Multi-modal Photo Upsampling via Latent Space & Exploration of StyleGAN

Zülal Nur Hıdıroğlu, Sarper Turan and Berk Saltuk Yılmaz, *Bilkent University*

***Abstract*—The main motivation behind the super-resolution task is to output images with high-resolution given images with low resolution. Even if there are many approaches for this task, the generated images generally occur to be blurry. However, PULSE (Photo Upsampling via Latent Space Exploration) brought a new perspective by introducing generative models to this task. PULSE explores latent space that is learned from the training data and upsamples images in an unsupervised manner. While being a game-changer approach, the controllability of outputs generated by PULSE is too low. With this project, we aim to bring diversity to outputs generated by PULSE by using StyleGAN2 as a generator. After completing this task, we aim to find semantic directions using an InterFaceGAN-based approach. This paper serves as a progress report to demonstrate our literature review, the future roadmap, and our implementation's current status.**

***Index Terms*—Multi-modal image upsampling, StyleGAN2, PULSE, InterFaceGAN, super-resolution**

I. Introduction

In the machine learning context, super-resolution refers to the task of transforming low-resolution images to high-resolution images while maintaining visual content and quality. Using a generative model for this task arose as an alternative as generative networks are evolving rapidly. PULSE is a self-supervised architecture offering a generative network-based solution with its underlying StyleGAN generator.

PULSE brought a novel approach to this task by using the power of generative adversarial networks, being independent of ground truths, introducing a novel optimization method that finds the latent code that gives the highest similarity between generated image and the original low-resolution image, and introducing perceptual losses and statistics to make output images sharper and more realistic. However, there are some deficiencies in this method. First of all, the model is able to output only one image at a time, and the generated images are not controllable and lack diversity. For example, as can be seen from Figure 1 the generated image, although similar to the original image, has different features and is one of the many possible high-resolution images that could be generated. This is because PULSE adopts a black box approach that does not provide room for users to edit the style of generated images. Moreover, the PULSE sometimes is not able to understand the style features of input images (specifically when the resolutions of images are very low) and generates images with high quality that are completely off from the input in terms of hair and skin colors. To overcome those problems, we offer changes to the PULSE model's generator by modifying it to function as a StyleGAN2 [4]. In this way, we are going to introduce diversity to the image generation pipeline by introducing multi-modality and making PULSE output multiple plausible high-resolution photos. Furthermore, we are going to adopt the techniques for finding directions in the latent space introduced in the InterFaceGAN model (which is a GAN-based method that disentangles different factors of variations in face images) to achieve even more diversity, as the super-resolution step disrupts the style space by editing the generated images with these directions.

In the beginning, we are going to have 32x32 low-resolution images, and we are going to achieve 256x256 images which are diverse and editable on the attributes such as smile, hair color, and bangs. To assess our success, we will use several metrics, including PSNR, SSIM, and LPIPS. In the rest of this report, detailed explanations of the aforementioned architectures, our roadmap for achieving what we have proposed, and the current implementation details can be found.

II. Covered Literature

The literature covered for the time being consists of Pulse: Self-supervised photo upsampling via latent space exploration of generative models and InterfaceGAN: Interpreting the disentangled face representation learned by GANs.

A. PULSE

Single-image super-resolution refers to the process of creating a high-resolution image from a low-resolution input. In the past, traditional supervised approaches have used pixel-wise average distances between the super-resolved and high-resolution images as training objectives, which lead to blurring because of smoothing in areas of high-variance. The PULSE algorithm proposes a different approach from the traditional ones that focuses on generating realistic super-resolved images that downscale correctly. It accomplishes this in a self-supervised way, without being limited to specific degradation ways used in training. Instead of starting with the low-resolution image and adding detail, PULSE starts with the high-resolution image manifold to find images that downscale to the original low-resolution image. This is guided by the "downscaling loss" that guides exploration through the latent space of a generative model. PULSE generates super-resolved images that are both realistic and downscaled correctly, and its effectiveness has been demonstrated in face super-resolution, outperforming state-of-the-art methods at higher resolutions and scale factors. The goal of PULSE (Photo Upsampling via Latent Space Exploration) is to find points that actually lie on the natural image manifold and also downscale correctly. The critical notion of correctness relies on how well the generated SR image corresponds to the LR input image [1].

For a proposed super-resolution image to represent the same information as a low-resolution image, it must downscale to that low-resolution image. PULSE achieves this by finding a latent vector z in the latent space L of a generative model such that the image generated by the generator G from z is

a good approximation of the target image $\mathbf{I_L R}$. The critical notion of correctness relies on how well the generated super-resolution image $\mathbf{I_S R}$ corresponds to $\mathbf{I_L R}$. This is formalized via downscaling loss mentioned above. Simply ensuring that z lies in L is not enough. More constraints are needed to ensure that G(z) is in the desired image manifold M [1].

To achieve this, PULSE adds a loss term for the negative log-likelihood of the prior distribution over the latent space. This encourages the latent vector z to be in a region of high probability under the prior distribution, which is typically assumed to be a high-dimensional spherical Gaussian distribution. However, this is not ideal because the mass of a high-dimensional Gaussian is located near the surface of a sphere with a radius of $\mathbf{sqrt}(\mathbf{d})$.

To overcome this limitation, PULSE uses a uniform prior distribution on the surface of a sphere with a radius of $\mathbf{sqrt}(\mathbf{d})$ instead. The new latent space L' is equivalent to the surface of a sphere in d-dimensional Euclidean space. By working in this L', the problem of finding a good latent vector z is reduced to a projected gradient descent problem where we want to find a point on the surface of the sphere that minimizes the distance between the generated image and the target image.

In conclusion, PULSE is covered for multi-modal photo upsampling, a powerful super-resolution imaging model that finds points on the natural image manifold and downscale correctly. It achieves this by using a downscaling loss and a uniform prior distribution on the surface of a sphere to find a good latent vector z that generates a high-resolution image from a low-resolution input image. By doing so, PULSE avoids the blurring effect that traditional methods often suffer due to smoothing high-variance areas and produces high-quality super-resolution images [1].

B. InterFaceGAN

The study proposes a system called InterFaceGAN, which aims to enable face editing without retraining Generative Adversarial Networks (GANs) by comprehending the face representation they generate. InterFaceGAN connects the latent space and the semantic space for representation analysis and uses commercially available classifiers to predict semantic scores for synthesized images. The paper also investigates how different semantics are encoded by GANs during training, separates them using subspace projection, and suggests a face editing pipeline. The work includes a thorough analysis of StyleGAN's taught face representation and a comparison with PGGAN, a quantitative assessment of the editing outcomes, an analysis of StyleGAN's learnt per-layer representation done layer by layer, and an identity analysis of the edited photos [2].

By using the synthetic data gathered by InterFaceGAN to train feed-forward models, the paper also suggests a new technique for actual face editing. The study applies InterFaceGAN to modify latent codes in StyleGAN's Z and W spaces. The findings show that W space outperforms Z space, particularly for long-distance manipulation, and that StyleGAN is capable of producing high-quality images with a variety of meanings. The study also discovers that certain visual attributes are associated with one another. Overall, InterFaceGAN satisfactorily functions on the style-based generator, allowing for simple modification of picture properties [2].

C. StyleGAN2

StyleGAN2 is first introduced in the "Analyzing and Improving the Image Quality of StyleGAN" paper as an extension of the original StyleGAN architecture. In this paper, the authors are introducing several modifications to StyleGAN. While the original StyleGAN produced a latent code fed through fully connected (FC) layers, the mapping network is deeper in the modified architecture. It maps random inputs to intermediate latent spaces. Moreover, this mapping network includes skip connections which provide better control over the style and provide more diversity. The authors also propose a novel regularization method to make the generator understand more structured representations. This new method, called "path length regularization," provides continuity (and therefore smoothness) of changes in the latent space. This regularization also reduces the noise and improves the quality of generated images [4].

In the new architecture, the detail level of generated images can also be controlled as the authors proposed style blocks with noise layers that add random noises. More importantly, each style block also contains a modulation layer that uses intermediate latent spaces and modulates activations of the convolution layers that allow the generator to produce higher-quality images that are highly editable. StyleGAN2 also has multi-resolution support. The authors also introduce some tricks to achieve progressive growth and equalized learning rate for stability in training [4].

III. Experiments

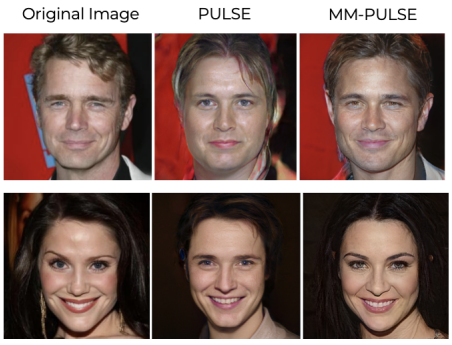

Fig. 1: PULSE vs MM-PULSE generated images

We fed our model images from the CelebA-HQ [3] dataset using images with 256x256 resolution. We downsampled all the images to 32x32 resolution using PULSE's downsampling function and then fed them onto both the original PULSE and our MM-PULSE in order to have a comparison as it can be seen from FIGURE X. We also generated images with different latents and displayed them in a matrix to see the differences of the generated images. We evaluated our model and PULSE with 3 different metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS). The

Model	PSNR			SSIM			LPIPS		
	Image 1	Image 2	Image 3	Image 1	Image 2	Image 3	Image 1	Image 2	Image 3
MM-PULSE	18.65	21.59	19.48	0.4577	0.5945	0.5409	0.379080	0.288116	0.242336
PULSE	20.65	21.84	20.67	0.5034	0.5570	0.5145	0.332250	0.246366	0.274747

TABLE I: PULSE vs MM-PULSE Evaluation Metrics

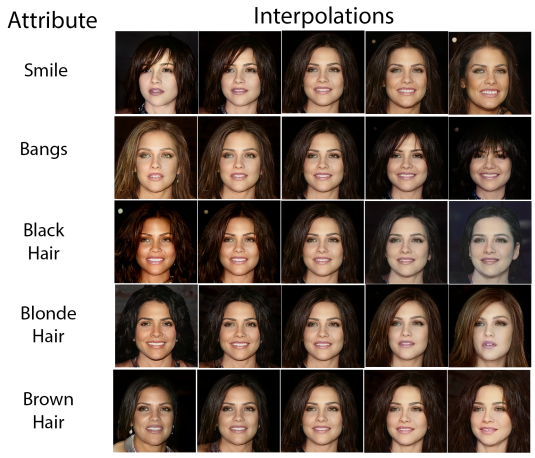

Fig. 2: Semantically Editing Upsampled Images

purpose behind those evaluations was to assess the perceptual similarity between the original image and the generated image. We also wanted to evaluate how well our model performs with the modifications comparing to the PULSE so both models gone through same steps throughout the evaluation. We got 100 images from CelebA-HQ dataset, downsampled them and fed them to MM-PULSE and PULSE, took the images and used our evaluation metrics to compare them to the original images. The results of these evaluations can be seen in tables 1 and 2, where table 2 shows the mean value found and table 1 shows values per-image for 3 images. A pattern that we observed throughout the evaluation metrics is that the metrics are low but for PULSE, the aim is not to optimizing pixel-average distances, so they have no meaningful implication so instead we focused on a comparison of MM-PULSE and PULSE.

Model	PSNR	SSIM	LPIPS
MM-PULSE	19.99	0.51774	0.2873294
PULSE	21.284	0.55098	0.279572

TABLE II: PULSE vs MM-PULSE Evaluation Metrics Mean for 100 Images

A. PSNR Evaluation

PSNR is a more traditional metric that primarily focuses on pixel-wise differences, it is a ratio between the maximum possible value of a signal and the power of distorting noise that affects the quality of the representation. A higher PSNR score means a higher quality image is generated. Here we observed that our model performed slightly worse comparing to the PULSE as it can be seen from table 1 and table 2.

B. SSIM Evaluation

SSIM metric focuses on 3 key features of an image which are luminance, contrast and the structure to measure the similarity between two images. The calculated value is called the Structural Similarity Index and it is between the values of -1 and +1, where +1 indicates the given images are extremely similar and -1 indicates the given images are extremely different. In some images MM-PULSE outperformed PULSE as it can be seen from table 2 images 2 and three, but as an overall evaluation the result indicated that our model is slightly less accurate according to SSIM as indicated in table 2.

C. LPIPS Evaluation

LPIPS is a perceptual metric used to measure similarity between images based on learned representations of human perception. It considers properties of visual perception such as color, texture and structure. This metric is an important metric in our case because it is the most similar to human perception. The LPIPS value indicates the distance between image patches which means that a higher LPIPS score represents more different images while lower LPIPS score means more identical the generated images are. We saw that our model had similar LPIPS values to PULSE as it can be seen from table 1. Looking at table 1, it can be seen that for image 1 it performed better than PULSE generating an image with higher similarity comparing to the image PULSE generated.

IV. Conclusion

In this project, we proposed MM-PULSE, a novel approach for multi-modal photo upsampling through the exploration of the latent space of StyleGAN. Building upon the PULSE method [1], we modified the architecture by integrating StyleGAN2. By leveraging a pre-trained StyleGAN2 model on the CelebA-HQ dataset, we were able to generate high-quality images with a resolution of 256x256. We achieved multi-modality in the generated images by introducing the concept of exploring multiple latents. We achieved this by generating k random latents and varying them across different iterations. To further enhance the flexibility and control over the generated images, we incorporated semantic direction-based editing, inspired by the methods employed in InterfaceGAN [2]. Lastly we evaluated our generated images on PSNR, SSIM and LPIPS metrics to compare it with PULSE, seeing whether our enhancements made any harm to the original PULSE architecture.

Our project contributes to the advancement of photo upsampling techniques by introducing MM-PULSE as a powerful tool for super-resolution multi-modal image generation through latent space exploration and semantic editing.

REFERENCES

[1] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition, pages 2437–2445, 2020.

[2] Y. Shen, C. Yang, X. Tang, and B. Zhou. Interfacegan: Interpreting the disentangled face representation learned by GANs. TPAMI, 2020.

[3] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[4] Karras, Tero, et al. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. p. 8110-8119.