Learning Images Across Scales Using Adversarial Training

KRZYSZTOF WOLSKI, Max-Planck-Institut für Informatik, Germany
ADRASH DJECOUMAR, 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.

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...
Coarse
Fine

(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.

We introduce a novel approach for the construction of a multi-scale 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 coherence 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. 20fps) 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 [Kuznetsova 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 restoration 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. 2023a,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 [Snellgrove 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, 2023; Rodriguez-Pardo and Garcés 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 in performing 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 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 \( \hat{I} \) of an image [Iijima 1959; Witkin 1987], which captures versions of that image with different upper bounds on frequency content. We write

\[
\hat{I}(\mathbf{x}, \omega) = c,
\]

where \( c \in \mathbb{R}^3 \) 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 \( \omega \) corresponds to an image with only little detail, while increasing \( \omega \) progressively reveals structures of higher frequencies (Fig. 4a).

We are concerned with scale spaces in which \( \omega \) 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:

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

We further define \( s_{\max} = s(\omega_{\max}) \) to denote the full dynamic range of scales for a particular \( \hat{I} \). 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 \( \mathbf{x}_p \). According to the Nyquist-Shannon sampling theorem [Antoniou 2006], a signal with bandwidth \( \omega \) needs to be sampled at a rate of at least \( 2\omega \) 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

\[
N = N(s) = \lceil \sqrt[6]{2} \omega_{\min}^{2^{s+1}} \rceil.
\]

The factor \( \sqrt[6]{2} \) 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 \( \hat{I} \), 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 \( \hat{I} \), which relies on an unstructured collection of ordinary, low-resolution images that constitute patch samples of \( \hat{I} \). Each patch has a fixed resolution \( N_p \times N_p \) 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 \( \hat{I} \) from the input data (Fig. 5). Specifically, we design a generator \( G \) that is able to learn a continuous orders-of-magnitude scale space \( \hat{I}_G \).
G takes as input a random vector \( z \), as well as a continuous patch location \( x_p \) and scale \( s_p \), and renders one \( N_p \times N_p \) image at a time. Continuously varying \( x_p \) and \( s_p \) produces corresponding slices of the learned \( I_G \) [Bora et al. 2018]. This allows interactive exploration of \( I_G \), 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 \( I_G \) to the distribution of data patches from \( I \). During training, \( z \), \( x_p \) 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 \( I \), i.e., multiscale observations of the same scene, \( I_G \) converges to a plausible, coherent approximation of \( I \). While we find providing random inputs \( z \) to \( G \) is necessary for stable training, a converged generator \( G \) disregards \( z \), producing negligible variations of the output. We refer to this solution as pseudo-reconstruction, as \( I_G \) might differ from \( I \) in the arrangement of details, but captures the overall multiscale structure well. Note that in addition to missing knowledge about the patch locations \( x_p \), the patches do not exhaustively cover \( I \) 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 \( p(I_G) \), where different \( z \) 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 \( I_G \) 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 \( I_G \). Importantly, while the output of \( G \) is a pixel grid, the model has to be intrinsically continuous to be a faithful representation of \( I \) 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 \( f \in \mathbb{R}^d \), i.e., directional 2D sinusoids evaluated on a regular grid \( x_i \), i.e., they can be represented in the spatial domain as a pixel grid with \( d_p \) channels (Fig. 6a):

\[
   f_j(x_i) = \sin(2\pi \omega^T x_i),
\]

where \( \omega_j \in \mathbb{R}^d \) are \( d \) different frequency vectors. The features \( f \) 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 \( z \) 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 \( f \) 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 \( x_i \) via

\[
   g(x_i; x_p, s_p) = 2^{-s_p} (x_i - x_p),
\]

and feed the corresponding shifted and scaled Fourier features \( f_j(g(x_i; x_p, s_p)) \) into \( S \). Unfortunately, choosing a high \( s_p \) stretches \( f \) 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 \( \| \omega_j \| \) 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 \( \omega_j \), but progressively blend them in only after \( g \) has stretched out the corresponding \( f_j \) far enough to stay below the Nyquist limit, using some blending function \( w \) (Fig. 6d). While this strategy provides
Progressively blending in different $S$ 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 by the Fourier features in Eq. 4: The effect of scaling $f$ because we are interfering with the positional encoding provided during zooming. This is met: At the scale $s_p$ where the feature is blended in completely $f_s$. 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 $f_j$ to neural features after each up-sampling layer. The assignment of $f_j$ 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 $f_j$ 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 $f_j$ 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 $z$, we also randomly sample the patch location $x_p$ 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. 2020]. However, we find that our multiscale setting poses two significant challenges: (i) training stability, and (ii)
consistency of the learned $I_G$ across scales. We address these items using a progressive patch sampling scheme and a scale consistency loss, respectively.

### 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 $I_G$, we follow ideas from Chai et al. [2022] and add an additional loss term that compares two generated patches, separated by a scale offset $\Delta s_p$.

Let

$$L_i = \mathbb{E}_{x, x_p, s_p, \Delta s_p} \left[ d \left(R_{\Delta s_p} (G(z, x_p, s_p), G(z, x_p, s_p - \Delta s_p)) \right) \right].$$

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 $\ell_1$-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 $x_p$ in Eq. 7, we find that the choice of sampling distribution for $\Delta s_p$ has a strong influence on the scale consistency of $I_G$ for our large $s_{\max}$. To make sure $I_G$ is globally consistent, we want $\Delta s_p$ 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 $\Delta s_p$ 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:

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

where $\alpha = \sqrt{\max(1, s_p)}$, 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 contain 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.

---

[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 $\Delta s_p$ for different patch scales $s_p$.]

Learning Images Across Scales Using Adversarial Training • 1317
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 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 [San-turkar 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

![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.](image-url)
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 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 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
Learning Images Across Scales Using Adversarial Training

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.

Table 3. Ablations.

<table>
<thead>
<tr>
<th>Config.</th>
<th>FID ↓</th>
<th>Scale Consistency</th>
<th>PSNR GT ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias ↓</td>
<td>Angle ↓</td>
<td>EMD</td>
</tr>
<tr>
<td>Uniformfull</td>
<td>9.5</td>
<td>0.20</td>
<td>1.6</td>
</tr>
<tr>
<td>Uniform2</td>
<td>14.6</td>
<td>0.19</td>
<td>1.8</td>
</tr>
<tr>
<td>250 Patches</td>
<td>48.4</td>
<td>0.14</td>
<td>1.74</td>
</tr>
<tr>
<td>500 Patches</td>
<td>28.9</td>
<td>0.21</td>
<td>1.7</td>
</tr>
<tr>
<td>1k Patches</td>
<td>23.5</td>
<td>0.15</td>
<td>1.26</td>
</tr>
<tr>
<td>3k Patches</td>
<td>14.4</td>
<td>0.23</td>
<td>5.1</td>
</tr>
<tr>
<td>12k Patches</td>
<td>10.6</td>
<td>0.17</td>
<td>2.1</td>
</tr>
<tr>
<td>Noisy Scales</td>
<td>14.4</td>
<td>0.21</td>
<td>1.6</td>
</tr>
<tr>
<td>Ours</td>
<td>9.00</td>
<td>0.21</td>
<td>1.5</td>
</tr>
</tbody>
</table>

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

Sam Bond-Taylor and Chris G. Willcocks. 2024. ∞-Diff: Infinite Resolution Diffusion with Subsampled Mollified States. In ICLR
131:12     Krzysztof Wolski, Adarsh Deacoumar, Alireza Javamandi, Hans-Peter Seidel, Christian Theobalt, Guillaume Codonier, Karol Myszkowski, G went Dettke, Xiangang Pan, and Thomas Leinikohl


95504595b6169131b6ed6cd72eb05616-Paper-Conference.pdf


