# PIE: Portrait Image Embedding for Semantic Control –Supplemental Document–

AYUSH TEWARI, MOHAMED ELGHARIB, and MALLIKARJUN B R, Max Planck Institute for Informatics, SIC
FLORIAN BERNARD, Max Planck Institute for Informatics, SIC and Technical University of Munich
HANS-PETER SEIDEL, Max Planck Institute for Informatics, SIC
PATRICK PÉREZ, Valeo.ai
MICHAEL ZOLLHÖFER, Stanford University
CHRISTIAN THEOBALT, Max Planck Institute for Informatics, SIC

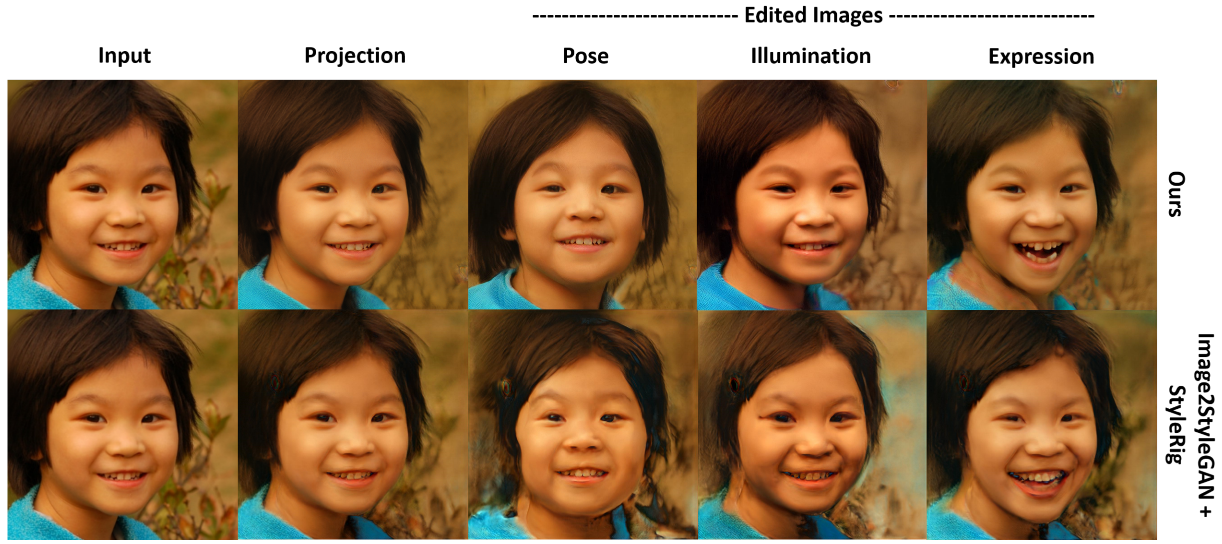
---------------------------- Edited Images ----------------------------
Input
Projection
Pose
Illumination
Expression
Ours
Image2StyleGAN + StyleRig

Fig. 1. We present an approach for embedding portrait images in the latent space of StyleGAN [Karras et al. 2019] (visualized as "Projection") which allows for intuitive photo-real semantic editing of the head pose, facial expression, and scene illumination using StyleRig [Tewari et al. 2020]. Our optimization-based approach allows us to achieve higher quality editing results compared to the existing embedding method Image2StyleGAN [Abdal et al. 2019].

In this supplementary document, we include more implementation details, ablative analysis and comparisons.

## 1 IMPLEMENTATION DETAILS

We first transform the 3DMM parameters before using them as input to RigNet. The euler angles are transformed into rotation matrices. Expression parameters are tranformed into per-vertex displacements due to expressions. A learnable linear encoder maps these displacements into a 32 dimensional vector, which is then used as an input to RigNet. This leads to better results, compared to directly using the expression parameters as input. We do not transform the illumination parameters. The spherical harmonic coefficients are directly used as inputs in the network. RigNet is implemented as a linear two-layer perceptron (MLP) [Tewari et al. 2020].

## 2 FEEDBACK

We update the target 3DMM parameters used as input to RigNet using a simple feedback loop explained in the main paper. Feedback allows us to obtain more accurate editing results, see Tab. 1. However, this comes at the cost of higher average recognition error.

## 3 COMPARISONS

We provide more qualitative comparisons for pose editing in Fig. 3 and illumination editing in Fig. 2.

### 3.1 User study

We evaluate the cross-identity pose editing and relighting capabilities of our approach through a user-study. We also evaluate the realism of Wiles et al. [2018], Wang et al. [2019a], Siarohin et al. [2019] and Zhou et al. [2019]. For every method, we use 7 images

| | with feedback | without feedback |
|---|---|---|
| Editing Error (rad) ↓ | **0.08** | 0.16 |
| Recognition Error ↓ | 42.82 | **25.97** |

Table 1. We quantitatively evaluate the importance of feedback for pose editing. All numbers are averaged over more than 2500 pose editing results. Editing error is measured as the angular difference between the desired and achieved face poses. Recognition error measures the value of the facial recognition error for the edited images. Feedback allows us to obtain more accurate editing, at the cost of higher recognition errors.

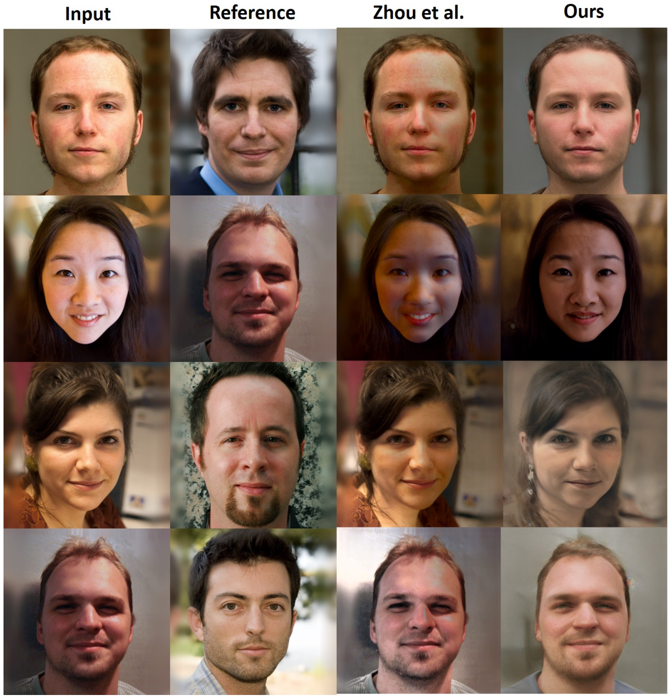

Fig. 2. Comparison of our relighting results with Zhou et al. [2019]. The illumination in the reference image is transferred to the input. Our results are more natural and achieve more accurate relighting. We can edit colored illumination while Zhou et al. [2019] can only edit monochrome light. In addition, we can also edit the head pose and facial expressions, while Zhou et al. [2019] is trained only for relighting.

each for pose editing and relighting. The participants were asked to rate their agreement to the statement "This image looks realistic to me" using either "strongly agree", "agree", "don't know", "disagree" or "strongly disagree" with realism scores from 5 to 1. Participants were asked to focus on the head region. We also included 13 real images in the study. In total, 61 subjects participated in the study. Table 2 summarizes the results. Our pose editing and relighting capabilities outperform related techniques significantly. Note that a baseline error exists where real images were rated real (realism score > 3) only 89.2% of the time with an average realism score of 4.37.

| | % of participants rated images as real | average realism score |
|---|---|---|
| Wiles et al. [2018] | 2.6 | 1.21 |
| Wang et al. [2019a] | 11.7 | 1.95 |
| Siarohin et al. [2019] | 2.6 | 1.24 |
| PIE pose editing | **44.3** | **3.02** |
| Zhou et al. [2019] | 42.8 | 2.96 |
| PIE relighting | **62.0** | **3.60** |
| real images | 89.2 | 4.37 |

Table 2. Summary of a user study examining the pose editing and relighting capabilities of our approach. Our approach outperforms state-of-the-art techniques significantly. Realism score is calculated on a discrete scale that reports the agreement to the statement "This image looks realistic to me". The scale is "strongly agree" (5), "agree" (4), "don't know" (3), "disagree" (2) or "strongly disagree" (1). Note that real images are rated real only 89.2% of times with an average realism score of 4.37. This highlights a baseline error.

## REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *The IEEE International Conference on Computer Vision (ICCV)*.

Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-guided GANs for single-photo facial animation. *ACM Trans. Graph.* 37 (2018), 231:1–231:12.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images, CVPR 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. 2019a. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2019b. Video-to-Video Synthesis. In *Proc. NeurIPS*.

O. Wiles, A.S. Koepke, and A. Zisserman. 2018. X2Face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. 2019. Deep Single-Image Portrait Relighting. In *The IEEE International Conference on Computer Vision (ICCV)*.

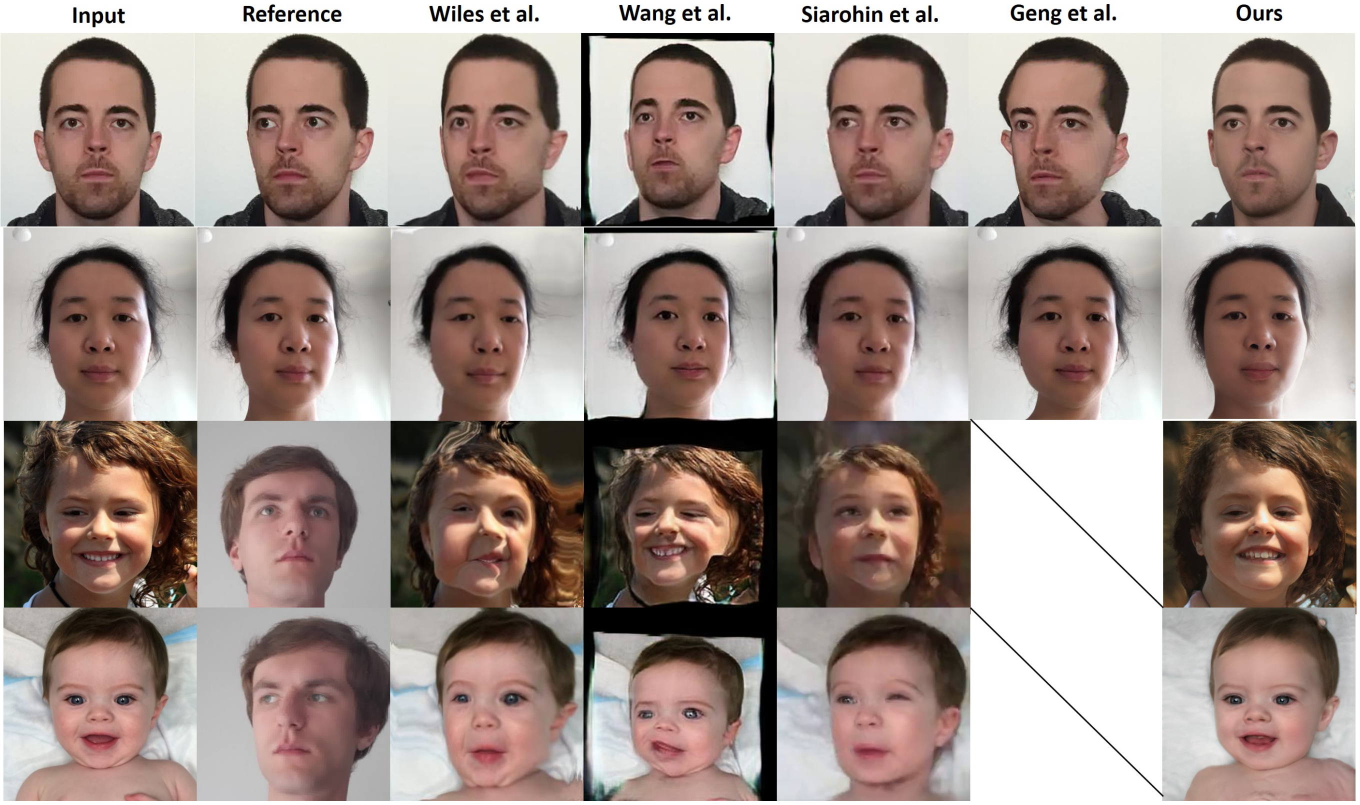

Fig. 3. Comparison of head pose editing for self-reenactment (first two rows) and cross-identity reenactment (last two rows). We compare our approach to Wiles et al. [2018], Wang et al. [2019b], Siarohin et al. [2019] and Geng et al. [2018]. The pose from the reference images is transferred to the input. Our approach obtains higher quality head pose editing results, specially in the case of cross-identity transfer. All approaches other than ours are incapable of *disentangled* edits, i.e., they cannot transfer the pose without also changing the expressions. The implementation of Geng et al. [2018] does not handle cross-identity reenactment. Note that while the three competing approaches require a reference image in order to generate the results, we allow for explicit control over the pose parameters.