Text-guided 3D Human Generation from 2D Collections

Tsu-Jui Fu[1], **Wenhan Xiong**[2], **Yixin Nie**[2]
Jingyu Liu[2], **Barlas Oğuz**[2], **William Yang Wang**[1]

[1]UC Santa Barbara [2]Meta

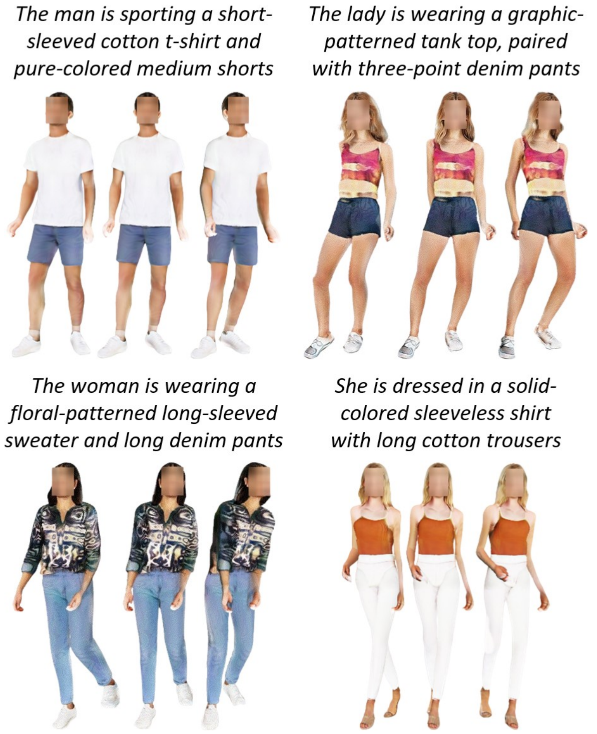
{tsu-juifu, william}@cs.ucsb.edu {xwhan, ynie, jingyuliu, barlaso}@meta.com

Abstract

3D human modeling has been widely used for engaging interaction in gaming, film, and animation. The customization of these characters is crucial for creativity and scalability, which highlights the importance of controllability. In this work, we introduce Text-guided 3D Human Generation (T3H), where a model is to generate a 3D human, guided by the fashion description. There are two goals: 1) the 3D human should render articulately, and 2) its outfit is controlled by the given text. To address this T3H task, we propose Compositional Cross-modal Human (CCH). CCH adopts cross-modal attention to fuse compositional human rendering with the extracted fashion semantics. Each human body part perceives relevant textual guidance as its visual patterns. We incorporate the human prior and semantic discrimination to enhance 3D geometry transformation and fine-grained consistency, enabling this to learn from 2D collections for data efficiency. We conduct evaluations on DeepFashion and SHHQ with diverse fashion attributes covering the shape, fabric, and color of upper and lower clothing. Extensive experiments demonstrate that CCH achieves superior results for T3H with high efficiency.

1 Introduction

Our world is inherently three-dimensional, where this nature highlights the importance of 3D applications in various fields, including architecture, product design, and scientific simulation. The capability of 3D content generation helps bridge the gap between physical and virtual domains, providing an engaging interaction within digital media. Furthermore, realistic 3D humans have vast practical value, especially in gaming, film, and animation. Despite enhancing the user experience, the customization of the character is crucial for creativity and scalability. Language is the most direct way of communication. If a system follows the description and establishes the 3D human model, it will significantly improve controllability and meet the considerable demand.



The man is sporting a short-sleeved cotton t-shirt and pure-colored medium shorts

The lady is wearing a graphic-patterned tank top, paired with three-point denim pants

The woman is wearing a floral-patterned long-sleeved sweater and long denim pants

She is dressed in a solid-colored sleeveless shirt with long cotton trousers

Figure 1: Text-guided 3D Human Generation (T3H).

We thus introduce Text-guided 3D Human Generation (T3H) to generate a 3D human with the customized outfit, guided via the fashion description. Previous works (Kolotouros et al., 2019; Gao et al., 2022) depend on multi-view videos to learn 3D human modeling, but these data are difficult to obtain and are not language controllable. Text-to-3D (Jain et al., 2022; Poole et al., 2023) has shown attractive 3D generation results through the success of neural rendering (Mildenhall et al., 2020). However, these methods apply iterative inference optimization by external guidance, which is inefficient for usage.

To tackle these above issues, we propose Compositional Cross-modal Human (CCH) to learn T3H from 2D collections. CCH divides the human body into different parts and employs individual volume rendering, inspired by EVA3D (Hong et al., 2023).

Project website: https://text-3dh.github.io

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4508–4520
December 6-10, 2023 ©2023 Association for Computational Linguistics

We extract the fashion semantics from the description and adopt cross-modal attention to fuse body volumes with textual features, where each part can learn to perceive its correlated fashion patterns. To support various angles of view, CCH leverages the human prior (Bogo et al., 2016) to guide the geometry transformation for concrete human architecture. Then these compositional volumes can jointly render a 3D human with the desired fashion efficiently. The semantic discrimination further considers compositional distinguishment over each human part, which improves the fine-grained alignment with its description through adversarial training.

We perform experiments on DeepFashion (Liu et al., 2016; Jiang et al., 2022) and SHHQ (Fu et al., 2022a), which contain human images with diverse fashion descriptions. The patterns include various types of shapes (*sleeveless*, *medium short*, *long*, etc.), fabrics (*denim*, *cotton*, *furry*, etc.), and colors (*floral*, *graphic*, *pure color*, etc.) for the upper and lower clothing. To study the performance of T3H, we conduct a thorough evaluation from both visual and semantic aspects. We treat overall realism, geometry measure, and pose correctness to assess the quality of generated 3D humans. For the alignment with the assigned fashion, we apply text-visual relevance from CLIP and fine-grained accuracy by a trained fashion classifier.

The experiments indicate that language is necessary to make 3D human generation controllable. Our proposed CCH adopts cross-modal attention to fuse compositional neural rendering with textual fashion as 3D humans, and semantic discrimination further helps fine-grained consistency. In summary, our contributions are three-fold:

- We introduce T3H to control 3D human generation via fashion description, learning from collections of 2D images.
- We propose CCH to extract fashion semantics in the text and fuse with 3D rendering in one shot, leading to an effective and efficient T3H.
- Extensive experiments show that CCH exhibits sharp 3D humans with clear textual-related fashion patterns. We advocate that T3H can become a new field of vision-and-language research.

2 Related Work

Text-guided Visual Generation. Using human-understandable language to guide visual generation can enhance controllability and benefit creative visual design. Previous works built upon adversarial training (Goodfellow et al., 2015; Reed et al., 2016) to produce images (Xu et al., 2018; El-Nouby et al., 2019; Fu et al., 2020, 2022c) or videos (Marwah et al., 2017; Li et al., 2018b; Fu et al., 2022b) conditioned on given descriptions. With sequential modeling from Transformer, vector quantization (Esser et al., 2021) can generate high-quality visual content as discrete tokens (Ramesh et al., 2021; Ding et al., 2021; Fu et al., 2023). The denoising diffusion framework (Ho et al., 2020; Ramesh et al., 2022; Saharia et al., 2022; Feng et al., 2023) gains much attention as its diversity and scalability via large-scale text-visual pre-training. Beyond images and videos, 3D content creation is more challenging due to the increasing complexity of the depth dimension and spatial consistency. In this paper, we consider text-guided 3D human generation (T3H), which has vast applications in animated characters and virtual assistants.

3D Generation. Different representations have been explored for 3D shapes, such as mesh (Gao et al., 2019; Nash et al., 2020; Henderson et al., 2020), voxel grid (Tatarchenko et al., 2017; Li et al., 2017), point cloud (Li et al., 2018a; Yang et al., 2019; Luo et al., 2021), and implicit field (Chen and Zhang, 2019; Park et al., 2019; Zheng et al., 2022). Neural Radiance Field (NeRF) (Mildenhall et al., 2020; Barron et al., 2022; Muller et al., 2022) has shown remarkable results in novel view synthesis (Schwarz et al., 2021; Chan et al., 2021; Skorokhodov et al., 2023) and 3D reconstruction (Yariv et al., 2021; Zhang et al., 2021). With the differentiable neural rendering, NeRF can be guided by various objectives. Text-to-3D draws appreciable attraction these days, which adopts external text-visual alignments (Jain et al., 2022; Khalid et al., 2022; Michel et al., 2022; Wang et al., 2022a; Hong et al., 2022) and pre-trained text-to-image (Wang et al., 2022b; Nam et al., 2022; Poole et al., 2023; Metzer et al., 2023; Tang et al., 2023; Seo et al., 2023). However, existing methods take numerous iterations to optimize a NeRF model, which is time-consuming for practical usage. Our CCH learns to extract fashion semantics with NeRF rendering and incorporates the human prior for a concrete human body, achieving effective and efficient T3H.

3D Human Representation. To reconstruct a 3D human, early works (Collet et al., 2015; Guo et al., 2019; Su et al., 2020) count on off-the-shelf tools to predict the camera depth. As mitigating the costly hardware requirement, they estimate a 3D human

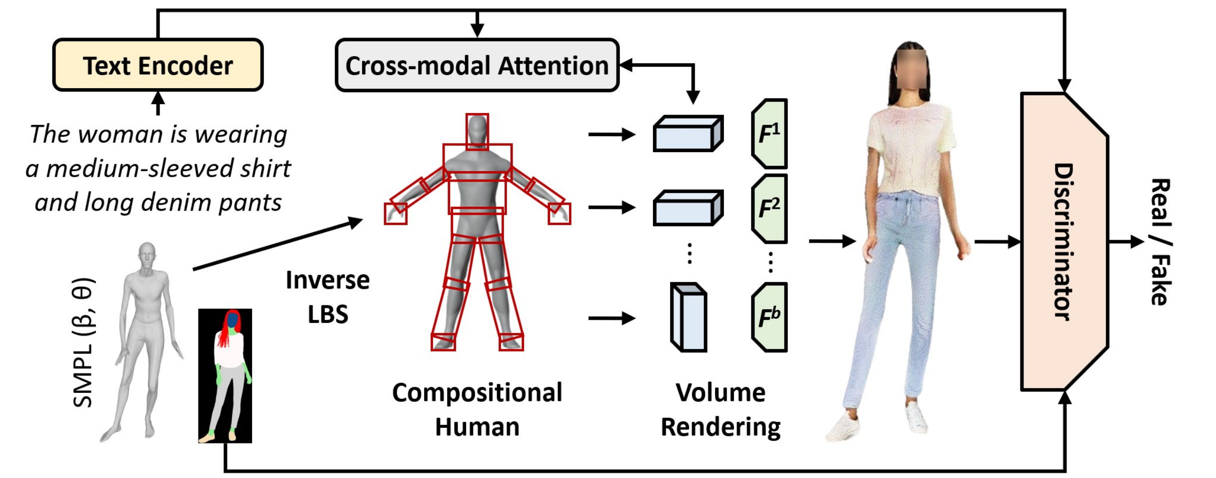


Figure 2: Compositional Cross-modal Human (CCH). CCH extracts fashion semantics from the description and adopts cross-modal attention in compositional body volumes for controllable 3D human rendering. The human prior (SMPL) provides robust geometry transformation, enabling CCH to learn from 2D collections for data efficiency. The semantic discrimination further helps find-grained consistency through adversarial training.

texture (Xu and Loy, 2021; Gomes et al., 2022) via the UV mapping (Shysheya et al., 2019; Yoon et al., 2021). With the promising success of NeRF, recent works (Peng et al., 2021b,a; Su et al., 2021) adopt volume rendering for 3D humans from multi-view videos (Weng et al., 2022; Chen et al., 2022). Since the data are difficult to collect, the 3D-aware generation (Chan et al., 2022; Gu et al., 2022; Noguchi et al., 2022) learns 3D modeling from the collection of human images (Yang et al., 2022; Hong et al., 2023). In place of arbitrary outputs, we introduce the first controllable 3D human generation that also learns from a 2D collection, and the presented fashion patterns should align with the description.

3 Text-guided 3D Human Generation

3.1 Task Definition

We present text-guided 3D human generation (T3H) to create 3D humans via fashion descriptions. For data efficiency, a 2D collection $\mathcal{D}$=$\{\mathcal{V}, \mathcal{T}\}$ is provided, where $\mathcal{V}$ is the human image, and $\mathcal{T}$ is its fashion description. Our goal is to learn the neural rendering that maps $\mathcal{T}$ into an articulate 3D human with the fashion patterns of $\mathcal{V}$.

3.2 Background

Neural Radiance Field (**NeRF**) (Mildenhall et al., 2020) defines implicit 3D as $\{c, \sigma\}$=$F(x, d)$. The query point x in the viewing direction d holds the emitted radiance c and the volume density σ. To get the RGB value $C(r)$ of certain rays $r(t)$, volume rendering is calculated along a ray r from the near bound t_n to the far bound t_f:

$$T(t) = \exp(-\int_{t_\text{n}}^{t} \sigma(r(s))ds),$$
$$C(r) = \int_{t_\text{n}}^{t_\text{f}} T(t)\sigma(r(t))c(r(t), d)dt, \qquad (1)$$

where $T(t)$ stands for their accumulated transmittance. StyleSDF (Or-El et al., 2022) then replaces σ with single distance field (SDF) $d(x)$ for a better surface, where $\sigma(x) = \alpha^{-1}\text{sigmoid}(\frac{-d(x)}{\alpha})$ and α is a learnable scalar that controls the tightness of the density around the surface boundary.

SMPL (Bogo et al., 2016) defines the human body as $\{\beta, \theta\}$, where $\beta \in \mathbb{R}^{10}$ and $\theta \in \mathbb{R}^{3\times 23}$ control its shape and pose. We consider Linear Blend Skinning (LBS) as the transformation from the canonical into the observation space for the point x to $\sum_{k=1}^{K} h_k H_k(\theta, J)x$, where $h_k \in \mathbb{R}$ is the scalar of the blend weight and $H_k \in \mathbb{R}^{4\times 4}$ is the transformation matrix of the kth joint. Inverse LBS transforms the observation back to the canonical space as a similar equation but with an inverted H.

3.3 Compositional Cross-modal Human

Following EVA3D (Hong et al., 2023), we split the human body into 16 parts. As shown in Fig. 2, each body part holds its own bounding box $\{o^b_\text{min}, o^b_\text{max}\}$. To leverage the human prior for a target pose θ, we transform these pre-defined bounding boxes with SMPL's transformation matrices H_k. Ray $r(t)$ is sampled for each pixel on the canvas. For a ray that intersects bounding boxes, we pick up its near and far bounds (t_n and t_f) and sample N points as follows: $t_i \sim \mathcal{U}\left[t_\text{n} + \frac{i-1}{N}(t_\text{f} - t_\text{n}), t_\text{n} + \frac{i}{N}(t_\text{f} - t_\text{n})\right]$.

We then transform these sampled points back to the canonical space with inverse LBS. For shape generalization, we consider not only pose transformation but also blend shapes ($B^P(\theta)$ and $B^S(\beta)$) (Zheng et al., 2021). $\mathbb{N}$ contains K nearest vertices v of the target SMPL mesh for the sample point ray $r(t_i)$:

$$
\begin{aligned}
g_k &= \frac{1}{||r(t_i) - v_k||}, \\
M_k &= \left(\sum_{k=1}^{K} g_k H_k\right) \begin{bmatrix} I & B_k^P + B_k^S \\ 0 & I \end{bmatrix}, \\
\begin{bmatrix} x_i \\ 1 \end{bmatrix} &= \sum_{v_k \in \mathbb{N}} \frac{g_k}{\sum_{v_k \in \mathbb{N}} g_k} (M_k)^{-1} \begin{bmatrix} r(t_i) \\ 1 \end{bmatrix},
\end{aligned} \tag{2}
$$

where $g_k \in \mathbb{R}$ is the inverse weight of the vertex v_k and $M_k \in \mathbb{R}^{4\times4}$ is the transformation matrix. The x_i can be used for further volume rendering.

Cross-modal Attention. During rendering, if the canonical point x_i with the viewing direction d_i is inside the bth bounding box, it will be treated as:

$$
\begin{aligned}
\hat{x}_i^b &= \frac{2x_i - (o_{\text{max}}^b + o_{\text{min}}^b)}{o_{\text{max}}^b - o_{\text{min}}^b}, \\
f_i^b &= \text{Linear}(\hat{x}_i^b, d_i),
\end{aligned} \tag{3}
$$

where a linear mapping is applied to acquire preliminary features $f^b \in \mathbb{R}^{16\times8\times128}$. To exhibit the desired fashion in the final rendering, we extract the word features by the text encoder as $\{w_l \in \mathbb{R}^{512}\}$ from $\mathcal{T}$. We then fuse the textual features with f_i^k via cross-modal attention:

$$
\begin{aligned}
p_l &= \frac{\exp(f_i^b\, W^b\, w_l^{\text{T}})}{\sum_{\iota=1}^{L} \exp(f_i^b\, W^b\, w_\iota^{\text{T}})}, \\
\text{CA}(f_i^b \mid \{w\}) &= \sum_{l=1}^{L} p_l w_l,
\end{aligned} \tag{4}
$$

where L is the length of $\mathcal{T}$ and W^b is the learnable matrix. In this way, each point can learn to perceive relevant textual guidance for the bth human body part and depict corresponding fashion patterns.

Each body part has its individual volume rendering F^b, which consists of stacked multilayer perceptrons (MLPs) with the SIREN activation (Sitzmann et al., 2020). Since the point x_i may fall into multiple boxes $\mathbb{B}_i$, we follow EVA3D to apply the mixture function (Lombardi et al., 2021):

$$
\begin{aligned}
\{c_i^b, \sigma_i^b\} &= F^b(\text{CA}(x_i^b, d_i \mid \{w\})), \\
u_b &= \exp(-m(\hat{x}_i^b(x)^n + \hat{x}_i^b(y)^n + \hat{x}_i^b(z)^n)), \\
\{c_i, \sigma_i\} &= \frac{1}{\sum_{b\in\mathbb{B}} u_b} \sum_{b\in\mathbb{B}} u_b \{c_i^b, \sigma_i^b\},
\end{aligned} \tag{5}
$$

Algorithm 1 Compositional Cross-modal Human

```
$\mathcal{D}$: 2D collection of human images / fashion descriptions

$G$, $D$: the generator / discriminator model for T3H
while TRAIN_CCH do
    $\mathcal{V}$, $\mathcal{T}$ ← sampled human / description from $\mathcal{D}$
    $\{\beta, \theta\}$ ← estimated SMPL parameters of $\mathcal{V}$
    $\{x_i\}$ ← canonical points via inverse LBS          ▷ Eq. 2
    $f_i^b$ ← rendering features inside the $b$th box     ▷ Eq. 3
    $\{w_l\}$ ← extracted textual features of $\mathcal{T}$
    CA ← fusion via cross-modal attention                 ▷ Eq. 4
    $\{c_i, \sigma_i\}$ ← mixture radiance / density       ▷ Eq. 5
    $R$ ← final rendering human                           ▷ Eq. 1

    $\mathcal{S}$ ← segmentation map of $\mathcal{V}$
    $Q$ ← fashion map between $\mathcal{S}$ and $\mathcal{T}$     ▷ Eq. 6
    $\mathcal{L}_{\text{adv}}$ ← adversarial loss from $D$        ▷ Eq. 8
    $\mathcal{L}_{\text{off}}$, $\mathcal{L}_{\text{eik}}$ ← offset and derivation loss   ▷ Eq. 9
    $\mathcal{L}_{\text{all}}$ ← overall training loss            ▷ Eq. 10
    Update $G$ by minimizing $\mathcal{L}_{\text{all}}$
    Update $D$ by maximizing $\mathcal{L}_{\text{all}}$
end while
```

where m and n are hyperparameters. With $\{c_i, \sigma_i\}$, we adopt Eq. 1 to render the RGB value of ray $r(t)$. Through all sampled rays r, we then have our final human rendering R, where the overall process can be simplified as $R = G(\beta, \theta \mid \mathcal{T})$. To summarize, CCH leverages the human prior and adopts inverse LBS to acquire the canonical space for the target pose. The human body is divided into 16 parts, and each of them fuses its correlated fashion semantics via cross-modal attention. Finally, compositional bodies jointly render the target 3D human.

Semantic Discrimination. With the SMPL prior, our CCH contains robust geometry transformation for humans and can learn from 2D images without actual 3D guidance. For a ground-truth $\{\mathcal{V}, \mathcal{T}\}$, we parse the 2D human image as the segmentation $\mathcal{S}$ (MMHuman3D, 2021), which provides the reliable body architecture. To obtain its fashion map Q, we apply cross-modal attention between $\mathcal{S}$ and $\mathcal{T}$:

$$
\begin{aligned}
\{e_{i,j}\} &= \text{Conv}(\mathcal{S}), \\
Q_{i,j} &= \sum_{l=1}^{L} \frac{\exp(e_{i,j} W w_l^{\text{T}})}{\sum_{\iota=1}^{L} \exp(e_{i,j} W w_\iota^{\text{T}})} w_l,
\end{aligned} \tag{6}
$$

where e is the same dimension as f, W is the learnable attention matrix, and Q perceives which human body part should showcase what fashion patterns. We concatenate the rendered human R (or the ground-truth $\mathcal{V}$) with Q and feed them into our discriminator D to perform binary classification:

$$
D(R \mid \mathcal{T}) = \text{BC}([\text{Conv}(R), Q]). \tag{7}
$$

In consequence, D can provide alignments of both

	DeepFashion					SHHQ				
Method	FID↓	Depth↓	PCK↑	CLIP-S↑	FA↑	FID↓	Depth↓	PCK↑	CLIP-S↑	FA↑
Latent-NeRF	69.654	0.0298	74.211	22.500	65.883	72.256	0.0381	73.401	22.210	67.427
TEXTure	37.058	0.0165	86.354	23.385	67.508	48.618	0.0216	85.502	24.456	68.233
CLIP-O	25.488	0.0133	87.892	21.887	61.964	34.212	**0.0164**	87.312	21.401	66.808
CCH	**21.136**	**0.0121**	**88.355**	**25.023**	**72.038**	**32.858**	0.0165	**87.624**	**27.855**	**76.194**

Table 1: Overall results of pose-guided `T3H`.

the human pose and fashion semantics, which improves the fine-grained consistency of our CCH.

Learning of CCH. We include the non-saturating loss with R1 regularization (Mescheder et al., 2018) for adversarial learning over the ground-truth $\{\mathcal{V}\}$:

$$\begin{aligned} U(u) &= -\log(1+\exp(-u)), \\ \mathcal{L}_{\text{adv}} &= U(G(\beta, \theta \mid \mathcal{T}) \mid \mathcal{T}) \\ &\quad + U(-D(\mathcal{V} \mid \mathcal{T})) + \lambda|\nabla D(\mathcal{V} \mid \mathcal{T})|^2. \end{aligned} \tag{8}$$

Following EVA3D, we also append the minimum offset loss $\mathcal{L}_{\text{off}}$ to maintain a plausible human shape as the template mesh. $\mathcal{L}_{\text{eik}}$ penalizes the derivation of delta SDFs to zero and makes the estimated SDF physically valid (Gropp et al., 2020):

$$\begin{aligned} \mathcal{L}_{\text{off}} &= ||\Delta d(x)||_2^2, \\ \mathcal{L}_{\text{eik}} &= ||\nabla(\Delta d(x))||_2^2. \end{aligned} \tag{9}$$

The learning process of our CCH is also illustrated as Algo. 1, where the overall optimization can be:

$$\begin{aligned} \mathcal{L}_{\text{all}} &= \mathcal{L}_{\text{adv}} + 1.5 \cdot \mathcal{L}_{\text{off}} + 0.5 \cdot \mathcal{L}_{\text{eik}}, \\ &\min_G \max_D \mathcal{L}_{\text{all}}. \end{aligned} \tag{10}$$

4 Experiemnts

4.1 Experimental Setup

Datasets. We coduct experiments on DeepFashion (Jiang et al., 2022) and SHHQ (Fu et al., 2022a) for `T3H`. DeepFashion contains 12K human images with upper and lower clothing descriptions. Since there are no annotations in SHHQ, we first fine-tune GIT (Wang et al., 2022c) on DeepFashion and then label for 40K text-human pairs. We follow OpenPose (Cao et al., 2019) and SMPLify-X (Pavlakos et al., 2019) to estimate the human keypoints and its SMPL parameters. The resolution is resized into 512x256 in our experiments. Note that all faces in datasets are blurred prior to training, and the model is not able to generate human faces.

Evaluation Metrics. We apply metrics from both visual and semantic prospects. Following EVA3D, we adopt Frechet Inception Distance (**FID**) (Heusel et al., 2017) and **Depth** (Ranftl et al., 2020) to calculate visual and geometry similarity, compared to the ground-truth image. We treat Percentage of Correct Keypoints (**PCK@0.5**) (Andriluka et al., 2014) as the correctness of the generated pose. To investigate the textual relevance of `T3H` results, we follow CLIP Score (**CLIP-S**) (Hessel et al., 2021) for the text-visual similarity. We fine-tune CLIP (Radford et al., 2021) on DeepFashion for a more accurate alignment in this specific domain. To have the fine-grained evaluation, we train a fashion classifier on DeepFashion labels[1] and assess Fashion Accuracy (**FA**) of the generated human.

Baselines. As a new task, we consider the following methods as the compared baselines.

- Latent-NeRF (Metzer et al., 2023) brings NeRF to the latent space and guides its generation by the given object and a text-to-image prior.
- TEXTure (Richardson et al., 2023) paints a 3D object from different viewpoints via leveraging the pre-trained depth-to-image diffusion model.
- CLIP-O is inspired by AvatarCLIP (Hong et al., 2022), which customizes a human avatar from the description with CLIP text-visual alignment. We apply the guided loss to optimize a pre-trained EVA3D (Hong et al., 2023) for faster inference.
- Texformer (Xu and Loy, 2021) estimates the human texture from an image. Text2Human (Jiang et al., 2022) predicts the target human image, and we treat Texformer to further build its 3D model.

For a fair comparison, all baselines are re-trained on face-blurred datasets and cannot produce identifiable human faces.

Implementation Detail. We divide a human body into 16 parts and deploy individual StyleSDF (Or-El et al., 2022) for each volume rendering, and two following MLPs then estimate SDF and RGB values. We adopt the same discriminator as StyleSDF over fashion maps to distinguish between fake rendered humans and real images. We sample N=28 points for each ray and set (m, n) to (4, 8) for mix-

[1] There are six targets for FA, including the shape, fabric, and color of the upper and lower clothing. The fashion classifier has +95% accuracy, which provides a precise evaluation.

Method	DeepFashion FID↓	 CLIP-S↑	 FA↑
Texformer	45.844	20.546	66.679
CLIP-O	25.579	20.112	61.298
CCH	**21.355**	**24.920**	**70.771**

Table 2: Overall results of pose-free T3H.

Ablation Settings Text	 CA	 SD	DeepFashion FID↓	 CLIP-S↑	 FA↑
✗	✗	✗	25.671	9.632	36.634
✓	✗	✗	24.624	21.079	69.173
✓	✓	✗	21.966	24.103	80.028
✓	✓	✓	**21.275**	**25.211**	**80.776**

Table 3: Ablation study (256x128) of Cross-modal Attention (CA) and Semantic Discrimination (SD).

ture rendering. The text encoder is initialized from CLIP and subsequently trained with CCH. We treat Adam (Kingma and Ba, 2015) with a batch size of 1, where the learning rates are 2e-5 for G and 2e-4 for D. We apply visual augmentations by randomly panning, scaling, and rotating within small ranges. All trainings are done using PyTorch (Paszke et al., 2017) on 8 NVIDIA A100 GPUs for 1M iterations.

4.2 Quantitative Results

Table 1 shows the pose-guided T3H results on DeepFashion and SHHQ, where we feed the estimated human mesh as the input object into Latent-NeRF and TEXTure. Although Latent-NeRF can portray body shapes in multiple angles from its latent NeRF space, the rendering is clearly counterfeit (higher FID and Depth). For TEXTure, the human architecture is constructed well by the given mesh (higher PCK). However, the estimated texture is still spatially inconsistent and contains inevitable artifacts (still higher FID). From the semantic aspect, Latent-NeRF and TEXTure borrow those trained diffusion models and depict the assigned appearance in the description (higher CLIP-S than CLIP-O). CLIP-O relies on EVA3D to produce feasible 3D humans (lower FID). While the external CLIP loss attempts to guide the fashion, the global alignment is insufficient to demonstrate detailed patterns (lower FA). Without those above drawbacks, our CCH learns to extract fashion semantics along with the compositional human generation, leading to comprehensive superiority across all metrics.

Train	FT.	DeepFashion R1	 R5	 R10
OpenAI-400M	✗	4.2	20.4	28.0
LAION-2B	✗	13.4	33.4	46.4
LAION-2B	✓	**45.0**	**83.0**	**93.8**

Table 4: Text-to-Fashion retrieval (sample 500 pairs) by CLIP with different fine-tunings (FT.).

Method	DeepFashion Quality	 Relevance
Latent-NeRF	1.82	2.37
TEXTure	2.38	2.51
CLIP-O	**2.93**	2.20
CCH	2.87	**2.92**

Table 5: Human evaluation for T3H with aspects of 3D quality and fashion relevance.

A similar trend can be found on SHHQ. Latent-NeRF and TEXTure exhibit related fashion patterns but are hard to present realistic humans (higher FID and Depth). CLIP-O produces a sharp human body with the correct pose, but not the assigned fashion (lower CLIP-S and FA) by the inexplicit alignment from CLIP. Table 2 presents the pose-free results. With the guided 2D image, Texformer contains the assigned clothing in the text (higher FA than CLIP-O). But the 3D reconstruction is unable to handle spatial rendering, resulting in low-quality humans (higher FID). With cross-modal attention and semantic discrimination, CCH exceeds baselines in both visual and textual relevance, making concrete human rendering with the corresponding fashion.

4.3 Ablation Study

We study each component effect of CCH in Table 3. Without the guided description, the model lacks the target fashion and results in a poor FA. This further highlights the importance of textual guidance for controllable human generation. When applying the traditional training (Reed et al., 2016), conditional GAN is insufficient to extract fashion semantics for effective T3H (not good enough CLIP-S). On the other hand, our cross-modal attention constructs a better fusion between fashion patterns and volume rendering, facilitating a significant improvement in depicting the desired human appearance. Moreover, semantic discrimination benefits fine-grained alignment and leads to comprehensive advancement.

Fine-tune CLIP-S as Evaluator. CLIP has shown promising text-visual alignment, which can calculate feature similarity between the generated human and the given text as CLIP-S (Hessel et al., 2021). Since our T3H is in a specific fashion domain, we consider the larger-scaled trained checkpoint from OpenCLIP (Ilharco et al., 2021) and fine-tune it as

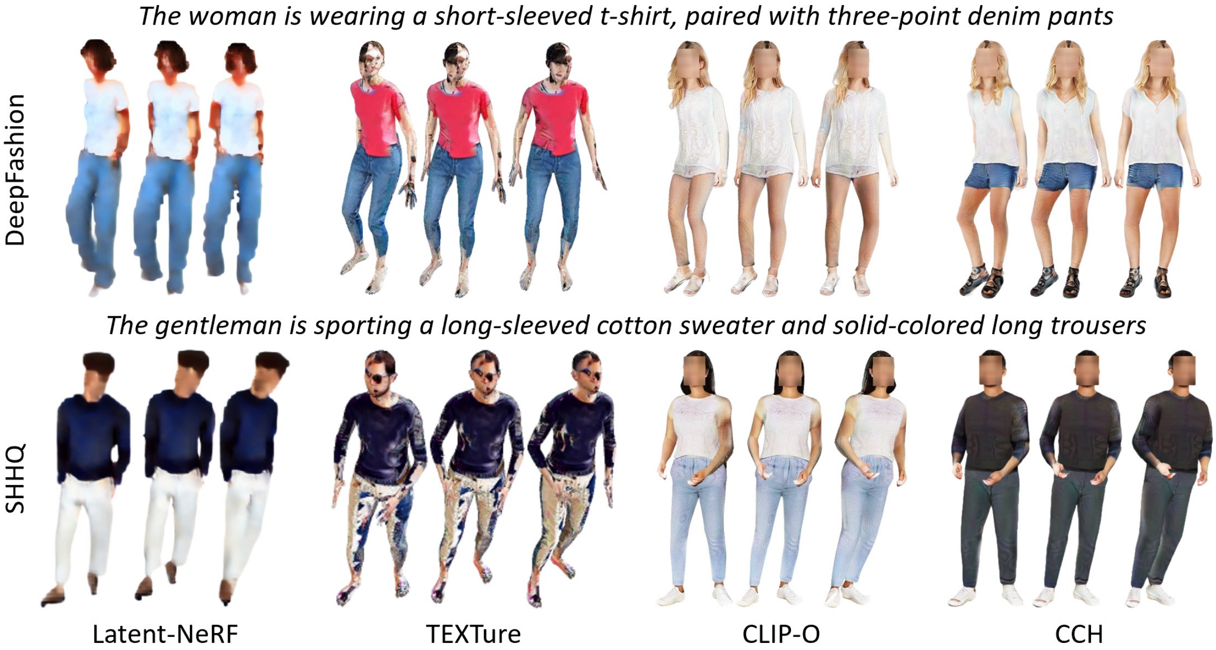

The woman is wearing a short-sleeved t-shirt, paired with three-point denim pants
DeepFashion
The gentleman is sporting a long-sleeved cotton sweater and solid-colored long trousers
SHHQ
Latent-NeRF
TEXTure
CLIP-O
CCH

Figure 3: Qualitative comparison of pose-guided T3H.

Method	Time (sec)	GPU (MB)
Latent-NeRF	755.7	11250
TEXTure	103.7	12530
CLIP-O	181.6	15988
CCH	**0.372**	**6258**

Table 6: Time and GPU cost to perform T3H.

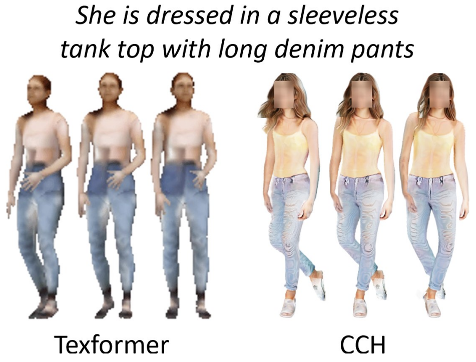

She is dressed in a sleeveless
tank top with long denim pants
Texformer
CCH

Figure 4: Qualitative comparison of pose-free T3H.

a more precise evaluator. Table 4 presents text-to-fashion retrieval results, where a higher recall leads to a better alignment. Whether the original CLIP or OpenCLIP, both result in poor performance and is insufficient for our evaluation. By perceiving DeepFashion, fine-tuning helps bring reliable alignment and is treated as the final evaluator.

Human Evaluation. Apart from automatic metrics, we conduct the human evaluation with aspects of 3D quality and fashion relevance. We randomly sample 75 T3H results and consider MTurk[2] to rank over baselines and our CCH. To avoid the potential ranking bias, we hire 3 MTurkers for each example. Table 5 shows the mean ranking score (from 1 to 4, the higher is the better). CLIP-O and CCH are built upon EVA3D, which provides an articulate human body for superior 3D quality. Even if Latent-NeRF and TEXTure take pre-trained diffusion models to acquire visual guidance, CCH exhibits more corresponding fashion via cross-modal fusion. This performance trend is similar to our evaluation, which supports the usage of CLIP-S and FA as metrics.

Inference Efficiency. In addition to T3H quality, our CCH also contains a higher efficiency. Table 6 shows the inference time and GPU cost on a single NVIDIA TITAN RTX. All baselines take more than 100 seconds since they require multiple iterations to optimize the 3D model from an external alignment. In contrast, we extract fashion semantics and carry out T3H in one shot. Without updating the model, we save the most GPU memory. In summary, CCH surpasses baselines on both quality and efficiency, leading to an effective and practical T3H.

4.4 Qualitative Results.

We demonstrate the qualitative comparison of pose-guided T3H in Fig. 3. Although Latent-NeRF can portray the 3D human based on the given mesh, it only presents inauthentic rendering. TEXTure generates concrete humans, but there are still obvious cracks and inconsistent textures from different angles of view. Moreover, both of them fail to capture "*three-point*", where the rendered lower clothing is

[2] Amazon MTurk: https://www.mturk.com

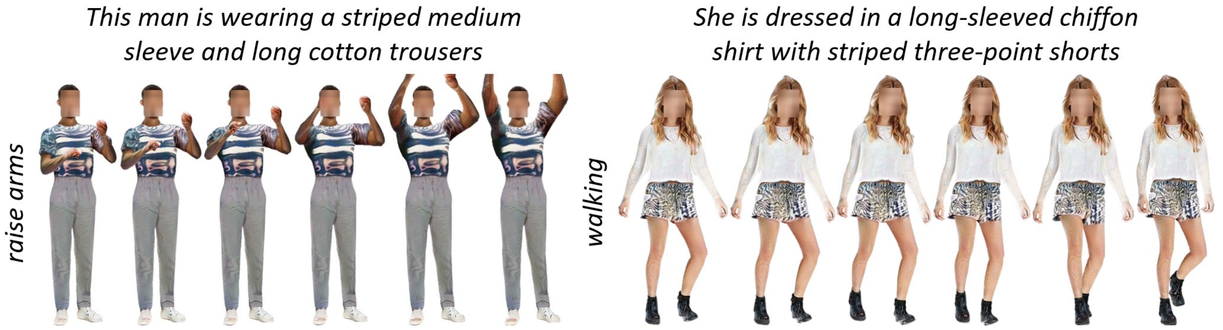

This man is wearing a striped medium sleeve and long cotton trousers
raise arms
She is dressed in a long-sleeved chiffon shirt with striped three-point shorts
walking

Figure 5: Qualitative examples of animatable T3H, where the motion is also controlled by the text.

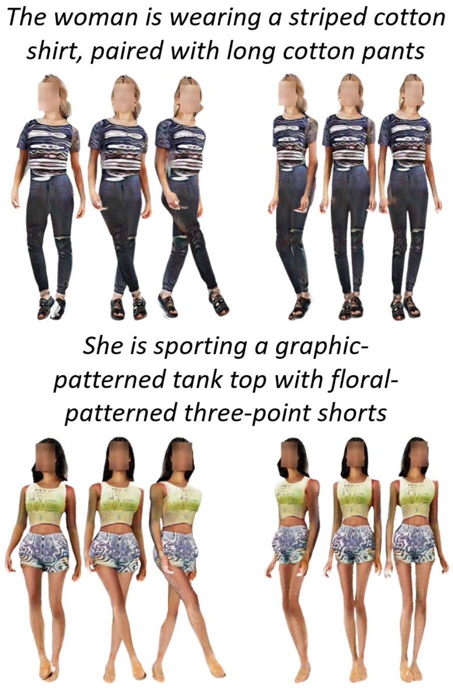

The woman is wearing a striped cotton shirt, paired with long cotton pants
She is sporting a graphic-patterned tank top with floral-patterned three-point shorts

Figure 6: Qualitative examples of pose-control T3H.

incorrectly depicted as long pants. Because CLIP provides an overall but inexplicit alignment to the description, CLIP-O is limited and exhibits vague "*denim*" or "*long-sleeved*". This observation further indicates the flaw of CLIP in detailed fashion patterns, even if it has been fine-tuned on the target dataset. In contrast, our CCH adopts cross-modal attention with NeRF, contributing to high-quality T3H with fine-grained fashion controllability. Fig. 4 shows the pose-free results. Texformer relies on a 2D image to estimate its 3D texture. Despite containing the assigned fashion, it is still restricted by the capability of 3D reconstruction, resulting in a low-resolution rendering. By learning text-to-3D directly, CCH can produce textual-related humans from random poses with clear visual patterns.

Pose-control T3H. Since our CCH is generating 3D humans from given SMPL parameters, as illustrated in Fig. 6, we can further control T3H with a specific pose. Different fashion descriptions make a human body present diverse appearances; different poses then guide the character to express rich body language. This flexibility in controlling appearance and pose allows for better practical customization.

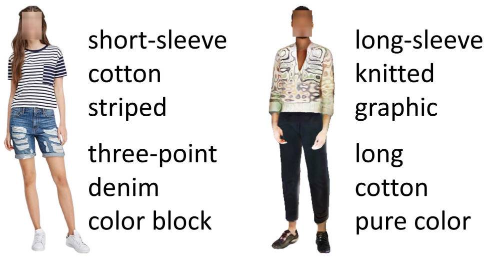


Figure 7: Qualitative examples of our fashion classifier, which provides fine-grained labels for real/fake humans.

Animatable T3H. In addition to static poses, CCH can benefit from dynamic motions to achieve animatable T3H. Fig. 5 adopts MotionDiffuse (Zhang et al., 2022) to create the assigned action also from the text and apply it to our produced 3D models. In this way, we prompt them to "*raise arms*" or "*walk*" for favorable dynamic scenarios.

5 Conclusion

We present text-guided 3D human generation (T3H) to create a 3D human by a fashion description. To learn this from 2D collections, we introduce Compositional Cross-modal Human (CCH). With cross-modal attention, CCH fuses compositional human rendering and textual semantics to build a concrete body architecture with the corresponding fashion. Experiments across various fashion attributes show that CCH effectively carries out T3H with high efficiency. We believe T3H helps advance a new field toward vision-and-language research.

Ethics Discussion and Limitation. Our work enhances the controllability of 3D human generation. To prevent identity leakage, we blur out faces prior to training and avoid risks similar to DeepFake (Korshunov and Marcel, 2018). Because we depend on SMPL parameters, an inaccurate estimation causes a distribution shift and quality degradation. For the datasets, they reveal narrow viewing angles, which results in visible artifacts of 3D consistency.

Acknowledgements. We appreciate the anonymous reviewers for constructive feedback. The research presented in this work was funded by Meta AI. The views expressed are those of the authors and do not reflect the official policy or position of the funding agency.

References

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2022. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *European Conference on Computer Vision (ECCV)*.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. 2022. Geometry-Guided Progressive NeRF for Generalizable and Efficient Neural Human Rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiqin Chen and Hao Zhang. 2019. Learning Implicit Fields for Generative Shape Modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality Streamable Free-viewpoint Video. In *Transactions on Graphics (TOG)*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering Text-to-Image Generation via Transformers. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W.Taylor. 2019. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In *International Conference on Computer Vision (ICCV)*.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *International Conference on Learning Representations (ICLR)*.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. 2022a. StyleGAN-Human: A Data-Centric Odyssey of Human Generation. In *European Conference on Computer Vision (ECCV)*.

Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. 2020. SSCR: Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. 2022b. M^{3}L: Language-based Video Editing via Multi-Modal Multi-Level Transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. 2022c. Language-Driven Artistic Style Transfer. In *European Conference on Computer Vision (ECCV)*.

Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. 2023. Tell Me What Happened: Unifying Text-guided Video Completion via Multimodal Masked Video Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin Gao, Jie Yang, Tong Wu, Yu-Jie Yuan, Hongbo Fu, Yu-Kun Lai, and Hao Zhang. 2019. SDM-NET: Deep Generative Network for Structured Deformable Mesh.

In *Special Interest Group on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia)*.

Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. 2022. MPS-NeRF: Generalizable 3D Human Rendering from Multiview Images. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Thiago L. Gomes, Thiago M. Coutinho, Rafael Azevedo, Renato Martins, and Erickson R. Nascimento. 2022. Creating and Reenacting Controllable 3D Humans with Differentiable Rendering. In *Winter Conference on Applications of Computer Vision (WACV)*.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2015. Generative Adversarial Networks. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. 2020. Implicit Geometric Regularization for Learning Shapes. In *International Conference on Machine Learning (ICML)*.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations (ICLR)*.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Matthew DuVall, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanellov, Graham Fyffe, Christoph Rhemannv, Jonathan Taylor, Paul Debevec, and Shahram Izadi. 2019. The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting. In *Transactions on Graphics (TOG)*.

Paul Henderson, Vagia Tsiminaki, and Christoph H. Lampert. 2020. Leveraging 2D Data to Learn Textured 3D Mesh Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. 2023. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *International Conference on Learning Representations (ICLR)*.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. Avatar-CLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. In *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP. https://github.com/mlfoundations/open_clip.

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2Human: Text-Driven Controllable Human Image Generation. In *Transactions on Graphics (TOG)*.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. 2022. CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. In *Special Interest Group on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In *International Conference on Computer Vision (ICCV)*.

Pavel Korshunov and Sebastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. In *arXiv:1812.08685*.

Chun-Liang Li, Manzil Zaheer, Yang Zhang, Barnabas Poczos, and Ruslan Salakhutdinov. 2018a. Point Cloud GAN. In *International Conference on Learning Representations (ICLR)*.

Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. 2017. GRASS: Generative Recursive Autoencoders for Shape Structures. In *Transactions on Graphics (TOG)*.

Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018b. Video Generation From Text. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. In *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*.

Andrew Luo, Tianqin Li, Wen-Hao Zhang, and Tai Sing Lee. 2021. SurfGen: Adversarial 3D Shape Synthesis with Explicit Surface Discriminators. In *International Conference on Computer Vision (ICCV)*.

Tanya Marwah, Gaurav Mittal, and Vineeth N. Balasubramanian. 2017. Attentive Semantic Video Generation using Captions. In *International Conference on Computer Vision (ICCV)*.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which Training Methods for GANs do actually Converge? In *International Conference on Machine Learning (ICML)*.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2022. Text2Mesh: Text-Driven Neural Stylization for Meshes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*.

MMHuman3D. 2021. OpenMMLab 3D Human Parametric Model Toolbox and Benchmark. https://github.com/open-mmlab/mmhuman3d.

Thomas Muller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. In *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*.

Gimin Nam, Mariem Khlifi, Andrew Rodriguez, Alberto Tono, Linqi Zhou, and Paul Guerrero. 2022. 3D-LDM: Neural Implicit 3D Shape Generation with Latent Diffusion Models. In *arXiv:2212.00842*.

Charlie Nash, Yaroslav Ganin, Ali Eslami, and Peter W. Battaglia. 2020. PolyGen: An Autoregressive Generative Model of 3D Meshes. In *International Conference on Machine Learning (ICML)*.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. 2022. Unsupervised Learning of Efficient Geometry-Aware Neural Articulated Representations. In *European Conference on Computer Vision (ECCV)*.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *International Conference on Learning Representations Workshop (ICLRW)*.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Ahmed A Timo Bolkart, A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *International Conference on Computer Vision (ICCV)*.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *International Conference on Learning Representations (ICLR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and

Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning (ICML)*.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. TEXTure: Text-Guided Texturing of 3D Shapes. In *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2021. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. 2023. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. In *arXiv:2303.07937*.

Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor Lempitsky. 2019. Textured Neural Avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. 2020. Implicit Neural Representations with Periodic Activation Functions. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 2023. 3D generation on ImageNet. In *International Conference on Learning Representations (ICLR)*.

Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. 2021. A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. 2020. RobustFusion: Human Volumetric Capture with Data-driven Visual Cues using a RGBD Camera. In *European Conference on Computer Vision (ECCV)*.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior. In *arXiv:2303.14184*.

Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. 2017. Octree Generating Networks: Efficient Convolutional Architectures for High-resolution 3D Outputs. In *International Conference on Computer Vision (ICCV)*.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022a. CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2022b. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *arXiv:2212.00774*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022c. GIT: A Generative Image-to-text Transformer for Vision and Language. In *Transactions on Machine Learning Research (TMLR)*.

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiangyu Xu and Chen Change Loy. 2021. 3D Human Texture Estimation from a Single Image with Transformers. In *International Conference on Computer Vision (ICCV)*.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. 2019. PointFlow: 3D Point Cloud Generation with Continuous Normalizing Flows. In *International Conference on Computer Vision (ICCV)*.

Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 2022. 3DHumanGAN: Towards Photo-Realistic 3D-Aware Human Image Generation. In *arXiv:2212.07378*.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume Rendering of Neural Implicit Surfaces. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. 2021. Pose-Guided Human Animation from a Single Image in the Wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. 2021. NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. In *arXiv:2208.15001*.

Xin-Yang Zheng, Yang Liu, Peng-Shuai Wang, and Xin Tong. 2022. SDF-StyleGAN: Implicit SDF-Based StyleGAN for 3D Shape Generation. In *Computer Graphics Forum (SGP)*.

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2021. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human Reconstruction. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.