CS585 Project Final Report: Multi-modal Photo Upsampling via Latent Space Exploration of StyleGAN

Musa Ege Ünalan and Abdallah Zaid Alkilani

***Abstract*—This report introduces several approaches for enhancing image resolution while enabling multi-modal multi-modal photo upsampling via latent space exploration of StyleGAN. Single-modal methods for image upsampling, traditional or otherwise, often lack diversity and richness in high-resolution representations. Our proposed methods leverage the power of StyleGAN to navigate the distribution of high-resolution images and explore their latent space, generating multiple plausible upscaled versions. By using a pre-trained StyleGAN synthesis notwork and developing latent space exploration tactics, we achieve fine results compared to single-modal methods in terms of objective metrics and subjective visual quality. Our multi-modal photo upsampling technique has potential applications in digital photography, medical imaging, and surveillance systems, offering visually appealing and realistic upscaled images without the burden of obtaining or storing large high-resolution data. Our implementation is available here.**

I. Introduction

IMAGE Super-Resolution (ISR) is the process of using one or more Low-Resolution (LR) images to produce a Super-Resolution (SR) image that is likely to come from a distribution of matching true High-Resolution (HR) images. This involves advanced algorithms that exploit the information contained in blurry LR images to transform them into sharp and realistic SR images. This process is particularly important in situations where high-quality imaging is required but the direct capturing of HR images is difficult, due to real-constraints such as hardware and storage limitations [12]. This is especially true in the era of big data, with consumer electronics demanding high-quality data at scale [6]. ISR techniques are also often employed to enhance the visual quality of images in various application domains of science and engineering, including medical imaging [10], satellite imaging [3], and video compression [5]. In recent years, ISR has made significant strides owing to advancements in deep learning techniques, with this renaissance of the field culminating in highly efficient SR algorithms.

II. Background and Related Work

While the benefits of ISR methods are clearly evident, the gap in information content between the LR and HR images can make it difficult to accurately recover a plausible SR image. LR images typically have degraded high-frequency information, which can cause blurring of important details and subsequently the inability to recover finer features of the underlying image content. Moreover, the ISR problem setup is inherently ill-posed, as there are many plausible HR images that correspond to the exact same starting LR image, with exponential growth of the problem to the scale of the image upsampling [1]. This entails that generating a set of potential HR images (i.e., SR images) is a much easier problem than accurately recovering the true HR image.

Traditional supervised ISR approaches have trained models to minimize the Mean Squared Error (MSE) of the matched training dataset of LR and HR images [8]. However, this was shown to perform poorly in terms of capturing fine-grained details such as texture, even with better architectures such as ResNet; this lead to a shift in literature towards Generative Adversarial Networks (GANs) [9]. GANs attempt to pull the solution towards the manifold of natural images where, in addition to the adversarial loss, perceptual loss and other regularizers were utilized in training. Be that as it may, issues of loss of high-frequency details still prevailed and the degradation from MSE-based approached persisted for the most part. Overall, MSE-based approaches result in an *averaging* of the plausible solutions (which could lie entirely outside the plausible HR images manifold), and modified-loss GAN-based approaches are not guaranteed to lie on the plausible HR images manifold (they merely approach it with an unguaranteed approximation).

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

To find images that both lie on the manifold of natural images and plausible HR images, a recent innovative idea is to explore the latent space of generative models. Menon *et al.* [11] demonstrated this approach using *Face StyleGAN* on face images (the so-called *face hallucination* problem; although the framework is generalizable to other domains). This approach is titled *PULSE* (Photo Upsampling via Latent Space Exploration). They introduce the downscaling loss to enforce consistency between the generated SR image and the input LR image, thereby addressing the plausible HR image manifold constraint. This is achieved by down-sampling the SR image and enforcing consistency by using a ℓ_p norm. Additionally, they put forward the idea of latent space exploration to find regions of the natural image manifold without using a differentiable parameterization of this manifold (as it is not implementable in practice). By using techniques from unsupervised learning and leveraging advances in generative modeling, pretrained models can be explored without the need

for additional training. This can be achieved via the guidance of a ℓ_p norm term between the LR image and the down-sampled generator output of the latent code, that as is limited to be arbitrarily small; by using a high dimensional uniform prior to replace the often-used spherical Gaussian prior, the latent space $\mathcal{L}$ can be confined to $\mathcal{L}' = \sqrt{d}S^{d-1}$ (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 latent code $z \in \mathcal{L}'$ that satisfies the ℓ_p norm term. This reduces the problem from gradient descent in the entire latent space to projected gradient descent on a sphere.

Evaluations on the well-known high-resolution face dataset CelebA HQ and comparison with various methods show that PULSE outperforms state-of-the-art in perceptual quality at higher resolutions and scale factors than previously achievable, taking 5 seconds to generate each image on a single GPU. This framework can be applied to any generative model, and can benefit from future advancements to architectures. However, the work is limited to a demonstration with bicubic downsampled images and remains to be evaluated in other interesting domains where degradation is not well-characterized. Moreover, the output is a single SR prediction, whereas ISR is a multi-modal problem and many images can satisfy the constraint of lying on the manifold of natural images and plausible HR images (see [1] for ill-posed ISR argument). Additionally, real-time efficient ISR is a challenge that could be addressed but is unachievable at the current rate of one image per 5 seconds due to the iterative spherical gradient descent search procedure at the time of inference.

B. Intermediate Layer Optimization for Inverse Problems using Deep Generative Models

Another approach that is different to PULSE from section II-A is Intermediate Layer Optimization (ILO) [4]. ILO is a framework for solving inverse problems by using deep generative models. Rather than optimizing exclusively over the initial latent code, the input layer is progressively modified to increase the expressibility of the generator. The generator G is considered in two consecutive parts, $G = G_1 \circ G_2$, where the split entails the intermediate layer for optimization. ILO searches for latent codes that lie within a small ℓ_1 ball around the manifold induced by the previous layer, effectively exploring the higher dimensional space for valid solutions. First, the latent code z^k from the latent distribution of the generator is initialized and optimized by matching the generator output to the input via a norm measure (after application of the general differentiable operator to both; down-sampling in the PULSE case) to produce $\hat{z}^k$, much like in the case of PULSE. $\hat{z}^p = G_1\left(\hat{z}^k\right)$ is initialized for the input of G_2. Then, the input space of G_2 is optimized via the norm measure over the aforementioned ℓ_1 ball that is centered around the previous $\hat{z}^p$ to produce the intermediate $\tilde{z}^p$. Finally, this $\tilde{z}^p$ is projected back to the range of the generator G_1 by optimizing via the norm measure and $\hat{z}^p = G_1\left(z^k\right)$ is updated for the next iteration. The final output is given as $G_2\left(\tilde{z}^p\right)$.

Theoretical analysis shows that deep generative models can achieve improved error bounds (cf. CSGM [2]) by keeping the radius of the ℓ_1 ball relatively small. Experiments are provided that verify ILO surpasses state-of-the-art methods such as in StyleGAN2 and PULSE for a wide range of inverse problems (including ISR); ILO achieves impressive results as an extension of PULSE beyond ISR (i.e., for all inverse problems with differentiable forward operators). This motivates the extension of PULSE to other generalizable approaches, in this case to the general framework of inverse problems. In this project, we intend to generalize PULSE to multi-modal image upsampling by introducing stochasticity to StyleGAN2.

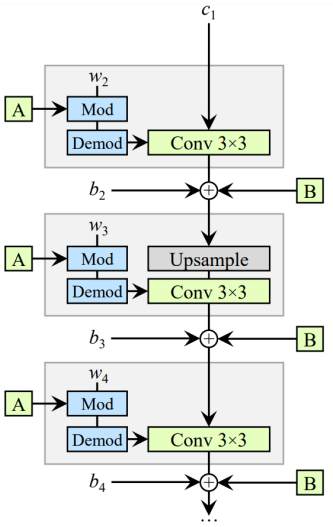


Fig. 1. The revised StyleGAN2 synthesis network architecture [7]

III. Methods

Our implementation uses the StyleGAN2 network and the pre-trained "stylegan2-ffhq-256x256.pkl" model from the official implementation[1]. The methodology we followed to achieve multi-modality in guided photo upsampling encapsulates several different ideas:

A. Sampling different initial latent codes

Starting the latent exploration process from multiple different initial latent codes, and optimizing them at the same time allows us to arrive at different plausible upsampled images that downsample to similar images.

This was implemented simply by increasing the batch size for the optimization process and sampling different latent codes in W+ space (14×512 for the 256×256 pre-trained model we are using) for each sample in the batch.

B. Applying user controlled noise on reference images

We explored applying Gaussian noise, with mean 0, on the whole image as [11] does, but we also implemented a way to only add noise on the parts of an image specified by a user provided mask (in our implementation this was generated from user scribbles), to explore the effect of applying localized noise to reference images on the optimization process.

[1]https://github.com/NVlabs/stylegan3

C. User controlled noise injected to StyleGAN2

We explored injecting noise to the synthesis layers in StyleGAN2's synthesis network (Figure 1). In contrast to the previous method of applying noise to reference images, this method applies noise to the generation process and directly guides it.

Since the StyleGAN2 implementation we used did not support injecting custom user-controlled noise to the synthesis layers, we needed to modify the implementation itself and the pre-trained model file we were using to be compatible with the modified implementation.

D. Enforcing inter-batch diversity

When generating multiple images in batch, we can introduce a negative-loss term in the loss function that favors pairwise pixel diversity, which results in diversity among the batch, while still downsampling to the same reference picture.

The negative loss term is defined as:

$$\mathcal{L}_{Batch} = \sum_{i,j \in Batch} -\|I_i - I_j\|_2 \tag{1}$$

We scale this negative-loss by a constant as higher amounts of negative-loss made the optimization process stray away from the original reference image.

IV. Results

We evaluated our implementation on the CelebA-HQ images that exist in the CelebA training set using the CelebA-HQ to CelebA mapping, on a total of 2824 images, with PSNR, SSIM, and LPIPS metrics (see subsection IV-E).

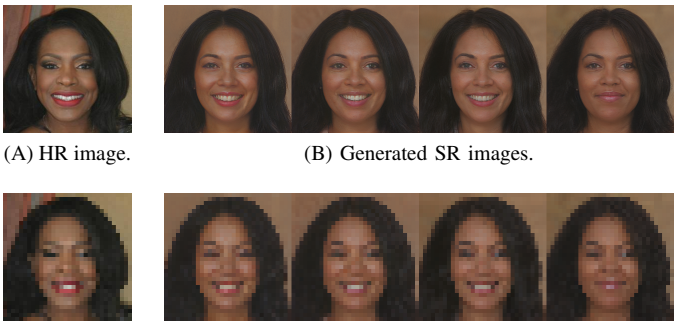
The images were generated using a learning rate of 0.4 with the Ranger21 optimizer[2] at 200 optimization steps, using an L2 loss constant of 405, geodesic loss constant of 0.058, inter-batch diversity loss constant of 0.02, with a batch size of 4, guided with 32×32 reference images downsampled from the CelebA-HQ images using bicubic downsampling, on the FFHQ-256x256 StyleGAN2 model.

A. Sampling different initial latent codes

This modification resulted in realistic looking upsampled images that downsample close to the reference 32×32 image, some results can be seen in Figure 2.

But we noticed that some initial latent codes for some images do not optimize well to the reference image, we attribute this to the optimization process getting stuck in a local optima.

[2]https://github.com/lessw2020/Ranger21



(A) HR image. (B) Generated SR images.

(C) LR image. (D) Downscaled SR images.

Fig. 2. An indicative example of sampling different initial latent codes. Each SR image in (B) was generated with a different initial latent code. The downscaled SR images show correspondence to the LR image, as expected.

B. Applying user controlled noise on reference images

Applying user controlled Gaussian noise on reference images allows the optimization process to arrive at diverse images by loosening the image reference constraint (L2 loss), while keeping the overall structure of the image same.

(A) LR images with random noise with differing seeds.

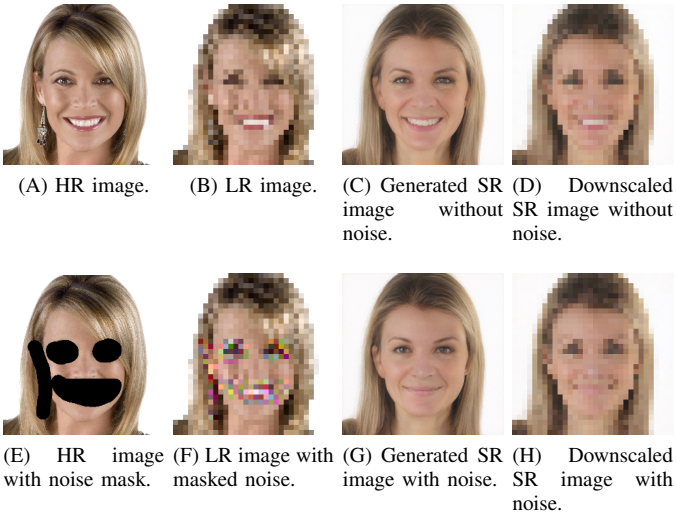
(B) Generated SR images by varying the random seed.

(C) Downscaled SR images.

Fig. 3. An indicative example of applying noise to the LR images. Each SR image in (B) was generated with a different seed for the random noise, but using the same initial latent code. The downscaled SR images show correspondence to each other, as expected from the application of noise, notwithstanding changes such as the volume of the hair and the smile which are high-level details that can easily be blocked out by the applied noise.

But applying stronger noise on the reference image leads to the optimization process straying away from the original

target, as the original information gets lost as more noise is applied. Figure 3 shows examples when noise is applied to the entire LR image, and Figure 4 shows the results when the noise is applied via a user-controlled mask.



(A) HR image. (B) LR image. (C) Generated SR image without noise. (D) Downscaled SR image without noise.

(E) HR image with noise mask. (F) LR image with masked noise. (G) Generated SR image with noise. (H) Downscaled SR image with noise.

Fig. 4. An indicative example of applying noise with a user-controlled mask to the reference images, while using the same initial latent code. As can be seen by the difference between (C) and (G), the noise mask affects the smile while the ear and eyes are largely maintained. This shows that diversity can be obtained while allowing the user creative control over the procedure.

C. User controlled noise injected to StyleGAN2

(A) HR image. (B) HR with face-isolating mask. (C) Generated SR image without diversity.

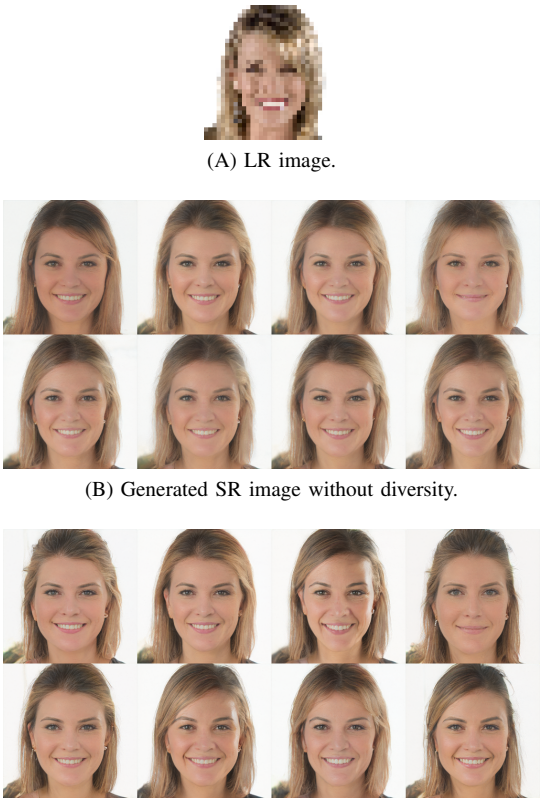
(D) Generated SR images with diversity.

Fig. 5. An indicative example of injecting noise to StyleGAN2. As can be seen by the difference amongst the SR images in (D) and with the SR image generated with the injected noise in (C), this method provides very subtle variety to the generated images. While this might be interpreted as insufficient multi-modality, it might be desirable in some cases where too much variation is not preferred (e.g., some medical imaging applications).

We found that applying too little to moderate amounts noise did not change the output of the generation process other than applying a film-grain like effect at the higher ends, while larger amounts of noise resulted in degraded images or images unrelated to the reference image. Some results can be seen in Figure 5.

D. Enforcing inter-batch diversity

The inter-batch loss term was successful in providing diversity among the batch while still being able to optimizing to the reference image. The results for a batch of 8 can be seen in Figure 6.



(A) LR image.

(B) Generated SR image without diversity.

(C) Generated SR image with diversity.

Fig. 6. An indicative example of enforcing inter-batch diversity on the generated SR images, using the same initial latent code for each image with some random StyleGAN2 noise applied. As can be seen by the difference between the SR images in (B) and in (C), this method provides very distinct variety to the generated images. Details such as the illumination of the face and the hairstyle are quite diverse, indicating that this method is a very viable and balanced means of introducing multi-modality.

E. Combination of all methods

Combining all of the previously mentioned methods require some careful fine tuning of the parameters to get realistic images. The results for a batch of 9, using reference image noise of mean 0 and standard deviation 0.108, and StyleGAN noise of mean 0 and standard deviation 0.02 can be seen in Figure 7. The quantitative results are summarized in Table I. We found our scores are comparable with the original PULSE [11] paper while also introducing multi-modality.

TABLE I
METRIC RESULTS FOR THE COMBINATION OF ALL METHODS DISCUSSED. THE RESULTS WERE AVERAGED OVER CELEBA-HQ MAPPED TO THE CELEBA TEST SET (2824 IMAGES). THESE RESULTS ARE INDICATIVE OF OBJECTIVE PERFORMANCE FOR THE MULTI-MODALITY APPROACHES.

Metric	Mean, μ	Standard deviation, σ
PSNR [dB]	18.28	1.65
SSIM [%]	49.61	8.41
LPIPS [%]	37.32	8.03

(A) HR image.

(B) LR image with applied random noise.

(C) Generated SR images.

(D) Downscaled SR images.

Fig. 7. An indicative example of the combination of all discussed methods for enforcing diversity on the generated SR images. As can be seen by the SR images in (B), a wide range of diversity is possible, including adding glasses and varying the hairline and skin tone. At the same time, the downscaled SR images in (D) show that consistency is still present between the generated diverse images.

V. DISCUSSION AND CONCLUSION

We explored several different methods of introducing diversity to photo upsampling using PULSE [11]. We found that there was a trade-off between diversity and consistency for the methods we implemented, with the exception of sampling different initial latent codes which does not have any adjustable parameters other than the batch size. Our experiments show that increasing the strength of the noise on reference images or the noise injected to the StyleGAN synthesis layers too much makes the optimization process stray away from the reference or result in unrealistic images. The most effective methods for introducing diversity to the upsampling process were enforcing inter-batch diversity and adding noise to the reference images with properly tuned parameters. Combining all of these methods requires fine tuning of the individual knobs of each method to achieve realistic but diverse images.

REFERENCES

[1] S. Baker and T. Kanade. “Limits on super-resolution and how to break them”. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. 2000, 372–379 vol.2.

[2] Ashish Bora et al. “Compressed Sensing using Generative Models”. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. 2017, pp. 537–546.

[3] Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis. “Open High-Resolution Satellite Imagery: The WorldStrat Dataset –With Application to Super-Resolution”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 25979–25991.

[4] Giannis Daras et al. “Intermediate Layer Optimization for Inverse Problems using Deep Generative Models”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. 2021, pp. 2421–2432.

[5] Zejiang Hou and Sun-Yuan Kung. “Multi-Dimensional Dynamic Model Compression for Efficient Image Super-Resolution”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 633–643.

[6] Andrey Ignatov et al. “Real-Time Quantized Image Super-Resolution on Mobile NPUs, Mobile AI 2021 Challenge: Report”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2021, pp. 2525–2534.

[7] Tero Karras et al. “Analyzing and Improving the Image Quality of StyleGAN”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[8] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. “Accurate Image Super-Resolution Using Very Deep Convolutional Networks”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1646–1654.

[9] Christian Ledig et al. “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 105–114.

[10] Y. Li, B. Sixou, and F. Peyrin. “A Review of the Deep Learning Methods for Medical Images Super Resolution Problems”. In: *IRBM* 42.2 (2021), pp. 120–133.

[11] Sachit Menon et al. “PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 2434–2442.

[12] Amanjot Singh and Jagroop Singh Sidhu. “Super Resolution Applications in Modern Digital Image Processing”. In: *International Journal of Computer Applications* 150 (2016), pp. 6–8.