Super-Resolution through StyleGAN Regularized Latent Search

Marzieh Gheisari, Auguste Genovesio

Super-Resolution through StyleGAN Regularized Latent Search

Marzieh Gheisari
IBENS, École Normale Supérieure
Paris, France
gheisari@bio.ens.psl.eu

Auguste Genovesio
IBENS, École Normale Supérieure
Paris, France
auguste.genovesio@ens.psl.eu

Abstract

This paper addresses the problem of super-resolution: constructing a highly resolved (HR) image from a low resolved (LR) one. Previous methods to solve this task were mainly supervised and consisted of approaching a real HR image from its down-sampled LR version, using a pixel-wise reconstruction loss. This often results in blurred images. Recognizing that the problem is ill-posed, recent self-supervised alternative approaches proposed to constrain it by searching the latent space of a StyleGAN pre-trained on HR images, for the image that best downscales to the query LR image. Here we demonstrate that such an optimization process can still shift out of the StyleGAN image manifold and produce out-of-domain images containing artifacts. We thus introduce a new regularizer to facilitate the search in the latent space, by leveraging a normalizing flow model for effective density estimation, which serves as a powerful image prior. This prior is incorporated into the reconstruction process, ensuring that the inverted code lies in the original latent space and thus allowing us to achieve a good trade-off between fidelity and realness.

1 Introduction

Super-resolution aims to reconstruct an unknown High Resolution (HR) image $\mathbf{x} \in \mathbb{R}^{n \times n}$ from a Low Resolution (LR) image $\mathbf{y} \in \mathbb{R}^{m \times m}$, related to one another by a down-sampling process described by $\mathbf{y} = \mathsf{D}(\mathbf{x}) + \delta$ with $\mathsf{D} : \mathbb{R}^{n \times n} \to \mathbb{R}^{m \times m}$ a down-sampling non-invertible forward operator and δ an independent noise with distribution p_δ.

As with many data generation tasks, super-resolution has largely benefited in recent years from the advent of GANs and especially from coupled image translation where a direct coupling of HR and LR images can be learnt [7, 17, 18]. However these approaches tend to produce fuzzy output, due to the averaging effect introduced by the pixel-wise reconstruction loss of the HR image, or alternatively tend to degrade the accuracy of the reconstruction, due to the adversarial loss. Importantly the problem as such is ill-posed as, for a non-invertible forward operator D with $m < n$, there are infinitely many HR images that match a given LR image. Thus the reconstruction procedure must be further constrained by prior information to better define the objective and lead to a stable solution.

One such prior consisted of considering the reconstructed HR image to be part of a given domain. First the distribution of HR images is learnt in an unsupervised fashion, thanks to a GAN, then the latent space of this trained GAN is searched to find the HR image that, once down-sampled, is the closest to the LR image. This idea was first introduced by Bora et *al* [2]. and further improved using StyleGAN [9, 10] by Menon et *al* in PULSE [13] by constraining the search to remain on the image manifold. To this end, PULSE and a few other studies [20, 19] use an invertible transformation of the intermediate latent space $\mathcal{W}^+$ which includes a leaky rectified linear unit (ReLU) [5] followed by an affine whitening transformation, so that transformed latent vectors approximately followed the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, where d is the dimensionality of the latent space. Sampled vectors are then constrained to lie around a hypersphere with radius $\sqrt{d}$ hypothesizing that most of

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the mass of a high-dimensional Gaussian distribution is located at or near $\sqrt{d}\mathcal{S}^{d-1}$, where $\mathcal{S}^{d-1}$ is the d-dimensional unit hypersphere. Constraining samples to lie in dense area of the StyleGAN style distribution resulted in increased realism of the generated images.

Although the above approach showed major improvements over previous work, it also presents two important caveats that in practice led to image artifacts. First, as we will show later, transforming the intermediate latent space this way does not lead to an accurate standard Gaussian distribution, and prevents proper regularization based on this hypothesis. Second, we argue that a search strictly limited to the spherical surface $\mathcal{S}^{d-1}$ restricts access to the whole variety of images a StyleGAN can generate, thus preventing a close reconstruction of the HR image to be reached.

In this work, we take advantage of normalizing flow to Gaussianize the latent style sample distribution and show that it leads to a much closer standard Gaussian distribution. We then use this revertible transformation to regularize the search in $\mathcal{W}^+$ such that it remains in high density area of the style vector distribution. We then show that the latter produces reconstructed images that are more realistic and that the approach is robust to noise and other image corruptions.

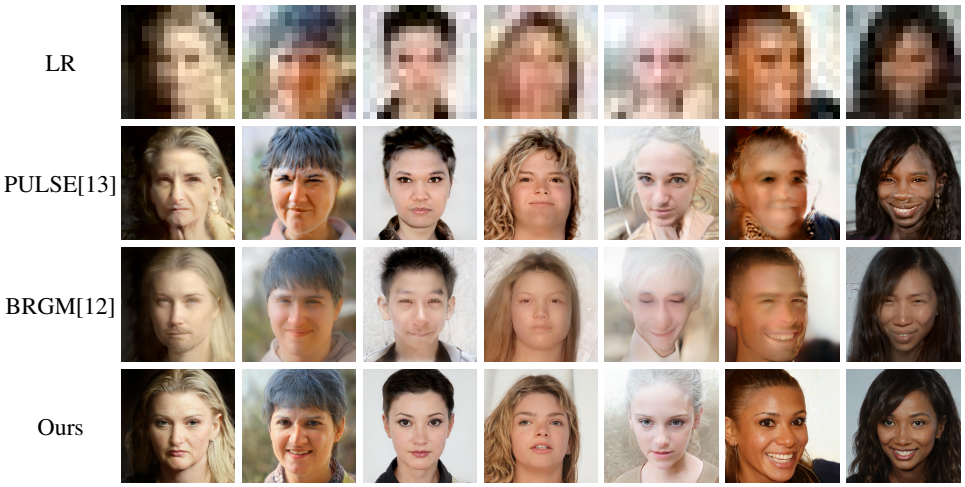

Figure 1: Comparison of the reconstructions of a high resolution face from CelebA

2 Proposed Method

MAP estimation We can formulate the super-resolution problem in terms of MAP estimation [12]. For a given LR image $\mathbf{y}$, we wish to recover the HR image $\mathbf{x}$ as the MAP estimate of the conditional distribution $p_{\mathsf{G}}(\mathbf{x}|\mathbf{y})$:

$$\begin{aligned} \arg\max_{\mathbf{x}} \log p_{\mathsf{G}}(\mathbf{x}|\mathbf{y}) &= \arg\max_{\mathbf{x}} [\log p(\mathbf{y}|\mathbf{x}) + \log p_{\mathsf{G}}(\mathbf{x}) + \log p(\mathbf{y})] \\ &= \arg\max_{\mathbf{x}} [\log p_{\delta}(\mathbf{y} - \mathsf{D}(\mathbf{x})) + \log p_{\mathsf{G}}(\mathbf{x})] \end{aligned} \tag{1}$$

Since the marginal density $\log p(\mathbf{y})$ is constant we drop it and rewrite $p(\mathbf{y}|\mathbf{x})$ as $p_{\delta}(\mathbf{y} - \mathsf{D}(\mathbf{x}))$. Now, the first term in Eq. (1) is the likelihood term which describes the image degradation process D and the second term is the image prior, which describes the manifold of real HR images.

Image prior Let G_s be the synthesis network of a StyleGAN [10] pretrained on the considered image domain. G_s takes as input $\mathbf{w}$, produced by the mapping network, and outputs an image (A.1). A change of variables can be used for a non-invertible mapping and we can thus express the image prior with respect to the latent variables $\mathbf{w}$ this way [3]:

$$\log p_{\mathsf{G}}(\mathsf{G}_s(\mathbf{w})) = \log p_{\mathbf{w}}(\mathbf{w}) + \log |\det J_{\mathsf{G}_s^{-1}}(\mathbf{w})|. \tag{2}$$

The second term can be dropped as the path length regularization in StyleGAN2 implies that the Jacobian determinant is constant for all $\mathbf{w}$. The first term $p_{\mathbf{w}}(\mathbf{w})$ is the image prior we define on $\mathbf{w} \in \mathcal{W}^+$ by:

$$\log p_{\mathbf{w}}(\mathbf{w}) = \lambda_w \mathcal{P}_w + \lambda_g \mathcal{P}_{gaussian} + \lambda_c \mathcal{P}_{cross} \tag{3}$$

where:

- $\mathcal{P}_w$ is a prior that keeps $\mathbf{w}$ in the area of high density in $\mathcal{W}^+$: $\mathcal{P}_w = \frac{1}{L}\sum_{i=1}^{L} \log p_{\mathsf{F}}(\mathbf{w}_i)$, where $p_{\mathsf{F}}(\mathbf{w})$ is estimated by a normalizing flow model F explained in further detail in Appendix A.2.
- $\mathcal{P}_{gaussian}$ is a gaussianization prior. Using the normalizing flow model F, a gaussianized latent vector $\mathbf{w}_n$ can be obtained and a $L2$ regularization applied on it to keep it near the surface of the hypersphere: $\mathcal{P}_{gaussian} = -\frac{1}{L}\sum_{i=1}^{L}(||\mathsf{F}(\mathbf{w})||_2 - \sqrt{d})^2$
- $\mathcal{P}_{cross}$ is a pairwise euclidean distance prior on $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_L] \in \mathcal{W}^+$ that ensures $\mathbf{w} \in \mathcal{W}^+$ remains close to the trained manifold in $\mathcal{W}$: $\mathcal{P}_{cross} = -\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} ||\mathbf{w}_i - \mathbf{w}_j||_2^2$

Optimization In the likelihood term in Eq. (1), we assume that the noise follows a Laplace distribution *i.e.* $\delta \sim Laplace(0, \lambda_l I)$, then the log-density of δ becomes: $\log p_\delta(\delta) = -||\delta||_1 - C$ for a constant C. So the problem in Eq. (1) can be recast as an optimization over $\mathbf{w}$, leading to the final objective function:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\mathbf{y} - \mathsf{D}(\mathsf{G}_\theta(\mathbf{w}))||_1 - \log p_{\mathbf{w}}(\mathbf{w}) \tag{4}$$

3 Results

Experimental setup. We used a StyleGAN2 generator [10] pre-trained on the FFHQ dataset [9] that includes 70,000 high quality face images of resolution 1024×1024. For evaluation we used the CelebA-HQ dataset [8] and simulated degraded faces from the HR images using bicubic downsampling. The regularization parameters λ_w, λ_c and λ_g were set to 0.0001, 0.01 and 0.0002, respectively. We used an Adam optimizer over 500 iterations with learning rate of 0.5 and initialized the search by the mean of 100,000 randomly generated latent vectors. We compared our algorithm with a simple Bicubic upsampling and two state-of-the-art image reconstruction methods based on StyleGAN inversion: PULSE [13] and BRGM [12]. We assumed PULSE had already shown superiority over former supervised super-resolution approaches. We provide additional results in Appendix B.

Quality of the Gaussianization process. We generated 5000 random samples $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and their associated style vector $\mathbf{w} = \mathsf{G}_m(\mathbf{z})$, then gaussianized distribution of $\mathcal{W}$ by PULSE and our method. We then computed the squared norm for all of these samples (see Fig. 2). As expected, the squared norm of the standard gaussian distribution $\mathcal{Z}$ approximately follows $||\mathbf{z}_n||_2^2 \sim \chi_d^2$ and thus forms a narrow distribution around $d = 512$ while the squared norm of the untransformed distribution $\mathcal{W}$ does not, which is inconsistent with the prior assumption held by BRGM [12]. Furthermore, PULSE[13] appears to produce a wider squared norm distribution while ours approaches the squared norm distribution of $\mathcal{Z}$ more closely.

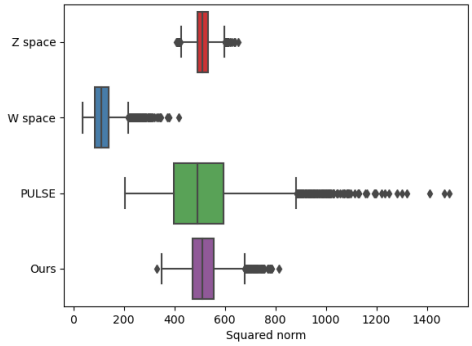

Figure 2: Evaluation of the gaussianization prior.

Method	FID↓	NIQE↓	PSNR↑	SSIM↑
Bicubic	51.38	14.17	17.62	**0.62**
PULSE [13]	5.15	5.09	**19.55**	0.58
BRGM [12]	5.41	4.98	15.11	0.52
Ours	**4.18**	**3.93**	19.23	0.57
GT	-	3.75	∞	1

Table 1: Quantitative comparison on CelebA for 64x super-resolution.

Preservation of domain integrity. Fig. 1 shows that the competing methods fail to produce reasonable facial details. They tend to produce an image that strictly downscales to the LR at the cost of generating distorted facial structures. In comparison, our approach maintains the overall structure of the face and generates visually realistic and faithful details (*e.g.* , eye, eyebrow, teeth, mouth, hair, etc.). To quantitatively assess this gain in performance, we used FID [6] to measure the discrepancy between the original HR image distribution and the reconstructed one, and NIQE [14] as a perceptual

metric. As expected, the scores of FID and NIQE listed in Tab. 1 show that our method performs significantly better both at reproducing the distribution of real HR images and at producing images that have better perceptual quality. This is in accordance with our strategy to impose the image so it belongs to a given domain.

Quality of the reconstruction. While the ill-posed nature of the super-resolution task induces the existence of many plausible solutions, we nevertheless used PSNR and SSIM to evaluate the reconstruction closeness to the real HR image. In Tab. 1 the corresponding scores show that our reconstruction is comparable to those of our competitors. However, the bicubic interpolator achieves the best results while it cannot restore facial details (see Appendix B, Fig. 5). The latter may illustrate the lack of satisfactory image metrics when it comes to pixel-wise image comparison.

Robustness to artifacts. In opposition to supervised methods that are sensitive to the input image domain, this approach is not restricted to a particular degradation operator that used during training. To evaluate this aspect we applied additional degradation operators to a bicubic downscaled image DS before reconstruction. Fig. 3 shows that the reconstruction is close to the DS image. The last legitimates the use of the bicubic downscaling operator during training instead of more complicated specific degradation.

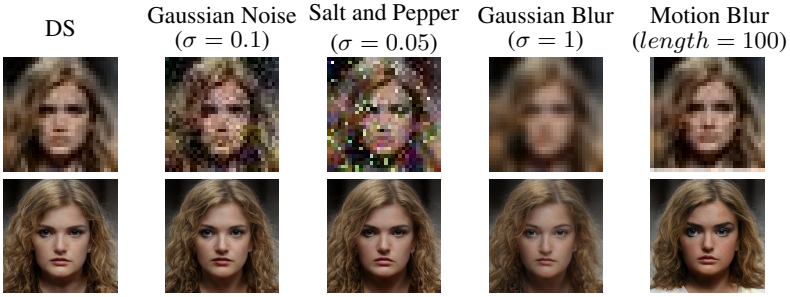

Figure 3: Robustness evaluation: degradation includes downscaling followed by corresponding operations, except for the last column, which consists of motion blur followed by downscaling.

Diversity of the output. As shown by Fig. 4, our method can produce multiple consistent and realistic HR images from a single LR image compared to PULSE. We hypothesize that this could be explained by the fact that PULSE limits the solutions to lie on hypersphere which does not cover the whole distribution of images learnt during StyleGAN training.

Figure 4: Generating multiple solutions for a given LR image. Apart from the first column that displays Low Resolution and Ground Truth images, top row is PULSE and bottom row is our approach.

4 Conclusion

Super-resolved images reconstructed by existing self-supervised approaches, exploiting a pretrained style-based generative model, seem to suffer from a lack of realism, especially when the original image is out of domain. With this work, we address this issue by introducing a new regularization of the latent space exploration. Our approach leverages a normalizing flow model, providing a more robust image prior to ensure that the latent code remains in the original generative model latent space. Doing so enables us to mitigate the fidelity-realness trade-off. We performed extensive experiments demonstrating that our method can generate high quality face images with clear facial details from severely degraded ones, outperforming prior work.

Acknowledgments and Disclosure of Funding

This work was supported by ANR–10–LABX–54 MEMOLIFE and ANR–10 IDEX 0001 –02 PSL* Université Paris, ANR Visualpseudotime and was granted access to the HPC resources of IDRIS under the allocation 2020- AD011011495 made by GENCI.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Int. Conf. Mach. Learn.*, pages 537–546. PMLR, 2017.

[3] Milan Cvitkovic and Günther Koliander. Minimal achievable sufficient statistic learning. In *ICML*, pages 1465–1474. PMLR, 2019.

[4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1125–1134, 2017.

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Int. Conf. Learn. Represent.*, 2018.

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019.

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020.

[11] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *PAMI*, 43(11):3964–3979, 2020.

[12] Razvan Marinescu, Daniel Moyer, and Polina Golland. Bayesian image reconstruction using deep generative models. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[13] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2437–2445, 2020.

[14] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2012.

[15] George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

[16] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[17] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8798–8807, 2018.

[18] 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 *Eur. Conf. Comput. Vis.*, pages 0–0, 2018.

[19] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020.

[20] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020.

Appendices

A Prior work

A.1 Style-based Generative Adversarial Networks

StyleGAN models are well known for generating highly realistic images. The StyleGAN architecture consists of two sub-networks: a mapping network $\mathsf{G}_m : \mathbb{R}^d \to \mathbb{R}^d$, and an L-layer synthesis network $\mathsf{G}_s : \mathbb{R}^{L\times d} \to \mathbb{R}^n$. The mapping network maps a sample $\mathbf{z} \in \mathbb{R}^d$ from a standard normal distribution to a vector $\mathbf{w} \in \mathcal{W}$. The $d \times L$ dimensional vector $\mathbf{w} \in \mathcal{W}^+$ containing L copies of $\mathbf{w}$ is fed to the L-layer synthesis network G_s. The i-th copy of $\mathbf{w}$ represents the input to the i-th layer of G_s, which controls the i-th level of detail in the generated image. In addition to these, G_s also takes as input a collection of latent noise vectors η that control minor stochastic variations of the generated image at each scale. The ability of a StyleGAN to control features of the generated image at different scales is partially due to this architecture, and partially due to the style-mixing regularization occurring during training [9, 10]. In addition to these basic characteristics, StyleGAN2 introduces path-length regularization, which helps in reducing the representation error. Prior works demonstrated that performing latent search in the extended latent space $\mathcal{W}^+$ led to a more accurate reconstruction but at cost of a loss of editability [19, 1]. In PULSE, GEOCROSS is introduced as a penalty term to force the embedding in the extended latent space $\mathcal{W}^+$ to remain close to the latent space $\mathcal{W}$, which in turn promotes the embedded object to be close to the range of the generator network G_s with the latent space $\mathcal{W}$.

A.2 Learning the prior with Normalizing Flow

Using a sequence of invertible mappings, a Normalizing Flow $\mathsf{F} : \mathbb{R}^d \to \mathbb{R}^d$ is a transformation of an unknown complex distribution into a simple probability distribution that is easy to sample from and whose density is easy to evaluate such as standard Gaussian [11].

Let $\mathbf{z} = \mathsf{F}(\mathbf{w})$ with probability density function $p(\mathbf{z})$. Using the change-of-variable formula, we can express the log-density of $\mathbf{w}$ by[15]:

$$\log p_{\mathsf{F}}(\mathbf{w}) = \log p(\mathbf{z}) + \log |\det J_{\mathsf{F}}(\mathbf{w})|, \qquad \mathbf{z} = \mathsf{F}(\mathbf{w}) \tag{5}$$

where $J_{\mathsf{F}}(\mathbf{w})$ is the Jacobian of F evaluated at $\mathbf{w}$.

In practice the Jacobian determinant in Eq. (5) should be easy to compute, so that the density $p_{\mathsf{F}}(\mathbf{w})$ can be evaluated. Furthermore, as a generative model, the invertibility of F allows new samples $\mathbf{w} = \mathsf{F}^{-1}(z)$ to be drawn through sampling from the base distribution.

In the literature, several flow models were proposed, such as Real Non-volume Preserving Flow (RealNVP) [4] and Masked Auto-regressive Flow (MAF) [16]. We use MAF, as it tends to work better than RealNVP for density estimation tasks. Five flow blocks are used in our model and all hidden dimensions are set to 1024.

B Additional results

B.1 Ablation Study

In order to better understand the role of each image prior terms, ablation experiments are presented in Fig. 8. The first variant is denoted "w/o $\mathrm{P}_{\mathrm{gaussian}}$", *i.e.* , the image prior does not include the gaussian prior term $\mathcal{P}_{gaussian}$. Similarly, the second and third variants are denoted "w/o P_{w}", "w/o $\mathrm{P}_{\mathrm{cross}}$", which refer to the suppression of $\mathcal{P}_w$ and $\mathcal{P}_{cross}$ terms, respectively. The last column denoted "w/o W^+" refers to an optimization in the $\mathcal{W}$ space rather than in the $\mathcal{W}^+$.

To evaluate the three variants, we use the same set of parameters used in Sec. 3. One can see that "w/o $\mathrm{P}_{\mathrm{gaussian}}$" achieves better qualitative measures than its variants. We can see that w/o W^+ can generate realistic HR face images; however, the appearance of the face is quite different from the ground truth. This is because the latent space $\mathcal{W}$ is less expressive than $\mathcal{W}^+$, reducing the range of images that can be reconstructed with high fidelity. Moreover, it results in less realistic face images when $\mathrm{P}_{\mathrm{cross}}$ is discarded. This implies that the cross prior plays an import role in generating realistic facial

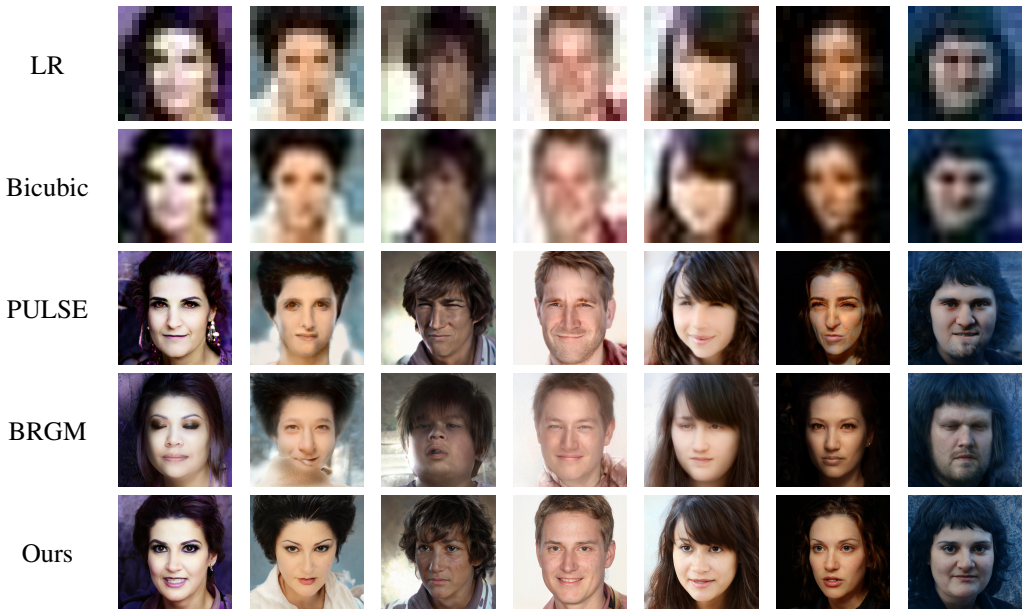

Figure 5: Additional comparative qualitative examples.

details. Overall, compared to its variants, our approach shows superior performance, demonstrating the effectiveness of our image prior.

Finally, we analyze the effect of the strength of the regularizer parameters λ_w and λ_g on the quality of reconstruction in Fig. 9. To do this, we consider two variants, in which we keep only the corresponding prior and unplug the rest, *e.g.* when we measure the effect of λ_w (*i.e.* $\lambda_w \uparrow$), λ_g and λ_c are set to zero. We can see that while both variants provide excellent reconstruction results for some parameter values, keeping only the prior $\mathcal{P}_w$ produces a reconstructed image that is closer to the ground-truth. This result is the consequence of the search not being restricted to the regions near the hyper-sphere. On the other hand, when λ_w is very large, the reconstruction quality is higher. We can thus control the trade-off between realism (with respect to the dataset the model was trained on) and reconstruction quality.

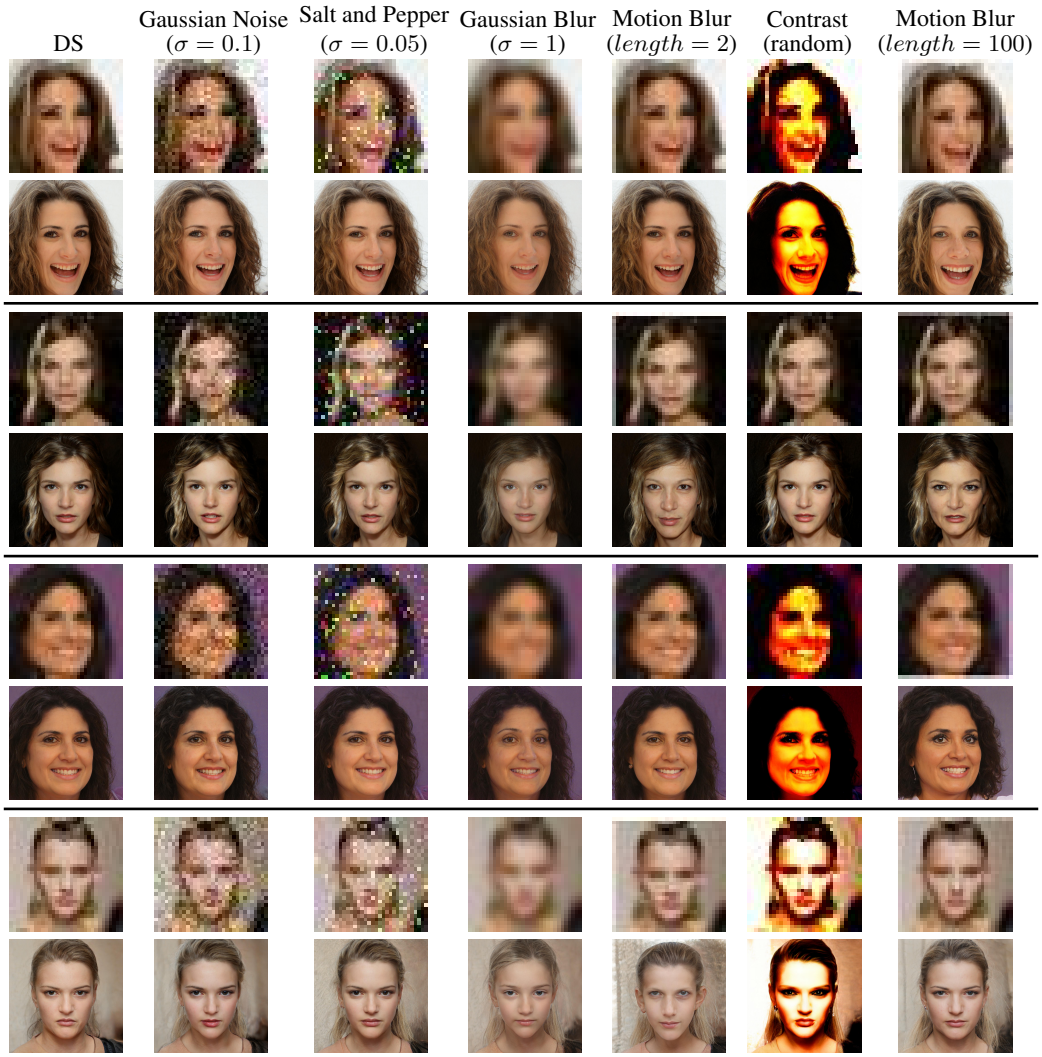

Figure 6: Additional results of robustness evaluation. After downscaling (DS) a HR image using bicubic downsampling, various degradation operators are applied, except for the last column where motion blur in a random direction with a length of 100 is first applied before downscaling.

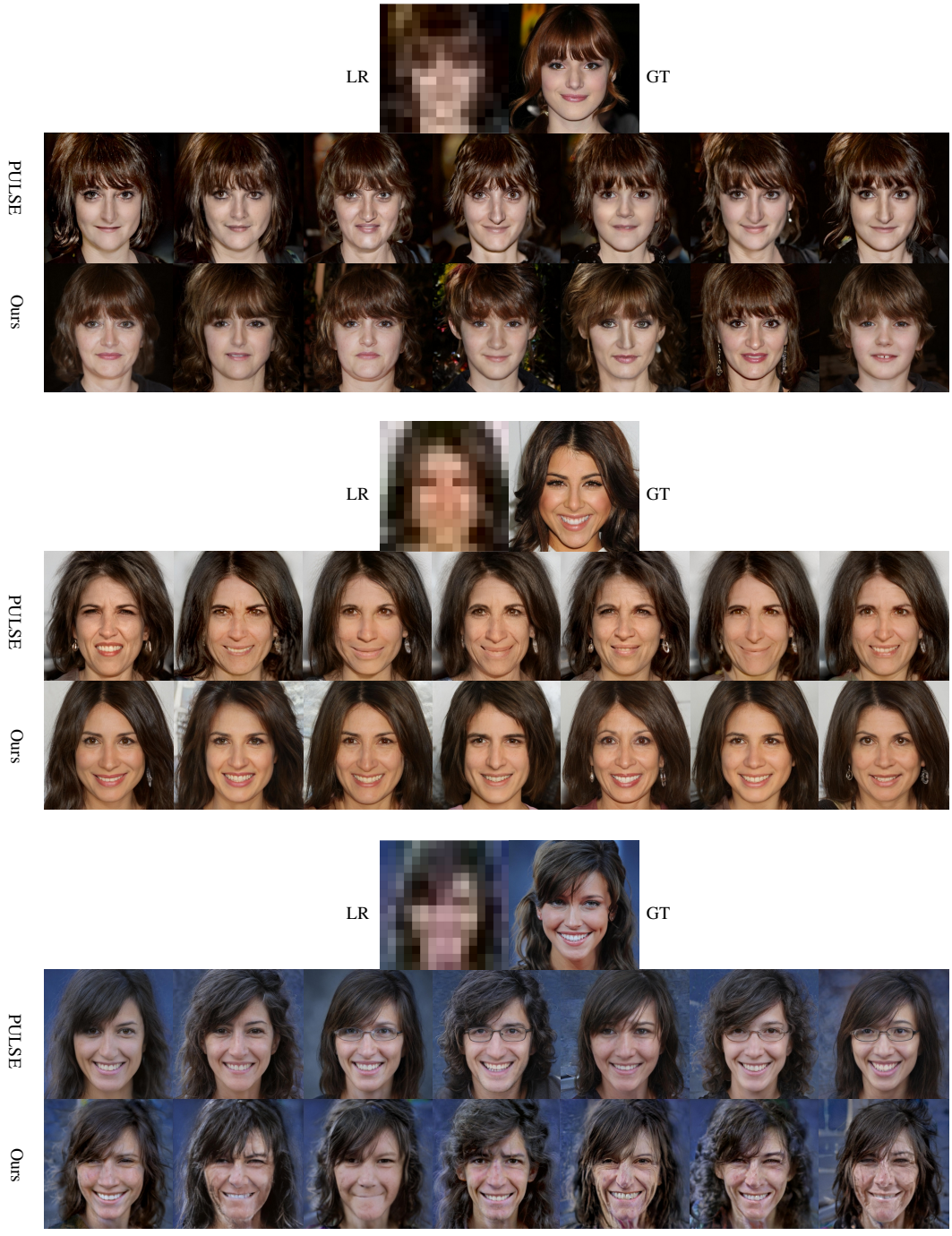

Figure 7: Additional examples of multiple solutions obtained for a given LR input image by using different initialization. Our approach leads to a diversity of generated images, whereas PULSE generates images with a more similar appearance.

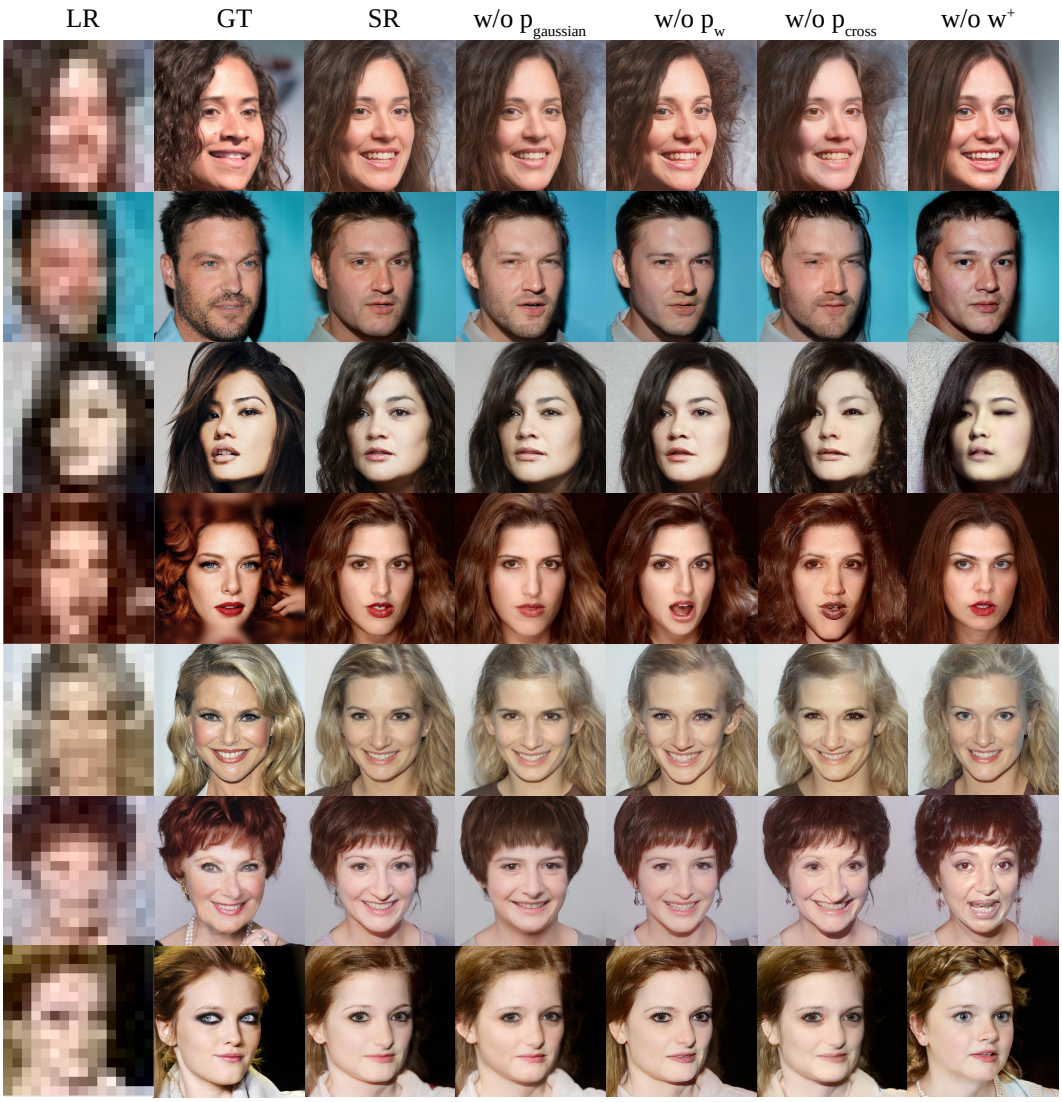

Figure 8: Qualitative comparison of four variants of our method to evaluate the effectiveness of our image prior

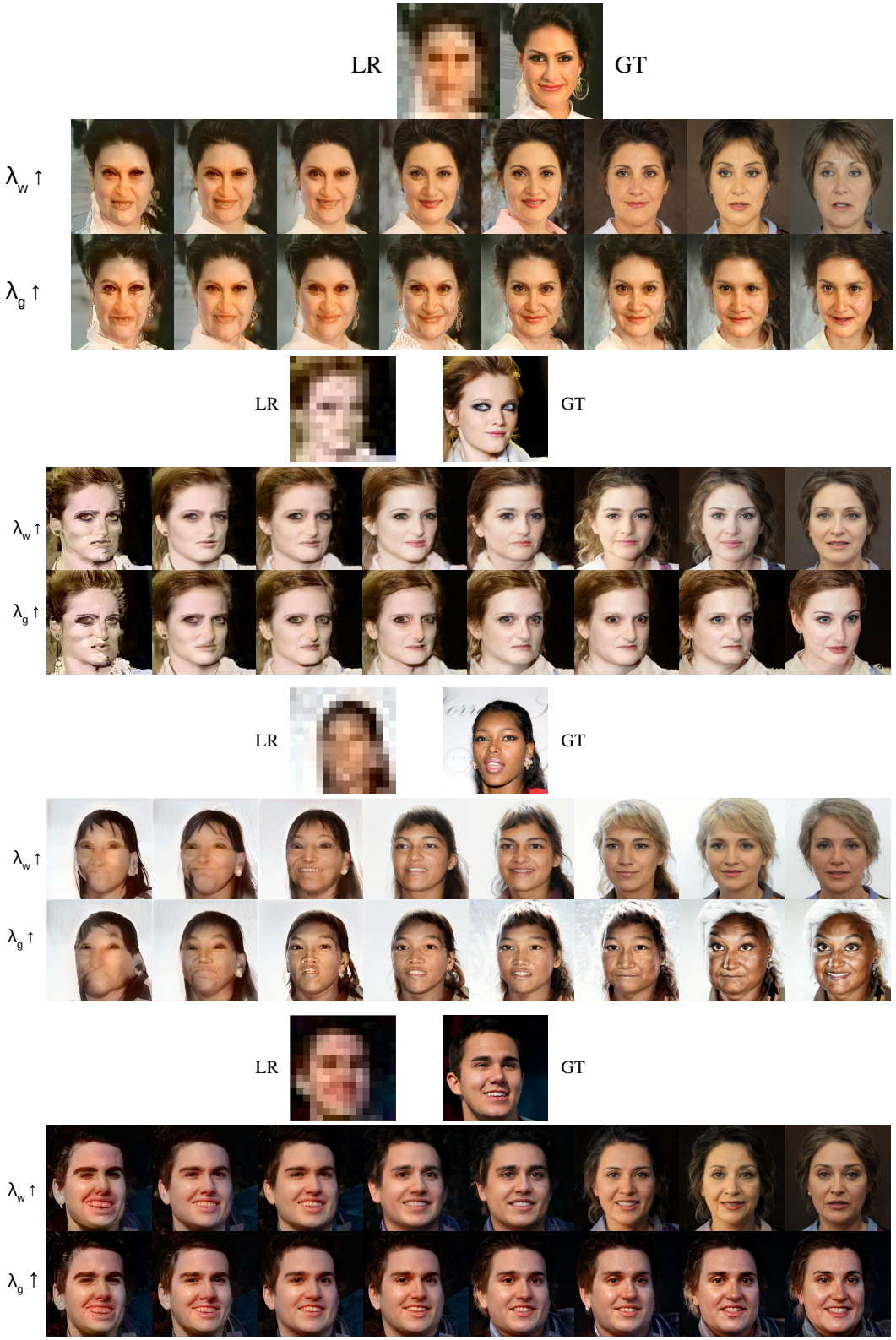

Figure 9: Balancing realism and fidelity by controlling the regularization parameters. $\lambda_w \uparrow$ indicates λ_w increases from left to right where $\lambda_g = 0$ and $\lambda_c = 0$ (Similarly for $\lambda_g \uparrow$)

.