PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models

Sachit Menon*, Alexandru Damian*, Shijia Hu, Nikhil Ravi, Cynthia Rudin
Duke University
Durham, NC
{sachit.menon,alexandru.damian,shijia.hu,nikhil.ravi,cynthia.rudin}@duke.edu

Abstract

The primary aim of single-image super-resolution is to construct a high-resolution (HR) image from a corresponding low-resolution (LR) input. In previous approaches, which have generally been supervised, the training objective typically measures a pixel-wise average distance between the super-resolved (SR) and HR images. Optimizing such metrics often leads to blurring, especially in high variance (detailed) regions. We propose an alternative formulation of the super-resolution problem based on creating realistic SR images that downscale correctly. We present a novel super-resolution algorithm addressing this problem, PULSE (Photo Upsampling via Latent Space Exploration), which generates high-resolution, realistic images at resolutions previously unseen in the literature. It accomplishes this in an entirely self-supervised fashion and is not confined to a specific degradation operator used during training, unlike previous methods (which require training on databases of LR-HR image pairs for supervised learning). Instead of starting with the LR image and slowly adding detail, PULSE traverses the high-resolution natural image manifold, searching for images that downscale to the original LR image. This is formalized through the "downscaling loss," which guides exploration through the latent space of a generative model. By leveraging properties of high-dimensional Gaussians, we restrict the search space to guarantee that our outputs are realistic. PULSE thereby generates super-resolved images that both are realistic and downscale correctly. We show extensive experimental results demonstrating the efficacy of our approach in the domain of face super-resolution (also known as face hallucination). Our method outperforms state-of-the-art methods in perceptual quality at higher resolutions and scale factors than previously possible.

* denotes equal contribution

1. Introduction

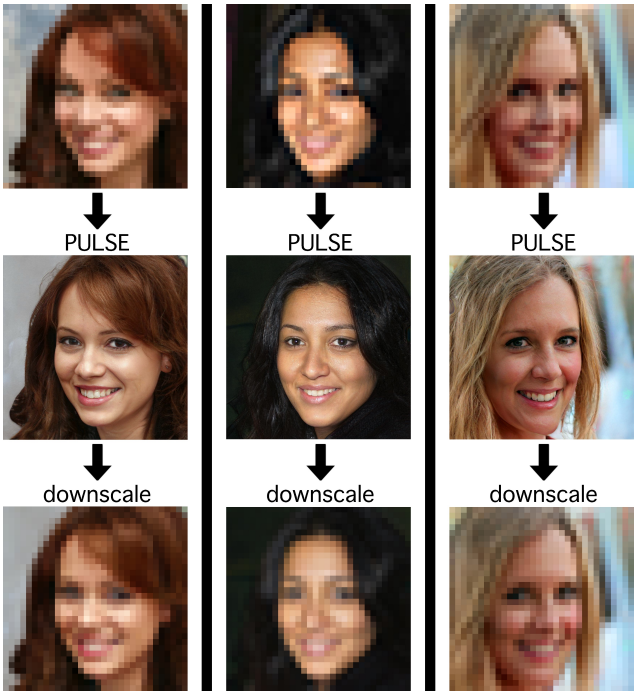

Figure 1. (x32) The input (top) gets upsampled to the SR image (middle) which downscales (bottom) to the original image.

In this work, we aim to transform blurry, low-resolution images into sharp, realistic, high-resolution images. Here, we focus on images of faces, but our technique is generally applicable. In many areas (such as medicine, astronomy, microscopy, and satellite imagery), sharp, high-resolution images are difficult to obtain due to issues of cost, hardware restriction, or memory limitations [20]. This leads to the capture of blurry, low-resolution images instead. In other cases, images could be old and therefore blurry, or even in a modern context, an image could be out of focus or a person could be in the background. In addition to being visually unappealing, this impairs the use of downstream anal-

ysis methods (such as image segmentation, action recognition, or disease diagnosis) which depend on having high-resolution images [17] [19]. In addition, as consumer laptop, phone, and television screen resolution has increased over recent years, popular demand for sharp images and video has surged. This has motivated recent interest in the computer vision task of *image super-resolution*, the creation of realistic high-resolution (henceforth HR) images that a given low-resolution (LR) input image could correspond to.

While the benefits of methods for image super-resolution are clear, the difference in information content between HR and LR images (especially at high scale factors) hampers efforts to develop such techniques. In particular, LR images inherently possess less high-variance information; details can be blurred to the point of being visually indistinguishable. The problem of recovering the true HR image depicted by an LR input, as opposed to generating a set of potential such HR images, is inherently ill-posed, as the size of the total set of these images grows exponentially with the scale factor [2]. That is to say, *many* high-resolution images can correspond to the exact same low-resolution image.

Traditional supervised super-resolution algorithms train a model (usually, a convolutional neural network, or CNN) to minimize the pixel-wise mean-squared error (MSE) between the generated super-resolved (SR) images and the corresponding ground-truth HR images [14] [7]. However, this approach has been noted to neglect perceptually relevant details critical to photorealism in HR images, such as texture [15]. Optimizing on an average difference in pixel-space between HR and SR images has a blurring effect, encouraging detailed areas of the SR image to be smoothed out to be, on average, more (pixelwise) correct. In fact, in the case of mean squared error (MSE), the ideal solution is the (weighted) pixel-wise average of the set of realistic images that downscale properly to the LR input (as detailed later). The inevitable result is smoothing in areas of high variance, such as areas of the image with intricate patterns or textures. As a result, MSE should not be used alone as a measure of image quality for super-resolution.

Some researchers have attempted to extend these MSE-based methods to additionally optimize on metrics intended to encourage realism, serving as a force opposing the smoothing pull of the MSE term [15, 7]. This essentially drags the MSE-based solution in the direction of the natural image manifold (the subset of $\mathbb{R}^{M \times N}$ that represents the set of high-resolution images). This compromise, while improving perceptual quality over pure MSE-based solutions, makes no guarantee that the generated images are realistic. Images generated with these techniques still show signs of blurring in high variance areas of the images, just as in the pure MSE-based solutions.

To avoid these issues, we propose a new paradigm for super-resolution. The goal should be to generate realistic images within the set of feasible solutions; that is, to find points which *actually lie on the natural image manifold and also downscale correctly*. The (weighted) pixel-wise average of possible solutions yielded by the MSE does not generally meet this goal for the reasons previously described. We provide an illustration of this in Figure 2.

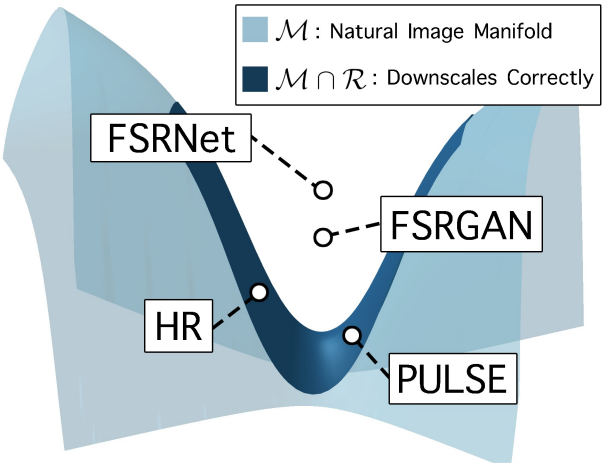

Figure 2. FSRNet tends towards an average of the images that downscale properly. The discriminator loss in FSRGAN pulls it in the direction of the natural image manifold, whereas PULSE always moves along this manifold.

Our method generates images using a (pretrained) generative model approximating the distribution of natural images under consideration. For a given input LR image, we traverse the manifold, parameterized by the latent space of the generative model, to find regions that downscale correctly. In doing so, we find examples of realistic images that downscale properly, as shown in 1.

Such an approach also eschews the need for supervised training, being entirely self-supervised with no 'training' needed at the time of super-resolution inference (except for the unsupervised generative model). This framework presents multiple substantial benefits. First, it allows the same network to be used on images with differing degradation operators even in the absence of a database of corresponding LR-HR pairs (as no training on such databases takes place). Furthermore, unlike previous methods, it does not require super-resolution task-specific network architectures, which take substantial time on the part of the researcher to develop without providing real insight into the problem; instead, it proceeds alongside the state-of-the-art in generative modeling, with zero retraining needed.

Our approach works with any type of generative model with a differentiable generator, including flow-based models, variational autoencoders (VAEs), and generative adversarial networks (GANs); the particular choice is dictated by the tradeoffs each make in approximating the data manifold. For this work, we elected to use GANs due to recent advances yielding high-resolution, sharp images [12, 11].

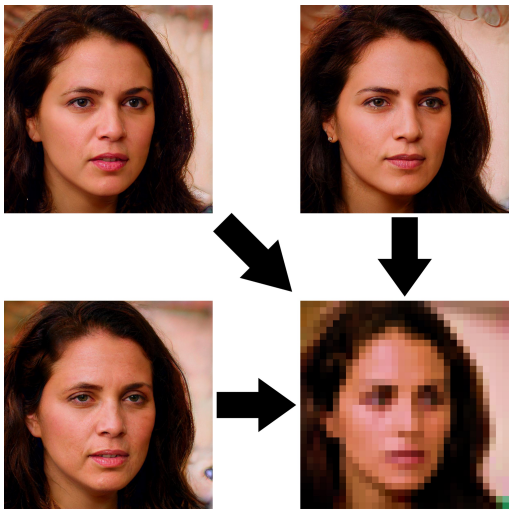

Figure 3. We show here how visually distinct images, created with PULSE, can all downscale (represented by the arrows) to the same LR image.

One particular subdomain of image super-resolution deals with the case of face images. This subdomain – known as *face hallucination* – finds application in consumer photography, photo/video restoration, and more [23]. As such, it has attracted interest as a computer vision task in its own right. Our work focuses on face hallucination, but our methods extend to a more general context.

Because our method always yields a solution that both lies on the natural image manifold and downsamples correctly to the original low-resolution image, we can provide a range of interesting high-resolution possibilities e.g. by making use of the stochasticity inherent in many generative models: our technique can create a *set* of images, *each* of which is visually convincing, yet look different from each other, where (without ground truth) *any* of the images could plausibly have been the source of the low-resolution input.

Our main contributions are as follows.

1. **A new paradigm for image super-resolution**. Previous efforts take the traditional, ill-posed perspective of attempting to 'reconstruct' an HR image from an LR input, yielding outputs that, in effect, average many possible solutions. This averaging introduces undesirable blurring. We introduce new approach to super-resolution: a super-resolution algorithm should create realistic high-resolution outputs that downscale to the correct LR input.

2. **A novel method for solving the super-resolution task**. In line with our new perspective, we propose a new algorithm for super-resolution. Whereas traditional work has at its core aimed to approximate the LR $\rightarrow$ HR map using supervised learning (especially with neural networks), our approach centers on the use of unsupervised generative models of HR data. Using generative adversarial networks, we explore the latent space to find regions that map to realistic images and downscale correctly. No retraining is required. Our particular implementation, using StyleGAN [12], allows for the creation of any number of realistic SR samples that correctly map to the LR input.

3. **An original method for latent space search under high-dimensional Gaussian priors.** In our task and many others, it is often desirable to find points in a generative model's latent space that map to realistic outputs. Intuitively, these should resemble samples seen during training. At first, it may seem that traditional log-likelihood regularization by the latent prior would accomplish this, but we observe that the 'soap bubble' effect (that much of the density of a high dimensional Gaussian lies close to the surface of a hypersphere) contradicts this. Traditional log-likelihood regularization actually tends to draw latent vectors away from this hypersphere and, instead, towards the origin. We therefore constrain the search space to the surface of that hypersphere, which ensures realistic outputs in higher-dimensional latent spaces; such spaces are otherwise difficult to search.

2. Related Work

While there is much work on image super-resolution prior to the advent of convolutional neural networks (CNNs), CNN-based approaches have rapidly become state-of-the-art in the area and are closely relevant to our work; we therefore focus on neural network-based approaches here. Generally, these methods use a pipeline where a low-resolution (LR) image, created by downsampling a high-resolution (HR) image, is fed through a CNN with both convolutional and upsampling layers, generating a super-resolved (SR) output. This output is then used to calculate the loss using the chosen loss function and the original HR image.

2.1. Current Trends

Recently, supervised neural networks have come to dominate current work in super-resolution. Dong *et al.* [8] proposed the first CNN architecture to learn this non-linear LR to HR mapping using pairs of HR-LR images. Several groups have attempted to improve the upsampling step by utilizing sub-pixel convolutions and transposed convolutions [18]. Furthermore, the application of ResNet architectures to super-resolution (started by SRResNet [15]), has yielded substantial improvement over more traditional convolutional neural network architectures. In particular, the use of residual structures allowed for the training of larger networks. Currently, there exist two general trends: one, towards networks that primarily better optimize pixel-wise

average distance between SR and HR, and two, networks that focus on perceptual quality.

2.2. Loss Functions

Towards these different goals, researchers have designed different loss functions for optimization that yield images closer to the desired objective. Traditionally, the loss function for the image super-resolution task has operated on a per-pixel basis, usually using the L2 norm of the difference between the ground truth and the reconstructed image, as this directly optimizes PSNR (the traditional metric for the super-resolution task). More recently, some researchers have started to use the L1 norm since models trained using L1 loss seem to perform better in PSNR evaluation. The L2 norm (as well as pixel-wise average distances in general) between SR and HR images has been heavily criticized for not correlating well with human-observed image quality [15]. In face super-resolution, the state-of-the-art for such metrics is FSRNet [7], which used a facial prior to achieve previously unseen PSNR.

Perceptual quality, however, does not necessarily increase with higher PSNR. As such, different methods, and in particular, objective functions, have been developed to increase perceptual quality. In particular, methods that yield high PSNR result in blurring of details. The information required for details is often not present in the LR image and must be 'imagined' in. One approach to avoiding the direct use of the standard loss functions was demonstrated in [21], which draws a prior from the structure of a convolutional network. This method produces similar images to the methods that focus on PSNR, which lack detail, especially in high frequency areas. Because this method cannot leverage learned information about what realistic images look like, it is unable to fill in missing details. Methods that try to learn a map from LR to HR images can try to leverage learned information; however, as mentioned, networks optimized on PSNR are still explicitly penalized for attempting to hallucinate details they are unsure about, thus optimizing on PSNR stills resulting in blurring and lack of detail.

To resolve this issue, some have tried to use generative model-based loss terms to provide these details. Neural networks have lent themselves to application in generative models of various types (especially generative adversarial networks–GANs–from [9]), to image reconstruction tasks in general, and more recently, to super-resolution. Ledig *et al.* [15] created the SRGAN architecture for single-image upsampling by leveraging these advances in deep generative models, specifically GANs. Their general methodology was to use the generator to upscale the low-resolution input image, which the discriminator then attempts to distinguish from real HR images, then propagate the loss back to both networks. Essentially, this optimizes a supervised network much like MSE-based methods with an additional loss term corresponding to how fake the discriminator believes the generated images to be. However, this approach is fundamentally limited as it essentially results in an averaging of the MSE-based solution and a GAN-based solution, as we discuss later. In the context of faces, this technique has been incorporated into FSRGAN, resulting in the current perceptual state-of-the-art in face super resolution at $\times 8$ upscaling factors up to resolutions of 128×128. Although these methods use a 'generator' and a 'discriminator' as found in GANs, they are trained in a completely supervised fashion; they do not use unsupervised generative models.

2.3. Generative Networks

Our algorithm does not simply use GAN-style training; rather, it uses a truly unsupervised GAN (or, generative model more broadly). It searches the latent space of this generative model for latents that map to images that downscale correctly. The quality of cutting-edge generative models is therefore of interest to us.

As GANs have produced the highest-quality high-resolution images of deep generative models to date, we chose to focus on these for our implementation. Here we provide a brief review of relevant GAN methods with high-resolution outputs. Karras *et al*. [11] presented some of the first high-resolution outputs of deep generative models in their ProGAN algorithm, which grows both the generator and the discriminator in a progressive fashion. Karras *et al*. [12] further built upon this idea with StyleGAN, aiming to allow for more control in the image synthesis process relative to the black-box methods that came before it. The input latent code is embedded into an intermediate latent space, which then controls the behavior of the synthesis network with adaptive instance normalization applied at each convolutional layer. This network has 18 layers (2 each for each resolution from 4×4 to 1024×1024). After every other layer, the resolution is progressively increased by a factor of 2. At each layer, new details are introduced stochastically via Gaussian input to the adaptive instance normalization layers. Without perturbing the discriminator or loss functions, this architecture leads to the option for scale-specific mixing and control over the expression of various high-level attributes and variations in the image (e.g. pose, hair, freckles, etc.). Thus, StyleGAN provides a very rich latent space for expressing different features, especially in relation to faces.

3. Method

We begin by defining some universal terminology necessary to any formal description of the super-resolution problem. We denote the low-resolution input image by I_{LR}. We aim to learn a conditional generating function G that, when applied to I_{LR}, yields a higher-resolution super-resolved image I_{SR}. Formally, let $I_{LR} \in \mathbb{R}^{m \times n}$. Then our desired

function SR is a map $\mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{M\times N}$ where $M > m$, $N > n$. We define the super-resolved image $I_{SR} \in \mathbb{R}^{M\times N}$

$$I_{SR} := SR(I_{LR}). \tag{1}$$

In a traditional approach to super-resolution, one considers that the low-resolution image could represent the same information as a theoretical high-resolution image $I_{HR} \in \mathbb{R}^{M\times N}$. The goal is then to best recover this particular I_{HR} given I_{LR}. Such approaches therefore reduce the problem to an optimization task: fit a function SR that minimizes

$$L := \|I_{HR} - I_{SR}\|_p^p \tag{2}$$

where $\|\cdot\|_p$ denotes some l^p norm.

In practice, even when trained correctly, these algorithms fail to enhance detail in high variance areas. To see why this is, fix a low resolution image I_{LR}. Let $\mathcal{M}$ be the natural image manifold in $\mathbb{R}^{M\times N}$, i.e., the subset of $\mathbb{R}^{M\times N}$ that resembles natural realistic images, and let P be a probability distribution over $\mathcal{M}$ describing the likelihood of an image appearing in our dataset. Finally, let R be the set of images that downscale correctly, i.e., $R = \{I \in \mathbb{R}^{N\times M} \ : \ DS(I) = I_{LR}\}$. Then in the limit as the size of our dataset tends to infinity, our expected loss when the algorithm outputs a fixed image I_{SR} is

$$\int_{\mathcal{M}\cap R} \|I_{HR} - I_{SR}\|_p^p \, dP(I_{HR}). \tag{3}$$

This is minimized when I_{SR} is an l_p average of I_{HR} over $M \cap R$. In fact, when $p = 2$, this is minimized when

$$I_{SR} = \int_{\mathcal{M}\cap R} I_{HR} \, dP(I_{HR}), \tag{4}$$

so *the optimal I_{SR} is a weighted pixelwise average of the set of high resolution images that downscale properly*. As a result, the lack of detail in algorithms that rely only on an l_p norm cannot be fixed simply by changing the architecture of the network. The problem itself has to be rephrased.

We therefore propose a new framework for single image super resolution. Let $\mathcal{M}$, DS be defined as above. Then for a given LR image $I_{LR} \in \mathbb{R}^{m\times n}$ and $\epsilon > 0$, our goal is to find an image $I_{SR} \in \mathcal{M}$ with

$$\|DS(I_{SR}) - I_{LR}\|_p \leq \epsilon. \tag{5}$$

In particular, we can let $\mathcal{R}_\epsilon \subset \mathbb{R}^{N\times M}$ be the set of images that downscale properly, i.e.,

$$\mathcal{R}_\epsilon = \{I \in \mathbb{R}^{N\times M} \ : \ \|DS(I) - I_{LR}\|_p^p \leq \epsilon\}. \tag{6}$$

Then we are seeking an image $I_{SR} \in \mathcal{M}\cap\mathcal{R}_\epsilon$. The set $\mathcal{M}\cap\mathcal{R}_\epsilon$ is the set of *feasible solutions*, because a solution is not feasible if it did not downscale properly and look realistic.

It is also interesting to note that the intersections $\mathcal{M}\cap\mathcal{R}_\epsilon$ and in particular $\mathcal{M} \cap \mathcal{R}_0$ are guaranteed to be nonempty, because they must contain the original HR image (i.e., what traditional methods aim to reconstruct).

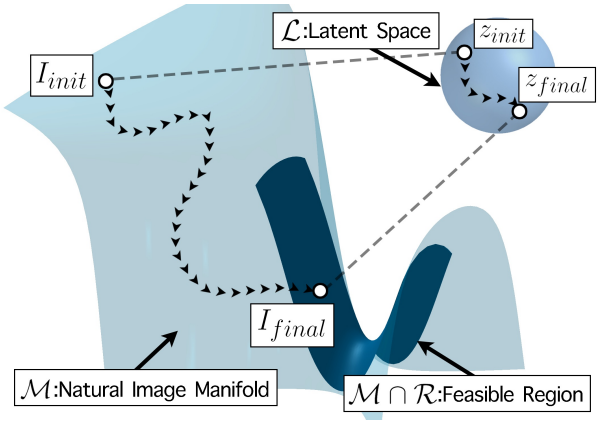

Figure 4. While traveling from z_{init} to z_{final} in the latent space $\mathcal{L}$, we travel from $I_{init} \in \mathcal{M}$ to $I_{final} \in \mathcal{M} \cap \mathcal{R}$.

3.1. Downscaling Loss

Central to the problem of super-resolution, unlike general image generation, is the notion of *correctness*. Traditionally, this has been interpreted to mean how well a particular ground truth image I_{HR} is 'recovered' by the application of the super-resolution algorithm SR to the low-resolution input I_{LR}, as discussed in the related work section above. This is generally measured by some l_p norm between I_{SR} and the ground truth, I_{HR}; such algorithms only look somewhat like real images because minimizing this metric drives the solution somewhat nearer to the manifold. However, they have no way to ensure that I_{SR} lies *close* to $\mathcal{M}$. In contrast, in our framework, we never deviate from $\mathcal{M}$, so such a metric is not necessary. For us, the critical notion of correctness is how well the generated SR image I_{SR} corresponds to I_{LR}.

We formalize this through the *downscaling loss*, to explicitly penalize a proposed SR image for deviating from its LR input (similar loss terms have been proposed in [1],[21]). This is inspired by the following: for a proposed SR image to represent the same information as a given LR image, it must downscale to this LR image. That is,

$$I_{LR} \approx DS(I_{SR}) = DS(SR(I_{LR})) \tag{7}$$

where $DS(\cdot)$ represents the downscaling function.

Our downscaling loss therefore penalizes SR the more its outputs violate this,

$$L_{DS}(I_{SR}, I_{LR}) := \|DS(I_{SR}) - I_{LR}\|_p^p. \tag{8}$$

It is important to note that the downscaling loss can be used in both supervised and unsupervised models for super-resolution; it does not depend on an HR reference image.

3.2. Latent Space Exploration

How might we find regions of the natural image manifold $\mathcal{M}$ that map to the correct LR image under the downscaling

operator? If we had a differentiable parameterization of the manifold, we could progress along the manifold to these regions by using the downscaling loss to guide our search. In that case, images found would be guaranteed to be high resolution as they came from the HR image manifold, while also being correct as they would downscale to the LR input.

In reality, we do not have such convenient, perfect parameterizations of manifolds. However, we can approximate such a parameterization by using techniques from unsupervised learning. In particular, much of the field of deep generative modeling (e.g. VAEs, flow-based models, and GANs) is concerned with creating models that map from some latent space to a given manifold of interest. By leveraging advances in generative modeling, we can even use pretrained models without the need to train our own network. Some prior work has aimed to find vectors in the latent space of a generative model to accomplish a task; see [1] for creating embeddings and [5] in the context of compressed sensing. (However, as we describe later, this work does not actually search in a way that yields realistic outputs as intended.) In this work, we focus on GANs, as recent work in this area has resulted in the highest quality image-generation among unsupervised models.

Regardless of its architecture, let the generator be called G, and let the latent space be $\mathcal{L}$. Ideally, we could approximate $\mathcal{M}$ by the image of G, which would allow us to rephrase the problem above as the following: find a latent vector $z \in \mathcal{L}$ with

$$\|DS(G(z)) - I_{LR}\|_p^p \leq \epsilon. \tag{9}$$

Unfortunately, in most generative models, simply requiring that $z \in \mathcal{L}$ does not guarantee that $G(z) \in \mathcal{M}$; rather, such methods use an imposed prior on $\mathcal{L}$. In order to ensure $G(z) \in \mathcal{M}$, we must be in a region of $\mathcal{L}$ with high probability under the chosen prior. One idea to encourage the latent to be in the region of high probability is to add a loss term for the negative log-likelihood of the prior. In the case of a Gaussian prior, this takes the form of l_2 regularization. Indeed, this is how the previously mentioned work [5] attempts to address this issue. However, this idea does not actually accomplish the goal. Such a penalty forces vectors towards 0, but most of the mass of a high-dimensional Gaussian is located near the surface of a sphere of radius $\sqrt{d}$ (see [22]). To get around this, we observed that we could replace the Gaussian prior on $\mathbb{R}^d$ with a uniform prior on $\sqrt{d}\mathbb{S}^{d-1}$. This approximation can be used for any method with high dimensional spherical Gaussian priors.

We can let $\mathcal{L}' = \sqrt{d}S^{d-1}$ (where $S^{d-1} \subset \mathbb{R}^d$ is the unit sphere in d dimensional Euclidean space) and reduce the problem above to finding a $z \in \mathcal{L}'$ that satisfies Equation (9). This reduces the problem from gradient descent in the entire latent space to projected gradient descent on a sphere.

4. Experiments

We designed various experiments to assess our method. We focus on the popular problem of face hallucination, enhanced by recent advances in GANs applied to face generation. In particular, we use Karras *et al.*'s pretrained Face StyleGAN (trained on the Flickr Face HQ Dataset, or FFHQ) [12]. For each experiment, we used 100 steps of spherical gradient descent with a learning rate of 0.4 starting with a random initialization. Each image was therefore generated in $\sim$5 seconds on a single NVIDIA V100 GPU.

4.1. Data

We evaluated our procedure on the well-known high-resolution face dataset CelebA HQ. (Note: this is not to be confused with CelebA, which is of substantially lower resolution.) We performed these experiments using scale factors of $64\times$, $32\times$, and $8\times$. For our qualitative comparisons, we upscale at scale factors of both $8\times$ and $64\times$, i.e., from 16×16 to 128×128 resolution images and 1024×1024 resolution images. The state-of-the-art for face super-resolution in the literature prior to this point was limited to a maximum of $8\times$ upscaling to a resolution of 128×128, thus making it impossible to directly make quantitative comparisons at high resolutions and scale factors. We followed the traditional approach of training the supervised methods on CelebA HQ. We tried comparing with supervised methods trained on FFHQ, but they failed to generalize and yielded very blurry and distorted results when evaluated on CelebA HQ; therefore, in order to compare our method with the best existing methods, we elected to train the supervised models on CelebA HQ instead of FFHQ.

4.2. Qualitative Image Results

Figure 5 shows qualitative results to demonstrate the visual quality of the images from our method. We observe levels of detail that far surpass competing methods, as exemplified by certain high frequency regions (features like eyes or lips). More examples and full-resolution images are in the appendix.

4.3. Quantitative Comparison

Here we present a quantitative comparison with state-of-the-art face super-resolution methods. Due to constraints on the peak resolution that previous methods can handle, evaluation methods were limited, as detailed below.

We conducted a mean-opinion-score (MOS) test as is common in the perceptual super-resolution literature [15, 13]. For this, we had 40 raters examine images upscaled by 6 different methods (nearest-neighbors, bicubic, FSRNet, FSRGAN, and our PULSE). For this comparison, we used a scale factor of 8 and a maximum resolution of 128×128, despite our method's ability to go substantially higher, due

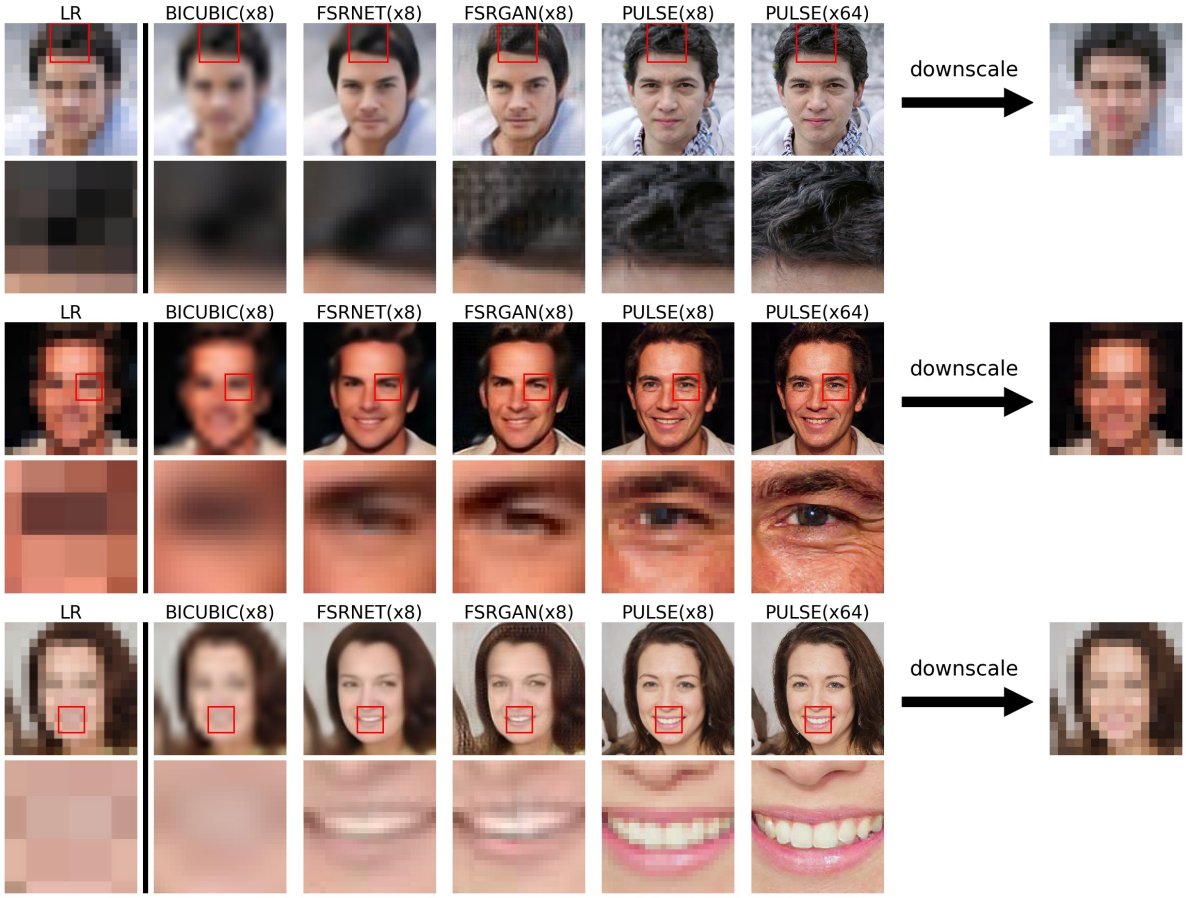

Figure 5. Comparison of PULSE with bicubic upscaling, FSRNet, and FSRGAN. In the first image, PULSE adds a messy patch in the hair to match the two dark diagonal pixels visible in the middle of the zoomed in LR image.

HR	Nearest	Bicubic	FSRNet	FSRGAN	PULSE
3.74	1.01	1.34	2.77	2.92	**3.60**

Table 1. MOS Score for various algorithms at 128×128. Higher is better.

to this being the maximum limit for the competing methods. After being exposed to 20 examples of a 1 (worst) rating exemplified by nearest-neighbors upsampling, and a 5 (best) rating exemplified by high-quality HR images, raters provided a score from 1-5 for each of the 240 images. All images fell within the appropriate $\epsilon = 1e-3$ for the downscaling loss. The results are displayed in Table 1.

PULSE outperformed the other methods and its score approached that of the HR dataset. Note that the HR's 3.74 average image quality reflects the fact that some of the HR images in the dataset had noticeable artifacts. All pairwise differences were highly statistically significant ($p < 10^{-5}$ for all 15 comparisons) by the Mann-Whitney-U test. The results demonstrate that PULSE outperforms current methods in generating perceptually convincing images that downscale correctly.

HR	Nearest	Bicubic	PULSE
3.90	12.48	7.06	**2.47**

Table 2. NIQE Score for various algorithms at 1024×1024. Lower is better.

To provide another measure of perceptual quality, we evaluated the Naturalness Image Quality Evaluator (NIQE) score [16], previously used in perceptual super-resolution [10, 4, 24]. This no-reference metric extracts features from images and uses them to compute a perceptual index (lower is better). As such, however, it only yields meaningful results at higher resolutions. This precluded direct comparison with FSRNet and FSRGAN, which produce images of at most 128×128 pixels.

We evaluated NIQE scores for each method at a resolution of 1024×1024 from an input resolution of 16×16, for a scale factor of 64. All images for each method fell within the appropriate $\epsilon = 1e-3$ for the downscaling loss. The results are in Table 2. PULSE surpasses even the CelebA HQ images in terms of NIQE here, further showing the perceptual quality of PULSE's generated images. This is possible as NIQE is a no-reference metric which solely consid-

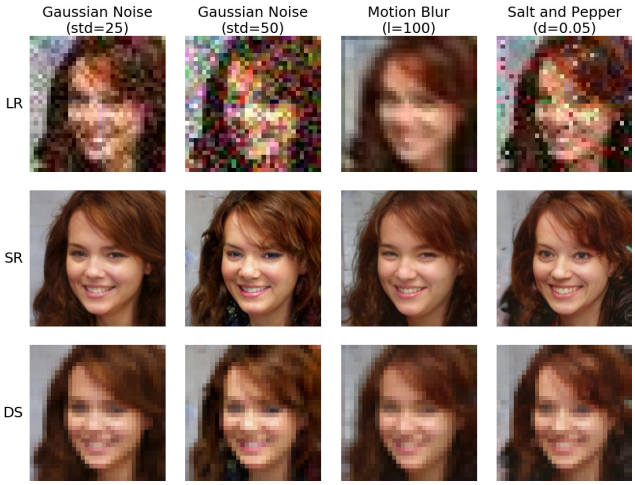

Figure 6. (x32) We show the robustness of PULSE under various degradation operators. In particular, these are downscaling followed by Gaussian noise (std=$25, 50$), motion blur in random directions with length 100 followed by downscaling, and downscaling followed by salt-and-pepper noise with a density of 0.05.

ers perceptual quality; unlike reference metrics like PSNR, performance is not bounded above by that of the HR images typically used as reference.

4.4. Image Sampling

As referenced earlier, we initialize the point we start at in the latent space by picking a random point on the sphere. We found that we did not encounter any issues with convergence from random initializations. In fact, this provided us one method of creating many different outputs with high-level feature differences: starting with different initializations. An example of the variation in outputs yielded by this process can be observed in Figure 3.

Furthermore, by utilizing a generative model with inherent stochasticity, we found we could sample faces with fine-level variation that downscale correctly; this procedure can be repeated indefinitely. In our implementation, we accomplish this by resampling the noise inputs that StyleGAN uses to fill in details within the image.

5. Robustness

The main aim of our algorithm is to perform perceptually realistic super-resolution with a known downscaling operator. However, we find that even for a variety of *unknown* downscaling operators, we can apply our method using bicubic downscaling as a stand-in for more substantial degradations applied–see Figure 6. In this case, we provide only the degraded low-resolution image as input. We find that the output downscales approximately to the *true*, non-noisy LR image (that is, the bicubically downscaled HR) rather than to the degraded LR given as input. This is desired behavior, as we would not want to create an image that matches the additional degradations. PULSE thus implicitly denoises images. This is due to the fact that we restrict the outputs to only realistic faces, which in turn can only downscale to reasonable LR faces. Traditional supervised networks, on the other hand, are sensitive to added noise and changes in the domain and must therefore be explicitly trained with the noisy inputs (e.g., [3]).

6. Discussion and Future Work

Through these experiments, we find that PULSE produces perceptually superior images that also downscale correctly. PULSE accomplishes this at resolutions previously unseen in the literature. All of this is done with unsupervised methods, removing the need for training on paired datasets of LR-HR images. The visual quality of our images as well as MOS and NIQE scores demonstrate that our proposed formulation of the super-resolution problem corresponds with human intuition. Starting with a pre-trained GAN, our method operates only at test time, generating each image in about 5 seconds on a single GPU.

One reasonable fear when using GANs for this purpose may be that while they generate sharp images, they need not cover the whole distribution as, e.g., flow-based models must. However, we did not observe any practical manifestation of this in our experiments. Advances in generative modeling will allow us to cover larger distributions (instead of just faces), and generate higher resolution images.

Another potential concern that may arise when considering this unsupervised approach is the case of an unknown downscaling function. In this work, we focused on the most prominent SR use case: on bicubically downscaled images. In fact, in many use cases, the downscaling function is either known analytically (e.g., bicubic) or is a (known) function of hardware. However, methods have shown that the degradations can be estimated in entirely unsupervised fashions for arbitrary LR images (that is, not necessarily those which have been downscaled bicubically) [6, 25]. Through such methods, we can retain the algorithm's lack of supervision; integrating these is an interesting topic for future work.

7. Conclusions

We have established a novel methodology for image super-resolution as well as a new problem formulation. This opens up a new avenue for super-resolution methods along different tracks than traditional, supervised work with CNNs. The approach is not limited to a particular degradation operator seen during training, and it always maintains high perceptual quality.

Acknowledgments: Funding was provided by the Lord Foundation of North Carolina and the Duke Department of Computer Science. Thank you to the Google Cloud Platform research credits program.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[2] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 372–379. IEEE, 2000.

[3] Yijie Bei, Alexandru Damian, Shijia Hu, Sachit Menon, Nikhil Ravi, and Cynthia Rudin. New techniques for preserving global structure and denoising with low information loss in single-image super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 874–881. IEEE Computer Society, 2018.

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 PIRM challenge on perceptual image super-resolution. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[5] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[6] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 185–200, 2018.

[7] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2018.

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 184–199, 2014.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[10] Seokhwa Jeong, Inhye Yoon, and Joonki Paik. Multi-frame example-based super-resolution using locally directional self-similarity. *IEEE Transactions on Consumer Electronics*, 61(3):353–358, 2015.

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2018. Appeared at the 6th Annual International Conference on Learning Representations.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[13] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. In *Proceedings of the 30th British Machine Vision Conference (BMVC)*, 2019.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[16] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, March 2013.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[18] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[20] Amanjot Singh and Jagroop Singh Sidhu. Super resolution applications in modern digital image processing. *International Journal of Computer Applications*, 150(2):0975–8887, 2016.

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] Roman Vershynin. *Random Vectors in High Dimensions*, page 38–69. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[23] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 106(1):9–30, 2014.

[24] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[25] Tianyu Zhao, Changqing Zhang, Wenqi Ren, Dongwei Ren, and Qinghua Hu. Unsupervised degradation learning for single image super-resolution. *arXiv preprint arXiv:1812.04240*, 2018.