Total Selfie: Generating Full-Body Selfies

BOWEI CHEN, University of Washington, USA
BRIAN CURLESS, University of Washington, USA
IRA KEMELMACHER-SHLIZERMAN, University of Washington, USA
STEVE SEITZ, University of Washington, USA

Fig. 1. Example of 4 pairs of training data rendered from one textured mesh. The left 4 columns are the input images, and the last column is their ground truth.

ACM Reference Format:
Bowei Chen, Brian Curless, Ira Kemelmacher-Shlizerman, and Steve Seitz. 2023. Total Selfie: Generating Full-Body Selfies. *ACM Trans. Graph.* 1, 1 (August 2023), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

A SELFIE UNDISTORTION

A common problem with on-site selfies, which typically focus on the facial region, is perspective distortion. This is caused by the camera being too close to the subject, resulting in facial features closer to the camera appearing larger and those farther appearing smaller, thereby creating an unnatural and distorted appearance.

Previous studies have addressed this issue either through single-image optimization [Shih et al. 2019; Wang et al. 2023] or training on a combined dataset of real and unrealistic synthetic images [Zhao et al. 2019]. For test-time efficiency, we follow the idea of large dataset training.

The first step is to create a high-quality realistic paired dataset for supervised training. To do this, we render the dataset using EG3D [Chan et al. 2021], a state-of-the-art textured 3D head generation method. EG3D uses a random noise vector and camera parameters to generate a set of tri-planes, which can then be used to produce color images and meshes through volumetric rendering.

Authors' addresses: Bowei Chen, University of Washington, 1410 NE Campus Pkwy, Seattle, WA, 98195, USA, boweiche@cs.washington.edu; Brian Curless, University of Washington, Seattle, USA, curless@cs.washington.edu; Ira Kemelmacher-Shlizerman, University of Washington, Seattle, USA, kemelmi@cs.washington.edu; Steve Seitz, University of Washington, Seattle, USA, seitz@cs.washington.edu.

0730-0301/2023/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Our goal is to render images of the same face captured at varying camera-subject distances. One straightforward idea is to fix the random noise vector and adjust the camera parameters to directly render desired RGB images. However, this is not feasible as EG3D is pre-trained on a dataset with a specific camera-subject distance. Consequently, rendering images with out-of-distribution camera-subject distances introduces noticeable artifacts. Instead, we generate a textured 3D mesh of the head and render the head images from different distances using traditional rendering techniques. Specifically, we first use the generated tri-planes to sample the volume to obtain a H x W x C cube of density and color value. Then we extract the surface of the scene as a mesh using Marching Cubes [Lorensen and Cline 1987]. Finally, for each 3D surface vertex, we obtain the vertex color by assigning the color value of the nearest point on the volume. Given this textured mesh, we can now render images from a different distances using traditional rendering techniques. Specifically, we fix the camera rotation matrix and only adjust camera distance d. To ensure the eye position unchanged in different images (for the same mesh), we compute focal length f based on the camera distance, given by:

$$f = df_0, \tag{1}$$

where $f_0 = 2.9$ is the focal length pre-defined to render the image without invalid pixels (*i.e.*, all camera rays can hit the mesh) when $d = 1$. We use PyTorch3D to render four input images with severe distortion by setting d to 1, 1.3, 1.6, and 1.9. We then render a shared ground-truth image by setting d to 10. Finally, we align all rendered images using the face alignment technique proposed by [Karras et al. 2019]. This alignment operation is defined as $A(\cdot)$ in the main paper. Fig. 1 shows 4 training pairs of one textured mesh. In total, we render 10K textured meshes, each with 4 training pairs, resulting in a dataset of 40K training pairs.

The next step is to train an undistortion network, $T(\cdot)$, using the rendered dataset. For this, we adapt an existing method, face-vid2vid [Wang et al. 2021]. This method uses a source image and

On-Site Selfie I_s Aligned $A(I_s)$ Undistored $T(A(I_s))$

Fig. 2. Results of our trained perspective undistortion network. Given an on-site selfie (left), we first align it (middle), and then correct the perspective distortion (right).

a driving image to synthesize a talking-head image with appearance and head pose derived from the source and driving images respectively. For our task, both the source and driving images are the image with severe distortion, and the output image is the undistorted image, which will be supervised by our rendered ground-truth. Facevid2vid consists of a couple of face feature extractors that can be applied to any face image regardless of the downstream task. In order to harness this power, we choose to fine-tune the pretrained model on our dataset instead of training from scratch.

Finally, given an on-site selfie I_s, we first align it to get $A(I_s)$, and then use the fine-tuned network to perform perspective undistortion to obtain $T(A(I_s))$. Fig. 2 shows results of perspective correction on the on-site selfies.

B IMPLEMENTATION DETAILS

We perform all the experiments on one single NVIDIA A40 GPU. We use stable diffusion 1.5 [Rombach et al. 2022] and ControlNet v1.1 for our pipeline. The output image resolution is 512x512. We set denoise timestep T to 50, and guidance scale m to 7.5. In the following, we illustrate the implementation details in each module.

B.1 Region-Aware Generation

We train the DreamBooth with ControlNet human pose model [Ruiz et al. 2022; Zhang and Agrawala 2023]. We keep the ControlNet frozen and fine-tune Unet and text encoder during the training. For hyperparameters, we set the learning rate as 1e-6, batch size as 2, prior preservation weight as 1, and the number of images generated for prior preservation as 50. We train DreamBooth for 6K iterations, which takes around 2 hours using a single NVIDIA A40 GPU.

We use a pretrained human parsing model [Yang et al. 2023] to obtain the semantic map of the target pose image. We set $s_w = \frac{250w_m}{1+250w_m}$, where $w_m = \frac{dst(M_g)}{\|dst(M_g)\|_\infty}$, which has been explained in the main paper.

B.2 Appearance Refinement

For perspective undistortion, following the original paper, we fine-tune the pretrained facevid2vid checkpoints using Adam optimizer with learning rate equal to $2e-4$. We fine-tune the network for 7 epochs.

We train the face-specific on-site DreamBooth without ControlNet since we find that this leads to better generalization when fine-tuning on a single image. The hyperparameters are set to be the same as we train the DreamBooth in the previous section. We train this on-site face-specific DreamBooth for 600 iterations, which takes around 8 minutes using a single NVIDIA A40 GPU.

For $AC(I_t)^l$, we first detect landmarks and pose from $AC(I_t)$, and we then filter out the landmarks corresponding to the eye, nose, and mouse because we need to generate the new expression in the on-site selfie. In other words, we only keep the landmarks describing the face shape. We dilate M_i for each body part by 5 pixels to account for inaccurate segmentation.

B.3 Image Harmonization

To fine-tune the pretrained decoder D on I_r, we optimize for the following objective function:

$$\min_{\gamma} L_1(D(z_r), I_r) + L_p(D(z_r), I_r), \tag{2}$$

where $z_r = E(I_r)$, E is the encoder. γ is the parameters in decoder D we aim to optimize. L_1 is L1 loss and L_p is perceptual loss. We use Adam optimizer with learning rate 1e-4, and optimize D for 400 iterations, which takes around 3 minutes on one single NVIDIA A40 GPU.

For null-text inversion, at each timestep t, we optimize $\{\tilde{\varnothing}_t\}$ using Adam optimizer with 50 iterations. The learning rate for this optimization is 0.01.

In the final stage of image harmonization, we obtain the final output I_h by blending the denoised output I'_h (using forward guidance) with the background image I_b. Specifically, we first estimate the semantic map of I'_h using [Yang et al. 2023] to obtain the person mask M_p and the shoe mask M_s. We dilate the shoe mask by 17 pixels to cover the region around the shoes. Then we apply Gaussian blur to M_p and M_s with kernel sizes equal to 3 and 21, respectively. Then we compute the blending mask $M_b = 1 - (1 - M_p) \cdot (1 - M_s)$. Now, M_b is a soft mask of a person that has shoe region dilated. We specially handle shoe regions because we want to preserve the shadow around the shoes (usually on the ground) in I'_h. Finally, we blend I'_h and I_b by:

$$I_h = \alpha \cdot (I'_h \cdot M_b + I_b \cdot (1 - M_b)) + (1 - \alpha) \cdot I'_h, \tag{3}$$

where $\alpha = 0.5$, and I_h is the final output of the pipeline.

(a) Overhead (b) Cloth (c) Pants (d) Shoes

Fig. 3. Sample images of different body parts of different users extracted from their selfie videos. The appearance of the same outfit can vary across different selfies, depending largely on factors such as spatially variable lighting conditions and diverse camera settings. For instance, when comparing the black top in row 1 (a) to row 1 (b), it is noticeable that the black top appears somewhat lighter in the latter image.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 Data Capturing

Each user is requested to capture a square selfie video showcasing four distinct components of his outfit: an overhead view, his upper cloth, pants, and shoes.

Firstly, the user is instructed to hold the camera above his head, using the front camera. This overhead view should encompass a full display of his clothing. The goal of overhead selfie is to provide a comprehensive look at the individual's overall outfit. Next, the user will shift the camera focus towards the upper cloth area. The majority of the image frame should cover this region, again using the front camera. Following this, the user needs to switch to the rear camera and adjust the camera angle to focus on the pants region. Finally, the user continues using the rear camera, moving it to cover the shoe area. Throughout the entire recording, the camera is set to wide-angle mode for easy capture. Each content capture should last approximately 15 to 20 seconds, resulting in a total video length of 60 to 80 seconds. We again show the sample images in Fig. 3 for convenience.

Upon arrival at a new location, the user is initially requested to capture a background image. Subsequently, the user is asked to take three distinct on-site selfies, each with a different expression. For the purpose of obtaining a comprehensive evaluation, following each on-site selfie, the user is asked to maintain the same facial expression. Another individual is then tasked with taking a full-body photo of the user, preserving both the expression and the same background as the user-captured image. We collected data from five individuals across two different scenes. Each scene will encompass three on-site selfies per individual. This results in 30 examples in total.

C.2 Results of Total Selfie

More results of the Total Selfie are shown in Fig. 4, 5, 6, 7, and 8. Total Selfie has the capability to handle complex facial expressions and generate realistic and accurate full-body images with plausible shading.

C.3 Ablation Study

Fig. 9 presents the ablation analysis of two variants: *Ours-IH-AR* and *Ours-IH*. The former struggles to maintain the correct attire and identity, whereas the latter generates accurate clothing and identity but introduces artifacts in the boundaries, shading, and hand areas. However, our complete pipeline successfully produces a full-body image against a specific background, upholding the correct identity, outfit, and credible shading.

Fig. 10 shows comparisons between *Ours-SU* and Total Selfie. Thanks to perspective distortion correction, Total Selfie can produce a more natural facial appearance that matches the on-site selfie.

Fig. 11 shows the ablation study of two variants: *Ours-Style* and *Ours-Lpips*. Without perceptual loss, *Ours-Lpips* fails to preserve the content in the image I_r. Compared to *Ours-Style*, our full pipeline, with the assistance of style loss, generates a full-body image with shading that is more consistent with the on-site selfie.

C.4 Comparison with Baselines

Fig. 12 compares Total Selfie with various baselines. It's clear that Total Selfie significantly outperforms all the evaluated baselines in terms of image realism, as well as the preservation of identity and outfit.

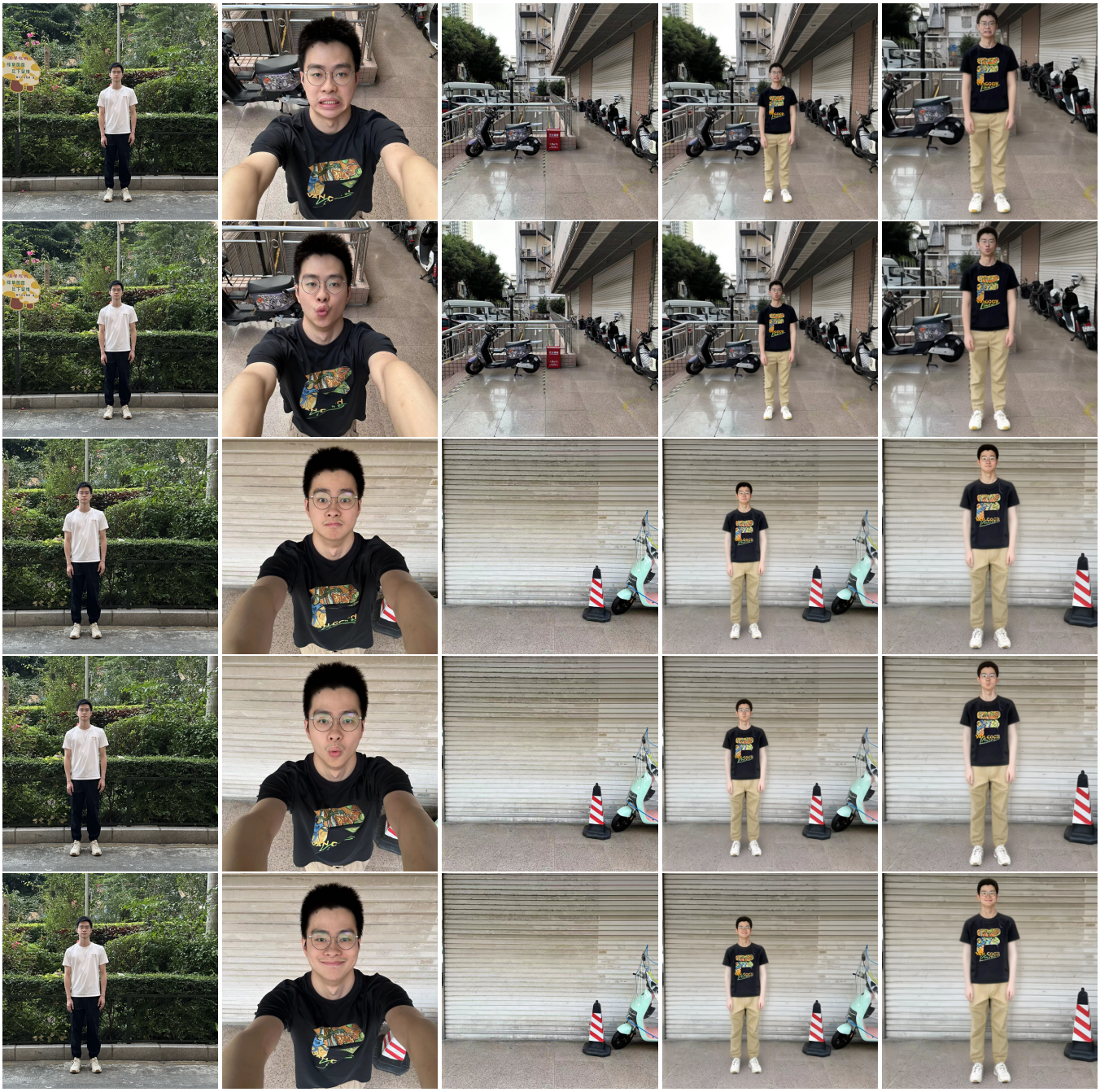
REFERENCES

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

William E Lorensen and Harvey E Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics* 21, 4 (1987), 163–169.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.



(a) Target Pose (b) On-Site Selfie (c) Background (d) Ours (e) Ours (Zoom In)

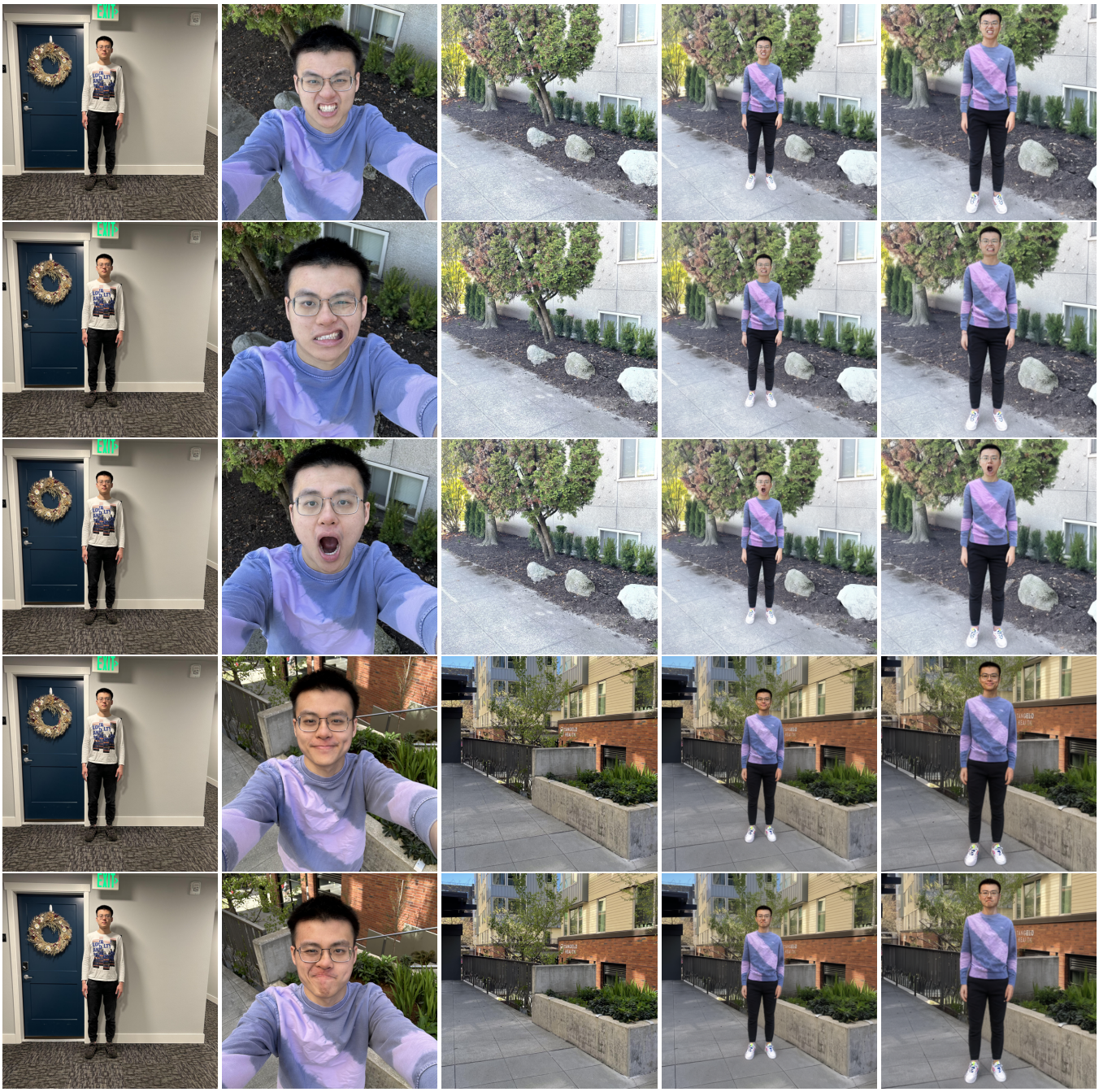
Fig. 4. Results of Total Selfie. The sample pre-captured images (from selfie video) are shown in Figure. 3.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242* (2022).

YiChang Shih, Wei-Sheng Lai, and Chia-Kai Liang. 2019. Distortion-free wide-angle portraits on camera phones. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Zhixiang Wang, Yu-Lun Liu, Jia-Bin Huang, Shin'ichi Satoh, Sizhuo Ma, Guru Krishnan, and Jian Wang. 2023. DisCO: Portrait Distortion Correction with Perspective-Aware 3D GANs. *arXiv preprint arXiv:2302.12253* (2023).



(a) Target Pose (b) On-Site Selfie (c) Background (d) Ours (e) Ours (Zoom In)

Fig. 5. Results of Total Selfie. The sample pre-captured images (from selfie video) are shown in Figure. 3.

Lu Yang, Wenhe Jia, Shan Li, and Qing Song. 2023. Deep Learning Technique for Human Parsing: A Survey and Outlook. *arXiv preprint arXiv:2301.00394* (2023).

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

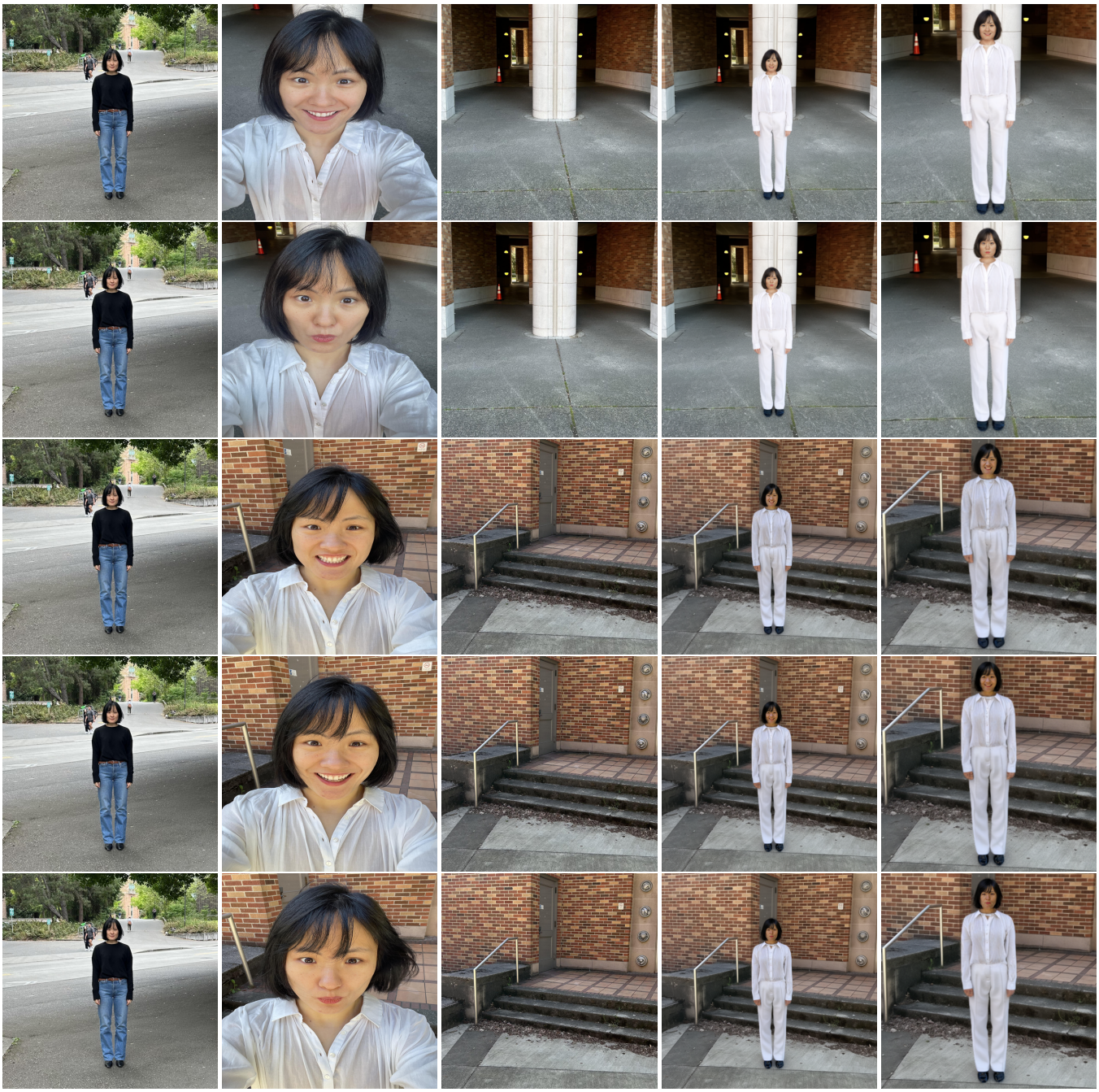
Yajie Zhao, Zeng Huang, Tianye Li, Weikai Chen, Chloe LeGendre, Xinglei Ren, Ari Shapiro, and Hao Li. 2019. Learning perspective undistortion of portraits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7849–7859.

(a) Target Pose (b) On-Site Selfie (c) Background (d) Ours (e) Ours (Zoom In)

Fig. 6. Results of Total Selfie. The sample pre-captured images (from selfie video) are shown in Figure. 3.

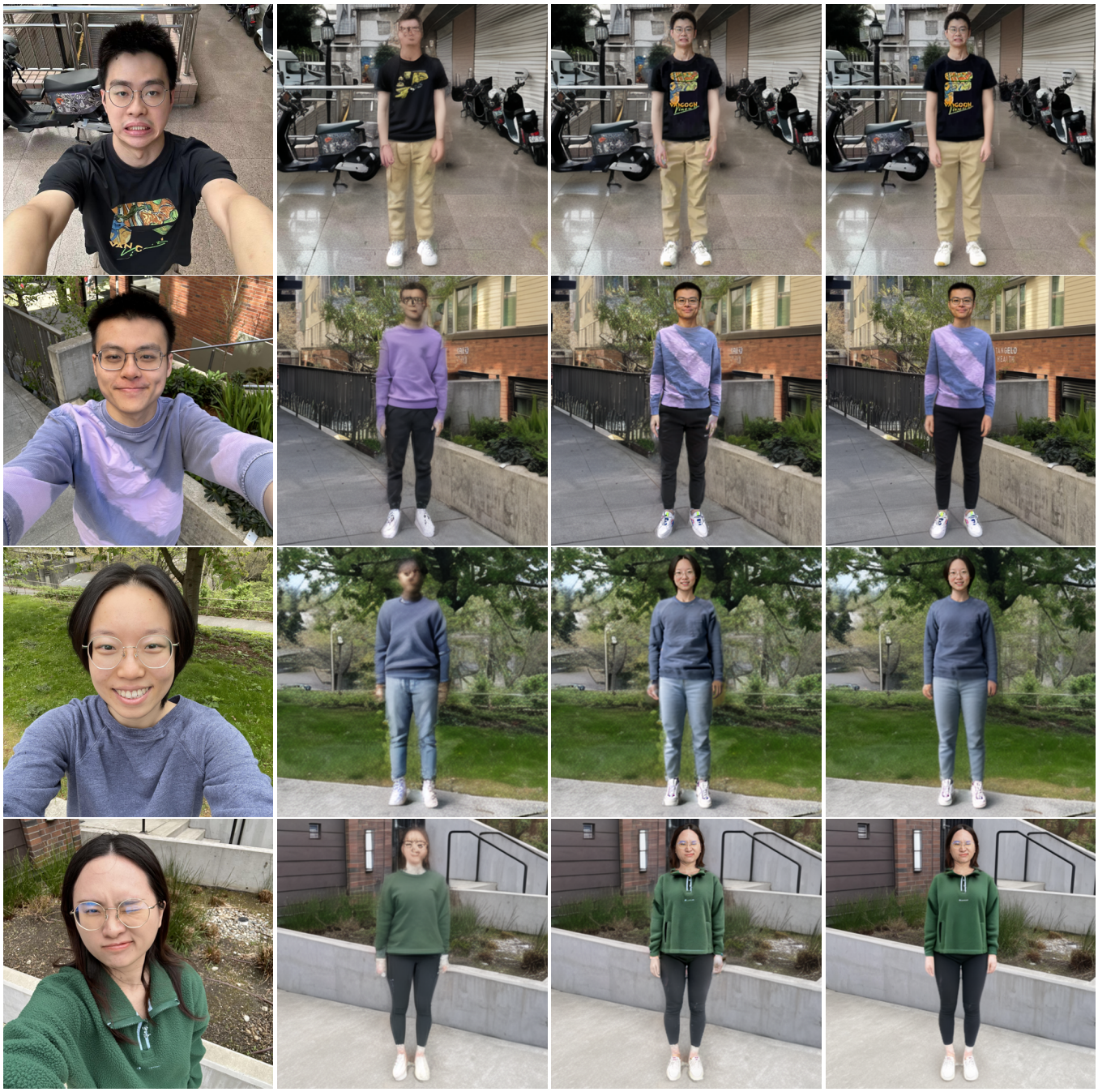
(a) Target Pose (b) On-Site Selfie (c) Background (d) Ours (e) Ours (Zoom In)

Fig. 7. Results of Total Selfie. The sample pre-captured images (from selfie video) are shown in Figure. 3.



(a) Target Pose (b) On-Site Selfie (c) Background (d) Ours (e) Ours (Zoom In)

Fig. 8. Results of Total Selfie. The sample pre-captured images (from selfie video) are shown in Figure. 3.



(a) On-Site Selfie (Input) (b) Ours-IH-AR (c) Ours-IH (d) Ours

Fig. 9. Ablation Study of our pipeline. All results are zoomed in for better visualization.

(a) On-Site Selfie (Input) (b) Ours-SU (c) Ours

Fig. 10. Comparison of the pipeline with and without selfie undistortion. All results are zoomed in for better visualization. With perspective undistortion, the pipeline produces a more natural-looking face.

(a) On-Site Selfie (Input) (b) Ours-Lpips (c) Ours-Style (d) Ours

Fig. 11. Ablation study of loss in our pipeline. All results are zoomed in for better visualization.

(a) On-Site Selfie (b) PIDM (c) Paint-By-Example (d) DB+ControlNet (e) Ours (f) Ground-Truth

Fig. 12. Qualitative Comparison with baselines. Sample pre-captured images (from the selfie video) are shown in Fig. 3. All results are zoomed in for clear visualization. In all methods, we use the ground-truth as the target pose to constrain the pose. Our pipeline clearly outperforms all baselines in terms of photo realism and faithfulness. Note that, despite being captured nearly at the same time, the color tone of the on-site selfie, background image, and ground-truth may not match due to differences in lighting conditions, auto exposure, and white balance *etc*.