# UMFuse: Unified Multi View Fusion for Human Editing applications

Rishabh Jain*
MDSR Adobe

Mayur Hemani
MDSR Adobe

Duygu Ceylan
Adobe Research

Krishna Kumar Singh
Adobe Research

Jingwan Lu
Adobe Research

Mausoom Sarkar
MDSR Adobe

Balaji Krishnamurthy
MDSR Adobe

## Abstract

*Numerous pose-guided human editing methods have been explored by the vision community due to their extensive practical applications. However, most of these methods still use an image-to-image formulation in which a single image is given as input to produce an edited image as output. This objective becomes ill-defined in cases when the target pose differs significantly from the input pose. Existing methods then resort to in-painting or style transfer to handle occlusions and preserve content. In this paper, we explore the utilization of multiple views to minimize the issue of missing information and generate an accurate representation of the underlying human model. To fuse knowledge from multiple viewpoints, we design a multi-view fusion network that takes the pose key points and texture from multiple source images and generates an explainable per-pixel appearance retrieval map. Thereafter, the encodings from a separate network (trained on a single-view human reposing task) are merged in the latent space. This enables us to generate accurate, precise, and visually coherent images for different editing tasks. We show the application of our network on two newly proposed tasks - Multi-view human reposing and Mix&Match Human Image generation. Additionally, we study the limitations of single-view editing and scenarios in which multi-view provides a better alternative. Datasplits and results can be found at [Project Webpage](#).*

## 1. Introduction

Automating person-image editing can be of great value to business applications for advertisements, commercial merchandising, as well as individual creativity. Pose-based

---

*\*rishabhj@adobe.com

image editing is a class of human image editing workflows where the end result depends on an input human pose. These include human reposing, garment virtual try-on [10, 40, 4] etc. Recent methods for person-image editing [28, 18, 33, 6, 3, 23, 34] have made great strides in producing accurate edit results from a single image of a person. However, they suffer from textural, shape, and color artifacts because a single image may not have sufficient information to produce an accurate rendition of the person in the desired target pose. In this work, we present a novel pose-guided human image generation method that can incorporate multiple images as well as leverage existing single-view reposing models to improve the editing results significantly.

In most contemporary methods for pose-guided human image reposing [28, 18, 33, 6, 3, 23], a deep neural network transforms a *source image* of a person into a *target pose* specified as a sequence of keypoints. The network takes a single source image of a person, the pose of the person in that image (*source pose*), and the *target pose* as input, and produces an image corresponding to the person in the target pose. The source image supplies the color and texture information for guiding the generation of the target pose image. However, the target pose can be significantly different from the source pose, such as when a part of the human body or clothing obscured in the original image might appear in the target pose. In such cases, the network must infer those regions of the output from the context. We posit that using multiple source images of the same person is an effective solution to this problem. This is also practical for real-world applications as companies usually have multiple photos of a human model in distinct poses. However, this approach requires mapping parts of the target image to parts of the source images based on how their textures match up geometrically, which is a non-trivial task. In this paper, we present a framework that can extend existing single-view networks for multi-view reposing task by effectively com-

bining data from multiple source images. We make the following contributions to the field of pose-guided person image editing:

- UMFuse, a novel plug-and-play framework to fuse multi-scale pose and appearance features from different source images for human image editing applications, using:
    - an appearance retrieval map for *interpretable* feature fusion predicted using the input source images and the corresponding pose information.
    - a visibility-informed pre-training task for initialization of the fusion module.
- The task of multi-view human reposing (MVHR, sec. 5), and a benchmark datasplit based on the DeepFashion [21] dataset for evaluating the quality of output.
    - We demonstrate its compatibility with two different single-view reposing networks [29, 14].
    - We also showcase its versatility by combining different fashion components from completely different persons (Mix-and-Match, sec. 6).
- Detailed quantitative and qualitative analysis, comparisons, and ablation studies indicating the effectiveness of the proposed method.

The next section discusses how the proposed method relates to existing methods for human-image editing.

## 2. Related work

**Pose-guided human image generation**   The single view reposing problem has been extensively studied in [33, 27, 18, 26, 29] .  These methods use a single source image as input to produce a reposed person image. They follow a two-stage strategy – warping the source image using a flow field to align with the target pose, and synthesizing the reposed output using a generator.  Variations of these methods include leveraging target segmentation mask[22], visibility maps[18], UV-maps[2], attention-based style distribution[43] etc. for the generation process. While these methods can generate realistic output, they are adversely affected by factors such as heavy occlusions and poor camera angles. We model different body parts in the generated output using multiple source photos where these parts are either clearly visible or have a better style transfer compatibility with the source pose, thereby alleviating the problems faced by single-view methods.

**NeRFs and Motion Transfer**   Recent work in neural radiance field (NeRF) based models [36, 37, 15, 25, 41, 7, 19] can reconstruct an entire 3D scene with humans using a relatively larger set of images from one or more videos. Some of them can be adapted to generate novel poses as

demonstrated by [15, 36]. Few-shot human motion transfer task[13, 35] also aims to transfer the motion of one person to another via continual frame synthesis.  However, apart from requiring a vast number of frames and expensive training, these methods fail to preserve fine-grained clothing texture and produce a coarse body shape output. Also, most of them can reproduce novel poses for only the person they are trained on.  Moreover, in contrast to single-view reposing networks, the NeRF-based models have a limited ability to predict the missing content. Our work bridges these two approaches and combines the advantages of using multiple source images with the warping and inpainting capabilities of single-view reposing architectures.

The next section details how UMFuse produces high-quality, pose-guided human images from multiple sources.

## 3. Methodology

We first describe single-view, pose-guided human image generation (PHIG) as a generic task, and motivate the need for utilizing multiple views using human reposing as an example.  Next, we describe the transformation of the single source PHIG network to a multi-view network utilizing multiple source views with UMFuse.

**Single-view Human Image Reposing:**   We begin by describing VGFlow[14], a recent human reposing network. It broadly consist of three stages (Fig 1, A) :

1. Warping the source image $I_s$ to match the target pose $P_t$ and encoding its texture.
2. Encoding the source and target pose key points $P_s$ and $P_t$.
3. Generation of the final reposed output $I_p$ from both pose and texture encoding.

The network takes $I_s$, $P_s$, and $P_t$ as input, and predicts two flow fields - $F_v$ and $F_i$. These flow fields are per-pixel 2D displacement fields to warp the source image. Here, $F_v$ represents displacements for regions of the source image that would remain visible when the pose changes to the target. $F_i$ are for the displacements that produce pixels that are invisible in the source pose but may be predicted using context from the source image. These flow fields are used to sample two warped ($I_w$) renditions of the source image – $I_w^v$ and $I_w^i$ – corresponding to the visible and invisible regions in the target pose respectively. Additionally, the network also produces a visibility segmentation map $V_t$ of the source image as it would appear in the target pose, indicating the regions that are visible (and invisible) in the source image. Intuitively, the visibility map ($V_t$) is an indicator of whether the network can obtain the correspondence for a pixel from the source image ($I_s$) directly, or it must make an informed prediction. The encoder stage of the network consists of two modules - the *texture-encoder* ($E_t$), and the *pose-encoder* ($E_p$). The texture encoder captures the color,
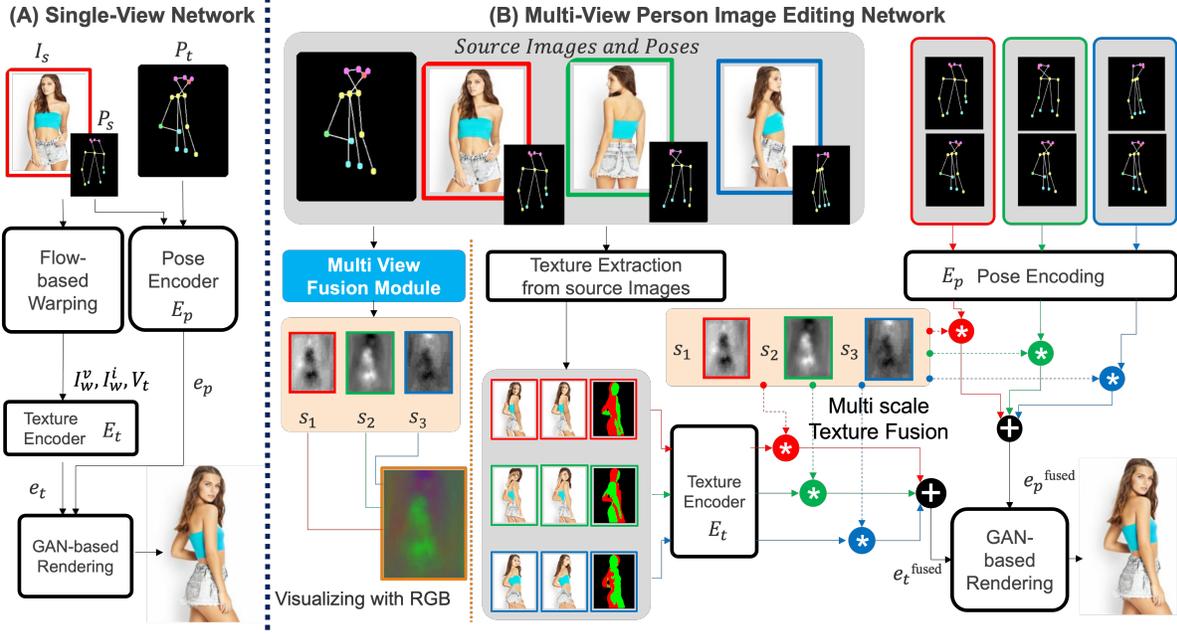
Figure 1. UMFuse framework: (A) is the single view PHIG network on top of which the UMFuse framework operates. The source image and its keypoints along with the target keypoints are used to produce warped images and a visibility map $(I_w^v, I_w^i, V_t)$. These are used to obtain the texture encoding at different scales $l$ $(e_{t,l})$, and the source and target poses are used to obtain the pose encoding $e_p$. Together they are used to render the output with a GAN-based renderer. (B) Shows the UMFuse adaptation for the single view network. The multiple source images and poses are passed individually to the warping and visibility prediction module to obtain multiple warped images and visibility maps, and from those, multiple texture-encoding vectors. Likewise, the source poses paired with the target pose are used to obtain multiple pose encoding vectors. These are merged in an affine combination using the predicted Appearance Retrieval Maps($s_{1-3}$). These maps are obtained using the Multi-view Fusion module and are a key contribution of the UMFuse framework.
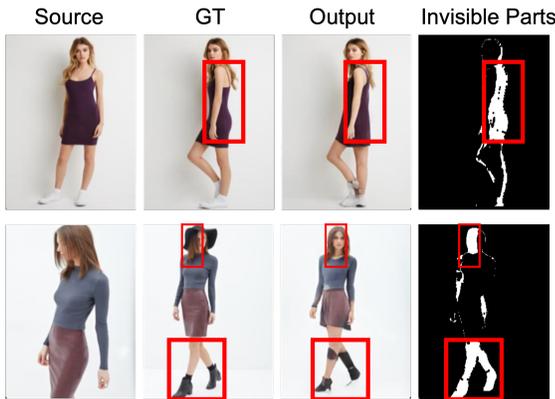


Figure 2. Examples indicating a high degree of overlap between the invisible regions in the visibility maps and the output errors.

shape, and style information for the source image $I_s$ at multiple scales($l$ $layers$) as it would appear in the target pose by combining information from $I_w^v$, $I_w^i$, and $V_t$. The pose encoder produces a latent representation for the source and target key points $(P_s, P_t)$ that can be used to condition the generation of a person-image to make it conform to the target pose.

$$e_{t,l} = E_t(I_s, P_s, P_t),$$
$$e_p = E_p(P_s, P_t) \tag{1}$$

Note that UMFuse is agnostic of the exact architecture used by VGFlow. Other networks like those described in [29, 2, 5] that have a similar broad architecture are also compatible. We share results for both VGFlow and GFLA [29], in sec. 5.

The decoder stage ($D$) transforms the encoded pose and texture information into an RGB output image $I_p$. One branch of the decoder processes the encoded pose information, while the activations of each layer of the second branch modulate the output of the corresponding layer of the first branch using style modulation[16] from texture encodings.

We take VGFlow[14] as our primary single-source PHIG network due to its superior performance. Next, we discuss the need for employing multiple inputs for the PHIG task.

**Is a single-source input adequate?** Fig 5 illustrates the limitations of using a single source input image for the human reposing task. We argue that if the target pose is significantly different from the source pose, a PHIG network

cannot accurately predict the pixel colors for those pixels that are in the invisible regions. This is expected in extreme cases where one pose is the front, and the other is the back. But even for moderate pose changes, we find that error regions for generated images overlap with the invisible areas in the corresponding ground-truth visibility maps (see Fig 2). Appearance artifacts can also be caused by non-rigid clothing deformation between poses. Even if the source and target poses are similar in terms of the relative arrangements of the key points, other factors such as the camera position and viewing angle contribute to the errors. Using multiple source images can help improve the output quality if we can learn to map regions in the source views which have the best appearance compatibility with the corresponding target pose regions. Table 1 shows that UMFuse is able to learn this mapping and produces better results by effectively fusing information from up to three source views. We discuss this fusion architecture next.

**Combining multiple inputs using an appearance retrieval map (ARMap):** UMFuse employs a feature blending approach for combining multiple source input images. The process is illustrated in Fig 1 (B). Our *Multi-View Fusion(MVF)* module predicts a 2D appearance retrieval map $s_{1:k} \in [0,1]^{H \times W}$, for each of the $k$ source images. The input to this module is a channel-wise concatenation of the source images ($I_s^{1:k}$), their corresponding key-point representations ($P_s^{1:k}$), and the target key-point representation ($P_t$). The output is a 2D mask with as many channels as the number of input images.

Hence, the combined soft-selection volume has dimensions $H \times W \times k$. For the present work, we fix the number of input source images to three(k=3). Intuitively, each $(row, col)$ position of the combined soft-selection volume represents the probability of deriving the output pixel at that position from one of the $k$ source images.

The problem of predicting such a per-pixel ARMap is posed as a joint, conditional soft attention over the input images (Eqn 2), where the conditioning is on the target pose. We model this as an ARMap prediction task to ensure that the produced weights represent the joint probability of the output pixel being derived from a source image. To this end, we use a network adapted from the Swin Transformer [20] with a UPer [38] head. As described in [20], the Swin transformer captures the inter-channel relationships in the input (i.e. between the input images and key-point representations) using self- and cross-attention between shifted windows. Its computational complexity is linear in the size of the output which makes it useful for high-resolution input and output. The UPer head has also been shown in [20] to be highly accurate for per-pixel segmentation due to its ability to merge information from multiple scales. The ARMap masks $s_{1:3}$ are obtained by applying *SoftMax* on the out-
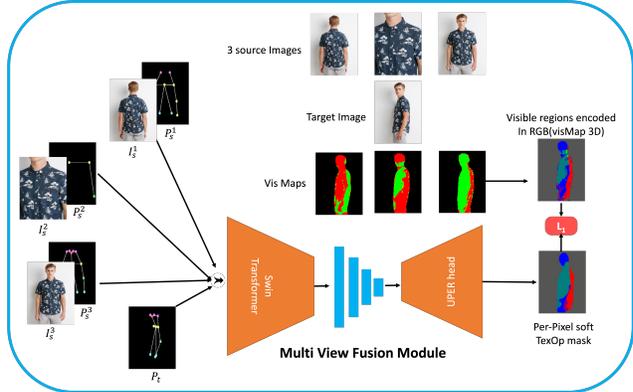


Figure 3. Pre-training of our Multi-View Fusion module for learning visibility correspondence with the target pose

put weights obtained from the UPer head at a resolution of 128x128. Formally:

$$s_{1:k} = Softmax(W_{1:k}|I_s^{1:k}, P_s^{1:k}, P_t; \theta) \qquad (2)$$

Here, $W_{1:k}$ are the computed weights for each of the input images and $\theta$ are the parameters of our MVF module. The ARMap masks ($s_{1:k}$) are used to fuse the pose- and texture features for the corresponding source images ($I_s^{1:k}$). The fusion is done with the features and not the images themselves to ensure that non-local information present in each image can be combined effectively into the resulting fused features ($e_{t,l}^{fused}$ and $e_p^{fused}$). The computation can be summarized by the following equations:

$$e_p^{fused}; e_{t,l}^{fused} = \sum_{\lambda=1}^{k} e_p^{\lambda} \odot s_{\lambda}; \sum_{\lambda=1}^{k} e_{t,l}^{\lambda} \odot Interp(s_{\lambda}, l)$$

$$I_p = D(e_p^{fused}, e_{t,l}^{fused})$$

Here, $Interp(s_{\lambda}^l, l)$ represents bilinear interpolation of ARMap to the resolution of the $l^{th}$ texture encoding layer and $\odot$ denotes pointwise multiplication of feature encodings. Fig 1 illustrates the entire process.

**Design & Pre-Training of fusion module** We essentially wish to model the joint probability of a pixel being derived from a source image using our ARMap. The architectures which match the closest to our desired objective are the ones that were proposed for performing semantic segmentation. The source images can be thought of as different classes and the task of the segmentation network is to predict the likelihood that a certain location should be retrieved from a given input image. However, there are two significant disparities between our intended objective and the output of standard semantic segmentation networks. First, conventional segmentation networks have aligned input and output, allowing them to rely heavily on proximity to establish the class. In our scenario, however, the source humans are all standing

in distinct poses. Thus, the network should be able to properly utilize global information from all viewpoints when predicting the probabilities for any pixel. Secondly, unlike semantic segmentation where the task is to predict one class per pixel, our use case requires a prediction of multi-class probability distribution to facilitate blending from different views. Guided by these considerations, our choice of architecture is to use Swin Transformer[20] backbone with UPer Head[38] and in Section 6 we show comparisons with other alternatives.

To improve its attention to visible regions in each input image (as per the target pose), the MVF module is first trained to predict the visibility map (Fig 3) using an L1 loss against ground truth generated using a separate 3D pose-estimation(visMap 3D) and projection mechanism based on DensePose [9]. We analyze the effect of this pre-training in Section 6. Finally, the pre-trained network is trained end-to-end for the human reposing task, using the following training objective:

1. Minimizing the following distances between the output and ground-truth images ($I_p$ and $I_{gt}$):
   - The pixel-wise mean of absolute difference ($L_1$) for the exact pattern and shape reproduction.
   - VGG-features based perceptual difference ($L_{vgg}$)
   - Style difference measured using the mean-squared difference between the Gram-matrices as described in [8] ($L_{sty}$).

   The perceptual and style losses help preserve semantic features taken from the input images, such as the identity of the person, and the garment style.
2. Minimizing an adversarial loss based on LSGAN [24] ($L_{adv}$) for the output image. This is useful to render realistic output, especially for regions where the decoder must guess the pixel colors.

The total loss is defined as:

$$L(I_p, I_{gt}) = \alpha_{rec} \|I_p, I_{gt}\|_1 + \alpha_{per} L_{vgg}(I_p, I_{gt})$$
$$+ \alpha_{sty} L_{sty}(I_p, I_{gt}) + \alpha_{adv} L_{adv}(I_p, I_{gt})$$

where $\alpha_{rec}, \alpha_{per}, \alpha_{sty},$ and $\alpha_{adv}$ are hyperparameters used to combine the losses.

We observe that the Multi-View Fusion module produces accurate region-specific ArMaps (see Fig 4), attributing the information required to produce accurate and realistic output images. This provides a way to explain the operation of the network as a blending of image regions for reproducing the correct output image. In the following section, we offer a series of experiments and ablations that demonstrate the efficacy of our technique.

## 4. Experiments

**Dataset.** We carry out our experiments using the Deepfashion dataset [21] which has a total of 52,712 high-quality person images with plain backgrounds. Following recent human reposing works [29, 5, 2, 43], we split the dataset into 48674 training images and 4038 testing images. For all our experiments, we resize the images to $256 \times 256$ resolution. We create new training and test quadruples consisting of 3 input source images and 1 output image. We wanted our network to be able to perform reposing task for any number of images $\leq 3$. For training, we utilize all permutations of input source images. We also repeat single source image multiple times in the input to ensure that the network works well if the number of input images are less than three. Note that if a single source image is repeated three times, UMFuse essentially turns into an identity operator and the network functionally becomes the same as its underlying backbone(GFLA[29], VGFlow[14] etc). The total number of training samples becomes($\Sigma^{n-1}P_1 + {}^{n-1}P_2 + {}^{n-1}P_3$), where n is the number of images of a single model in different poses and the sum is over the whole dataset. One of the advantages of performing multi-view training is the polynomial increase in dataset tuples compared to single-view reposing where the samples are limited to $\Sigma^{n-1}P_1$. The testing split consists of one entry for a distinct 3-view combination($\Sigma^{n-1}C_3$) to consistently evaluate the performance of our model for different number of source images(Sec 5). The order of the testing quadruples was randomly chosen to ensure fairness. Finally, we are left with 720,949 training and 5,247 testing quadruples

**Evaluation metrics.** Following previous works in supervised person image generation learning, we use the Structural Similarity (SSIM) [31], Peak Signal to Noise Ratio (PSNR) [12], Learned Perceptual Image Patch Similarity(LPIPS)[42] and Frechet Inception Distance (FID) [11] metrics to evaluate the results of our multi-view human reposing output. PSNR measures the exact reconstruction accuracy compared to the ground-truth image, while SSIM [32] measures fine-grained similarity based on the local contrast, luminance, and structure. LPIPS computes the patch-wise similarity between two images by comparing the AlexNet activations and FID score measures the realness of the images by computing the 2-Wasserstein distance between the latent space statistics of generated and the ground truth data.

**Implementation details** All experiments are carried out on $8 \times 3090$ RTX Nvidia GPUs. For the multi-view fusion module pre-training, we train the model for 3 epochs using a batch size of 32, a learning rate of $10^{-4}$, and the Adam optimizer [17]. For the final training of the Multi-view reposing task, we use a batch size of 24, a learning rate of $3 \times 10^{-4}$,
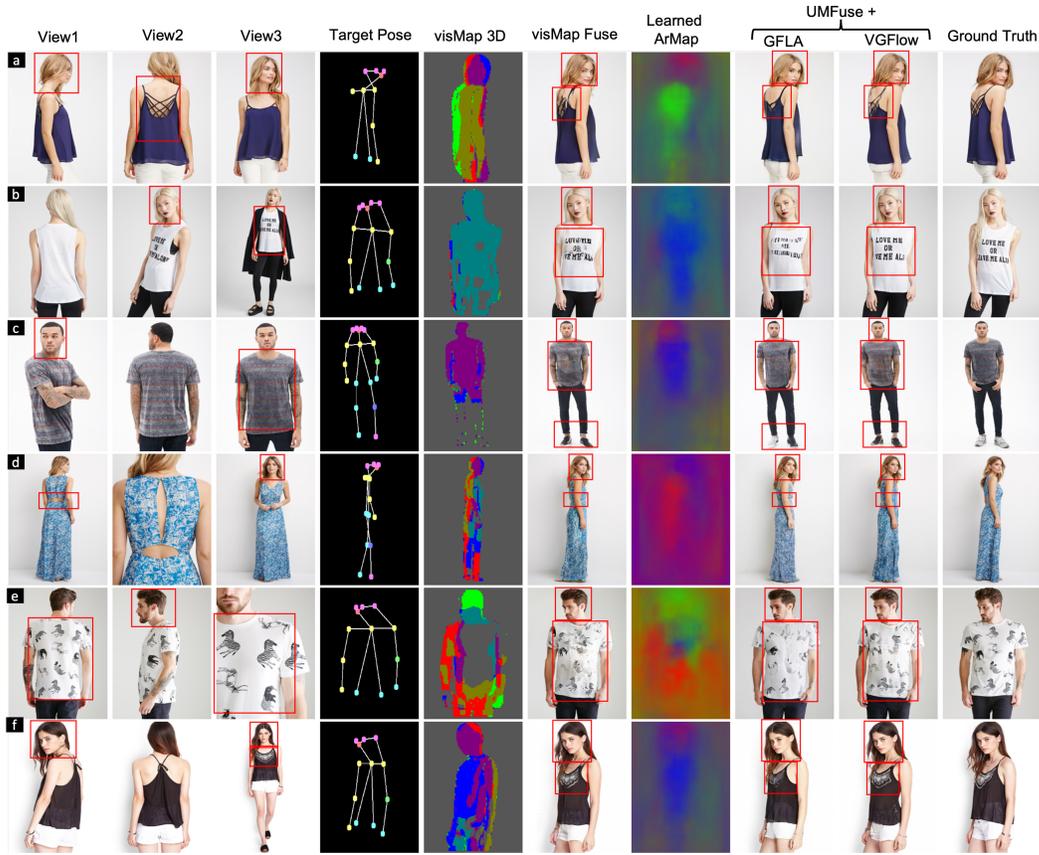
Figure 4. Qualitative Results for Multi-View reposing capturing improvements in (a) handling occlusions, (b) text preservation, (c) inpainting missing information, (d) preserving cloth designs, (e) style transfer from invisible regions, and, (f) pattern reproduction

train for 10 epochs and then finetune only the generator.

## 5. Results

Our experiments highlight the advantages of using multiple views (MVHR) over the single view reposing task. The test dataset for the multi-view tests is prepared as discussed in Section 4, into quadruples of three source images and one target image (pose). We report the qualitative improvements (Fig 4), as well as the enhancement across a wide set of metrics. Table 1 presents the quantitative improvements in output quality obtained by using 2 and 3 unique source views over a single source view (Details in Sec 6). All the image quality metrics(SSIM, PSNR, LPIPS and FID) show a consistent increase with increasing the number of views and are easily attributed to our framework's ability to fill in the missing information or find the best texture correspondence by sampling from distinct views of the same person. We also show that using a single image is fundamentally worse compared to 3 views by testing a source pose that is closest to the target(1-view closest) in terms of the Object-Keypoint Similarity score [1] (using a rigid

pose-alignment step). To show the generalizability of our method, we freeze our multi-view fusion module and train GFLA[29] generator using our methodology. The results show clear improvement over single-view reposing as the network becomes capable of combining texture from different viewpoints. For establishing a baseline, we also perform fusion by visMap3D generated using Densepose[9]. This approach confuses the network in areas where a single image would be better(Fig 4 Row b) or an invisible region can provide better correspondence (Fig 4 Row e). The network also suffers from mistakes due to incorrect 3d pose estimation by Densepose[9]. An end-to-end learned approach performs significantly better in establishing the best pixel-level correspondence from source images.

Qualitatively, we see that providing multiple source views of a person helps in improving the transfer of different body and clothing characteristics to the output. In Fig 4, row a-c shows the mixing of the model face from one image with the back/side garment view from another. This is further confirmed by visualizing the ARMap. To read the RGB mask - the *red* value represents selection weights for the features of the first source image, the *green* value
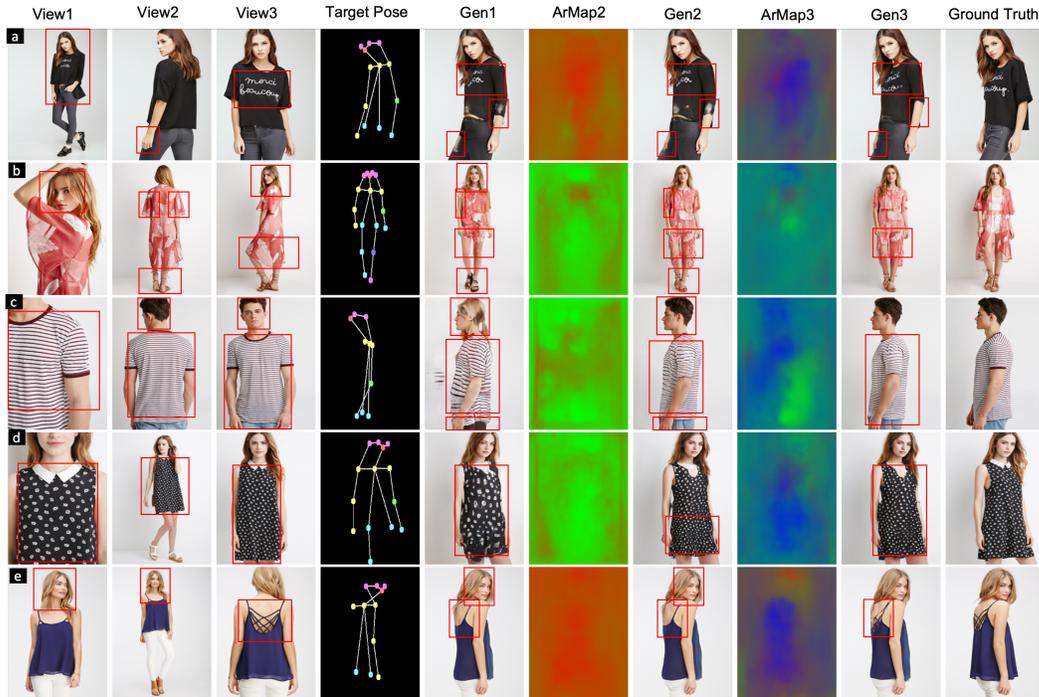
Figure 5. Here, we show incremental improvements gained by the network after the addition of multiple images. Gen1 is generated using only View1 (repeated thrice). Gen2 is generated using View1 and View2 pair with ArMap2 showing the attribution map using red and green colors respectively. Gen3 is obtained using all the 3 images and ArMap3 encodes the weights in RGB respectively.

| Method | SSIM ↑ | PSNR ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| VGFlow + UMFuse for < 3 Views | | | | |
| 1-view | 0.716 | 17.16 | 0.204 | 13.52 |
| 1-view(closest) | 0.729 | 17.88 | 0.181 | 13.15 |
| 2-view | 0.730 | 17.96 | 0.179 | 12.16 |
| 3-View comparisons | | | | |
| GFLA+UMFuse | 0.731 | 18.21 | 0.176 | 13.48 |
| VGFlow+visMap | 0.731 | 18.16 | 0.177 | 12.23 |
| VGFlow+UMFuse | **0.737** | **18.32** | **0.168** | **12.00** |

Table 1. Multiple views help in improving SSIM, PSNR and LPIPS metrics significantly over a single view.

for the second image, and the *blue* value for the third. For example, in row a, the face is derived from the combination of View1 (red channel) and View3 (blue channel) while the back design of the garment is obtained using the second view (green channel). Row b highlights accurate skin generation and conservation of text geometry while generation. Row c shows the network's capability of in-painting the missing portions which are still not visible in all the source views and Row d depicts the accurate reproduction of the texture by fusing appearance appropriately from a desirable scale image(view1). Row e captures the appropriate style transfer to the invisible regions and Row f displays proper pattern reproduction.

# 6. Ablation Studies and Analysis

**Benefits of using multiple views** We highlight the incremental improvements done by UMFuse in Fig 5 and Tab 1. Gen1 shows the generated output for a single image (View1) repeated 3 times as input. Gen2 uses the pair (View1, View2), repeating View1 for the 3 views. Gen3 is produced by using all 3 views. $ARMap1$ is redundant as the final encoding(after merging using softmax weights) would remain the same regardless of the map. $ARMap2$ is re-normalized to encode the weights assigned to View1 and View2 as Red and Green respectively. We see that in (a), better geometric correspondence for text was obtained after using View3. In (b), the model produced the wrong shoes and dress length in Gen1 which were corrected consecutively in Gen2 and Gen3. In (c), because the face was not visible, Gen1 has the image of a woman and Gen2 was able to correct it but the front region of the cloth still has blurred lines. Gen3 produces the closest match to the ground truth. In (d), we see that the zooming of the source image also has a huge impact on the generated output and the best texture spacing is obtained using a pose which is at a similar zoom level from the Target pose. In (e), we see that even though Gen1 & Gen2 looks realistic, they still lack the wired pattern at the back of model.
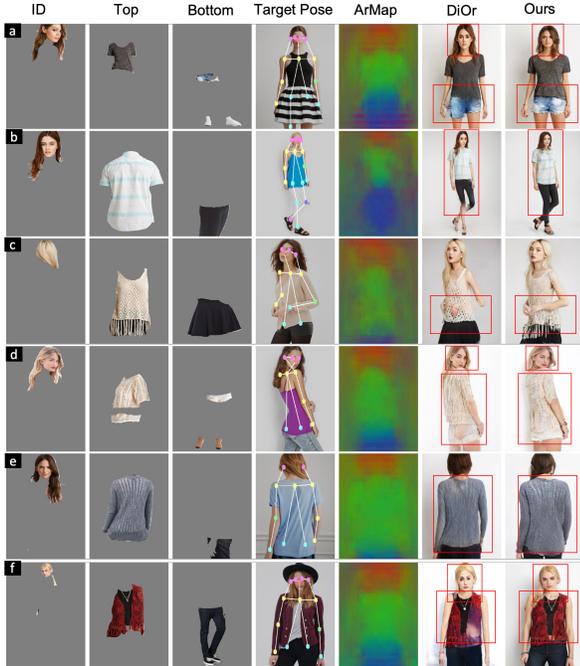
Figure 6. Qualitative Results for Mix & Match Human Image Generation showing realistic generation quality while preserving bodyshape(a), modelling complex pose(b), complicated design and texture of cloth(c), heavy occlusions(d), missing information(e), multiple clothing garments with accessories(f). Additional results can be found in the supplemental material.

**Extension to Mix and match task**  Many single view reposing models can automatically perform virtual try-on as an additional task[5, 29, 27] as there is a natural compatibility between the two. Similarly, multi view reposing has a natural correspondence with the Mix and match task.  In Mix & Match Human Image generation(MMHIG), given a tuple (id, top, bottom, pose) from different views, we model all of these jointly to synthesize the final output. We use an off-the-shelf human body parsing network[39] to segment an image into different fashion components($I_{id}, I_{top}, I_{bottom}$) before using the images as input to our network.  We then train our network to reconstruct a novel view(guided by $P_t$) which combines all the fashion elements. By learning to disentangle these elements in a supervised fashion, UMFuse is able to generate accurate rendition of the combined images.  More details about the training can be found in supplementary.  Note that editing only the pose element will result in single-view human reposing task, Changing top or bottom of the tuple is referred to as virtual try-on and changing id results in identity swapping. In both MVHR and MMHIG tasks, we observe that the Multi-View fusion module produces accurate ArMaps (see Fig 4,6), attributing the information required to produce accurate and realistic output images to

the appropriate source views. We compare our output quantitatively and qualitatively to that produced from DiOr[5] which performs sequential edits on a person image, as it closely matches the Mix-and-Match task. Performing edits sequentially on different parts of the tuple results in inferior output quality as these networks are not able to handle the intersection regions between clothing items and body parts successfully which often results in bleeding and bad texture reproduction. Also, the compounding of errors introduced by these sequential edits result in identity deterioration and inadequate occlusion handling. UMFuse is a robust, joint learning variant of the error-prone sequential reposing and virtual try-on approach. For quantitative comparison, we prepare 10,000 tuples selected from the test set randomly. We find that UMFuse significantly outperforms DiOr[5] with an FID score of **14.71** vs DiOr's **21.73**.

| Fusion Method | SSIM ↑ | PSNR ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|---|
| visMap 3D | 0.731 | 18.16 | 0.177 | 12.23 |
| Swin+UPer(-PT)[20] | 0.732 | 18.05 | 0.178 | 12.77 |
| UNet (+PT)[30] | 0.735 | 18.30 | 0.172 | 12.37 |
| Swin+UPer(+PT)[20] | **0.737** | **18.32** | **0.168** | **12.00** |

Table 2. The Swin+UPer(UMFuse) combination with pre-training(PT) works best for the reposing task.

**Multi-View Fusion Module Ablation**  In Tab 2, we compare the performance for different methods of latent space fusion. In the first method, we compare against the baseline $visMap3D$ obtained through Densepose[9] to fuse the latent space features. However, direct fusion result in sub optimal performance(Sec 5).  In the second method, we use a Unet[30] instead of Swin+Uper combination to predict the selection mask but it wasn't able to produce the desired accuracy due to limited field of vision & lack of global context in ConvNets. Finally, we ran an experiment of directly training the Swin Transformer without any visibility region Pretraining. As can be seen, there is a significant jump in SSIM(0.732→0.737), PSNR(18.05→18.32), LPIPS(0.178→0.168) and FID(12.77→12.00) metric due to our pretraining(qualitative benefits in supplementary).

# 7. Conclusion

We present UMFuse- a novel framework for effectively combining information from multiple source images in a pose-guided human image generation pipeline. The approach consists of using a multi-view fusion module that explicitly produces a 2D ARMap for combining features from multiple source images. We demonstrate its effectiveness on two newly proposed tasks: Multi-View human reposing and Mix & Match Human Image Generation with state-of-the-art results. UMFuse showcases a promising new and pragmatic approach to produce better results in pose-guided human image generation tasks.

# References

[1] Coco keypoints evaluation. https://cocodataset.org/#keypoints-eval. 6

[2] 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, 5

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 1

[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. 1

[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. 3, 5, 8

[6] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in neural information processing systems*, pages 474–484, 2018. 1

[7] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[8] 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. 5

[9] 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. 5, 6, 8

[10] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 1

[11] 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. 5

[12] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 5

[13] Zhichao Huang, Xintong Han, Jia Xu, and Tong Zhang. Few-shot human motion transfer by personalized geometry and texture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2297–2306, 2021. 2

[14] Rishabh Jain, Krishna Kumar Singh, Mayur Hemani, Jingwan Lu, Mausooom Sarkar, Duygu Ceylan, and Balaji Krishnamurthy. Vgflow: Visibility guided flow network for human reposing. *arXiv preprint arXiv:2211.08540*, 2022. 2, 3, 5

[15] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022. 2

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] 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. 1, 2

[19] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 2

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4, 5, 8

[21] 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. 2, 5

[22] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. *CoRR*, abs/2104.06650, 2021. 2

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017. 1

[24] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076, 2016. 5

[25] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 2

[26] 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

[27] 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, 8

[28] 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 (CVPR)*, June 2020. 1

[29] 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, 5, 6, 8

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 8

[31] Kalpana Seshadrinathan and Alan C Bovik. Unifying analysis of full reference image quality assessment. In *2008 15th IEEE International Conference on Image Processing*, pages 1200–1203. IEEE, 2008. 5

[32] De Rosal Igantius Moses Setiadi. Psnr vs ssim: impercepti-bility quality assessment for image steganography. *Multimedia Tools and Applications*, 80(6):8423–8444, 2021. 5

[33] 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. 1, 2

[34] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsuper-vised person image generation with semantic parsing trans-formation. *CoRR*, abs/1904.03379, 2019. 1

[35] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019. 2

[36] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable per-son synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020. 2

[37] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Hu-mannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 2

[38] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understand-ing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 4, 5

[39] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and effi-cient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 8

[40] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wang-meng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020. 1

[41] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 2

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shecht-man, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recogni-tion*, pages 586–595, 2018. 5

[43] 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 Con-ference on Computer Vision*, pages 161–178. Springer, 2022. 2, 5