VGFlow: Visibility guided Flow Network for Human Reposing

Rishabh Jain*
MDSR Adobe

Krishna Kumar Singh
Adobe Research

Mayur Hemani
MDSR Adobe

Jingwan Lu
Adobe Research

Mausoom Sarkar
MDSR Adobe

Duygu Ceylan
Adobe Research

Balaji Krishnamurthy
MDSR Adobe

Abstract

The task of human reposing involves generating a realistic image of a person standing in an arbitrary conceivable pose. There are multiple difficulties in generating perceptually accurate images, and existing methods suffer from limitations in preserving texture, maintaining pattern coherence, respecting cloth boundaries, handling occlusions, manipulating skin generation, etc. These difficulties are further exacerbated by the fact that the possible space of pose orientation for humans is large and variable, the nature of clothing items is highly non-rigid, and the diversity in body shape differs largely among the population. To alleviate these difficulties and synthesize perceptually accurate images, we propose VGFlow. Our model uses a visibility-guided flow module to disentangle the flow into visible and invisible parts of the target for simultaneous texture preservation and style manipulation. Furthermore, to tackle distinct body shapes and avoid network artifacts, we also incorporate a self-supervised patch-wise "realness" loss to improve the output. VGFlow achieves state-of-the-art results as observed qualitatively and quantitatively on different image quality metrics (SSIM, LPIPS, FID). Results can be downloaded from Project Webpage

1. Introduction

People are frequently featured in creative content like display advertisements and films. As a result, the ability to easily edit various aspects of humans in digital visual media is critical for rapidly producing such content. Changing the pose of humans in images, for example, enables several applications, such as automatically generating movies of people in action and e-commerce merchandising. This paper presents a new deep-learning-based framework for reposing humans guided by a target pose, resulting in high-quality and realistic output.

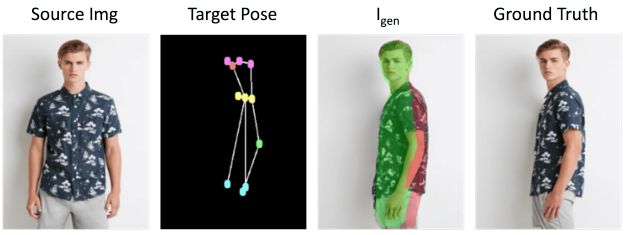

Figure 1. Human reposing involves changing the orientation of a source image to a desired target pose. To get accurate results, we learn to preserve the region visible (green) in the source image and transfer the appropriate style to the invisible region (red)

Recent approaches for human-image reposing based on deep-learning neural networks, such as [19,26,39], require a person image, their current pose, represented as a sequence of key-points or a 2D projection of a 3D body-pose map, and the target pose represented similarly. These methods fail to reproduce accurate clothing patterns, textures, or realistic reposed human images. This mainly happens when either the target pose differs significantly from the current (source) pose, there are heavy bodily occlusions, or the garments are to be warped in a non-rigid manner to the target pose. Many of these failures can be attributed to the inability of these networks to discern regions of the source image that would be visible in the target pose from those that would be invisible. This is an important signal to determine which output pixels must be reproduced from the input directly and which must be predicted from the context. We present VGFlow, a framework for human image reposing that employs a novel visibility-aware detail extraction mechanism to effectively use the visibility input for preserving details present in the input image.

VGFlow consists of two stages - encoding the changes

*rishabhj@adobe.com

in appearance and pose of the source image required to achieve the new pose and decoding the encoded input to the re-posed human image. The encoding stage includes a pose-based warping module that takes the source image and the source and target pose key-points as input and predicts two 2D displacement fields. One corresponds to the visible region of the source image in the target pose, and the other to the invisible areas. It also predicts a visibility mask indicating both visible and invisible regions in the source image, as they should appear in the target pose. The displacement fields, known as *appearance flows*, are used to sample pixels from the source image to produce two warped images. These warped images and the visibility masks are then encoded into the appearance features, a multi-scale feature pyramid. The encoding stage also tries to capture the relationship between the source and target poses by encoding their respective key-points together. The encoded pose key-points are translated into an image during the decoding stage, with the appearance features modulating the translation at each scale. This appearance-modulated pose to image decoding provides the final reposed output, which is then subjected to multiple perceptual and reconstruction losses during training.

The vast majority of existing methods [5, 26, 27, 39] are trained using paired source and target images. However, in terms of output realism, we observe various artifacts and a lack of generalization in these methods to unpaired inputs, especially when the source image differs significantly in body shape or size [20]. To that end, VGFlow is trained with a self-supervised patch-wise adversarial loss on unpaired images alongside the pairwise supervised loss to ensure a high level of realism in the final output. In summary, this paper proposes a new human reposing network VGFlow, based on:

- A novel visibility-aware appearance flow prediction module to disentangle visible and invisible regions of the person image in the target pose.
- An image decoder employing multi-scale texture modulated pose encoding.
- And, a patch-wise adversarial objective to improve the realism of the produced images leading to fewer output artifacts.

Our method achieves state-of-the-art on image quality metrics for the human reposing task. We present extensive qualitative and quantitative analysis with previous baselines, as well as ablation studies. Next, we discuss work related to the proposed method.

2. Related work

Human Reposing In recent years, several works have tried to generate a person in the desired target pose [1, 16, 23, 25, 32, 37, 39]. One of the initial work was PG^2 [21], which concatenated the source image with the target pose to generate the reposed output. Their work produced erroneous results due to the misalignment between the source and target image. Follow-up works tried to mitigate the problem by using a deformable skip connection [30] or progressive attention blocks [40] to achieve better alignment. However, modeling complex poses and bodily occlusions were still challenging. Recently, there have been some attention-based approaches [26, 39] proposed to learn the correspondence between the source and target pose. Although these attention [26] and style distribution-based [39] methods have shown impressive results, they need improvement for handling complex transformations. Apart from being computationally expensive, they do not preserve spatial smoothness and are inefficient in modeling fine-grained texture. In contrast, our approach is a flow-based method that can naturally handle complex and large transformations by using flow to warp the person in the source image to the target pose while preserving geometric integrity.

Flow-based methods Flow estimation aids in learning the correspondence between the content of two images or video frames and has been used in a variety of computer vision tasks such as optical flow estimations [33, 36], 3D scene flow [15], video-to-video translation [3], video inpainting [14], virtual-try-on [4, 9], object tracking [6] etc. Flow-based methods are also heavily explored for the human reposing task [1, 5, 19, 27] in which pixel-level flow estimates help to warp the texture details from the source image to the target pose. Still, as they don't incorporate visibility cues in their architecture, the network often relies more on in-painting rather than preserving the source image content. DIF [13] utilized a visibility mask to refine their flow estimation by splitting the appearance features into visible and invisible regions and applying a convolution on top. However, we show that visible and invisible information contain crucial complementary details and should be treated separately in the network pipeline. Most current networks are trained with paired poses, in which the target image is of the same person as the source image. The inference is usually performed for a different human, with the target pose derived from someone with a different body shape and size. This results in unusual distortions in the output. FDA-GAN [20] proposed a pose normalization network in which they added random noise to the target pose using SMPL [18] fitting to make reposing robust to noise. However, adding random noise would not lead to the imitation of real human pose distributions. We tackle this problem by introducing a self-supervised adversarial loss using unpaired poses during training.

3. Methodology

Our human reposing network requires three inputs, source image, source pose keypoints, and target pose key-

points. The task is to use the appearance and pose information from inputs to generate an image in the target pose. We subdivide this task into two key sub-parts. A detail extraction task, conditioned on the target pose, followed by a generation task where the extracted details are used to create a realistic image. Our detail extraction task is performed by a flow module $FlowVis$, which warps and aligns the texture and body shape information from the source image with the target pose. The outputs of $FlowVis$ are passed through our generator module, where texture and latent vectors from different resolutions are merged with the pose encoding using 2D style modulation. We also fine-tune the network using self supervised patch-wise realness loss to remove the generator artifacts. More details can be found in subsequent sections.

3.1. Flow & Visibility module

CNNs are well suited for image-to-image transformation tasks where there is a per-pixel mapping between input and output. An absence of such pixel-to-pixel mapping requires incorporation of other methods to aid learning. Therefore, many human reposing networks [1,5,19,27] usually employ some warping to deform source images to align them with output pose to get a better spatial correspondence between input and output. Warping an image requires a per pixel displacement flow field of size $2 \times H \times W$, where H and W are the height and width of the image respectively. This flow field can be learnable or obtained using a 3d model of humans such as SMPL [18] or UV maps [8]. The flow generated by existing works are not able to preserve the intricate geometry and patterns present on real clothes. These inadequacies get highlighted in the presence of complex poses and occlusions in the source image. To alleviate these issues, we propose a novel visibility aware flow module.

Our $FlowVis$ Module (Fig 2) takes in the source image I_s, source pose keypoints K_s, and target pose keypoints K_t as inputs and generates a visibility map $VisMap$, two flow field pyramids f_v^l, f_i^l(for visible and invisible regions respectively) and two warped image pyramids $I^l_{Vis}(I^l_v)$, $I^l_{Invis}(I^l_i)$ using the generated flow fields. The $VisMap$, f_v^l and f_i^l are generated by a Unet [28] like architecture FlowPredictor(FP). $VisMap$ segments the target image I_t into visible and invisible regions in the source image (Fig 2). The visible region(green in $VisMap$) corresponds to an area that is visible in I_s, and the invisible part (red) is the area that is occluded in I_s. The two separate per-pixel displacement flow-field pyramids f_v^l and f_i^l are predicted at different resolutions l. These flows are used to warp the source image to align with the target pose and generate the target's visible I_v^l and invisible I_i^l regions. The insight for predicting two flow fields comes from the observation that prediction for both the visible I_v^l and invisible I_i^l target regions may require pixels from the same location in the source. Therefore we need two flow fields to mitigate this issue because a single flow field can only map a source pixel to either one of the two regions. Flows from multiple resolutions are combined using the Gated aggregation [4] (GA) technique which filters flow values from different radial neighborhoods to generate a composite flow. This allows the network to look at multiple scales sequentially and refine them at each step. To construct the flow at the final 256×256 level, we upsample the flow from the previous layer using convex upsampling. Convex upsampling [34] is a generalization of bilinear upsampling where the upsampling weights for a pixel are learnable and conditioned on the neighborhood features. Convex upsampling aids in the preservation of fine-grained details and sharpens the warped output. Moreover, employing a single decoder to generate both flow pyramids helps preserve consistency and coherence between visible and invisible warped images. The module is summarised in the following equations.

$$\begin{aligned} f_v^l, f_i^l, VisMap &\leftarrow FP(I_s, K_s, K_t) \\ f_v^{agg}, f_i^{agg} &\leftarrow GA(f_v^l, f_i^l) \\ f_v^o, f_i^o &\leftarrow ConvexUpsample(f_v^{agg}, f_i^{agg}) \\ I_v^l, I_i^l &\leftarrow Warp(I_s, f_v^l), Warp(I_s, f_i^l) \end{aligned} \tag{1}$$

Losses We train the flow module seperately before passing the output to the generator. $FlowVis$ is trained to generate I_v and I_i by minimizing the losses on the visible and invisible regions of the human model respectively. The $VisMap$ can be broken down to visible area mask m_v and invisible area mask m_i by comparing the per-pixel class. For the visible region, we can find an exact correspondence between predicted and the target image and hence we utilize L1 and perceptual loss L_{vgg} [31] on the masked visible part. The loss on the visible area minimizes texture distortion and loss in detail(L1). For the invisible region, there is no exact correspondence between I_i and I_t but we can still optimize using the target image style. For example, in-case a person needs to be reposed from a front pose to back pose then the entire body will be invisible in I_s. However, there is a very strong style similarity that we can leverage for reposing. Hence, we use perceptual [31] and style loss L_{sty} [7] to capture the resemblance for these regions on the masked invisible regions. We also minimize tv norm [35] on the flow pyramids to ensure spatial smoothness of flow and the losses are computed for the entire flow pyramid. $Vismap$ is optimized by imposing cateogrical cross entropy loss, L_{cce}. The ground truth for the visibility map $Vismap_{gt}$, was obtained by fitting densepose [8] on I_s, I_t and then matching the acquired UV coordinates to generate the visible and invisible mask. We further use teacher forcing technique [24] for training $FlowVis$, in which the ground truth VisMap is used with 50% probability for the warping losses. The losses guiding the flow module can be summarized as follows. Here $\odot$ indicates per pixel multiplication.

$$\begin{aligned} m_v, m_i &\leftarrow VisMap \\ L_{wrp} = \sum_l & L_{vis}(I_v^l, m_v, f_l) + L_{invis}(I_i^l, m_i, f_l) \\ & + L_{cce}(VisMap, VisMap_{gt}) \end{aligned} \tag{2}$$

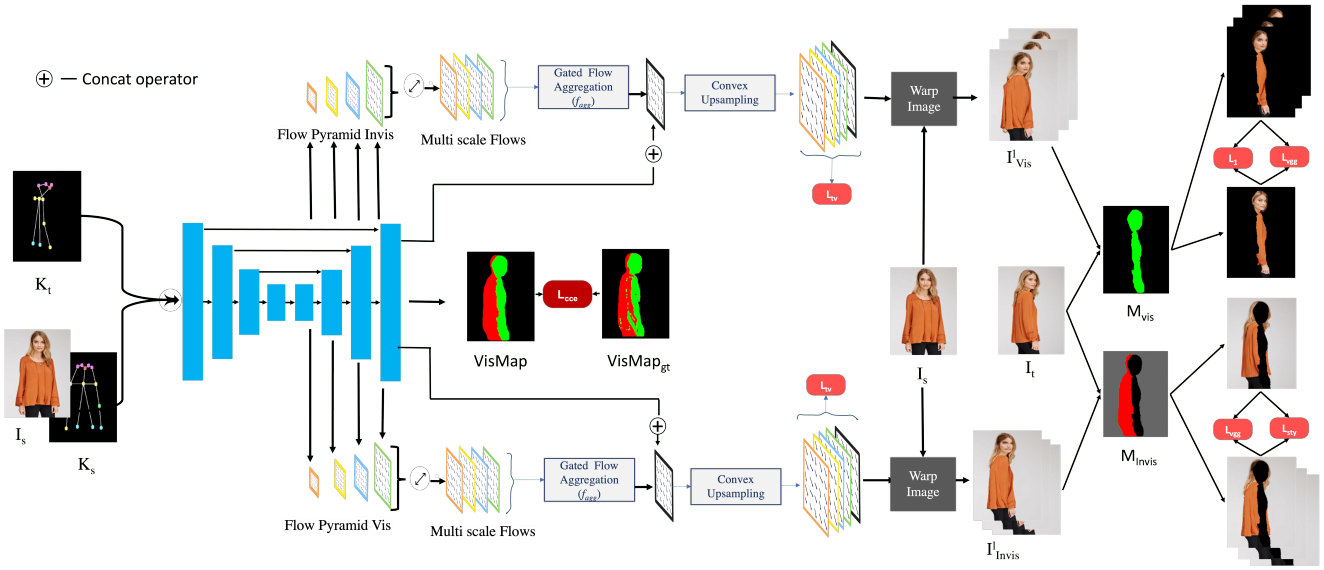

Figure 2. Our *FlowVis* module takes in concatenated I_s, K_s & K_t as input. We predict two flow pyramids at multiple scales for different visibility regions using a Unet architecture. Subsequently, these flow pyramids are combined using *Gated Aggregation* [4] and upsampled via *Convex upsampling* [34]. Losses L_1, L_{vgg} are imposed on the visible areas and L_{vgg}, L_{sty} are imposed on the invisible areas.

where,

$$\begin{aligned} L_{vis}(I,m,f) &= \beta_1 L_{vgg}(I \odot m, I_t \odot m) + \\ &\quad \beta_2 \|I \odot m, I_t \odot m\|_1 + \beta_3 L_{tv}(f) \\ L_{invis}(I,m,f) &= \beta_1 L_{vgg}(I \odot m, I_t \odot m) \\ &\quad + \beta_4 L_{sty}(I \odot m, I_t \odot m) + \beta_3 L_{tv}(f) \end{aligned} \tag{3}$$

3.2. Generator module

The generator module (Fig 3) takes as input the source and target poses and the output of $FlowVis$ module. It is important to point out that only the final level of the transformed image pyramid I^l_v, I^l_i, i.e. the level at resolution 256×256 referred now as I_v, I_i, is used in our generator.

Pose encoder Majority of the previous networks [5, 27, 39] encoded the pose information only as a function of 18 channel target keypoints. However, single view reposing is fundamentally an ill defined task. The network has to hallucinate the body and clothing region whenever it is invisible in the source image. Hence, it should be able to distinguish between the portions of the target image for which it can obtain corresponding texture from the source image and those for which it must inpaint the design of the respective clothing item. Therefore, to model the correlation between the source(K_s) and target poses(K_t), we pass both source and target keypoints to a Resnet architecture $PoseEnc$ to obtain a 16×16 resolution pose feature volume.

$$e_p = PoseEnc(K_s, K_t) \tag{4}$$

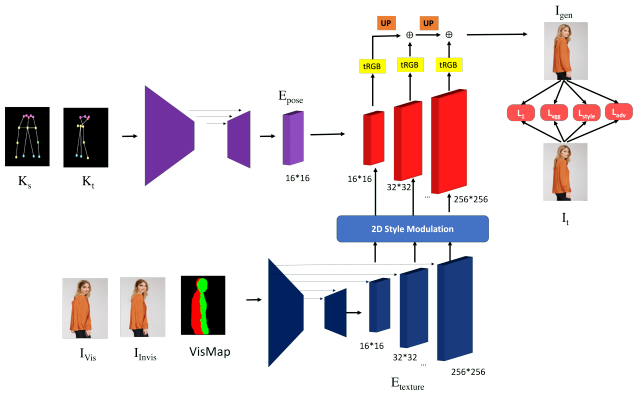

Figure 3. Our *Generator* module consumes the FlowVis outputs to generate the final reposed output. It utilizes 2D style modulation [1] to inject Multi-scale Appearance features into the pose encoding for the generation process

Texture injection & Image Generation The texture encoder takes in I_v, I_i and $Vismap$ as input and uses a ResNet architecture similar to pose encoder to obtain texture encodings at different hierarchical scales. The low resolution layers are useful for capturing the semantics of the clothing items, identity of the person and the style of individual garments while the high resolution layers are useful for encapsulating the fine grained details in the source image. We also add skip connections in our texture encoder which helps in combining low and high resolution features and

capture different levels of semantics.

The image decoder module takes in the pose encoding e_p as input and up samples them to higher resolutions. The texture is injected into these pose encodings at different scales by using 2D style modulation [1]. After the modulation, features are normalized such that they have zero mean and unit standard deviation. Similar to [1] which is based on StyleGAN2 architecture, RGB images are predicted at multiple resolutions and sequentially lower resolution image is added to next higher resolution image after upsampling it to obtain the final output image (Fig 3). As the network has to fill in the invisible region (I_i) by generating new content similar to neighbourhood pixels at inference, we perform an auxiliary task of inpainting 20% of the training time, similar to [5]. Here, a random mask is applied on the target image and given as input to the model. The generator is then asked to output the complete target image. This teaches our network to complete any missing information of warped images in a visually convincing manner. VGFlow achieves SOTA in the single view human reposing task.

$$\begin{aligned} e^l_{tex} &= TexEnc(I_{vis}, I_{Invis}, VisMap) \\ I_{gen} &= ResNetDec(e_p, 2DStyleMod(e^l_{tex})) \end{aligned} \tag{5}$$

Losses We enforce L1, VGG, Style and LSGAN [22] losses between I_{gen} and I_t. L1 loss helps to preserve the identity, body pose and cloth texture with pixel level correspondence. Vgg and Style loss are useful in conserving high level semantic information of garments present on source human and bringing the generated image perceptually closer to target. For the LSGAN loss, we pass target pose(K_t) along with generated image(I_{gen}) to the discriminator for better pose alignment. Adverserial loss assist in removing small artifacts and making the image more sharper. We utilize LSGAN loss as it has been shown to produce better output than traditional GAN loss [22] and is more stable during training. Overall, the loss function can be defined as:

$$\begin{aligned} L_{sup} = \alpha_{rec}\|I_{gen}, I_t\|_1 + \alpha_{per}L_{vgg}(I_{gen}, I_t) \\ + \alpha_{sty}L_{sty}(I_{gen}, I_t) + \alpha_{adv}L_{LSGAN}(I_{gen}, I_t, K_t) \end{aligned} \tag{6}$$

The α's are the weight for the different losses. L_{sup} refers to the supervised loss used when the input and output are of same person in the same attire.

Self supervised *realness* Loss Even though the network is able to produce perceptually convincing results with the above supervised training, there are still bleeding artifacts present that occur when there is a complicated target pose or occluded body regions. There is also a discontinuity present between the clothing and skin interface at some places. Moreover, during training, the target pose is taken from paired image where the input and output are for the same person wearing the same apparel. This would introduce a bias in the network as it would not be robust to alterations in body types(e.g. fat, thin, tall) between K_s, K_t [20]. During inference this could deteriorate results when, for example, the body shape of the person from which the target pose is extracted vary significantly from the source image human body shape. To alleviate these issues, we fine tune our network with an additional patch wise adversarial loss [11] whose task is to identify if a particular patch is real or not. Therefore during fine-tuning, we choose unpaired images with 50% probability(Sec 4) from a single batch. Only an adversarial loss is applied on the unpaired images and L_{sup} loss is present on the paired images(Fig 4).

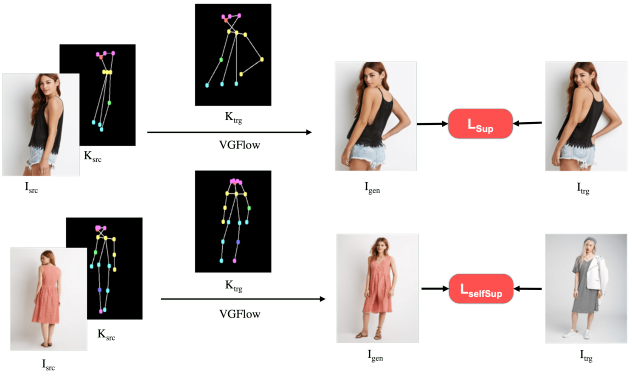

Figure 4. Addition of patch-wise Self Supervised loss(L_{SS}) helps in enhancing the image quality and increases *realness*

$$L_{SS} = L_{PatchGAN}(I_{gen}, I_t) \tag{7}$$

All these losses i.e. L_{wrp}, L_{sup} and L_{SS} are used to fine-tune the networks in an end-to-end fashion.

4. Experiments

Dataset We perform experiments on the In-shop clothes benchmark of Deepfashion [17] dataset. The dataset consists of 52,712 pairs of high resolution clean images with 200,000 cross-pose/scale pairs. The images are divided into 48,674 training and 4038 testing images and resized to a resolution of 256×256. The keypoints of all images are extracted using OpenPose [2] framework. We utilize the standard 101,966 training and 8,570 testing pairs, following previous works [1, 5, 26]. Each pair consists of source and target image of the same person standing in distinct pose. It is also worth noting that the identity of training and testing pairs are separate from one another in the split.

Evaluation metrics We compute structural similarity index(SSIM), Learned Perceptual Image Patch Similarity (LPIPS) and Fretchet Inception Distance(FID) for compar-

ison with previous baselines. SSIM quantifies image degradation by employing luminance, contrast and and structure present in the image [29]. LPIPS [38] compute patchwise similarity using features from a deep learning model. It has been shown to correlate well with human perception of image similarity. FID [10] works by comparing the 2-wasserstein distance between the InceptionNet statistics of the generated and ground truth dataset. This provides a good metric for estimating the *realness* of our generated results.

Implementation details All the experiments were carried out using pytorch framework on 8 A100 gpus. We first train our reposing network using only supervised losses for 30 epochs with Adam [12] optimizer, batch size of 32 and learning rate of 1e-4. Afterwards, we fine tune the model using self supervised loss. For the self supervised training of our model, we randomly choose a target sample or pose image from the complete training dataset during training. We also choose the patch size of 16×16 for the discriminator which provides a good tradeoff between high level and low level feature preservation. We found that imposing the self supervised loss works better when we impose it intra-batch. So, supervised and self supervised losses were propagated together in the same mini-batch and we used loss masking to stop gradients for the respective branches. In contrast, the inpainitng auxilliary task was carried out among different mini-batches. The intra-batch strategy provides an anchoring to the network with supervised losses in a single backward pass as we don't have a pairwise supervision for a random target pose. Without intra-batch training, the network starts hallucinating details which are not consistent with the source image but looks plausible. The fine-tuning was carried out with a batch size of 32 and a reduced learning rate of 5e-5. We highlight additional advantages of using self supervised learning technique in sec 6.

5. Results

We compare our method with several strong previous works [1, 5, 19, 26, 27, 39]. Among these, Gfla [27], Spgnet [19], Dior [5] and PWS [1] leverage flow-based warping in their network to align source image with the target pose. The flow obtained by them is used to warp the features obtained from the source image and move them to the k_t orientation. PWS [1] takes additional information in the form of UV maps of the target image as input(more discussion in supplementary). Note that the densepose UV maps of the target image contain a lot more information than simple keypoints. PWS [1] exploits the UV map and symmetry to directly calculate the flow and inpaint the occluded part using a coordinate completion model. However, in the case of unpaired images at inference, obtaining the ground truth UV maps is a difficult task, and the body shape problem is worsened further. Moreover, UV map contains information about the per-pixel mapping of the body regions between I_s & I_t. Such extensive information won't be accurate if we use a UV map from an unpaired pose and hence, keypoints-based methods are not directly comparable with UV-based methods. We include qualitative results from [1] for completeness and to highlight the issues in UV based warping in Sec.6. In contrast to flow-based techniques, NTED [26] and CASD [39] utilize semantic neural textures and cross-attention-based style distribution respectively to diffuse the texture of the source image onto the target pose. However, disentangling the texture and reconstructing it in a different orientation is a challenging task without warping assistance, especially for intricate patterns and text. We further corroborate this phenomenon in our qualitative analysis.

Method	SSIM ↑	FID ↓	LPIPS ↓
Intr-Flow [13]	-	16.31	0.213
GFLA [27]	0.713	10.57	0.234
ADGAN [23]	0.672	14.45	0.228
SPGNet [19]	0.677	12.24	0.210
Dior [5]	0.725	13.10	0.229
CASD [39]	0.724	11.37	0.193
Ours	**0.726**	**9.29**	**0.185**

Table 1. Our network outperforms all the previous baselines for quantitative image metrics at 256×256 resolution

Quantitative comparison for human reposing We compare our method with previous works which have generated images with 256×256 resolution as their output size. As can be seen from Table 1, VGFlow performs significantly better in SSIM, LPIPS and FID compared to other baselines. The improvement in LPIPS and SSIM metric, which compare images pairwise, can be attributed to better modelling of the visible regions of the model. On the other hand, FID, which measures the statistical distance between the latent space distribution of real and generated images, becomes better due to our superior style transfer for the invisible region.

Qualitative comparison for human reposing We highlight improvements in human reposing from several qualitative dimensions in Fig 5. In (a), VGFlow generates images with accurate pattern reproduction and seamless intersection of top and bottom designs. Other works were either not able to maintain coherence during generation or produced unnatural artifacts in their final outputs. In (b), our network was able to reproduce the back ribbon at the correct target location along with the strap. Whereas, PWS [1] is not able to model the ribbon as the UV map is a dense representation of only the human body and does not capture the details of loose clothes properly. (c) and (d) highlight texture preser-

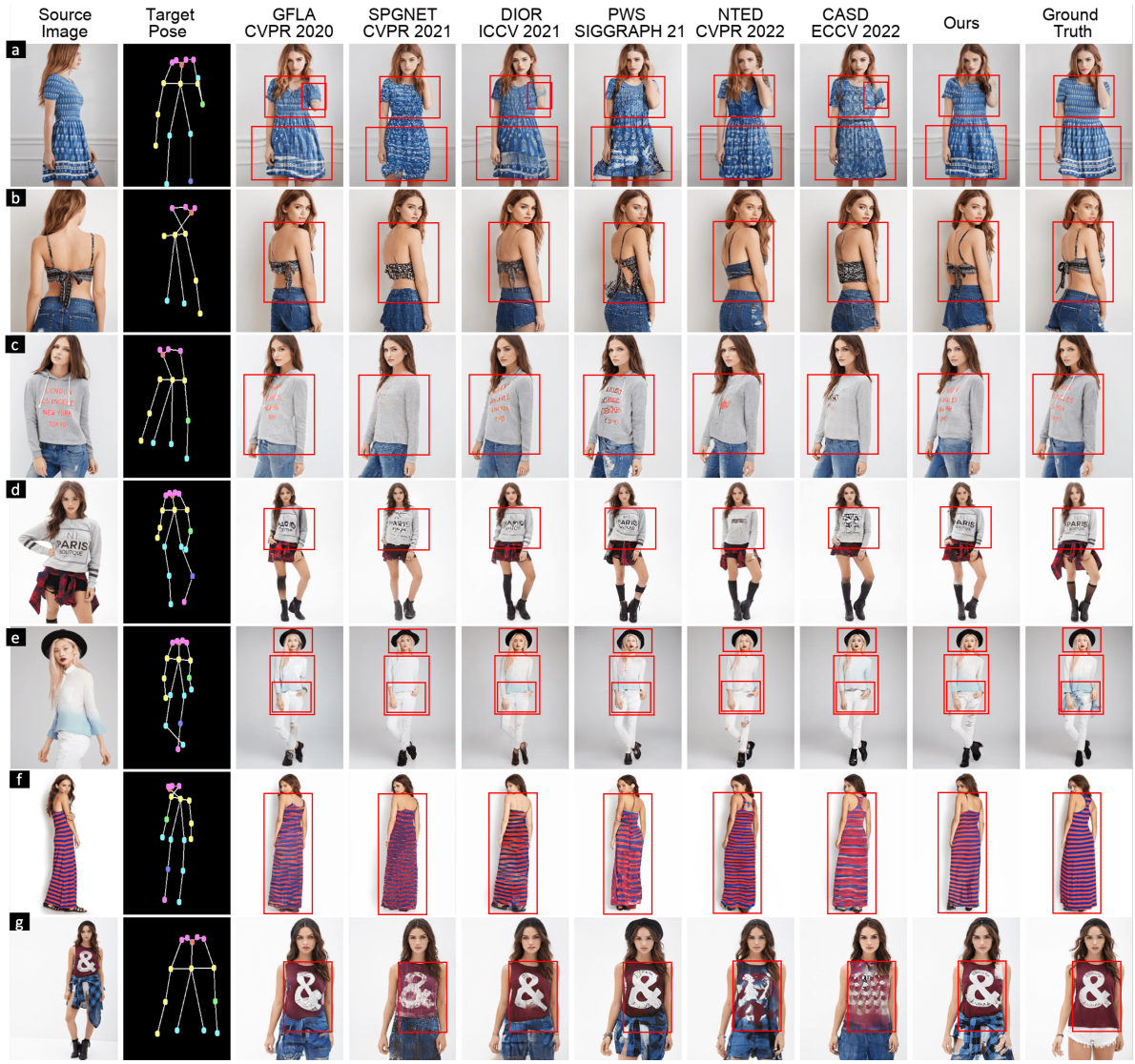

Figure 5. In this figure, we underscore improvements along different qualitative aspects over previous works. We emphasize enhancement in preserving pattern & segmentation (a), handling complex poses (b), design consistency (c), conserving text readability (d), reducing bleeding, skin generation & identity reproduction (e), maintaining geometric integrity (f) and pattern coherence (g). (Best viewed in zoom)

vation while reposing where the text maintains its relative orientation during warping. The words in the original pose are only intelligible for VGFlow. In (e), the blue color is bleached onto the white shirt, and face reproduction is not accurate for multiple other reposed outputs. (f) shows the faithful geometric construction of parallel lines and (g) emphasizes pattern coherence while global deformation of texture. We see that only the warping-based methods were able to preserve the rough texture in (g) while NTED [26] and CASD [39] completely dismantled the structural integrity of the"&" symbol on the t-shirt. Additional results can be found in supplementary material.

6. Ablation and Analysis

Warping functions Flow estimation is integral to many reposing pipelines [1, 5, 27], including ours. Therefore we perform an analysis of the warping capabilities of different flow modules. The images in Fig 6 show the quality of warping of the source image based on the flow predicted by the respective methods. As flow warping moves pixels, it can only hallucinate within the bound of the source image's content. We see that VGFlow performs significantly better than previous flow-based warping techniques in preserving the semantic layout of the source human. The fine-grained texture details are transferred seamlessly along with

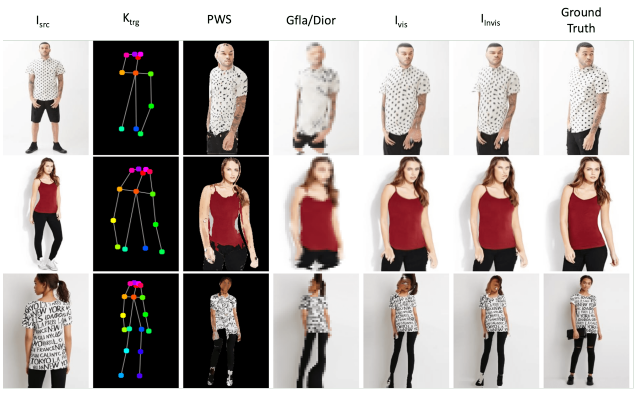

Figure 6. Qualitative comparison between different warping functions for PWS [1], Gfla/Dior [5, 27] and our (I_{vis}, I_{Invis}) warp. Gfla and Dior used the same flow prediction module

the overall body shape. PWS is limited due to mistakes in estimating UV maps by off-the-shelf components [8], and GFLA [27] had a sub-optimal flow estimation module. By letting our network predict different flow fields for visible and invisible regions while simultaneously constraining it to use the same latent features, we were able to achieve a good consistency between I_v and I_i.

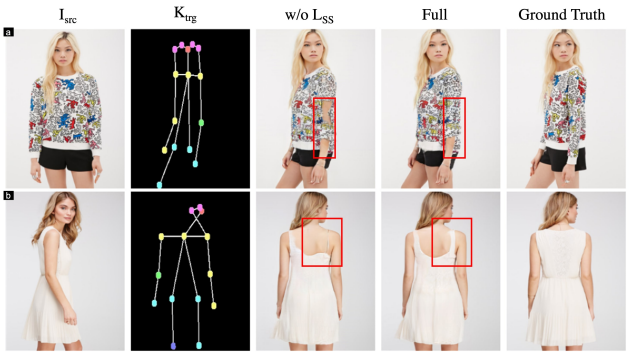

Figure 7. Introducing PatchWise Self-Supervised Loss alleviates unnatural segmentation (a) and increases *realness* (b)

Ablation study To gauge the effectiveness of different components of our pipeline, we perform various ablations of our network in Table 2. Note that the Self Supervised loss is excluded in all the ablations. The result of ablations are as follow:-

- **Zflow I_{warp}** We replace our flow network with Zflow [4] architecture, which includes a flow pyramid with GA(Gated Aggregation) to predict a single 2D displacement flow field and finetune our VGFlow generator. This model produced sub-optimal flow and lacked the ability to preserve high-frequency components like text & design. The quantiative metrics significantly deteriorate against VGFlow(Tab 2) and we show the qualitative comparison in supplementary.
- **w/o VisMap, I_i, L_{SS}** The result of removing the Vismap and I_i from the inputs of texture encoder and only passing I_v indicates that the visibility map plays an integral role in capturing the appropriate relationship for the texture encodings.
- **w/o I_i, L_{SS}** The degradation of quantitative image metrics on the removal of only the I_i from the input of texture encoder shows that even though I_v and I_i produce similarly warped images, they do provide crucial complementary information.
- **w/o k_s, L_{SS}** The deterioration in FID score(9.70 $\rightarrow$ 9.90) on removing the K_s input from the pose encoding indicates that passing in K_s helps in modeling the correlation between source and target pose.
- **w/o L_{SS}** We also study the effect of finetuning with self-supervised loss L_{SS}. L_{SS} plays a major role in improving the FID(9.70$\rightarrow$ 9.29) and marginally improving SSIM(0.725$\rightarrow$0.726) and LPIPS(0.186$\rightarrow$0.185). We also show qualitative improvements of integrating L_{SS} in Fig 7
- **Full** This model contains all the components presented in the paper. These ablations show that our configuration of VGFlow produces the best output.

Method	SSIM ↑	FID ↓	LPIPS ↓
ZFlow I_{warp}	0.719	11.70	0.205
w/o VisMap, I_i, L_{SS}	0.719	9.89	0.196
w/o I_i, L_{SS}	0.724	9.93	0.190
w/o K_s, L_{SS}	0.726	9.90	0.186
w/o L_{SS}	0.725	9.70	0.186
Full	**0.726**	**9.29**	**0.185**

Table 2. We perform extensive ablations to gauge the importance of each component in our network

We also include some failure cases due to warping errors and target segmentation faults in supplementary material.

7. Conclusion

We propose VGFlow, a visibility-guided flow estimation network for human reposing that generates the reposed output by leveraging flow corresponding to different visibility regions of the human body. We propose an additional self-Supervised Patchwise GAN loss to reduce artifacts and improve the network's ability to adapt to various body shapes. VGFlow achieves SOTA in the pose-guided person image generation task, and we demonstrate the significance of our contributions through extensive ablation studies.

References

[1] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 2, 3, 4, 5, 6, 7, 8

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 5

[3] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. Mocycle-gan: Unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 647–655, 2019. 2

[4] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5433–5442, 2021. 2, 3, 4, 8

[5] 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, October 2021. 2, 3, 4, 5, 6, 7, 8

[6] Long Fan, Tao Zhang, and Wenli Du. Optical-flow-based framework to boost video object detection performance with object enhancement. *Expert Systems with Applications*, 170:114544, 2021. 2

[7] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 3

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3, 8

[9] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 6

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR, 2017*, 2017. 5

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[13] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 2, 6

[14] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17562–17571, 2022. 2

[15] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[16] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 2

[17] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[19] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10806–10815, 2021. 1, 2, 3, 6

[20] Liyuan Ma, Kejie Huang, Dongxu Wei, Zhao-Yan Ming, and Haibin Shen. Fda-gan: Flow-based dual attention gan for human pose transfer. *IEEE Transactions on Multimedia*, 2021. 2, 5

[21] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[22] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 5

[23] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020. 2, 6

[24] Mahardhika Pratama, Choiru Za'in, Edwin Lughofer, Eric Pardede, and Dwi AP Rahayu. Scalable teacher forcing network for semi-supervised large scale data streams. *Information Sciences*, 576:407–431, 2021. 3

[25] 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

[26] 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. 1, 2, 5, 6, 7

[27] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 2, 3, 4, 6, 7, 8

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 3

[29] De Rosal Igantius Moses Setiadi. Psnr vs ssim: imperceptibility quality assessment for image steganography. *Multimedia Tools and Applications*, 80(6):8423–8444, 2021. 6

[30] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[32] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2357–2366, 2019. 2

[33] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV 2020*, 2020. 3, 4

[35] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. An improved algorithm for tv-l 1 optical flow. In *Statistical and geometrical approaches to visual motion analysis*, pages 23–45. Springer, 2009. 3

[36] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 2

[37] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[39] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision*, pages 161–178. Springer, 2022. 1, 2, 4, 6, 7

[40] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 2