This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Total Selfie: Generating Full-Body Selfies

Bowei Chen Brian Curless Ira Kemelmacher-Shlizerman Steven M. Seitz
University of Washington

{boweiche, curless, kemelmi, seitz}@cs.washington.edu

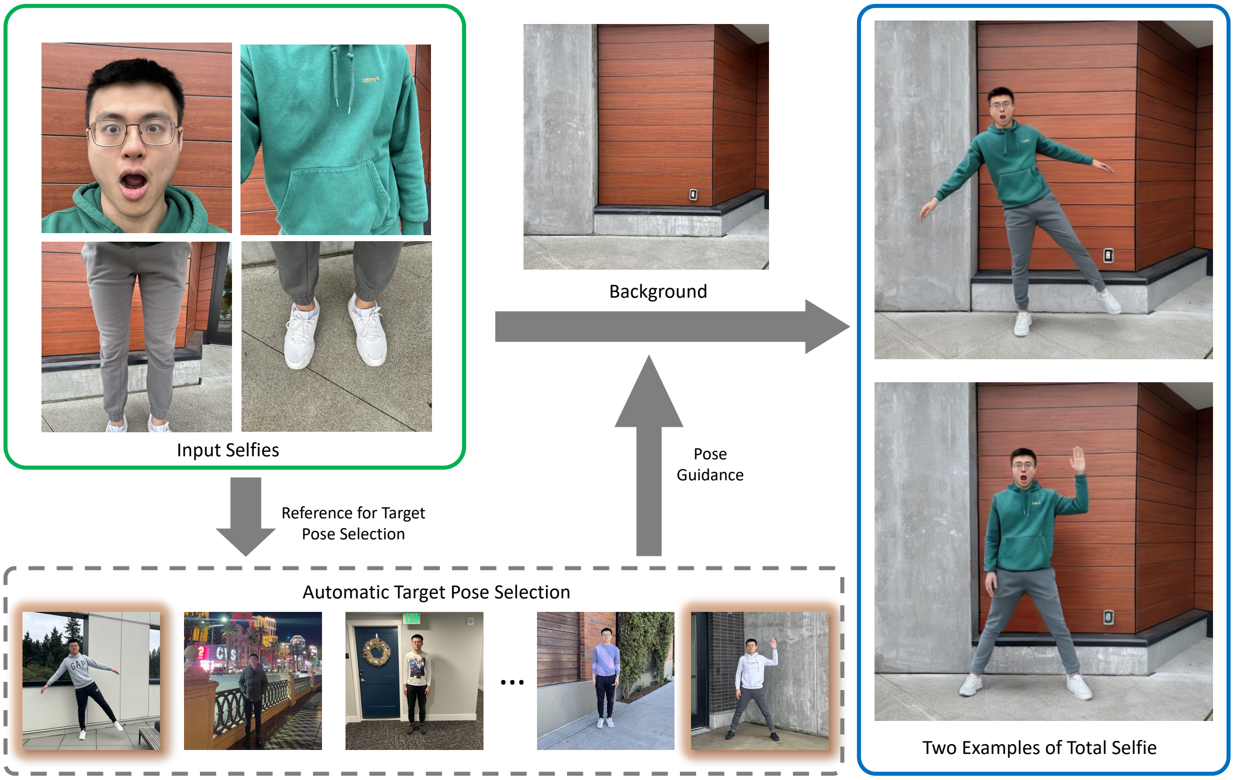

Input Selfies
Background
Reference for Target Pose Selection
Pose Guidance
Automatic Target Pose Selection
...
Two Examples of Total Selfie

Figure 1. We generate full-body selfies of you (right), from self-captured images of your face and body (top left) and background. You can choose any target pose from a reference photo — we auto-select a set of good candidates from your photo collection (bottom).

Abstract

We present a method to generate full-body selfies from photographs originally taken at arms length. Because self-captured photos are typically taken close up, they have limited field of view and exaggerated perspective that distorts facial shapes. We instead seek to generate the photo some one else would take of you from a few feet away. Our approach takes as input four selfies of your face and body, a background image, and generates a full-body selfie in a desired target pose. We introduce a novel diffusion-based approach to combine all of this information into high-quality, well-composed photos of you with the desired pose and background.

1. Introduction

The prevalence of selfies has skyrocketed in recent years, with an estimated 93 million taken each day. Despite their popularity, they suffer from multiple shortcomings: (1) they capture only the upper portion of the subject, (2) the close-up camera viewpoint distorts faces and requires awkward poses (e.g., with arm reaching out), and (3) it is difficult to compose a shot that optimally captures both the subject and the scene.

Instead, what if you could capture the full-body image that *someone else would take of you* in the scene? We call this a *total selfie*. As input, we require four selfies to cover different parts of your body, and a photo of the background that you would like to be composited into (Fig. 1). Based on this information, we generate convincing full-body photos of you in a specified target pose in the desired scene (Fig. 1 right). In practice, we automatically select candidate reference photos from your photo collection, allowing you to choose one or more of them to determine the target pose.

Solving this problem requires addressing a number of challenges. First, we must render a complete and accurate image of your body, piecing together separate close-up images of your face, upper torso, legs, and shoes. Second, we must reproject you to a virtual viewpoint from several feet away – far enough to compose your full body within the scene. And third, we need to render you with a desired target pose (where you're not holding the camera), which can be completely different from the one you used to take the selfies. The target pose can be specified by any full-body image from your photo collection. To facilitate target pose selection, we auto-detect photos from your collection where you are wearing similar clothing to the input selfie set, leading to results with more accurate body shapes for a given type of clothing. Most importantly, the resulting composite must retain your identity, expression, and clothing, but be composited realistically into the target scene with the desired pose and appearance.

One approach to this problem would be to collect a paired dataset of selfies and full-body images of many people, and train a generative model on it. However, acquiring such a dataset would be time and cost intensive. Instead, we train a selfie to full-body model using a paired *synthetic* dataset, and further perform per-capture fine-tuning to narrow the gap between real and synthetic data. Specifically, we first introduce a diffusion-based inpainting model trained in a self-supervised manner. This model takes four selfies as input and inpaints a full-body subject into a masked background. Given a set of input images, we first remove perspective distortion of the face selfie, and then fine-tune the trained model on these images to enhance the fidelity of generated full-body photo further.

Our contributions can be summarized as follows:

- We introduce a novel type of self-captured photo – *total selfie* – that captures your entire body, as if some one was taking a photo of you in the scene.
- We propose a diffusion-based full-body generation model, followed by per-capture fine-tuning techniques, to generate *total selfie* from four selfies covering the body, a background image, and an auto-selected reference image as target pose.
- We demonstrate results for twelve individuals in various scenes (*e.g.*, indoor and sunny outdoor) and clothing (*e.g.*, skirts) with a wide range of poses and expressions. Our experimental results outperform existing methods in generating realistic and accurate full-body images.

2. Related Work

Full-Body Image Generation. Extensive research has been dedicated to generating full-body images, either without specific conditions [15–17, 56] or with conditions such as pose [62], shape [49], or text prompts [22]. One related area of research to our task is human reposing [3, 10, 13, 14, 19, 21, 24, 26, 29, 32, 34, 42–44, 48, 53, 66, 68]. These methods transform a full-body (or partial-body) human image from one pose to another, with a target pose provided. For example, DisCo [53] designed a diffusion-based framework to achieve this by using a person image (in any pose), a background image, and a target pose. However, these methods are tailored for single image input and fall short when applied to our task due to the inherent limitation of a single selfie in capturing the full view of the person.

Another related line of work is Virtual Try-On [9, 11, 12, 20, 28, 31, 57, 59, 69], where the goal is to generate a visualization of how a garment might appear on a person, given the person image (in different clothing) and the garment image. For instance, LaDI-VTON [36] introduced the first Latent Diffusion textual Inversion-Enhanced model to synthesize an image of the person wearing a specified clothing item. However, these approaches usually assume simple backgrounds, posing challenges in generating realistic shading within complex backgrounds. They also cannot alter footwear and facial expressions, crucial for our task.

Despite these challenges, both streams of work assume input from third-person view images, not selfies.

Selfie-Related Techniques. Numerous studies have explored selfies for applications like reposing [33], face recognition [6, 27], style transfer [30, 52], novel view synthesis [1, 4, 23, 37, 38], relighting [7], and video stabilization [63, 64]. For example, [33] proposed a coordinate-based method to transform a typical selfie, mainly focusing on the face, into a neutral-pose portrait. Nevertheless, their approach was limited to upper body selfies and could not generate full-body selfies capturing both the subject and the surroundings. Our work is the first to propose and generate *total selfie* from arm-captured selfies.

Diffusion Models. Diffusion models have recently demonstrated their success in various tasks such as text-to-image [41, 45, 47] and image-to-image translation [18, 60]. DreamBooth [46] was proposed to personalize a text-to-image model by fine-tuning the model on a few reference images. RealFill [51] introduced an image completion technique to outpaint an input photo using reference images from the same scene. It fine-tuned a text-to-image inpainting model using reference images and applied it to com-

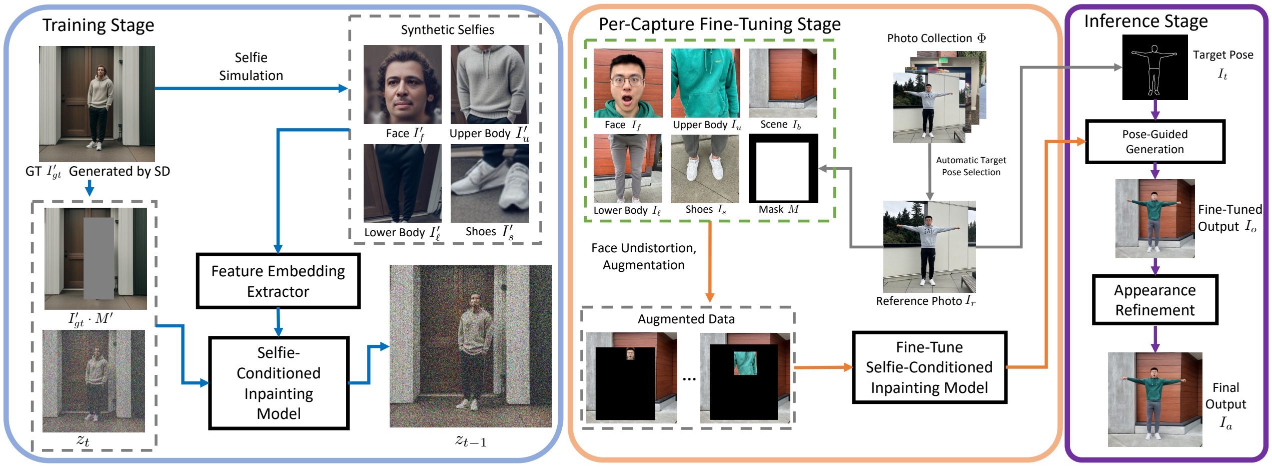


Figure 2. Overview of Total Selfie. First, we train a selfie-conditioned inpainting model based on a synthetic selfie to full-body dataset (blue box). Second, we fine-tune the trained model on a specific capture (orange box), and use it to produce a full-body selfie with the help of modified ControlNet (for pose) and appearance refinement (for face and shoes), visualized in the purple box. Note that, images in the green dashed box (inside the orange box) serve as input and conditional signals to the inpainting model, arrows omitted for simplicity.

plete the input photo. However, RealFill assumed third-person view images as input and mainly focused on generating scene content, rather than full-body subjects. Another relevant work, Paint-By-Example [60], introduced an image-conditioned inpainting model to inpaint masked scenes with content specified in a reference photo. Similarly, we frame our problem as an exemplar-based inpainting problem. Thus we adapt this model to suit our settings.

3. Total Selfie

We refer to our pipeline as Total Selfie, and define our task setting more formally. As input, a user captures four selfies, including face I_f, upper body I_u, lower body I_ℓ, and shoes I_s, as well as the background image I_b. Total Selfie inpaints the full-body individual into I_b, with the target pose I_t and inpainting mask M (where 1 indicates the region to be inpainted) specified by a reference image I_r. Here, I_r is automatically selected from the user's photo collection Φ.

Total Selfie has two main steps. As depicted in Fig. 2 left, we first generate a large paired dataset comprising four selfies as input and a corresponding full-body image as ground truth. Then we train a selfie-conditioned inpainting model on this dataset. Given an input capture, we perform preprocessing on the input images, including face undistortion and automated pose selection. These preprocessed images are then used to fine-tune the trained model (Fig. 2 middle). The fine-tuned model is employed to generate an initial output I_o, which is further refined to produce the final output I_a (Fig. 2 right).

3.1. Training Selfie-Conditioned Inpainting Model

Training the selfie-conditioned inpainting model involves two steps: (1) generating a large paired dataset and (2) training an image-conditioned diffusion model on this dataset.

Dataset Generation. We define one training pair as $\{(S', I'_{gt} \cdot M', M'), I'_{gt}\}$, where $S' = \{I'_f, I'_u, I'_\ell, I'_s\}$ is a set of four synthetic selfies for face, upper body, lower body, and shoes respectively. I'_{gt} is the ground-truth full-body image, and M' is the inpainting mask.

To create a pair, we start by generating I'_{gt} using Stable Diffusion [45], with the pose guided by OpenPose ControlNet [65]. The person's bounding box in I'_{gt} is then scaled up (following [60]) to generate the inpainting mask M'. We choose a bounding box representation instead of a more detailed shape for the mask since we anticipate changes in nearby regions, like those affected by shadows, when integrating the individual into the scene. Simulating selfie set S' from the third-person view I'_{gt} is non-trivial. One possible idea is to estimate 3D geometry of I'_{gt} using depth estimation [2, 5, 40] or human reconstruction [58] methods, and then re-render it from the perspective of a selfie camera. However, this is not practical due to inaccuracy of the estimated 3D geometry. Instead we propose a simpler yet effective way for obtaining S' using homography transformation. To create I'_u, we follow these steps. (1) Identify the *typical* positions of upper body OpenPose keypoints from pre-captured real upper body selfies (see supplementary). (2) Detect the upper body OpenPose keypoints from I'_{gt}. (3) Compute a homography transformation that maps keypoints in I'_{gt} to their typical positions. (4) Use this transformation to warp I'_{gt} and obtain I'_u. We apply the same approach to generate I'_ℓ and I'_s. Finally, we use FFHQ face alignment [25] to extract the face selfie I'_f. The resultant selfie set S' is visualized in the top right part of Fig. 2 blue box. We repeat this process to create multiple training pairs with I'_{gt} in diverse poses. Although the selfie simulation process primarily focuses on viewpoint correction and does not consider pose differences (*e.g.*, using one arm to hold camera), we observed that this is sufficient for training an inpainting model and obtaining a reasonable selfie-to-full-body prior.

Training. We start with an existing image-conditioned, diffusion-based inpainting model called Paint-By-Example [60] and tailor our specific task. We refer to our adapted model as selfie-conditioned inpainting model.

To train our model, we first apply the forward diffusion process to I'_{gt}. Starting from a clean latent $z_0 = E(I'_{gt})$, the forward process yields a noisy latent $z_t = \alpha_t z_0 + (1-\alpha_t)\epsilon$, where E is the encoder of Variational AutoEncoder used in Latent Diffusion [45]. Here, t is a randomly sampled timestep, ϵ represents Gaussian noise, and α_t is a weight parameter determined by t. We then use a diffusion model to denoise z_t and update the model parameters by minimizing the following loss function:

$$\mathcal{L} = \mathbb{E}_{t,z_0,\epsilon} ||\epsilon_\theta(z_t, I'_{gt} \cdot M', c', t) - \epsilon||_2^2, \tag{1}$$

where c' is the condition for the diffusion model, which, in our case, is the image embeddings extracted from S'. In Paint-By-Example, c' is implemented as $L \cdot F(I)$, where I is a single image, $F(\cdot)$ is a pretrained CLIP image encoder [39] followed by a multi-layer perceptron (MLP), and $L(\cdot)$ is a linear layer. In our model, we extend this idea and implement it as:

$$c' = L([F(I'_f), F(I'_u), F(I'_\ell), F(I'_s)]), \tag{2}$$

where $[\cdot]$ is the operation of concatenation. L is modified to adapt to the dimension of the concatenated embeddings. One advantage of $F(\cdot)$ is that it transforms an image into a highly compressed vector. This forces the network to understand the clothing type and color in S', preventing it from reaching optimal results by simply applying homography transformation to S' in training. The training process is visualized in the blue box in Fig. 2. In practice, we train our model by updating parameters (θ, weights in F's MLP and L) using Eq. 1. We initialize weights (excluding those in layer L) with pretrained weights from Paint-By-Example. This allows us to leverage the generative prior in this pretrained model and significantly reduce our training time.

3.2. Per-Capture Preprocessing and Fine-Tuning

We start by presenting two test time preprocessing steps: *face undistortion* and *automatic target pose selection*. Then we introduce how to generate results aligning with the target pose, and discuss the issue of generated results to motivate the per-capture fine-tuning.

At test time, the set of user-captured selfies $S = \{I_f, I_u, I_\ell, I_s\}$ has a different distribution from the simulated selfies S' in the training set. We note two particular differences: (1) real selfie I_f often exhibits significant face distortion, which is not present in the simulated selfie I'_f (obtained from full-body image), (2) upper body selfies I_u and I'_u have different poses, arising from the need to hold the camera out front (using an upper body arm) when taking a real selfie (see Fig. 2). To address the first issue, we propose a face undistortion strategy as a preprocessing step to help reduce the domain gap. We address the second issue, pose differences, with a fine-tuning step which we discuss later, as it is non-trivial to resolve during preprocessing.

Face Undistortion. Existing methods alleviate selfie distortion by either optimizing on a single image [50, 55] or by training a model on a combination of an unrealistic synthetic dataset and a small real dataset [67]. For test-time efficiency, we follow the latter idea. We render a large paired dataset using a method that generates realistic, textured 3D heads using 3D GANs [8]. Then we fine-tune a talking-head synthesis network [54] to perform perspective undistortion using the rendered dataset. See supplementary for more details and results. Additionally, we roughly align the shoes selfie I_s with I'_s by cropping it based on the bounding box of the shoes. For simplicity, we reuse I_f and I_s to represent corrected face selfie and cropped shoes selfie, respectively.

Automatic Target Pose Selection. Another preprocessing step is to obtain the target pose I_t to guide the full-body selfie generation. We require I_t to convey two types of information: (1) the desired pose and (2) the actual body shape. To achieve this, we represent I_t as the contour of the user's body, derived from a reference photo I_r of the *same* person. We first discuss how to obtain I_r and address the process of deriving I_t from I_r in the next section.

We develop an automatic selection strategy to help obtain I_r from the users' photo collection Φ. The selection criteria are based on the similarity between the clothing types in the input selfies and a candidate image in Φ. This is because the more similar the clothing type is, the more accurately the body shape (in this particular type of outfit) can be extracted from I_r. Specifically, we first use a pretrained human parsing model [61] to detect the clothing types (*e.g.*, hoodie) in the selfies I_u, I_ℓ, and I_s. We then apply the detection to each full-body photo in Φ. A list of candidate reference photos is suggested based on the number of matched clothing types (higher is preferred) between the selfies and the image from Φ. Then the user can choose a reference image I_r from the list. Finally, the inpainting mask M is obtained using the scaled bounding box of the person in I_r.

Pose-Guided Generation. After preprocessing, to generate a full-body selfie, we follow the standard diffusion denoising process with the guidance of ControlNet [65] to ensure that the generated image aligns with the pose in I_r. The key difference is that we modify the ControlNet architecture to enable it to possess similar guiding capabilities for our image-conditioned diffusion model as it does for text-conditioned models. Specifically, we replace the

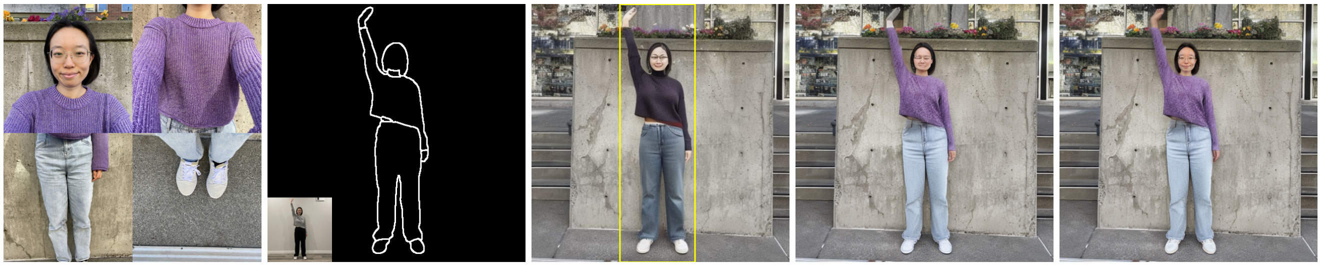


Figure 3. Results for different modules of our pipeline. Background image omitted due to space; regions inside bounding box (c) are to be inpainted. The Canny Edge image in (b) is detected from the reference image, inset. Generating without fine-tuning (c) produces inaccurate outfit and identity. Through fine-tuning, the pipeline (d) generates correct outfit with reasonable shading and clothing details (*e.g.*, wrinkles on upper cloth), but with wrong identity. With appearance refinement, the full pipeline (e) yields high-quality full-body selfies.

text embeddings with c, which is computed using Eq. 2 by replacing simulated selfie I'_k with real selfie I_k, where $k \in \{f, u, \ell, s\}$. This modification allows any pretrained ControlNet to be plugged into our model, providing the desired guidance. To meet our requirements, we specifically opt for a Canny Edge ControlNet with target pose I_t as the control signal. To obtain I_t, we segment various body parts in I_r using [61], producing a semantic map. The Canny Edge I_t is then detected from this semantic map (Fig. 2 top right), not directly from I_r itself. This ensures that the pipeline is not influenced by the outfit details in I_r, such as cloth texture. Note that, as we will discuss, one can always use OpenPose ControlNet, and obtain I_t (skeleton) from any human image, albeit sacrificing accurate body shape.

To obtain the full-body selfie, starting from a random noise z_T at timestep T, we iteratively perform one-step denoising using $\epsilon_\theta(z_t, I_b \cdot M, c, t)$ with modified ControlNet for T steps. This yields the clean latent z_0, which is further decoded by decoder D to generate the full-body selfie I_n. However, as shown in Figure 3 (c), directly using trained model in Sec. 3.1 produces inaccurate outfit and identity due to the gap between synthetic and real data.

Fine-Tuning. To improve results, we fine-tune the selfie-conditioned inpainting model on the specific capture. By leveraging the full-body generative prior in the model, this strategy is robust to distribution differences (*e.g.*, pose variations in the upper body) between S and S', enhancing the preservation of clothing texture and generating pose-specific details (wrinkles on upper clothing in Fig. 3 (d)).

To implement this strategy, we generate a "ground truth" for fine-tuning by resizing and placing a randomly selected selfie image from the set S into the masked area of the image I_b (see supplementary for details). The bottom left part of Fig. 2 orange box visualizes examples of two augmented images produced using I_f and I_u. We create 200 augmented images and employ them to fine-tune $\epsilon_\theta(z_t, I_b \cdot M, c, t)$ supervised by the loss computed similar to Eq. 1. While fine-tuning on close-up selfies, the full-body generative prior in the trained model helps prevent overfitting, especially for I_u with significant pose differences. Finally, we utilize the fine-tuned model to produce full-body selfie I_o using the pose-guided generation. Fig. 3 (d) shows an example of fine-tuned output, which has reasonable outfit and shading but wrong identity. The identity problem arises because the VAE used in Latent Diffusion fails to generate details of the small face in the full-body photo, a well-known limitation. Similar challenges arise with shoes because they are too small in I_o.

Appearance Refinement. To further enhance the details of I_o, we employ several post-processing strategies. First, for refining face and shoes, we augment I_f and I_s by resizing them to various resolutions and zero-padding the border. Subsequently, we train a DreamBooth model [46] with two concepts (shoes and face) using these augmented images. Finally, we crop the face region from I_o and then employ SDEdit [35], based on the trained DreamBooth, to refine the face. We subsequently compose the refined face back into I_o. A similar operation is applied to the shoes region. Second, we also apply the similar operation to hands region, but with a pretrained Stable Diffusion model [45] since hands are usually not shown in the input selfies. As shown in Fig. 3 (e) and Fig. 2 purple box, these post-processing steps culminate in the creation of the final output, denoted as I_a. This final output faithfully preserves the identity and outfit, and exhibits realistic shading, along with the desired pose.

At the end, the user can switch to a new reference photo and use the current fine-tuned model for generation, provided the person in the new reference photo is within the inpainting mask M. If not, fine-tuning for the new reference photo is required. See the supplementary for execution time and implementation details, including strategy to help preserve background content inside the inpainting mask.

4. Experiments

Results. We begin by showcasing the results of Total Selfie across five different captures, as illustrated in Figure 4. The goal is to retain the user's facial expression and clothing, but realistically retargeted onto the background with the desired pose; these results demonstrate this capability.

In particular, the results show compelling full-body selfies with a wide range of poses, even if the desired poses significantly deviate from those in the input selfies (row 2 to 5). Importantly, the method is able to add realistic wrinkles in the clothing to fit the new pose, with consistent shading (*e.g.*, raised arm in rows 3 to 5). Observe that the technique works with a range of clothing, from short sleeves to jackets, and both pants and skirts. Finally, the method is able to convincingly fill in missing details (*e.g.*, hands, which are often missing in selfies), that fit the individual and are realistically shaded in the target scene. More results are in the supplementary.

Evaluation Data. For evaluation, we collected a dataset of twelve people wearing various outfits with selfies taken in a variety of scenes and lighting conditions, resulting in a total of seventeen captures. Additionally, we captured real, "ground truth" full-body photos of each subject, maintaining the same clothing, nearly identical facial expressions and background.

Ablation Study. We study the effects of different parts of the pipeline and design three variants: (1) *Ours-FT-AR*: full pipeline without per-capture fine-tuning and appearance refinement. (2) *Ours-AR*: full pipeline without appearance refinement, (3) *Ours-FU*: full pipeline without face undistortion. As discussed in Sec. 3, Fig. 3 and Table. 1 show that the full pipeline outperforms all tested variants.

Comparison to Baseline. To the best of our knowledge, there are no existing papers solving the same task. We therefore adapt four existing methods to create four baselines. (1) *Paint-By-Example* [60] was designed to inpaint a masked image using a single source image. Here, we concatenate four input selfies vertically as the source since this works better than only using one selfie. We use Canny Edge ControlNet [65] to guide the pose. (2) *DisCo* [53] reposes an input human image using diffusion models, given a background photo and a target pose. We opt for this over other reposing approaches because it can handle complex backgrounds. Similar as before, we concatenate four input selfies vertically to create the input human image for better performance. Following the official implementation, we use OpenPose Skeleton as target pose since their provided model does not work well with Canny Edge input. (3) *LaDI-VTON* [36] is a diffusion-based Virtual Try-On method. We opt for this method over other similar methods since it can handle both upper- and lower-body garments, and the code is available. This method specifically requires a full-body RGB image as input. Thus we use Stable Diffusion and Canny Edge ControlNet to generate the full-body input aligning with target pose. Then we apply garment editing on this input image using upper and lower body selfies, excluding the face and shoes, as this method does not support editing those elements. Finally, we blend the edited image with the input background for the final output. (4) *DreamBooth* [46] customizes a pretrained text-to-image diffusion model by fine-tuning with a few reference images. Here, we use four selfies (with augmentation) as reference images and generate outputs using Canny Edge ControlNet.

Method	LPIPS ↓	SSIM ↑	PSNR ↑	FID ↓
Paint-By-Example	0.252	0.621	12.37	184.6
DisCo	0.287	0.586	12.13	245.7
LaDI-VTON	0.263	0.563	10.35	172.8
DreamBooth	0.224	0.663	13.61	159.2
Ours-FT-AR	0.218	0.691	14.19	154.1
Ours-AR	0.190	0.703	16.89	143.4
Ours-FU	0.188	0.706	16.98	141.6
Ours	**0.187**	**0.708**	**17.01**	**139.4**

Table 1. Comparison with baselines and our ablation variants. The metrics are evaluated only in the foreground. We employ ground truth Canny Edge to define the target pose for all rows.

Fig. 5 and Table. 1 show the qualitative and quantitative results of the compared methods, respectively. Please note that real photos may have contrast and color tone variations (e.g., due to auto-exposure and white balance), leading to differences between real and generated outputs. Among the compared methods, *DisCo* performs worst because it is guided by skeleton and assumes a third-person view input. *Paint-By-Example* is also ineffective since it is not designed specifically for full-body generation. Further, both methods can only consider one reference image, leading to sub-optimal results. *LaDI-VTON* exhibits artifacts on the clothes (Fig. 5 (c)), which may be attributed to the reference garment images being selfies, distinct from those in its training set. *DreamBooth* produces inaccurate outfits (Fig. 5 (d)). This is due to the pretrained text-to-image model lacking strong priors for understanding clothes in selfies and linking them together as a coherent full-body image. Conversely, Total Selfie, which is initially trained on a synthetic dataset and then fine-tuned per capture, excels in realistically generating full-body selfies.

Discussion of Target Pose Options. We explore the trade-offs of different target pose options, as shown in Fig. 6. *No Condition* is the simplest option, requiring no target pose input. Our pipeline automatically determines pose, body shape, and scale from the masked background and input selfies. However, it generates inaccurate body shape and scale (row 1(d)), and lacks pose controllability. *OpenPose Skeleton* enables users to specify the target pose (skeleton) using any human photo. Our pipeline integrates this control signal using a modified OpenPose ControlNet. While this option produces reasonable poses, it struggles with correct body shapes (row 1(e)). *Canny Edge* offers accurate body shape and pose but may yield sub-optimal results when the reference pose photo has different clothing than the input selfies. For instance, the hem of the jacket in row 2(f) appears less natural. In such cases, *OpenPose Skeleton* (row

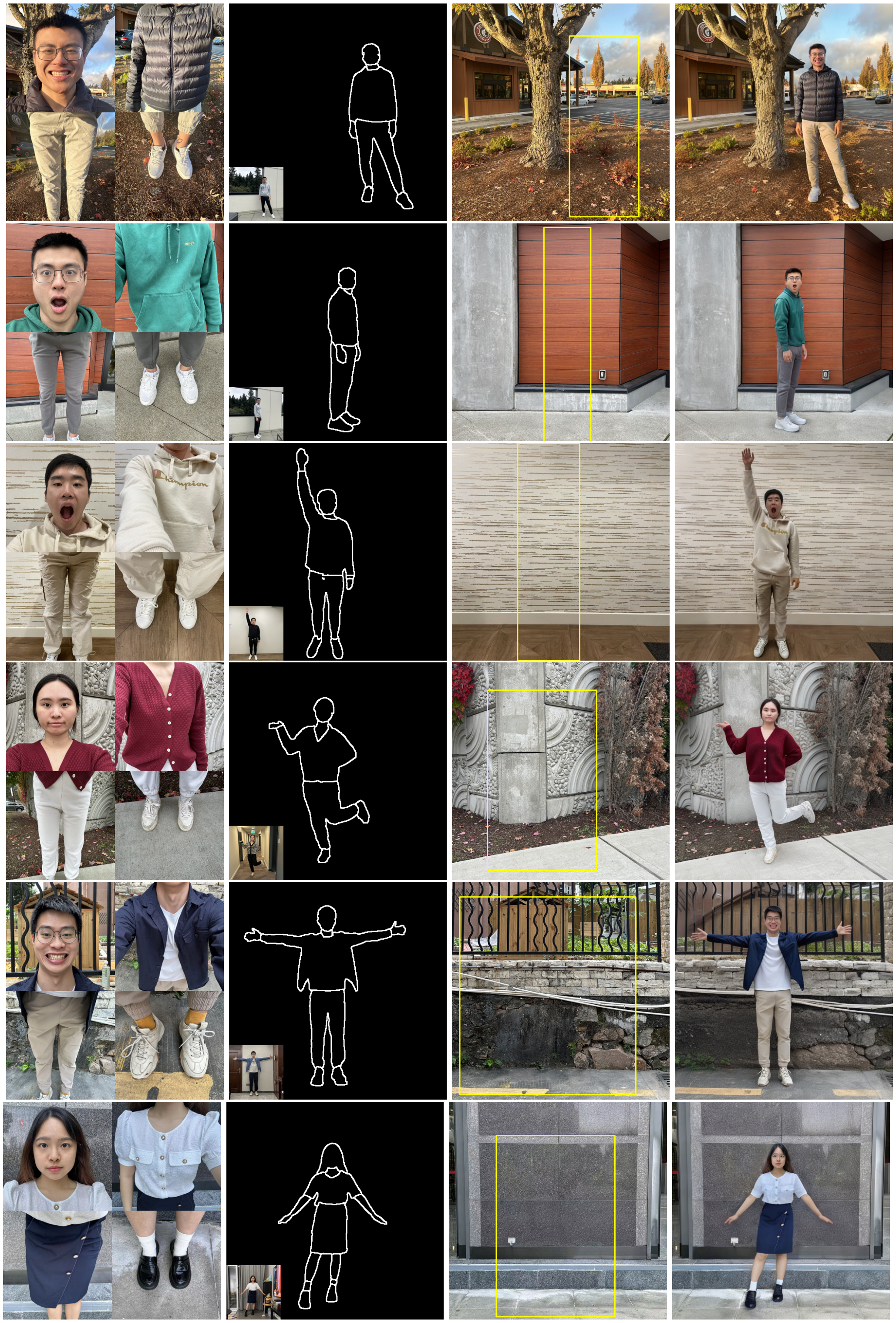
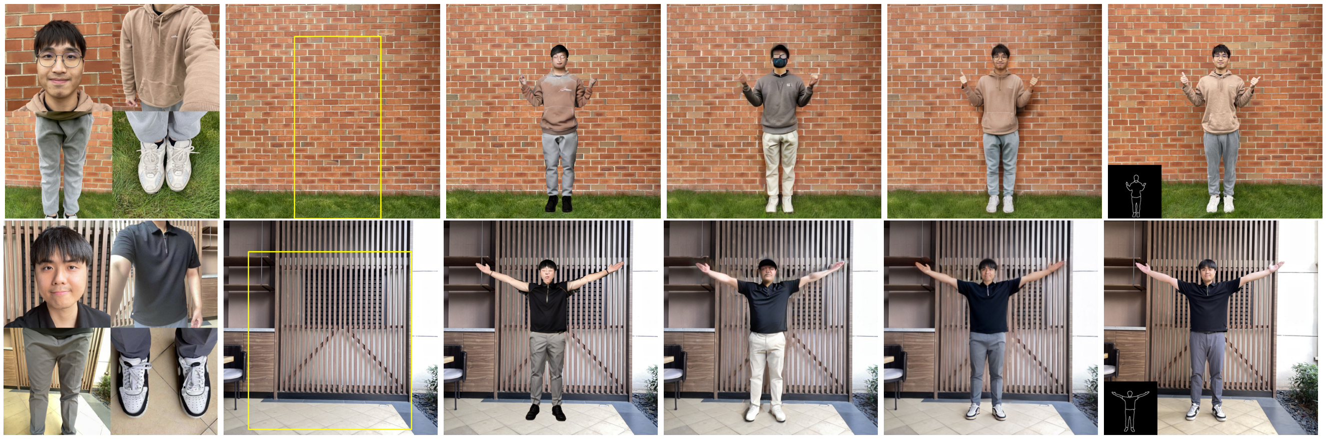
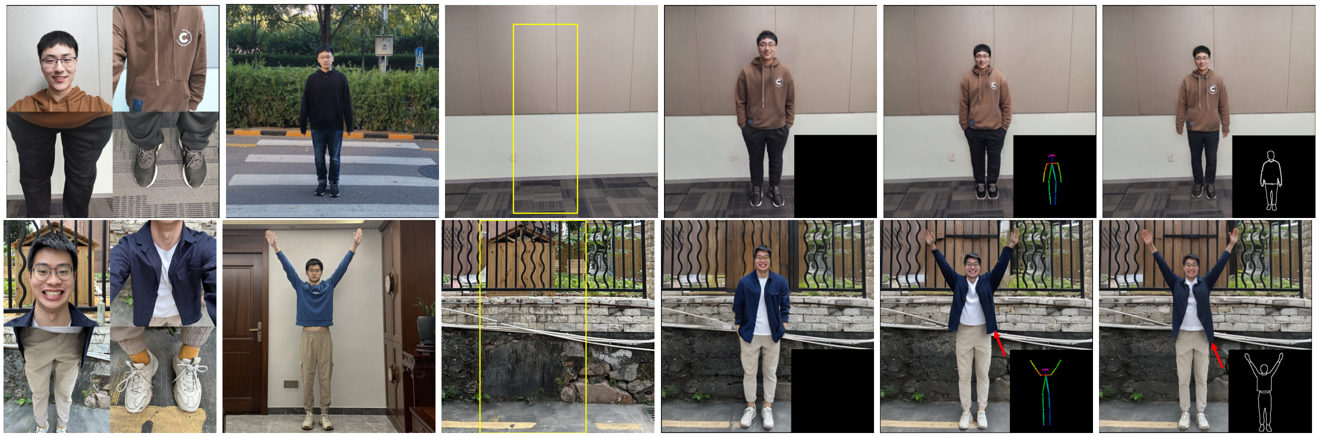

(a) Input Selfies (b) Target Pose (c) Masked Background (d) Total Selfie

Figure 4. Results. The second column shows the Canny Edge images detected from reference images (shown as insets). Regions inside yellow box of (c) are the masked regions. Total Selfie generates realistic, full-body images of different individuals with diverse poses and expressions against a variety of backgrounds, while preserving facial expression and clothing. The results are robust to selfies captured in different ways, such as those with one or two hands involved or from a downward-looking perspective (row 5), and with target pose images in outfits that differ somewhat from the input selfies.



(a) Input Selfies (b) Background (c) LaDI-VTON (d) DreamBooth (e) Ours (f) Real Photo

Figure 5. Qualitative comparison with two best-performing baselines. For this comparison, we used Canny Edge of the real photo as target pose (inset of (f)). Our pipeline clearly outperforms baselines in terms of photorealism and faithfulness (zoom in for details, including faces and shoes). Note that, while the selfies, background image, and real photo were captured in the same session, variations in lighting conditions, auto exposure, white balance, and other factors may result in intensity and color tone differences.

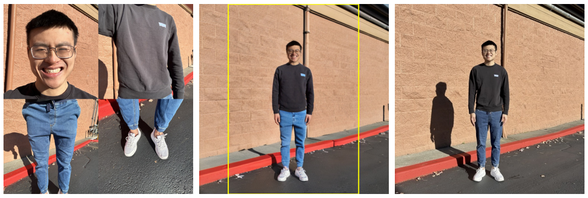


(a) Input Selfies (b) Reference Photo (c) Background (d) No Condition (e) OpenPose Skeleton (f) Canny Edge

Figure 6. Discussion of target pose options. Insets of (d)-(f) show the conditional signals. In the first row, Canny Edge (f) exhibits the closest body shape to the reference photo (b) compared to (d) and (e). In the second row, when the reference photo has a different clothing type, Canny Edge (f) produces an unnatural hem of the jacket (highlighted by a red arrow). In such cases, the OpenPose skeleton (e) may offer a more natural result despite a slight fattening of the waist compared to reference.

2(e)) can do a better job with clothing contours. In summary, each option has its advantages and limitations, and there is no one-size-fits-all choice. The selection of the target pose type depends on users' needs and available data.

Limitations and Future Work. Fig. 7 shows two limitations of Total Selfie: (1) While our method generally yields a harmonized output (b), the shading of the body may not precisely align with that in the actual photo (c). (2) Our method cannot accurately generate hard shadows of a person under strong sunlight since inferring the sun's direction and scene geometry solely from the background is difficult.



(a) Input Selfies (b) Ours (c) Real Photo

Figure 7. Failure case of shadow generation under strong sunlight. Masked regions are shown in (b), and target pose is from (c).

Topics of future work include explore how to effectively infer body shape and scale from input selfies, and automatically suggesting good target poses.

Conclusions. We introduce a new selfie type called *total selfie*, and propose a diffusion-based framework to generate it from four selfies, a background image, and a target pose. Our method generates faithful and realistic full-body selfies, outperforming existing techniques.

Acknowledgement. This work was supported by the UW Reality Lab, Meta, Google, OPPO, and Amazon.

References

[1] ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. 2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[3] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. *arXiv preprint arXiv:2211.12500*, 2022. 2

[4] Jia-Wang Bian, Huangying Zhan, and Ian Reid. Nvss: High-quality novel view selfie synthesis. In *2021 International Conference on 3D Vision (3DV)*, pages 1085–1094. IEEE, 2021. 2

[5] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 3

[6] Cristian Botezatu, Mathias Ibsen, Christian Rathgeb, and Christoph Busch. Fun selfie filters in face recognition: Impact assessment and removal. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):91–104, 2022. 2

[7] Nicola Capece, Francesco Banterle, Paolo Cignoni, Fabio Ganovelli, Roberto Scopigno, and Ugo Erra. Deepflash: Turning a flash selfie into a studio portrait. *Signal Processing: Image Communication*, 77:28–39, 2019. 2

[8] 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. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 4

[9] Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7513–7522, 2023. 2

[10] Junyang Chen, Xiaoyu Xian, Zhijing Yang, Tianshui Chen, Yongyi Lu, Yukai Shi, Jinshan Pan, and Liang Lin. Open-world pose transfer via sequential test-time adaption. *arXiv preprint arXiv:2303.10945*, 2023. 2

[11] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 2

[12] Zheng Chong, Xujie Zhang, Fuwei Zhao, Zhenyu Xie, and Xiaodan Liang. Fashion matrix: Editing photos by just talking. *arXiv preprint arXiv:2307.13240*, 2023. 2

[13] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14638–14647, 2021. 2

[14] Haipeng Fang, Zhihao Sun, Ziyao Huang, Fan Tang, Juan Cao, and Sheng Tang. Dance your latents: Consistent dance generation through spatial-temporal subspace attention guided by motion flow. *arXiv preprint arXiv:2310.14780*, 2023. 2

[15] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7723–7732, 2022. 2

[16] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 1–19. Springer, 2022.

[17] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Wayne Wu, and Ziwei Liu. Unitedhuman: Harnessing multi-source data for high-resolution human generation. *arXiv preprint*, arXiv:2309.14335, 2023. 2

[18] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[19] Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22768–22777, 2023. 2

[20] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, 2022. 2

[21] Rishabh Jain, Mayur Hemani, Duygu Ceylan, Krishna Kumar Singh, Jingwan Lu, Mausoom Sarkar, and Balaji Krishnamurthy. Umfuse: Unified multi view fusion for human editing applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7182–7191, 2023. 2

[22] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 2

[23] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18623–18632, 2022. 2

[24] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. 2

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[26] Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3d features. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 1044–1045, 2020. 2

[27] Laxman Kumarapu, Shiv Ram Dubey, Snehasis Mukherjee, Parkhi Mohan, Sree Pragna Vinnakoti, and Subhash Karthikeya. Wsd: Wild selfie dataset for face recognition in selfie images. *arXiv preprint arXiv:2302.07245*, 2023. 2

[28] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 2

[29] Nannan Li, Kevin J Shih, and Bryan A Plummer. Collecting the puzzle pieces: Disentangled self-driven human pose transfer by permuting textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7126–7137, 2023. 2

[30] Yuanming Li, Youngsaeng Jin, Jeonggi Kwak, Dongsik Yoon, David Han, and Hanseok Ko. Adaptive content feature enhancement gan for multimodal selfie to anime translation. 2021. 2

[31] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22788–22797, 2023. 2

[32] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. 2021. 2

[33] Liqian Ma, Zhe Lin, Connelly Barnes, Alexei A Efros, and Jingwan Lu. Unselfie: Translating selfies to neutral-pose portraits in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 156–173. Springer, 2020. 2

[34] Liyuan Ma, Tingwei Gao, Haitian Jiang, Haibin Shen, and Kejie Huang. Waveipt: Joint attention and flow alignment in the wavelet domain for pose transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7215–7225, 2023. 2

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 5

[36] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. *arXiv preprint arXiv:2305.13501*, 2023. 2, 6

[37] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[38] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[40] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 3

[41] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: flaws in word-to-concept mapping in text2image models. *arXiv preprint arXiv:2210.10606*, 2022. 2

[42] Jian Ren, Menglei Chai, Oliver J Woodford, Kyle Olszewski, and Sergey Tulyakov. Flow guided transformable bottleneck networks for motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10795–10805, 2021. 2

[43] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing*, 2020.

[44] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022. 2

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 4, 5

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2, 5, 6

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[48] Soubhik Sanyal, Alex Vorobiov, Timo Bolkart, Matthew Loper, Betty Mohler, Larry S Davis, Javier Romero, and Michael J Black. Learning realistic human reposing using cyclic self-supervision with 3d shape, pose, and appearance consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11138–11147, 2021. 2

[49] Soubhik Sanyal, Partha Ghosh, Jinlong Yang, Michael J Black, Justus Thies, and Timo Bolkart. Sculpt: Shape-conditioned unpaired learning of pose-dependent clothed and textured human meshes. *arXiv preprint arXiv:2308.10638*, 2023. 2

[50] YiChang Shih, Wei-Sheng Lai, and Chia-Kai Liang.

Distortion-free wide-angle portraits on camera phones. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 4

[51] Luming Tang, Nataniel Ruiz, Chu Qinghao, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, and Michael Rubinstein. Realfill: Reference-driven generation for authentic image completion. *arXiv preprint arXiv:2309.16668*, 2023. 2

[52] Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, and Yihui Ren. Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 702–712, 2023. 2

[53] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 2, 6

[54] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4

[55] Zhixiang Wang, Yu-Lun Liu, Jia-Bin Huang, Shin'ichi Satoh, Sizhuo Ma, Gurunandan Krishnan, and Jian Wang. Disco: Portrait distortion correction with perspective-aware 3d gans. *arXiv preprint arXiv:2302.12253*, 2023. 4

[56] Zhenzhen Weng, Laura Bravo-Sánchez, and Serena Yeung. Diffusion-hpc: Generating synthetic images with realistic humans. *arXiv preprint arXiv:2303.09541*, 2023. 2

[57] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023. 2

[58] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[59] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17194–17204, 2023. 2

[60] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 2, 3, 4, 6

[61] Lu Yang, Wenhe Jia, Shan Li, and Qing Song. Deep learning technique for human parsing: A survey and outlook. *arXiv preprint arXiv:2301.00394*, 2023. 4, 5

[62] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3dhumangan: 3d-aware human image generation with 3d pose mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23008–23019, 2023. 2

[63] Jiyang Yu and Ravi Ramamoorthi. Selfie video stabilization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 551–566, 2018. 2

[64] Jiyang Yu, Ravi Ramamoorthi, Keli Cheng, Michel Sarkis, and Ning Bi. Real-time selfie video stabilization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12036–12044, 2021. 2

[65] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3, 4, 6

[66] Yahui Zhang, Shaodi You, Sezer Karaoglu, and Theo Gevers. Pose guided human motion transfer by exploiting 2d and 3d information. In *2022 International Conference on 3D Vision (3DV)*, pages 587–595. IEEE, 2022. 2

[67] Yajie Zhao, Zeng Huang, Tianye Li, Weikai Chen, Chloe LeGendre, Xinglei Ren, Ari Shapiro, and Hao Li. Learning perspective undistortion of portraits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7849–7859, 2019. 4

[68] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. *arXiv preprint arXiv:2208.00712*, 2022. 2

[69] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 2