GAN Inversion for Out-of-Range Images with Geometric Transformations Supplementary Material

Kyoungkook Kang
POSTECH CSE
kkang831@postech.ac.kr

Seongtae Kim
POSTECH GSAI
seongtae0205@postech.ac.kr

Sunghyun Cho
POSTECH CSE & GSAI
s.cho@postech.ac.kr

1. Encoder Architecture Details

Table 1 presents a detailed architecture of our encoder network. The network has 10 convolution blocks, three average pooling layers and one convolution layer. Each convolution block consists of a convolution layer, a Leaky ReLU layer and a Batch Normalization layer.

$I_{\downarrow}(3ch)$
3×3 ConvBlock $(128ch)$
3×3 ConvBlock $(128ch)$
3×3 ConvBlock $(128ch)$
Average Pooling
3×3 ConvBlock $(256ch)$
3×3 ConvBlock $(256ch)$
3×3 ConvBlock $(256ch)$
Average Pooling
3×3 ConvBlock $(512ch)$
3×3 ConvBlock $(512ch)$
3×3 ConvBlock $(512ch)$
Average Pooling
3×3 ConvBlock $(512ch)$
3×3 Conv $512ch$

Table 1: Detailed architecture of our encoder network. Each ConvBlock has a convolution layer, a Leaky ReLU activation, and a Batch Normalization layer. We use average pooling to downsample feature maps by the scale factor of 0.5.

2. Reconstruction Quality Evaluation using the FFHQ dataset [2]

Table 2 shows the full result of Table 1 in the main paper including SSIM [7] and RMSE, which compares the reconstruction qualities of state-of-the-art GAN inversion methods and ours on the test set generated from the CelebA-HQ dataset [2]. As shown in the table, ours (16×16) outperforms all the other methods in all the metrics. Our 8×8 version also shows better reconstruction results than the previous methods except for Im2StyleGAN [1].

3. Editing Quality Evaluation using the FFHQ Dataset [2]

We compare the editing qualities of previous GAN inversion methods and ours using the FFHQ dataset [2] both qualitatively and quantitatively. To quantitatively measure the editing quality, we compare the editing result of an in-range image, and the editing result of an out-of-range image obtained from the in-range image using a geometric transformation. Specifically, we sample 50 latent codes $\mathbf{z}$ and generate images using a StyleGAN2 [3] model pre-trained on the FFHQ dataset. Then, to each latent code, we apply editing operations and synthesize ground-truth editing results. For the editing operation, we prepare three editing vectors using SeFa [6]: pose change, aging and expression change and apply them to images both in the positive and negative directions resulting in six editing results for each image. We also prepare out-of-range images corresponding to the in-range images by applying a random geometric transformation uniformly sampled from translation, rotation and scaling. To each transformed image, we apply previous GAN inversion methods and ours and obtain its latent codes. Then, we apply the same editing operation to the latent codes and synthesize their resulting images. Finally, we measure the difference between the ground-truth editing results and the editing results of the GAN inversion methods.

Table 3 reports the quantitative comparison result. For the in-range images without geometric transformations (Translation 0 in the table), our 16×16 version performs slightly better than PSP [5], but worse than the others as our method uses only the fine-scale latent codes for image editing. Our method with the base code $\mathbf{f}$ of size 8×8 performs better than StyleGAN2 inversion [3] and PSP [5], but worse than Im2StyleGAN [1] and P-norm^{+} [10]. On the other hand, for the out-of-range images transformed with translation and rotation, our 16×16 version outperforms the other methods showing no performance degradation

Models	Metric		Translation 0	50	100	150	Rotation 10	20	30	Scaling 7/8 ↓	3/4 ↓	9/8 ↑	5/4 ↑
Im2StyleGAN [1]	PSNR	↑	25.63	25.06	24.53	23.92	25.76	24.65	23.87	25.82	25.25	26.17	26.27
	FID	↓	48.37	45.73	52.52	58.64	50.06	56.63	65.76	33.80	34.24	38.02	36.78
	SSIM	↑	0.73	0.71	0.71	0.70	0.72	0.71	0.71	0.72	0.71	0.74	0.74
	RMSE	↓	**0.05**	**0.05**	0.06	0.06	**0.05**	0.06	0.06	**0.05**	**0.05**	**0.05**	0.05
P-norm$^+$ [10]	PSNR	↑	21.79	20.94	19.78	18.54	20.70	18.91	17.93	21.53	19.41	22.07	21.85
	FID	↓	58.69	64.52	78.56	98.53	77.93	86.16	110.48	46.89	60.38	52.76	49.06
	SSIM	↑	0.72	0.71	0.69	0.67	0.71	0.68	0.66	0.71	0.67	0.73	0.74
	RMSE	↓	0.08	0.09	0.10	0.12	0.09	0.12	0.13	0.09	0.11	0.08	0.08
StyleGAN2 inv. [3]	PSNR	↑	18.73	18.29	17.31	16.71	17.95	17.22	16.02	18.65	18.43	19.12	19.43
	FID	↓	65.49	70.36	78.32	87.70	79.31	82.25	96.23	52.26	50.23	60.64	60.24
	SSIM	↑	0.68	0.67	0.66	0.65	0.67	0.65	0.64	0.67	0.65	0.70	0.71
	RMSE	↓	0.12	0.13	0.14	0.15	0.13	0.14	0.16	0.12	0.12	0.12	0.11
PSP [5]	PSNR	↑	20.54	19.03	17.59	16.50	19.14	17.78	16.99	19.02	17.78	20.63	20.15
	FID	↓	78.53	84.85	99.66	118.50	108.13	115.46	142.09	84.87	96.29	70.16	68.32
	SSIM	↑	0.67	0.65	0.63	0.61	0.65	0.63	0.62	0.64	0.60	0.68	0.68
	RMSE	↓	0.10	0.11	0.13	0.15	0.11	0.13	0.14	0.11	0.13	0.09	0.10
Ours (8×8)	PSNR	↑	23.69	23.35	23.74	23.50	23.30	22.06	21.35	23.37	22.72	23.93	24.22
	FID	↓	49.68	49.47	46.05	49.00	60.84	60.52	71.71	37.51	38.34	44.11	37.43
	SSIM	↑	0.75	0.75	0.75	0.74	0.75	0.73	0.72	0.75	0.72	0.76	0.77
	RMSE	↓	0.07	0.07	0.07	0.07	0.07	0.08	0.09	0.07	0.08	0.07	0.06
Ours (16×16)	PSNR	↑	**26.47**	**26.30**	**26.37**	**26.43**	**26.48**	**26.49**	**26.33**	**26.44**	**26.28**	**26.98**	**27.26**
	FID	↓	**30.27**	**32.16**	**30.68**	**31.58**	**37.01**	**33.96**	**33.98**	**24.92**	**24.29**	**27.61**	**23.84**
	SSIM	↑	**0.78**	**0.78**	**0.78**	**0.78**	**0.79**	**0.78**	**0.78**	**0.78**	**0.77**	**0.80**	**0.81**
	RMSE	↓	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.04**

Table 2: Reconstruction quality of different methods on geometrically transformed images measured using the FFHQ dataset [2]. This table reports SSIM and RMSE values in addition to the PSNR and FID values already reported in Table 1 in the main paper.

Models	Metric		Translation 0	50	100	150	Rotation 10	20	30	Scaling 7/8 ↓	3/4 ↓	9/8 ↑	5/4 ↑
Im2StyleGAN [1]	PSNR	↑	**24.86**	14.98	13.81	13.43	13.96	13.30	12.74	14.60	14.70	16.96	16.06
	FID	↓	38.56	80.05	119.95	173.25	124.15	170.51	199.50	123.91	169.67	53.57	59.07
	SSIM	↑	**0.81**	0.60	0.57	0.57	0.58	0.57	0.55	0.63	0.66	0.65	0.63
	RMSE	↓	**0.08**	0.19	0.21	0.22	0.21	0.22	0.24	0.19	0.19	0.15	0.17
P-norm$^+$ [10]	PSNR	↑	21.54	14.67	13.84	13.55	14.03	13.65	13.36	14.39	14.25	16.59	15.71
	FID	↓	42.97	77.75	90.13	116.43	122.41	150.66	158.63	104.00	148.66	52.80	58.48
	SSIM	↑	0.78	0.61	0.60	0.61	0.61	0.61	0.61	0.65	0.68	0.66	0.65
	RMSE	↓	0.10	0.19	0.21	0.22	0.20	0.21	0.22	0.19	0.20	0.16	0.17
StyleGAN2 inv. [3]	PSNR	↑	19.73	16.91	15.31	15.01	16.27	14.76	14.43	14.94	15.84	**18.79**	**18.04**
	FID	↓	**31.24**	54.02	64.97	80.16	98.08	129.90	142.86	84.06	111.83	**32.78**	40.65
	SSIM	↑	0.78	0.66	0.65	0.66	0.67	0.65	0.65	0.68	0.75	**0.70**	**0.70**
	RMSE	↓	0.11	**0.15**	0.17	0.18	0.16	0.19	0.19	0.18	0.16	**0.12**	**0.13**
PSP [5]	PSNR	↑	17.55	16.62	16.01	15.54	16.52	15.92	15.70	**16.74**	17.16	17.71	17.57
	FID	↓	49.77	67.54	75.96	90.12	101.88	127.27	138.04	122.91	154.81	46.71	47.58
	SSIM	↑	0.69	0.64	0.63	0.65	0.65	0.65	0.65	0.68	0.73	0.67	0.67
	RMSE	↓	0.14	**0.15**	0.16	0.17	0.15	0.16	0.17	**0.15**	**0.14**	0.13	**0.13**
Ours (8×8)	PSNR	↑	20.18	**16.97**	16.94	16.89	16.16	15.57	14.82	16.21	16.78	18.10	17.72
	FID	↓	41.93	61.66	59.76	72.59	100.98	120.94	135.73	96.33	135.91	43.22	42.81
	SSIM	↑	0.77	**0.67**	**0.68**	0.69	0.67	0.66	0.65	0.69	0.74	0.69	0.69
	RMSE	↓	0.11	**0.15**	**0.15**	0.15	0.16	0.17	0.18	0.16	0.15	0.13	0.14
Ours (16×16)	PSNR	↑	17.84	16.83	**17.02**	**17.04**	**16.59**	**16.46**	**16.33**	16.68	**17.60**	17.35	17.46
	FID	↓	39.76	**51.54**	**54.67**	**64.87**	**64.36**	**87.26**	**88.68**	**78.11**	**111.00**	36.32	**34.25**
	SSIM	↑	0.73	**0.67**	**0.68**	**0.70**	**0.68**	**0.68**	**0.68**	**0.70**	**0.77**	0.68	0.69
	RMSE	↓	0.14	**0.15**	**0.15**	**0.14**	**0.15**	**0.15**	**0.16**	**0.15**	**0.14**	0.14	0.14

Table 3: Editing quality of different methods on geometrically transformed images measured using the FFHQ dataset [2].

while the performances of the previous methods drop significantly. In the case of scaling with scale factors smaller than 1, our 16×16 version still outperforms the other methods in most metrics. Finally, in the case of scaling with scale factors larger than 1, our 8×8 version outperforms the other methods except for StyleGAN2 inversion, which interestingly shows highly accurate editing results similar to the results of the in-range images in this case.

Fig. 1 shows a qualitative comparison of the image editing qualities on the FFHQ dataset. The top row shows the

ground-truth source images and their ground-truth editing results. The input images in the left, middle, and right are rotated, translated, and scaled, respectively. As shown in the figure, all the previous methods fail to produce accurate editing results especially in the cases of rotation and translation. On the other hand, our method produces visually more pleasing results than the others. The figure also shows the effect of the scale of the base code $\mathbf{f}$. Our 8×8 version handles the pose change operation better than our 16×16 version, while it produces less accurate results for the aging operation. In the case of scaling, all the methods generally tend to produce natural-looking editing results, but our results still look closer to the ground-truth images.

4. Reconstruction and Editing of Natural Images

Table 4 shows the full result of Table 2 in the main paper including SSIM [7] and RMSE, which compares the reconstruction qualities of different methods on natural images collected from the internet as described in the main paper. The results of IDinvert [9] are unavailable for the cat dataset because IDinvert does not provide a pre-trained encoder for the cat dataset. As shown in the table, ours (16×16) outperforms all the other methods in all the metrics for the bedroom and cat test sets, while performs comparably to Im2StyleGAN [1] for the tower test set. Figs. 2, 3, 4, 5, 6, 7, 8 and 9 show additional reconstruction and editing examples of the natural images. As shown in the figures, our method produces higher quality editing results consistently for natural images.

5. Editing Quality Evaluation using the LSUN Dataset [8]

We quantitatively and qualitatively compare the editing qualities of different GAN inversion methods on geometrically transformed natural images. We compare the methods using three categories of synthetic images: bedroom, tower and cat. For each category, we generate ground-truth editing results and the editing results of different GAN inversion methods in a similar way to the evaluation in Sec. 3. Specifically, for each of the bedroom and tower categories, we sample 50 in-domain latent codes $\mathbf{z}$ and synthesize their corresponding in-domain images using a StyleGAN [2] model pre-trained on the LSUN dataset [8] of the corresponding category. Then, we apply editing operations to each latent code and generate ground-truth editing results. For the bedroom and tower categories, we use the editing operations provided by [9]. Specifically, we use cloth, wood and indoor lighting for the bedroom category, vegetation, cloud and sunny for the tower category, resulting in three editing results for each image in each category.

We also prepare out-of-range images corresponding to the in-range images by applying a random geometric transformation. We uniformly sample a random geometric transformation from 10 transformations listed in Table 2 except for Translation with 0 pixels. To each transformed image, we apply different GAN inversion methods and obtain its latent codes. We then apply the same editing operation to the latent codes and synthesize their resulting images. We measure the difference between the ground-truth editing results and the editing results of different GAN inversion methods for the quantitative evaluation. For the cat category, we generate the ground-truth images and editing results similarly except for that we use a StyleGAN2 model pre-trained on the LSUN cat dataset and two semantic editing vectors that we found using SeFa [6].

			Dataset		
Models	Metric		Bedroom	Tower	Cat
Im2StyleGAN	PSNR	↑	19.88	**20.64**	22.90
	FID	↓	111.73	58.14	71.19
	SSIM	↑	0.62	**0.63**	0.66
	RMSE	↓	0.11	**0.10**	0.08
IDinvert	PSNR	↑	19.27	20.02	-
	FID	↓	80.21	75.59	-
	SSIM	↑	0.53	0.57	-
	RMSE	↓	0.11	**0.10**	-
Ours (16x16)	PSNR	↑	**20.21**	20.37	**24.67**
	FID	↓	**49.92**	**42.89**	**31.74**
	SSIM	↑	**0.64**	**0.63**	**0.73**
	RMSE	↓	**0.10**	**0.10**	**0.06**

Table 4: Reconstruction quality of different methods on natural images. This table reports SSIM and RMSE values in addition to the PSNR and FID values already reported in Table 2 in the main paper.

Table 5 shows a quantitative comparison. The results of IDinvert [9] are unavailable for the cat dataset because IDinvert does not provide a pre-trained encoder for the cat dataset. The table shows that our method achieves superior editing quality to previous methods for all categories, proving the effectiveness of our approach. Fig. 13 shows qualitative examples of semantic image editing on the LSUN dataset. While the other methods produce unnatural editing results, ours successfully produces high-quality natural-looking results.

6. Ablation Study - Quantitative Evaluation

We report a quantitative result of our ablation study in Table 6. In this analysis, we measure the reconstruction and editing qualities of different variants of our method. For the quantitative evaluation, we use the test set of 50 images, a StyleGAN2 [3] model pre-trained on the FFHQ dataset, and the three editing vectors used in Sec. 3. As the table shows all the variants show similar reconstruction qualities. However, in the aspect of editing quality, the variants with only the reconstruction loss in Table 6(a), and with the reconstruction loss and regularization on detail code $\mathbf{w}_{M+}$ in

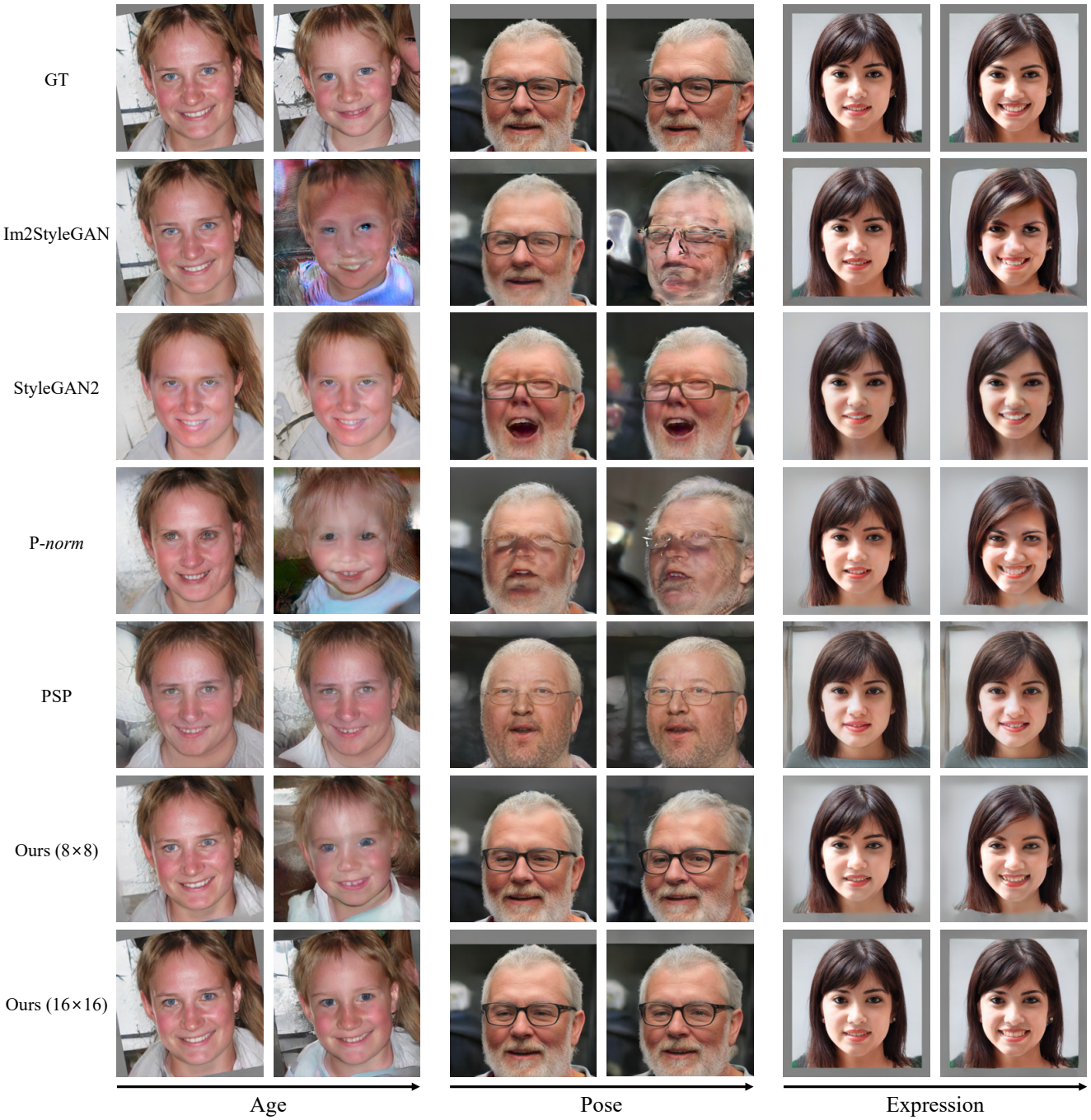

Figure 1: Qualitative examples of semantic image editing using different methods. Best viewed in zoom for details.

Table 6(b) perform poorly due to their out-of-domain latent codes. The variants with an encoder in Table 6(c) and (d) achieve better editing qualities. Table 6(c) and (d) show that our encoder loss L_{enc} based on image reconstruction performs better than the loss $\|E(I_{\downarrow}) - \mathbf{f}^{gt}\|^2$ based on the latent code distance.

7. Comparison with the alignment-and-inversion approach

One naïve approach to consider for dealing with out-of-range images is to align the input image with training distribution before inversion. Here, we compare our approach with this naïve approach with more challenging real-world examples.

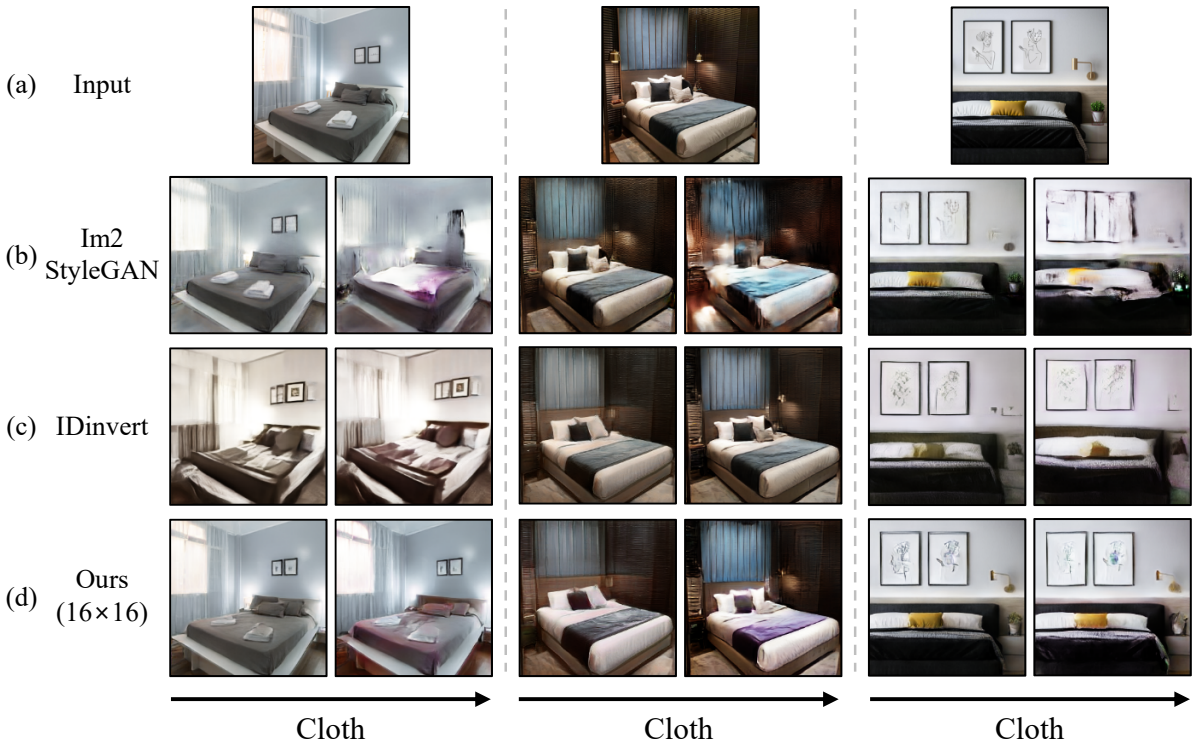

Figure 2: Qualitative comparison of the reconstruction and semantic editing (cloth) quality of different methods on natural bedroom images. The input images on the top row are collected from the internet. We use a StyleGAN [2] model pre-trained on the LSUN bedroom dataset [8].

			Dataset		
Models	Metric		Bedroom	Tower	Cat
Im2StyleGAN	PSNR	↑	15.46	13.34	12.20
	FID	↓	172.17	216.63	274.18
	SSIM	↑	0.49	0.38	0.37
	RMSE	↓	0.17	0.22	0.26
IDinvert	PSNR	↑	16.35	14.30	-
	FID	↓	92.25	127.95	-
	SSIM	↑	0.50	0.42	-
	RMSE	↓	0.16	0.20	-
Ours (16x16)	PSNR	↑	**17.50**	**16.17**	**19.10**
	FID	↓	**49.58**	**79.12**	**129.87**
	SSIM	↑	**0.57**	**0.48**	**0.62**
	RMSE	↓	**0.14**	**0.16**	**0.12**

Table 5: Editing quality of different methods on geometrically transformed images measured using the LSUN dataset [8]. For the cat dataset, the results of IDinvert [9] are not available as IDinvert does not provide pre-trained weights for its encoder network.

Our approach has a couple of clear advantages compared to the naïve approach. First, our approach is more robust as it does not require accurate alignment of an image, which can be sometimes difficult or impossible. Fig. 10 shows examples of real-world face images that are difficult to align. For the input images in Fig. 10, a face alignment method [4] used for generation of the FFHQ dataset [2] completely fails due to severe cropping and occlusion. Furthermore, simply applying previous inversion approaches without alignment leads to unacceptable reconstruction and editing results (2nd and 3rd rows in Fig. 10). Nevertheless, as our approach does not rely on the alignment, it can still successfully reconstruct the input images and provide visually appealing editing results (4th row in Fig. 10).

Second, our approach provides better reconstruction and editing quality even for input images that can be accurately aligned. This is because the base code $\mathbf{f}$ supports images not only with geometric transformations but also with more diverse local variations as the features in $\mathbf{f}$ support locally different information in contrast to detail code $\mathbf{w}^+ \in \mathcal{W}^+$. Moreover, thanks to this, the domain in the $\mathcal{F}/\mathcal{W}^+$ space that supports semantic editing is also larger than that in the $\mathcal{W}+$ space. Fig. 10 shows real-world examples to verify this. For the previous inversion methods, we first aligned the input images using a face alignment method [4]. Then, we performed inversion and editing, and transformed back the results. As shown in Fig. 10, both reconstruction and editing results of the previous methods are worse than ours due to their limited latent space and domain. For example, our 'gender' editing result looks more consistent with the input image. This result is also consistent with Table 1 in our main paper, which shows a large margin between the previous methods and ours even when the translation is 0.

	Models				Metric	Reconstruction	Editing
	Recon. loss	Reg. on $\mathbf{w}_{M+}$	Reg. on $\mathbf{f}$ w/ E trained with $\|E(I_{\downarrow}) - \mathbf{f}^{gt}\|^2$	Reg. on $\mathbf{f}$ w/ E trained with L_{enc}			
(a)	✓				PSNR ↑	<u>26.81</u>	15.43
					FID ↓	43.58	180.61
					SSIM ↑	0.80	0.65
					RMSE ↓	<u>0.05</u>	0.17
(b)	✓	✓			PSNR ↑	26.73	15.40
					FID ↓	44.10	177.44
					SSIM ↑	0.80	0.65
					RMSE ↓	<u>0.05</u>	0.17
(c)	✓	✓	✓		PSNR ↑	26.54	<u>16.67</u>
					FID ↓	<u>41.81</u>	<u>66.32</u>
					SSIM ↑	<u>0.81</u>	<u>0.68</u>
					RMSE ↓	<u>0.05</u>	<u>0.15</u>
(d)	✓	✓		✓	PSNR ↑	**27.90**	**16.86**
					FID ↓	**38.49**	**64.96**
					SSIM ↑	**0.82**	**0.69**
					RMSE ↓	**0.04**	**0.15**

Table 6: Ablation study of our method in terms of both reconstruction and editing quality.

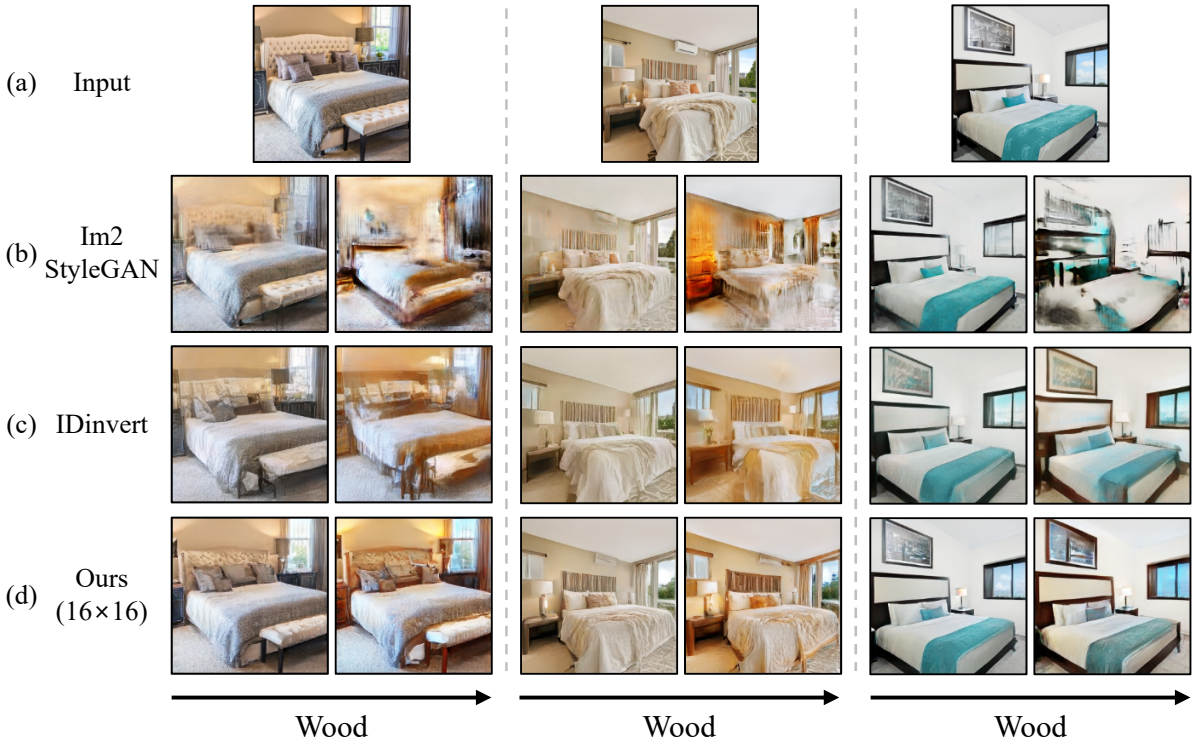

Figure 3: Qualitative comparison of the reconstruction and semantic editing (wood) quality of different methods on natural bedroom images. The input images on the top row are collected from the internet. We use a StyleGAN [2] model pre-trained on the LSUN bedroom dataset [8].

8. Limitations

Fig. 12 shows examples of the limitations discussed in the main paper. We perform reconstruction and semantic

Figure 4: Qualitative comparison of the reconstruction and semantic editing (indoor lighting) quality of different methods on natural bedroom images. The input images on the top row are collected from the internet. We use a StyleGAN [2] model pre-trained on the LSUN bedroom dataset [8].

editing on real-world face images. Fig. 12(a) shows an example with severe rotation, and Fig. 12(b) shows an example with severe non-planar rotation. The input images of both examples deviate significantly from the original training dataset. In both case, because of the superior expressive power of the $\mathcal{F}/\mathcal{W}+$ space, our method can reconstruct input image, but fails to synthesize appropriate semantic editing results.

9. Additional Examples

Fig. 14 shows additional examples of our method with various editing operations on the CelebA-HQ dataset [2]. In these examples, we use a StyleGAN2 [3] model pre-trained on the FFHQ dataset [2]. As shown in the figure, our method successfully supports various editing operations such as gender, age, race, lighting, hair color, mustache, and expression despite not using low-scale latent codes $\mathbf{w}$ for semantic image editing.

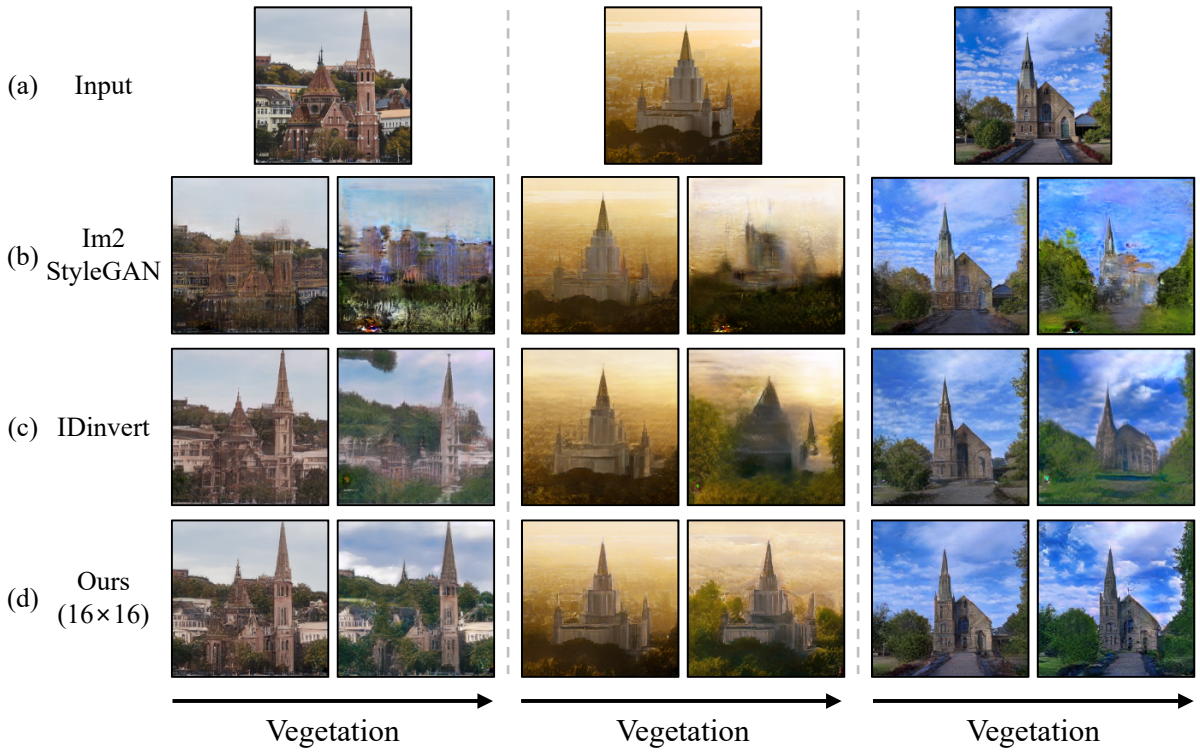

Figure 5: Qualitative comparison of the reconstruction and semantic editing (vegetation) quality of different methods on natural tower images. The input images on the top row are collected from the internet. We use a StyleGAN [2] model pre-trained on the LSUN tower dataset [8].

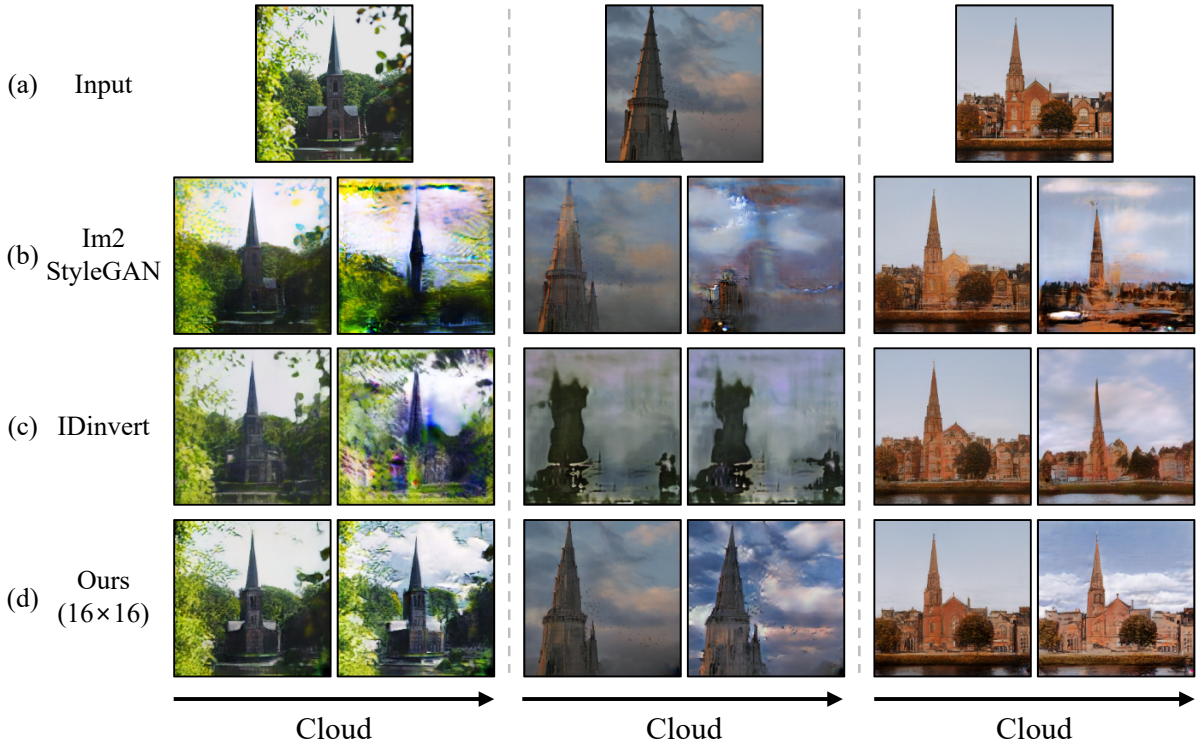

Figure 6: Qualitative comparison of the reconstruction and semantic editing (cloud) quality of different methods on natural tower images. The input images on the top row are collected from the internet. We use a StyleGAN [2] model pre-trained on the LSUN tower dataset [8].

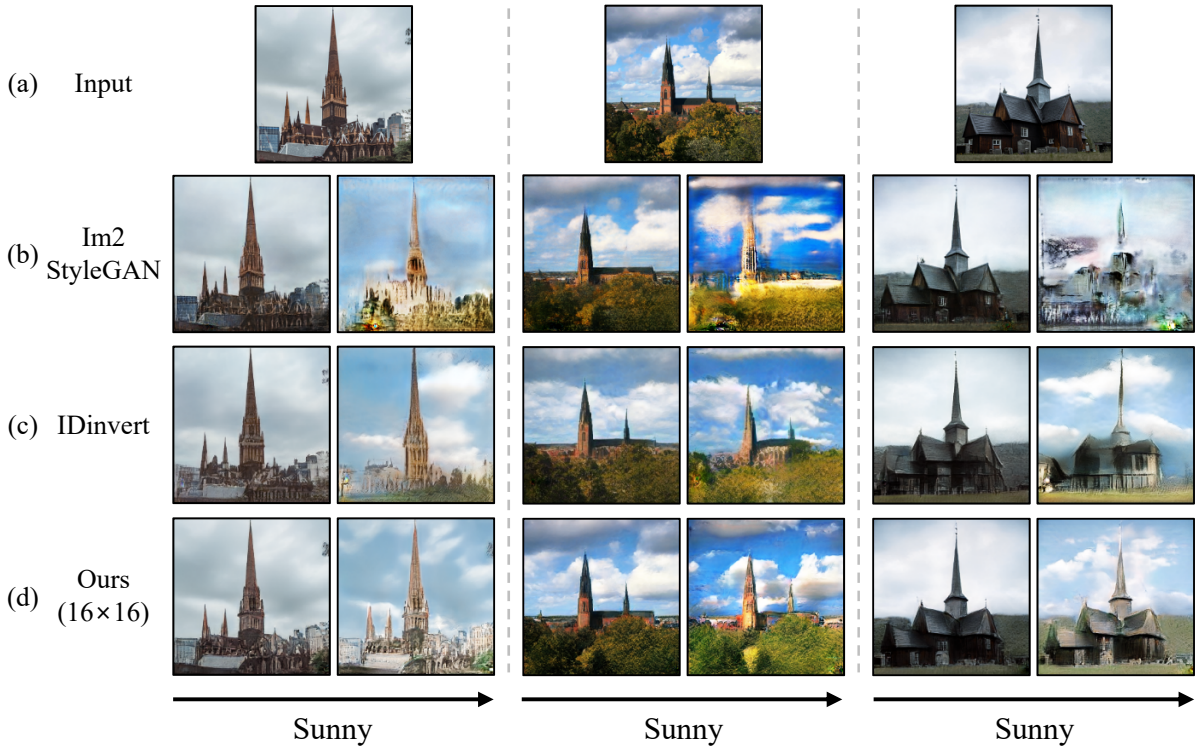

Figure 7: Qualitative comparison of the reconstruction and semantic editing (sunny) quality of different methods on natural tower images. The input images on the top row are collected from the internet. We use a StyleGAN [2] model pre-trained on the LSUN tower dataset [8].

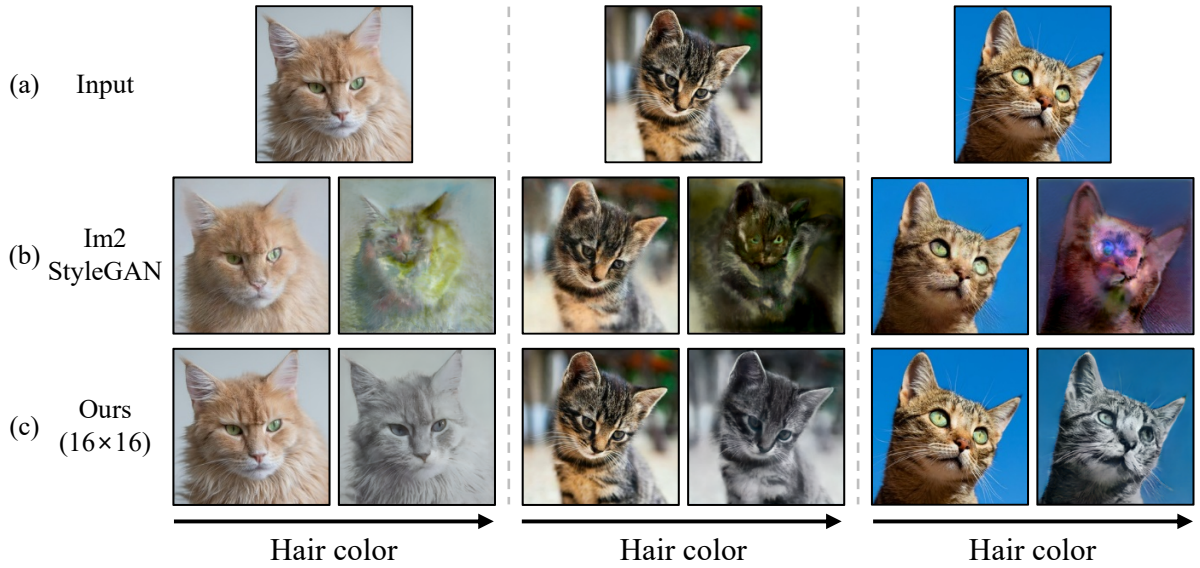

Figure 8: Qualitative comparison of the reconstruction and semantic editing (hair color) quality of different methods on natural cat images. The input images on the top row are collected from the internet. We use a StyleGAN2 [3] model pre-trained on the LSUN cat dataset [8]. For the cat dataset, the results of IDinvert [9] are not available as IDinvert does not provide pre-trained weights for its encoder network.

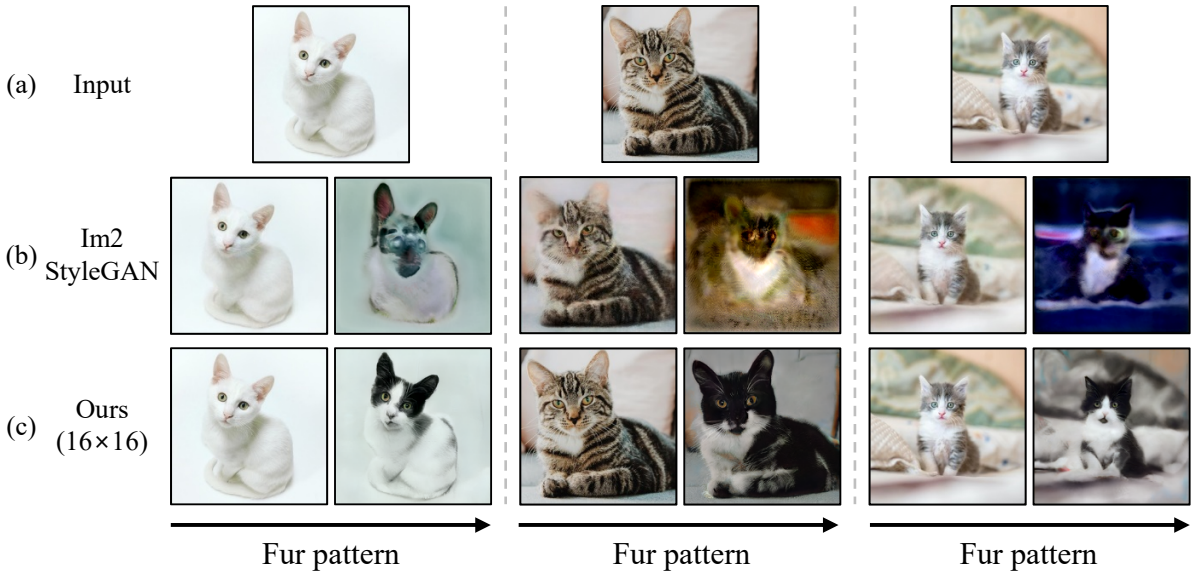

Figure 9: Qualitative comparison of the reconstruction and semantic editing (fur pattern) quality of different methods on natural cat images. The input images on the top row are collected from the internet. We use a StyleGAN2 [3] model pre-trained on the LSUN cat dataset [8]. For the cat dataset, the results of IDinvert [9] are not available as IDinvert does not provide pre-trained weights for its encoder network.

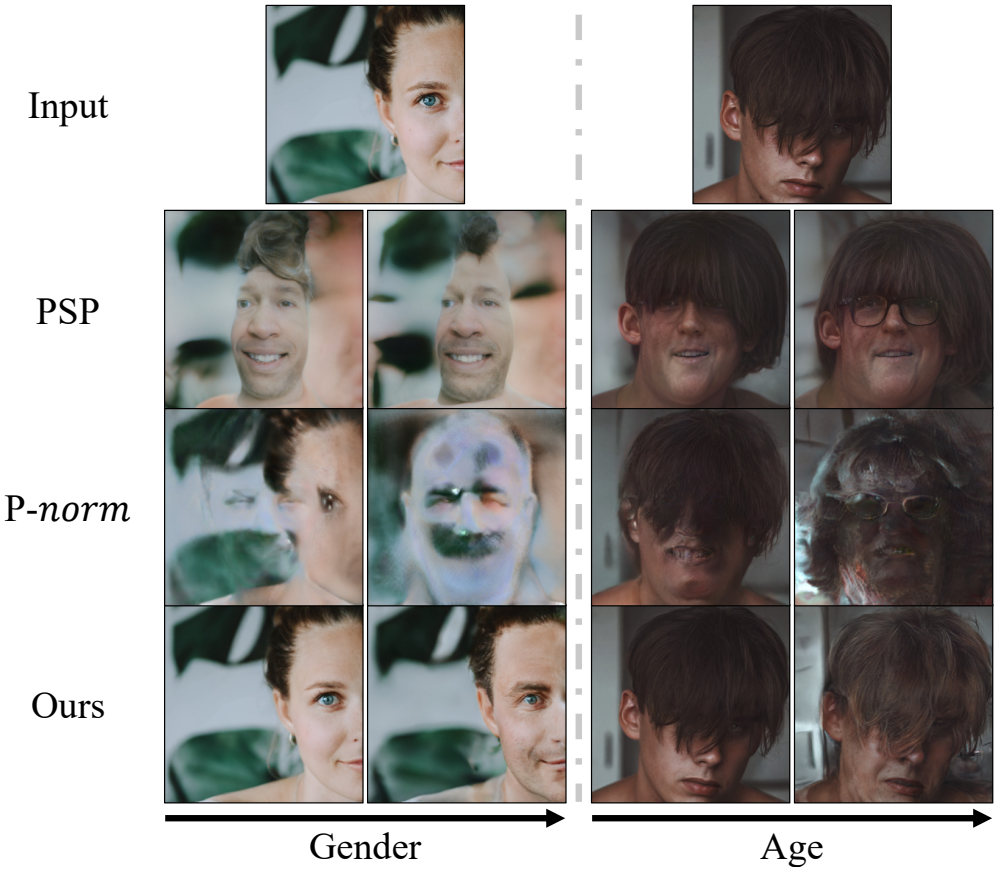

Figure 10: Examples of in-the-wild images that an alignment method fails on. For each method, the left and right images show the reconstruction and editing results, respectively.

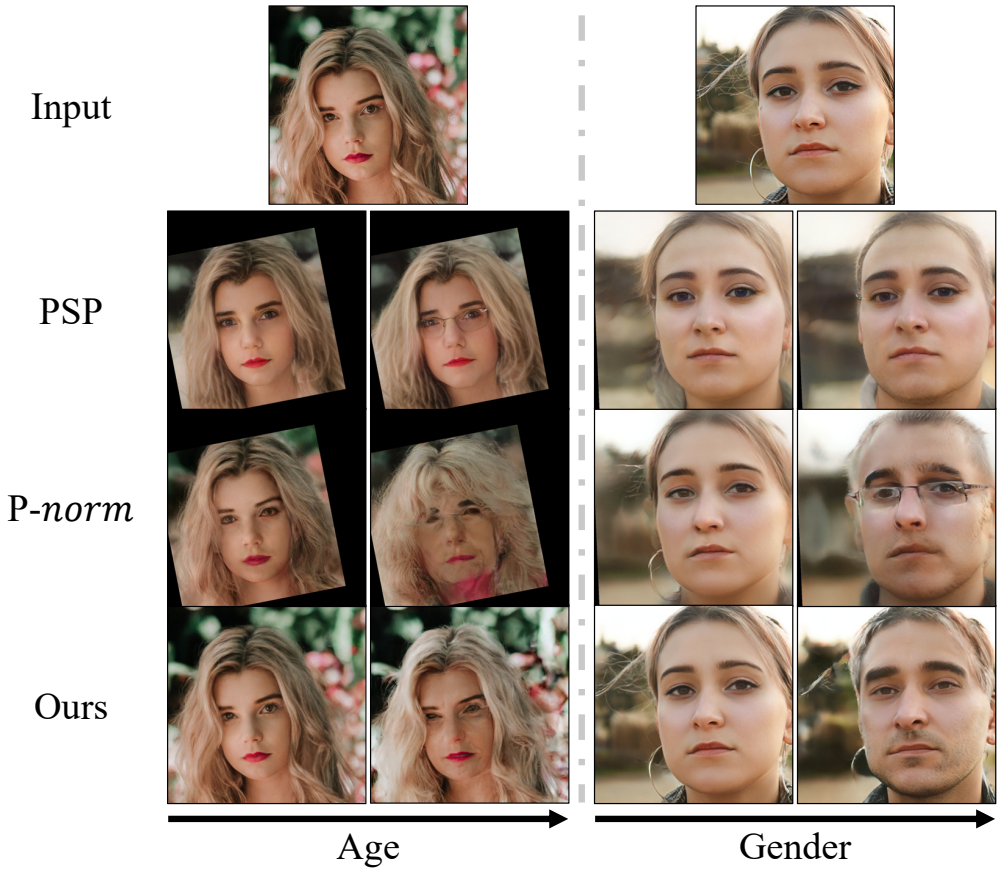

Figure 11: Examples of in-the-wild images. For each method, the left and right images show the reconstruction and editing results, respectively.

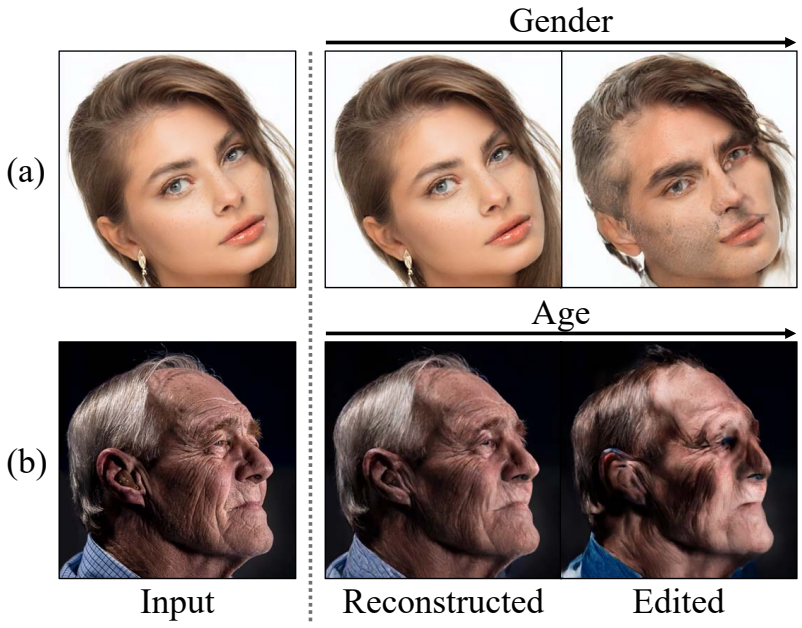

Figure 12: Failure examples: our method fails for images that deviate significantly from the original training dataset. For each row, the left most image is an input image, and the other images are its reconstruction and semantic editing results of our method, respectively.

Figure 13: Qualitative comparison of the editing quality of different methods on geometrically transformed images. The first row shows the ground-truth target images and the ground-truth editing results. The other rows show the results of inversion and manipulation of inverted latent codes using different methods. For the cat image, the results of IDinvert [9] are not available as IDinvert does not provide pre-trained weights for its encoder network.

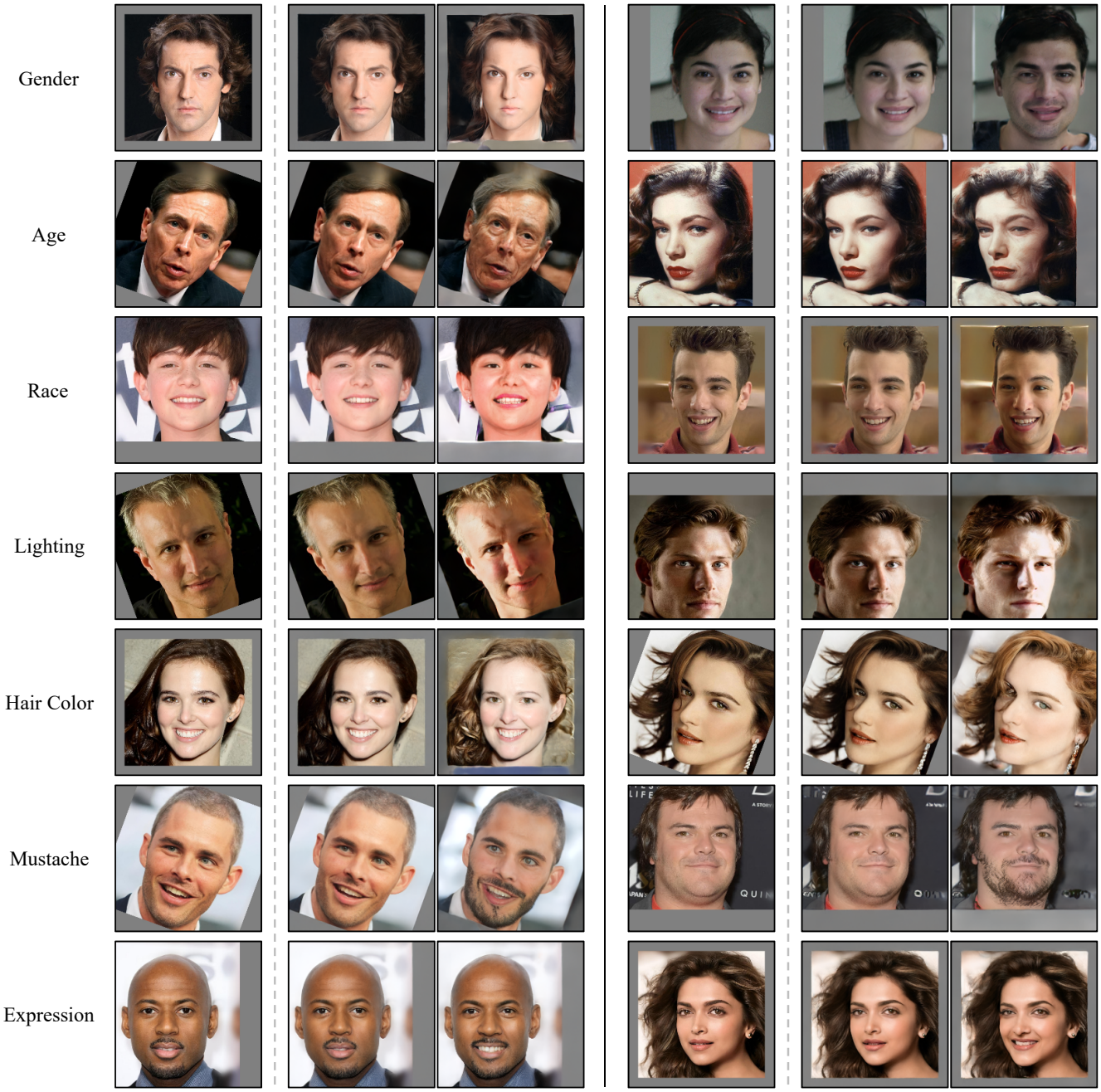

Figure 14: Diverse semantic editing results. Each row shows input images and their reconstruction and editing results obtained by our method.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *CVPR*, 2019. 1, 2, 3

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 3, 5, 6, 7, 8, 9

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 2, 3, 7, 9, 10

[4] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874, 2014. 5

[5] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 1, 2

[6] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 1, 3

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 3

[8] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3, 5, 6, 7, 8, 9, 10

[9] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 3, 5, 9, 10, 12

[10] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 1, 2