PULSE: Supplementary Material

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

1. Appendix A: Additional Figures

Here we provide further samples of the output of our super-resolution method for illustration in Figure 1. These results were obtained with $\times 8$ scale factor from an input of resolution 16×16. This highlights our method's capacity to illustrate detailed features that we did not have space to show in the main paper, such as noses and a wider variety of eyes. We also present additional examples depicting PULSE's robustness to additional degradation operators in Figure 2 and some randomly selected generated samples in Figures 3, 4, 5, and 6.

2. Appendix B: Implementation Details

2.1. StyleGAN

In order to generate experimental results using our method, we had to pick a pretrained generative model to work with. For this we chose StyleGAN due to its state-of-the-art performance on high resolution image generation.

StyleGAN consists of two components: first, a mapping network $M : \mathbb{R}^{512} \to \mathbb{R}^{512}$, a tiling function $T : \mathbb{R}^{512} \to \mathbb{R}^{18\times 512}$, and a synthesis network $S : \mathbb{R}^{18\times 512} \times \mathcal{N} \to \mathbb{R}^{1024\times 1024}$ where $\mathcal{N}$ is collection of Euclidean spaces of varying dimensions representing the domains of each of the noise vectors fed into the synthesis network. To generate images, a vector z is sampled uniformly at random from the surface of the unit sphere in $\mathbb{R}^{512}$. This is transformed into another 512-dimensional vector by the mapping network, which is replicated 18 times by the tiling function T. The new 18×512 dimensional vector is input to the synthesis network which uses it to generate a high-resolution, 1024×1024 pixel image. More precisely, the synthesis network consists of 18 sequential layers, and the resolution of the generated image is doubled after every other layer; each of these 18 layers is re-fed into the 512-dimensional output of the mapping network, hence the tiling function. The synthesis network also takes as input noise sampled from the unit Gaussian, which it uses to stochastically add details to the generated image. Formally, η is sampled from the Gaussian prior on $\mathcal{N}$, at which point the output is obtained by computing $S(T(M(z)), \eta)$.

2.2. Latent Space Embedding

Experimentally, we observed that optimizing directly on $z \in S^{511} \subset \mathbb{R}^{512}$ yields poor results; this latent space is not expressive enough to map to images that downscale correctly. A logical next step would be to use the expanded 18×512-dimensional latent space that the synthesis network takes as input, as noted by Abdal, Qin, and Wonka [1]. By ignoring the mapping network, S can be applied to any vector in $\mathbb{R}^{18\times 512}$, rather than only those consisting of a single 512-dimensional vector repeated 18 times. This expands the expressive potential of the network; however, by allowing the 18 512-dimensional input vectors to S to vary independently, the synthesis network is no longer constrained to the original domain of StyleGAN.

2.3. Cross Loss

For the purposes of super-resolution, such approaches are problematic because they void the guarantee that the algorithm is traversing a good approximation of $\mathcal{M}$, the natural image manifold. The synthesis network was trained with a limited subset of $R^{18\times 512}$ as input; the further the input it receives is from that subset, the less we know about the output it will produce. The downscaling loss, defined in the main paper, is alone not enough to guide PULSE to a realistic image (only an image that downscales correctly). Thus, we want to make some compromise between the vastly increased expressive power of allowing the input vectors to vary independently and the realism produced by tiling the input to S 18 times. Instead of optimizing on downscaling loss alone, we need some term in the loss discouraging straying too far in the latent space from the original domain.

To accomplish this, we introduce another metric, the "cross loss." For a set of vectors $v_1, ..., v_k$, we define the cross loss of $v_1, ..., v_k$ to be

$$CROSS(v_1, ..., v_k) = \sum_{i<j} |v_i - v_j|_2^2$$

The cross loss imposes a penalty based on the Euclidean

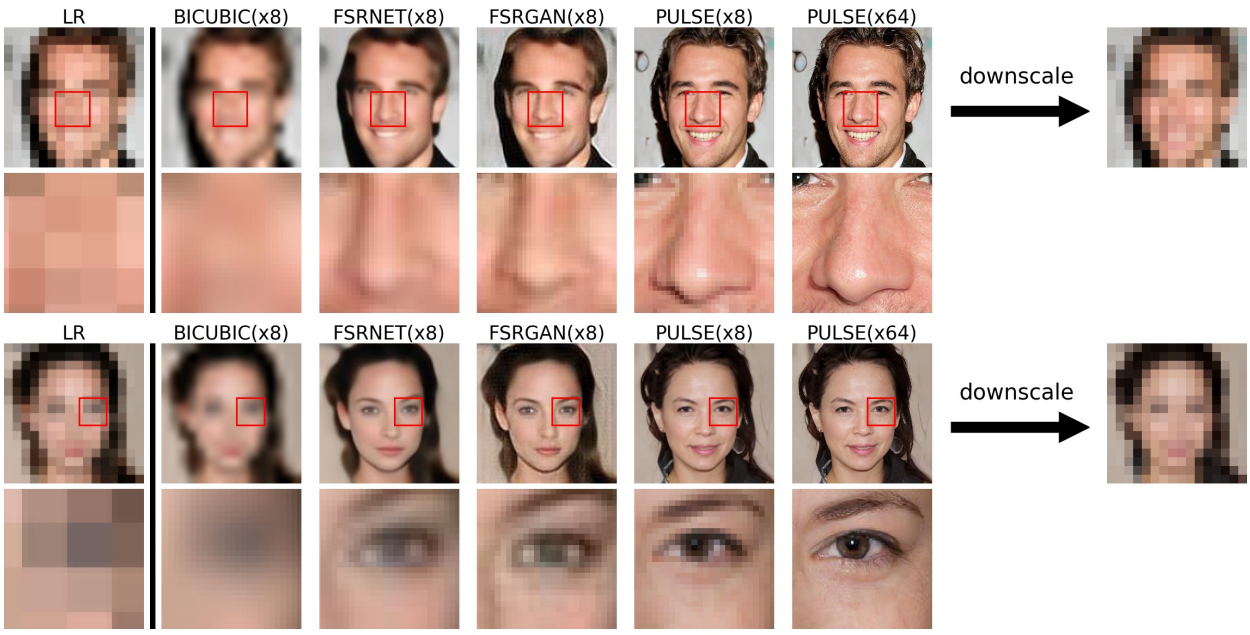

Figure 1. Further comparison of PULSE with bicubic upscaling, FSRNet, and FSRGAN.

distance between every pair of vectors input to S. This can be considered a simple form of relaxation on the original constraint that the 18 input vectors be exactly equal.

When $v_1, ..., v_k$ are sampled from a sphere, it makes more sense to compare geodesic distances along the sphere. This is the approach we used in generating our results. Let $\theta(v, w)$ denote the angle between the vectors v and w. We then define the geodesic cross loss to be

$$GEOCROSS(v_1, ..., v_k) = \sum_{i<j} \theta(v_i, v_j)^2$$

.

Empirically, by allowing the 18 input vectors to S to vary while applying the soft constraint of the (geo)cross loss, we can increase the expressive potential of the network without large deviations from the natural image manifold.

2.4. Approximating the input distribution of S

StyleGAN begins with a uniform distribution on $S^{511} \subset \mathbb{R}^{512}$, which is pushed forward by the mapping network to a transformed probability distribution over $\mathbb{R}^{512}$. Therefore, another requirement to ensure that $S([v_1, ..., v_{18}], \eta)$ is a realistic image is that each v_i is sampled from this push-forward distribution. While analyzing this distribution, we found that we could transform this back to a distribution on the unit sphere without the mapping network by simply applying a single linear layer with a leaky-ReLU activation– an entirely invertible transformation. We therefore inverted this function to obtain a sampling procedure for this distribution. First, we generate a latent w from S^{511}, and then apply the inverse of our transformation.

2.5. Noise Inputs

The second parameter of S controls the stochastic variation that StyleGAN adds to an image. When the noise is set to 0, StyleGAN generates smoothed-out, detail-free images. The synthesis network takes 18 noise vectors at varying scales, one at each layer. The earlier noise vectors influence more global features, for example the shape of the face, while the later noise vectors add finer details, such as hair definition. Our first approach was to sample the noise vectors before we began traversing the natural image manifold, keeping them fixed throughout the process. In an attempt to increase the expressive power of the synthesis network, we also tried to perform gradient descent on both the latent input and the noise input to S simultaneously, but this tended to take the noise vectors out of the spherical shell from which they were sampled and produced unrealistic images. Using a standard Gaussian prior forced the noise vectors towards 0 as mentioned in the main paper. We therefore experimented with two approaches for the noise input:

1. *Fixed noise:* Especially when upsampling from 16×16 to 1024×1024, StyleGAN was already expressive enough to upsample our images correctly and so we did not need to resort to more complicated techniques.

2. *Partially trainable noise:* In order to slightly increase the expressive power of the network, we optimized on the latent and the first 5-7 noise vectors, allowing us to slightly modify the facial structure of the images we generated to better match the LR images while main-

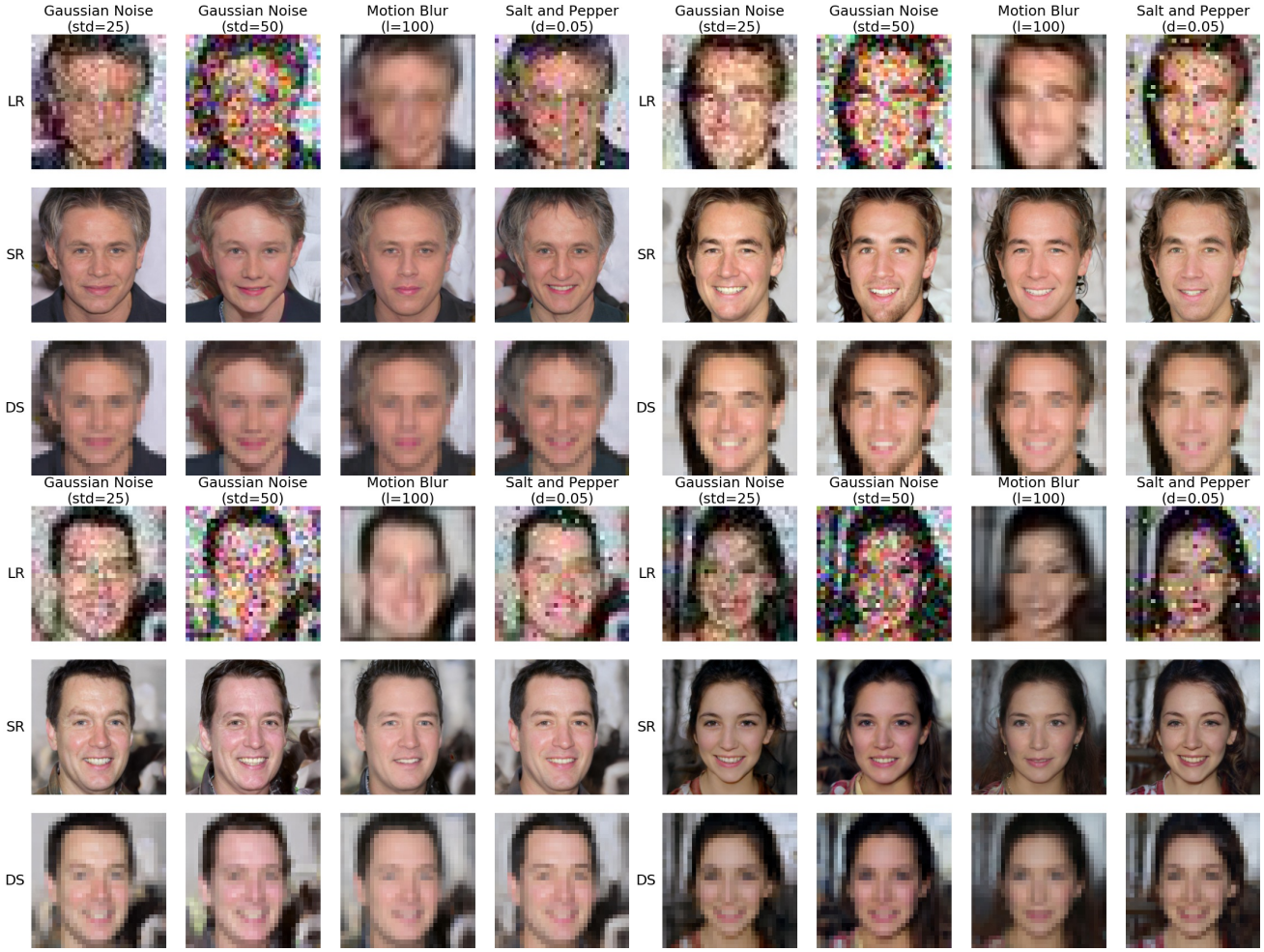

Figure 2. (x32) Additional robustness results for PULSE under additional degradation operators (these are downscaling followed by Gaussian noise (std=$25, 50$), motion blur in random directions with length 100 followed by downscaling, and downscaling followed by salt-and-pepper noise with a density of 0.05.)

	Nearest	Bicubic	FSRNet	FSRGAN	PULSE
PSNR	21.78	23.40	25.93	24.55	22.01
SSIM	0.51	0.63	0.74	0.66	0.55

taining image quality. This was the approach we used to generate the images presented in this paper.

3. Appendix C: Alternative Metrics

For completeness, we provide the metrics of PSNR and SSIM here. These results were obtained with $\times 8$ scale factor from an input of resolution 16×16. Note that we explicitly do not aim to optimize on this pixel-wise average distance from the high-resolution image, so these metrics do not have meaningful implications for our work.

4. Appendix D: Robustness

Traditional supervised approaches using CNNs are notoriously sensitive to tiny changes in the input domain, as any perturbations are propagated and amplified through the network. This caused some problems when trying to train FSRNET and FSRGAN on FFHQ and then test them on CelebAHQ. However, PULSE never feeds an LR image through a convolutional network and never applies filters to the LR input images. Instead, the LR image is only used as a term in the *downscaling loss*. Because the generator is not capable of producing "noisy" images, it will seek an SR image that downscales to the closest point on the LR natural image manifold to the noisy LR input. This means that PULSE outputs an SR image that downscales to the projection of the noisy LR input onto the LR natural image manifold, and if the noise is not too strong, this should be close to the "true" unperturbed LR . This may explain why PULSE had

no problems when applied to different domains, and why we could demonstrate robustness when the low resolution image was degraded with various types of noise.

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. 1

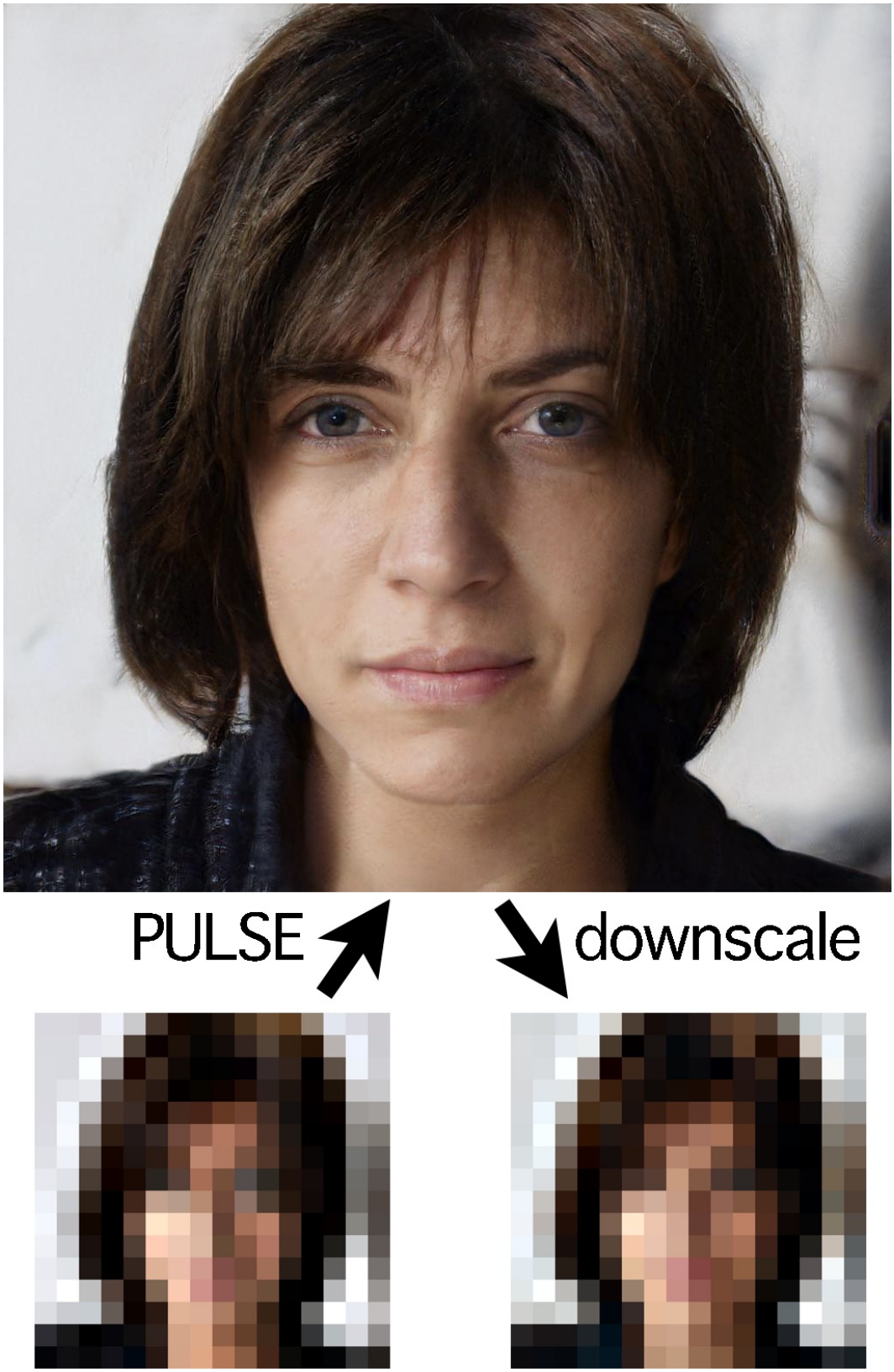

Figure 3. (64x) Sample 1

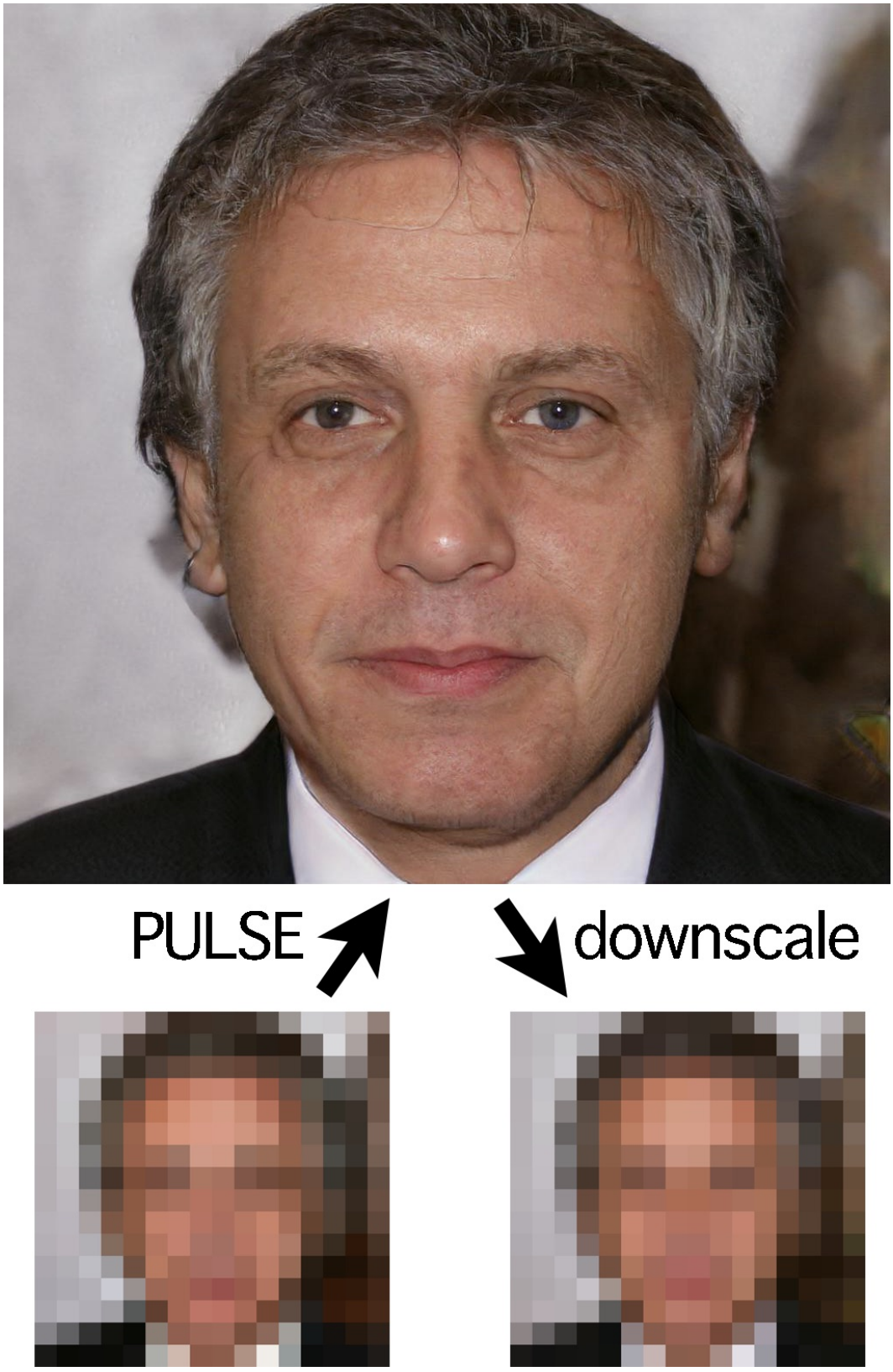

Figure 4. (64x) Sample 2

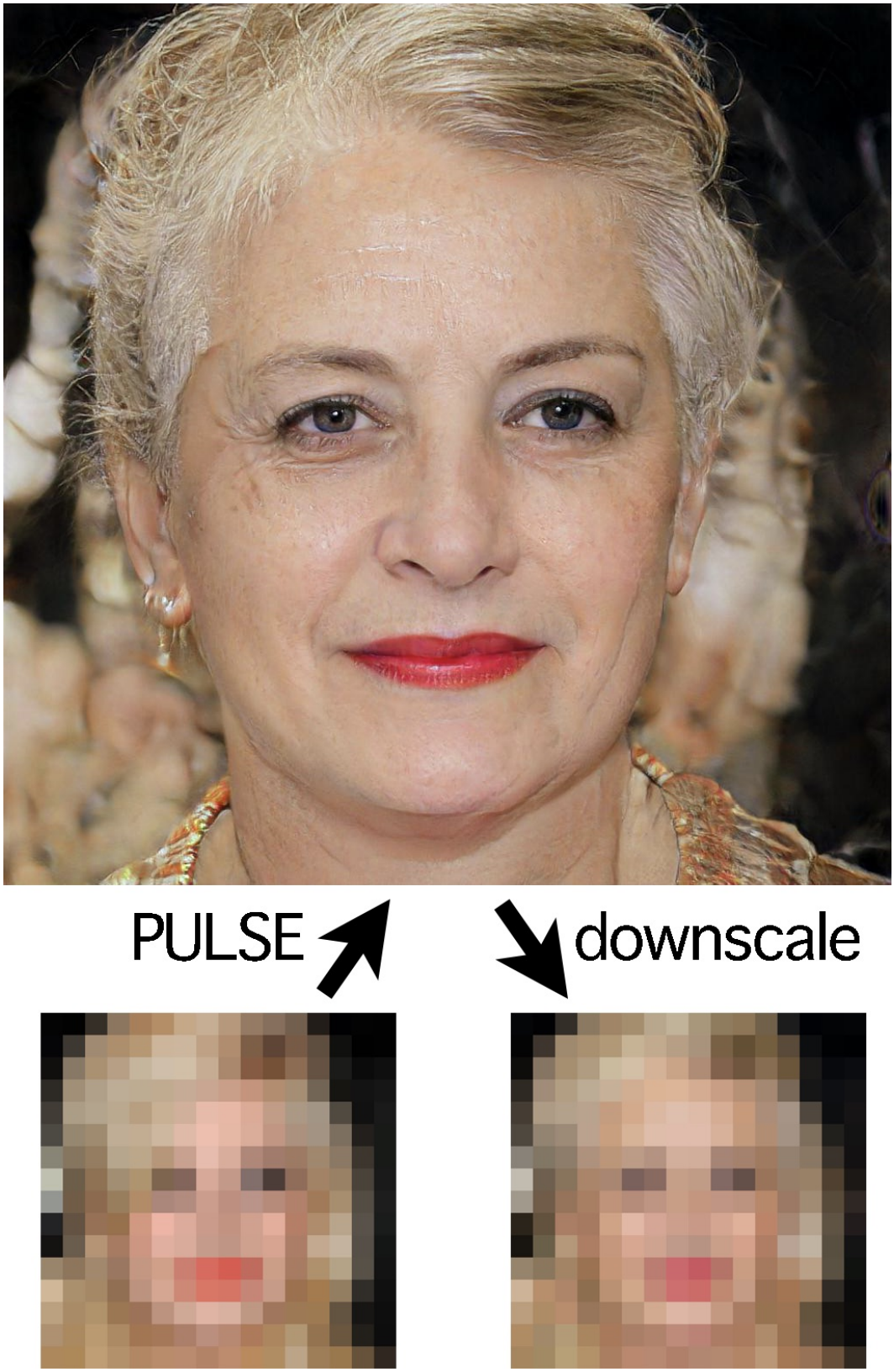

Figure 5. (64x) Sample 3

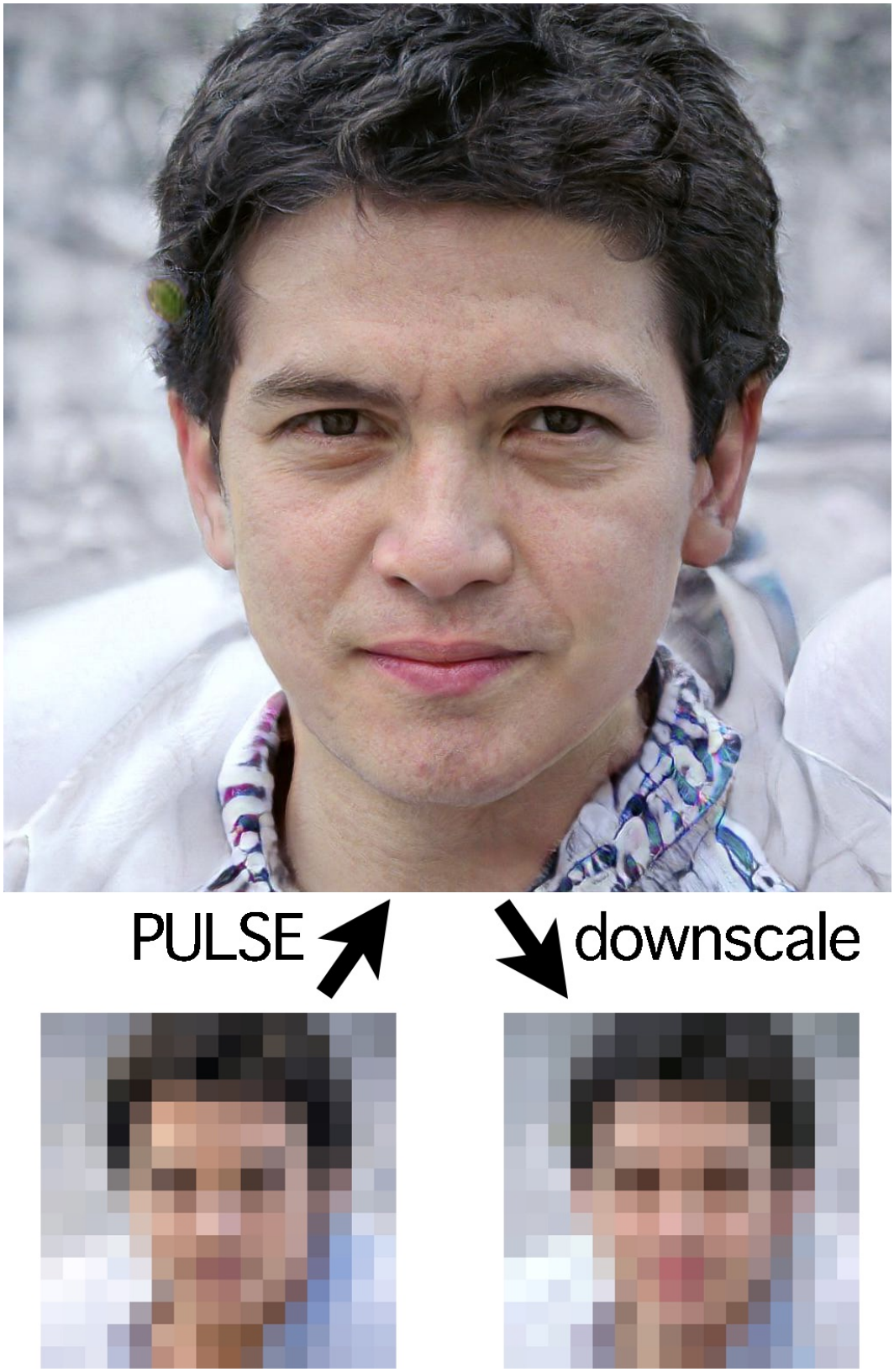

Figure 6. (64x) Sample 4