Image2StyleGAN++: How to Edit the Embedded Images? -Supplementary Material-

Rameen Abdal
KAUST
rameen.abdal@kaust.edu.sa

Yipeng Qin
Cardiff University
qiny16@cardiff.ac.uk

Peter Wonka
KAUST
pwonka@gmail.com

1. Additional Results

1.1. Image Inpainting

To evaluate the results quantitatively, we use three standard metrics, SSIM, MSE loss and PSNR score to compare our method with the state-of-the-art Partial Convolution [2] and Gated Convolution [4] methods.

As different methods produce outputs at different resolutions, we bi-linearly interpolate the output images to test the methods at three resolutions 1024×1024, 512×512 and 256×256 respectively. We use 7 masks (Fig. 1) and 10 ground truth images (Fig. 2) to create 10 defective images (*i.e.* images with missing regions) for the evaluation. These masks and images are chosen to make the inpainting a challenging task: i) the masks are selected to contain very large missing regions, up to half of an image; ii) the ground truth images are selected to be of high variety that cover different genders, ages, races, *etc*.

Table 1 shows the quantitative comparison results. It can be observed that our method outperforms both Partial Convolution [2] and Gated Convolution [4] across all the metrics. More importantly, the advantages of our method can be easily verified by visual inspection. As Fig. 3 and Fig. 4 show, although previous methods (*e.g.* Partial convolution) perform well when the missing region is small, both of them struggle when the missing region covers a significant area (*e.g.* half) of the image. Specifically, Partial Convolution fails when the mask covers half of the input image (Fig. 3); due to the relatively small resolution (256×256) model, Gated Convolution can fill in the details of large missing regions, but of much lower quality compared to the proposed method (Fig. 4).

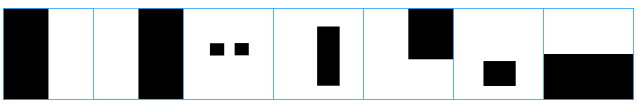

Figure 1: Masks used in the quantitative evaluation of image inpainting methods.

In addition, our method is flexible and can generate different inpainting results (Fig. 5), which cannot be fulfilled by any of the above-mentioned methods. All our inpainting results are of high perceptual quality.

Limitations Although better than the two state-of-the-art methods, our inpainting results still leave room for improvement. For example in Fig. 3, the lighting condition (first row), age (second row) and skin color (third and last row) are not learnt that well. We propose to address them in the future work.

1.2. Image Crossover

To further evaluate the expressibility of the Noise space, we show additional results on image crossover in Fig. 6. We show that the space is able to crossover parts of images from different races (see second and third column). Fig. 7 highlights the difference between the third and fourth images in the second row of Fig. 5 (main paper).

1.3. Local Edits using Scribbles

In order to evaluate the quality of the local edits using scribbles, we evaluate the face attribute scores [3] on edited images. We perform some common edits of adding baldness, adding a beard, smoothing wrinkles and adding a moustache on the face images to evaluate how photo-realistic the edited images are. Table 2 shows the average change in the confidence of the classifier after a particular edit is performed. We also show additional results of the Local edits in Fig. 8. For our method, one remaining challenge is that sometimes the edited region is overly smooth (e.g. first row).

1.4. Attribute Level Feature Transfer

We show a video in which attribute interpolation can be performed on the base image by copying the content from an attribute image. Here different attributes can be taken from different images embedded in the W^+ space and applied to the base image. These attributes can be independently interpolated and the results show that the blending quality of the framework is quite high. We also show additional results on LSUN Cars and LSUN Bedrooms in the video (also see Fig. 9). Notice that in the LSUN bedrooms, for instance, the style and the position of the beds can be customized without changing the room layout.

Figure 2: Images used in the quantitative evaluation of image inpainting methods.

Method	Image Resolution (1024 × 1024)			Image Resolution (512 × 512)			Image Resolution (256 × 256)		
	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR
Partial Convolution [2]	0.8957	199.39	21.83	0.8865	98.83	21.92	0.8789	48.39	22.17
Gated Convolution [4]	0.8693	246.46	19.65	0.8568	121.98	19.77	0.8295	61.82	19.41
Ours	**0.9176**	**180.69**	**22.35**	**0.9104**	**89.25**	**22.48**	**0.9009**	**43.85**	**22.65**

Table 1: Evaluation results of image inpainting methods using SSIM, MSE and PSNR score.

In order to evaluate the perceptual quality of attribute level feature transfer, we compute perceptual length [1] between the images produced by independently interpolated attributes (called masked interpolation). StyleGAN [1] showed that the metric evaluates how perceptually smooth the transitions are. Here, perceptual length measures the changes produced by feature transfer which may be affected especially by the boundary of the blending. The boundary may tend to produce additional artifacts or introduce additional features which is clearly undesirable.

We compute the perceptual length across 1000 samples using two masks shown in Fig. 1 (First and Seventh column). In Table 3 we show the results of the computation of the perceptual length (both for masked and non-masked interpolation) on FFHQ, LSUN Cars and LSUN Bedrooms pretrained StyleGAN. We compare these scores as the non-masked interpolation gives us the upper bound of the perceptual length for a model (in this case there is no constraint on what features of the face should change). As a particular area of the image is interpolated rather than the whole image, note that our results on FFHQ pretrained StyleGAN produce lower score than the non-masked interpolation. The low perceptual length score suggests that there is a less drastic change. Hence, we conclude that the output images have comparable perceptual quality with non-masked interpolation.

LSUN Cars and LSUN Bedrooms produce relatively higher perceptual length score. We attribute this result to the fact that the images in these datasets can translate and the position of the features is not fixed. Hence, the two images produced at random might have different orientation in which case the blending does not work as good.

1.5. Channel wise feature average

We perform another operation denoted by $I_{att}(1, 0, w_x, , n_{ini}, 6)$, where w_x can be the W^+ code for images I_1 or I_2. In Fig. 10, we show the result of this operation which is initialized with two different W^+ codes. The resulting faces contain the characteristics of both faces and the styles are modulated by the input W^+ codes.

References

[1] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018. 2

[2] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. *Lecture Notes in Computer Science*, page 89–105, 2018. 1, 2, 3

[3] Microsoft. Microsoft azure face. https://azure.microsoft.com/en-us/services/cognitive-services/face/. 1

[4] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018. 1, 2, 4

Figure 3: First column: original image; Second column: defective image; Third column: inpainted image via Partial Convolutions [2]; Fourth column: inpainted image using our method.

Figure 4: First column: original image; Second column: defective image; Third column: inpainted image via Gated Convolutions [4]; Fourth column: inpainted image using our method.

Edit	Attribute	Change in confidence
Wrinkle Smoothing	age	21%
Adding Baldness	bald	75%
Adding Beard	beard	42%
Adding Moustache	moustache	49%

Table 2: Changes in confidence scores of classifier after user edits.

Figure 5: Inpainting results using different w_{ini} initializations.

Pretrained model	Interpolation	Perceptual length (full)	Perceptual length (end)
FFHQ	Non-Masked	227.1	191.1
	Masked	112.1	89.8
LSUN Cars	Non-Masked	12388.1	6038.5
	Masked	4742.3	3057.9
LSUN Bedrooms	Non-Masked	2521.1	1268.7
	Masked	1629.8	938.1

Table 3: Perceptual length evaluation for masked and non-masked interpolation.

Figure 6: (a) and (b): input images; (c): the "two-face" generated by naively copying the left half from (a) and the right half from (b); (d): the "two-face" generated by our Image2StyleGAN++ framework.

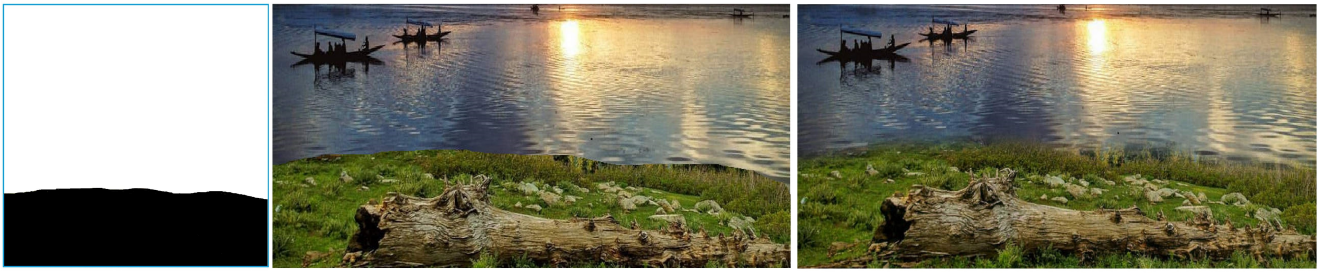

Figure 7: Mask and zoomed-in images from Fig. 5 in the main paper.

Figure 8: Column 1 & 4: base image; Column 2 & 5: scribbled image ; Column 3 & 6: result of local edits.

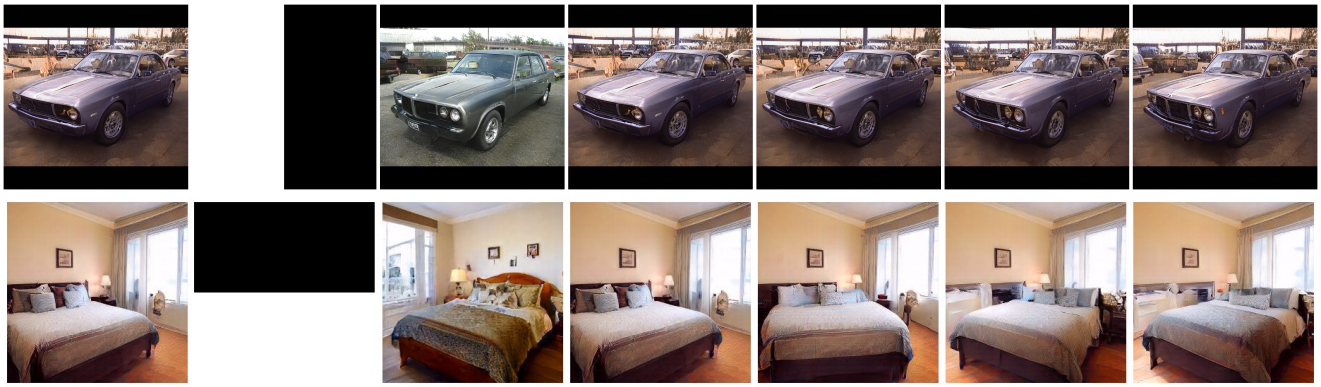

Figure 9: First column: base image; Second column: mask area; Third column: attribute image; Fourth to Eighth column: image generated via attribute level feature transfer and masked interpolation.

Figure 10: First column: First Image; Second Column: Second Image; Third Column: Feature averaged image using W^+ code of first image; Fourth Column: Feature averaged image using W^+ code of second image.