UnitedHuman: Harnessing Multi-Source Data for High-Resolution Human Generation

Jianglin Fu[1*], Shikai Li[1*], Yuming Jiang[2], Kwan-Yee Lin[1,3], Wayne Wu[1†], Ziwei Liu[2†]
[1] Shanghai AI Laboratory [2] S-Lab, Nanyang Technological University [3] CUHK
{fujianglin, lishikai}@pjlab.org.cn
{yumingj80, wuwenyan0503}@gmail.com ziwei.liu@ntu.edu.sg junyilin@cuhk.edu.hk

Figure 1: **UnitedHuman.** We present our method UnitedHuman that **left)** integrates multi-source datasets related to human into the full-body image space and **right)** generates full-body human images at multiple resolutions, and captures high-quality details of the body-parts across different resolutions.

Abstract

*Human generation has achieved significant progress. Nonetheless, existing methods still struggle to synthesize specific regions such as faces and hands. We argue that the main reason is rooted in the training data. A holistic human dataset inevitably has insufficient and low-resolution information on local parts. Therefore, we propose to use multi-source datasets with various resolution images to jointly learn a high-resolution human generative model. However, multi-source data inherently **a)** contains different parts that do not spatially align into a coherent human, and **b)** comes with different scales. To tackle these challenges, we propose an end-to-end framework, **UnitedHuman**, that empowers continuous GAN with the ability to effectively utilize multi-source data for high-resolution human generation. Specifically, **1)** we design a Multi-Source Spatial Transformer that spatially aligns multi-source images to full-body space with a human parametric model. **2)** Next, a continuous GAN is proposed with global-structural guidance and CutMix consistency. Patches from different datasets are then sampled and transformed to supervise the training of this scale-invariant generative model. Extensive experiments demonstrate that our model jointly learned from multi-source data achieves superior quality than those learned from a holistic dataset. Project page: https://unitedhuman.github.io/.*

[*] Equal contribution.
[†] Equal advising.

1. Introduction

Human generation tasks have been intensively explored recently, as synthesized photo-realistic human images can benefit various related applications, such as virtual try-on, movie production, etc. Despite the great success of face generation since the emergence of StyleGAN [17], early attempts at human generation tasks [10, 11, 15] exhibit limited generative capabilities, especially in producing high-resolution full-body humans.

We argue this is due to intricately articulated human structures and limited training datasets. Specifically, since local parts like hands and faces only occupy a small portion of the entire image, and as a result, they cannot provide the model with sufficient texture information. Unfortunately, to the best of our knowledge, a comprehensive dataset capable of representing highly-detailed visual information on various human body parts is sorely lacking. Also, collecting such a dataset from scratch is time-consuming and labour-intensive.

Despite the scarcity of holistic human datasets, a vast quantity of human partial-body data is accessible to assist scholars in finishing multifarious human-related tasks [11, 12, 22, 23]. A trivial solution is to supplement the generation process with multiple datasets of human body parts. This simple idea is promising since the existing human-related datasets are expected to enhance details of local body parts, and these multi-source datasets constructed from different groups maintain a high degree of diversity in several aspects, including image scale, illumination, and body part position. Also, datasets of body components offer more ample texture details compared to full-body datasets. Both of these satisfy our needs for human generation and motivate us to utilize multiple human-related datasets to generate high-resolution human bodies.

To unite these multi-source datasets to push the limit of human generation, we analyze the difficulty of this endeavor and find that the main obstacle is twofold. 1) It is challenging to align disparate body parts into a coherent, realistic human since the scales and locations of the body components in each dataset have different distributions. Merely aligning with 2D keypoints [3] is feasible for rigid objects but suboptimal for hinged human structures, as it disregards depth information as well as the body shape. A reliable alignment mechanism is therefore required to connect these body components. 2) Image resolution varies among multi-source datasets, and this work requires training with these datasets to synthesize results at different resolutions. Multi-scale generation is another necessity that needs to be addressed because GAN models are typically trained on identical-resolution images and can only synthesize images of a fixed resolution.

To tackle the above two challenges, we propose **UnitedHuman**. This end-to-end framework, consisting of *Multi-source Spatial Transformer* module for spatial alignment and *Continuous GAN* module for arbitrary-scale training, leverages multiple datasets to synthesize higher-resolution full-body images. Fig. 2 illustrates our entire working pipeline. Specifically, the *Multi-Source Spatial Transformer* employs the parametric human model as a prior. This transformation serves to convert the partial-body image into the full-body image space, leading to a unified spatial distribution. With different sampling parameters in the full-body image space and latent code from the prior distribution, the *Continuous GAN* takes the transformed Fourier feature as input and generates the fixed-resolution patches at corresponding positions and scales. Finally, the generated patches over the full-body space are stitched to form high-resolution full-body images.

Compared to the existing cutting-edge human-GAN models, *UnitedHuman* demonstrates the ability to incorporate human-related datasets from multiple sources to produce high-resolution humans. The humans generated by our model, with zoomed-in details from 256px to 2048px, are shown on the right side of Fig. 1. During experiments, we demonstrate that by leveraging only 10% of the high-resolution images needed by the SOTA methods, along with the incorporation of various partial datasets from multiple sources, our technique can outperform the existing state-of-the-art outcomes. Furthermore, the model has the potential to scale up to any resolution by introducing additional partial-human datasets.

In summary, our main contributions are listed: 1) We propose UnitedHuman, the first work to explore the usage of multi-source data for high-resolution human generation tasks. 2) We design *Multi-source Spatial Transformer* to assist in aligning body parts from diverse datasets, based on the articulated human structure. 3) We design the *Continuous GAN* to achieve multi-resolution and scale-invariant training.

2. Related work

Human Image Generation. Scholars in the 2D human generation field commonly adopt GAN models as paradigms: they either generate humans unconditionally [11] or are conditioned on pose/semantic priors [1, 25, 28, 32, 38, 39, 40], which explicitly alleviates the entanglement between pose and appearance. Other image-based researchers exploit diffusion models to achieve human-associate tasks, such as head swapping [34] and person image synthesis [2]. Although the above methods can generate full-body humans, they have certain limitations when it comes to reproducing intricate details of local parts. One recent work, InsetGAN [10], proposed a novel method for integrating multiple GAN models to alleviate artifacts in the generated human images. This pipeline iteratively searches and combines latent codes from the pretrained GAN mod-

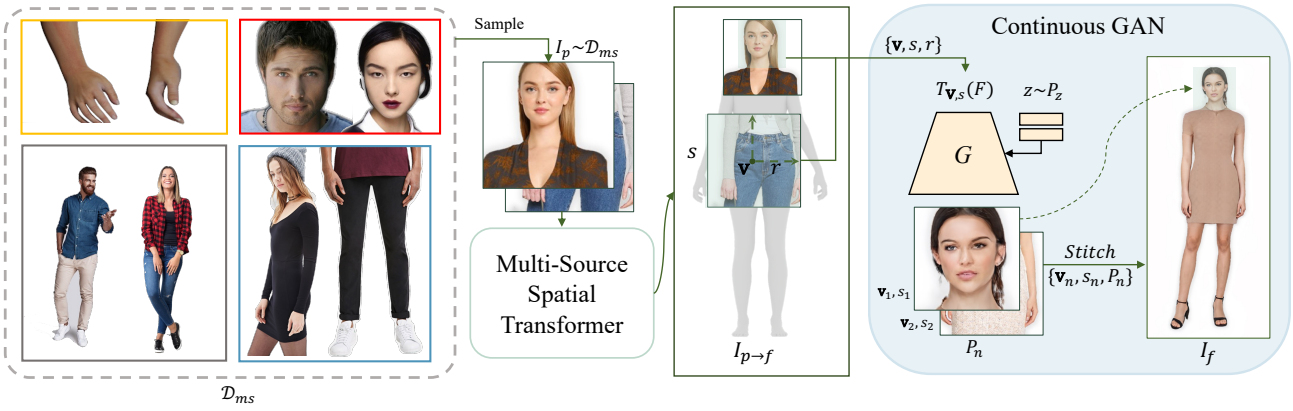

Figure 2: **Overview of UnitedHuman.** Given the images I_p from the multi-source datasets D_{ms}, the Multi-Source Spatial Transformer puts the partial-body image into the full-body image space as $I_{p \to f}$ for a unified spatial distribution. With sampling parameters $\mathbf{v}, s, r$ and latent code z from the prior distribution, our Continuous GAN generates the patches P_n at center $\mathbf{v}$ with scale s. The patches over the full-body space are stitched to form the high-resolution full-body images I_f.

els, and the found latent codes can produce plausible full-body humans. This work demonstrates the possibility of approximating the latent codes of two or more pretrained models from different distributions to each other to generate a coherent full human body. Unlike InsetGAN, we focus on how to leverage different existing datasets to establish an end-to-end framework for high-resolution human generation.

Multi-scale Generation. Requiring models to produce images in multi-resolution or even higher-resolution raises the topic of multi-scale generation. The prevalent approach is to stitch the generated independent patches to construct a full-size image [4, 21, 27]. The leading challenges of this task are generating patches according to specific positions and maintaining global structure across scales. The former is often solved by employing positional encoding [7, 21, 27, 36]; consistency loss [4] or global discriminator [27] are often employed as a method to overcome the later challenge. Consequently, AnyRes [4] suggest a novel continuous-scale training architecture and exploit datasets with arbitrary resolutions. Although AnyRes can be trained on datasets of arbitrary size, it still requires the training set to have some similarity in distribution. In Sec. 4, we show AnyRes cannot generate plausible full-body humans with local-part datasets, which means it cannot take advantage of partially human datasets. To tackle this issue, we design a unique alignment module targeting human structure.

Human-centric datasets. Human-centric tasks are long-standing and always attract everyone's attention. Although the tasks related to humans are diverse and complicated, the solution to these downstream tasks can greatly improve productivity and contribute to the entire community. Researchers in academia and industry construct large-scale human-related datasets [6, 8, 11, 12, 14, 17, 22, 23, 26, 33, 35, 37, 41] through collection or synthesis methods to solve various human-related problems. For example, face generation commonly employs FFHQ [17] and CelebA [23], hands synthesis has DART [12] and FreiHAND [41], DeepFashion [22] and Viton-HD [8] are released for fashion item generation/recommendation, SHHQ [11] is constructed for full-body generation. We select four representative datasets to conduct experiments. This paper first adopts the downsampled SHHQ as a benchmark to provide global information of human bodies, as SHHQ delivers the highest-resolution full-body images. Then, to offer high-definition textures, the high-resolution versions of DeepFashion and SHHQ are also included. DART and CelebA are added to enhance the local details of hands and faces respectively.

3. UnitedHuman

Our aim is to generate full-body images from multi-source datasets of varying distributions and resolutions. As shown in Fig. 2, Multi-Source Spatial Transformer aligns the spatial distributions among the multi-source datasets within the full-body image space (Sec. 3.1). Given the spatial and scale parameters in the image space, the proposed Continous GAN will generate the patches to form the full-body images with global-structural guidance and CutMix consistency (Sec. 3.2).

3.1. Multi-Source Spatial Transformer

Previous works [17, 18] require full-content images with an aligned spatial distribution such as FFHQ and AFHQ. While the SHHQ dataset [11] provides full-body images for human generation, it can only reach a resolution of 1024, which is insufficient for generating specific regions such as faces and hands. Therefore, we set our sights on vast human partial-body data with various spatial distributions

but provide high-resolution parts. To unite these datasets for human generation, we propose the Multi-Source Spatial Transformer that transforms different partial-human images into a defined full-body image space using the parametric human model as prior.

We first define the full-body image space as a bounded region that represents the spatial distribution of the human from the well-aligned full-body dataset. Note that the partial-body images can also be placed in this image space with appropriate transformation. To represent both the full-body and partial-body human in a unified manner, we choose the parametric model SMPL [24] as the geometry prior. Specifically, given a full-body dataset, we first employ a model-based human reconstruction method [19] to estimate the body parameters of pose θ, shape β, and a weak-perspective camera model α with scale and translation parameters. As for the issue of scale ambiguity in monocular images, we keep the camera intrinsics fixed and regard the estimated body heights as approximate, heavily constrained by the shape prior. Finally, our full-body image space S can be simply defined by rendering the bounded region with the camera parameters $\bar{\alpha}$.

Proper camera parameters of partial images are also required for placing the images within the full-body image space. However, inaccurately predicted parameters result from the lack of suitable training data under the partially-observed setting for human mesh recovery. To address this, we follow the regression-optimization hybrid manner [5] with the assistance of a full-body dataset. Specifically, we decompose partial images into visible and invisible parts by the predicted 2D keypoints [3]. After the initial regression on partial images, we optimize the pose parameters of visible parts by minimizing the error between projected 3D keypoints and estimated 2D keypoints. For unseen parts, we use a variational autoencoder (VAE) trained on pose parameters of the full-body dataset as the pose prior for optimization. We also propose an additional orientation regularization to regularize the pitch value of the global orientation o to prevent the problem of depth ambiguity. Although the optimization focuses on pose parameters, the weak-perspective camera parameters can also be more reasonable. The overall loss of optimization is then defined as

$$\begin{aligned} E(\alpha_{opt}) &= L_{vis} + L_{invis} + L_{reg} \\ &= \sum_{i}^{\text{vis}} \mathcal{L}(J_i, J_i') + \sum_{i}^{\text{invis}} \mathcal{L}(\theta_i, \hat{\theta}_i) + \mathcal{L}(o, \hat{o}) \end{aligned} \tag{1}$$

where the J and J' denotes the projected keypoints with camera α_{opt} and estimated 2D keypoints respectively while the θ and the $\hat{\theta}$ denotes the axis-angle of joints from the predicted SMPL and VAE. The mean squared error (MSE) loss is used in Eq. 1. Finally, the partial images I_p are transformed into the full-body image space by

$$I_{p\to f} = H(\alpha, \bar{\alpha}) \times I_p \tag{2}$$

where H is the matrix that calculated by the optimized camera α_{opt} and average camera $\bar{\alpha}$ of the full-body dataset.

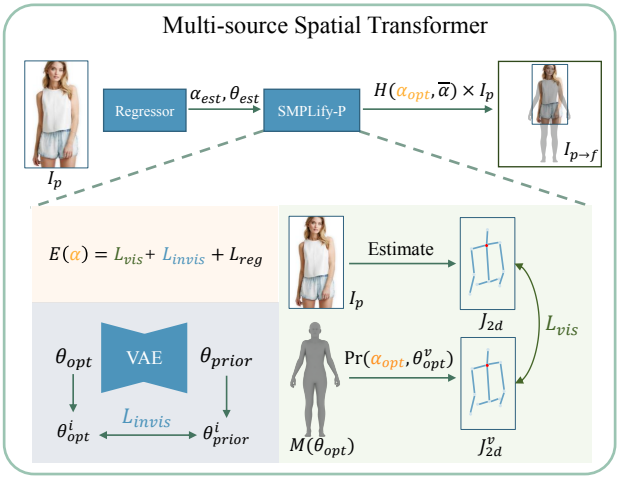

Figure 3: **Multi-source Spatial Transformer.** Given a image I_p, we first predict the camera α_{est} and pose θ_{est} of SMPL. Optimized by SMPLify-P on both visible and invisible parts, the camera α_{opt} is used to calculate the matrix H that transforms the patch into the full-body image space.

3.2. Continous GAN

Preliminary. Generative adversarial network [13] is proposed to synthesize images with a fixed resolution. Continuous-resolution generative models [4, 27] extend the approach to arbitrary-scale training by transforming the input embedding and applying supervisions between different scales. The generation of the image can be viewed as sampling values at discrete positions within a bounded region of continuous image space ranging from $[0, 0]$ to $[1, 1]$. In the sampling process, we define the center $\mathbf{v}$ as the coordinate center of the sampled patch while scale s and resolution r represent sampling frequency and the number of sampling points, respectively. For image sampling in the image space, the continuous-resolution generator G synthesizes the patch's pixel values at the sampled coordinates as:

$$\begin{aligned} G(z, \text{emb}, \mathbf{v}, s, r) &= G(T_{\mathbf{v},s,r}(\text{emb}), z) \\ &= T_{\mathbf{v},s,r}(G(\text{emb}, z)) \end{aligned} \tag{3}$$

where T is the transform function based on the sampling parameters $(\mathbf{v}, s, r)$ and z is the latent code sampled from the prior distribution. As seen in Equation 3, the continuous-resolution generator also owns the following properties: spatial equivariance and scale consistency.

Continuous Network. We constructed our network based on StyleGAN3-T [16] with a pixel-wise discriminator to fully utilize the multi-source dataset. Although the architecture of the generator is proposed for anti-aliasing and spatial equalvariance, it is able to generate the images at a slightly larger scale by directly modifying the sampling frequency of the Fourier feature. With additional training and supervision at different scales, we could utilize the transformation matrix T to control the image generation in the full-body image space. Specifically, given the sampling parameters $(\mathbf{v}, s)$ with a fixed resolution r, the transformation matrix T of input Fourier function is as follows:

$$T_{\mathbf{v},s} = \begin{pmatrix} \frac{1}{s} & 0 & \mathbf{v_x} - 0.5 \\ 0 & \frac{1}{s} & \mathbf{v_y} - 0.5 \\ 0 & 0 & 1 \end{pmatrix} \tag{4}$$

where T is an identity matrix that samples the full-body image in the continuous space when $(\mathbf{v}, s) = ([0.5, 0.5], 1)$. As for image patch generation with large scale s or different position $\mathbf{v}$, the two parameters of the Fourier feature, frequency f and phase p, can be transformed by T to represent the transformed sampling patch in the image space. Notably, the Nyquist–Shannon sampling theorem [31] is still satisfied when $s > 1$ for anti-aliasing generation [16]. The Fourier embedding $F(f, p)$ will be fed to the generator G to synthesize the image x as follows:

$$x = G(T_{\mathbf{v},s}(F)), z) \tag{5}$$

where z denotes the latent code from the prior distribution. A pixel-wise loss between generated patches at different scales is applied to achieve scale consistency.

So far our network supports the scale-invariant training on a holistic multi-resolution human dataset with a typical discriminator. However, the images from the partial-body datasets cannot occupy the entire continuous image space, resulting in partial-body image generation. One trivial solution is to sample the image patches inside the subregion of image space, where these patches can only be seen at specific ranges of position and scale. To overcome this issue, we apply the pixel-wise discriminator with a proposed CutMix consistency for multi-source dataset training. Following [30], we alter the architecture of D to an encoder-decoder network by applying the upsampling blocks and skip connections, resulting in performing pixel-wise classification of input image. As for partial-body image patches, we apply the CutMix operation by mixing the real patches and the generated ones by the mask of the subregion as:

$$\text{CutMix}(x, \hat{y}, M) = x \odot (1 - M) + \hat{y} \odot M \tag{6}$$

where M is the mask inside the subregion and $\odot$ is the element-wise production. Therefore the partial-body data can be sent to discriminator at all scales during training, bringing a more powerful discriminator.

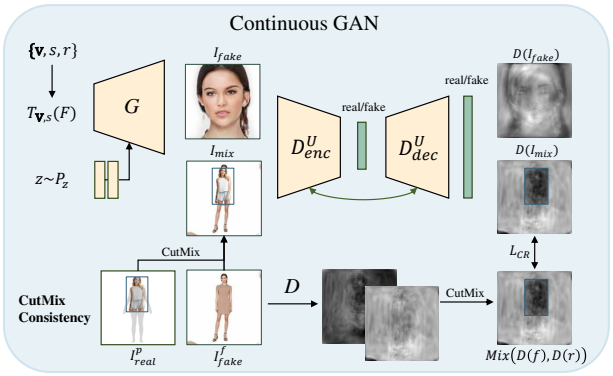

Figure 4: **Continous GAN**. Given sampling parameters (v, s), the transformed Fourier feature $T_{\mathrm{v},s}(F)$ is used to generate patches with latent code z. We use the U-Net discriminator for global and pixel-wise adversarial training. The proposed CutMix consistency makes the partial-body images I^p_{real} trained at all scales.

Multi-source dataset training. We implement the training among multi-source datasets in a two-stage manner. During the *Stage 1*, we train the basic model on a low-resolution full-body dataset as global structural guidance. We apply both global and pixel-wise non-saturating GAN loss with R1 regularization for *Stage 1* as:

$$\begin{aligned} \mathcal{L}_{adv,D} &= \mathbf{E}_{\boldsymbol{z}\sim P_z}[f(D(G(T_{\mathbf{v},s}(F)), z))] \\ &\quad + \mathbf{E}_{\boldsymbol{I}\sim P_{data}}[f(D(\boldsymbol{I})) + \lambda||\nabla_{\boldsymbol{I}} D(\boldsymbol{I})||_2^2] \\ \mathcal{L}_{adv,G} &= \mathbf{E}_{\boldsymbol{z}\sim P_z}[f(-D(G(T_{\mathbf{v},s}(F)), z))] \end{aligned} \tag{7}$$

where $f(x) = -log(1+e^{-x})$ and $T_{\mathbf{v},s}$ is an identity matrix.

In *Stage 2*, we use the Multi-source Spatial Transformer to sample the image patches from multi-source high-resolution datasets as real images. For training efficiency, we also apply the keypoint-based sampling since the human body occupies the horizontally center region mostly. As for scale-invariant training schema, the pixel-wise loss between the transformed patches and the full-body image generated by the basic model is also applied for same latent code z as follows:

$$\mathcal{L}_{pixel} = (\mathcal{L}_{lpips}(H(x), I_p) + \mathcal{L}_1(H(x), I_p)) \odot M_T \tag{8}$$

where H is another operation that transforms the generated patches to another scale and M_T is the corresponding mask. To achieve consistent training on partial images, we utilize the CutMix consistency by regularizing the decoder outputs of mixed image and the mixed of decoder outputs of real partial images and fake full images. The regularization is:

$$\mathcal{L}_{cr} = D_d(\text{Mix}(x, I_p)) - \text{Mix}(D_d(x), D_d(I_p)) \tag{9}$$

where D_d denotes the decoder of pixel-wise discriminator. Therefore, the overall loss of generator in *Stage 2* can be

formulated as follows:

$$\begin{aligned} \mathcal{L}_G &= \mathcal{L}_{adv,G} + \lambda_p \mathcal{L}_{pixel} \\ \mathcal{L}_D &= \mathcal{L}_{adv,D} + \lambda_{cr} \mathcal{L}_{cr} \end{aligned} \tag{10}$$

where λ_p and λ_{cr} are 5.0 and 1.0 respectively.

4. Experiments

4.1. Experimental Setup

The whole work is trained successively through two phases as discussed before. *Stage 1* is trained with full-body images in a resolution of 256×256 pixels, whereas *Stage 2* employs *Stage 1* model as the teacher model and further refines body-part details with multi-source datasets in higher-resolutions.

Datasets. In this work, experiments are conducted on the following human-centric datasets: SHHQ [11], DeepFashion [22], CelebA [23], and DART [12] that is rendered by Blender's Eevee [9]. Refer to Tab. 2 for more details.

Evaluation Metrics. Fréchet Inception Distance (FID) is one common indicator for accessing GAN models. However, studies [10, 11] argues that FID is more sensitive to diversity in data distribution but struggles to accurately evaluate the visual quality of a human body. In this paper, we jointly utilize two variants of FID to better quantify the model performance. 1) **patch-FID (pFID)** is proposed in AnyRes GAN [4] to better assess the local textures in high-resolution images. It randomly samples image patches from high-resolution datasets and stores each transformation matrix. The stored matrixes are then injected into the generator to produce image patches of correlated scales and positions. 2) **keypoint-FID (kFID)** can be treated as a special case of pFID. It calculates pFID patches around each body keypoint at a specific scale. This kFID is more closely connected to human keypoints, and is able to more clearly reveal the intricate texture representation surrounding the joint points with various degrees of freedom. To be precise, we employ the sampling parameters $\mathbf{v}$ and s to extract patches from both the generated full-body images and the training dataset. In particular, we fix the scale s at 8 to ensure the images are generated with a resolution of 2048 pixels. Then we randomly select a central point $\mathbf{v}$ ranging from $[0, 0]$ to $[1, 1]$ and crop patches around this center point $\mathbf{v}$.

Comparison Methods. To demonstrate the efficacy of our approach for generating plausible humans in high resolution by uniting multi-source datasets, we compare UnitedHuman with three baseline models: StyleGAN-Human [11], InsetGAN [10] and AnyRes GAN [4]. Tab. 2 depicts detailed data composition that is employed to train each model. Since StyleGAN-Human does not support training images with multiple resolutions and different body parts, we train it with $100K$ SHHQ$^{\mathbf{HR}}$ images in 1024×1024 pixels. InsetGAN serves as a multi-GAN optimization technique for merging face and body images created by separate GAN models. In our study, we adopt the public pre-trained CelebA [23] model with a resolution of 256×256 and the StyleGAN-Human model we previously trained as the two fundamental models for InsetGAN. Lastly, AnyRes allows continuous-scale training using a two-stage approach but lacks the ability to generate coherent individuals by combining body parts from multi-source datasets. As a result, we train its first stage using $100K$ SHHQ$^{\mathbf{LR}}$, and then employ $10K$ SHHQ$^{\mathbf{HR}}$ for the second stage of training.

Fairness of Comparison. The cross-method comparison is conducted fairly and impartially. To ensure a fair comparison, all the compared SOTAs are fully trained till they converged using the configurations of their reported best models. The main difference is the data composition (Tab. 2). Comprehensive training information for all the compared methods, encompassing training parameters and the time it takes for inference, can be found in the supplementary materials.

4.2. Main Results

Fig. 5 depicts the results of the visual comparison among our model against the three baseline methods. In the figure, we show the full-body humans (1024×1024 pixels) and the crops of the face and hand (obtained from 1024px and 2048px full-body humans), respectively. In addition, the full-body images in the figure are derived from the mean latent code of each model. As described in the above section, StyleGAN-Human is trained with a single holistic dataset, and it only supports the generation of fixed-size human images; therefore, the cropped faces and hands are up-scaled from the images in 1024×1024 pixels using bi-cubic interpolation. The same approach is used for InsetGAN to obtain image patches from 2048px. In contrast, AnyRes and our model are able to synthesize images with varied resolutions, and we show the patches cropped from the generated 2048px images. The figure indicates that the StyleGAN-Human model, which is trained on $100K$ high-resolution full-body images, is capable of producing plausible human images at 1024 pixels. However, attempting to up-sample these images to higher resolutions brings in artifacts. InsetGAN delivers less desirable results when the distribution of face and body is far apart. The results of AnyRes show that even trained on multi-scale data from the same distribution, it cannot effectively map articulated human body structures to different resolutions. Besides, UnitedHuman generates human images in high-resolution with greater details while maintaining the overall human structure. More results generated by UnitedHuman are illustrated in Fig. 6.

To quantitatively measure the advancement, Tab. 1 records the numerical numbers of different evaluation metrics. We utilize the SHHQ$^{\mathrm{HR}}$ dataset as the real data in FID

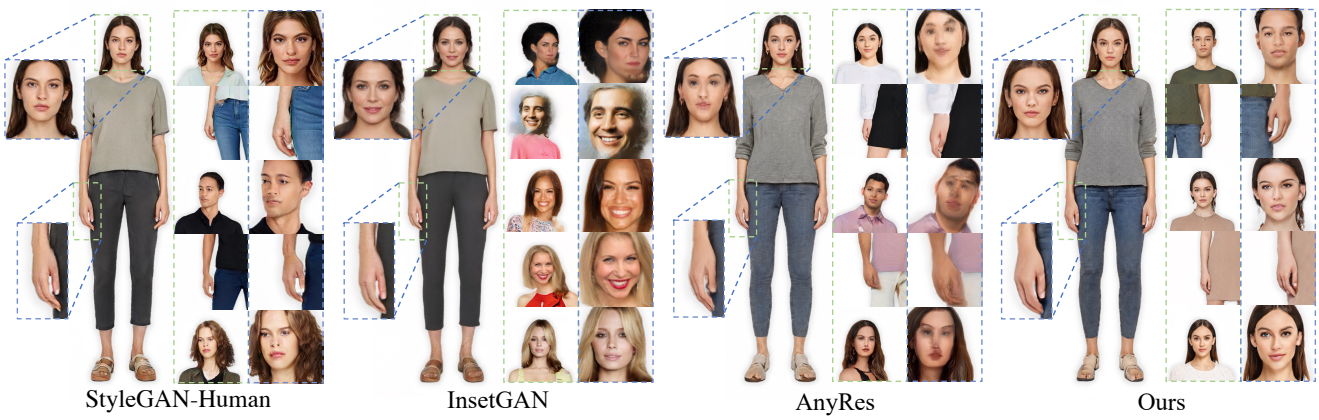

Figure 5: **Results of baseline comparison**: StyleGAN-Human [11], InsetGAN [10], AnyRes [4], and UnitedHuman. We exhibit the full-body human images generated from each experiment at a resolution of 1024 (- - -), as well as the face and hand patches cut from the 2048px images (- - -).

Figure 6: More results. The top row illustrates more UnitedHuman results in 1024px (- - -) and 2048px (- - -). An example of interpolation between latent codes is given in the bottom row.

calculations to guarantee equitable comparisons, since all four methods have seen this dataset. In addition, we do not have a ground-truth dataset that allows us to evaluate model performance in 2048 pixels. Therefore, we down-sampled all 2048-pixel images back to 1024px before commencing evaluation. As demonstrated in the table, UnitedHuman outperforms InsetGAN and AnyRes models in terms of kFID at each body keypoints as well as pFID. Comparing our approach to StyleGAN-Human, we are not surprised that the evaluation outcomes of UnitedHuman do not show a noteworthy advantage in terms of kFID and pFID. We argue that the reason for this phenomenon is that StyleGAN-Human is trained on the single, holistic SHHQ dataset only, the model encounters a simpler and more homogeneous data distribution. Furthermore, the model is assessed using the same SHHQ dataset, which possesses the inherent benefit of computing FIDs. For UnitedHuman, incorporating diverse datasets in the training process presents a more sig-

	kFID							$\overline{\text{kFID}}$	pFID	Precsion	Recall
	Face	Neck	Shoulder	Elbow	Hand	Hip	Knee				
SG-Human [11]	22.21	21.33	19.77	**19.42**	18.47	20.01	**20.56**	20.25	18.96	0.71	**0.65**
InsetGAN [10]	40.84	37.26	33.60	31.99	25.44	29.90	28.49	32.50	27.22	0.70	0.39
AnyRes [4]	33.26	31.20	28.55	33.57	33.02	37.97	34.26	33.12	30.49	0.67	0.48
Ours	**21.49**	**19.48**	**17.88**	19.66	**17.80**	**19.62**	20.96	**19.56**	**18.94**	**0.74**	0.61

Table 1: **Quantitative results for baseline comparison.** $\overline{\text{kFID}}$ denotes the average value derived from all kFIDs. All the metrics are measured in the images at 2048×2048 pixels. Our method delivers the overall best results, though seeing only 1/10 high-resolution full-body data compared to StyleGAN-Human [11] and InsetGAN [10].

	LR^{256}	HR^{1024}	SR^{2048}			
	SHHQ [11]	SHHQ	CelebA ^{256} [23]	DF_p [22]	DART [12]	SHHQ
SG-Human [11]	✗	$100K$	✗	✗	✗	✗
InsetGAN [10]	✗	$100K$	$30K$	✗	✗	✗
AnyRes [4]	$100K$	$10K$	✗	✗	✗	✗
Ours	$100K$	$10K$	$10K$	$10K$	$7K$	$5K$

Table 2: **Data composition** seen by baseline models and the proposed UnitedHuman. The datasets can be divided into three categories based on resolution: 256px (LR), 1024px (HR), and 2048px (SR). We treat CelebA in 256 pixels as an SR dataset because the faces in CelebA are of similar size to the faces in the full-body images at 2048px. DF_p refers to the partial human images from DeepFashion dataset. SHHQ^{SR} is obtained by up-sampling SHHQ^{HR} and taking only the lower half of the images. It is used to supplement the human lower-body dataset. We show that UnitedHuman can digest datasets from different data sources in diverse resolutions.

nificant difficulty. Besides, we only leverage $10K$ SHHQ^{HR} images along with other partial datasets to train the model. In this scenario, we are still able to achieve comparable and slightly better results than StyleGAN-Human.

In addition, our model achieves an FID score of 15.37, slightly higher than StyleGAN-Human's 13.81. Once again, we posit that this difference can be ascribed to StyleGAN-Human being trained on a large-scale holistic dataset. Moreover, it is crucial to note that FID itself may not effectively capture the perceptual quality. We further evaluate the four models with the improved precision and recall metric[20], as shown in Tab 1. As depicted in the table, UnitedHuman achieves the best precision but obtains marginally lower recall. According to the previous study [20], recall quantifies the fraction of the training data manifold covered by the generator, and we employ a subset of the training dataset (SHHQ^{HR}) used by StyleGAN-Human as our ground truth dataset. Consequently, the StyleGAN-Human model possesses an advantage in recognizing a greater number of true positive instances. On the other hand, our model, having encountered diverse datasets, has generated a data distribution that significantly deviates from the ground truth. This discrepancy has led to some of the generated images being considered negative cases during computation.

In sum, the above results reveal the effectiveness of our model in uniting multi-source datasets with various distributions and improving the details of the generated images. It also indicates that UnitedHuman can synthesize a decent full-body human utilizing a limited amount of high-resolution partial-body datasets.

4.3. Ablation Study

Ablation on Dataset. We begin with $10K$ $\text{SHHQ}^{\textbf{HR}}$ and progressively introduce additional datasets to train UnitedHuman in order to investigate the influence of data on the Multi-source Spatial Transformer. All the experiments, with the exception of data compositions, share the same *Stage 1* teacher model and training hyper-parameters. The results are shown in the first section of Tab. 3. Start from SHHQ^{HR}, we add $5K$ SHHQ^{SR} and $10K$ DeepFashion data into the model. As mentioned in Tab. 2, the partial images from DeepFashion are mainly focused on the upper-body, while SHHQ^{SR} supplements the lower-body. We observe a considerable enhancement in kFID and pFID after adding these two datasets. Subsequently, the inclusion of CelebA in the pipeline results in a reduction of approximately 1.5 points in kFID around the face, neck and shoulder, while causing an average increase of 0.5 points in other regions. Next, the introduction of the DART dataset leads to a notable enhancement in the kFID around hands. This ablation reflects that the benefits of incorporating body-part datasets outweigh the negative impact.

Ablation on Alignment. We conduct experiments to probe the impact of alignment strategy (see the second section of Tab. 3). When we align human poses, we preemptively take

	kFID							$\overline{\text{kFID}}$	pFID
	Face	Neck	Shoulder	Elbow	Hand	Hip	Knee		
Ablation on Datasets									
SHHQ^{HR}	25.56	24.94	22.28	25.24	25.07	28.15	27.13	25.48	25.58
+ SHHQ^{SR} +DF_p	23.43	22.02	20.14	21.75	19.89	22.03	24.61	21.98	21.87
+ CelebA	21.57	20.28	18.99	21.87	20.60	22.86	24.32	21.50	21.12
+ DART (Ours)	**21.49**	**19.48**	**17.88**	**19.66**	**17.80**	**19.62**	**20.96**	**19.56**	**18.94**
Ablation on Alignment									
Keypoint	22.56	21.89	19.77	22.52	20.46	22.91	27.04	22.45	22.88
Pose-mapping	51.24	22.49	27.01	24.41	19.47	21.54	28.55	27.81	25.32
SMPL (Ours)	**21.49**	**19.48**	**17.88**	**19.66**	**17.80**	**19.62**	**20.96**	**19.56**	**18.94**

Table 3: **Ablation on datasets and alignment.** All the metrics are measured in the images at 2048×2048 pixels.

the human images generated from the mean latent code of the teacher model as a reference and compute keypoints by OpenPose [3], which we denote as *apose*. The "Keypoint" experiments simply sample the patches of the associated keypoints from the real dataset according to *apose* and use the same transformation matrix to produce the corresponding patches. This naive approach ignores that not all poses of the generated images during training are in line with *apose*. A small offset in pose can cause the training data pairs to be misaligned. The second "Pose-mapping" method trains an auxiliary MLP network to enhance the accuracy when querying the keypoint locations in the generated images during the sampling process. This approach, however, only improves local alignment precision but still cannot cope with the articulated human body structure that contains multiple postures and body shapes. Compared to the above two experiments, UnitedHuman incorporated with SMPL achieves better alignment results, as shown in Tab. 3.

Ablation on Loss. Studies are performed to investigate how different loss components affect the outcome. Initially, the full set of the multi-source datasets is trained on StyleGAN3-T with standard adversarial loss. Following that, pixel loss is applied to supervise the global human structure, leading to a substantial reduction in the mean kFID score by 20.7 points and a corresponding drop in pFID by 17.96 points. Afterwards, the pixel-wise discriminator with CutMix consistency loss is added and both the kFID and pFID are dropped to 19.56 and 18.94, respectively. This set of experiments demonstrates that low-resolution full-body images can effectively provide structural guidance to the generation of high-resolution humans. Additionally, the use of a pixel-wise discriminator with CutMix consistency allows for the partial-body data to be distributed across different scales and further improves the model performance.

5. Discussion

Limitation and Future Work. As we use StyleGAN3, we argue that the underlying architecture may have limitations in representing high-frequency information. This becomes particularly evident when progressively enlarging the image. This limitation is consistent with observations from prior works [4, 29]. Furthermore, we've identified that the StyleGAN3 architecture leads to the emergence of a circular grid-like moire pattern. We will further investigate how to alleviate this issue in the future.

We also conduct an analysis of the constrained variety of poses and garments in the generated results, which can be found in the supplementary. To amplify the generative capabilities of the model and promote diversity in poses and garments, we anticipate that incorporating data augmentation techniques and more varied datasets could serve as the subsequent steps.

Conclusion. This work proposes an end-to-end training pipeline with the goal of orchestrating multi-source human-centric datasets with various distributions and scales into a full-body image space and achieving high-resolution human synthesis. The Multi-source Spatial Transformer, in particular, copes with articulated human structure, while the Continuous GAN module enables producing images at different resolutions. UnitedHuman breaks the technical barrier of being unable to generate high-fidelity human bodies in the absence of adequate HD full-body images and opens up new research directions to accelerate the process of human-body generation.

Acknowledgements. This study is supported by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

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 TOG*, 2021. 2

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

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 2, 4, 9

[4] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *ECCV*, 2022. 3, 4, 6, 7, 8, 9

[5] Xiyi Chen and Sergey Prokudin. Towards robust 3D body mesh inference of partially-observed humans. https://github.com/xiyichen/smplify-x-partial. 4

[6] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint*, arXiv:2204.11798, 2022. 3

[7] Jooyoung Choi, Jungbeom Lee, Yonghyun Jeong, and Sungroh Yoon. Toward spatially unbiased generative models. *arXiv preprint*, arXiv:2108.01285, 2021. 3

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-HD: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021. 3

[9] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[10] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. InsetGAN for full-body image generation. In *CVPR*, 2022. 2, 6, 7, 8

[11] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *ECCV*, 2022. 2, 3, 6, 7, 8

[12] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. DART: Articulated hand model with diverse accessories and rich textures. *arXiv preprint*, arXiv:2210.07650, 2022. 2, 3, 6, 8

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 4

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 3

[15] Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM TOG*, 2022. 2

[16] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 5

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 3

[19] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 4

[20] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019. 8

[21] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. InfinityGAN: Towards infinite-pixel image synthesis. *arXiv preprint*, arXiv:2104.03963, 2021. 3

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2, 3, 6, 8

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale CelebFaces attributes (CelebA) dataset. *Retrieved August*, 2018. 2, 3, 6, 8

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 4

[25] 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 *CVPR*, 2021. 2

[26] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, 2020. 3

[27] Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Radu Timofte, Martin Danelljan, and Luc Van Gool. Arbitrary-scale image synthesis. In *CVPR*, 2022. 3, 4

[28] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-xinggan. In *ECCV*, 2018. 2

[29] Haonan Qiu, Yuming Jiang, Hang Zhou, Wayne Wu, and Ziwei Liu. Stylefacev: Face video generation via decomposing and recomposing pretrained stylegan3. *arXiv preprint*, arXiv:2208.07862, 2022. 9

[30] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *CVPR*, 2020. 5

[31] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 1949. 5

[32] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020. 2

[33] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 3

[34] Qinghe Wang, Lijie Liu, Miao Hua, Qian He, Pengfei Zhu, Bing Cao, and Qinghua Hu. HS-Diffusion: Learning a semantic-guided diffusion model for head swapping. *arXiv preprint*, arXiv:2212.06458, 2022. 2

[35] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 3

[36] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in GANs. In *CVPR*, 2021. 3

[37] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *CVPR*, 2020. 3

[38] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, 2022. 2

[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 *ECCV*, 2022. 2

[40] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 2

[41] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019. 3