Unselfie: Translating Selfies to Neutral-pose Portraits in the Wild

Liqian Ma[1], Zhe Lin[2], Connelly Barnes[2], Alexei A Efros[2,3], and Jingwan Lu[2]

[1]KU Leuven   [2]Adobe Research   [3]UC Berkeley

Abstract. Due to the ubiquity of smartphones, it is popular to take photos of one's self, or "selfies." Such photos are convenient to take, because they do not require specialized equipment or a third-party photographer. However, in selfies, constraints such as human arm length often make the body pose look unnatural. To address this issue, we introduce *unselfie*, a novel photographic transformation that automatically translates a selfie into a neutral-pose portrait. To achieve this, we first collect an unpaired dataset, and introduce a way to synthesize paired training data for self-supervised learning. Then, to *unselfie* a photo, we propose a new three-stage pipeline, where we first find a target neutral pose, inpaint the body texture, and finally refine and composite the person on the background. To obtain a suitable target neutral pose, we propose a novel nearest pose search module that makes the reposing task easier and enables the generation of multiple neutral-pose results among which users can choose the best one they like. Qualitative and quantitative evaluations show the superiority of our pipeline over alternatives.

Keywords: Image Editing, Selfie, Human Pose Transfer.

1   Introduction

Smartphone cameras have democratized photography by allowing casual users to take high-quality photos. However, there still remains a tension between the ease of capture and the photograph's quality. This is particularly apparent in the case

Fig. 1: We automatically *unselfie* selfie photos into neutral-pose portraits.

of personal portraits. On one hand, it is easy to take a photo of oneself (a selfie) by using the front camera of a smartphone. On the other hand, one can usually take a much higher-quality photograph by relying on an extra photographer, or equipment such as a tripod or selfie stick. While less convenient, these avoid the compositional problem that tends to occur in selfies: an unnatural body pose.

In this paper, we introduce a new photographic transformation that we call *unselfie*. This transformation aims to make selfie photos look like a well-composed portrait, captured by a third party photographer, showing a neutral body pose with relaxed arms, shoulders and torso. We call this desired result a "neutral-pose portrait." The unselfie transform moves any raised arms downward, adjusts the pose of the shoulder and torso, tweaks the details of the clothing and then fills in any exposed background regions (see Figure 1).

There are three main challenges that we need to tackle in order to be able to *unselfie* photos: (1) Paired (selfie, neutral-pose portrait) training data do not exist, so we need to train a model without such data; (2) The same selfie pose can reasonably correspond to multiple plausible neutral poses, so we need to be able to handle this multiplicity; (3) Changing the pose creates holes in the background, so we need to fill in the holes while maintaining a smooth transition between the background and the human body.

We first tried out several previous methods to see if they could address challenge (1). We collected separate sets of selfie and neutral-pose portraits and used the unsupervised approach CycleGAN [62] for unpaired translation. CycleGAN excels at appearance-level translation that modifies textures and colors, but cannot perform large geometric transformations which are often needed for reposing the complex human body. It also produces unnatural poses with artifacts that result in more noticeable artifacts later in our pipeline. We also tried unsupervised person generation approaches [32,10]. Though better than CycleGAN, these are not designed for our Unselfie task and produce lower quality results than seen in their papers. As shown in our experiments, these methods result in noticeable artifacts on the generated person images, and texture details are missing because appearance information is compressed heavily.

Due to these reasons, we instead propose to synthesize (selfie, neutral-pose portrait) pairs and use a self-supervised learning approach. In particular, we propose a way to synthesize paired selfie images from neutral-pose portraits by using a non-parametric nearest pose search module to retrieve the nearest selfie pose given a neutral-pose portrait, and then synthesize a corresponding selfie. We also adopt a nearest pose search module during inference. Given an input selfie pose, we retrieve the best matching neutral poses, which we use to synthesize the final portraits. This addresses challenge (2) by enabling diverse outputs to be synthesized and allowing users to choose among them.

The synthesized paired data mentioned above can be directly used to train a supervised person image generation network like [33,50,63], but there still exist noticeable artifacts in the results as shown in our experiments. These methods are sensitive to the pixel-level domain gap between our synthetic paired training data and the real selfies testing data (see Fig. 3). Inspired by [14], we use the
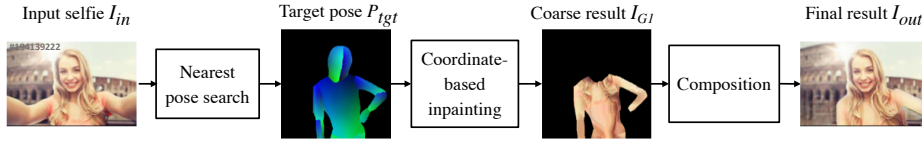
Fig. 2: Our three-stage pipeline. Based on the input selfie I_{in}, we extract its pose information using DensePose [1]. We perform nearest neighbour search on the pose representation to find the target neutral pose P_{tgt} that has the most similar pose configuration in the upper torso region. Using DensePose, we map the pixels in the input selfie to the visible regions of the target pose and then use coordinate-based inpainting [14] to synthesize a coarse human body. We then use a composition step to refine the coarse result by adding more details and composite it into the original background.

coordinate-based inpainting method to inpaint the body texture in UV space. This space is mostly invariant to the original body pose, and is therefore more robust to imperfections in the synthesized data. Additionally, the coordinate-based inpainting method can reuse visible pixels and thus give sharper results.

To address challenge (3), we adopt a gated convolutional layer [58] based composition network to jointly refine the body appearance, fill the background holes, and maintain smooth transitions between the human body and background.

Overall, to address the unselfie task, we propose a three-stage pipeline shown in Fig. 2: we first search for a nearest neutral pose in a database, then perform coordinate-based inpainting of the body texture, and finally use a composition module to refine the result and composite it on the background. We conducted several experiments and compared our method with alternatives to demonstrate the effectiveness of our pipeline.

To the best of our knowledge, this work is the first to target the problem of selfie to neutral-pose portrait translation, which could be a useful and popular application among casual photographers. Our contributions include: 1) We collect a new dataset of unpaired selfies and neutral-pose portraits and introduce a way to synthesize paired training data for self-supervised learning; 2) We introduce a three-stage pipeline to translate selfies into neutral-pose portraits; 3) We propose a novel nearest pose search module to obtain suitable multi-modal target neutral poses; 4) We design an all-in-one composition module to refine the foreground, complete the background, and compose them together seamlessly.

2   Related work

Image Generation and Translation. Generative models, such as VAEs [22] and GANs [13,45,34,2,5,21] can synthesize realistic-looking images from noise. To allow more control during the generation process, much research has been devoted to conditional image generation, with class labels [37], attributes [7], text [60,48], key-points [47,32] and images [20,62] as conditioning signals. Image-to-image

translation networks are conditioned on images, such as semantic segmentation maps, edge maps and RGB images [20]. To alleviate the need of collecting paired data, researchers have introduced unsupervised methods based on the ideas of cycle-consistency [62] and shared latent space [28]. Subsequent works [24,19,31] further extended the unsupervised approach to solve multi-modal image-to-image translation problems by disentangling content and style representations. These methods mainly focus on appearance manipulation. Recently, research efforts have also been extended to modify the underlying geometry in image translation tasks [53,44]. In general, unsupervised image manipulation is quite challenging [26], especially if the goal is to modify the underlying geometry.

Image Completion. The goal of image completion is to fill in missing regions of an image. Applications of image completion include image inpainting and image out-painting. Traditional patch-based texture synthesis approaches [3,55,8,18] work well for images containing simple and repetitive structures, but may fail to handle images of complex scenes. To tackle this problem, modern approaches apply deep learning for image completion due to its ability to gain semantic understanding of the image content [41,57,56,39,58]. One of the first deep learning methods for image inpainting is context encoder [41], which uses an encoder-decoder architecture. In order to achieve better results, some prior works apply the PatchMatch idea [3] at the feature level, such as Contextual Attention [57] and ShiftNet [56]. Recently, Yu *et al.* [58] introduce a gated convolutional layer, which learns a dynamic feature gating mechanism for each channel and each spatial location. In our framework, we also use the gated convolutional layer [58] in both the coordinate-based inpainting network and the composition network to fill in holes in the UV map and the background.

Person Image Generation. Person image generation is a challenging task, as human bodies have complex non-rigid structures with many degrees of freedom [38], Previous works in this space usually generate person images by conditioning on these structures. Ma *et al.* [33] propose to condition on image and pose keypoints to transfer the human pose. [23] and [9] generate clothed person by conditioning on fine-grained body and clothing segmentation. Recent works [36,52] also extend the conditions to attributes. To model the correspondences between two human poses explicitly, recent works introduce flow-based techniques which improve the appearance transfer result quality significantly [50,63,40,14,15,29,49]. Siarohin *et al.* [50] propose deformable skip-connections to warp the feature maps with affine transformations. Grigorev *et al.* [14] propose to inpaint the body texture in the UV space which is mostly invariant to the pose in the image space. Although these methods achieve good results, they need paired training data, *i.e.* two images containing the same individual in two different body poses, which may be difficult to collect in many applications. To address this, several unpaired methods have been proposed [32,10,42]. [32] and [10] propose to decompose a person image into different factors which are then used to reconstruct the original image. Other works also adopt human parsing algorithms to help out on the difficult unpaired setting [46,51]. Raj *et al.* [46] generate training pairs from a single image via data augmentation. Inspired by [46], we synthesize selfie

data from neutral-pose portrait data to construct paired training data. These human synthesis approaches focus on generating realistic human appearance in a relatively simple background environment (*i.e.* fashion or surveillance datasets) given a target pose. Our work on the other hand handles selfie photos captured in the wild that contain a wide variety of backgrounds, lighting conditions, identities and poses. Compared to fashion photos, the background pixels in selfies are of greater importance.

3   Our Method

Our goal is to convert a selfie I_{in} into a neutral-pose portrait I_{out} as shown in Fig. 2. We collect separate sets of selfie and portrait photos and synthesize paired training data for self-supervised learning (Sect.3.1). Due to the complexity of the problem, we solve it in three stages. In the first stage, we use a non-parametric nearest pose search module (Sect.3.2) to find the target neutral pose P_{tgt} that closely matches the pose in the upper torso region of the selfie. We then map the pixels in the selfie to regions of the target pose based on the correspondences between the two pose representations. This design makes the remaining problem easier, since most of the pixels can be directly borrowed from the input selfie and thus fewer pixels need to be modified by the remaining steps. In the second stage, inspired by the previous work [14], we train a coordinate-based inpainting model to synthesize the coarse appearance of the human body (Sect.3.3). In the final stage, we train a composition model to synthesize details and fix artifacts in the body region caused by pose changes, fill in holes in the background, and seamlessly compose the synthesized body onto the original photo (Sect.3.4).

We use DensePose [1] in all three stages of our *unselfie* pipeline. Unlike keypoint based representations [6], which predict a limited amount of pose information described by sparse keypoints, DensePose provides a dense UV-map for the entire visible human body. The UV-map is an ideal pose representation for our purposes, because it provides body shape and orientation information, which are useful for the pose search module. Color values in UV space are also relatively invariant to the person's pose, so this also enables the coordinate-based inpainting step to produce sharp results.

3.1   Datasets

We are not aware of any datasets that contain selfie photos and their corresponding portrait photos taken at the exact same place and time. However, there are many unpaired selfies and neutral-pose portraits online.

Unpaired selfies and neutral-pose portraits. We collect 23169 photos of people in frontal and neutral poses from the following public datasets: DeepFashion [30], DeepFashion2 [11], and ATR [25]. We apply the DensePose algorithm to extract the pose information from all the images. The extracted DensePose representations form a neutral pose database $\{P^i_{neutral}\}$. Because many images in DeepFashion dataset have clean backgrounds, they are not diverse enough for

Ground truth portrait image I_{tgt} Portrait pose P_{tgt} Nearest selfie pose P_{src} From $\{P^i_{selfie}\}$ Synthesized selfie image I_{src}
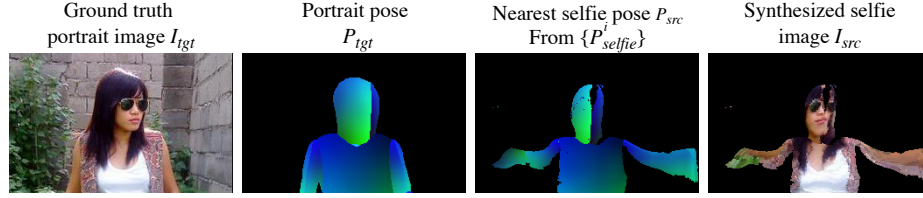
Fig. 3: Synthesized (portrait, selfie) pairs in image space. Given a neutral-pose portrait I_{tgt}, we use DensePose to extract its pose P_{tgt}. We perform nearest neighbour search to find the closest pose from the selfie pose database. The selfie image I_{src} is synthesized from pixels in I_{tgt} by the correspondence between P_{tgt} and P_{src}. The displayed I_{tgt} and P_{tgt} are cropped due to alignment (Sect. 3.2)

Ground truth portrait texture-map T_{tgt} Synthesized selfie coordinate-map C_{src} Synthesized selfie texture-map T_{src}
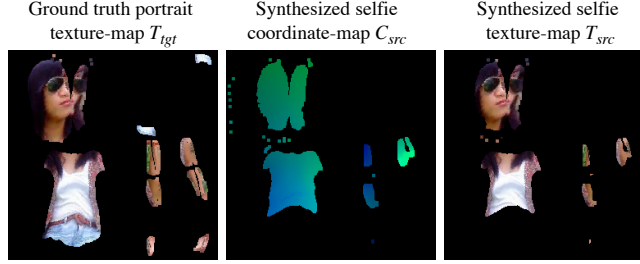
Fig. 4: Synthesized (portrait, selfie) pairs in UV space to train our coordinate-based inpainting network.

our composition network to learn proper background inpainting. Therefore, we apply a state-of-the-art matting technique [54] to extract the foreground humans and paste them into random background images to increase the data diversity.

We collect 4614 selfie photos from the Internet using the following strategy. We first search with keywords like "selfie girl," "selfie boy," etc. Many photos returned by the search engines contain cellphones. These are not the selfies we desire but are third-person view photos of someone taking selfies. Since Mask R-CNN [16] is pretrained on the COCO dataset [27], which contains person and cell phone classes, we use it to select photos that contain a single person without any cell phones. We then eliminate photos that have disconnected body parts or have any of the frontal upper body parts missing in the DensePose representation. We also use this strategy to create the neutral-pose portrait dataset. Finally, we manually clean up the remaining data in case any of the previous filters fail. We create a 4114/500 split for training and testing. We denote the DensePose representation of selfie photos as $\{P^i_{selfie}\}$.

Synthesized paired training data. To allow self-supervised training, based on the collected neutral-pose portraits, we synthesize their corresponding selfie data using DensePose. As shown in Fig. 3, given a neutral pose P_{tgt}, we first search for a selfie pose P_{src} from $\{P^i_{selfie}\}$ that matches the input neutral pose the best in the upper torso region. Through DensePose correspondences, we

Input selfie pose P_{in} Source pose P_{src} Pose database $\{P^i\}_N$ Target pose P_{tgt}
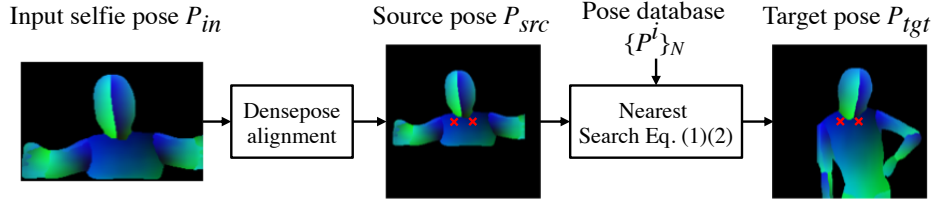

Fig. 5: Nearest pose search module. The detected selfie DensePose P_{in} is first aligned so that the two shoulder points are centered in the image P_{src}. Then we perform nearest neighbour search to find a target pose P_{tgt} from the neutral pose database that closely matches the input selfie pose in the upper torso region.

map the portrait image pixels to the nearest selfie pose (see Sect. 3.2 for more details). Due to the pose change and self-occlusion, the synthesized selfie might contain holes caused by mistakes of DensePose or by pixels visible in the selfie but not visible in the original portrait such as the armpits or the backside of arms. Though not perfect, these paired images can be used to train supervised human synthesis models like [33,50,63]. However, the results have noticeable artifacts as shown in our experiments in Sect. 4.1.

Instead of using the synthetic paired images to synthesize pixels directly, we convert them into the UV space (see Fig. 4) and perform texture inpainting in the UV coordinate space by building on [14]. In particular, we first obtain ground truth portrait texture map T_{tgt} from I_{tgt} with DensePose mapping. Then, we obtain the selfie coordinate map C_{src} from the nearest selfie pose P_{src} masked by the visible region of T_{tgt}. Finally from C_{src}, we sample the pixels from T_{tgt} to synthesize the selfie texture map T_{src}. C_{src} and T_{src} are used as input to train the coordinate-based inpainting model (see Sect.3.3).

3.2   Nearest Pose Search

Because our goal is to turn a selfie into a neutral-pose portrait, it is important to define what we desire for the target neutral pose. Motivated by the success of retrieval-based image synthesis [43], we propose a retrieval-based approach where given a selfie pose at testing time, we perform non-parametric nearest pose search to find the neutral poses in the $\{P^i_{neutral}\}$ database that match the input selfie pose the best. Compared to pixel-level pose translation using approaches like CycleGAN [62], our approach has several advantages: (1) It is simpler and more explainable since no training is needed; (2) the retrieved target poses are guaranteed to be natural since they come from real photos; (3) we can generate multiple unselfie results by choosing the top-K most similar poses as the target poses and we can allow users to choose their favorite result; (4) the retrieved poses are similar to the input pose which makes pose correction easier since fewer pixels need to be modified.

During inference, given an input selfie, we search for the k-nearest target neutral poses. At training time, we reverse the search direction: given a target

neutral pose, we search for a matching selfie. This allows us to synthesize synthetic data, which are used to self-supervise the inpainting and composition networks. The procedure is otherwise the same at both training and inference time.

In the remainder of this subsection, we describe the details of our pose search module. As shown in Fig. 5, we first align the input selfie pose P_{in} by putting two selected shoulder points in the center of the image to obtain the source pose P_{src}. All neutral poses are also aligned in the same way. We calculate the pose similarity in the frontal torso region excluding the head, since the later stages of our pipeline keep the head region intact and only correct the body pose. The DensePose representation P is an IUV map which contains three channels. The P^I channel contains indices of body parts to which pixels belong, and the P^{UV} channels contain the UV coordinates.

Based on P^I and P^{UV}, we propose a two-step search strategy to calculate pose similarity. First, we search for suitable target poses based on global information such as body shape and position. To determine the global similarity between two poses, we use the following equation:

$$d^I(P_1, P_2) = \sum_{x \in R_1 \cup R_2} \mathbb{1}(P_1^I(x) \neq P_2^I(x)), \qquad (1)$$

where R refers to the front torso regions of the body. We iterate over all pixels in both torso regions and count the number of pixels that belong to different body parts in the two poses. If there is large body part index mismatch in the torso regions, the two poses are dissimilar at a global level.

Among the top-K pose candidates selected based on d_I, we further improve the ranking by leveraging local pose information given by the UV coordinates. In particular, for pixels belonging to torso regions in both poses, we calculate the sum of the distances of their UV coordinates:

$$d^{UV}(P_1, P_2) = \sum_{x \in R_1 \cap R_2} \|P_1^{UV}(x) - P_2^{UV}(x)\|_2. \qquad (2)$$

3.3   Coordinate-based Inpainting

Inspired by self-supervised image inpainting work [41,58] and human synthesis work [14], we learn to reuse the visible body pixels to fill in the invisible body parts. As illustrated in Fig. 6 left, we first use an Image-to-UV (I2UV) mapping to translate pose P_{src} and the color image I_{src} from the image domain to the UV domain. Defined in the UV domain, C_{src} stores the associated $\{x, y\}$ coordinates of pixels in the original image space. Likewise in the UV domain, T_{src} contains the RGB colors of the associated pixels in the original image I_{src}: these are looked up by using bilinear sampling via $T_{src} = I_{src}(C_{src})$.

After the I2UV mapping, we use an inpainting neural network G_1 to inpaint the coordinate-map C_{src}. We concatenate C_{src} and T_{src} as input to the network. The network outputs the inpainted coordinate-map $C_{G_1} = G_1(C_{src}, T_{src})$. We then perform bilinear sampling to obtain the inpainted texture-map $T_{G_1} = I_{src}(C_{G_1})$.
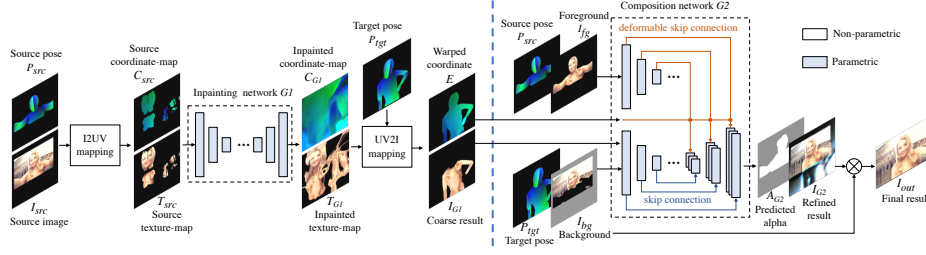
Fig. 6: Left: coordinate-based inpainting stage. Right: composition stage.

Finally, we map C_{G_1} and T_{G_1} back to the image space with UV-to-Image (UV2I) mapping using the bilinear sampling operations $E = C_{G_1}(P_{tgt})$, $I_{G_1} = T_{G_1}(P_{tgt})$. To train G_1, we use three loss functions, identity loss $L_{idt}^{G_1}$, reconstruction loss $L_1^{G_1}$ and perceptual loss $L_P^{G_1}$ [61] as follows,

$$L_{idt}^{G_1} = \mathbb{E}\big[\|C_{G_1} - C_{src}\|_2^2 V_{src}\big], \tag{3}$$

$$L_1^{G_1} = \mathbb{E}\big[\|T_{G_1} - T_{tgt}\|_1 V_{tgt}\big], \tag{4}$$

$$L_P^{G_1} = \mathbb{E}\big[\|\phi(T_{G_1}) - \phi(T_{tgt})\|_2^2 V_{tgt}\big], \tag{5}$$

V_{src} and V_{tgt} are binary masks that select the non-empty regions in the coordinate map C_{src} and T_{tgt}, respectively. T_{tgt} is the ground truth texture mapped from image domain to UV domain (Fig. 4 left), that is, $T_{tgt} = I2UV(I_{tgt})$. The identity loss encourages the existing coordinates to stay unchanged while the network synthesizes coordinates elsewhere. The reconstruction loss and perceptual loss are performed in the pixel space instead of the coordinate space and use the ground truth image for supervision. The overall loss for G_1 is

$$\min_{G_1} L^{G_1} = L_1^{G_1} + \lambda_1 L_P^{G_1} + \lambda_2 L_{idt}^{G_1}, \tag{6}$$

3.4   Composition

The advantage of doing inpainting in the coordinate space is that the network can copy and paste original pixels to fill in missing regions based on body symmetry and therefore the synthesized pixels tend to look sharp. However, in some cases, a small number of visible pixels get copied into a much larger region resulting in flat and unrealistic appearance. In addition, when arms are moved down, holes will appear in the background due to dis-occlusion.

To address these problems, we use an all-in-one composition network to add details and fix artifacts in the body region and fill in the gaps between the body and the background by synthesizing a natural transition. As illustrated in Fig. 6 right, we use a U-net architecture equipped with gated convolutional layers [58] for G_2. The U-net architecture helps preserve the high resolution features through skip-connections. The gated convolutional layer improves the

inpainting result quality when dealing with holes of arbitrary shape. To keep the body appearance more consistent, we also use deformable skip connections [50,14] to propagate the appearance information from the source image I_{src} to the result despite large changes in poses. The network synthesizes missing foreground pixels, fills in background holes and also produces an alpha mask A_{G_2}. A_{G_2} is used to blend the synthesized pixels I_{G_2} into the background image I_{bg}, i.e. $I_{out} = I_{G_2}A_{G_2} + I_{bg}(1 - A_{G_2})$. We add the original head and neck regions into I_{bg} so that after blending the head regions will remain untouched.

To train G_2, we apply reconstruction loss $L_1^{G_2}$, perceptual loss $L_P^{G_2}$ [61], alpha loss $L_A^{G_2}$ and adversarial loss $L_{adv}^{G_2,D}$,

$$L_1^{G_2} = \mathbb{E}\big[\|I_{out} - I_{tgt}\|_1(1 + H)\big], \tag{7}$$

$$L_P^{G_2} = \mathbb{E}\big[\|\phi(I_{out}) - \phi(I_{tgt})\|_2^2(1 + H)\big], \tag{8}$$

$$L_A^{G_2} = \mathbb{E}\big[\|A_{G_2} - H\|_1\big], \tag{9}$$

$$\min_{G_2} \max_D L_{adv}^{G_2,D} = \mathbb{E}\big[(D(I_{tgt}))^2(1 + H)\big]$$
$$+ \mathbb{E}\big[(D(I_{out}) - 1)^2(1 + H)\big], \tag{10}$$

where I_{tgt} denotes the ground truth neutral-pose portrait. $H \in [0,1]$ is a binary spatial mask to encourage the network to focus more on synthesizing foreground and filling dis-occluded holes and the details are explained later. When applied to different spatial size, H will be resized to the corresponding spatial size by nearest-neighbor scaling accordingly. As to the adversarial learning, we use the same residual discriminator as that of [63]. The overall loss for G_2 is

$$\min_{G_2} \max_D L^{G_2} = \lambda_3 L_1^{G_2} + \lambda_4 L_P^{G_2} + \lambda_5 L_{adv}^{G_2,D} + L_A^{G_2}. \tag{11}$$

There is a big domain gap between the training and testing data. During testing, arms in real selfies are moved downward revealing a large hole in the background. During training, we also mimic the dis-occluded background holes. In particular, we calculate a binary mask $H = H_{selfie} \cup H_{neutral}$, which is also used in Eq. 7 to 10. H_{selfie} and $H_{neutral}$, which are estimated using an off-the-shelf DeepMatting model [54] and binarized with threshold 0.1, denote the body regions from the selfie and the neutral-pose portrait, respectively. The synthesized hole mask H is then applied to I_{bg} to mimic dis-occluded background holes.

4   Experiments

We compare our approach with several prior work through a qualitative evaluation, a user study and a quantitative evaluation[1]. Note that none of the previous approaches address exactly our unselfie problem, so we cannot compare our approach with previous work using their datasets and result quality for previous work on our dataset is worse than the result quality in those papers. We present

---

[1] More results and implementation details are reported in the supplementary materials.

ablation studies to validate the effectiveness of different algorithm choices. Finally, we discuss the limitations and future work. If not otherwise specified, we use the top-1 retrieved neutral pose as target pose. Note that there may be JPEG compression artifacts in the input selfies.

4.1   Comparisons with existing methods

Since we defined a brand new unselfie application, there is no prior work to compare to that addresses the exact same problem. Nevertheless, we introduce some modifications to two state-of-the-art human synthesis methods, DPIG [32] and PATN [63], so that we can compare to them directly in our new application setting. Note that these methods synthesize pixels based on a pre-specified target pose. To make their approaches work, we need to perform our proposed nearest pose search module to calculate P_{tgt} and then use their approaches to synthesize the final pixels. DPIG is a key-points based unsupervised approach. For fair comparison, we replace their key-points with the DensePose representation. We also made various other improvements for it to produce comparable results to ours (see supplementary material). PATN is a key-points based supervised method, so we use our synthesized paired images for self-supervised training by using DensePose IUV map as the input pose representation and feeding I_{src}, I_{bg} as input to their model. In the supplementary material, we also compare our approach with another keypoint-based unsupervised approach VUNET [10]. Due to low result quality and space limitations, we do not show those results here.

Qualitative evaluation. Fig. 7 shows that our method synthesizes more photo-realistic body and background appearance compared to prior art. We manually picked the best target pose out of results using our top-5 retrieved poses. The multi-modal results are reported in the supplementary material. In the top row, the short-sleeved clothing style is better preserved. In the second row, the stripe pattern is better preserved, and the synthesized arms are sharper. In the bottom two rows, our method synthesizes better clothing and background details and produces more natural transitions between foreground and background. The reasons that our method outperforms the baselines are: 1) unsupervised methods, like DPIG, encode images into heavily compressed features, which results in loss of details and texture inconsistency between the generated output and the input. They perform well in more constrained settings (clean backgrounds and simple texture), while our task involves complex images in the wild; 2) these baseline methods are more sensitive to the domain gap between training and testing data since they directly synthesize image pixels. Our method performs foreground inpainting in the coordinate space and then uses a composition module to refine details and fill in background holes and thus is less sensitive to the domain gap between imperfect synthesized selfies at training and perfect selfies at testing.

User study. For a useful real-world application, we believe qualitative perceptual evaluation is more important. Thus, we perform a user study on Amazon Mechanical Turk (AMT). Similar to previous works [19,24], given the input selfie and a pair of results generated by our approach and one of the baseline approaches, users are asked to pick one that looks better than the other. Within

Input selfie   DPIG [32]   PATN [63]   Ours

Fig. 7: Comparisons with state-of-the-art methods. Please zoom in to see details.

each Human Intelligence Task (HIT), we compare our method with the same baseline method. We randomly generate 200 result pairs including 10 sanity pairs where the answers are obvious. After filtering out careless users based on their answers on the sanity pairs, we calculate the user study statistics using the remaining 190 questions. We have three HITs for three baseline methods. Each HIT is done by 20 users. As shown in Table. 1, our method is preferred over others. We assume a null hypothesis that on average, half the users prefer ours for a given paired comparison. We use a one-sample permutation t-test [12] to measure p using 10^6 permutations and find $p < 10^{-6}$ for the 3 baselines.

Quantitative evaluation. Since we do not have the ground truth neutral portraits corresponding to input selfies, we cannot use metrics like SSIM. To quantitatively compare our result quality with other baselines, we report Frechet

Model	Human Prefers Ours	FID↓	KID↓
DPIG [32]	0.798	88.27	0.026
VUNET [10]	0.851	135.90	0.077
PATN [63]	0.822	104.74	0.041
Ours	N/A	**71.93**	**0.014**

Table 1: User study and, FID/KID scores.

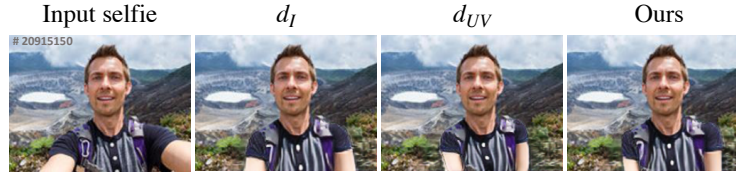| Model	FID↓	KID↓
Ours w/o $L_P^{G_2}$	82.09	0.019
Ours w/o Deform	73.87	0.017
Ours w/o Gated	72.89	**0.014**
Ours	**71.93**	**0.014**

Table 2: Ablation study.

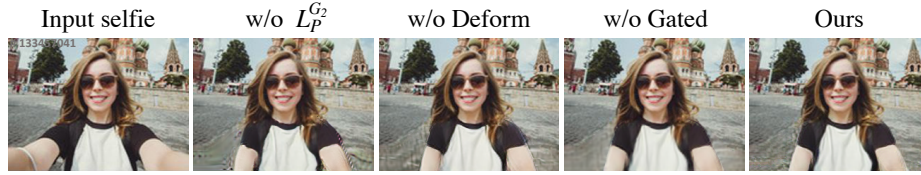Fig. 8: Ablation study results for nearest pose search.

Fig. 9: Ablation study for composition network.

Inception Distance (FID) [17] and Kernel Inception Distance (KID) [4] as shown in Table. 1. We aim to translate the body into a neutral pose while keeping the rest of the image intact. Therefore, a good translation method should have low FID and KID values when compared to both the portraits and the selfie domains. As suggested by [35], we combine both real selfie and real portrait images into the real domain, and compute the FID and KID values between our synthesized results (*i.e.* fake domain) and the real domain. The mean FID and KID values are averaged over 10 different splits of size 50 randomly sampled from each domain. The trend of FID and KID is consistent with the user study result, and our method outperforms others significantly.

4.2   Ablation Study

Analysis of the pose search module. We compare three ablation settings: 1) using body index distance only (*i.e.* d_I); 2) using UV distance only (*i.e.* d_{UV}); 3) using our two-step strategy. As shown in Fig. 8, the neutral poses retrieved by d_I have a reasonable body shape and size but the local coordinates mismatch resulting in the cloth regions being distorted (see the elongated buttons). For the neutral poses retrieved by d_{UV}, the body shape and size are not compatible with the head region. The retrieved body part is too small. Our two-step strategy combines the benefits of d_I and d_{UV} and retrieves better poses.

Analysis of the compositing network. We compare three ablation settings. **w/o** $L_P^{G_2}$**:** removing perception loss $L_P^{G_2}$. **w/o Deform:** removing deformable skip-connection. **w/o Gated:** use normal conv layer instead of gated conv layer [58]. As shown in Fig. 9, removing any of the components will result in noticeable result degradation. Our full setting synthesizes more details in the foreground, smoother transition between the foreground and the background and also better fills the large background holes. The quantitative results reported in Table. 2 are consistent with the above observations.
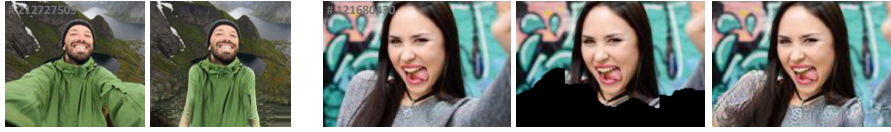
Fig. 10: Failure cases. Left: input, result. Right: input, foreground mask, result.

Fig. 11: Results of using inpaited background.

4.3   Limitations

Our approach has several limitations. First, as shown in Fig. 10 left, for challenging non-frontal selfie poses/viewpoints, our nearest pose search module might struggle with finding compatible neutral poses, which results in the synthesized result containing arms or shoulders that are too slim or wide compared to the head region. This problem happens less than 10% of the time in our top-1 result, and users can usually find a good compatible pose from our top-5 results. Second, this example also reveals the limitation of our background synthesis. We also show one example obtained by inpainting the background with an extra off-the-shelf model [59] in Fig. 11 to demonstrate the benefits from the image inpainting model trained on large-scale datasets. Finally, our system is prone to errors in DensePose detection. Fig. 10 right, DensePose failed to detect her left arm as foreground. Therefore the composition module retained her left arm in the result.

5   Conclusion

In this work, we introduce a novel "unselfie" task that translates a selfie into a portrait with a neutral pose. We collect an unpaired dataset and introduce a way to synthesize paired training data for self-supervised learning. We design a three-stage framework to first retrieve a target neutral pose to perform warping, then inpaint the body texture, and finally fill in the background holes and seamlessly compose the foreground into the background. Qualitative and quantitative evaluations demonstrate the superiority of our framework over other alternatives.

References

1. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. In: ICLR (2017)
3. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on graphics (TOG) **28**(3), 24 (2009)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: ICLR (2018)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
7. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
8. Darabi, S., Shechtman, E., Barnes, C., Goldman, D.B., Sen, P.: Image melding: Combining inconsistent images using patch-based synthesis. ACM Transactions on graphics (TOG) **31**(4), 82–1 (2012)
9. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. In: NeurIPS (2018)
10. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: CVPR (2018)
11. Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X., Luo, P.: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images (2019)
12. Good, P.: Permutation tests: a practical guide to resampling methods for testing hypotheses. Springer Science & Business Media (2000)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
14. Grigorev, A., Sevastopolsky, A., Vakhitov, A., Lempitsky, V.: Coordinate-based texture inpainting for pose-guided image generation. In: CVPR (2019)
15. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: ICCV (2019)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
18. Huang, J.B., Kang, S.B., Ahuja, N., Kopf, J.: Image completion using planar structure guidance. ACM Transactions on graphics (TOG) **33**(4), 129 (2014)
19. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
20. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
23. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model of people in clothing. In: ICCV (2017)

24. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: ECCV (2018)
25. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. IEEE transactions on pattern analysis and machine intelligence (TPAMI) **37**(12), 2402–2414 (2015)
26. Liang, X., Zhang, H., Lin, L., Xing, E.: Generative semantic manipulation with mask-contrasting gan. In: ECCV (2018)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
28. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS (2017)
29. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV (2019)
30. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016)
31. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation with semantic consistency. In: ICLR (2019)
32. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: CVPR (2018)
33. Ma, L., Xu, J., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: NIPS (2017)
34. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV (2017)
35. Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I.: Unsupervised attention-guided image-to-image translation. In: NeurIPS (2018)
36. Men, Y., Mao, Y., Jiang, Y., Ma, W.Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed gan. In: CVPR (2020)
37. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
38. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding (CVIU) **104**(2), 90–126 (2006)
39. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: ICCV Workshops (2019)
40. Neverova, N., Alp Guler, R., Kokkinos, I.: Dense pose transfer. In: ECCV (2018)
41. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
42. Pumarola, A., Agudo, A., Sanfeliu, A., Moreno-Noguer, F.: Unsupervised person image synthesis in arbitrary poses. In: CVPR (2018)
43. Qi, X., Chen, Q., Jia, J., Koltun, V.: Semi-parametric image synthesis. In: CVPR (2018)
44. Qian, S., Lin, K.Y., Wu, W., Liu, Y., Wang, Q., Shen, F., Qian, C., He, R.: Make a face: Towards arbitrary high fidelity face manipulation. In: ICCV (2019)
45. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
46. Raj, A., Sangkloy, P., Chang, H., Hays, J., Ceylan, D., Lu, J.: Swapnet: Image based garment transfer. In: ECCV (2018)
47. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NIPS (2016)

48. Reed, S.E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
49. Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: CVPR (2020)
50. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable gans for pose-based human image generation. In: CVPR (2018)
51. Song, S., Zhang, W., Liu, J., Mei, T.: Unsupervised person image generation with semantic parsing transformation. In: CVPR (2019)
52. Weng, S., Li, W., Li, D., Jin, H., Shi, B.: Misc: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. In: CVPR (2020)
53. Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: Geometry-aware unsupervised image-to-image translation. In: CVPR (2019)
54. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: CVPR (2017)
55. Xu, Z., Sun, J.: Image inpainting by patch propagation using patch sparsity. IEEE transactions on image processing (TIP) **19**(5), 1153–1165 (2010)
56. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: Image inpainting via deep feature rearrangement. In: ECCV (2018)
57. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: CVPR (2018)
58. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV (2019)
59. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. In: ECCV (2020)
60. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
62. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
63. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019)