SemanticHuman-HD: High-Resolution Semantic Disentangled 3D Human Generation

Peng Zheng[1], Tao Liu[1], Zili Yi[3], and Rui Ma[1,2,⋆]

[1] School of Artificial Intelligence, Jilin University, Changchun, China
[2] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China
[3] School of Intelligence Science and Technology, Nanjing University, Suzhou, China

Abstract. With the development of neural radiance fields and generative models, numerous methods have been proposed for learning 3D human generation from 2D images. These methods allow control over the pose of the generated 3D human and enable rendering from different viewpoints. However, none of these methods explore semantic disentanglement in human image synthesis, i.e., they can not disentangle the generation of different semantic parts, such as the body, tops, and bottoms. Furthermore, existing methods are limited to synthesize images at 512^2 resolution due to the high computational cost of neural radiance fields. To address these limitations, we introduce SemanticHuman-HD, the first method to achieve semantic disentangled human image synthesis. Notably, SemanticHuman-HD is also the first method to achieve 3D-aware image synthesis at 1024^2 resolution, benefiting from our proposed 3D-aware super-resolution module. By leveraging the depth maps and semantic masks as guidance for the 3D-aware super-resolution, we significantly reduce the number of sampling points during volume rendering, thereby reducing the computational cost. Our comparative experiments demonstrate the superiority of our method. The effectiveness of each proposed component is also verified through ablation studies. Moreover, our method opens up exciting possibilities for various applications, including 3D garment generation, semantic-aware image synthesis, controllable image synthesis, and out-of-distribution image synthesis. Our project page is at `https://pengzheng0707.github.io/SemanticHuman-HD/`

Keywords: Generative models · 3D-aware human image synthesis · Compositional image synthesis

1 Introduction

Human image synthesis plays a crucial role in the field of artificial intelligence. This area holds significant potential for applications in virtual reality, virtual try-on, video games, and more. Traditional 2D generative models can only synthesize single-view images. On the other hand, recent advancements, such as the development of the neural radiance field (NeRF) [35], have led to a surge

⋆ Corresponding author

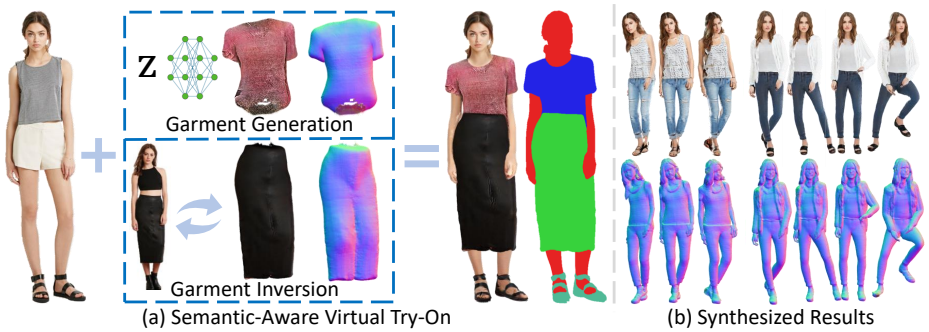

Fig. 1: (a) Semantic-aware virtual try-on. Given a real image, we first employ GAN inversion to obtain its semantic latent code. Subsequently, we replace the top and bottom garment by manipulating the semantic latent code. Here, the top is randomly generated by our model, and the bottom is disentangled from another GAN inversion result. (b) Controllable image synthesis. Our method allows for generating the same person in different poses as well as rendering them from different viewpoints.

of interest in 3D-aware image synthesis. These methods [5, 37, 52] allow precise control over the viewpoint of synthesized images. While many 3D generative models [5, 22, 33, 37, 42, 43] focus primarily on portrait synthesis, there is a growing body of work dedicated to full-body human image synthesis. However, none of the existing human image synthesis models fully address semantic disentanglement during generation.

To achieve semantic disentanglement, some methods [33, 60] employ K local 3D generators to model K NeRFs. Each NeRF corresponds to a specific semantic part in the synthesized image. In the case of CNeRF [33], each generator outputs the color, density and semantic value for a sampled point. The colors outputted by each generator are then weighted and summed, with the weights corresponding to semantic values. While this approach successfully disentangles colors, the geometry of different semantic parts remains entangled. This limitation arises because the densities output by each generator are simply summed. In contrast, 3D-SSGAN [31] effectively disentangles both texture and geometry. However, it maps 2D feature maps into 3D space, limiting its ability to model complex geometric structures such as full-body humans. Furthermore, the methods mentioned above are specifically designed for portrait synthesis and cannot be naively applied to full-body image synthesis. This limitation arises due to the intricate poses and geometries inherent in the human body. As for full-body image synthesis, AttriHuman-3D [54] proposes a framework for semantic-aware human image synthesis, in which decomposed feature planes corresponding to distinct semantic parts are generated using a single 2D generator. While such design of using one generator to generate all semantic parts makes AttriHuman-3D more efficient than previous methods [33, 60], entanglement between different semantic parts still exist from their results.

On the other hand, synthesizing high-resolution images using NeRF-based methods [35] poses challenges such as high computational cost, each pixel requires sampling numerous points for the accurate integration of colors. Some works [12, 31, 33, 57] employ a super-resolution module to circumvent direct rendering of high-resolution images. However, this strategy might impact the 3D consistency. Other works [1, 8, 19] propose more efficient way to render high-resolution (512^2) images without a super-resolution module. Nevertheless, this resolution may still not satisfy the demands of users, e.g., the need of 1024^2 images.

To address these issues, we propose SemanticHuman-HD, a novel method for high-resolution human image synthesis with semantic disentanglement. Unlike previous methods, our method generates each semantic part in a completely independent way. Specifically, we propose a two-stage training process. In the first stage, we synthesize human images, depth maps, semantic masks, and normal maps at 256^2 resolution. In the second stage, we employ a novel 3D-aware super-resolution module to synthesize 1024^2 resolution images. This module leverages the depth map and semantic mask synthesized in the first stage as guidance, significantly reducing the computational cost in volume rendering. To demonstrate the superiority of our method, we conduct quantitative and qualitative comparison experiments with state-of-the-art (SOTA) methods. Meanwhile, the effectiveness of each component proposed in this paper is verified in the ablation studies. In summary, our contributions are as follows:

1. We propose SemanticHuman-HD, the first method to achieve semantic disentanglement in 3D-aware human image synthesis. In our method, the underlying representation of each part is independent from other parts, leading to exciting applications such as 3D garment generation, semantic-aware virtual try-on, garment-level image editing and out-of-distribution image synthesis.
2. Leveraging our 3D-aware super-resolution module, SemanticHuman-HD attains 1024^2 resolution image synthesis. Importantly, our proposed super-resolution module preserves 3D consistency throughout the synthesis.
3. Comparing to SOTA human image synthesis methods, our SemanticHuman-HD demonstrates clear superiority in both quantitative measures (e.g., FID) and qualitative evaluation.

2 Related Work

2.1 3D-Aware Image Synthesis

Generative adversarial networks (GANs) [25–27] have demonstrated impressive results on image synthesis tasks. While certain GAN-based methods [39, 44] achieve pose control in image synthesis, they suffer from a lack of 3D-consistency due to their reliance on 2D feature representations. The advent of neural radiance fields (NeRF) [2,35,36,46,56] have opened the door to learn 3D-aware image synthesis from 2D image datasets. Numerous works [5,6,14,16,17,37,40,51,52] combine NeRF with GAN to achieve 3D-aware image synthesis. Notably, EG3D [5]

proposed a tri-plane representation as an efficient alternative to the computationally expensive point-based MLP. Beyond GAN-based generative models, diffusion models have gained prominence in recent years. Several diffusion model-based 3D-aware image synthesis methods [7, 21, 30, 34, 38, 47, 50] have emerged. However, these methods either are general models or only focus on portrait synthesis, making them incapable of synthesizing high quality human images.

2.2 3D-Aware Human Image Synthesis

3D-aware human image synthesis faces significant challenges, primarily because humans exhibit articulation and appear in diverse poses and clothing. gDNA [49] introduces a multi-subject forward skinning module for 3D human generation supervised by human scans. Some works [13, 15, 23, 53, 55, 57, 59] leverage human prior [4] and NeRF [35] to learn 3D-aware human image synthesis from 2D image datasets. AG3D [12] proposes to model the deformation of loose clothing using Fast-SNARF [9]. Additionally, it introduces a normal discriminator to improve geometric details in the generated results. EVA3D [19] adopts a compositional human NeRF representation for high-resolution (512^2) 3D-aware human image synthesis, all without relying on super-resolution modules. By leveraging vertex-based radiance fields, VeRi3D [8] allows local editing of generated results by replacing features at specified vertices. GSM [1] is an efficient framework for 3D human generation, which employs Gaussian shell maps to model feature volumes. Similar to VeRi3D, GSM also achieves local editing by a similar way. While some methods [20, 24, 28, 45] focus on 3D human generation using diffusion models, they struggle to synthesize photorealistic images due to limitations inherent in the diffusion model. Notably, none of these methods explore semantic-aware image synthesis. i.e., they cannot edit specific semantic parts of synthesized images while keeping other regions unchanged.

2.3 Semantic-Aware Image Synthesis

Some methods [10, 11, 22, 29, 31, 33, 41–43, 60] explore semantic-aware image synthesis, supervised by semantic masks. To translate a single-view semantic mask into a NeRF, Sem2NeRF [10] encodes the mask into latent code, controlling the 3D scene representation of a pre-trained decoder. Unlike Sem2NeRF, which requires a semantic mask as input, NeRFaceEditing [22] and IDE-3D [42] aim to achieve 3D-aware paired semantic mask and image synthesis by learning a semantic mask volume. Specifically for human image synthesis, 3D-SGAN [58] proposes a semantic-guided architecture comprising two generators: one for 3D-aware semantic mask synthesis and the other for translating the semantic mask into the corresponding image. Nevertheless, in these methods, different semantic parts are entangled during synthesis.

Several methods [31, 33, 41] have explored semantic disentangled synthesis. SemanticStyleGAN [41] uses K local generators to generate K semantic parts in synthesized image. These generators are supervised by paired portraits and

semantic masks. The design of K local generators ensures semantic disentanglement, enabling shape and texture changes in specific semantic regions while preserving others. CNeRF [33] and LC-NeRF [60] extend SemanticStyleGAN into the realm of 3D-aware image synthesis by learning compositional NeRFs. 3D-SSGAN [31] lifts the 2D generator into 3D space for efficiency and stronger disentanglement. Unfortunately, the aforementioned semantic disentangled methods can only be used in portrait synthesis due to the complexities of human poses.

For human image synthesis, both VeRi3D [8] and GSM [1] can achieve coarse-grained disentanglement during inference, as their features are closely tied to SMPL [4] vertices, although these vertices do not align with semantic masks. On the other hand, AttriHuman-3D [54] achieves semantic-aware human image synthesis, but it relies on a single generator to produce all semantic parts, lacking true semantic disentanglement. In summary, no existing method achieves 3D-aware human image synthesis with full semantic disentanglement. Furthermore, all these methods, whether utilizing the super-resolution module or not, can only synthesize 512^2 resolution images. In contrast, our proposed method achieves 3D-aware human image synthesis by independently generating each semantic part. Additionally, we introduce a 3D-aware super-resolution module, enabling the synthesis of 1024^2 resolution images.

3 Method

The overview of our method is depicted in Fig. 2, and it comprises two stages. In the first stage, given the human pose P and semantic label L_s, each generator G_k generates a tri-plane representation, which models one semantic part. The K tri-plane representations are further rendered into a 256^2 resolution image, depth map, semantic mask and normal map using the semantic renderer. Moving to the second stage, we feed the K tri-plane representations into the 3D-aware super-resolution module. This module enhances the resolution of each tri-plane representation, resulting in higher-quality outputs. Specifically, these refined tri-plane representations can be rendered into a 1024^2 resolution image, with the depth map and semantic mask serving as guidance. The details of each component are introduced below.

3.1 Semantic Disentangled Neural Radiance Field

Semantic Mapper: Given random noise z sampled from a Gaussian distribution, the Semantic Mapper maps it to latent code W, conditioned on human pose P and semantic label L_s. Here, the pose P corresponds to the parameter of SMPL [4], while semantic label L_s indicates whether the image contains specific types of garments, e.g., dress, skirt or hat. The latent code W is further extended into a semantic latent code W^+, where $W^+ = W^1 \times W^2 ... \times W^K$ and each W^k controls the generation of the k_{th} semantic part. In theory, K can take any value, but in our method, we set K to 6, corresponding to the body, tops, outer, bottoms, shoes and accessories, respectively. During training, we enforce

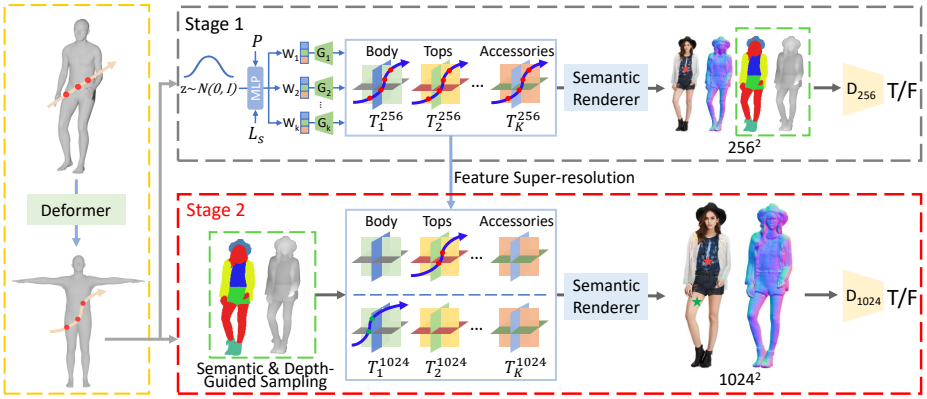

Fig. 2: Pipeline of SemanticHuman-HD. In stage 1, given random noise z, the Semantic Mapper maps it to K latent code W_k, conditioned on human pose P and semantic label L_s. Each local generator G_k then maps W_k into a tri-plane representation T_k^{256}. For each pixel in the synthesized image, we sample 72 points in posed space, which are subsequently deformed into canonical space using the deformer. These sampled points allow us to interpolate within the tri-plane representation, obtaining color and density information for each point. Finally, the Semantic Renderer renders the image, depth map, semantic mask, and normal map at 256^2 resolution. In stage 2, we employ a convolutional network to obtain high-resolution tri-plane representations, denoted as T_k^{1024}. To enhance efficiency, we significantly reduce the number of sampling points per pixel using semantic and depth-guided sampling. Ultimately, we render the image and normal map at 1024^2 resolution.

$W^1 = W^2 ... = W^K$ to ensure that different local generators generate consistent parts. For example, men typically do not wear skirts or dresses. During inference, simply modifying W^k allows us to edit the synthesized image in the k_{th} semantic part.

Local Generator: Similar to CNeRF [33], we employ K local 3D generators to model K semantic parts. However, unlike CNeRF, the generation of different semantic parts in our method are entirely independent. This independence is the key idea behind disentangling both geometry and texture. Conditioned on pose P and semantic label L_s, each local generator G_k maps the latent code W_k into a tri-plane representation. For each sampled point x, we calculate the density $\sigma(x)$, color $c(x)$, normal $n(x)$ and semantic value $s(x)$ as follows:

$$\sigma(x) = \sum_{k=1}^{K} (\sigma(x)^k), \;\; c(x) = \sum_{k=1}^{K} c(x)^k \times s(x)^k, \tag{1}$$

$$s(x) = \text{Concatenate}(s(x)^1, s(x)^2, ..., s(x)^K), \tag{2}$$

$$n(x) = \nabla_x(d(x)^{SMPL} + \sum_{k=1}^{K} \Delta d(x)^k), \tag{3}$$

$$s(x)^k = \frac{\sigma(x)^k}{\sum_{k=1}^{K} \sigma(x)^k},\ \sigma(x)^k = Sigmoid(d(x)^{SMPL} + \Delta d(x)^k). \tag{4}$$

Here, $\Delta d(x)^k$ and $c(x)^k$ are sampled from k_{th} tri-plane representation, while $d(x)^{SMPL}$ is sampled from the signed distance field (SDF) of canonical SMPL model [4]. Notably, as demonstrated in Eq. 4, we initially convert local SDF $\Delta d(x)^k$ to local density $\sigma(x)^k$ and then sum them to obtain the final density $\sigma(x)$. This approach differs from [33, 54, 60], which first sum local SDF $\Delta d(x)^k$ to obtain final SDF $d(x)$ and subsequently convert it to density $\sigma(x)$. This difference allows us to obtain local density $\sigma(x)^k$ and then map it into semantic value $s(x)^k$, thereby enabling the disentanglement of geometry across different semantic parts. For further details on this distinction, please refer to [33, 54, 60].

Semantic Renderer: Similar to NeRF [35], we cast a ray r for each pixel along its view direction v from camera center o: $\mathrm{r}(t) = \mathrm{o} + t\mathrm{v}$. The color C(r), semantic mask S(r), depth D(r) and normal N(r) of each ray r can be rendered as follows:

$$\Phi(\mathrm{r}) = \int_{t_n}^{t_f} T(t) \cdot \sigma(\mathrm{r}(t)) \cdot \phi(\mathrm{r}(t))dt, \tag{5}$$

$$\mathrm{D}(\mathrm{r}) = \int_{t_n}^{t_f} T(t) \cdot \sigma(\mathrm{r}(t)) \cdot tdt, \tag{6}$$

$$\text{where } T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathrm{r}(s))ds). \tag{7}$$

In the equations above, (Φ, ϕ) represents universal symbols, that can correspond to (C, c), (S, s) or (N, n). For clarity, we intentionally omitted details about the deformer, so in Eq. 5 and Eq. 6, each point $\mathrm{r}(t)$ is actually transformed from posed space to canonical space. For further information about the deformer, please refer to AG3D [12].

3.2 3D-Aware Super-Resolution Module

In stage 2, we train a 3D-aware super-resolution module to synthesize 1024^2 resolution images, building upon the K local generator pre-trained in stage 1. The core concept behind this module lies in leveraging semantic mask and depth map synthesized during stage 1 to significantly reduce the number of sampling points. Theoretically, our method achieves a remarkable reduction, i.e., from 432 to 11, where $432 = 72 \times 6$. Here, 72 represents 36 points for uniform sampling and 36 points for importance sampling, while 6 corresponds to the number of local generators. Even in the general case, without relying on our specialized semantic disentangled generation, this module can still reduce the sampling points from 72 to 11. The rationale behind this specific number will be explained below.

Depth-Guided Sampling: Given a 256^2 resolution depth image D_{origin}, we perform depth aggregation by considering neighboring pixels for each individual pixel. Depth aggregation serves the purpose of preventing incorrect depth values in regions with depth discontinuities when upsampling the depth maps. The formulation is as follows:

$$D(x,y,i) = \begin{cases} D_{origin}(x+\delta_x^i, y+\delta_y^i) & , \quad i \in \{1,...,9\} \\ D_{origin}(x,y) + \tau & , \quad i \in \{10\} \\ D_{origin}(x,y) - \tau & , \quad i \in \{11\}, \end{cases} \tag{8}$$

$$\delta_x^i = (i-1)/3 - 1, \ \delta_y^i = (i-1)\%3 - 1. \tag{9}$$

Here, $D(x,y,i)$ represents the i_{th} channel of depth value for pixel (x,y). For $i \in \{1,2,...,9\}$, $D(x,y,i)$ contains the depths of neighboring pixels. In other cases, it contains depths sampled around $D_{origin}(x,y)$, following Eq. 8. To enhance the resolution of D, we first sort the aggregated depth values for each pixel and then upsample them to 1024^2 resolution.

Semantic-Guided Sampling: In stage 2, we address the computational cost associated with having K local generators. Since our goal is to increase the resolution of the synthesized image while preserving its structure, we focus on the most important semantic part for each pixel. Specifically, we upsample the semantic mask synthesized by the generator (as described in Eq. 5) to 1024^2 resolution. By querying the weights of different semantic parts for each pixel based on the upsampled semantic mask, we can mask out those parts whose weights fall below the threshold δ.

3.3 Training

Loss Function: In stage 1, a low-resolution discriminator D_{256}, which comprises image discriminator D_{image}, semantic discriminator $D_{semantic}$, normal discriminator D_{normal} and face discriminator D_{face}, is employed to train K local generators. In stage 2, we freeze the K generators and focus on training the 3D-aware super-resolution module, and only one high-resolution image discriminator D_{1024} is used in this stage. To ensure consistency between I_{256} and I_{1024}, we introduce an upsample loss term: $\mathcal{L}_{upsample} = \|\text{Downsample}(I_{1024}) - I_{256}\|$. The loss functions for both stages are as follows:

$$\mathcal{L}_1 = \mathcal{L}_{256} + \mathcal{L}_{AG3D}, \ \mathcal{L}_2 = \mathcal{L}_{1024} + \mathcal{L}_{upsample} + \mathcal{L}_{AG3D}, \tag{10}$$

where $\mathcal{L}_{256}$ and $\mathcal{L}_{1024}$ represent GAN loss for D_{256} and D_{1024} respectively. $\mathcal{L}_{AG3D}$ is a loss function adopted from AG3D [12].

Implementation Details: The models are trained on 4 NVIDIA A40 GPUs for 9 days. The training in stage 1 takes 6 days, and stage 2 takes 3 days. Additionally, all the experiments mentioned in the paper were also conducted using the A40. During the training in stage 1, the normal discriminator is used only for the last 3 days of training.

4 Experiments

Datasets: The DeepFashion dataset [32] comprises 12,701 pairs of human images and corresponding semantic masks. For our training and evaluation, we utilize 8,037 pairs from this dataset. During training, we leverage a pre-trained model [48] to obtain normal maps, and we convert the semantic masks from 24 categories into 6 simplified categories. For instance, dresses and rompers are grouped under the broader category of "tops". Additionally, the SMPL [4] parameters for each image are provided by AG3D [12]. All experiments in this paper are conducted using the DeepFashion dataset, and all models are trained on this dataset.

Metrics: We evaluate our method using two key metrics: Frechet Inception Distance (FID) [18] and Kernel Inception Distance (KID) [3]. These metrics assess the diversity and quality of synthesized images by measuring their similarity to real images. Given that different methods yield relatively random results, all FID and KID scores reported in this paper are based on 50,000 synthesized images to ensure fair comparison.

Baselines: To demonstrate the superiority of our method, we compare against several baselines: AG3D [12] is a SOTA method for 3D-aware human image synthesis. EVA3D [19] and GSM [1] achieve 512^2 resolution image synthesis without a super-resolution module. AttriHuman-3D [54] enables semantic-aware generation and VeRi3D [8] allows local editing of synthesized images. Other works [23, 53, 55, 57] have not released their training code or pre-trained models, so we do not include them in our comparison.

4.1 Comparison

Human Image synthesis: Table 1 presents quantitative comparisons between our method and several SOTA methods. Our method consistently produces superior image quality at both 512^2 and 1024^2 resolutions. While AG3D [12] ranks as the second-best method in terms of image quality, it relies on a 2D super-resolution module, which compromises 3D consistency. EVA3D [19], capable of synthesizing 512^2 resolution images without a super-resolution module, unfortunately suffers from circular artifacts due to its network design. On the other hand, GSM [1] and VeRi3D [8] allow local editing, but their lack of semantic awareness results in artifacts. AttriHuman-3D [54], while enabling semantic-aware synthesis, entangles different semantic parts during generation. In summary, our method uniquely achieves a comprehensive set of capabilities: local editing, semantic disentangled synthesis, and 3D garment generation. Moreover, leveraging our 3D-aware super-resolution module, we stand out as the sole method capable of synthesizing 1024^2 resolution images, as demonstrated qualitatively in Fig. 3.

Table 1: Quantitative comparisons. * denotes the use of a super-resolution module that is not 3D-aware. Some results are marked with *, indicating that these results are quoted from other papers because the authors did not release their training code or pre-trained model. A: Local editing. B: Semantic-aware synthesis. C: Semantic disentangled synthesis. D: 3D Garment generation.

Method	A	B	C	D	Resolution	FID↓	1000×KID↓
AG3D	✘	✘	✘	✘	512*	11.33	5.75
EVA3D	✘	✘	✘	✘	512	15.89	9.25
GSM	✔	✘	✘	✘	512	15.78*	/
VeRi3D	✔	✘	✘	✘	512	21.4*	/
AttriHuman-3D	✔	✔	✘	✘	512*	16.85*	/
SemanticHuman-HD	✔	✔	✔	✔	512	10.04	5.02
					1024	**8.70**	**4.04**

Fig. 3: Qualitative comparison. To better assess the detailed quality of the generated results, we zoom in on the face and clothing areas in the synthesized images. Notably, the image synthesized by our method is at 1024^2 resolution, whereas the results from other methods are only at 512^2 resolution.

Local Editing: Both AttriHuman-3D [54] and our method exhibit semantic awareness. However, AttriHuman-3D employs a single generator to generate triplane representations corresponding to different semantic parts. Unfortunately, this design entangles different semantic parts. Consequently, when editing a specific semantic part, it may not seamlessly match other regions. GSM [1] and VeRi3D [8] allow local editing by manipulating features of specified vertices. Although SMPL [4] provides category labels for these vertices, the resulting edits lack semantic awareness, leading to suboptimal image quality. A comparison of the editing capabilities across different methods is illustrated in Fig. 4. For additional evidence of semantic disentangled synthesis, refer to Fig. 5.

Computational Efficiency: Our proposed super-resolution module significantly reduces the number of required sampling points, thereby minimizing computational costs. Comparative experiments on computational resources, specifically GPU memory usage during training, are detailed in Table 2. The results clearly demonstrate that our method outperforms other methods in terms of computational efficiency. As a consequence, we can successfully synthesize 1024^2 resolution images—a feat that other methods may struggle to achieve.

Fig. 4: Comparison for local editing. For each edited image, we zoom in on key areas to demonstrate the editing capabilities.

Fig. 5: Semantic disentangled image synthesis. By modifying the latent code of a specified semantic part, we can alter that specified part in the synthesized image.

4.2 Ablation Study

3D-Aware Super-Resolution Module: Our proposed super-resolution module enhances the resolution of synthesized images from 256^2 to 1024^2 while preserving 3D consistency. Quantitative results in Table 3 demonstrate that this super-resolution module significantly improves the quality of synthesized images. Qualitative results in Fig. 6 showcase the effectiveness of this module.

Depth Aggregation: Within our super-resolution module, we introduce depth aggregation to address discontinuities in depth maps. Consider two adjacent regions with depth values around 1 and 5, respectively. Direct upsampling would yield a depth value of 3 at the boundary. Obviously, the pixels on the boundary should belong to one of the two parts, so the correct depth value should align with either 1 or 5. Our solution involves aggregating the depth values of neighboring points, ensuring that both depth values (1 and 5) are preserved in our aggregated depth map, thereby facilitating accurate rendering. The qualitative results in Fig. 6 validate the effectiveness of depth aggregation.

Upsample Loss: During super-resolution module training, we employ an upsample loss to ensure consistency between the original image and the image after super-resolution. Quantitative results reported in Table 3 confirm the effectiveness of this upsample loss.

Table 2: Efficiency comparisons. EVA3D [19] and AG3D [12] are unable to synthesize 1024^2 resolution images.

Methods	Resolution	Mem
EVA3D	512	34G
AG3D	512	21G
Ours	512	10G
Ours	1024	31G

Table 3: Quantitative ablation studies. The results validate the effectiveness of our proposed components. SR: Super-Resolution. DA: Depth Aggregation. UL: Upsample Loss.

Methods	Resolution	FID ↓	1000×KID↓
w/o SR	256	13.47	9.13
w/o DA	1024	9.38	4.56
w/o UL	1024	13.52	8.18
Ours	1024	**8.70**	**4.04**

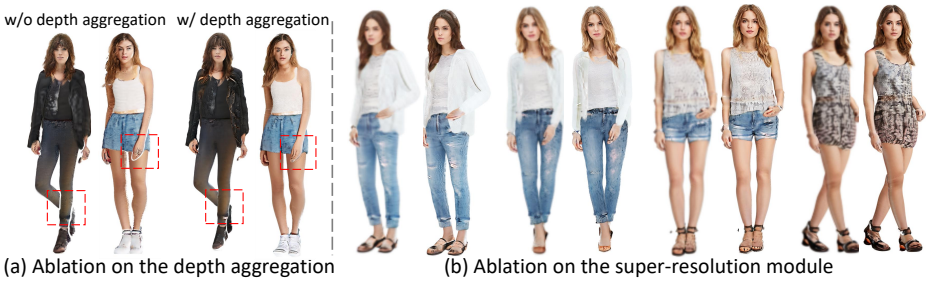

Fig. 6: Qualitative ablation studies. (a) The results on the left are synthesized by the model that dose not use depth aggregation, while the images on the right are synthesized by the opposite approach. (b) For each paired set of images, the original image from stage 1 is on the left, and the image after super-resolution is on the right.

4.3 Applications

Our method can achieve many interesting applications, some of which are showcased below. Additional results are provided in the supplementary material.

Semantic-Aware Virtual Try-On: To further demonstrate the capabilities of our method, we combine GAN inversion to achieve semantic-aware virtual try-on, as shown in Fig. 1. Notably, our method can disentangle specific garments from the results obtained by GAN inversion and even place these garments into new images. As far as we know, we are the only method that achieves this, leading to a special application: you can combine the bottoms from one image and randomly generated tops with your own image, all while controlling the pose and viewpoint of the newly synthesized image.

3D Garment Generation: Our method disentangles both geometry and texture, enabling 3D garment generation. Specifically, by setting the density of other semantic parts (excluding the specified part) to 0, we obtain the generation of specific items such as dresses, shoes, and hats. Refer to Fig. 7 for results.

Fig. 7: Garment generation. We independently generate 3D garments by setting the density of other semantic parts (excluding the specified part) to 0. The corresponding normal map for each synthesized image is also displayed to demonstrate the geometric quality of the results. (a) Results randomly generated by our model. (b) Results obtained from GAN inversion.

Out-of-distribution Image Synthesis: Leveraging our disentangled synthesis, we can create out-of-distribution images (e.g., think of a man wearing a dress) by manipulating the semantic latent code. Fig. 8 showcases some intriguing out-of-distribution image synthesis results that defy typical dataset or daily life representations.

Controllable Image Synthesis: The synthesis of our method is conditioned on pose P and semantic label L_s. The results of conditional image synthesis are shown in Fig. 9. Furthermore, our method allows to control the pose of the generated 3D human and render it from various viewpoints, as shown in Fig. 1.

5 Limitation

Our method faces certain limitations, which we discuss below. Dataset constraints: the quality of synthesized results suffer when dealing with poses or viewpoints that are rarely encountered in the dataset. Unfortunately, optimizing the network alone does not fully address this issue. To overcome this limitation, we require higher-quality datasets. Challenges with 2D supervision: while obtaining 2D human images is relatively easy, training a model only based on 2D images proves challenging when aiming for results with accurate geometries. A potential solution could involve training a model supervised by both 3D human models and 2D images, leveraging complementary information from both domains. Hand generation challenges: existing methods struggle with hand generation, and achieving realistic hand deformations remains difficult. Addressing this limitation is an ongoing area of research.

Fig. 8: Out-of-distribution image synthesis. To achieve out-of-distribution image synthesis, we assign different semantic labels to various semantic parts, e.g., if we set the semantic label corresponding to the body as "male", and the label corresponding to the tops as "dress", we can synthesize an image of a man wearing a dress.

Fig. 9: Conditional image synthesis. For each paired set of images, the image on the right is synthesized conditioned on the semantic label L_s and human pose P of the real image on the left.

6 Conclusion

In this paper, we introduce SemanticHuman-HD, the pioneering method for achieving semantic disentangled human image synthesis. By leveraging our proposed 3D-aware super-resolution module, our method is also the first to successfully synthesize images at an impressive 1024^2 resolution. Notably, the proposed 3D-aware super-resolution can be easily employed in NeRF-based generative models. Our experiments consistently demonstrate the superiority of our proposed method. Furthermore, we showcase a range of interesting applications, including 3D garment generation, semantic-aware virtual try-on, controllable image synthesis, garment-level image editing and out-of-distribution image synthesis. Looking ahead, we will consider addressing the limitations mentioned in Section 5: dataset constraints, challenges with 2D supervision, and hand generation challenges. Specifically, we are particularly focused on overcoming the challenges related to 2D supervision. A generative model capable of achieving high-quality geometric details in 3D human generation is highly needed in the areas of virtual reality, video games and beyond.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62202199).

References

1. Abdal, R., Yifan, W., Shi, Z., Xu, Y., Po, R., Kuang, Z., Chen, Q., Yeung, D.Y., Wetzstein, G.: Gaussian shell maps for efficient 3d human generation. arXiv preprint arXiv:2311.17857 (2023)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
6. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5799–5809 (2021)
7. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
8. Chen, X., Huang, J., Bin, Y., Yu, L., Liao, Y.: Veri3d: Generative vertex-based radiance fields for 3d controllable human image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8986–8997 (2023)
9. Chen, X., Jiang, T., Song, J., Rietmann, M., Geiger, A., Black, M.J., Hilliges, O.: Fast-snarf: A fast deformer for articulated neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
10. Chen, Y., Wu, Q., Zheng, C., Cham, T.J., Cai, J.: Sem2nerf: Converting single-view semantic masks to neural radiance fields. In: European Conference on Computer Vision. pp. 730–748. Springer (2022)
11. Deng, K., Yang, G., Ramanan, D., Zhu, J.Y.: 3d-aware conditional image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4434–4445 (2023)
12. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: Ag3d: Learning to generate 3d avatars from 2d image collections. arXiv preprint arXiv:2305.02312 (2023)
13. Fu, T.J., Xiong, W., Nie, Y., Liu, J., Oğuz, B., Wang, W.Y.: Text-guided 3d human generation from 2d collections. arXiv preprint arXiv:2305.14312 (2023)

14. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems **35**, 31841–31854 (2022)
15. Grigorev, A., Iskakov, K., Ianina, A., Bashirov, R., Zakharkin, I., Vakhitov, A., Lempitsky, V.: Stylepeople: A generative model of fullbody human avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5151–5160 (2021)
16. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. arXiv preprint arXiv:2110.08985 (2021)
17. He, H., Yang, Z., Li, S., Dai, B., Wu, W.: Orthoplanes: A novel representation for better 3d-awareness of gans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22996–23007 (2023)
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
19. Hong, F., Chen, Z., Lan, Y., Pan, L., Liu, Z.: Eva3d: Compositional 3d human generation from 2d image collections. arXiv preprint arXiv:2210.04888 (2022)
20. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. arXiv preprint arXiv:2205.08535 (2022)
21. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
22. Jiang, K., Chen, S.Y., Liu, F.L., Fu, H., Gao, L.: Nerffaceediting: Disentangled face editing in neural radiance fields. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
23. Jiang, S., Jiang, H., Wang, Z., Luo, H., Chen, W., Xu, L.: Humangen: Generating human radiance fields with explicit priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12543–12554 (2023)
24. Jiang, S., Luo, H., Jiang, H., Wang, Z., Yu, J., Xu, L.: Mvhuman: Tailoring 2d diffusion with multi-view sampling for realistic 3d human generation. arXiv preprint arXiv:2312.10120 (2023)
25. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021)
26. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
27. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
28. Kolotouros, N., Alldieck, T., Zanfir, A., Bazavan, E., Fieraru, M., Sminchisescu, C.: Dreamhuman: Animatable 3d avatars from text. Advances in Neural Information Processing Systems **36** (2024)
29. Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8300–8311 (2021)

30. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
31. Liu, R., Zheng, P., Wang, Y., Ma, R.: 3d-ssgan: Lifting 2d semantics for 3d-aware compositional portrait synthesis. arXiv preprint arXiv:2401.03764 (2024)
32. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016)
33. Ma, T., Li, B., He, Q., Dong, J., Tan, T.: Semantic 3d-aware portrait synthesis and manipulation based on compositional neural radiance field. arXiv preprint arXiv:2302.01579 (2023)
34. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023)
35. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
36. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5589–5599 (2021)
37. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13503–13513 (2022)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
39. Sarkar, K., Golyanik, V., Liu, L., Theobalt, C.: Style and pose control for image synthesis of humans from a single monocular view. arXiv preprint arXiv:2102.11263 (2021)
40. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems **33**, 20154–20166 (2020)
41. Shi, Y., Yang, X., Wan, Y., Shen, X.: Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11254–11264 (2022)
42. Sun, J., Wang, X., Shi, Y., Wang, L., Wang, J., Liu, Y.: Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. ACM Transactions on Graphics (ToG) **41**(6), 1–10 (2022)
43. Sun, J., Wang, X., Zhang, Y., Li, X., Zhang, Q., Liu, Y., Wang, J.: Fenerf: Face editing in neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7672–7682 (2022)
44. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6142–6151 (2020)
45. Wang, J., Liu, Y., Dou, Z., Yu, Z., Liang, Y., Li, X., Wang, W., Xie, R., Song, L.: Disentangled clothed avatar generation from text descriptions. arXiv preprint arXiv:2312.05295 (2023)

46. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
47. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36** (2024)
48. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022)
49. Xu, C., Jiang, T., Song, J., Yang, J., Black, M.J., Geiger, A., Hilliges, O.: gdna: Towards generative detailed neural avatars. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society (2022)
50. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20908–20918 (2023)
51. Xu, X., Pan, X., Lin, D., Dai, B.: Generative occupancy fields for 3d surface-aware image synthesis. Advances in Neural Information Processing Systems **34**, 20683–20695 (2021)
52. Xu, Y., Peng, S., Yang, C., Shen, Y., Zhou, B.: 3d-aware image synthesis via learning structural and textural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18430–18439 (2022)
53. Xu, Z., Zhang, J., Liew, J.H., Feng, J., Shou, M.Z.: Xagen: 3d expressive human avatars generation. arXiv preprint arXiv:2311.13574 (2023)
54. Yang, F., Chen, T., He, X., Cai, Z., Yang, L., Wu, S., Lin, G.: Attrihuman-3d: Editable 3d human avatar generation with attribute decomposition and indexing. arXiv preprint arXiv:2312.02209 (2023)
55. Yang, Z., Li, S., Wu, W., Dai, B.: 3dhumangan: Towards photo-realistic 3d-aware human image generation. arXiv preprint arXiv:2212.07378 (2022)
56. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems **34**, 4805–4815 (2021)
57. Zhang, J., Jiang, Z., Yang, D., Xu, H., Shi, Y., Song, G., Xu, Z., Wang, X., Feng, J.: Avatargen: a 3d generative model for animatable human avatars. In: European Conference on Computer Vision. pp. 668–685. Springer (2022)
58. Zhang, J., Sangineto, E., Tang, H., Siarohin, A., Zhong, Z., Sebe, N., Wang, W.: 3d-aware semantic-guided generative model for human synthesis. In: European Conference on Computer Vision. pp. 339–356. Springer (2022)
59. Zhang, X., Zhang, J., Chacko, R., Xu, H., Song, G., Yang, Y., Feng, J.: Getavatar: Generative textured meshes for animatable human avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2273–2282 (2023)
60. Zhou, W., Yuan, L., Chen, S., Gao, L., Hu, S.: Lc-nerf: Local controllable face generation in neural randiance field. arXiv preprint arXiv:2302.09486 (2023)