HyperInverter: Improving StyleGAN Inversion via Hypernetwork — Supplementary Material —

Tan M. Dinh Anh Tuan Tran Rang Nguyen Binh-Son Hua

VinAI Research, Hanoi, Vietnam

{v.tandm3,v.anhtt152,v.rangnhm,v.sonhb}@vinai.io

In this supplementary material, we first discuss the potential negative social impacts of our research. Then, we present another user study for the churches domain. This survey and the previous one presented in the main paper for human faces show that our method works well for both domains under human assessment. Moreover, we present the difference map visualizations to analyze the image residuals from our Phase II update. We also include additional dataset and implementation details for reproducibility. Finally, we provide extensive visual examples for further qualitative evaluation.

1. Discussion on Negative Societal Impacts

Besides some fancy and potential applications that could be commercial in the future to create numerous profits and have a large impact on society, our method can not avoid being used in ways harmful to society in some cases. For example, an attacker can use our model to create deep fake examples from reconstructing and editing a real photograph of a human face. Our inversion method can potentially yield realistic results, making them indistinguishable from real faces. Our method might also lead to the caveat of creating deep fake videos at real time since our neural network performs very fast predictions. Despite such, we believe that our method could contribute positively to the society via inspiring and advocating the development of more sophisticated detection methods to mitigate deep fakes.

2. User Study

2.1. The Churches Domain

In the main paper, we have shown the user study results for the human faces domain. Here, we also conduct an additional user study on the churches domain to verify the effectiveness of our method on this domain. The results of the user study for churches domain can be found in Figure 1. As can be seen, our method outperforms significantly e4e [10] and ReStyle [2] with a very large gap. Our method is favored in 84.3% and 77.2% of the reconstruction and editing tests, respectively, which is about 5.3× and 3.3× better than e4e and ReStyle combined for each test.

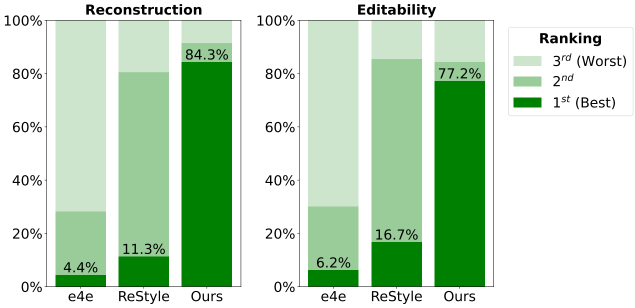


Figure 1. **User study results for Churches domain.** We reported the percentage of times testers rank the method at 1^{st} (best), 2^{nd}, and 3^{rd} (worst) based on two criteria, which are reconstruction and editing quality. As can be seen, our method outperforms e4e and ReStyle significantly.

2.2. Details of User Study

We now turn to describe the detailed setup of our user survey. Particularly, there are two separate tests, which are *reconstruction* and *editability* tests. For each test, we ask each human subject to rank the image from the scale of 1^{st} (best) to 3^{rd} (worst) based on some criteria, which depend on the corresponding domain and would be described below.

Human faces. We first choose 30 images from the images of the CelebA-HQ [4, 6] test set and the collected high-resolution images from the Internet as the input images. Then, we use the candidate methods to reconstruct and edit these input images. Each question is a record of input image and the reconstructed/edited images by the candidate methods. For reconstruction, we ask each participant to rank the images on: (1) the ability to preserve identity; (2) the ability to reconstruct the details such as background, makeup, shadow, lipstick, hat, etc; (3) image aesthetics. For editability, we request the participant to rank: (1) the level of identity preservation compared to the input image; (2) the ability to preserve the details (except for the editing attribute) of the original photo in the edited image – the more details the better. We recruited a total of 38 and 30 participants to cast $1,140$ and 900 votes for the reconstruction and editing tests, respectively.

Churches. We choose randomly 30 images from the test set of LSUN Churches [11] dataset as the input images. Then, we use the candidate methods to reconstruct and edit these input images. Each question is a record of the input image and the reconstructed/edited images by the candidate methods. For reconstruction, we ask each human subject to rank on: (1) the ability to restore as much as details of the input image in the reconstructed image; (2) image aesthetics. For editability, we request the participant to rank on the ability to preserve as much detail of the input image as possible (except for the editing attribute). We recruited 35 and 26 participants, which results in $1,050$ and 780 votes for the reconstruction and editing tests, respectively.

3. Visualization and Analysis

3.1. Distribution of the predicted residual weights

In the main paper, we have shown an analysis on the distribution of residual weights predicted by the hypernetworks in Phase II for the human facial domain. Here, we also provide an equivalent analysis for the churches domain. Figure 2 presents this visualization. As can be seen, the churches domain results are not completely the same as those for the human facial domain. The observation for human faces still preserved in the churches domain is that the *main conv* weights contribute significantly compared to weights of *torgb* block. The difference here is that the residual weight updates occur at most layers in churches instead of concentrating at the last layer as in human faces. This aligns with the fact that the church domain is more diverse and challenging to reconstruct than the human facial domain, thus requiring updates at both low- and high-frequency signals. For face images, the initial images after Phase I are already very close to the inputs, and Phase II focuses on restoring the fined-grained details.

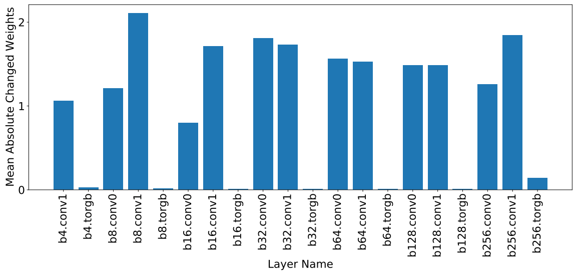
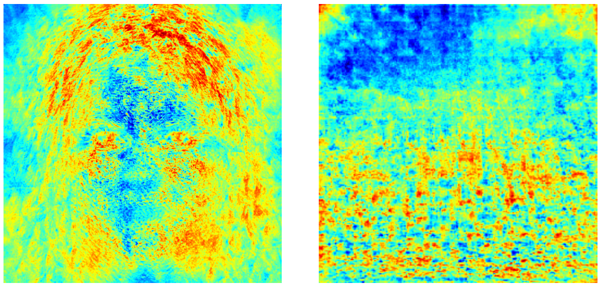


Figure 2. Visualizing the statistic of residual weights predicted by the hypernetworks on the churches domain. Compared to faces, our hypernetworks provides more uniform weight updates across layers, which means updates are required on both low- and high-frequency signals of the images.



(a) **Human Faces** (b) **Churches**

Figure 3. **Heat-map visualization** of the difference between the output images in Phase I and the final images after both phases of our method, averaged over images from the test set. For faces, large changes are focused around eye and hair. For churches, low changes are in the sky region. Blue indicates a small change, whereas red denotes a large change.

3.2. Difference maps

We also conduct an experiment to visualize the difference between the initial image from Phase I and the final image from both phases to analysis which regions change most in the refinement process of Phase II. Specifically, given the input image $x^{(i)}$, the Phase I's output image $\hat{x}_w^{(i)}$, and the final reconstructed image $\hat{x}^{(i)}$, where $i \in \{1..N\}$, N is the number of test images, we compute the difference map $m^{(i)}$ by subtracting two reconstructed images and take absolute values, which means $m^{(i)} = |\hat{x}^{(i)} - \hat{x}_w^{(i)}|$. We then compute the mean difference map $\overline{m}$ by averaging all difference maps $\{m^{(1)}, m^{(2)}, , m^{(N)}\}$. Next, we convert $\overline{m}$ from the RGB image to the grayscale image and visualize it as the heat map. In this experiment, we use all $2,824$ images from the CelebA-HQ test set and all 300 images from LSUN Church test set to analyze for human faces and churches domains, respectively. Figure 3 presents the results. As can be seen, for human faces, the heat map in Figure 3-a reveals that our method focuses mainly on refining the regions having many fine-grained details such as hair, cheek, beard, eye in Phase II. For the churches domain, since the images from this domain are not aligned and very diverse in terms of structures, its heat map in Figure 3-b appears more random. However, we can see that the top left region of the map does not change much, which is often the sky that is already well reconstructed from Phase I.

To gain better insights on individual cases, we also provide heat-maps for each image in Figure 4 and 5.

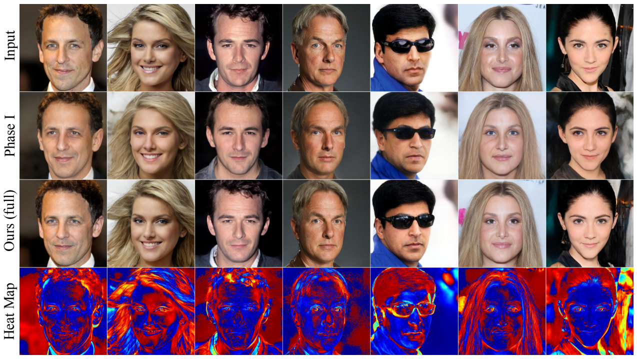


Figure 4. Visualizing the effectiveness of our phase II in bringing back the information of input image missed in the initial image on the human facial domain. We also include the difference map for reference which region change most in the image after phase II. Blue indicates a small change, while red denotes a large change. Best viewed in zoom.

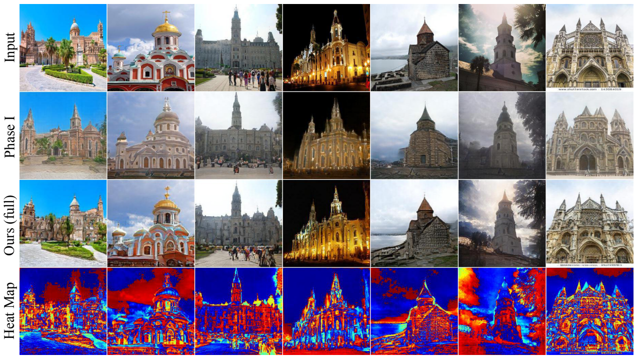


Figure 5. Visualizing the effectiveness of our phase II in bringing back the information of input image missed in the initial image on the churches domain. We also include the difference map for reference which region change most in the image after phase II. Blue indicates a small change, while red denotes a large change. Best viewed in zoom.

4. Additional Experimental Details

4.1. Datasets

In this section, we provide more details about the datasets we employ in conducting our experiments.

Human faces. We use $70,000$ images from the FFHQ [5] dataset as our training set and $2,824$ images from the official test set of CelebA-HQ [4, 6] as our test set. These datasets contain high-quality real-world face images at resolution 1024x1024. All faces are aligned to the center of the images.

Churches. We choose the Churches domain to test our method on images of natural outdoor scenes. These images are more diverse than human faces and thus considered more challenging. We use LSUN Church [11] in this task. The resolution of images is 256×256. We use $126,227$ and 300 images from the official train/test split of LSUN Church for training and testing, respectively.

4.2. Detailed architecture of E_1 and E_2 encoders.

As mentioned previously in the main paper, for E_1 and E_2 encoders, we adopt the design of [7, 10] as the main backbone with some modifications since the superior performance of the original network. For E_1 encoder, we utilize the $\mathcal{W}$ encoder of these networks without modifications. For E_2 encoder, since our network outputs the intermediate features having the similar size with the output of the $\mathcal{W}^+$ encoder of [7, 10]. Therefore we also leverage this design for E_2 with some modifications. Particularly, the architecture of this encoder is the FPN-based design [7]. We modify the *map2style* block to output the feature tensor with the dimension of $512 \times 8 \times 8$ instead of a 512 vector of original backbone. Figure 6 gives the network design of FPN-based network and our modifications.

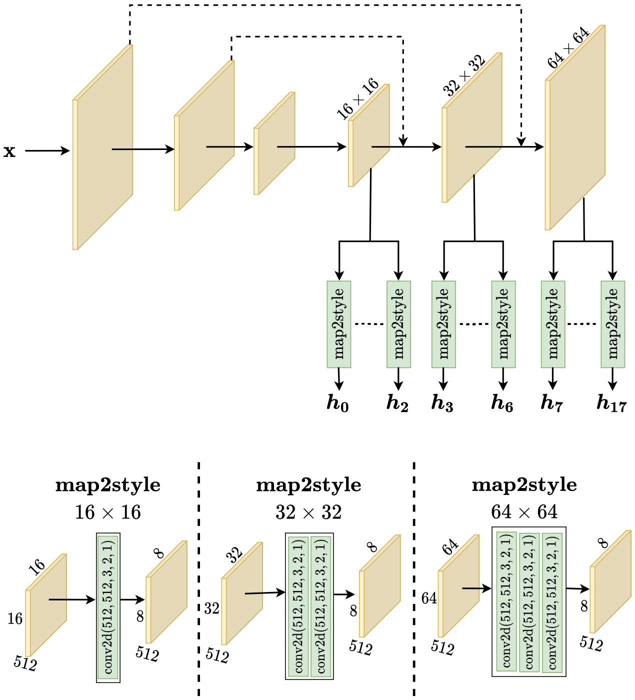


Figure 6. The FPN-based encoder network proposed by Richardson et al. [7], which has been used popularly by many previous encoder-based GAN inversion works, and our modifications in *map2style* block to output $512 \times 8 \times 8$ tensor instead of 512 vector as the original backbone. Diagram notation: conv2d(in_channels, out_channels, kernel_size, stride, padding). For simplicity, we omit the LeakyReLU activations after each conv2d layer in the figure.

5. Additional Qualitative Results

We now turn to provide more qualitative examples on reconstruction, editability and also interpolation to further demonstrate the superiority of our method. The short descriptions for the figures are shown below.

- Figure 7 compares the **reconstruction** results of our method with existing state-of-the-art inversion techniques, including encoder-based [2, 7, 10], optimization-based [1] and two-stage methods [8] for the *human facial* domain on the input images taken from the CelebA-HQ [4, 6] test set.
- Figure 8 and 9 compare the **editing** results of our method with the existing state-of-the-art encoder-based [2, 7, 10] inversion techniques for the *human facial* domain on the input images taken from the CelebA-HQ [4, 6] test set.
- Figure 10 compares the **editing** results of our method with PTI [8] and SG2 $\mathcal{W}^+$ [1] for the *human facial* domain on the input images taken from the CelebA-HQ [4, 6] test set.
- Figure 11 compares the **reconstruction** results of our method with existing state-of-the-art inversion techniques, including encoder-based [2, 7, 10], optimization-based [1] and two-stage methods [8] for the *churches* domain on the input images taken from the LSUN Church [11] test set.
- Figure 12 compares the **editing** results of our method with the existing state-of-the-art encoder-based [2, 7, 10] inversion techniques for the *churches* domain on the input images taken from the LSUN Church [11] test set.
- Figure 13 compares the **editing** results of our method with PTI [8] and SG2 $\mathcal{W}^+$ [1] on the *churches* domain on the images taken from the LSUN Church [11] test set.
- Figure 14, and 15 show additional results for **real-world image interpolation** of our proposed pipeline, which interpolates both latent codes and generator weights compared to the common-used pipeline, which interpolates latent codes only.

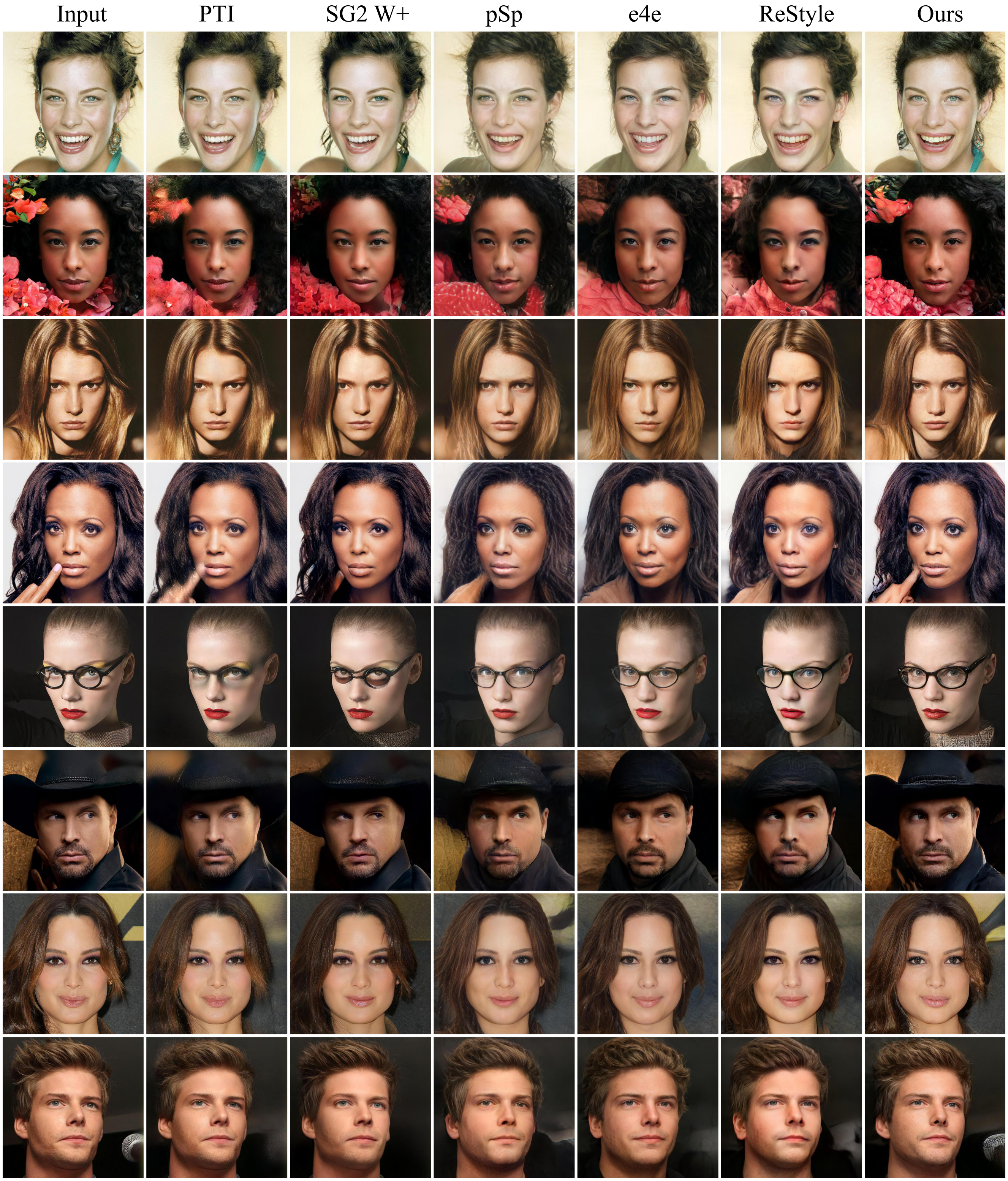


Figure 7. More visual examples shown the **reconstruction comparison** of our method with other *encodered-based approaches*: pSp [7], e4e [10], ReStyle [2]; *optimization-based methods*: SG2 $\mathcal{W}^+$ [1]; *two-stage works*: PTI [8] on the human facial domain. The input images are taken from the CelebA-HQ [4, 6] test set. Best viewed in zoom.

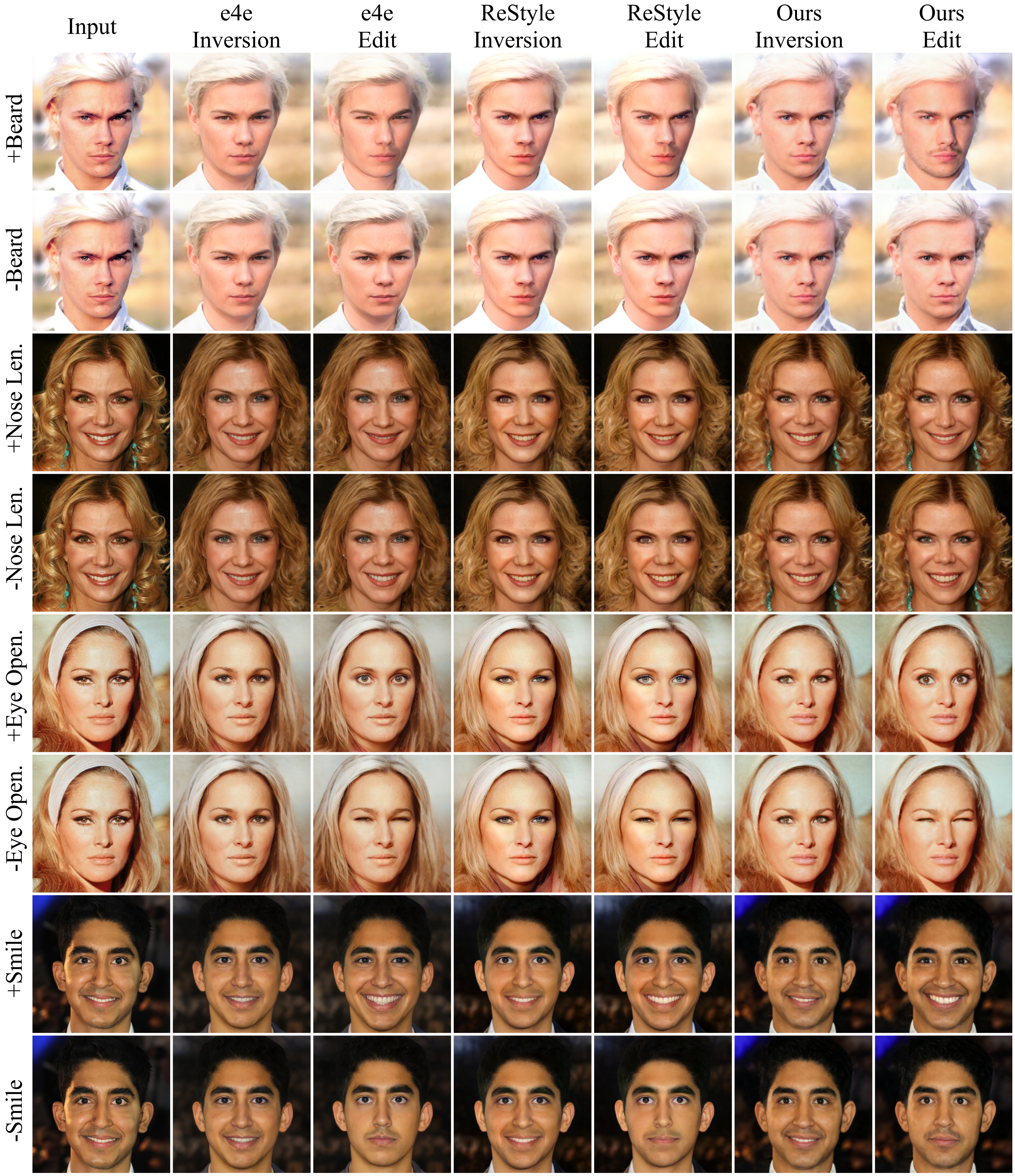


Figure 8. More visual examples shown the **editability comparison** of our method with the existing *encodered-based approaches*, which are e4e [10], ReStyle [2] on the human facial domain. The input images are taken from the CelebA-HQ [4, 6] test set. The smile direction is obtained from InterFaceGAN [9], whereas other directions are borrowed from GANSpace [3]. Best viewed in zoom.

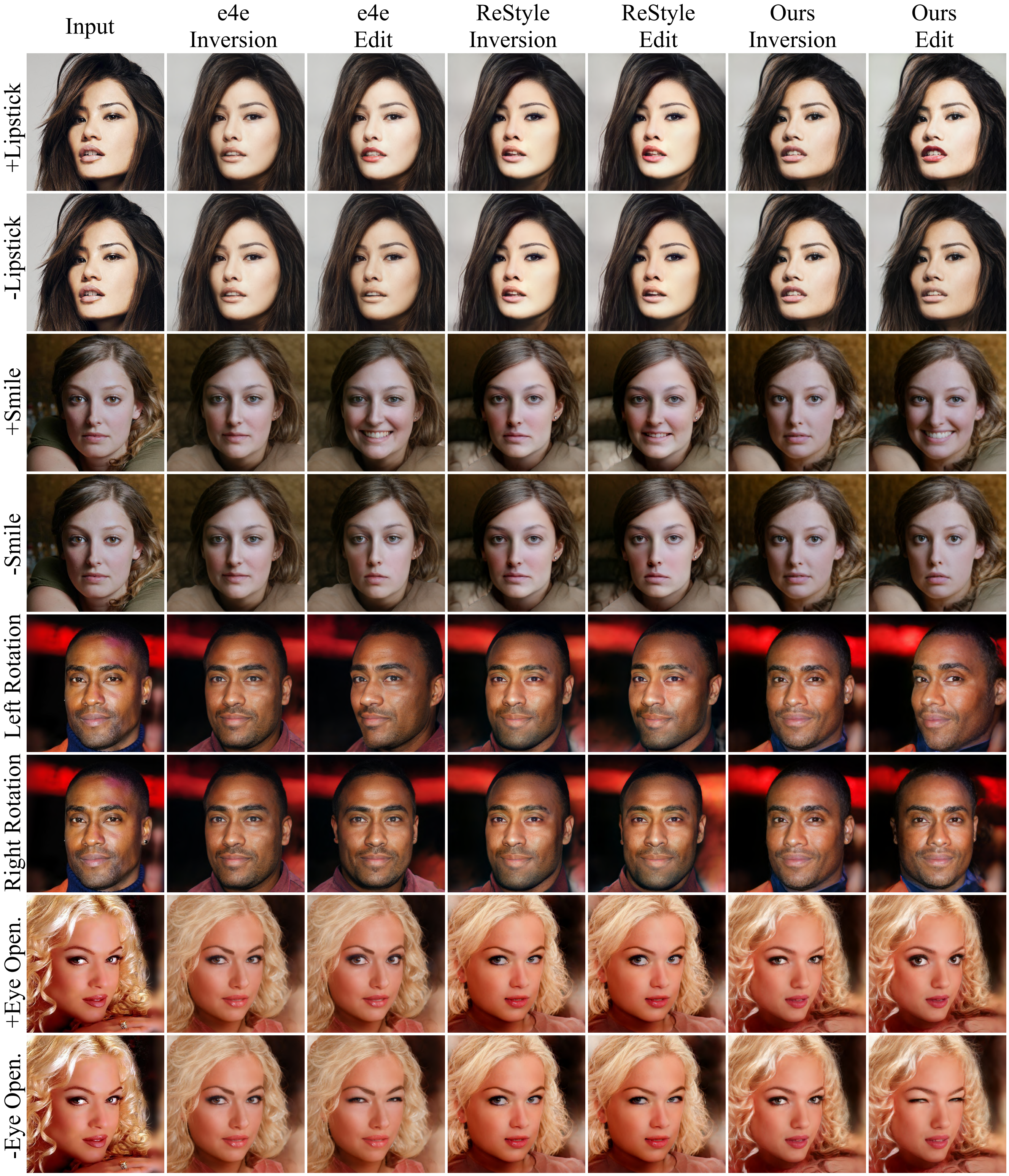


Figure 9. More visual examples shown the **editability comparison** of our method with the existing *encodered-based approaches*, which are e4e [10], ReStyle [2] on the human facial domain. The input images are taken from the CelebA-HQ [4, 6] test set. The rotation and smile directions are obtained from InterFaceGAN [9], whereas other directions are borrowed from GANSpace [3]. Best viewed in zoom.

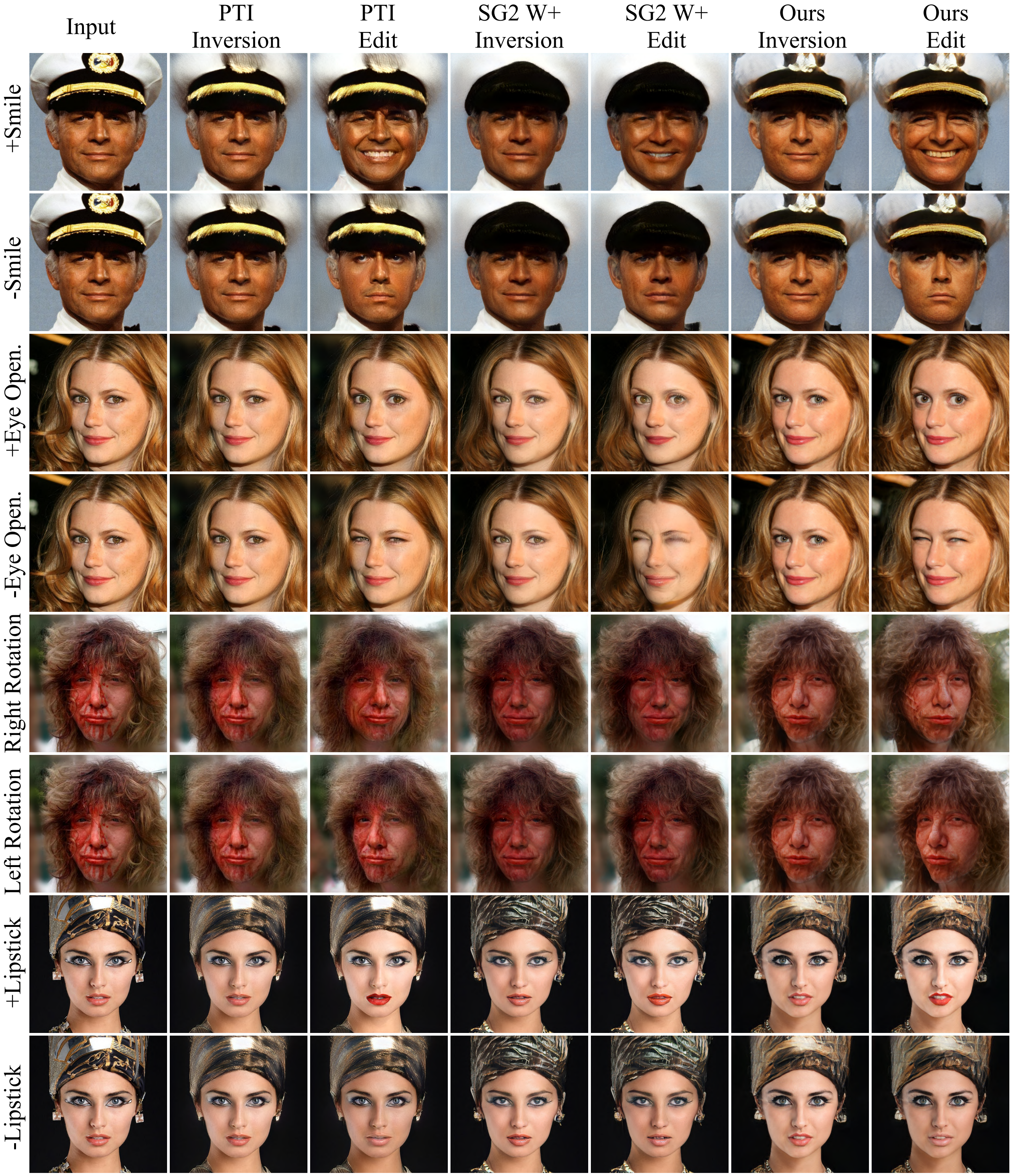


Figure 10. More visual examples shown the **editability comparison** of our method compared to PTI [8] and SG2 $\mathcal{W}^+$ [1] on the human facial domain. Recall that such two methods require the optimization process and/or generator fine-tuning in the inference time, therefore, they run very slow. The input images are taken from the CelebA-HQ [4, 6] test set. The rotation and smile directions are obtained from InterFaceGAN [9], whereas other directions are borrowed from GANSpace [3]. Best viewed in zoom.

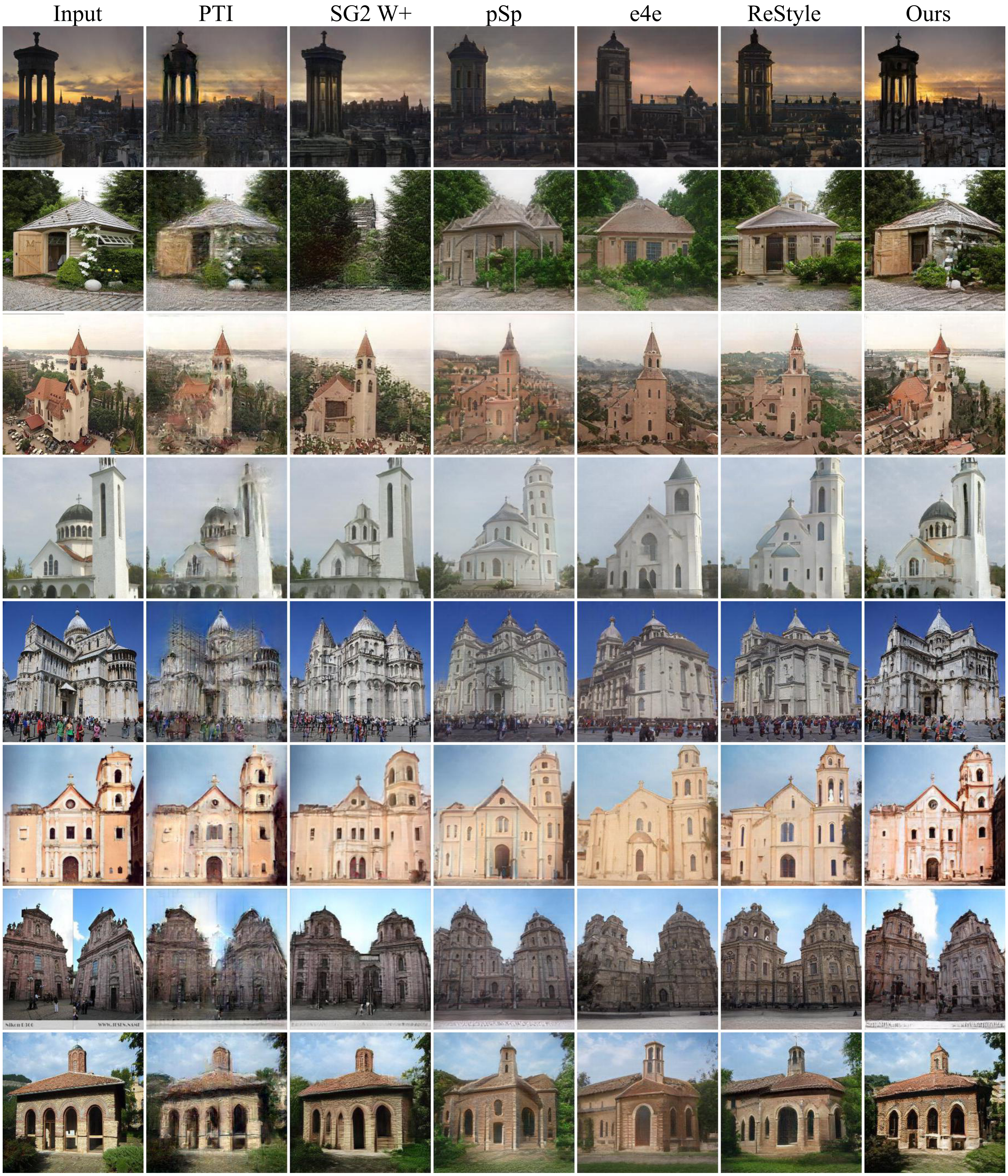


Figure 11. More visual examples shown the **reconstruction comparison** of our method with other *encodered-based approaches*: pSp [7], e4e [10], ReStyle [2]; *optimization-based methods*: SG2 $\mathcal{W}^+$ [1]; *two-stage works*: PTI [8] on the churches domain. The input images are taken from the LSUN Church [11] test set. Best viewed in zoom.

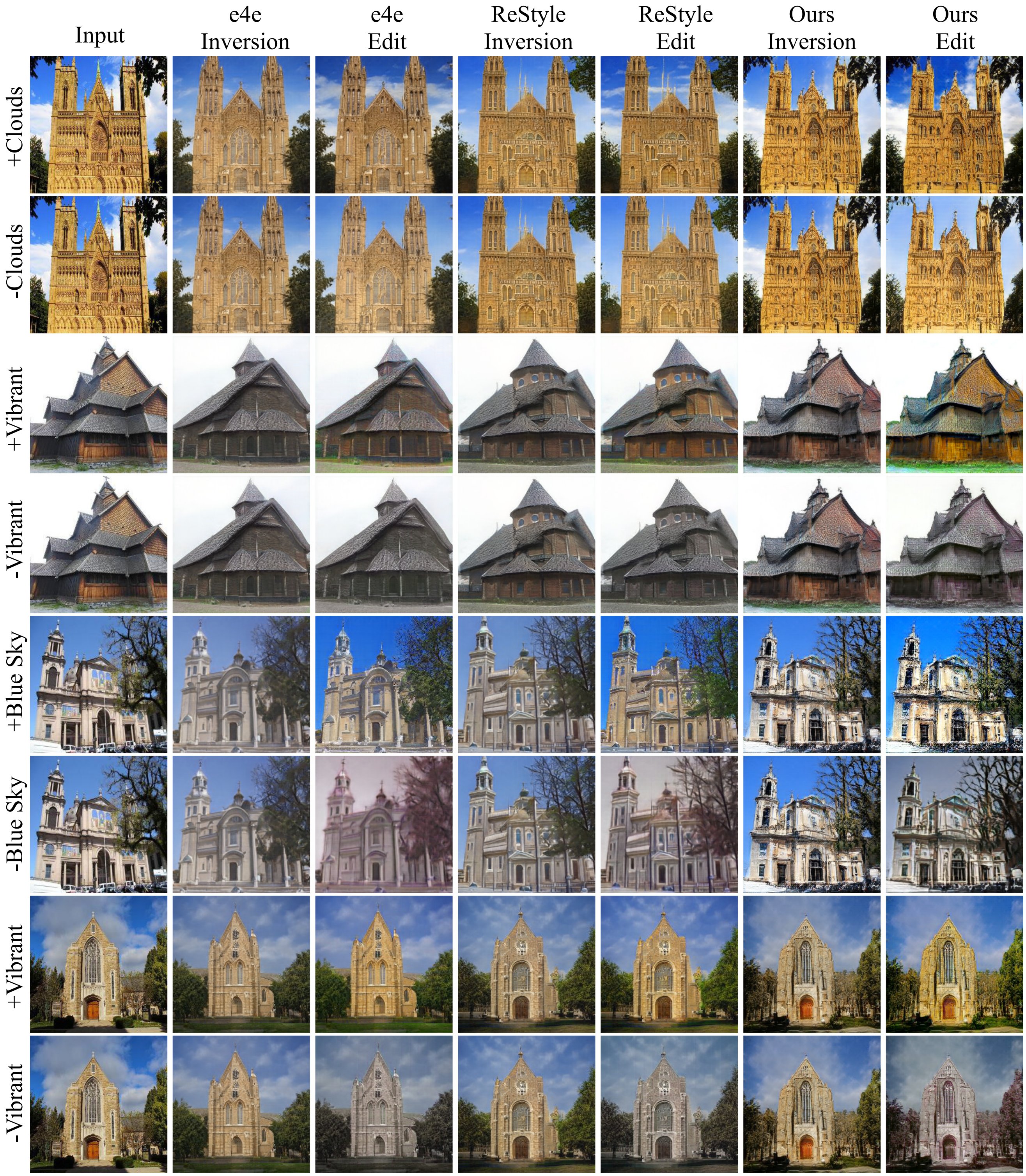


Figure 12. More visual examples shown the **editability comparison** of our method with the existing *encodered-based approaches*, which are e4e [10], ReStyle [2] on the churches domain. The input images are taken from the LSUN Church [11] test set. The editing directions are obtained from GANSpace [3]. Best viewed in zoom.

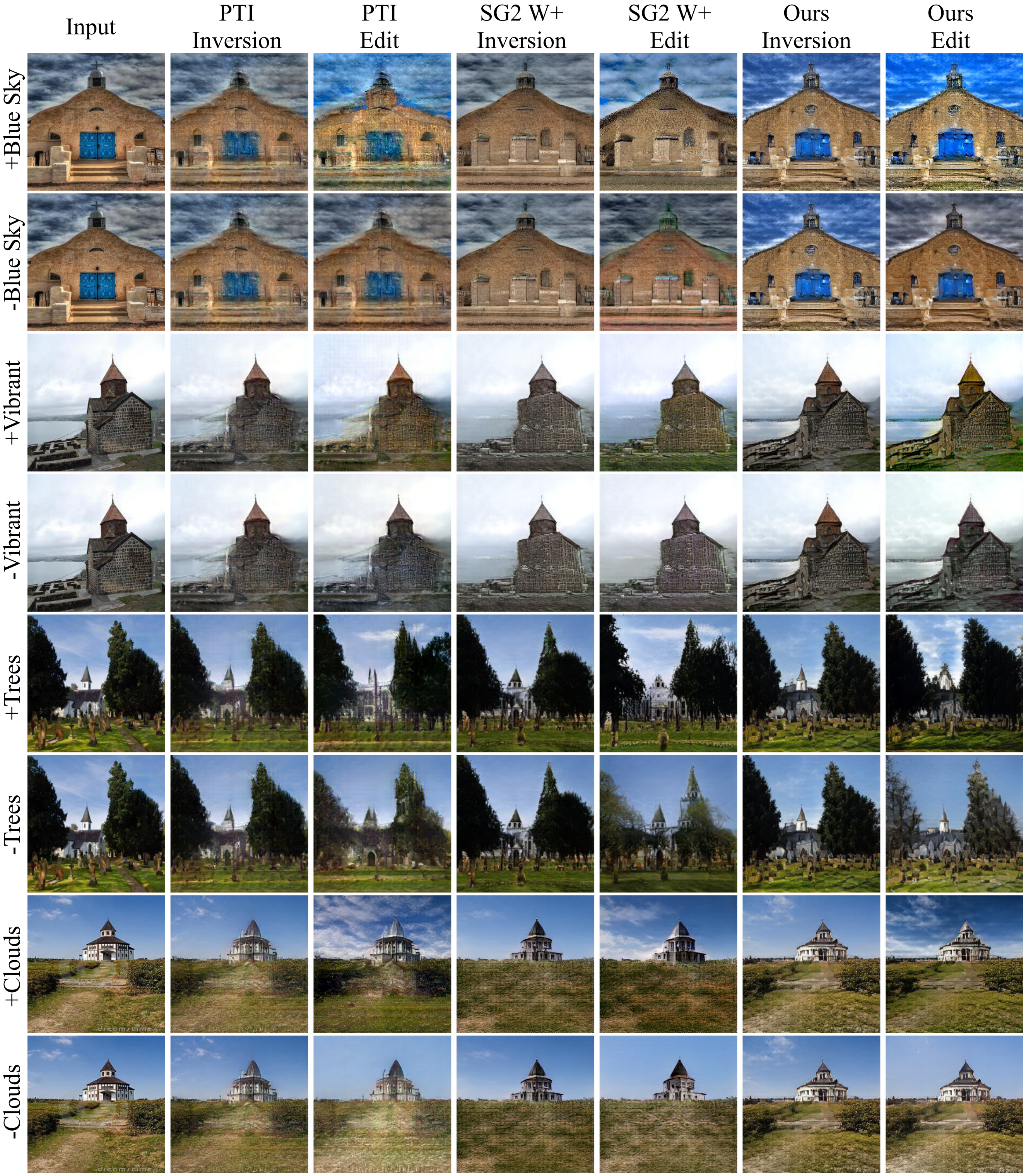


Figure 13. More visual examples shown the **editability comparison** of our method compared to PTI [8] and SG2 $\mathcal{W}^+$ [1] on the churches domain. Recall that such two methods require the optimization process and/or generator fine-tuning in the inference time, therefore, they run very slow. The input images are taken from the LSUN Church [11] test set. The editing directions are obtained from GANSpace [3]. Best viewed in zoom.

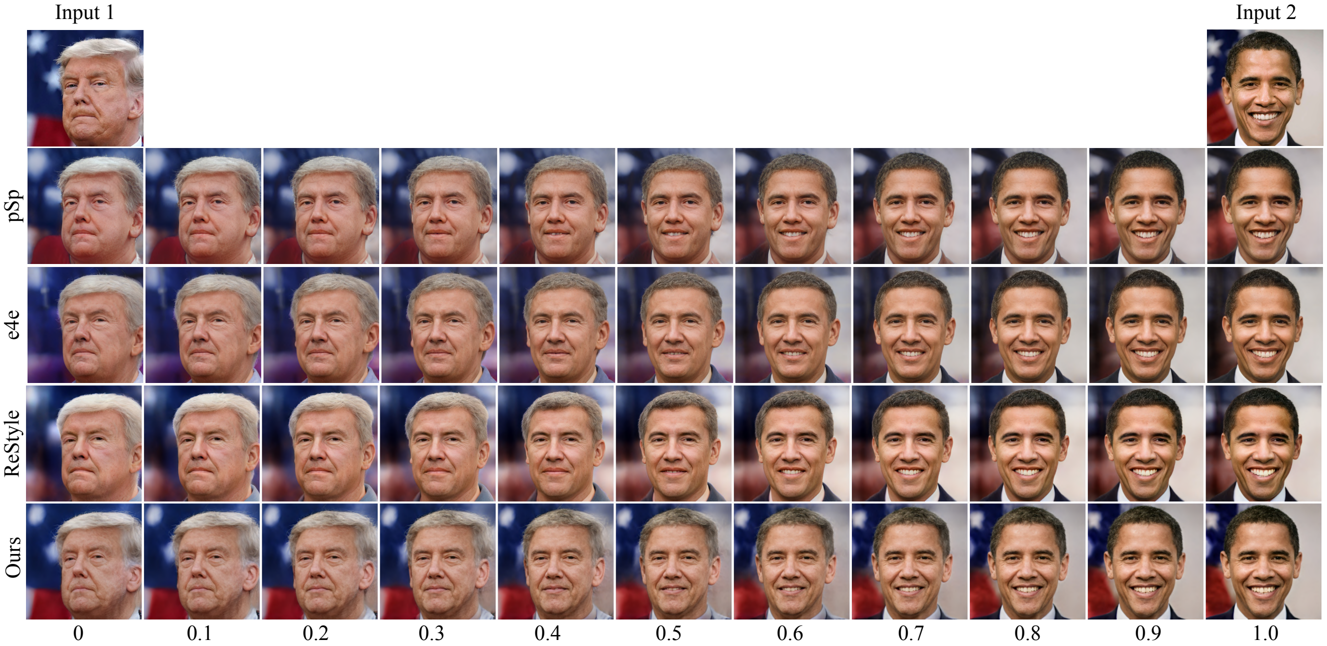


Figure 14. The additional results for **real-world image interpolation** of our method compared to the existing state-of-the-art encoder-based inversion techniques. The input images are taken from the Internet. Best viewed in zoom.

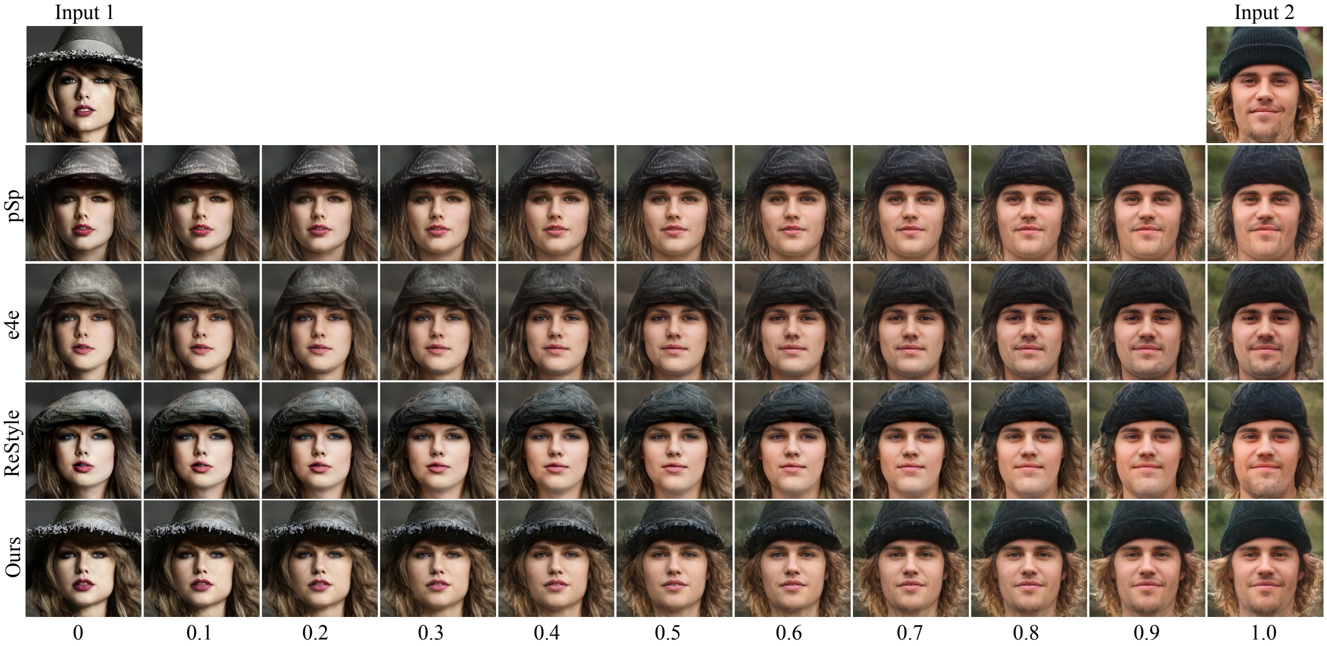


Figure 15. The additional results for **real-world image interpolation** of our method compared to the existing state-of-the-art encoder-based inversion techniques. The input images are taken from the Internet. Best viewed in zoom.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 4, 5, 8, 9, 11

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *ICCV*, 2021. 1, 4, 5, 6, 7, 9, 10

[3] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020. 6, 7, 8, 10, 11

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 1, 3, 4, 5, 6, 7, 8

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

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 3, 4, 5, 6, 7, 8

[7] 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. 3, 4, 5, 9

[8] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021. 4, 5, 8, 9, 11

[9] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 6, 7, 8

[10] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *TOG*, 2021. 1, 3, 4, 5, 6, 7, 9, 10

[11] 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. 2, 3, 4, 9, 10, 11