Ground Truth, right). Compared to the baselines, our method provides trajectories closest to the ground truth.

the JPEG image, indicating that storing pixels of gigapixel images is inferior to our continuous generative approach.

Ablations. In Tab. 3, we report the results of ablation studies on the pseudo-reconstruction task using Milkyway.

We first consider alternatives to our Beta sampling in Sec. 4.2.2. We compare against a uniform sampling of the full-scale range, as well as two scales. We observe that our Beta sampling improves all relevant metrics.

We further study the effect of reducing dataset size. We observe that, unsurprisingly, quality and consistency are highest with the full dataset containing 96k patches, but even a significant reduction in dataset size does not have a dramatic negative effect on our model. Interestingly, our method still converges when using only 250 images distributed across the six scales. This, however, comes with a severe degradation in image quality, while scale consistency improves. Training on only 100 images diverges. Fig. 14 shows corresponding qualitative results.

Finally, we investigate the reliance of our method on exact scale labels by adding uniform random noise with an interval of two scales to the labels. We observe a minor drop in FID score, while scale consistency metrics are barely affected.

5.4 Limitations

Our distribution of Fourier features across generator layers (Sec. 4.1) comes with a disadvantage: Compared to a vanilla setup, the generator network has less capacity to turn procedural frequency content into final image output. This can occasionally lead to regularity artifacts in the generated patches. As illustrated in Fig. 15, some images are faintly overlaid with parallel lines. We observe that these artifacts mostly appear at the finest scales. Additionally, we occasionally observe saturated colorful blobs in our generated scale spaces. The origin of these artifacts can also be traced back to the late injection of Fourier features.

As with many generative approaches, training times of our method are substantial. The (manual) effort our approach allows to save

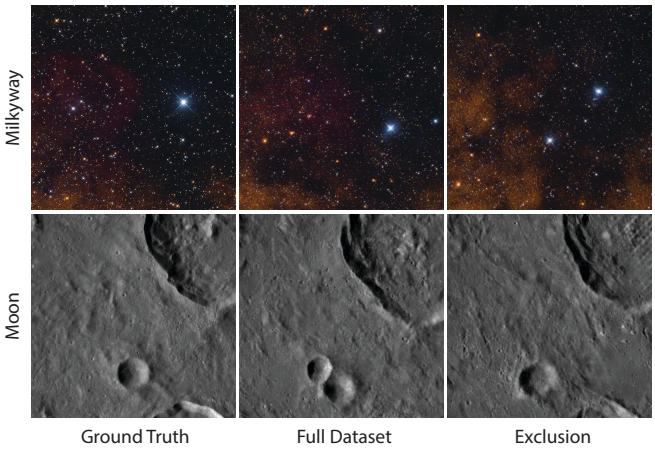

Fig. 13. Pseudo-reconstruction of an area not present in the training dataset. Our approach synthesizes plausible and coherent details, which do not necessarily match the reference, e.g., stars are placed at different locations. Milkyway image courtesy of Bartosz Wojczyński [2021].

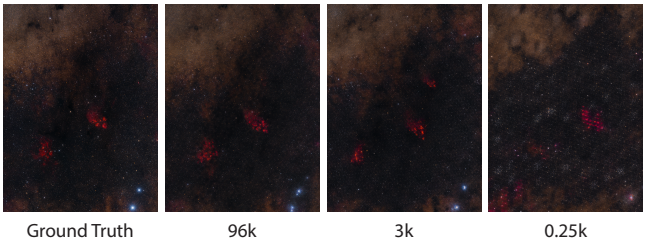

Fig. 14. Influence of dataset size on reconstruction quality. Structures gradually dissolve into regular patterns as training data becomes more sparse. Ground truth image courtesy of Bartosz Wojczyński [2021].

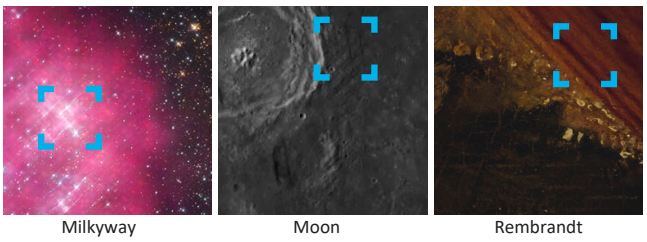

Fig. 15. Our solution occasionally produces artifacts in the form of overlays with parallel lines in certain sparse areas.

during data capture has to be paid by increased processing time. While we are not aware of any other method that is capable of handling the unstructured inputs our approach can process, more research at the foundations of generative modeling are necessary to allow end users with only a single workstation to fully benefit from our technology.

Table 3. Ablations.

Config.	FID ↓	Scale Consistency				PSNR$_{GT}$↑
		Bias ↓	Angle ↓	EMD ↓	PSNR$_{inter}$ ↑	
Uniform$_{full}$	9.5	0.20	1.6	2.77	25.1	23.8
Uniform$_2$	14.6	0.19	1.8	2.97	24.9	23.6
250 Patches	48.4	0.14	1.74	1.71	26.6	21.7
500 Patches	28.9	0.21	1.7	1.45	26.6	21.1
1k Patches	23.5	0.15	1.26	1.57	26.3	22.8
3k Patches	14.4	0.23	5.1	8.46	25.0	23.5
12k Patches	10.6	0.17	2.1	3.00	25.7	23.8
Noisy Scales	14.4	0.21	1.6	1.87	25.5	23.2
Ours	9.00	0.21	1.5	2.08	25.7	23.9

6 CONCLUSION

We have presented a novel approach for learning a multiscale image representation from a collection of low-resolution, unstructured images. Our method enhances an alias-free generator with progressive Fourier features distributed across various layers. Furthermore, we have developed techniques to stabilize training and guarantee scale consistency. The strength of our method is demonstrated in both multi-scale generative modeling and pseudo-reconstruction of scale spaces from unstructured patches. For the first time, our neural representation achieves zoom-in factors of up to 256x, opening up a new way for efficient modeling of multi-scale images.

ACKNOWLEDGMENTS

The authors thank Bartosz Wojczyński for providing the Milkyway data, as well as Joachim Weickert and Pascal Peter for early discussions. This research was partially funded by the ERC Advanced Grant FUNGRAPH (https://fungraph.inria.fr), No 788065 and an academic gift from Meta.

REFERENCES

Andreas Antoniou. 2006. *Digital signal processing*. McGraw-Hill.
Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *ICCV*. 5855–5864.
Yash Belhe, Michaël Gharbi, Matthew Fisher, Iliyan Georgiev, Ravi Ramamoorthi, and Tzu-Mao Li. 2023. Discontinuity-Aware 2D Neural Fields. *ACM Trans. Graph.* 42, 6, Article 217 (dec 2023), 11 pages. https://doi.org/10.1145/3618379
Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. 2019. Blind Super-Resolution Kernel Estimation using an Internal-GAN. In *NeurIPS*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/5fd0b37cd7dbbb00f97ba6ce92bf5add-Paper.pdf
Sergei Belousov. 2021. Mobilestylegan: A lightweight convolutional neural network for high-fidelity image synthesis. *arXiv preprint arXiv:2104.04767* (2021).
Sam Bond-Taylor and Chris G. Willcocks. 2024. ∞-Diff: Infinite Resolution Diffusion with Subsampled Mollified States. In *ICLR*.
Ashish Bora, Eric Price, and Alexandros G Dimakis. 2018. AmbientGAN: Generative models from lossy measurements. In *ICLR*.
David J Brady, Michael E Gehm, Ronald A Stack, Daniel L Marks, David S Kittle, Dathon R Golish, EM Vera, and Steven D Feller. 2012. Multiscale gigapixel photography. *Nature* 486, 7403 (2012), 386–389.
Peter J Burt. 1981. Fast filter transform for image processing. *Computer graphics and image processing* 16, 1 (1981), 20–51.
Lucy Chai, Michaël Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. 2022. Any-Resolution Training for High-Resolution Image Synthesis. In *ECCV*. 170–188.
Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. 2021. GLEAN: Generative Latent Bank for Large-Factor Image Super-Resolution. In *CVPR*.

14245–14254.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021b. Pre-Trained Image Processing Transformer. In *CVPR*. 12299–12310.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. 2021a. Learning Continuous Image Representation With Local Implicit Image Function. In *CVPR*. 8628–8638.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Trans. Graph.* 39, 4, Article 75 (aug 2020), 13 pages. https://doi.org/10.1145/3386569.3392457

Copernicus. 2024. Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A. Copernicus Sentinel data. Accessed 2024-01-24.

Oliver S Cossairt, Daniel Miau, and Shree K Nayar. 2011. Gigapixel computational imaging. In *ICCP*. 1–8.

Ingrid Daubechies. 1988. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics* 41, 7 (1988), 909–996.

Stavros Diolatzis, Jan Novak, Fabrice Rousselle, Jonathan Granskog, Miika Aittala, Ravi Ramamoorthi, and George Drettakis. 2023. MesoGAN: Generative Neural Reflectance Shells. In *Computer Graphics Forum*. Wiley Online Library.

Charles Eames and Ray Eames. 1968. Powers of Ten (film). In *Pyramid Films*.

Alexei A. Efros and William T. Freeman. 2001. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. Association for Computing Machinery, New York, NY, USA, 341–346. https://doi.org/10.1145/383259.383296

Alexei A. Efros and Thomas K. Leung. 1999. Texture synthesis by non-parametric sampling. In *ICCV*, Vol. 2. 1033–1038. https://doi.org/10.1109/ICCV.1999.790383

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*. 12873–12883.

Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. 2020. Multiplicative filter networks. In *ICLR*.

Anna Frühstück, Ibraheem Alhashim, and Peter Wonka. 2019. TileGAN: synthesis of large-scale non-homogeneous textures. *ACM Trans. Graph.* 38, 4, Article 58 (jul 2019), 11 pages. https://doi.org/10.1145/3306346.3322993

Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. 2023. Implicit Diffusion Models for Continuous Super-Resolution. In *CVPR*. 10021–10030.

Daniel Glasner, Shai Bagon, and Michal Irani. 2009. Super-resolution from a single image. In *ICCV*. 349–356. https://doi.org/10.1109/ICCV.2009.5459271

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*. 2672–2680.

Sarkis Halladjian, Haichao Miao, David Kouřil, M Eduard Gröller, Ivan Viola, and Tobias Isenberg. 2019. Scale Trotter: Illustrative visual travels across negative scales. *IEEE TVCG* 26, 1 (2019), 654–664.

Charles Han, Eric Risser, Ravi Ramamoorthi, and Eitan Grinspun. 2008. Multiscale texture synthesis. *ACM Trans. Graph.* 27, 3 (aug 2008), 1–8. https://doi.org/10.1145/1360612.1360650

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS* 30 (2017).

Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. 2019. Meta-SR: A Magnification-Arbitrary Network for Super-Resolution. In *CVPR*. 1575–1584.

Taizo Iijima. 1959. Basic theory of pattern observation. *Technical Group on Automata and Automatic Control* (1959), 3–32.

Michal Irani and Shmuel Peleg. 1991. Improving resolution by image registration. *CVGIP: Graphical models and image processing* 53, 3 (1991), 231–239.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training generative adversarial networks with limited data. *NeurIPS* 33 (2020), 12104–12114.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *NeurIPS* 34 (2021), 852–863.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*. 8110–8119.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising Diffusion Restoration Models. In *NeurIPS*, Vol. 35. Curran Associates, Inc., 23593–23606. https://proceedings.neurips.cc/paper_files/paper/2022/file/95504595b6169131b6ed6cd72eb05616-Paper-Conference.pdf

Staffan Klashed, Per Hemingsson, Carter Emmart, Matthew Cooper, and Anders Ynnerman. 2010. Uniview - Visualizing the Universe. In *Eurographics 2010 - Areas Papers*, Matthew Cooper and Kari Pulli (Eds.). The Eurographics Association, 37–43. https://doi.org/10.2312/ega.20101005

Jan J Koenderink. 1984. The structure of images. *Biological cybernetics* 50, 5 (1984), 363–370.

Johannes Kopf, Matt Uyttendaele, Oliver Deussen, and Michael F. Cohen. 2007. Capturing and viewing gigapixel images. *ACM Trans. Graph.* 26, 3 (jul 2007), 93–es. https://doi.org/10.1145/1276377.1276494

Alexandr Kuznetsov, Krishna Mullia, Zexiang Xu, Miloš Hašan, and Ravi Ramamoorthi. 2021. NeuMIP: multi-resolution neural materials. *ACM Trans. Graph.* 40, 4, Article 175 (jul 2021), 13 pages. https://doi.org/10.1145/3450626.3459795

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. 2023. SyncDiffusion: Coherent Montage via Synchronized Joint Diffusions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. SwinIR: Image Restoration Using Swin Transformer. In *ICCV Workshops*. 1833–1844.

Jian Liang, Chenfei Wu, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. 2022. NUWA-Infinity: Autoregressive over Autoregressive Generation for Infinite Visual Synthesis. In *NeurIPS*, Vol. 35. Curran Associates, Inc., 15420–15432. https://proceedings.neurips.cc/paper_files/paper/2022/file/6358cd0cd6607fdf4870595795eb1710-Paper-Conference.pdf

Cody Licorish, Noura Faraj, and Brian Summa. 2021. Adaptive Compositing and Navigation of Variable Resolution Images. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 138–150.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *CVPR Workshops*. 136–144.

Chieh Hubert Lin, Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, and Ming-Hsuan Yang. 2022. InfinityGAN: Towards Infinite-Pixel Image Synthesis. In *ICLR*. https://openreview.net/forum?id=ufGMqIM0a4b

Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. 2023b. InfiniCity: Infinite-Scale City Synthesis. In *ICCV*. 22808–22818.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. 2023a. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070* (2023).

Tony Lindeberg. 2013. *Scale-space theory in computer vision*. Vol. 256. Springer Science & Business Media.

Tony Lindeberg. 2022. Scale-covariant and scale-invariant Gaussian derivative networks. *Journal of Mathematical Imaging and Vision* 64, 3 (2022), 223–242.

David B. Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. 2022. BACON: Band-Limited Coordinate Networks for Multiscale Scene Representation. In *CVPR*. 16252–16262.

Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. 2021. MASA-SR: Matching Acceleration and Spatial Adaptation for Reference-Based Image Super-Resolution. In *CVPR*. 6368–6377.

Stephane G. Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE PAMI* 11, 7 (1989), 674–693.

Benoit B. Mandelbrot. 1982. *The fractal geometry of nature*. Vol. 1. WH freeman New York.

David Marr and Ellen Hildreth. 1980. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 207, 1167 (1980), 187–217.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *CVPR*. 2437–2445.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which Training Methods for GANs do actually Converge?. In *ICML (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 3481–3490. https://proceedings.mlr.press/v80/mescheder18a.html

Tomer Michaeli and Michal Irani. 2014. Blind deblurring using internal patch recurrence. In *ECCV*. Springer, 783–798.

Haneen Mohammed, Ali K Al-Awami, Johanna Beyer, Corrado Cali, Pierre Magistretti, Hanspeter Pfister, and Markus Hadwiger. 2017. Abstractocyte: A visual tool for exploring nanoscale astroglial cells. *IEEE TVCG* 24, 1 (2017), 853–861.

Brian B. Moser, Federico Raue, Stanislav Frolov, Sebastian Palacio, Jörn Hees, and Andreas Dengel. 2023. Hitchhiker's Guide to Super-Resolution: Introduction and Recent Advances. *IEEE PAMI* (2023), 1–21. https://doi.org/10.1109/TPAMI.2023.3243794

Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Radu Timofte, Martin Danelljan, and Luc Van Gool. 2022. Arbitrary-Scale Image Synthesis. In *CVPR*. 11533–11542.

Jim R Parker. 2010. *Algorithms for image processing and computer vision*. John Wiley & Sons.

Hallison Paz, Tiago Novello, Vinicius Silva, Guilherme Schardong, Luiz Schirmer, Fabio Chagas, Helio Lopes, and Luiz Velho. 2022. Multiresolution Neural Networks for Imaging. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Vol. 1. 174–179. https://doi.org/10.1109/SIBGRAPI55357.2022.9991765

Rijksmuseum. 2019. Operation Nightwatch.

Carlos Rodriguez-Pardo and Elena Garces. 2022. SeamlessGAN: Self-Supervised Synthesis of Tileable Texture Maps. *IEEE TVCG* (2022).

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *ICCV*. IEEE, 59–66.

Shibani Santurkar, David Budden, and Nir Shavit. 2018. Generative compression. In *2018 Picture Coding Symposium (PCS)*. IEEE, 258–262.

Vishwanath Saragadam, Jasper Tan, Guha Balakrishnan, Richard G. Baraniuk, and Ashok Veeraraghavan. 2022. MINER: Multiscale Implicit Neural Representation. In *ECCV*. 318–333.

Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 49, 10 pages. https://doi.org/10.1145/3528233.3530738

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. SinGAN: Learning a Generative Model From a Single Natural Image. In *ICCV*. 4570–4580.

Shayan Shekarforoush, David Lindell, David J Fleet, and Marcus A Brubaker. 2022. Residual Multiplicative Filter Networks for Multiscale Reconstruction. In *NeurIPS*, Vol. 35. Curran Associates, Inc., 8550–8563. https://proceedings.neurips.cc/paper_files/paper/2022/file/38e491559eb9e4cf31b8cd3a4e222436-Paper-Conference.pdf

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *CVPR*. 1874–1883.

Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. 2019. InGAN: Capturing and Retargeting the "DNA" of a Natural Image. In *ICCV*. 4492–4501.

Assaf Shocher, Nadav Cohen, and Michal Irani. 2018. "Zero-Shot" Super-Resolution Using Deep Internal Learning. In *CVPR*. 3118–3126.

Xavier Snelgrove. 2017. High-resolution multi-scale neural texture synthesis. In *SIGGRAPH Asia 2017 Technical Briefs* (Bangkok, Thailand) *(SA '17)*. Association for Computing Machinery, New York, NY, USA, Article 13, 4 pages. https://doi.org/10.1145/3145749.3149449

Sanghyun Son and Kyoung Mu Lee. 2021. SRWarp: Generalized Image Super-Resolution under Arbitrary Transformation. In *CVPR*. 7782–7791.

Gaochao Song, Qian Sun, Luo Zhang, Ran Su, Jianfeng Shi, and Ying He. 2023. OPE-SR: Orthogonal Position Encoding for Designing a Parameter-Free Upsampling Module in Arbitrary-Scale Image Super-Resolution. In *CVPR*. 10009–10020.

EJ Speyerer, MS Robinson, BW Denevi, LROC Science Team, et al. 2011. Lunar Reconnaissance Orbiter Camera global morphological map of the Moon. In *42nd Annual Lunar and Planetary Science Conference*. 2387.

Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. 2021. Neural Geometric Level of Detail: Real-Time Rendering With Implicit 3D Shapes. In *CVPR*. 11358–11367.

Wenbo Tao, Xiaoyu Liu, Yedi Wang, Leilani Battle, Çağatay Demiralp, Remco Chang, and Michael Stonebraker. 2019. Kyrix: Interactive pan/zoom visualizations at scale. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 529–540.

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*. Springer, 402–419.

Cristina N. Vasconcelos, Cengiz Oztireli, Mark Matthews, Milad Hashemi, Kevin Swersky, and Andrea Tagliasacchi. 2023. CUF: Continuous Upsampling Filters. (June 2023), 9999–10008.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. 2023c. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015* (2023).

Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. 2021a. Learning a Single Network for Scale-Arbitrary Super-Resolution. In *ICCV*. 4801–4810.

Xiaojuan Wang, Janne Kontkanen, Brian Curless, Steve Seitz, Ira Kemelmacher, Ben Mildenhall, Pratul Srinivasan, Dor Verbin, and Aleksander Holynski. 2023a. Generative Powers of Ten. *arXiv preprint* (2023).

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021b. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. In *ICCV Workshops*. 1905–1914.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *ECCV Workshops*.

Yinhuai Wang, Jiwen Yu, Runyi Yu, and Jian Zhang. 2023b. Unlimited-Size Diffusion Restoration. In *CVPR*. 1160–1167.

Zhihao Wang, Jian Chen, and Steven CH Hoi. 2020. Deep learning for image super-resolution: A survey. *IEEE PAMI* 43, 10 (2020), 3365–3387.

Min Wei and Xuesong Zhang. 2023. Super-Resolution Neural Operator. In *CVPR*. 18247–18256.

Norbert Wiener, Norbert Wiener, Cyberneticist Mathematician, Norbert Wiener, Norbert Wiener, and Cybernéticien Mathématicien. 1949. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. Vol. 113. MIT press Cambridge, MA.

Lance Williams. 1983. Pyramidal parametrics. In *Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques* (Detroit, Michigan, USA) *(SIGGRAPH '83)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/800059.801126

Andrew P Witkin. 1987. Scale-space filtering. In *Readings in Computer Vision*. Elsevier, 329–332.

Bartosz Wojczynski. 2021. 2.2 Gigapixel Milky Way. https://artuniverse.eu

Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. 2022. BungeeNeRF: Progressive Neural Radiance Field for Extreme Multi-scale Scene Rendering. In *ECCV*. 106–122.

Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2021. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716* (2021).

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. Learning Texture Transformer Network for Image Super-Resolution. In *CVPR*. 5791–5800.

Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. 2023. DiffCollage: Parallel Generation of Large Content With Diffusion Models. In *CVPR*. 10188–10198.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. 586–595.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018b. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *ECCV*. 286–301.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018c. Residual Dense Network for Image Super-Resolution. In *CVPR*. 2472–2481.

Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2018. Non-stationary texture synthesis by adversarial expansion. *ACM Trans. Graph.* 37, 4, Article 49 (jul 2018), 13 pages. https://doi.org/10.1145/3197517.3201285

Jialin Zhu and Tom Kelly. 2021. Seamless Satellite-image Synthesis. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 193–204.

Maria Zontak and Michal Irani. 2011. Internal statistics of a single natural image. In *CVPR 2011*. 977–984. https://doi.org/10.1109/CVPR.2011.5995401

U. Çoğalan, M. Bemana, HP. Seidel, and K. Myszkowski. 2023. Video frame interpolation for high dynamic range sequences captured with dual-exposure sensors. *Computer Graphics Forum* 42, 2 (2023), 119–131.