Text2Fashion

, ,

***Abstract*—The creation of diverse and high-quality human images is a crucial yet difficult undertaking in vision and graphics. The wide variety of clothing designs and textures, existing generative models are often not sufficient for the end user.** **In this work, we introduce Text2Fashion, a text-driven controlled framework for the development of varied human models and high-quality 3D models and images with a range of postures.** **We use two distinct procedures to create full-body 2D human photographs starting from a predetermined human posture. 1) The provided human pose is first converted to a human parsing map with some sentences that describe the shapes of clothing. 2) The system is then given further information about the textures of clothing as an input in order to produce the final human image. We plan to employ Inverse Graphic for the 3D model generation because it attempts to reconstruct 3D models from 2D observations.** **Regarding the creation of 2D images, a multi-scale neural codebook that is aware of hierarchy and textures is used.** **There are two sections in the codebook: a coarse level codebook and a fine level codebook. The codebook at the fine level concentrates on the minutiae of textures, whereas the codebook at the coarse level covers the structural representations of textures.** **A diffusion-based transformer sampler with a mixture of experts is first used to sample indices from the coarsest level of the codebook, which is then used to predict the indices of the codebook at finer levels in order to use the learnt hierarchical codebook to synthesis desired images. The decoder trained together with hierarchical codebooks converts the anticipated indices at various levels to human images. The created image can be dependent on the fine-grained text input thanks to the utilization of a blend of experts. The quality of clothing textures is refined by the prediction for finer level indices.** **Implementing these strategies can result in more diversified and realistic human images than state-of-the-art procedures, according to numerous quantitative and qualitative evaluations. These generated photographs will be converted into a 3D model, resulting in a number of postures and outcomes, or you may just make a 3D model from a dataset that produces a variety of stances.** **NeRF compositional human representation, which splits the human body into regional components, is used to create 3D models.** **A separate volume represents each component.**

**With the use of this compositional representation, 1) intrinsic human priors, 2) flexible network parameter allocation, and 3) effective training and rendering**

***Index Terms*—GAN (Generative Adversarial Networks) ,VAE (Variational autoencoders ), NeRF ,VQ-VAE (Vector Quantized Variational AutoEncoder )**

# I. INTRODUCTION

The need in variety of images of high quality is often in demand since the introduction of Generative Adversarial Networks (GANs) [Goodfellow et al. 2014], image production has advanced quickly. Today, we can quickly create a variety of faces with high fidelity using a pretrained StyleGAN [Karras et al. 2020], which also supports a number of downstream tasks, such as face stylization [Pinkney and Adler 2020; Song et al. 2021; Yang et al. 2022] and facial attribute editing [Abdal et al. 2021; Jiang et al. 2021; Patashnik et al. 2021]. Another sort of human-related media are full-body photographs of people, which have a richer, more varied, and finer-grained content. Additionally, there are several uses for human image generation, such as human pose transfer, virtual try-on, and animations [Cui et al. 2021; Lewis et al. 2021] as well as [Frühstück et al. 2019; Hong et al. 2022; Yoon et al. 2021]. In terms of applications and interactions, it is even preferable to enable lay users to easily control the synthesised human full-body images of people are another type of human-related media that have richer, more diversified, and finer-grained material. Human posture transfer, virtual try-on, and animations are a few other applications for the creation of human images [Cui et al. 2021; Lewis et al. 2021] as well as [Frühstück et al. 2019; Hong et al. 2022; Yoon et al. 2021]. In terms of interactions and applications, it is even ideal to produce high-fidelity human images while allowing lay users to simply manage them. images, in addition to producing high-fidelity human images. Due to the following difficulties, controlled human body image production with high accuracy and diversity is seldom researched despite its enormous potential: 1) When compared to faces, human body images are more complex due to a variety of factors, such as the variety of human stances, intricate clothing silhouettes, and various clothing textures; 2) Current methods for creating human body images [Sarkar et al. 2021b; Weng et al. 2020; Yildirim et al. 2019] do not produce a variety of clothing because they frequently produce items with basic patterns, such as solid colors, and they do not offer fine-grained control over the textures of the garments. 3) Additional fine-grained annotations are necessary for the production of clothing with textual controls. It is difficult to handle all involving aspects in a single generative model since human body images are so complex. Based on the provided human position and user-specified phrases specifying the clothes shapes, Stage I builds a human parsing mask with a variety of clothing shapes. Then Stage II enhances the human parsing mask with a variety of clothing textures based on texts that describe the textures of the clothing. Additionally, it is possible to create 3D models. with the aid of 2) generative network training techniques and 1) 3D human representation. Considering the articulated Due to the specific nature of human bodies, a desirable human portrayal should be able to 3D human stance or form. It is necessary to construct an articulated 3D human image rather than the static volume modelling used in 3D-aware GANs currently available. Using a clear illustration, The canonical pose (also known as "canonical space") of a 3D human is modelled, and several positions and forms (observation space). Additionally, the effectiveness of the portrayal is significant in

exceptional 3D human generation. prior techniques (Noguchi et al., 2022; Bergman et al., 2022) due to the ineffective human representations they use, high resolution production is not achieved. In this paper, we offer EVA3D, a 3D human generative model that is unconditionally high-quality. Only limited 2D human image sets. To make that easier, we suggest a compositional human. NeRF representation to boost model performance. The human body is divided into 16 keypoints/sections, and  assign a unique network to each component that represents the relevant local volume. As a human, three benefits are primarily offered through representation. 1) It already includes a description of the human body, this encourages direct control over the positions and contours of the human body. 2) It facilitates distributing compute resources in an adaptive manner. Heads, for example, are more complicated bodily parts that can be given more  parameters. 3) The compositional representation achieves high-resolution production while enabling efficient rendering. Thus this is how the approach is going to be to come up with an solution..

# II RELATED WORK

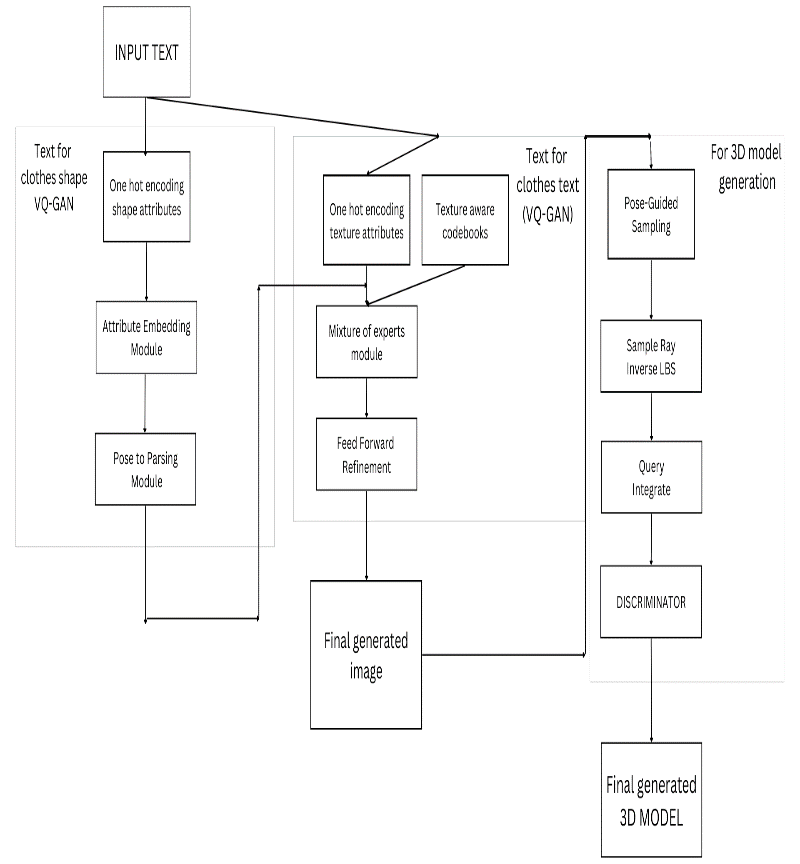
**The Generative Adversarial Network** (GAN) has proven to be extremely effective in producing high-fidelity images. Different variations of GAN have been proposed since [Goodfellow et al. 2014] first proposed the first generative model in 2014 [Brock et al. 2019; Chai et al. 2022; Karras et al. 2021, 2019, 2020]. Along with unconditional generation, conditional GANs [Mirza and Osindero 2014] were proposed to produce images based on criteria such as segmentation mask [Isola et al. 2017; Park et al. 2019; Wang et al. 2018] and natural language [Surya et al. 2020; Xu et al. 2018]. Our suggested Text2Human system creates conditional images using human poses and phrases as inputs. VAE [Kingma and Welling 2013] is an alternative picture generating paradigm to GAN.

**3D representations of humans:** For tasks involving humans, 3D human representations are essential tools. Parametric models are developed by Loper et al. (2015), Pavlakos et al. (2019b), and Hong et al. (2021) for the explicit modelling of 3D humans. Habermann et al. (2021); simulate human appearances **.**UV maps are further introduced by Shysheya et al. (2019); Yoon et al. (2021); and Liu et al. (2021). Although less realistic, parametric modelling offers reliable control over the human model. Implicit functions are used by Palafox et al. (2021) to create accurate 3D human body forms. The number of publications on human NeRF has also skyrocketed in tandem with the growth of NeRF (Peng et al., 2021b; Zhao et al., 2021; Peng et al., 2021a; Xu et al., 2021; Noguchi et al., 2021; Weng et al., 2022; Chen et al., 2021; Su et al., 2021; Ji For a variety of down-stream tasks, Hong et al. (2022a) suggest learning modal-invariant human representations. Several large-scale multi-modal 4D human datasets are provided by Cai et al. in 2022.

**Human Image Manipulation and Synthesis**. Pose transfer's purpose The goal of [Balakrishnan et al. 2018; Liu et al. 2020, 2019; Ma et al. 2017, 2018; Tao et al. 2022] is to maintain the same person's appearance in different poses. A styleGAN framework with pose conditioning was suggested by [Albahar et al. 2021]. After being twisted to the desired position, the original image's details are employed to spatially modulate the features for synthesis. An approach for the text-guided posture transfer challenge was suggested by [Zhou et al. 2019]. ADGAN was suggested by [Men et al. 2020] for controlled person image creation.

**Human Generation** : Despite the enormous progress made in the creation of human faces, the intricacy of human positions and appearances makes it difficult to create human images (Sarkar et al., 2021b; Lewis et al., 2021; Sarkar et al., 2021a; Jiang et al., 2022c). Recently, the dataset was scaled up by Fu et al. (2022); Fruhst uck et al. (2022), who produced excellent 2D human generation findings. Chen et al. (2022) used a 3D human dataset to build 3D human geometry. Some people also try to train 3D human GANs using just 2D human image libraries. The CNN-based neural renderers used by Grigorev et al. (2021) and Zhang et al. (2021) cannot ensure 3D consistency. Human NeRF, which only trains at low resolution, is used for this purpose by Noguchi et al. (2022) (Noguchi et al., 2021). Bergman et al. (2022); Zhang et al. (2022) suggest boosting the resolution by super-resolution, although this still doesn't yield excellent outcomes. From text inputs, Hong et al. (2022b) produce 3D avatars.

**3D-aware GAN**. In terms of creating 2D images, the Generative Adversarial Network (GAN) (Goodfellow et al., 2020) has achieved remarkable success (Karras et al., 2019; 2020). The generation of 3D awareness has also received a lot of attention. Voxels are used by Nguyen-Phuoc et al. (2019), Henzler et al. (2019), and meshes are used by Pan et al. (2020) to help the 3D-aware generation. Many people have developed 3D-aware GANs based on NeRF thanks to recent advancements in the technology (Mildenhall et al., 2020; Tewari et al., 2021). (Schwarz et al., 2020; Niemeyer & Geiger, 2021; Chan et al.,



2021; Deng et al., 2022). Gu et al. (2021); Or-El et al. (2022); and Chan et al. (2022) employ 2D decoders for super resolution to boost generation resolution. Furthermore, for more accurate geometry and better 3D consistency, it would be preferable to increase the raw resolution by increasing rendering efficiency (Skorokhodov et al., 2022; Xiang et al., 2022).We also suggest a powerful 3D human representation to enable training at high resolution.

## A. Abbreviations and Acronyms

GAN (Generative Adversarial Networks) ,VAE (Variational autoencoders ), NeRF ,VQ-VAE (Vector Quantized Variational Autoencoder ).

# III. Proposed work

The proposed system is used to generate 3D models from text so that poses of various ranges can be obtained along with High Quality images. The datasets used in this proposal are as follows

**Deepfashion** (Liu et al., 2016) dataset collects fashion images from the internet. We only use images that contain the full body and not wearing dresses, which results in 8,036 images for training,

**SHHQ** (Fu et al., 2022) dataset collects a larger-scale fashion dataset from the internet. It is processed similarly as DeepFashion, which results in 120,865 images in resolutions of 512 × 256. ,

**UBCFashion** (Zablotskaia et al., 2019) is a fashion video dataset containing 500 sequences of models posing in front of the camera.

Most models wear dresses in this dataset. AIST (Tsuchida et al., 2019) is a multi-view human dancing video dataset that provides rich poses and accurate SMPL(Skinned Multi-Person Linear Model).Apart from this if there is need of further new, high-quality images of various designs , styles and patterns we can still use the technique of the Text2human (Yuming Jiang et al ,2022) which is the most optimized till date as the use of VQVAE and Nvidia Tesla V100 GPU’s were used in training the models.

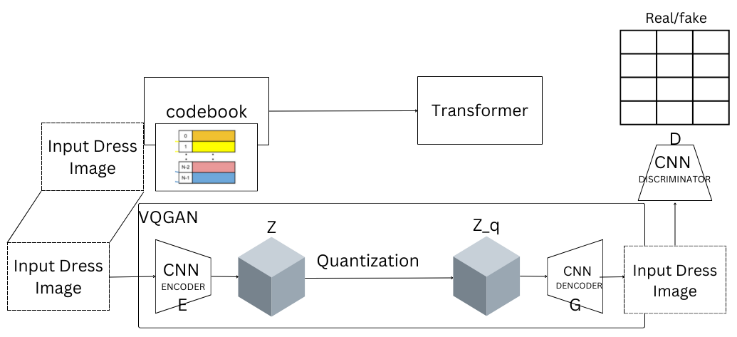
The only variation that we can is instead of using VQVAE and making the training process fast we can go for VQ-GAN in spite of it being slow it can provide accurate , high-quality images in comparison to VQVAE . Now once the images generated from this can be directly fed to generate 3D models. The 3D model generation is done by method of inverse graphics as it aims to recover 3D models from 2D observations.

A. PREREQUISITES

Before getting into VQ-GAN its important to know about **GAN**

and its **algorithms**

|  |
| --- |
| ***Algorithm 1Generative Adversarial Networks (GANs)*** |
| 1. for number of training iterations do 2. for k steps do 3. Sample minibatch of m noise samples   ***{z(1), . . ., z(m)}*** from noise prior ***pg(z).***   1. Sample minibatch of m examples   ***{x (1) , . . . , x (m)}*** from data generating distribution ***pdata(x).***   1. Update the discriminator by ascending its stochastic gradient: 2. end for 3. Sample minibatch of m noise samples   {z(1), . . . , z(m)}   1. Update the generator by descending its stochastic gradient:   The gradient-based updates can use any standard gradient-based learning rule. | |

**VQ-GAN** The main idea is that, to utilize the computationally costly self-attention operation in high-resolution synthesis, the images have to be in some way expressed as sequences. Instead of using pixels or patches as tokens, the encoder extracts an encoding ***Z\_hat*** which is then quantized to ***Z\_q*** using the closest codebook entry (as in VQ-VAE). The decoder can then reconstruct the image starting from the quantization. This part of training, needed to learn the codebook representation, is almost identical to the VQ-VAE one, with two main differences in the loss used. Indeed, as already said at the end of the last section, the loss used in VQ-VAE was composed of three terms, the Mean Squared Error (MSE ) plus the two alignment losses. Here, the MSE is substituted by perceptual loss, which is basically the MSE computed not on two images, but on some internal representations of the images. For example, two images are passed through a CNN, and the n-th layer features are extracted and compared with Mean Squared Error(MSE). In addition, they add an adversarial loss (the typical loss used in GANs, where a generator and a discriminator compete in a minimax game) to solve VQ-VAE blurring problem, with a prediction real/fake not for the entire image but for single patches. 

**NeRF** (Mildenhall et al., 2020) is an implicit 3D representation that can synthesize unique views that are lifelike. NeRF has the following definition: fc***;***

***g = F(x; d),………(3)***

where x is the query point, d is the viewing direction, ***c*** is the emitted radiance (RGB value), and is the volume density. We have the following formulation to determine the RGB value ***C(r)*** of some ray

***r(t) = o + td……….(4)***,

or volume rendering:

***)………(5)*** is the accumulated transmittance along the ray r from ***tn*** to ***t,*** where ***tn*** and ***tf*** denote the near and far limits.

It is discretized as follows to obtain the estimation of C(r):

***………(6)***

**SMPL (Loper et al., 2015)** A parametric human model, SMPL (Loper et al., 2015), is specified as *M*(***β****;* ***θ***), where ***β****;* ***θ***; governs body forms and positions. In this study, we transition from canonical space to observation spaces using the Linear Blend Skinning (LBS) algorithm of SMPL.

Formally,

defines how point x in the canonical space is changed to an observation space described by pose, where ***K*** is the joint number, is the blend weight of x against joint k, and is the transformation matrix of joint k. A similar technique using inverted transformation matrices is used for the inverse LBS transformation from observation spaces to the canonical space.

# 1.generation of human images

Our goal is to build 3D human images, but before we can do so, we must first construct human images based on words that describe the characteristics of clothing (clothes shapes and clothes textures).

The appropriate human image I where **𝐼 ∈ R𝐻×𝑊 ×3** should be produced using a human pose **𝑃 ∈ R𝐻×𝑊** texts for clothing shapes *Tshape*, and texts for clothes textures *Ttexture*.

We want to synthesis the human parsing map 𝑆 ∈ R𝐻×𝑊 . using a human pose P and words on clothing forms.

Texts are converted to a series of clothing shape properties called {𝑎1, ..., 𝑎𝑖, ..., 𝑎𝑘 }, where ai is represented by the values 0 through 1 and class number Ci. The Attribute Embedding Module receives the attributes after which a shape attribute embedding is produced.

**𝑓𝑠ℎ𝑎𝑝𝑒 = 𝐹𝑢𝑠𝑖𝑜𝑛([𝐸1(𝑎1), 𝐸2(𝑎2), ..., 𝐸𝑖 (𝑎𝑖 ), ..., 𝐸𝑘 (𝑎𝑘 )]) *………(9)*,**

Once this is complete further move to implementation using

VQ-GAN The VQGAN (Vector Quantized Generative Adversarial Network) is a GAN architecture that may be used to learn from prior data and produce new images. Esser, Rombach, and Ommer initially discussed it in their paper "Taming Transformers" (2021). In order to encode the feature map of the visual portions of the images, the feature map of the image data is first directly supplied to a GAN. Then, this picture data is vector quantized, a type of signal processing that organizes vector groups into clusters that may be accessed by a representative vector designating the centroid and is referred to as a "codeword." The vector quantized data is encoded and stored as a codebook, or dictionary of codes.

The loss produced in GAN is LGAN:

***LGAN(N,D)=[logD(x)+log(1−D(^x))]. ………(10)***

Vector quantization also happens between the encoder and decoder networks. After encoding the input x into , i.e., =E(x), we perform an element-wise operation **q** to obtain a discrete version of the input:

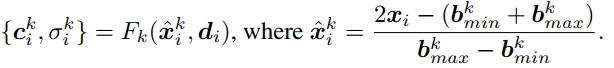
**=q():=arg min zk∈Z ||ij − zk||*…(11)***

Once after all this done further sent to mixture of experts and feed forward refinement where high-quality images are obtained next step is to send to 3D modelling.

# 2.Generation of 3d models.

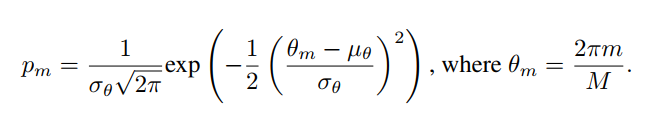
A collection of local bounding boxes B are used to define the compositional human NeRF representation as FΦ. We employ a subnetwork FΦ to model the local boundaries for each body part k.as seen in Fig. 2 b), box {***b****k min;* ***b****k max}*. Regarding a specific point xi in the canonical coordinate system,

The matching radiance ck I and density ik are in the direction di and falling inside the k-th bounding box, respectively. is challenged by

***………(12)***

Now comes the important enhancing phase,

***Pose-directed Sampling:*** In addition to using a 3D human template, we suggest balancing the input 2D photos based on human positions to help 3D information learning from sparse 2D image collections. The idea behind pose-guided sampling is that in order to effectively teach geometry, different viewing angles should be sampled more uniformly. Empirically, we use the angle of the head to direct the sample among all human joints. Additionally, the face holds more data than other portions of the brain. Angles from the front should be sampled more often than other angles. As a result, we decide to employ a Gaussian distribution with a standard deviation of and a centering at the front-view angle of. On the circle, M bins are segregated specifically.

 ***………(13)***

***Delta SDF Forecast***. Most 2D real-world human image collections, particularly fashion datasets, feature an unbalanced distribution of poses. We suggest introducing a strong human prior by basing our human representation on the SMPL template geometry dT (x). We forecast an SDF offset d(x) from the template rather than the SDF value d(x) directly (Yifan et al., 2022). The real SDF value of point x is then calculated using **dT (x) + d(x).**

After all this done and loss functions calculated an 3D model will be obtained finally but still there’s scope for improvement.

References

1. *Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. Styleflow: Attributeconditioned exploration of stylegan-generated images using conditional continuous*
2. *normalizing flows. ACM Transactions on Graphics (TOG) 40, 3 (2021), 1–21.*
3. *Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. 2021. Pose with Style: Detail-preserving pose-guided image synthesis with conditional stylegan. ACM Transactions on Graphics (TOG) 40, 6 (2021), 1–11.*
4. *Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. 2018.Synthesizing images of humans in unseen poses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8340–8348.*
5. *Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks.*
6. *2021. Unleashing Transformers: Parallel Token Prediction with Discrete AbsorbingDiffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes. arXiv preprint arXiv:2111.12701 (2021).*
7. *Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In International Conference on Learning Representations. ttps://openreview.net/forum?id=B1xsqj09Fm*
8. *Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. 2022. HuMMan: Multi-Modal 4D Human Dataset for*
9. *Versatile Sensing and Modeling. arXiv preprint arXiv:2204.13686 (2022). Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. 2022. Any-resolution Training for High-resolution Image Synthesis. arXiv preprint arXiv:2204.07156 (2022).*
10. *Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on*
11. *computer Vision and Pattern Recognition, pp. 3686–3693, 2014.*
12. *Alexander W Bergman, Petr Kellnhofer, Yifan Wang, Eric R Chan, David B Lindell, and Gordon*
13. *Wetzstein. Generative neural articulated radiance fields. arXiv preprint arXiv:2206.14314, 2022.*
14. *Mikołaj Binkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd ´*
15. *gans. arXiv preprint arXiv:1801.01401, 2018.*
16. *Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang*
17. *Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing*
18. *and modeling. arXiv preprint arXiv:2204.13686, 2022.*
19. *Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on*
20. *computer Vision and Pattern Recognition, pp. 3686–3693, 2014.*
21. *Alexander W Bergman, Petr Kellnhofer, Yifan Wang, Eric R Chan, David B Lindell, and Gordon*
22. *Wetzstein. Generative neural articulated radiance fields. arXiv preprint arXiv:2206.14314, 2022.*
23. *Mikołaj Binkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd ´*
24. *gans. arXiv preprint arXiv:1801.01401, 2018.*
25. *Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang*
26. *Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing*
27. *and modeling. arXiv preprint arXiv:2204.13686, 2022.*
28. *Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic*
29. *implicit generative adversarial networks for 3d-aware image synthesis. In Proceedings of the*
30. *IEEE/CVF conference on computer vision and pattern recognition, pp. 5799–5809, 2021.*
31. *Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio*
32. *Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware*
33. *3d generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer*
34. *Vision and Pattern Recognition, pp. 16123–16133, 2022*