Multi-modal Photo Upsampling via Latent Space Exploration of StyleGAN

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

***Abstract*—Super-resolution, the task of generating high-resolution images from low-resolution inputs, has long been challenged by blurry outputs from existing approaches. In that sense, PULSE (Photo Upsampling via Latent Space Exploration) produces high-resolution images with high accuracy; however, it outputs only one image. Therefore in this project, we address and solve the limitations of PULSE by integrating StyleGAN2 into the PULSE to enhance the quality of generated images and using different random latents for multi-modality. Additionally, we employ an InterFaceGAN-based approach to determine semantic directions for attributes like 'smile', 'bangs' and 'hair color' and utilize them to perform targeted edits on the generated images. This final report provides a comprehensive literature review that constitutes the related work, presents our methodology, and the evaluation results of our approach.**

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

I. Introduction

The field of computer vision has seen a lot of interest in generating high-resolution images from low-resolution inputs, known as super-resolution. There is a growing demand for high-quality images with more details and better visual quality, which has led to extensive research in this area. However, existing methods struggle to create realistic and clear high-resolution images because they have difficulty capturing fine details and producing accurate results.

This study aims to improve the process of increasing image resolution by introducing multi-modal upsampling and enabling semantic editing. We accomplish this by utilizing two techniques PULSE and InterFaceGAN, proposing Multi-modal PULSE (MM-PULSE). PULSE brings a fresh approach to super-resolution by incorporating generative models, which learn from training data to enhance image quality. However, one limitation of PULSE is its restricted control over the generated outputs, as it produces only one image per input. On the other hand, InterFaceGAN focuses on separating different aspects of images within generative models, allowing us to precisely control specific image characteristics.

The aim of MM-PULSE is to combine two techniques to achieve two main objectives: multimodal super-resolution and image attribute editing. To enhance the capabilities of PULSE, we incorporated the StyleGAN2 generator. This improves the quality of the images generated by PULSE. Additionally, by utilizing the InterFaceGAN-based approach, explore and utilize semantic directions associated with specific image features like smile, bangs, and hair color. This allows us to make precise edits to the generated images, providing more control and customization.

In this report, we start by presenting related work, then our methodology where we show our approach and lastly the evaluations of our model. By harnessing the unique advantages of PULSE, StyleGAN2, and InterFaceGAN, we strive to enhance the abilities of generating high-resolution images, opening up new avenues for more advanced and controllable image synthesis.

Overall, this study aims to push the boundaries of super-resolution image generation and opens the way for advancements in generating realistic, high-resolution images with enhanced details and customizable attributes.

II. Related Work

As it is obvious that the related work starts with the basics of image-to-image translation, generative adversarial networks, and superresolution tasks, we have decided to focus on three works directly related to our proposed method. The related papers regarding our model are "PULSE: Self-supervised photo upsampling via latent space exploration of generative models", "StyleGAN2", 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 led 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. Method

A. Multi-Modal Image Upsampling

We started with modifying the PULSE architecture before starting on multi-modality. To enhance the quality and realism of the generated images, we replaced the original StyleGAN generator used in PULSE with the more advanced StyleGAN2 architecture. StyleGAN2 offers improved image synthesis capabilities, allowing us to generate more visually appealing results. We did this by downloading a pre-trained StyleGAN2

Fig. 1: MM-PULSE generated images

model that is trained on a 256x256 CelebA-HQ dataset. While PULSE originally operated at 1024x1024 resolution, we

adapted the model to produce images at 256x256 resolution as a part of the project. Furthermore, we updated the loss function accordingly to optimize for this resolution conversion. We also incorporated the mapping network provided in this pre-trained model.

To introduce diversity in the generated images, we implemented multi-modality within our MM-PULSE framework. Instead of relying on fixed random seeds, we adopted an approach generating multiple random latents with different random seeds at each iteration of the model. By doing this in the input latents, we were able to have a wider exploration of the latent space, therefore unlocking a broader range of possibilities and having better diversity in the generated images. This approach to random seed generation allowed us to break away from the constraints of fixed seeds. In traditional settings, using a fixed seed would result in generating similar or nearly identical images every time the model is run. However, by generating multiple random latents with different seeds, we introduced a form of controlled randomness. Looking at figure 1, we can see that the generated images vary although keeping the same overall structure, which was one of our goals.

Having multi-modality on MM-PULSE not only enhances the variety of the generated images but also provides greater flexibility. The generated images can exhibit a wider range of interpretations of the given input or target attributes. This technique expanded the creative potential of our model and offers a more diverse output compared to the traditional approach of PULSE.

In summary, by generating multiple random latents with different random seeds on top of PULSE, we achieved multi-modality in MM-PULSE that enables a wider exploration of the latent space, leading to greater diversity in the generated images.

B. Finding Semantic Directions

To find semantic directions, we first experimented with the official implementation of InterfaceGAN [3]. We have generated images using our multi-modal PULSE model and saved the latent codes of those images. Using the boundary vectors provided in the implementation, we tried to interpolate our latent vectors and achieve semantic editing. For this purpose, we have used the smile boundary vector for the StyleGAN model. However, since the pre-trained boundary did not fit into our latent space, the interpolations resulted in changes in the gender of the generated images. To overcome this problem, we have decided to train our own boundary vectors.

To train boundary vectors, we have decided to use the method provided in the InterfaceGAN architecture. As the authors of InterfaceGAN proposed, there exists a hyperplane in the latent space that acts as a separation boundary for attributes; we could also train a support vector machine that finds this hyperplane using the saved latent vectors and their attribute scores. As we already saved the latent vectors for each image we generated, we only needed the score vectors for the attributes we wanted to edit.

To achieve the score vectors, we have used the annotated CelebA-HQ dataset provided in [3]. This dataset contains 202599 images, each labeled for 40 attributes, including smile, bangs, and hair colors, which are our aims in semantic editing. In addition to this dataset, we have used pre-trained ResNext50 (32x4d) that is provided by PyTorch. As implemented in rgkannan676's repository[1], we have incorporated a fully connected layer that has an output size of 40 since we had 40 features to predict. This implementation further uses MSE and Adamax optimizer and image augmentation techniques such as horizontal flipping, cropping, and resizing that improve the accuracy. We have trained this ResNext50 model with 10129 CelebA-HQ images for two epochs and 202 batches each. This trained model achieved 90.3% accuracy.

After having the attribute classifier, we fed each generated image to this network and saved the confidence scores for the attributes smile, bangs, blond hair, brown hair, and black hair, along with the latent vectors. By using 467 latent vectors, we have trained an SVM for each attribute and obtained five different boundary vectors as proposed in [2]. By using the obtained boundaries, we linearly interpolated our latent vectors in five steps and chose the start distance as -35 and the end distance as 35. By generating images from those interpolated latent vectors, we have achieved semantic attribute editing, as can be seen in Figure 2. However, in many cases, even if we

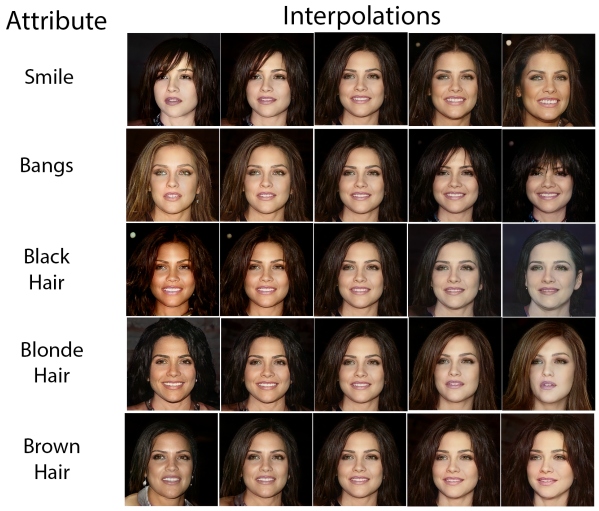

Fig. 2: Semantically Editing Upsampled Images

interpolated for a specific attribute, some other attributes are also changed with that specific attribute. For example, when we change the smile attribute, the pose and skin color also change. When we added bangs, the hair color tended to be darker, and the hair color seemed to be correlated with the skin color. These partial semantic failures can also be seen in Figure 2. The reason for those semantic failures might be due to the relatively small number of images we have trained our attribute classifier. This classifier might not be confidently spotting each attribute. This lack of confidence in attribute spotting can lead to errors or inconsistencies when training the boundary vectors. Another reason might be the number of generated images we used during the boundary training.

[1]https://github.com/rgkannan676/Recognition-and-Classification-of-Facial-Attributes

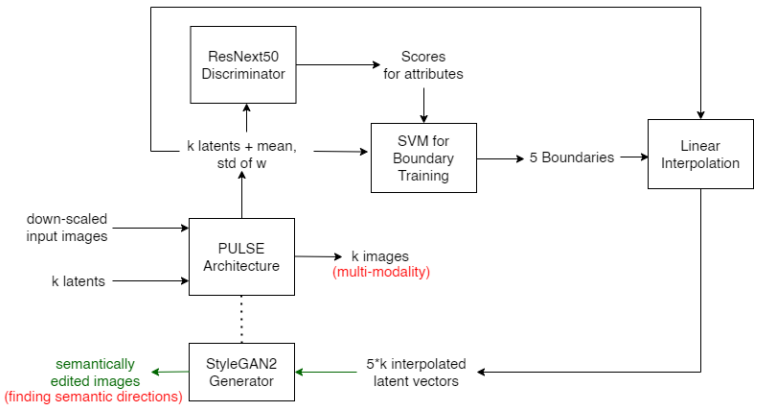

Fig. 3: Our Architecture

Four hundred sixty-seven might be too less in number to have a hyperplane that exactly separates the presence of an attribute. The final reason might be the superresolution step. The superresolution process, while enhancing the resolution and details of the images, may introduce unintended changes to the overall style or appearance. These changes may disrupt our style space in a way that the attributes become correlated with each other and not separable as before.

The disruption of the style space is also observed qualitatively by us; it is obvious that the superresolution step tends to add a smile to generated images by editing the style space in the smile direction since the vast majority of generated images have the smile attribute. Nevertheless, our method achieved semantic editing by composing the PULSE and InterfaceGAN architectures. For trying the semantic editing further, our repository[2] can be visited, and more boundaries can be created.

C. Our Architecture

As outlined in previous subsections, our architecture consists of the original PULSE architecture enriched by StyleGAN2 generator (synthesis and mapping networks), ResNext50 discriminator specialized for attribute classification of forty labels of CelebA-HQ, SVMs that are trained for five chosen attributes, and linear interpolation module. Images can be first fed to the align-face module of PULSE to downscale or directly the down-scaled images can be fed to PULSE. Along with the down-scaled images, we feed k latents to have k different outputs for each image, this forms the multi-modality part of our model. During the image generation, we add the mean and standard deviation of w vectors to our latent vectors and save them to input the ResNext50 discriminator. The discriminator gives the confidence scores for different attributes in the generated images. Using the saved latents and confidence scores, an SVM for each of the five attributes is trained. The resulting boundary vectors are fed to the interpolation module. The interpolated vectors are fed to the synthesis network of the StyleGAN2 model to achieve edited images, this outlines finding the semantic directions objective of our model. A rough sketch of our architecture can be seen in Fig. 3.

[2]https://github.com/berksaltuk/mm-pulse-with-semantic-editing

IV. Experiments

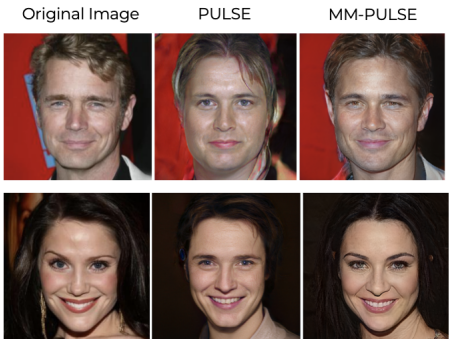

Fig. 4: PULSE vs MM-PULSE generated images

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 4. 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 three different metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS). The 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 go through the same steps throughout the evaluation. We got 100 images from the 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		
	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

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.

V. 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 using ResNext50 and SVM inspired by the methods employed in InterfaceGAN [2]. We achieved to edit the images we generated with five different attributes: smile, bangs, brown hair, black hair, and blond hair. Lastly, we evaluated our generated images on PSNR, SSIM, and LPIPS metrics to compare them with PULSE, seeing whether our enhancements did 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.