



Text-driven clothed human image synthesis with 3D human model estimation for assistance in shopping

S. Karkuzhali¹ · A. Syed Aasim¹ · A. StalinRaj¹

Received: 23 March 2024 / Revised: 18 July 2024 / Accepted: 26 August 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Online shopping has become an integral part of modern consumer culture. Yet, it is plagued by challenges in visualizing clothing items based on textual descriptions and estimating their fit on individual body types. In this work, we present an innovative solution to address these challenges through text-driven clothed human image synthesis with 3D human model estimation, leveraging the power of Vector Quantized Variational AutoEncoder (VQ-VAE). Creating diverse and high-quality human images is a crucial yet difficult undertaking in vision and graphics. With the wide variety of clothing designs and textures, existing generative models are often not sufficient for the end user. In this proposed work, we introduce a solution that is provided by various datasets passed through several models so the optimized solution can be provided along with high-quality 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 model developed is then given further information about the textures of clothing as an input to produce the final human image. The model is split into two different sections the first one being a codebook at a coarse level that deals with overall results and a fine-level codebook that deals with minute detailing. As mentioned previously at fine level concentrates on the minutiae of textures, whereas the codebook at the coarse level covers the depictions of textures in structures. 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 forecast for finer-level indexes. Implementing these strategies can result in more diversified and high-quality 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 several postures and outcomes, or you may just make a 3D model from a dataset that produces a variety of stances. The application of the PIFu method uses the Marching cube algorithm and Stacked Hourglass method to produce 3D models and realistic images respectively. This results in the generation of high-resolution images based on textual description and reconstruction of the generated images as 3D models. The inception score and Fréchet Intercept Distance, SSIM, and PSNR that was achieved was 1.64 ± 0.20 and 24.64527782349843, 0.642919520, and 32.87157744102002 respectively. The implemented method scores well in comparison with other techniques.

Extended author information available on the last page of the article

This technology holds immense promise for reshaping the e-commerce landscape, offering a more immersive and informative means of exploring clothing options.

Keywords Generative Adversarial Networks · Variational autoencoders · Neural Radiance Fields(NeRF) · Vector Quantized Variational Autoencoder · Hour glass

1 Introduction

The need for a variety of images of high quality is often in demand of Generative Adversarial Networks (GANs) [1], and image production has advanced quickly. Today, we can quickly create a variety of faces with high fidelity using a pre-trained Style GAN, which also supports several downstream tasks, such as face stylization and facial attribute editing. Another sort of human-related media is full-body photographs of people, which have richer, more varied, and finer-grained content. The vast majority of techniques currently in use in this field base their editing models on instances of the intended clothing, which is commonly twisted and sewn into the given input image. Although these methods allow for the modification of images through more natural high-level language-based descriptions of the intended apparel, text-conditioned fashion-image editing is still favored. This is because they offer a convincing alternative to example-based editing techniques. Additionally, there are several uses for human image generation, such as human pose transfer, virtual try-on, and animations [2] In terms of applications and interactions, it is even preferable to enable lay users to easily control the synthesized human full-body images of people are another type of human-related media that have richer, more diversified, and finer-grained material. Current methods for creating human body images 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. Additional fine-grained annotations are necessary for the production of clothing with textual controls. It is difficult to handle all involved 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, the first process builds a human parsing mask with a variety of clothing shapes. The second process 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. We implement the Pixel-aligned Implicit Function (PIFu) representation for 3D deep learning from a single or a large number of input photographs for the difficult task of textured surface inference of clothed 3D people. Although the majority of efficient deep learning methods for processing 2D images (such as semantic segmentation, 2D joint detection, etc.) use "fully-convolutional" network designs that keep the input and output's spatial alignment in place, this is particularly challenging when processing 3D images[3]. Voxel representations can be fully convolutionalized, however, because of their intrinsic memory needs, they are unable to create fine-scale detailed surfaces.

2 Related work

The GAN is incredibly efficient for creating high-fidelity images. Since the first researcher suggested the first generative model in 2014, other modifications of GAN have been developed [4]. As an alternative to unconditional generation, conditional GANs [5] were suggested to

create images based on standards such as segmentation masks [6–8] and natural language. The previous system that has been proposed uses human gestures and language as inputs to produce conditional visuals. The Variational Autoencoder (VAE) picture generation paradigm is an alternative to GAN. 3D representations of humans: For tasks involving humans, 3D human representations are essential tools. Parametric models are developed by [9] for the explicit modeling of 3D humans.; simulate human appearances. Although less realistic, parametric modeling offers reliable control over the human model. The number of publications on human Neural Radiance Fields (NeRF) has also skyrocketed in tandem with the growth of NeRF. For a variety of down-stream tasks suggest learning modal-invariant human representations. Several large-scale multi-modal 4D human datasets. Human Image Manipulation and Synthesis. Pose transfer's purpose The goal of this system 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. Anomaly Detection Generative Adversarial Network (ADGAN) was suggested by [10] for controlled person image creation.

2.1 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. This deals with 3D human dataset to build 3D human geometry. Some people also try to train 3D human GANs using just 2D human image libraries. The Convolution Neural Network (CNN)-based neural renderers used by [11–13] cannot ensure 3D consistency. Human NeRF, which only trains at low-resolution images, is used for this purpose suggesting boosting the resolution by super-resolution, although this still doesn't yield excellent outcomes. 3D-aware GAN. In terms of creating 2D images, the GAN has achieved remarkable success. The generation of 3D awareness has also received a lot of attention [14, 15]. Voxels are used by meshes are used by the researcher to help the 3D-aware generation. Many researchers have developed 3D-aware GANs based on NeRF thanks to recent advancements in the technology. employ 2D decoders for the 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 [16]. We proposed a powerful 3D human representation to enable training at high resolution.

3 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 work are as follows.

3.1 Dataset used

Deepfashion-multimodal can be used for text-driven human image creation, text-guided human image manipulation, skeleton-guided human image creation, human

Table 1 Deepfashion-multimodal dataset overview

| S.No | Parameters | Values |
|------|--|------------------------|
| 1 | Size | 12 GB |
| 2 | Number of high-resolution human images | 44,096 |
| 3 | Classes | 24 |
| 4 | Resolution | 1024×512 |
| 5 | Key points | 21 |
| 6 | Annotations | Shape, Fabric, Clothes |

Table 2 BUFF dataset overview

| Detail | Description |
|-----------------|--|
| Dataset Name | BUFF |
| Type | 4D dataset |
| Evaluation | Enables quantitative evaluation |
| Performance | Outperforms previous state-of-the-art qualitatively and quantitatively |
| Subjects | 5 (3 male, 2 female) |
| Clothing Styles | a) t-shirt and long pants b) soccer outfit |

posture estimation, human image captioning, multi-modal learning for human images, recognition of human attributes, and prediction of human parsing. Link: drive.google.com/drive/folders/1An2c_ZCkeGmhJg0zUjtZF46vyJgQwlr2. Table 1 shows the description of the dataset.

BUFF(Bodies under flowing fashion, 4D dataset) (Zhang et al. [5]) High-quality 4D dataset containing ground truth 3D shape of humans wearing apparel. Five subjects, three men and two women are wearing two different outfits make up the BUFF dataset. They move their hips, bend their heads to the left, twist their shoulders, and grind their shoulders. Link: <https://buff.is.tue.mpg.de/>. Table 2 shows the description of the BUFF dataset.

3.1.1 SMPL

SMPL is a detailed 3D model of the human body that was created using thousands of 3D body scans. (Skinned Multi-Person Linear Model) It is built on skinning and blending shapes. This website offers learning materials for SMPL, including code for utilizing SMPL in Python, Maya, and Unity, as well as sample FBX files containing animated SMPL models. Link: <https://star.is.tue.mpg.de/>. Table 3 shows the SMPL dataset description.

FashionMNIST is a fashion video dataset containing 500 sequences of models posing in front of the camera. Link:-(<https://github.com/zalandoresearch/fashion-mnist>). Table 4 shows the FashionMNIST Dataset overview.

Table 3 SMPL dataset overview

| Field Name | Description |
|--------------|---|
| Dataset Name | SMPL |
| Description | The SMPL (Skinned Multi-Person Linear) model is a statistical model of the human body that can represent a wide range of body shapes and poses. The SMPL dataset contains 10,000 SMPL models, each with a different set of body shape and pose parameters |
| Source | Max Planck Institute for Intelligent Systems |
| Publication | "SMPL: A Skinned Multi-Person Linear Model" by Loper et al., in ACM Transactions on Graphics, 2015 |
| Data Format | Each SMPL model is represented as a set of 6890 vertices and corresponding face indices, along with body shape and pose parameters. The data is stored in the OBJ file format |
| Data Size | Approximately 10 GB |
| Use Cases | The SMPL dataset is widely used in computer vision and graphics research for tasks such as 3D human pose estimation, tracking, and animation |
| Citation | Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, 34(6), 248:1–248:16 |

Table 4 FashionMNIST dataset

| S.no | Parameters | Values |
|------|-------------------|-----------------|
| 1 | Size | 30 MB |
| 2 | Number of Records | 60,000 |
| 3 | Classes | 10 |
| 4 | Resolution | 28×28 Greyscale |
| 5 | Accuracy attained | 97 – 99.7% |

Table 5 Overall dataset overview

| Dataset | Resolution | Training | | | | |
|------------------------|------------|----------|------------------|-------|---------|-------|
| | | Class | Number of images | Media | Mean | Max |
| Deepfashion-multimodal | 1024×512 | 24 | 44,096 | 12 | 6,000 | 300 |
| Fasion-MNIST | 28×28 | 10 | 60,000 | 10 | 72.9403 | 255 |
| SMPL | 100×101 | 48 | 293,008 | 887 | 1,721 | 5,600 |

Saeur et al. [17] 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 a need for further new, high-quality images of various designs, styles, and patterns we can still use the technique of Text2human [18] which is the most optimized to date as the Nvidia Tesla V100 GPUs were used in training the models. The only variation that we will be implementing is 3D model generation to give so many poses and help in an accurate way to increase the views of different poses. Now to attain this images have to be generated from text once the images generated from this can be directly fed to generate 3D models [19]. The 3D model generation is done by several methods of inverse graphics as it aims to recover 3D models from 2D observations. Table 5 shows the

description of all Dataset used in this work. Figure 1 shows the system architecture for Text to fashion image generation. Figure 2 denotes the overall system design.

The various modules involved in the work are:

1. Data pre-processing
2. Text to clothes texture, shape analyzer
3. 2D Image to 3D human model generator

3.2 Data preprocessing

Refine the dataset by removing any missing, damaged, or irrelevant photos and annotations. Augment the dataset by applying transformations like rotation, flipping, scaling, and shearing to the images, enhancing its size and diversity to better reflect real-world

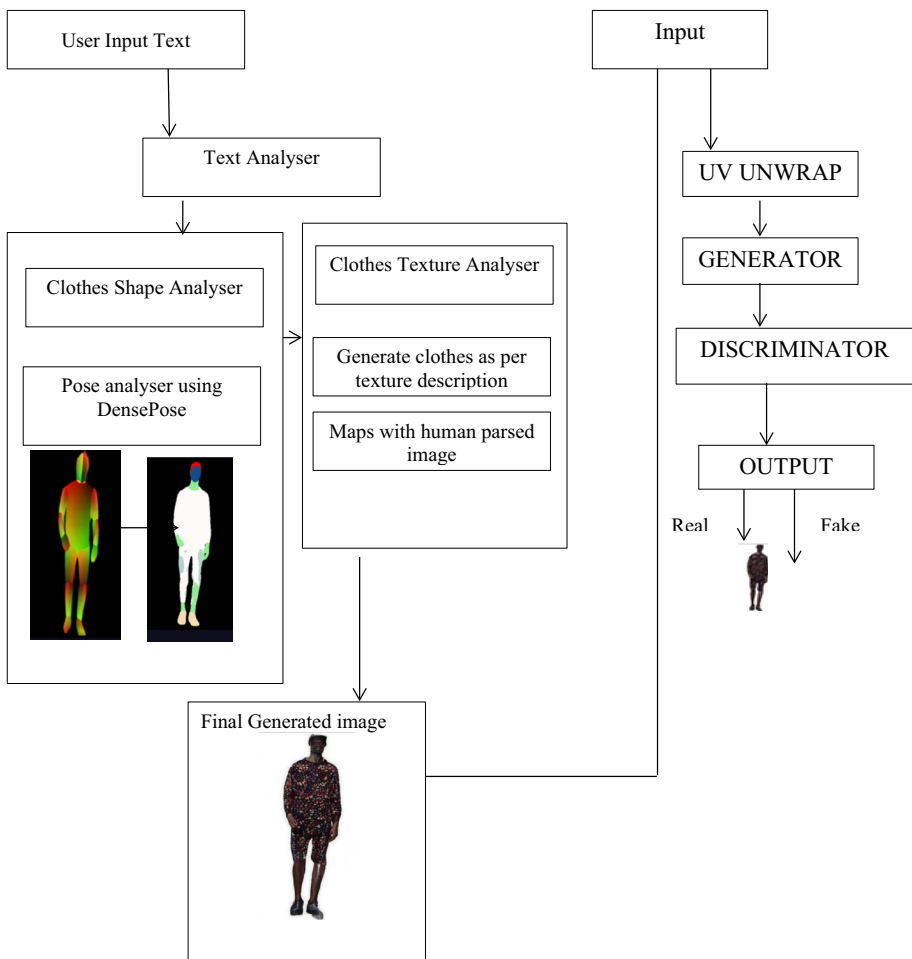


Fig. 1 System architecture diagram for text to fashion image generation

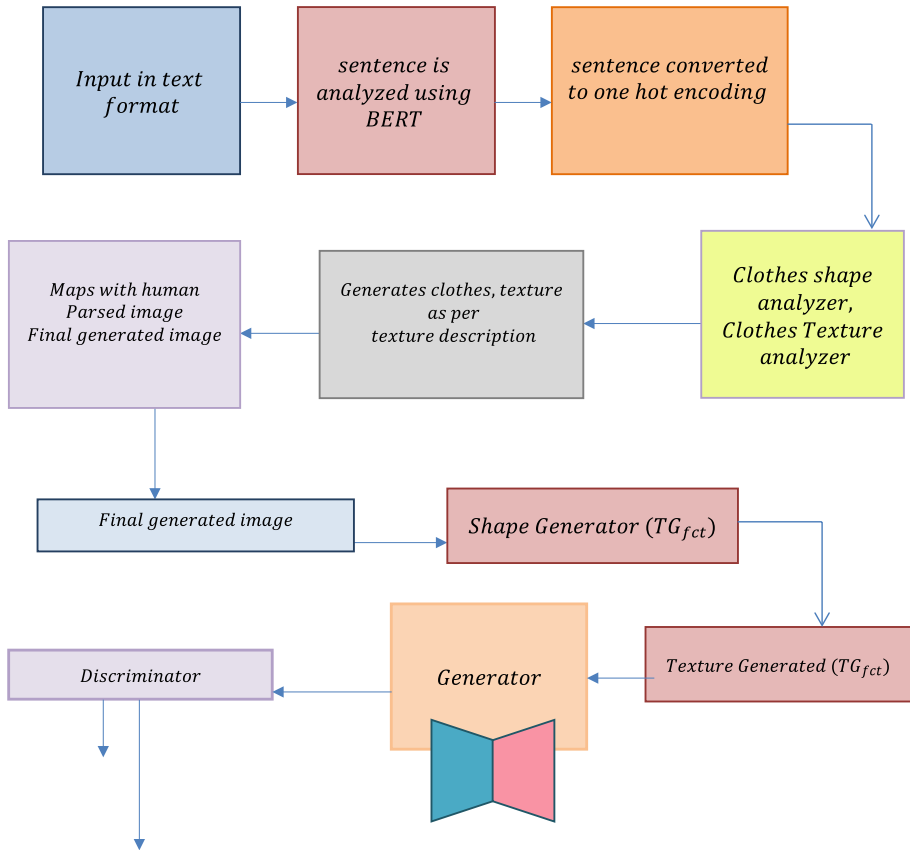


Fig. 2 Overall system design

scenarios. Standardize the photographs to a uniform scale, such as 256×256 , and crop them to eliminate any extraneous or distracting elements, thereby improving the effectiveness and efficiency of machine learning models. Normalize and standardize the pixel values of the photos to minimize variations in lighting and color, enhancing the data's comparability and consistency. Additionally, remove incomplete and half-body images from the Dense-Pose dataset.

3.3 Text to clothes texture, shape analyzer

The majority of techniques currently employed in this field construct their editing models using examples of the desired clothing, typically manipulating and fitting them onto the input image. While these methods enable image modification through more natural language-based descriptions of the intended attire, text-conditioned fashion-image editing remains popular due to its ability to provide a compelling alternative to example-based editing approaches. Moreover, there are numerous applications for generating human images, including human pose transfer, virtual try-on, and animations. Particularly in terms of applications and user interactions, facilitating the easy control of synthesized full-body

human images is desirable, as they represent a form of human-related media with richer, more varied, and finely detailed material. However, current methods for generating human body images often lack diversity in clothing, frequently producing items with basic patterns and limited control over garment textures. To address this, additional detailed annotations are required for producing clothing based on textual specifications. Given the complexity of human body images, it is challenging to encompass all relevant aspects within a single generative model. Stage I of our approach involves constructing a human parsing mask with diverse clothing shapes based on the provided human pose and user-specified descriptions of clothing shapes. Subsequently, Stage II enhances this mask by incorporating a variety of clothing textures derived from text descriptions. Figure 3 denotes the flow diagram and Fig. 4 denotes the System Design of Text-driven clothed human image synthesis with 3D human model estimation. Figure 5 shows the System Design of user input text to Human image synthesis. Figure 6 denotes the GAN Architecture.

3.4 2D image to 3D human model generator

The creation of 3D models can be achieved through the utilization of both generative network training techniques and 3D human representation. We employed the Pixel-aligned Implicit Function (PIFu) representation for deep learning in 3D, utilizing either a single photograph or a multitude of images to tackle the complex task of inferring textured surfaces of clothed 3D individuals. While many effective deep learning methods for analyzing 2D images, such as semantic segmentation and 2D joint detection, utilize "fully convolutional" network architectures that maintain spatial alignment between input and output, this presents a significant challenge in the context of 3D image processing. Voxel representations can indeed be fully convolutionalized, but due to their inherent memory requirements, they often fail to produce detailed surfaces at fine scales.

Figure 2 shows the system design. The block diagram states and shows the description regarding the clothes humans are to be fed into and this is further analyzed by the BERT and the words are one-hot-encoded which are mapped with the textual descriptions of the dataset's annotations. Once all these are done the Dense-Pose dataset that contains various images of human poses in thermal image format is also passed after feeding the textual

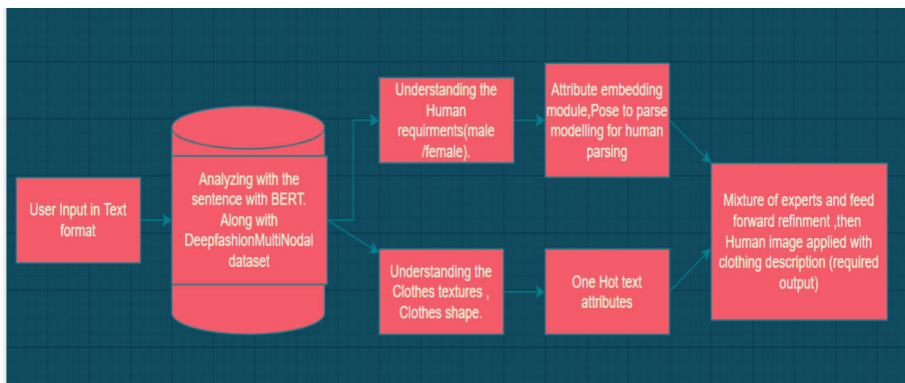


Fig. 3 Overall flow diagram

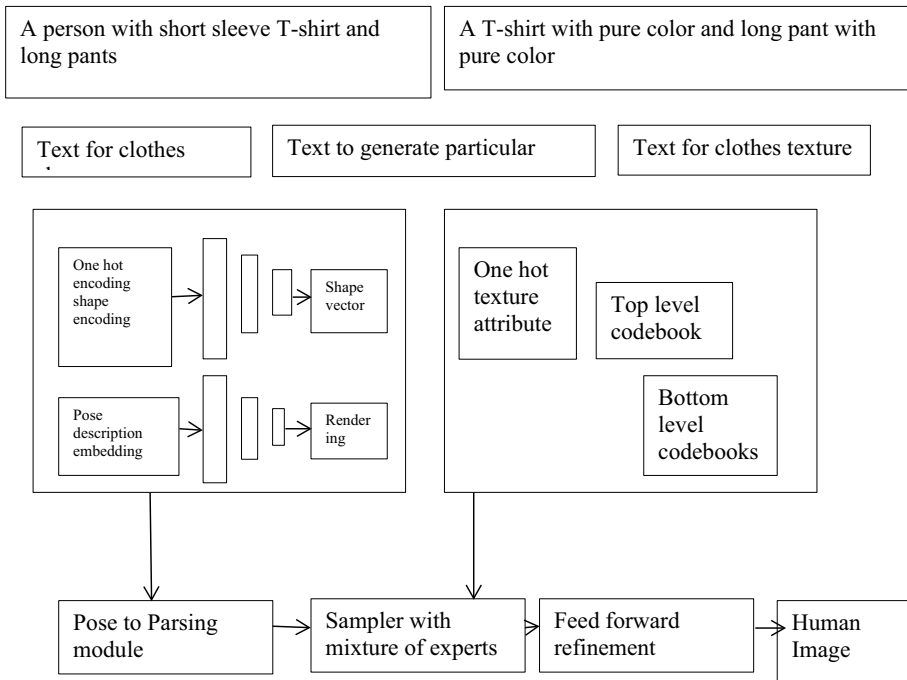


Fig. 4 System design of text-driven clothed human image synthesis with 3D human model estimation

descriptions into the model this helps in mapping the shape of the clothes that must be embedded on the human along with the clothes texture that must be embedded as well is carried out. Once the required image is generated further proceed in the conversion of 2D image to 3D image. The module that deals with generating 3D models is composed of 2 parts namely the shape construction and texture/ surface construction, once the image is passed the required space to build the 3D model is calculated, it's the probability that is calculated not the exact space view. Once all the procedures are completed the required 3D model is finalized and produced as the final result.

3.5 3D-generative adversarial network

The working of 3D GAN involves carrying out the transformation of the 2D image input into a 3D model. The following image shows that the Z_{img} that is passed as input in the next step is attached with the latent space $4 \times 4 \times 4$ then goes from $512 \times 4 \times 4 \times 4$ then transforms to $256 \times 8 \times 8 \times 8$ then to $128 \times 16 \times 16 \times 16$ further proceeds to $64 \times 32 \times 32 \times 32$ finally forming the $64 \times 64 \times 64 \times 64$. Figure 7 describes the working of 3D-GAN.

Before getting into VQ-GAN it is important to know about GAN and its algorithms (Algorithm 1 and 2). Figure 8 shows the loss curve for each epoch. It shows the result of Curvature Regularized VAE that shows more and more steps increase (epochs) the Loss decreases. Figure 9 shows the DCGAN Loss curve for each epoch. It shows the result of DCGAN that shows more and more steps increase (epochs) the Loss decreases.

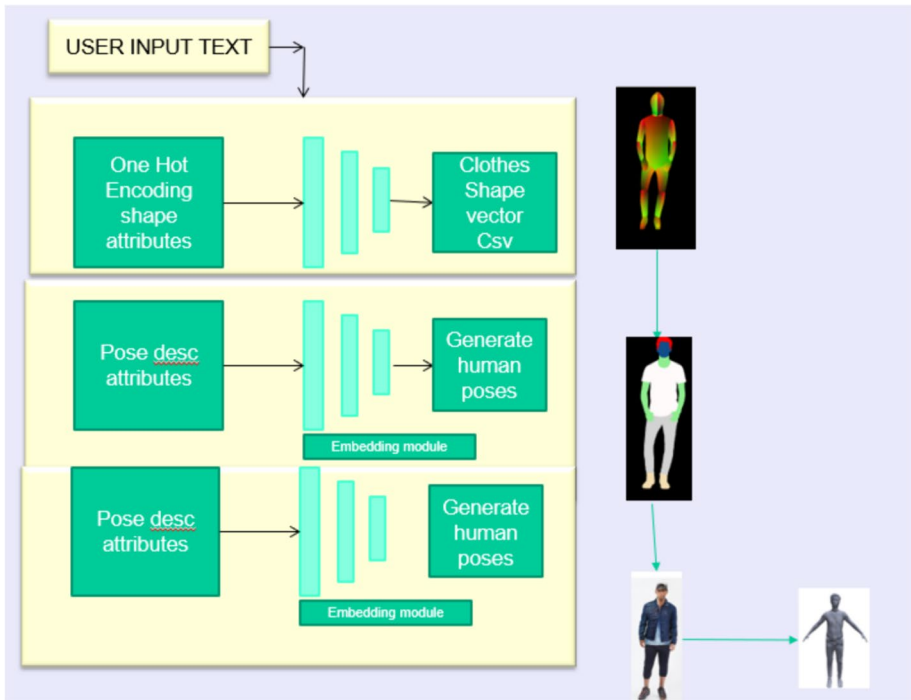


Fig. 5 System design of user input text to human image synthesis

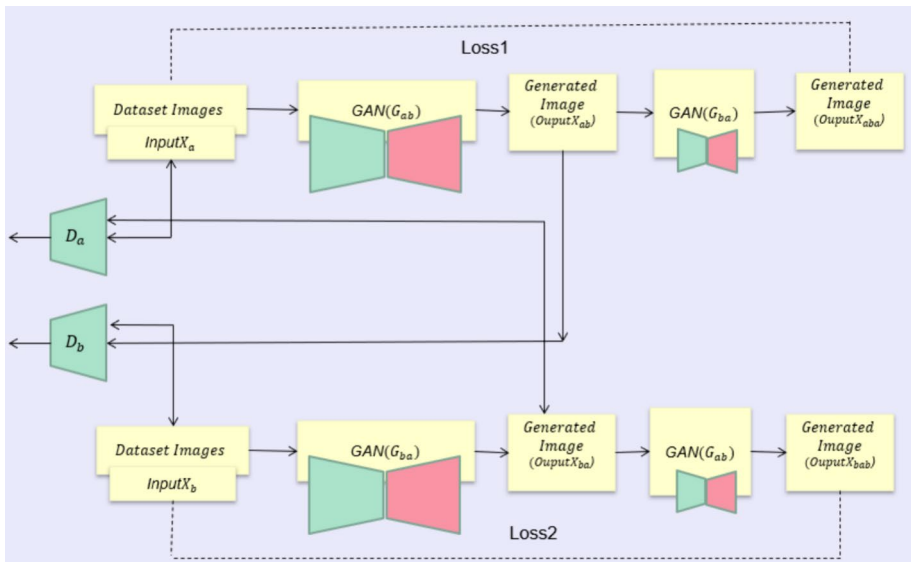


Fig. 6 GAN architecture

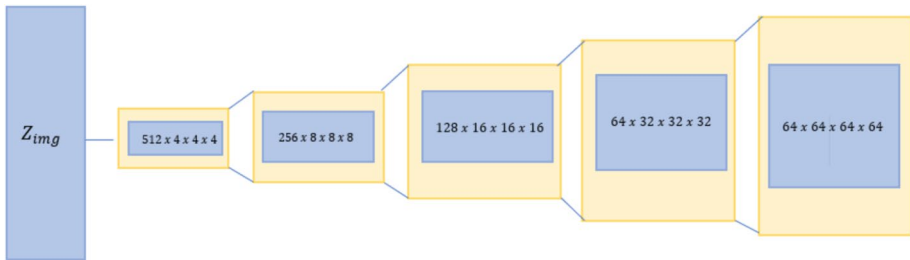


Fig. 7 3D-GAN implementation

Input: Number of training iterations

Output: Trained GAN model

Procedure:

1. For each training iteration:
2. For a fixed number of steps (k):
3. Sample a minibatch of m noise samples $\{z(1), \dots, z(m)\}$ from the noise prior distribution $pg(z)$.
4. Sample a minibatch of m examples $\{x(1), \dots, x(m)\}$ from the data generating distribution $pdata(x)$.
5. Update the discriminator by ascending its stochastic gradient:

$$\nabla\theta d \frac{1}{m} \sum_{i=1}^m [\log \log D(x^{(i)}) + \log \log (1 - D(G(z^i)))] \dots \dots (1)$$

6. Sample a new minibatch of m noise samples $\{z(1), \dots, z(m)\}$ for sampling.
7. Refresh the generator by lowering its stochastic gradient:

$$\nabla\theta d \frac{1}{m} \sum_{i=1}^m [\log \log (1 - D(G(z^i)))] \dots \dots (2)$$

3.6 Vector quantized variational autoencoder (VQ-VAE)

In VQ-VAE, the encoder transforms input data into continuous latent vectors, which are then quantized to discrete representations using a predefined codebook. The important point is that the images must somehow be expressed as sequences to make use of the computationally expensive self-attention process in high-resolution synthesis. The encoder extracts an encoding Z hat and quantizes it to Z q using the closest codebook entry rather than utilizing pixels or patches as tokens (as in VQ-VA) [20]. The image can then be rebuilt by the decoder starting with the quantization. Except for two key modifications in the loss applied, this portion of training required to acquire the codebook representation is nearly identical to the VQ-VAE one. As was stated in the

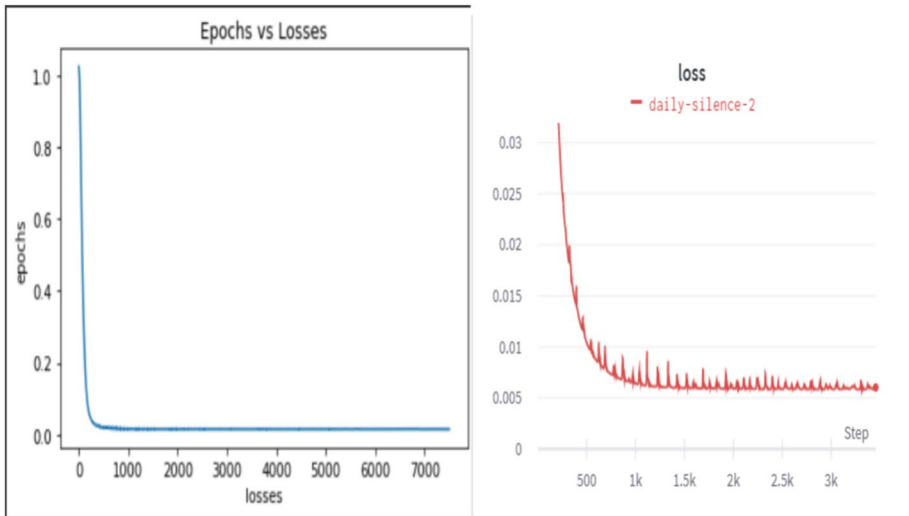


Fig. 8 Loss curve for each epoch (Curvature Regularized VAE) graph obtained 50 epochs

previous section, the Mean Squared Error (MSE) plus the two alignment losses made up the three terms that made up the loss employed in VQ-VAE [21]. Perceptual loss, which is essentially the MSE computed on internal representations of the images rather than two images, is used in place of the MSE in this situation. For instance, when

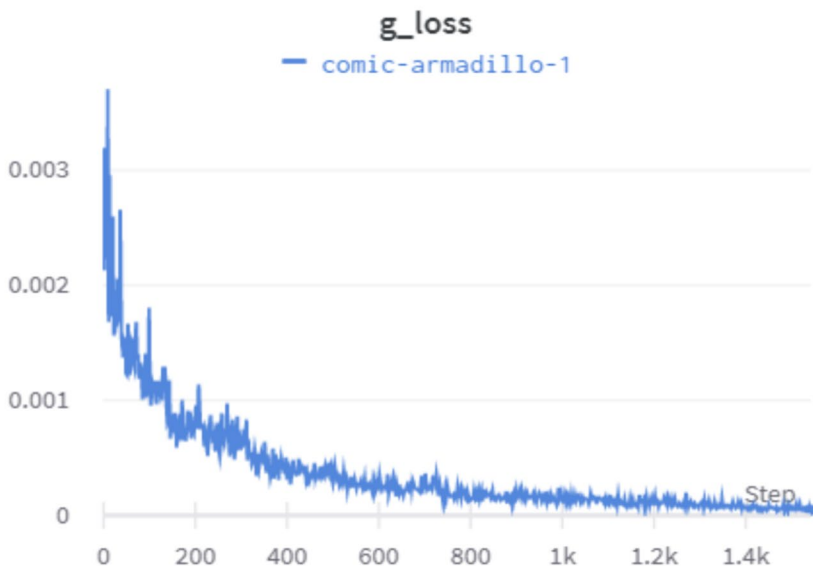


Fig. 9 DCGAN loss curve for each epoch

two images are put through a CNN, the n-th layer features are extracted and the MSE is compared [22]. To overcome the VQ-VAE blurring problem, they also include an adversarial loss (the traditional loss used in GANs, where a generator and a discriminator play in a minimax game), with a prediction of real/fake not for the entire image but rather for individual patches [23]. Figure 10 shows the VQGAN implementation and Fig. 11 shows the detailed VQGAN implementation. Figure 12 shows the Stack-hour-glassed structure. Figure 13 shows the Cloth texture Construction and Shape construction.

Require: Functions $Encode_{top}, E_{bottom}, D$, (batch of training images)

1: $top_code \leftarrow Encode_top(x)$ $h_{top} \leftarrow E_{top}(x)$

▸ *quantize with top codebook eq 1*

2: $e_{top} \leftarrow Quantize(h_{top})$

3: $bottom_code \leftarrow Encode_bottom(x, e_{top})$ $h_{bottom} \leftarrow E_{bottom}(x, e_{top})$

▸ *quantize with bottom codebook eq 1*

4: $e_{bottom} \leftarrow Quantize(h_{bottom})$

5: $x_hat \leftarrow Decode(e_{top}, e_{bottom})$ $\hat{x} \leftarrow D(e_{top}, e_{bottom})$ ▸ *Loss according to eq 2*

6: $\theta \leftarrow Update(L(x, x_hat))$ $\theta \leftarrow Update(L(x, \hat{x}))$

3.7 Unified fashion generative adversarial network(uFashGAN)

The uFashGAN architecture consists of a GAN that takes an input image and generates a corresponding image, which is then fed into another GAN layer. This generated image is compared to the original input image to compute the loss, and the first generated

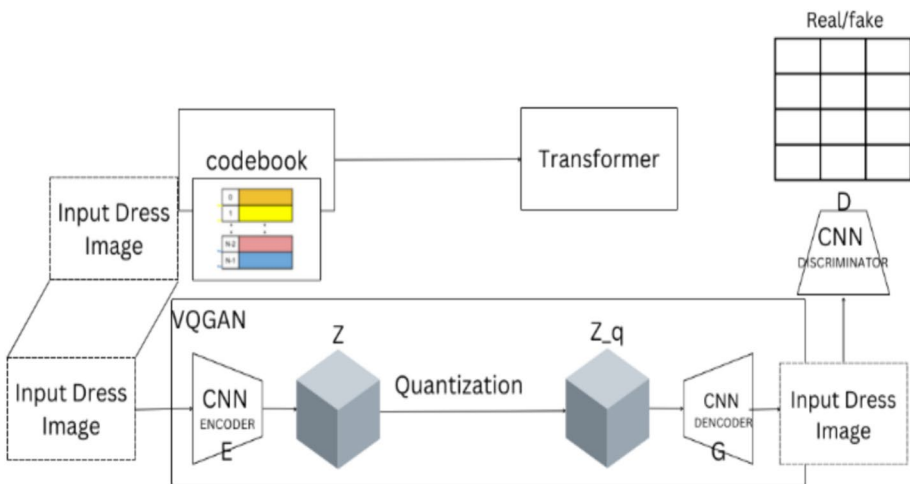


Fig. 10 VQGAN implementation

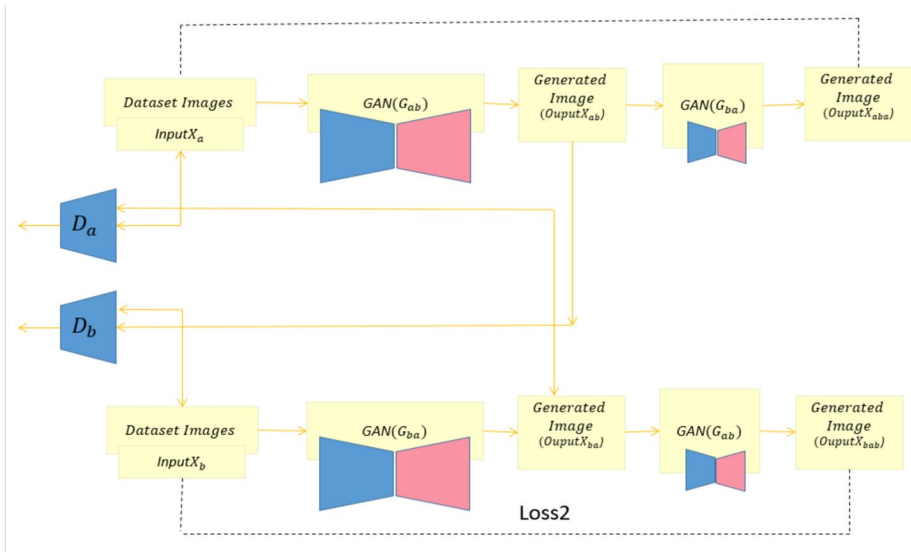


Fig. 11 VQGAN implementation detail

image is sent back to the discriminator along with other parts of the GAN module. This process is repeated several times to produce a diverse set of results.

Input: Number of epochs, number of batches

Output: Trained Generators

for each epoch:

for each batch:

Sample Image $\leftarrow \{x^{(i)}\}_{i=1}^m$ where X is input image

Generate m_1 image samples using $G_{ab}(x)$

Feed m_1 to Generator $G_{ba}(x)$

Generate image X_{aba}

Generate m_2 samples using $G_{ba}(x)$

Feed m_2 samples to Generator $G_{ab}(x)$

Generate image X_{bab}

Calculate the Loss between Input X and generate images X_{aba}, X_{bab}

Update the Discriminators D_x and D_y

$$\max_{D_x} L_{GAN}(F, D_x, X, Y)$$

$$\max_{D_y} L_{GAN}(G, D_y, X, Y)$$

(3)

Update the Generators G and F

$$\min_{G, F} L(G_{ab}(x), G_{ba}(x), D_x, D_y)$$

3.8 Stacked-hourglass

Hourglass module is a symmetrical encoder-decoder architecture with skip connections that help in preserving spatial information. The Hourglass module uses residual connections to improve the flow of gradients during training and allows the network to learn deeper features without the vanishing gradient problem. Stacked Hourglass Network (SHN) architecture is designed as a multi-stage process, where the output of one Hourglass module is used as input for the next module. The multi-stage approach helps to refine the estimates of human pose over time and improve the overall accuracy. The SHN uses intermediate supervision to train the network at each stage of the Hourglass module. Overall, the SHN is a powerful architecture for human pose estimation. Figure 12 shows Stacked hourglass Architecture. This Stacked-hour-glass structure here represents how the input image passed will generate the real human model. For example, to generate a real human 3D model of every hourglass that contains 1×1 convolution networks help in the overall reconstruction, here the first block tries to reconstruct the head followed by hands then followed by the body, and ends with legs thus the whole body the difficulty that stacked hourglass faces whilst developing of the 3D model is it must develop all features by guessing what it would look like and generate the complete model. The stacked hourglass is composed of the following.

Input: The Stack Hourglass architecture is commonly used for image segmentation, with the input usually being an image that requires segmentation. The size of the input image can differ based on the specific task and the model's needs.

Down-sampling: In the Stack Hourglass architecture, the input image is typically passed through multiple down-sampling layers, which are commonly composed of convolutional layers followed by max pooling. This process is used to decrease the spatial resolution of the image, allowing the network to capture high-level features and coarse details more effectively.

Up-sampling: The Stack Hourglass architecture incorporates several up-sampling layers, which often come in the form of deconvolutional or transposed convolutional layers, to boost the spatial resolution of the feature maps that were generated by the down-sampling layers. This is crucial for the network to capture small-scale details and preserve spatial information.

Intermediate output: The Stack Hourglass architecture produces multiple intermediate outputs, which are segmentation maps with different spatial resolutions. These intermediate outputs are generated at different stages of the network, typically after each down-sampling and up-sampling block. These intermediate outputs capture features at different scales and provide different levels of detail.

Intermediate supervision: To ensure effective training of the Stack Hourglass architecture, intermediate supervision is applied at each stage to provide feedback to the network. This is achieved by computing the loss between the intermediate output and the ground truth segmentation map, and then backpropagating the error through the network. By doing this, the network can learn and improve the accuracy and detail of the segmentation maps at each stage of the architecture. In summary, the Stack Hourglass architecture is a highly effective and adaptable technique for image segmentation that incorporates down-sampling, up-sampling, and intermediate supervision to capture features at various scales and generate precise segmentation maps.

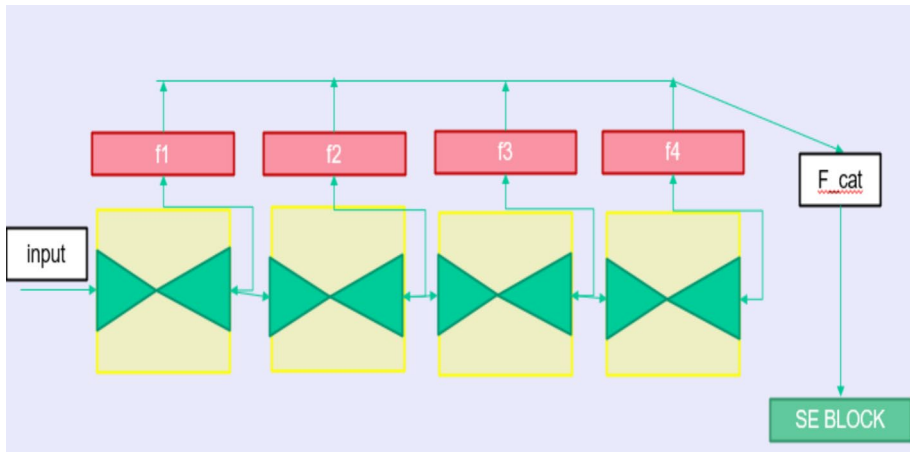


Fig. 12 Stack-hour-glassed structure

Additionally, the architecture may use a Squeeze-and-Excitation (SE) block to produce the final output (Fig. 13). The entire module is divided into 2 parts that hold the of the image and shape construction obtained from the input image. The image encoder passes the respective information here the information of the cloth texture along with the color is also obtained and this results in $TextureGenerated(TG_{fct})$ and the shape of the human in the image is passed to shape construction where the respective 3D model generation for that given pose is generated. Then both the Cloth Texture construction and Shape construction are combined to form the complete 3D clothed human model. Figure 14 describes the Overview of image-to-3D model conversion based on Z space.

- cloth texture construction
- shape construction

3.9 Single and multi-viewpoint

To begin with, the single-view surface reconstruction involves carrying out the probability field over the 3D space and finding the chance iso-surface. Field with the help of the Marching cube algorithm. Numerous views of the person can provide additional coverage of that person. The Pixel-Aligned Implicit function helps in providing more views so that the shape and texture can be enhanced much well and the detailing can be perfect alongside. Figure 15 shows Single and multi-view points.

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) [24]. The appropriate human image I where $I \mathbf{RH} \times \mathbf{W} \times \mathbf{3}$ should be produced using a human pose $P \mathbf{RH} \times \mathbf{W}$ texts for clothing shapes $Tshape$, and texts for clothes textures $Texture$. We wanted to synthesis the human parsing map $S \mathbf{RH} \times$. using a human pose P and words on clothing forms. Texts are converted to a series of clothing shape properties called $\{a1, \dots, ai, \dots, ak\}$, where ai is represented by

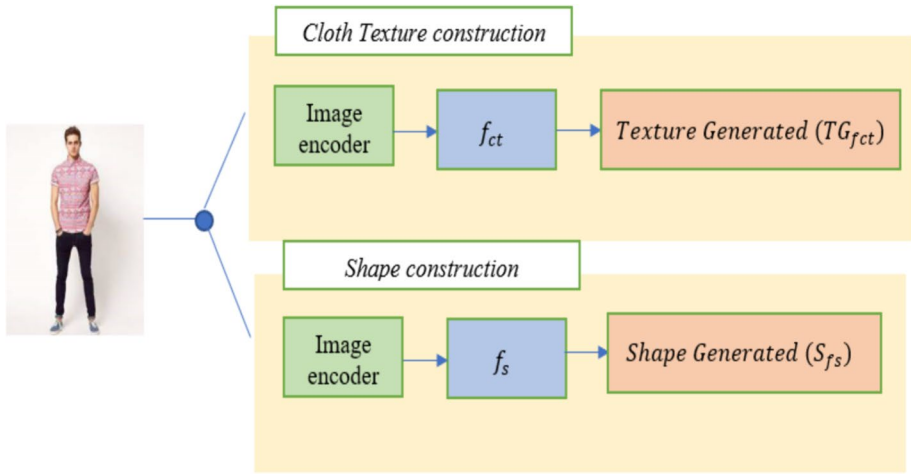


Fig. 13 Cloth texture construction and shape construction

the values 0 through 1 and class number C_i . The Attribute Embedding Module receives the attributes after which a shape attribute embedding is produced.

$$f_{sh} = Fusion([E1(a1), E2(a2), \dots Ei(ai), \dots Ek(ak)]) \tag{4}$$

Once this is complete further move to implementation using VQ-GAN The VQGAN is a GAN architecture that may be used to learn from prior data and produce new images. The feature map of the image data is initially directly fed to a GAN to encode the feature map of the visual sections of the images [25]. A codebook, or dictionary of codes, is created from the vector quantized data and stored.

The loss produced in GAN is LGAN:

$$LGAN(N, D) = [\log D(x) + \log(1 - D(x))] \tag{5}$$

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

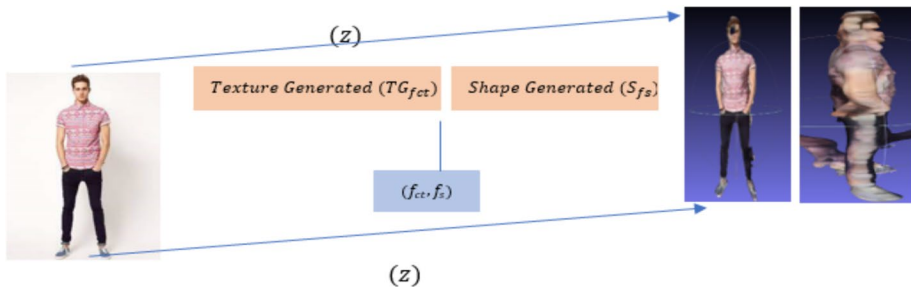


Fig. 14 Overview of the image to 3D model conversion based on Z space

$$z_q = q(\hat{z}) := \operatorname{argmin}_{z \in Z} \| \hat{z}_{ij} - z_k \| \tag{6}$$

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

4 Generation of 3D models

As for the construction of 3D models, the pixels of 2D images must be aligned with the global context of their corresponding 3D object with a stacked hourglass approach and for image construction by applying the method of image encoding CycleGAN. A collection of local bounding boxes B is used to define the compositional human NeRF representation as $F\Phi$ [26]. We employ a subnetwork $f_k \in F\Phi$ to model the local boundaries for each body part k , as seen in Fig. 16, box $\{bk \text{ min}; bk \text{ max}\}$. Regarding a specific point x_i in the canonical coordinate system,

The matching radiance c_k and density σ_k are in the direction d_i and falling inside the k -th bounding box, respectively, is challenged by

$$\{c_i^k, \sigma_i^k\} = F_k(x_i^k, d_i), \text{ where } x_i^k = \frac{2x_i - \{b_{min}^k + b_{max}^k\}}{b_{max}^k - b_{min}^k} \tag{7}$$

All this can lead to the highly detailed clothed humans that accurately replicate all the texture and geometry from a single image in comparison with existing 3D deep learning models it provides a highly detailed result [27]. This is developed using a Deep fashion dataset that contains a variety of clothing types of huge range. Now comes the important implementation along with the enhancing phase, The input image of a clothed human is fed into the model that carries out the reconstruction of the given image as a 3D model while replicating and preserving the geometry and texture details present in the image [28]. The result ends up as a clothed human that is an accurate representation of the given 2D clothed human image. For this, we imbibe the techniques of the Pixel-Aligned Implicit function. This tends to be heavily memory efficient. The proposed function contains a

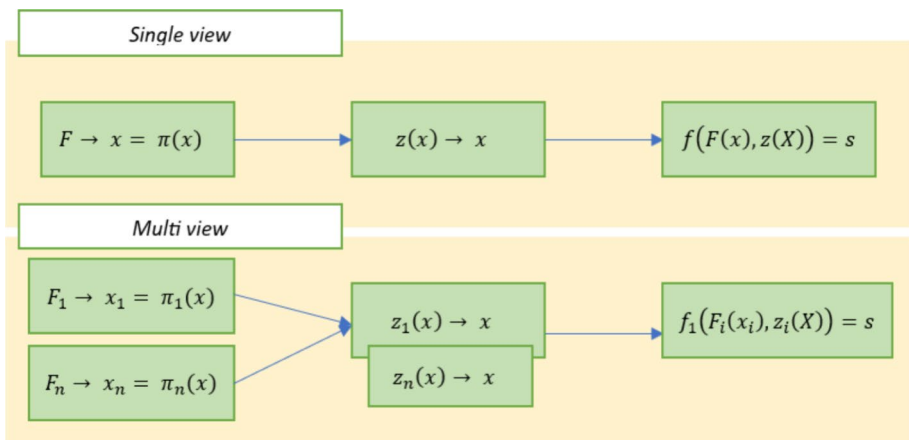


Fig. 15 Single and multi-view of point

full convolution image encoder g that is combined with the continuous implicit function f that has its representation using several (Multi-Layer Perceptrons) MLP. All this is implemented to carry out the conversion of a 2D clothed human image to a 3D model Here X is assumed as a 3D point,

$$f(F(x);z(X)) = s : s \in \mathbb{R} \tag{8}$$

happens to be the surface representation, the $x = \pi(X)$ is the 2D representation, and projection of the clothed human image and $z(x)$ is used as the multiple view angle that helps in capturing multiple views of the same point X . Further the $g(I(x))$ acts as the image feature at point x [29]. As the entire module is continuous and not broken down into several parts this helps in the reconstruction of shape and texture with a high level of detailing along with memory efficiency.

4.1 Single-view and multiple-view surface reconstruction

To begin with, the single-view surface reconstruction involves carrying out the space probability for the 3D space and obtaining the iso-surface of the probability field with the help of the Marching cube algorithm. Numerous views of the person can provide additional coverage of that person. The Pixel-Aligned Implicit function helps in providing more views so that the shape and texture [30–37]. states the different methodologies for processing video Fig. 16 shows the Shows Clockwise Adam optimizer, RMSProp, SDG, Adagrad, and Adamax) detailing can be perfect alongside. With the help of this model, we can directly predict the RGB colors of the surface geometry. This favors self-occlusion and arbitrary topology in the texturing of shapes. It is challenging to extend a model to forecast colors since RGB colors are only defined on the surface of the 3D space, unlike the 3D occupancy

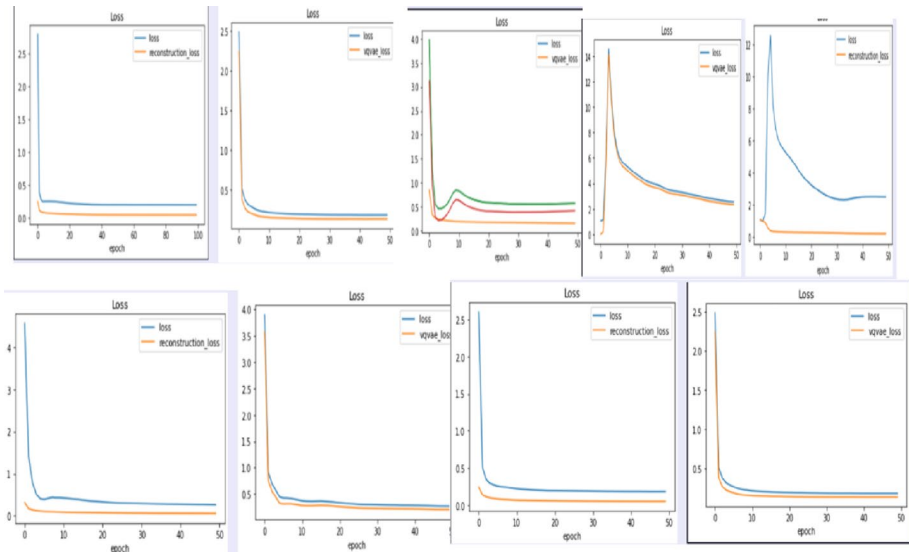


Fig. 16 Shows Clockwise Adam optimizer, RMSProp, SDG, Adagrad, Adamax)

field, which is known throughout the entire 3D area. Here, we emphasize the evolution of the model's network architecture and training methodology.

4.2 Marching cube algorithm

To begin with, the single-view surface reconstruction involves carrying out the probability field over the 3D space and finding the chance iso-surface. field with the help of the Marching cube algorithm. Numerous views of the person can provide additional coverage of that person. The Pixel-Aligned Implicit function helps in providing more views so that the shape and texture can be enhanced much better and the detailing can be perfect alongside.

Algorithm Marching cube algorithm (V, h):

Input: Volume data V iso-value h

Output: List of vertices and their respective normals for rendering

// \mathbf{tv} - triangle vertices, \mathbf{vn} - vertex normal

For each voxel (cube) v of V do:

1. Compute the cube's index ci by comparing its eight possible values with the iso-value h .
2. Verify the edge list from a lookup table using the calculated index.
3. Find the intersections of surfaces and edges using linear interpolation based on the scalar values of each edge vertex.
4. For each cube vertex, compute a unit normal using the central difference approach.
5. Add the normal interpolation to each triangle vertex.
6. Return the list of triangle vertices \mathbf{tv} and their corresponding vertex normals \mathbf{vn}

End for

5 Experimental evaluation

The performance metrics used are the Inception score, Fréchet Inception Distance, and Structural Similarity Index.

5.1 Inception score

The Inception Score (IS) is a metric for evaluating the quality of generated images. It measures the balance between the diversity and quality of the generated images. The IS is defined

as the exponential of the entropy of the conditional class distribution $p(y|X)$ (i.e., the predicted class probabilities given the generated image), multiplied by the marginal entropy of the class distribution $p(y)$. In mathematical terms, the IS can be given in Eq. (1):

$$IS(G) = \exp\left(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|X) || p(y))\right) \tag{9}$$

5.2 Fréchet inception distance score

For assessing the caliber of generated images, another statistic is the Fréchet Inception Distance (FID). It calculates the separation in feature space between the distributions of the generated images and the actual images. In particular, the FID determines the separation between the mean and covariance of the activations of the pre-trained Inception-v3 network on ImageNet. A definition of the FID is:

$$\left| \mu_r - \mu_g \right|^2 + T_r(\sum_r + \sum_g - 2(\sum_r \sum_g)^{1/2}) \tag{10}$$

where μ_r and μ_g represent mean of real and generated image feature representation. $\left| \mu_r - \mu_g \right|^2$ represent squared Euclidean distance (or L2 norm) between the means of the real and generated images' feature representations. T_r represent trace of a matrix, which is the sum of the elements on the main diagonal of the matrix. \sum_r and \sum_g covariance matrix of the real and generated images feature representations. $(\sum_r \sum_g)^{1/2}$ represent the matrix square root of the product of the covariance matrices of the real and generated images' feature representations.

5.3 Structural similarity index SSIM

Structural Similarity Index (SSIM). SSIM is used as a metric to measure the similarity between two given images. Given in equation (3)

$$SSIM(x, y) = (2\mu_x\mu_y + c1)(2\sigma_{xy} + c2) / (\mu_x^2 + \mu_y^2 + c1)(\sigma_x^2 + \sigma_y^2 + c2) \tag{11}$$

with:

- μ_x the average of x;
- μ_y the average of y,
- σ_x^2 the variance of x;
- σ_y^2 the variance of y;
- $2\sigma_{xy}$ the covariance of x and y.

C1 and C2 are two variables to stabilize the division with a weak denominator; the dynamic range of the pixel-values (typically this is 2 bits per pixel -1); =0.01 and k2=0.03 by default.

Peak Signal-to-Noise Ratio (PSNR) is a technique to figure out the difference between a signal's maximum potential value (power) and the strength of the noise that distorts it and lowers the quality of its representation.

5.4 MSE of X channel

$$MSE_x = Nlm - ln - l, N = 1m * n \quad (12)$$

5.5 Total MSE

$$MSE_t = MSER + MSEG + MSE \quad (13)$$

5.6 Calculate PSNR

$$PSNR = 10 * \log_{10}(\text{MAXI})^2 / MSE_t \quad (14)$$

5.7 Fréchet inception distance: FID

$$d^2 = \|\mu_1 - \mu_2\|^2 + \text{Tr}(C_1 + C_2 - 2 * \text{sqrt}(C_1 * C_2)) \quad (15)$$

“ μ_1 ” and “ μ_2 ” refer to the mean of the individual features of the real and generated images, C_1 and C_2 are the covariance matrix for the real and generated feature vectors, represented as sigma.

5.8 Chamfer distance

The Chamfer Distance is a metric for measuring the similarity between two sets of points in a metric space. It calculates the average distance between the points in one set and their nearest neighbor in the other set. The Chamfer Distance can be written as:

$$D_{\text{chamfer}}(T, I) = \frac{1}{|T|} \sum_{t \in T} d_t(t) \quad (16)$$

5.9 SSIM index

A metric for determining how similar two photographs are is called SSIM. It gauges similarity in terms of structure, brightness, and contrast. The SSIM is expressed as follows: Luminance: By averaging over all of the pixel values, luminance is determined. Its symbol is (μ), and the formula is shown below.

$$\mu_x = \frac{1}{N} \sum_{i=1}^N x_i \quad (17)$$

5.10 Earth mover’s distance

The Earth mover’s distance, normalized by the total weight of the lighter distribution, is the smallest amount of work required to match x and y . However, because both distributions have the same total weight in this situation, there is no lighter distribution. Simply divide the labor by one distribution’s overall weight.

$$EMD = \sum_{i=1}^m \sum_{j=1}^n M_{ij} d_{ij} \quad (18)$$

The performance evaluation is carried out for the VQ-SEG using the FashionMNIST dataset. *VQ-SEG/VQ-VAE* is Specifically created for semantic segmentation tasks, VQ-SEG is a modified version of the VQ-VAE method. The addition of a segmentation head to the network, which enables it to anticipate the class labels of each pixel in an image, is the primary distinction between VQ-SEG and the original VQ-VAE. The encoder network is used by the VQ-SEG technique to first encode an input image into a lower-dimensional feature space. A vector quantization module is then used to quantize the feature space into discrete vectors known as embeddings. The decoder network then creates a reconstruction of the original image using the embeddings as its input. The performance of VQ-SEG on semantic segmentation tasks has been the most optimized, and it is a promising methodology for other computer vision problems requiring high-resolution predictions. Table 6 denotes the Comparison of various methods to generate 2D images. Table 7 denotes the Comparison of various methods generated by 3D models, Table 8 shows the score table for GAN models and accuracy at various Epochs and Table 9 shows the overall loss measurement. From the above GAN models and carrying out comparisons using various optimizers along with the given performances are compared. Table 4 shows the various outputs. Figure 17 shows the output of various inputs FashionMNIST.

The (Table 10) output shows various images for various poses using the Dense-Pose image. The input text once given is matched with the image name and generates the parsing image the evaluation metrics that have been carried for images and the 3D models are mentioned earlier now the comparison of how the proposed model performs can be inferred from the below tables along with side to side comparison of 3D models generated from the input 2D clothed image further on the introduction of new scores that has not been compared earlier also will be mentioned.

MISC is a causal discovery algorithm that integrates the maximal information coefficient (MIC) and Bayesian network structure learning to identify causal relationships between variables in complex systems. The algorithm incorporates a search strategy to identify the optimal causal network structure using a combination of greedy hill-climbing and tabu search. MISC outperforms other state-of-the-art causal discovery algorithms and has been used in real-world issues in various fields. A pose-conditioned VAE-based technique called Human-GAN generates a variety of human looks by sampling from a given distribution.

Table 11 denotes the Comparison of Various Methods for Generated Images. Table 12 denotes the Output of generated 3D models. Body-Net is a deep learning-based method for creating an approximate 3D mesh representation of the human body from a single RGB photograph. The model, which was developed using a sizable dataset of 3D body scans, can precisely predict body shape, position, and garment distortion from a single image. The scores are mentioned in (Table 13). SiCloPe: SiCloPe is a technique for 3D object reconstruction from a single RGB picture. The method relies on a mix of generative and discriminative models, where the generative model generates a 3D mesh and the discriminative model checks that the mesh produced is compatible with the input image. The scores are mentioned in (Table 13). VRN (Volumetric Regression Network): A neural network architecture was created to regress 3D shapes from 2D photographs. By regressing a 3D volume representation of the item, it can determine the 3D form of an object from a

single 2D image. The architecture processes the 3D volume representation using 3D convolutional neural networks (CNNs). The scores are mentioned in (Table 13).

The above result shows the reconstruction of the 3D model from the given 2D clothed human image that was generated from the textual descriptions. The mask-generated column refers to the inverse of the given image and also mentions within which the 3D model must be generated. The application of the Marching cube algorithm and stacked hourglass aids in the development of the 3D model. The only outliers in this are the half-body images and low-resolution images. Figure 18 shows the Adam optimizer displaying reconstruction loss, vq_vae loss where epoch = 100. Figure 19 represents the SDG optimizer displaying reconstruction loss, vq_vae loss where epoch = 50. Figure 20 mentions the RMSProp optimizer displaying reconstruction loss, vq_vae loss where epoch = 50. Figure 21 denotes the AdaGrad optimizer displaying reconstruction loss, vq_vae loss where epoch = 50.

The proposed approach combines several state-of-the-art techniques, including text-driven image synthesis, 3D model estimation, and dataset utilization, to advance the field of clothed human image synthesis and 3D model estimation. While individual components of the system, such as VQGAN for image generation, Pixel-Aligned Implicit function for 3D model reconstruction, and Marching Cube algorithm for mesh creation, are well-established. The novelty lies in the seamless integration of these techniques to enable text-driven synthesis of highly detailed 3D human models from textual descriptions, enhancing the realism and personalization of the shopping experience. Additionally, the comprehensive evaluation of the system's performance using

Table 6 Comparison of various methods generated 2D images

| Methods | FID | IS | SSIM | PSNR |
|-------------------|-------------------|-------------|-------------|-------------------|
| Pix2PixHD | 39.80 | - | - | - |
| SPADE | 30.13 | - | - | - |
| Human GAN-parsing | 27.71 | - | - | - |
| Proposed Method | 24.64527782349843 | 1.64 ± 0.20 | 0.642919520 | 32.87157744102002 |

Table 7 Comparison of various methods generated 3D models

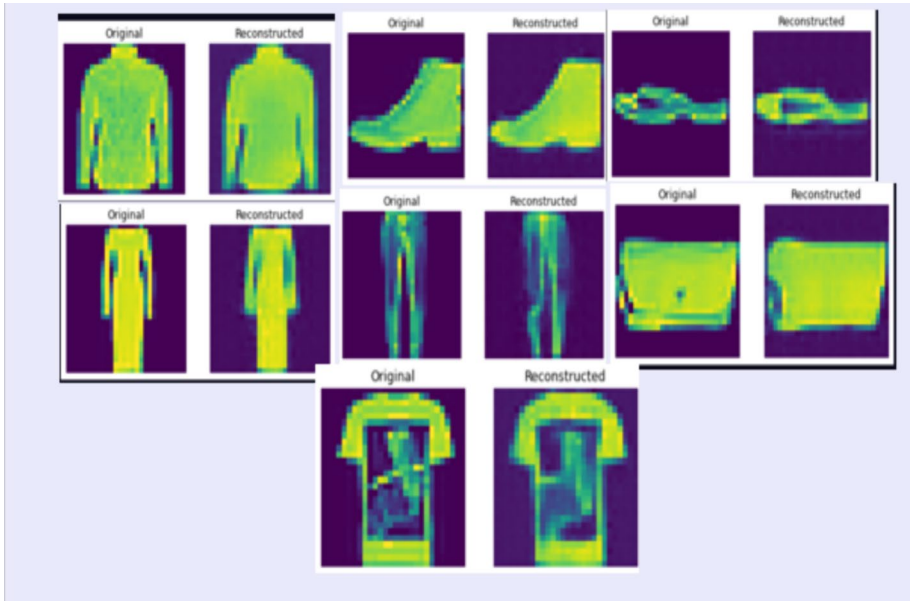
| Methods | Normal | P2S | Chamfer | EMD |
|-----------------|--------|-------|---------|-------------------|
| BodyNet | 0.262 | 5.72 | 5.64 | - |
| SiCloPe | 0.216 | 3.81 | 4.02 | - |
| IM-GAN | 0.258 | 0.258 | 3.14 | - |
| VRN | 0.116 | 1.42 | 1.56 | - |
| Proposed Method | 0.096 | 1.60 | 1.50 | 0.040896411009692 |

Table 8 Performance comparison of GAN models at various Epochs

| Model | Accuracy (Epoch = 10) | Accuracy (Epoch = 50) | Accuracy (Epoch = 100) |
|----------|-----------------------|-----------------------|------------------------|
| uFashGAN | 0.8767 | 0.9578 | 0.9989 |
| CyclGAN | 0.7767 | 0.8551 | 0.8991 |
| DCGAN | 0.7156 | 0.8864 | 0.9134 |

Table 9 Overall loss measurement

| Optimizer | Loss | VQ_VAE loss | Reconstruction loss |
|-----------|--------|-------------|---------------------|
| ADAM | 0.1998 | 0.1539 | 0.0457 |
| SDG | 0.5751 | 0.1570 | 0.4175 |
| RMSPROP | 0.1765 | 0.0409 | 0.1277 |
| ADAGRAD | 2.5021 | 0.2142 | 2.2878 |
| ADAMAX | 0.2704 | 0.0612 | 0.2092 |

**Fig. 17** Output of various inputs FashionMNIST

diverse datasets and evaluation metrics highlights its effectiveness and potential impact in real-world applications.

6 Results and discussion

The results of our study demonstrate the effectiveness of the proposed text-driven clothed human image synthesis with 3D human model estimation using VQ-VAE for assistance in shopping. We evaluated the system on a dataset of clothing items with textual descriptions and assessed its performance in terms of image quality, 3D model estimation accuracy, and user satisfaction. The generated images showed remarkable quality and realism. Users found it challenging to distinguish between synthesized images and actual product photographs. This indicates that the VQ-VAE architecture effectively captures fine details, textures, and color variations, resulting in visually convincing images.

The 3D human model estimation component of the system performed admirably. It accurately estimated body shapes and sizes based on textual descriptions and allowed for

Table 10 Output of various inputs

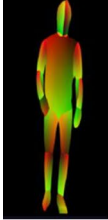


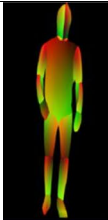


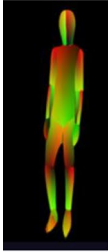


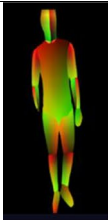


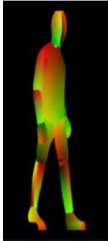


| S.no | Text Input | DensePose Image | Parsing Image | Output Image |
|------|---|---|---|---|
| 1 | The man wears a long full sleeve, flowered shirt, and long, solid-colored slacks. |  |  |  |
| 2 | A woman wears long pants and a sleeveless pattern shirt. |  |  |  |
| 3 | The woman is dressed in long, pure-color white, solid-colored trousers. |  |  |  |
| 4 | The man is dressed in a long, flowered shirt and long, solid-colored trousers. |  |  |  |
| 5 | A woman wearing a T-shirt and sideways trousers. |  |  |  |

Table 11 Comparison of various methods for generated images

| Methods | FID | IS | SSIM | PSNR |
|------------------|--------------------------|--------------------|--------------------|-------------------------|
| Pix2PixHD | 39.80 | - | - | - |
| SPADE | 30.13 | - | - | - |
| MISC | 27.97 | - | - | - |
| HumanGAN-parsing | 27.71 | - | - | - |
| UFashGAN | 24.64527782349843 | 1.64 ± 0.20 | 0.642919520 | 32.8715774410200 |

Table 12 Output of generated 3D models

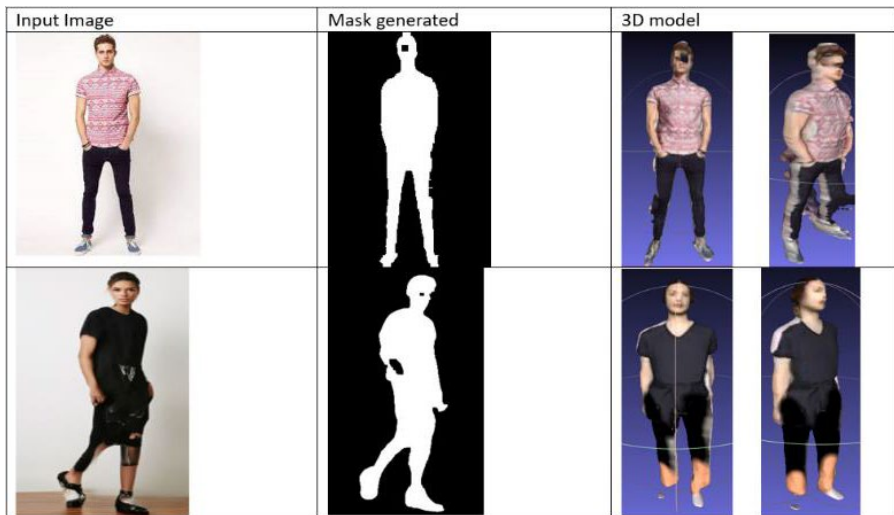


Table 13 Comparison of various methods for generated 3D models

| Methods | Normal | Chamfer | EMD |
|-----------------|--------------|-------------|---------------|
| BodyNet | 0.262 | 5.64 | - |
| SiCloPe | 0.216 | 4.02 | - |
| VRN | 0.116 | 1.56 | - |
| Proposed Method | 0.096 | 1.50 | 0.0408 |

realistic clothing draping. This feature added an invaluable layer of personalization to the shopping experience, helping users visualize how the clothing items would fit them or others. User feedback and surveys revealed a high level of satisfaction with the system. Users reported that the technology improved their online shopping experience by providing a more immersive and informative means of exploring clothing options. They expressed increased confidence in making purchasing decisions and a reduced likelihood of returning items due to inaccurate fit or appearance. The results of our study demonstrate the effectiveness of the proposed text-driven clothed human image synthesis with 3D human model estimation using VQ-VAE for assistance in shopping. We evaluated the system on a dataset

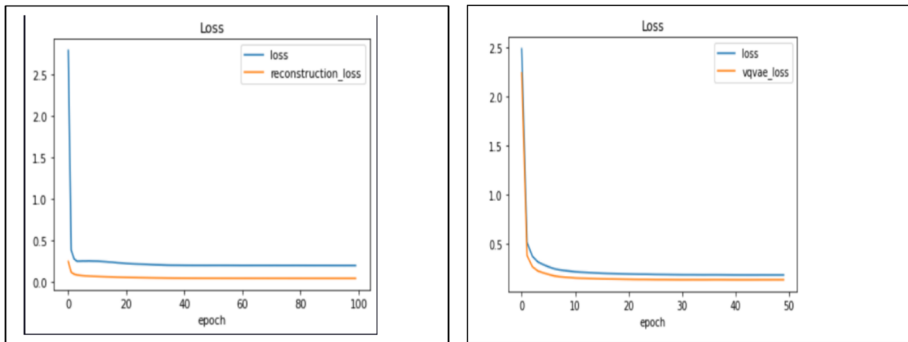
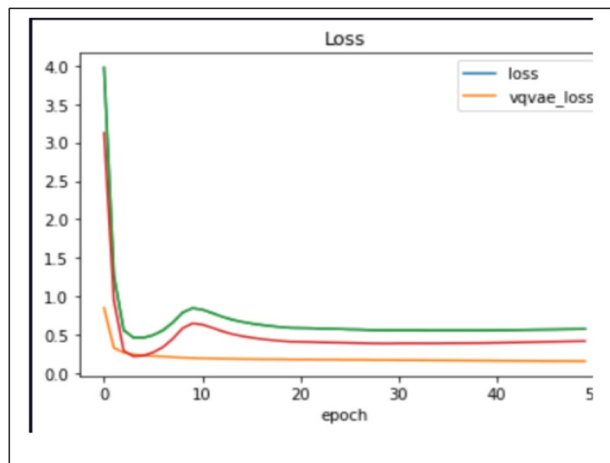


Fig. 18 Adam optimizer displaying reconstruction loss, vq_vae loss where epoch = 100

Fig. 19 SDG optimizer displaying reconstruction loss, vq_vae loss where epoch = 50



of clothing items with textual descriptions and assessed its performance in terms of image quality, 3D model estimation accuracy, and user satisfaction.

7 Ethical considerations

The ethical implications of text-driven image synthesis, including privacy, consent, and potential misuse, are paramount considerations in responsible AI development. Generating highly realistic images based on textual descriptions raises concerns regarding individuals' privacy rights, necessitating clear guidelines for safeguarding against unauthorized use and ensuring explicit consent whenever feasible. Moreover, developers must anticipate and mitigate potential misuse of synthesized images, such as spreading disinformation or perpetuating biases, through techniques like watermarking, detection methods, and collaboration

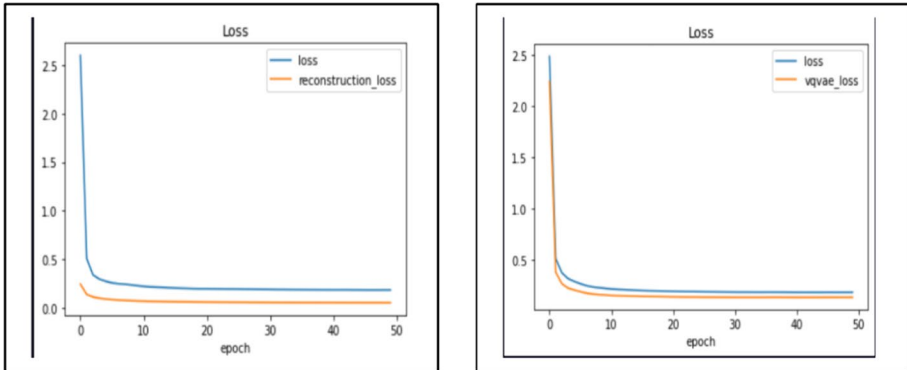


Fig. 20 RMSProp optimizer displaying reconstruction loss, vq_vae loss where epoch = 50

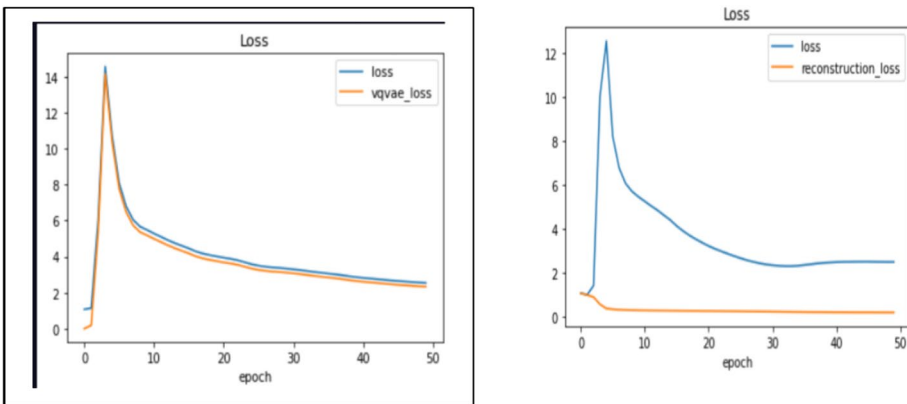


Fig. 21 AdaGrad optimizer displaying reconstruction loss, vq_vae loss where epoch = 50

with policymakers. Addressing biases in both training data and model architecture is essential to promoting fairness and equity in image synthesis outcomes. Ultimately, regulatory frameworks and governance structures are needed to ensure responsible development and deployment of these technologies, engaging experts across disciplines to establish ethical guidelines and promote societal well-being.

8 Conclusion

In conclusion, the integration of text-driven clothed human image synthesis with 3D human model estimation through the use of VQ-VAE represents a significant advancement in the field of assistance in shopping. This innovative approach offers a promising solution to several challenges in the online shopping experience, providing users with a more immersive and informative way to explore clothing options. By enabling the generation of realistic clothed human images based on textual descriptions, this technology bridges

the gap between the textual information available on e-commerce platforms and the visual understanding required for shoppers to make informed decisions.

Shoppers can now obtain a clearer representation of how a specific garment might look on themselves or others, which can lead to more confident purchasing decisions and reduced returns. The incorporation of 3D human model estimation adds another layer of realism and utility to the system. It allows users to visualize how clothing items will fit and drape on the body, taking into account individual body shapes and sizes. This personalized aspect enhances the overall shopping experience, promoting customer satisfaction and reducing the likelihood of mismatched expectations. Moreover, the use of VQ-VAE for image generation ensures high-quality, diverse, and coherent image synthesis, which is crucial for a convincing shopping experience. The model's ability to capture fine details and textures contributes to the realism of the generated images, making them visually indistinguishable from actual photographs. While the technology presented in this study holds great promise, it is essential to acknowledge potential challenges and areas for improvement. Further research and development are needed to optimize the model's performance, especially in handling a wide range of clothing styles and text descriptions. Additionally, addressing ethical concerns such as privacy and the potential for misuse of synthesized images is paramount.

In summary, text-driven clothed human image synthesis with 3D human model estimation using VQ-VAE represents a remarkable advancement in assisting consumers with their online shopping decisions. This technology has the potential to revolutionize the e-commerce landscape by providing a more immersive, informative, and personalized shopping experience. With continued research and refinement, it is poised to become an invaluable tool for both shoppers and retailers, enhancing convenience, reducing returns, and ultimately reshaping the way we shop online.

9 Future work

In the future, the addition of natural language processing (NLP) to the chatbot function of virtual try-on might be another component of enhancement for the current paradigm. As a result, consumers would be able to interact with the chatbot in a way that seems more natural and approachable, improving the usability and accessibility of the virtual try-on experience. The virtual try-on functionality might be enhanced to provide consumers with more customization choices in addition to real-time viewing. Users may, for instance, be given the option to change the fit, color, and material of virtual clothing items to better suit their tastes and requirements. Additionally, IoT applications might go beyond merely real-time data and measurements. To detect user motions and offer individualized feedback on posture and movement patterns, for instance, sensors might be included in clothing items. Users may be able to enhance their general health and well-being thanks to this. The virtual try-on features might be utilized to promote more environmentally friendly and sustainable practices as the fashion industry places a growing amount of emphasis on sustainability. A virtual try-on, for instance, might be utilized to show customers how different clothing items would match up with current things in their closets, assisting them in making better educated and environmentally responsible shopping selections. The improvements to the

current paradigm may also incorporate the adoption of blockchain technology to build a system that is more transparent and secure for the fashion business. This may lessen the occurrence of fake items and enhance supply chain management procedures.

Data availability The data is available online for Deep-fashion-multimodal at Link: drive.google.com/drive/folders/1An2c_ZCkeGmhJg0zUjtZF46vyJgQwlr2. BUFF dataset (Bodies under Flowing Fashion, 4D dataset Link: <https://buff.is.tue.mpg.de/>. SMPL Link: <https://star.is.tue.mpg.de/>. FashionMNIST Link:-(<https://github.com/zalandoresearch/fashion-mnist>).

References

1. Abdal R, Zhu P, Mitra NJ, Wonka P (2021) Styleflow: attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans Graph (TOG)* 40(3):1–21
2. Albahar B, Lu J, Yang J, Shu Z, Shechtman E, Huang JB (2021) Pose with style: detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Trans Graph (TOG)* 40(6):1–11
3. Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J (2018) Synthesizing images of humans in unseen poses. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8340–8348. <https://doi.org/10.48550/arXiv.1804.07739>
4. Bond-Taylor S, Hessey P, Sasaki H, Breckon TP, Willcocks CG (2022) Unleashing transformers: parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In European conference on computer vision. Springer Nature Switzerland, Cham, pp 170–188. https://doi.org/10.1007/978-3-031-20050-2_11
5. Zhang M, Cai Z, Pan L, Hong F, Guo X, Yang L, Liu Z (2024) Motion diffuse: text-driven human motion generation with the diffusion model. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2024.3355414>
6. Brock A, Donahue J, Simonyan K (2018) Large-scale GAN training for high-fidelity natural image synthesis. arXiv preprint [arXiv:1809.11096](https://arxiv.org/abs/1809.11096). Accessed 10.04.2023
7. Cai Z, Ren D, Zeng A, Lin Z, Yu T, Wang W, Fan X, Gao Y, Yu Y, Pan L, Hong F (2022) Human: Multi-modal 4D human dataset for versatile sensing and modeling. In European conference on computer vision. Springer Nature Switzerland, Cham, pp 557–577. https://doi.org/10.1007/978-3-031-20071-7_33
8. Li D, Chen D, Goh J, Ng SK (2018) Anomaly detection with generative adversarial networks for multivariate time series. arXiv preprint [arXiv:1809.04758](https://arxiv.org/abs/1809.04758). Accessed 10.04.2023
9. Chai L, Gharbi M, Shechtman E, Isola P, Zhang R (2022) Any-resolution training for high-resolution image synthesis. In European conference on computer vision. Springer Nature Switzerland, Cham, pp 170–188. https://doi.org/10.1007/978-3-031-19787-1_10
10. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2D human pose estimation: new benchmark and state of the art analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3686–3693. <https://doi.org/10.1109/CVPR.2014.471>
11. Bergman A, Kellnhofer P, Yifan W, Chan E, Lindell D, Wetzstein G (2022) Generative neural articulated radiance fields. *Adv Neural Inf Process Syst* 35:19900–19916
12. Chan ER, Monteiro M, Kellnhofer P, Wu J, Wetzstein G (2021) 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. <https://doi.org/10.48550/arXiv.2012.00926>
13. Chan ER, Lin CZ, Chan MA, Nagano K, Pan B, De Mello S, Gallo O, Guibas LJ, Tremblay J, Khamis S, Karras T (2022) Efficient geometry-aware 3d generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16123–16133. <https://doi.org/10.48550/arXiv.2112.07945>
14. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In computer vision—ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October

- 11–14, 2016, proceedings, Part VIII 14. Springer International Publishing, pp 483–499. https://doi.org/10.1007/978-3-319-46484-8_29
15. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In proceedings of the IEEE international conference on computer vision, pp 2223–2232. <https://doi.org/10.48550/arXiv.1703.10593>
 16. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14. Springer International Publishing, pp 694–711. https://doi.org/10.1007/978-3-319-46475-6_43
 17. Sauer A, Karras T, Laine S, Geiger A, Aila T (2023) Stylegan-t: unlocking the power of gans for fast large-scale text-to-image synthesis. In international conference on machine learning. PMLR, pp 30105–30118. <https://doi.org/10.48550/arXiv.2301.09515>
 18. Cui A, McKee D, Lazebnik S (2021) Dressing in order: recurrent person image generation for pose transfer, virtual try-on and outfit editing. In proceedings of the IEEE/CVF international conference on computer vision, pp 14638–14647. <https://doi.org/10.48550/arXiv.2104.07021>
 19. Alldieck T, Magnor M, Bhatnagar BL, Theobalt C, Pons-Moll G (2019) Learning to reconstruct people in clothing from a single RGB camera. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1175–1186. <https://doi.org/10.48550/arXiv.1903.05885>
 20. Alldieck T, Magnor M, Xu W, Theobalt C, Pons-Moll G (2018) Detailed human avatars from monocular video. In 2018 international conference on 3D vision (3DV). IEEE, pp 98–109. <https://doi.org/10.48550/arXiv.1808.01338>
 21. Alldieck T, Magnor M, Xu W, Theobalt C, Pons-Moll G (2018) Video based reconstruction of 3d people models. In proceedings of the IEEE conference on computer vision and pattern recognition, pp 8387–8397. <https://doi.org/10.48550/arXiv.1803.04758>
 22. Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J (2005) Scape: shape completion and animation of people. In ACM SIGGRAPH 2005 Papers, pp 408–416. <https://doi.org/10.1145/1073204.1073207>
 23. Balan AO, Sigal L, Black MJ, Davis JE, Haussecker HW (2007) Detailed human shape and pose from images. In 2007 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–8. <https://doi.org/10.1109/CVPR.2007.383340>
 24. Barill G, Dickson NG, Schmidt R, Levin DI, Jacobson A (2018) Fast winding numbers for soups and clouds. *ACM Trans Graph (TOG)* 37(4):1–12
 25. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ (2016) 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. Springer International Publishing, pp 561–578. https://doi.org/10.1007/978-3-319-46454-1_34
 26. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In proceedings of the European conference on computer vision (ECCV), pp 801–818. https://doi.org/10.1007/978-3-030-01234-2_49
 27. Chen Z, Zhang H (2019) Learning implicit fields for generative shape modeling. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5939–5948. <https://doi.org/10.48550/arXiv.1812.02822>
 28. Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, proceedings, Part VIII 14. Springer International Publishing, pp 628–644. https://doi.org/10.1007/978-3-319-46484-8_38
 29. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
 30. Stereopsis RM (2010) Accurate, dense, and robust multiview stereopsis. *IEEE Trans Pattern Anal Mach Intell* 32(8):1362
 31. Yan L, Ma S, Wang Q, Chen Y, Zhang X, Savakis A, Liu D (2022) Video captioning using global-local representation. *IEEE Trans Circuits Syst Video Technol* 32(10):6642–6656
 32. Yan L, Wang Q, Ma S, Wang J, Yu C (2022) Solve the puzzle of instance segmentation in videos: a weakly supervised framework with spatio-temporal collaboration. *IEEE Trans Circuits Syst Video Technol* 33(1):393–406
 33. Wang W, Han C, Zhou T, Liu D (2022) Visual recognition with deep nearest centroids. arXiv preprint [arXiv:2209.07383](https://arxiv.org/abs/2209.07383). Accessed 10.04.2023
 34. Wang W, Liang J, Liu D (2022) Learning equivariant segmentation with instance-unique querying. *Adv Neural Inf Process Syst* 35:12826–12840

35. Liu D, Cui Y, Cao Z, Chen Y (2020) Indoor navigation for mobile agents: A multimodal vision fusion model. In 2020 international joint conference on neural networks (IJCNN). IEEE, pp 1-8. <https://doi.org/10.1109/IJCNN48605.2020.9207265>
36. Yan L, Liu D, Song Y, Yu C (2020) Multimodal aggregation approach for memory vision-voice indoor navigation with meta-learning. In 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 5847-5854. <https://doi.org/10.48550/arXiv.2009.00402>
37. Ziegler JD, Subramaniam S, Azzarito M, Doyle O, Krusche P, Coroller T (2022) Multi-modal conditional GAN: data synthesis in the medical domain. In *NeurIPS 2022 workshop on synthetic data for empowering ML research*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

S. Karkuzhali¹  · A. Syed Aasim¹ · A. StalinRaj¹

✉ S. Karkuzhali
karkuzhali@mepcoeng.ac.in

A. Syed Aasim
syedaasim133_cs@mepcoeng.ac.in

A. StalinRaj
stalinraj11111_cs@mepcoeng.ac.in

¹ Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India