A Novel Technique for Optimizing Volumetric Avatars: From 2D Images to Lightweight 3D Models

Kinza Shahzad , Asma Naseer , Aamir Wali , Maria Tamoor *

Posted Date: 20 August 2024

doi: 10.20944/preprints202408.1391.v1

Keywords: 3D; avatars; PIFu; CycleGAN; Generative AI; CNN

 doi:10.20944/preprints202408.1391.v1

Article

A Novel Technique for Optimizing Volumetric Avatars: From 2D Images to Lightweight 3D Models

Avatar Construction with PIFu + CycleGAN

Kinza Shahzad [1], Asma Naseer [1], Aamir Wali [1] and Maria Tamoor [2,*]

[1] FAST School of Computing, National University of Computer and Emerging Science, Faisal Town, Lahore, Pakistan; l217320@lhr.nu.edu.pk (K.S.); asma.naseer@nu.edu.pk (A.N.); aamir.wali@nu.edu.pk (A.W.)
[2] Department of Computer Science, Forman Christian College, Lahore Pakistan, Lahore, Pakistan
* Correspondence: mariatamoor@fccollege.edu.pk

Abstract: In the world of digital avatars, three-dimensional (3D) human models are utilized to replicate real-world appearance and movements with a higher degree of realism compared to traditional 2D representations. Currently, extensive person-specific data capturing and 3D artists are required to create photorealistic avatars of existing people. Determining the 3D pose and articulation of an avatar from 2D images requires complex algorithms, a capability commonly found in the Visual Effects (VFX) industry but not readily available elsewhere. Hence, this study proposes an approach to generate 3D avatar from a 2D image of a person with uncanny resemblance and an accurate depiction of the subject's likeness. The proposed approach, PIFu+CycleGAN, combines Pixel-aligned Implicit Function (PIFu) and cycleGAN for textured avatar construction. PIFu is specifically used to capture intricate details and arbitrary topology. Additionally, the hourglass architecture-based module is utilized for T-pose estimation for predicting the initial geometry and shape. Evaluating the proposed approach on the benchmark RenderPeople dataset, it outperforms the state-of-the-art models with values of 1.53 and 1.50 for the Chamfer and P2S distance, respectively. This indicates the creation of high-quality 3D meshes with promising textures, which are of exceptional quality and suitable for animation.

Keywords: 3D; avatars; PIFu; CycleGAN; Generative AI; CNN

1. Introduction

A 3D Avatar is a digital representation of a person created with the help of 3D modeling software. It has numerous applications in video games, augmented reality, virtual reality environments, chatbots and social media. They are constructed to provide a more humanized or personalized interaction. In education, 3D avatars can be used to create virtual classrooms with student avatars allowing a more immersive and interactive environment. Virtual field trips with 3D avatars of classmates or 3D avatars of historical or prominent individuals are some other interesting uses of 3D avatars. Avatars can also be employed for simulations and interactive lessons which help students learn tasks by actually performing them.

Similarly, in the workplace, 3D avatars can be used in the work-from-home setting or in virtual meetings for improved communication and teamwork. 3D avatars can help to develop a sense of connectivity and community for workers working remotely otherwise, these workers may feel disconnected from other colleagues.

Avatars can be completely fictitious or customized to look like the user. Fictitious avatars can be used in customer service, education, training, etc. A customized avatar that resembles the facial features and expressions, or bodily characteristics of a user is an interesting and challenging task.

A human face is like a serial number that is used as an identification marker [1–3]. People have become very sensitive to faces as a result of evolution, especially familiar faces because faces are the primary social display [4]. Hence, the creation of a digital avatar has become an important challenge because even a little discrepancy from a subject's appearance, facial structure or motion, can result in an uncanny effect [5,6], leading to diminished utility of the avatar for communication facilitation and apparent authenticity. Traditionally, this issue has been handled by using costly and time-consuming

 doi:10.20944/preprints202408.1391.v1

data captures that are extensive, person-specific, and driven by manual processing done by artists. The process of creating avatars with low latency, low data capture, and acceptable quality is the intended focus of this research. One of the interesting problems in the area of computer vision, graphics and machine learning is the construction of 3D facial structure and other intrinsic components from single images (e.g., albedo images). This problem is encountered in numerous applications such as interactive face editing, XR, video-conferencing, Metaverse applications, etc. The main challenge is encountered while creating an automatic avatar using limited data.

3D avatars and facial images have been generated using neural network-based approaches such as PIFuHD [4,7], Generative Adversarial Networks (GANs) [8–13] or other contemporary approaches. High-resolution and photorealistic images have been attributed to GANs more than neural networks[14]. However, due to a lack of photorealistic 3D training data, these GANs frequently only work in two dimensions; therefore, are not capable of synthesizing multiple perspectives of an object. Regardless of which model is used [15–17], the 3D-aware image synthesizers allows face representations to be learned independently of 2D images. New camera poses provide view-consistent images, which can be rendered using the learned representations [18–21]. Contemporary approaches such as leverage neural implicit representations [19,22–24] are also not suitable for 3D face reconstruction as they manifest implicit representations which, so far, have been incapable of expressing refined details.

1.1. Background

Recently, many approaches have been proposed to create 3D objects but most of them require data from high-quality sensors and cameras hence the focus of this study is to overcome this constraint and use data from simple mobile phones to construct 3D avatars. There are several state-of-the-art approaches that are being used to construct 3D avatars from a 2D image.

Stacked Hourglass Network

is a widely used architecture for human pose estimation. It was introduced by Newell et al. in 2016 [25,26] and has since become popular in accurately predicting the locations of body joints or key points in images. It consists of multiple hourglass modules that are stacked on top of each other. By leveraging the encoding-decoding structure of the hourglass modules and incorporating skip connections, stacked hourglass networks can effectively capture multi-scale information and fuse it with fine-grained details to estimate human poses accurately. These networks have achieved remarkable results in human pose estimation and are considered state-of-the-art in the field. Their ability to capture multi-scale information and leverage skip connections has made them highly effective in accurately estimating human poses from images. Using this model, T-pose estimation is performed for human models as illustrated in Figure 1

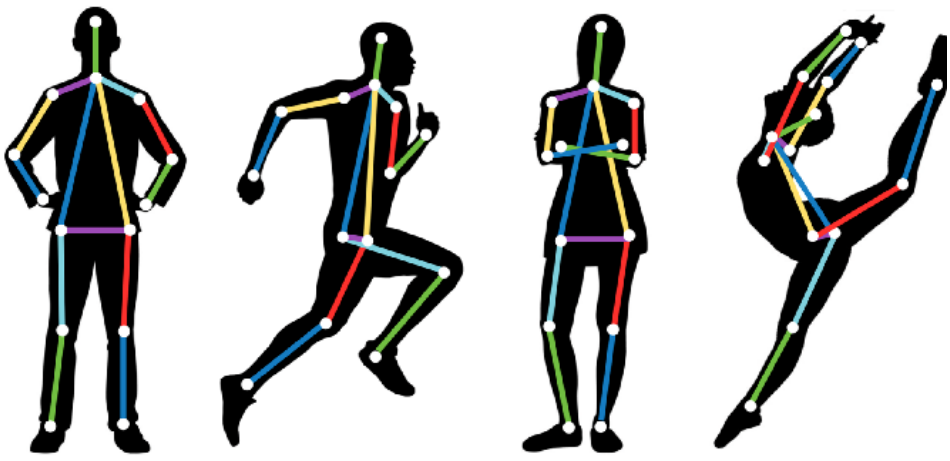

Figure 1. Some examples of estimating T-pose from 2D human images [27].

 doi:10.20944/preprints202408.1391.v1

Marching Cubes algorithm

is a popular method used in computer graphics and scientific visualization to create a three-dimensional surface mesh based on scalar values defined on a three-dimensional grid. It was developed by Lorensen and Cline in 1987. This algorithm utilizes mathematical functions to interpolate and generate vertices. The Marching Cubes algorithm is widely used due to its efficiency in extracting surfaces from volumetric data. It has found applications in fields such as medical imaging, scientific visualization, and computer-aided design. Over time, various variations and enhancements of the algorithm have been proposed to address specific challenges and improve its accuracy and efficiency.

CycleGAN

[28] is a type of deep learning model that falls under the category of generative adversarial networks (GANs). CycleGAN was introduced in 2017 at the University of California, Berkeley. A 3D variant of cycleGAN called 3D CycleGAN [29] was developed to process 3D data and has found applications in various tasks such as shape deformation, video-to-video translation, style transfer within 3D scenes, etc. By harnessing the capabilities of deep learning and adversarial training, 3D CycleGAN contributes to the field of 3D data processing, enabling creative transformations and manipulations of 3D content.

1.2. Major Research Contributions

This paper introduces a novel approach to 3D avatar construction. This approach combines PIFu and CycleGAN. Additionally, the hourglass architecture-based module is also utilized for T-pose (person standing with arms stretched forming a 'T') estimation. This is used for predicting the initial geometry and shape. The major contributions of this study are listed next.

- In this research, a technique for creating lightweight volumetric avatars has been proposed that only need a 2D image of human model as input. The proposed approach allows for more efficient avatar creation thus reducing the computational demands on devices.
- The research incorporates the CycleGAN algorithm, a state-of-the-art technique for texture generation, with the trained PIFu model. This combination enhances the visual quality of the generated avatars by producing realistic and appealing textures.
- A novel combination of techniques for textured avatar construction is employed, avoiding the complex process of modeling multiple aspects of human appearance which leads to more efficient and realistic avatar creation.
- The proposed approach, PIFu+CycleGAN, combines a fully connected convolutional neural network in context of image encoder with a continuous implicit function (PIFu). This combination, along with the hourglass architecture-based module for T-pose estimation, contributes to accurate and intricate avatar geometry and shape prediction.
- The proposed PIFu+CycleGAN model outperforms the state-of-the-art techniques and enhances the qualitative and quantitative values of avatar construction significantly.

The integration of advanced algorithms and diverse approaches provides a platform that makes high-quality 3D avatars more accessible. The rest of the paper is organized as follows. Section 2 presents a literature review on 3D images and avatar reconstructions. Section 4 highlights related concepts. Methodology is presented in section 4 while results are stated in section 5. Discussion and future studies follow in section 6.

2. Literature Review

The process of constructing 3D models from 2D images has been an active area of research in computer vision and graphics for many years. Traditional methods for 3D modeling require specialized equipment such as laser scanners or depth cameras, which can be expensive and time-consuming. In contrast, using 2D images as input data for 3D modeling is more convenient and cost-effective, and can be used in a wide range of applications such as virtual reality, gaming, and architecture.

 doi:10.20944/preprints202408.1391.v1

Due to the popularity of deep networks it is widely being used for solving multiple problems [12,16, 24]. The use of neural networks for 3D model reconstruction is one of the recent advancements in this area. One such method, "Pixel-aligned Implicit Function" of PIFu, is employed for predicting the 3D shape of an object from a single 2D image. The 3D volume representing the predicted shape of the object is generated by the neural network, and is so well-aligned with the input image that it makes the 3D model suitable for tasks like texture mapping and virtual try-on.

The effectiveness of PIFu has been demonstrated for a range of applications, including human body reconstruction, object modeling, and scene reconstruction. However, there are some limitations of this model such as the inability to reconstruct concave shapes and the sensitivity to occlusions and lighting conditions.

To address these limitations, a follow-up method called PIFuHD [4] has been proposed. PIFuHD has been used as an improved version of PIFu that uses a higher-resolution input image and a deeper neural network architecture to achieve higher accuracy and better handling of occlusions. PIFuHD has been demonstrated to produce more accurate and detailed 3D models compared to PIFu, particularly for complex shapes and textures.

Another area of research related to 3D model reconstruction from 2D images using PIFu is face reconstruction. Promising results have been reported by some studies in terms of accuracy and degree of details in the 3D face models. Potential applications of face reconstruction include facial recognition, virtual reality, and character animation. Unfortunately, when it comes to full-bodied, high fidelity and photorealistic 3D face reconstruction of existing people almost all approaches require intricate hardware setups such as multiple-view, light stage etc. and a lot of person-specific data. A novel technique has been proposed by [30] that relies on mobile phone images to capture the person specific data. This approach has three main components: a universal prior, a registration method and an inverse rendering-based method to adjust personalized model on additional expression of data. The high-quality state-of-the-art results are given by this approach.

In addition to PIFu and PIFuHD, other neural network-based methods have been proposed such as Neural Radiance Fields (NeRF) that predict scene and 3D structure from a set of input images. The Other method is Deep Signed Distance Function (DeepSDF) which has been used to predict the signed distance function of a 3D shape from a single 2D image.

In 3D model reconstruction from 2D images, Structure from Motion (SfM) and Multi-View Stereo (MVS) have also been used. SfM has been used to used estimate the camera poses and reconstruct the 3D geometry of the scene by using multiple images of a scene taken from different viewpoints [31]. MVS, on the other hand uses multiple images to compute the depth map of each pixel and then combines the depth maps to generate a 3D model as in [32].

Another interesting application of 3D model reconstruction is virtual try-on, where a user can visualize how a piece of clothing or accessory will look on them without physically trying it on. PIFu has been shown to be effective for this application, with some studies reporting high accuracy in terms of predicting the fit and appearance of clothing on the human body. PIFuHD could potentially further improve the accuracy of virtual try-on by producing more detailed and realistic 3D models.

While neural network-based methods for 3D model reconstruction from 2D images have shown promising results, they still face challenges. These challenges include the need for large amounts of training data which can be difficult to obtain for certain applications. Another challenge is the sensitivity of these methods to variations in lighting and viewpoint, which can affect the accuracy of the resulting 3D models.

In order to overcome the challenges of neural network-based methods, GANs have also been employed. One of the initial works on face reconstruction using GANs is paGAN [33] that needs a well illuminated, frontal, neutral pose of the face. Unfortunately, paGAN does not perform well in situations where the subject's face has a three-quarter view, is passing a smile, has a tongue sticking out, wear glasses, has a dark and hard shadow casted on face, etc. A more innovative GAN framework has been proposed called InterFaceGAN [1] to interpret the GAN models' latent face representation. In this

 doi:10.20944/preprints202408.1391.v1

approach, standard classifiers have been used to predict the semantic scores of GAN-generated images. A connection between latent space (Z) and the semantic space (W) is created in InterFaceGAN which is then utilized for representation analysis. The theoretical and empirical encoding of a single semantic in the latent space is the focus of analysis. It is evident that a linear subspace of the latent space corresponds to a true-or-false facial attribute. The age, gender, the presence of eyeglasses, expression, and even the facial pose of the created image can be successfully altered by InterFaceGAN by only modulating the semantic latent code.

By adjusting the latent codes in both Z space and W space, InterFaceGAN has been applied to the StyleGAN [34] model. StyleGAN is able to develop a deep understanding of various semantics. For example, a younger version of an older person can be produced by StyleGAN. This characteristic of styleGAN can used for photorealistic face reconstruction potentially overcoming all of the challenges faced by the current state-of-the-art.

In decoupling identity from structure, impressive results have been achieved through current solutions, allowing to render a single object from multiple poses. Fine details or consistency are usually missing in these approaches. Voxel-based approaches[35] generate interpretable, true three-dimensional representations, but have limitations due to computational complexity to coarse detail and low resolutions. With the deep-voxel representations, convolutional approaches [18,36] get benefits of the up-to-date progressed convolutional GANs and can generate fine detailed images. Though, these approaches are unable to guarantee multi-view consistency due to their dependence on learned black-box rendering, and cannot generalize beyond the inference-training distribution of camera poses.

In summary, 3D model reconstruction from 2D images using PIFu and PIFuHD is a promising area of research with potential applications in virtual try-on, 3D face reconstruction, and other areas. While these methods have shown promising results, there are still challenges to be addressed, such as the need for large amounts of training data and the sensitivity to variations in lighting and viewpoint. These methods have shown promising results for a range of applications and have the potential to significantly reduce the cost and complexity of 3D modeling.

3. Methodology

The objective of this study is to generate a high-fidelity 3D face avatar using single 2D images. The primary focus is to optimize the topology of these avatars, making them clean, lightweight and suitable for real-time applications. Existing methods often rely on high-quality images from specialized devices, typically accessible to the VFX industry, or generate heavy meshes that are impractical for portable devices or certain applications. To address this issue, a novel approach called PIFu is employed, which enhances the avatar reconstruction through a deep architecture consisting of multiple stages. The PIFu framework does not impose any restrictions on the network architecture and allows for the utilization of any fully convolutional neural network as the image encoder, providing flexibility in choosing the network architecture. The high-level architecture of the proposed approach is given in Figure 2.

As illustrated in Figure 2, the proposed scheme for generating 3D avatars from 2D models involves a series of steps such as data preparation, network Construction for implicit function, piFu structuring, training, T-pose estimation, 3D-Surface Mesh generation, mapping between domains, iterative refinement and finally, texture inference. Once trained, the model is assessed for its qualitative and quantitative performance metrics.

 doi:10.20944/preprints202408.1391.v1

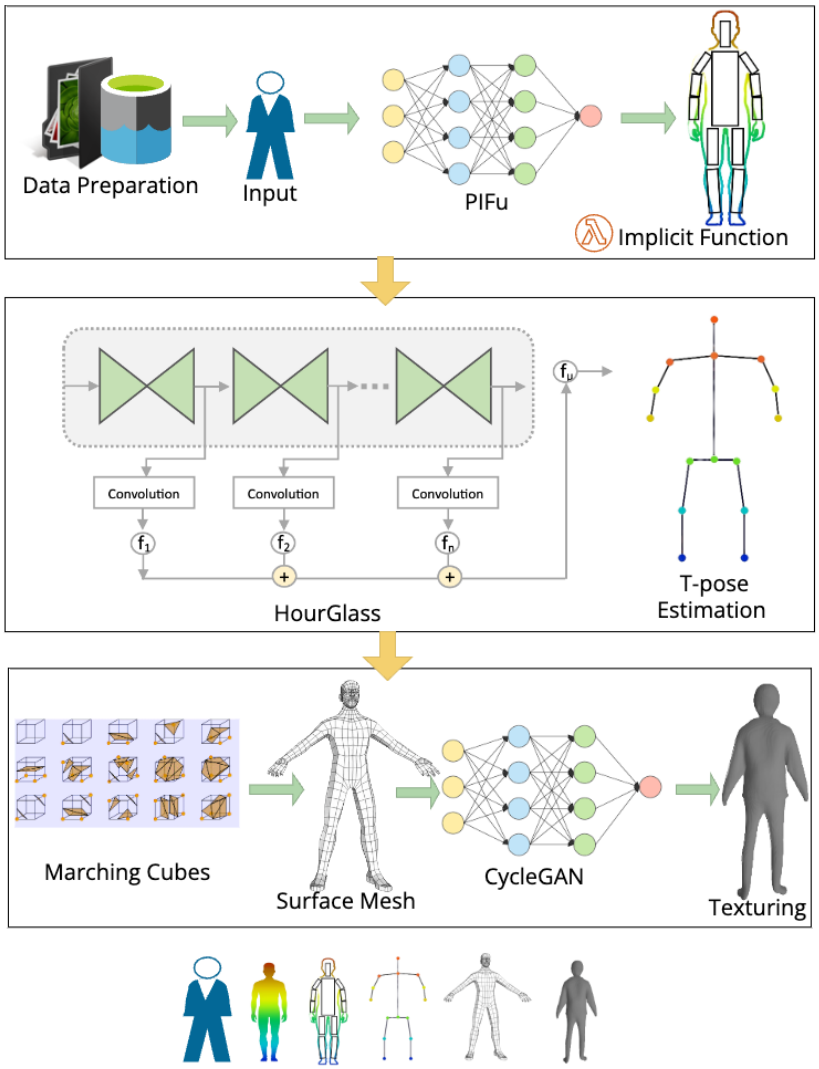

Figure 2. The High level architecture of the proposed approach, encompassing all the sub-modules (Hourglass module, PIFu, Marching Cubes and CycleGAN).

3.1. Data Preparation

The 3D conversion from 2D requires several crucial stages, with the initial step being data preparation. This involves carefully collecting, organizing, pre-processing and refining the relevant datasets in a systematic manner. In the data preparation step, the 3D models and mesh are loaded into tensors.

3.2. Network Construction for Implicit Function and PIFu Structure

We used PIFu implicit function to represent and generate the clothed human form. The advantage of using implicit function is that it can model human form without any knowledge of the actual mathematical model. We also devised a crucial spatial sampling strategy that plays a key role in achieving accurate and high-quality inference results.

This approach is composed of two primary components: a fully convolutional image encoder that uses multi-layer perceptions (MLPs) and a continuous implicit function. The purpose of the fully connected CNN is to estimate the implicit function, which represents the 3D geometry of the human body and clothing.

Functioning as an image encoder, the CNN processes the input RGB image, extracting meaningful features through a series of convolutional layers, non-linear activations (relu and sigmoid on the last layer), pooling, and potentially incorporating residual connections to capture intrinsic and crucial

 doi:10.20944/preprints202408.1391.v1

features. Ultimately, the CNN takes a single RGB image as input and outputs the implicit function, characterizing the 3D structure of the subject's body and attire.

The surface reconstruction approach of PIFu utilizes a multi-layer perceptron (MLP) with varying numbers of neurons in each layer: (257, 1024, 512, 256, 128, 1). Non-linear activations, specifically leaky ReLU, are applied to all layers, except the last one which employs sigmoid activation. In order to effectively propagate depth information, each layer of the MLP incorporates skip connections from the image feature. For Tex-PIFu, surface reconstruction is achieved by combining $Fn(I) \in \mathbb{R}^{256}$ with the image feature $Fn(P) \in \mathbb{R}^{256}$, accomplished by adjusting the number of neurons in the first layer of the MLP to 513 instead of 257. Additionally, the last layer of PIFu is replaced with 3 neurons, for the 3 RGB values, with tanh as the activation function.

3.3. T-Pose Estimation

The hourglass architecture, a module primarily employed in pose estimation tasks, is utilized for T-pose estimation to predict the initial geometry and shape parameters of the subject. It consists of multiple hourglass modules that are stacked on top of each other as illustrated in Figure 3. Each module acts as an encoding-decoding network, capturing multi-scale information by utilizing convolutional and pooling layers for feature extraction and up-sampling layers for spatial resolution recovery. These modules enable effective pose estimation.

To preserve both low-level and high-level features, skip connections are incorporated within the hourglass modules. These connections allow the network to combine fine-grained details with global context, leading to more precise pose estimation.To guide the learning process, intermediate supervision is applied at various stages of the hourglass architecture. This involves introducing loss functions to the output of each module, facilitating gradient updates and error back-propagation at multiple scales.

For constructing the whole human geometry, main body points (head, shoulder, hands, feet etc) should be identified. As the hourglass topology keeps symmetry hence, for each layer in the downward flow, there is a corresponding layer in the upward flow. The convolutional layers are followed by max pooling layers where the network splits and constructs further convolutions at the original resolution before pooling. After achieving the lowest resolution, the network initiates the top-down process of up-sampling and fusing features across scales. Information from two adjacent resolutions are combined which helps estimate the implicit function and capture multi-scale information to create animatable volumetric avatars from 2d planer.

The stacked hourglass networks undergo an iterative refinement process. With each iteration, the initial pose key point estimations are refined, gradually improving the accuracy of the final pose estimation. The output of stacked hourglass networks consists of heat maps, which indicate the likelihood or confidence of key points at each spatial location. These heat maps can be post-processed to extract the precise locations of body joints, enabling the estimation of human poses in diverse scenarios.

The operations of stacked hourglass networks can be described by Equation 1.

$$H(I) = F_{\text{refine}}(F_{\text{hg}_n}(\ldots(F_{\text{hg}_2}(F_{\text{hg}_1}(F_{\text{encode}}(I))))))) \tag{1}$$

Where input image is denoted by I, and $H(I)$ represents the function outcome of the stacked hourglass networks. F_{encode} represents the encoding process, which extracts hierarchical features from the input image I. F_{Hg_i} denotes the i^{th} hourglass module where each module performs encoding and decoding operations on the encoded features to capture multi-scale information. F_{Refine} represents the refinement step, which iteratively refines the pose estimations by processing the outputs of the hourglass modules.

It's important to note that the specific components (encode,hg, refine) depend on the architecture of the stacked hourglass networks. These expressions involve convolutional layers, pooling, skip

 doi:10.20944/preprints202408.1391.v1

connections, and other techniques, aiming at capturing and refining the features for accurate pose estimation.

The stacked hourglass networks function is designed to estimate human poses, shown in Figure 3, from input images for which the following steps are observed.

Input: The stacked hourglass networks function takes an input(I) in Eq. (1) image containing a human subject as its primary input.

Encoding: The input image is processed through a series of convolutional layers, which extract hierarchical features at different scales. This encoding stage Encode() in Eq. (1) captures low-level and high-level features, providing a rich representation of the image.

Hourglass Modules: The encoded features are then passed through a stack of hourglass modules Hgi() in Eq. (1). Each hourglass module consists of multiple encoding and decoding blocks, allowing the network to capture multi-scale information while maintaining spatial resolution.

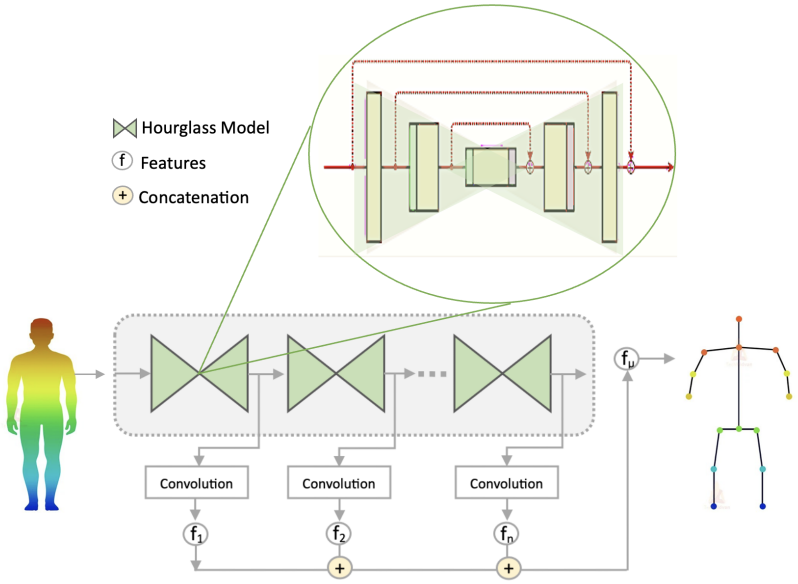

Figure 3. An illustration of Hourglass Module Architecture. The magnified part depicts the symmetry between the input and output layers of the module. At first the scale is reduced with the help of pooling layers and skip connections which are later on up-sampled in a symmetric way.

Skip Connections: Skip connections are utilized within the hourglass modules. These connections enable the network to fuse information from multiple scales, preserving fine-grained details while incorporating global context.

Intermediate Supervision: To aid in training and guide the learning process, intermediate supervision is employed at various stages of the hourglass architecture. Intermediate supervision involves adding loss functions to the outputs of the hourglass modules, facilitating gradient updates and error backpropagation at multiple scales.

Iterative Refinement: The outputs of the hourglass modules are refined iteratively. Each refinement step Refine() in Eq. (1) involves refining the initial pose estimations by further processing them through additional hourglass modules. This iterative refinement enhances the accuracy of the final pose estimation.

Output: The final output of the stacked hourglass networks function is a set of heatmaps or confidence maps. These heatmaps represent the likelihood or confidence of keypoints' presence at each spatial location in the input image. Post-processing techniques can then be applied to extract the precise locations of body joints from these heatmaps, providing an estimation of the human pose. Hence, stacked hourglass networks effectively capture multi-scale information and fuse it with fine-

 doi:10.20944/preprints202408.1391.v1

grained details to accurately estimate human poses by utilizing the encoding-decoding structure of hourglass modules and integrating skip connections.

3.4. 3D Surface Mesh Generation

The Marching Cubes algorithm is employed for generating a 3D surface mesh from scalar values defined on a 3D grid. Using this algorithm some mathematical functions are employed to interpolate and produce vertices.

Linear Interpolation: It estimates the vertex positions along the cube's edges where they intersect the isosurface. It involves two vertices, $V1$ and $V2$, with respective scalar values $S1$ and $S2$, along with a target isovalue T. The linear interpolation is estimated via vertex positions along the intersected edges, as described in Equation 2.

$$V = V_1 + \frac{T - S_1}{S_2 - S_1} \times (V_2 - V_1) \tag{2}$$

where V is the interpolated vertex position, V_1 and V_2 are the two vertices, S_1 and S_2 are their corresponding scalar values, and T is the desired isovalue.

The given equation determines the vertex V's position along the edge by considering the relative difference between the scalar values $S1$ and $S2$ and the desired isovalue T. In terms of vertex generation, the Marching Cubes algorithm utilizes interpolation and position calculations to create vertices. After obtaining the interpolated positions, these vertices are assigned coordinates within a three-dimensional space.

The representation of the vertex coordinates in three-dimensional is depicted in Equation 3.

$$V = (x, y, z) \tag{3}$$

where x, y, and z are the Cartesian coordinates of the vertex.

The surface mesh generation in the Marching Cubes algorithm involves connecting the interpolated vertices with triangles using the mathematical functions mentioned earlier, in addition to the utilization of the Marching Cubes lookup table.

The algorithm converts volumetric data, such as medical scans or fluid simulations, into a geometric representation that can be easily rendered or analyzed. The working of Marching Cubes algorithm is based on thresholding, cube classification, vertex generation and surface reconstruction.

Initially, a three-dimensional grid is provided as and input where each grid cell contains a scalar value representing a property of interest, like intensity or velocity.

Thresholding: A threshold is applied to the scalar values, dividing the volume into regions of interest.

Cube Classification: Each grid cell is analyzed to determine its configuration based on the scalar values at its eight corners by using Equation 2, distinguishing regions inside and outside the area of interest.

Vertex Generation: The vertices are generated, by using Equation 3, along the edges of each cell where the surface intersects. The positions of these vertices are interpolated using the scalar values at the cell's corners.

Surface Reconstruction: Triangles are created by connecting the generated vertices, forming a surface mesh. Lookup tables provide connectivity information based on the cell's configuration.

Output: The resulting surface mesh consists of interconnected triangles that approximate the desired surface. It is rendered using standard 3D graphics technique for deep analysis. Figure 4 provides an illustration of converting the implicit function into discrete triangulated mesh.

 doi:10.20944/preprints202408.1391.v1

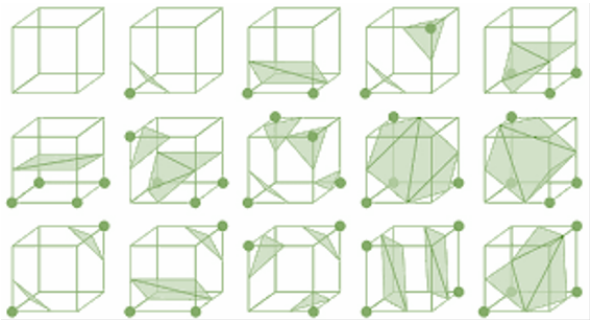

Figure 4. The Marching Cubes algorithm: An illustration of converting the implicit function into discrete triangulated mesh.

3.5. Iterative Refinement

PIFu incorporates an iterative refinement strategy to progressively enhance the reconstructed geometry. This refinement process refines the estimated implicit function and improves the fidelity of the reconstructed geometry. We can fine tune it by adjusting parameters as the geometry details are contingent upon the specific parameter values selected such as the value of batch size is directly proportional to the quality of mesh and inversely proportional to the performance of model.

3.6. Architecture for Surface Reconstruction

After estimating the implicit function, the Marching Cubes algorithm[37] is employed for surface reconstruction. The Marching Cubes algorithm Figure 4 converts the implicit function into a discrete triangulated mesh by extracting the surface geometry. It incorporates both local and global information for better accuracy and detail preservation.

3.7. Texture Inference

For texture inference, the image encoder in PIFu adopts the architecture of CycleGAN, which consists of residual blocks. The CycleGAN-based image encoder is used to generate texture information for the reconstructed geometry. It is a separate image-to-image translation model which can be used in PIFu for texturing. It can learn mappings between different visual styles or domains without paired training data. CycleGAN can be employed to transform textures from one domain (2D) to another(3D), aligning them with the requirements of UV mapping and subsequent texture mapping onto the 3D model.

Figure 5. Input images (2d images + masked version).

 doi:10.20944/preprints202408.1391.v1

Figure 6. Output images (constructed 3d mesh + textured).

3.8. Mapping between Domains

The main purpose of CycleGAN is to learn the mapping between two different domains without the need for paired data during training. It is primary applied in image translation to convert the images from one domain (2D) to another (3D). CycleGAN, used in this research, consists of two generator networks and two discriminator networks. The original CycleGAN is extended to operate in the 3D domain, enabling the translation of 3D data between different domains. This allows the transformation of 3D objects and volumetric data from one representation to another without requiring paired data during training.

The core architecture of CycleGAN retains similarities to the original 2D version. However, adjustments are made to the generators and discriminators to accommodate 3D data. The generators are responsible for learning the mapping between 3D domains, facilitating tasks such as converting the shape of 3D objects or modifying properties within 3D scenes. The discriminators evaluate the realism of the generated 3D outputs compared to real 3D data from the target domain.

Generator Function: The generator function Equation 4 and Equation 5 aims to generate a corresponding image in the target domain when provided with an input image from one domain. It consists of two generator functions, G_A and G_B, for mapping from domain A to domain B and for mapping from domain B to domain A respectively. These functions are typically implemented using convolutional neural network (CNN) architectures. It also contains two discriminator functions i.e., D_A and D_B.

$$L_{\text{adv\_G\_A}} = -\log(D_B(G_A(A))) \tag{4}$$

$$L_{\text{adv\_G\_B}} = -\log(D_A(G_B(B))) \tag{5}$$

Discriminator Function: The realism of an image is evaluated by the discriminator function, as shown in Equation 6 and Equation 7. This function determines whether the image belongs to the real target domain or if it is a generated image from the generator function. In CycleGAN, there are two discriminator functions: one for distinguishing real images in domain B and generated images in domain B, and another one for domain A. CNN architectures are employed to implement these discriminator functions.

$$L_{D_A} = -\log(D_A(A)) - \log(1 - D_A(G_A(A))) \tag{6}$$

$$L_{D_B} = -\log(D_B(B)) - \log(1 - D_B(G_B(B))) \tag{7}$$

Adversarial Loss: The loss function, given in Equation 8 is employed to train the generator functions. It quantifies how effectively the generators can deceive the discriminators by generating

 doi:10.20944/preprints202408.1391.v1

images that are indistinguishable from real images. The adversarial loss incentivizes the generators to produce more realistic outputs.

$$L_{\text{adv\_G\_A}} = -\log(D_B(G_A(A))) \tag{8}$$

Training 3D CycleGAN involves an adversarial process, where the generators strive to produce 3D outputs that deceive the discriminators, while the discriminators aim to distinguish between real and generated 3D data. Through iterative optimization, the generators learn to generate more realistic and meaningful 3D translations.

3.9. Algorithm

For a better understanding, Algorithm 1, highlights the mathematical working behind the proposed methodology. The input to the proposed model is a 2D image of a human (full form) of dimensions $I : m \times n$ which is converted int a 3D avatar of dimensions $O : i \times j \times k$. A convolutional network is used to extract features $F(p)$ from the 2D image I which is followed by the the projection of $3Dp$ onto 2D plane, where p captures the depth in spatial coordination space. While training the deep generative network, the loss L_u and L_v are considered for weight adjustments.

Algorithm 1: 3D Avatar Construction

Input:
I: $m \times n$ matrix – 2D image

Output:
O: $i \times j \times k$ matrix – 3D volumetric representation

Begin:
Features $\leftarrow$ CNN will Extract features from image I
$F(p) \leftarrow F(\text{Features}, p)$
$F(x) \leftarrow f(I(x))$
$x \leftarrow \pi(X)$: Projecting 3D p onto 2D plane
$z(X)$: Depth of p in camera coordinate space
$f(F(x), z(X)) \leftarrow s : s \in \mathbb{R}$
$F_n^*(p) \leftarrow \begin{cases} 1, & \text{if p is inside mesh surface} \\ 0, & \text{else} \end{cases}$
$L_V \leftarrow \frac{1}{n} \sum_{i=1}^{P} |f_v(F_v(x_i), z(X_i)) - f_v^*(X_i)|^2$
L_v : Calculating Loss
P : Points in Sampled space
$L_T = \frac{1}{n} \sum_{i=1}^{n} |f_t(F_T(x_i, F_V), X_i, z) - T(X_i)|$
L_T : Calculating Loss for Texture
$X_{i^n} \leftarrow X_i + \varepsilon \cdot N_i$
$V(x, y, z) \leftarrow F(x, y, z)$
L_T : Output matrix of voxels

4. Experimental Setup

This section presents the details of the experiments conducted to evaluate the proposed model.

4.1. DataSet

For human mesh construction, RenderPeople dataset [38] is used which comprises of 500 high-quality textured human meshes, encompassing a diverse range of clothing, shapes, and poses as illustrated in Figure 7. Each mesh consists of approximately 100,000 triangles. It is the most diverse dataset consisting of scanned 3D models of People. These models are very realistic in three-dimensional (3D) form and provide an authentic and feasible solution to create a 3D environment. The benchmark dataset Renderpeople is widely used in diverse applications such as gaming, AR, VR, etc. The data set is split into training and testing sets randomly according to the ratio 450:50 respectively.

 doi:10.20944/preprints202408.1391.v1

Figure 7. 3D samples from RenderPeople dataset, featuring high-quality textures, and a wide variety of clothing styles, poses and body shapes of human models.

4.2. Training and Hyper-Parameters

At first, PIFu is trained for surface reconstruction and then for texture inference, utilizing the learned image features FI as a condition. For surface reconstruction, an MLP is employed, while for texture inference, CycleGAN is used. The training process involves a learning rate of 1×10^{-3}, a batch size of 3 for surface reconstruction and 5 for texture inference, 12 epochs for surface reconstruction and 6 epochs for texture inference, and a number of sampled points of 5000 and 10000 per object in each training batch, respectively. The learning rate is decayed by a factor of 0.1 at the 10^{th} epoch. The training of PIFu for single-view surface reconstruction took approximately 4 days, while texture inference took another 2 days.

4.3. Evaluation Metrics

Whenever digital data is generated using AI, it needs to be evaluated at two levels: quantitative and qualitative. Quantitative evaluations use mathematical metrics to assign a numeric value to the generated avatars or meshes. The best approach for qualitative evaluation is subjective evaluation [39].

To comprehensively evaluate the ouput of our proposed scheme, several quantitative metrics were employed such as point-to-surface Euclidean distance (P2S), Chamfer distance, and L1 pixel loss. All of these metrics provide insights into different aspects of the reconstruction quality and are breifly discussed next.

1. **Point-to-surface Euclidean distance:** In the model space, the average point-to-surface Euclidean distance (P2S) is calculated in centimeters. This metric measures the proximity of the vertices on the reconstructed surface to their corresponding positions on the ground truth surface. By quantifying the spatial discrepancy between the two surfaces, the P2S metric offers a measure of accuracy in terms of geometric alignment. It also determines the quality of shape reconstruction.
2. **Chamfer distance:** For further analysis, the Chamfer distance (CD) is calculated for the evaluation of proposed model, which serves as a quantitative measure of dissimilarity between the reconstructed and ground truth surfaces. This metric captures the overall disparity between the surfaces, taking into account both global and local deviations. The Chamfer distance provides valuable insights into the fidelity of the reconstruction in capturing intricate details and fine

 doi:10.20944/preprints202408.1391.v1

structures. The formula used for Chamfer distance between two subsets A and B is defined in Equation 9:

$$CD(A,B) = \frac{1}{|A|}\sum_{x\epsilon A} \underset{y\epsilon B}{min}\|x-y\|_2^2 + \frac{1}{|B|}\sum_{x\epsilon B} \underset{y\epsilon A}{min}\|x-y\|_2^2 \quad (9)$$

3. **L1 Pixel Loss:** L1 pixel loss measures the pixel-wise difference and is similar to mean absolute error. L1 pixel loss is usually used to evaluate a target texture in terms of pixel values and visual appearance [40].

Each metric focuses on different aspects, including geometric alignment, overall surface dissimilarity, and the consistency of local details and projections. This multifaceted evaluation enables us to assess the performance of this approach across various dimensions, facilitating robust analysis and comparison with other methods in the field.

5. Results

The section highlights the results obtained by the proposed model both quantitatively and qualitatively. The model has been trained on the Render People dataset [38], which encompasses a diverse collection of 3D human models with various characteristics. However, for the qualitative evaluation of the proposed approach, more test images were obtained from the internet. These were processed to create masks which were subsequently fed to the trained model to convert them from 2D to 3D. To enhance the texturing process, CycleGAN is utilized. For quantitative evaluations, 50 samples that used that were obtained by splitting the dataset into 450:50 training and test sets.

5.1. Qualitative Evaluations

We begin with Figure 8 which demonstrates the stage-wise output of the proposed scheme. As depicted in Figure 8, the proposed method first successfully converts the 2D input image into smooth 3D meshes, and then each mesh is converted into 3D textured output. It can be seen that the model accurately captures the subtle details as well, such as the creases in the shirt in the original input image. The output of this model is a .obj (object) file, which is a standard format for 3D models. To further visualize and manipulate the textured output file, the generated output can be opened in various 3D editing software such as Blender, MeshLab, or 3Ds Max. The textured images generated and presented in this paper were visualized using such software. In addition to 50 test images, several other input images are also used for the experiments, analysis and evaluation. The output of these images provides evidence of the qualitative capabilities of our approach.

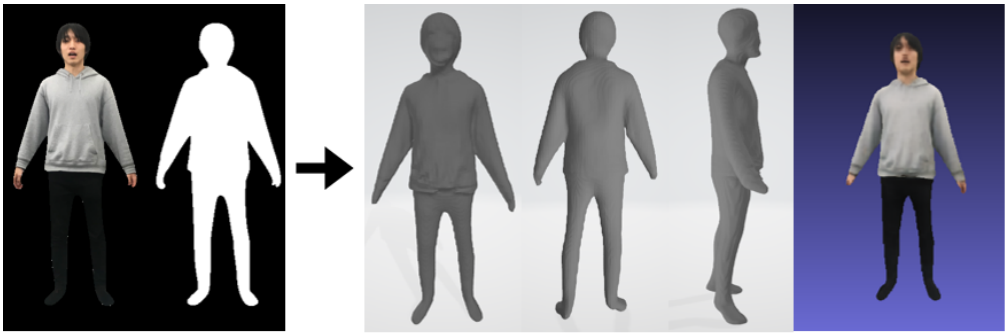

Figure 8. Stage-wise outputs generated by the proposed PIFu and CycleGAN model.

Similarly, Figure 9 illustrates another sample example demonstrating the output of the model. Once again, the details captured by PIFu and CycleGAN are realistic.

 doi:10.20944/preprints202408.1391.v1

Figure 9. Input image of different dressing style (2d images + masked version).

One of the most impressive aspects of our method is its ability to extract complete textures from just a single input image. By doing so, it enables a comprehensive 360-degree view of our 3D models, offering a holistic representation that captures the essence and intricacies of the original image. This breakthrough opens up exciting possibilities for virtual try-on experiences, online shopping, and fashion design, as it allows users to explore and interact with virtual models in unprecedented detail. The combination of our technique's versatility, precision, and ability to capture the true essence of input 2D image.

These results illustrate the model's ability to accurately reconstruct high-fidelity 3D representations of human subjects. For a more comprehensive model evaluation, a variety of images were used based on different sets of factors (or features) such as gender, age, hairstyle, clothing, etc.

Clothing

Figure 10 shows the same person as Figure 9 but with different attire. The color of the dress is blue with a yellow, floral head scarf instead of plain maroon. 3D meshes from different angles are also shown followed by the final output. The notable part of the textured output is the floral scarf which shows that the model is able to capture such high precision and attention to detail.

Figure 10. Input image with head piece and different dressing style.

 doi:10.20944/preprints202408.1391.v1

Hairstyle

Figure 11 shows the 3D meshes generated of a person with a different hairstyle. The interesting aspect of this image is the shoes. Note how the model is able to determine that the girl is wearing a heel from just the raised position of the feet.

Figure 11. Input image with different hairstyle.

Age

Figure 12 shows the 3D meshes generated of an input image of a boy. The purpose of this demonstration is to show that although the training data consisted of mostly elders, the PIFu + CycleGAN was able to compensate and generalize. Also, the large wrinkle on the original input image is captured with high precision.

Figure 12. Input image of a young boy.

Object in Hand

Finally, Figure 13 shows the human avatar generated by the proposed scheme with the person holding a cup. The way the fingers are holding the cup is once again very intricate. However, if we carefully examine the side view, (third mesh), it can be seen that the image of the cup is distorted.

 doi:10.20944/preprints202408.1391.v1

Figure 13. Input with foreign object in hand.

Figures 9–13 present the remarkable digitization results obtained from real-world input images using this innovative approach. The effectiveness of the proposed method becomes evident as it flawlessly handles a wide range of clothing types, including skirts, jackets, and dresses. Notably, this technique goes beyond mere replication, as it excels in generating intricate high-resolution local details and accurately infer plausible 3D surfaces even in regions that were previously unseen or challenging to recreate.

5.2. Quantitative Evaluations

Quantitative evaluation of reconstructed models is done using the metics: P2S and Chamfer distance. Comparative results of our proposed approach and existing state-of-the-art algorithm of 2D mesh reconstruction are shown in Table 1. These results show that the proposed method performs better than other state-of-the-art techniques in terms of distance (P2S) and error (CD) calculations.

Table 1. Quantitative evaluation for reconstructed avatar mesh.

Algorithm	P2S	Chamfer
IM-GAN	2.87	3.14
BodyNet	5.72	5.64
FaceScape	1.56	1.58
SiCloPe	3.81	4.02
[41]	-	1.53
Proposed	**1.53**	**1.5**

L1 pixel loss is used to measure the difference between the generated textures and the ground truth. Notably, this method demonstrates the ability to generate textures that exhibit sensitivity to various lighting conditions. Comparative results of the proposed approach and state-of-the-art methods of CycleGAN [28] are shown in Table 2.

Table 2. Quantitative comparison of texture.

Algorithm	L1 pixel loss
FaceScape [42]+StyleGAN	0.205
Proposed	0.196

 doi:10.20944/preprints202408.1391.v1

6. Discussion and Future Scope

In contrast to other methods, this proposed approach overcomes the limitations of high memory requirements in volumetric representations, which allows the generation of high-resolution outputs. Furthermore, the natural extension of this method to infer complete textures on a person even from partial observations is demonstrated. Unlike existing approaches that synthesize back regions based on frontal views in the image space, this method directly predicts colors in unseen, concave, and side regions on the surface. By generating textured 3D surfaces of clothed individuals using only a single RGB camera, a new direction is explored towards monocular reconstructions of dynamic scenes from input images without the need for a template model. Consequently, this method's ability to handle additional arbitrary views makes it highly practical and efficient for 3D modeling purposes.

The major highlighted contributions of this works can be summarised as:

- In this research, a technique for creating lightweight volumetric avatars has been proposed that only need a 2D image of human model as input.
- The proposed approach allows for more efficient avatar creation and reduces the computational demands on devices.
- It overcomes the limitations of high memory requirements in volumetric representations and generates high-resolution images.
- the method predicts colors directly in unseen, concave, and side regions and can inpaint the textures for shapes with arbitrary topology.
- A novel combination of techniques for textured avatar construction is employed which leads to more efficient and realistic avatar creation.
- The proposed approach, PIFu+CycleGAN, combines a fully connected convolutional neural network along with the hourglass architecture-based module which contributes to accurate and intricate avatar geometry and shape prediction.
- The proposed model outperforms the state-of-the-art techniques and enhances the qualitative and quantitative values of avatar construction significantly.

For future work, other factors like lighting conditions, multiple poses of the input images, and facial expressions can be incorporated to improve the quality of the model and textures. Future work may include achieving more reconstruction accuracy. Furthermore, the current study doesn't consider occlusions which can be added as the extension of this research.

Another major observation is that in some cases the model could not estimate the T pose correctly. This can be seen in Figure 14 where there is a pose misalignment. The position of the feet is different. Even in such cases where the hourglass architecture could not perform well, CycleGAN performed well and created the texture even when the topology was terrible.

Figure 14. Input image with unaligned pose.

In precise, further advancing the field of 3D face reconstruction in terms of higher-resolution texture inference, scale estimation, and addressing occlusions holds great promise for enhancing the

 doi:10.20944/preprints202408.1391.v1

practicality and effectiveness of the technology. By focusing on these areas, researchers can improve the visual realism, accuracy, and applicability of 3D face models, enabling their integration into a wide range of real-world applications while also addressing ethical concerns and ensuring the responsible use of this technology, ethical issues are further discussed here in detail.

7. Conclusion

In this study a novel approach in presented, in which a deep architecture of multiple stages refines the avatar reconstruction progressively. This framework consists of a combination of two state-of-the-art techniques (PIFu+ CycleGAN) to reconstruct high fidelity textured avatars.

The process involves estimating the implicit function utilizing a combination of the multi-layer perceptron and hour glass architecture, which is specifically employed for T-pose estimation. Following this step, the Marching Cubes algorithm is incorporated, which converts the obtained implicit function into a discrete triangulated mesh by extracting the surface geometry. The proposed approach not only examines the qualitative aspects of the results but also delves into the quantitative measures, highlighting the outcomes achieved through the training of the model on the Rendered People 3D dataset.

While PIFu+ CycleGAN exhibited superior quality results, it is worth noting that further enhancements are required in the domain of texture work. The discussion surrounding the qualitative and quantitative results sheds light on the effectiveness of the methodology, and the areas that necessitate further refinement to achieve optimal performance in terms of texture representation.

Ethical Approval: Not applicable.

Competing Interests: There is no competing interests between the authors.

Authors' Contributions: KS proposed the methodology, implemented it and prepared the first draft of the article. AN refined the proposed methodology, supervised and improved the write-up of the article. AW co-supervised, and evaluated and refined the article. MT improved the technique and paper write-up.

Funding: Not applicable.

Availability of Data and Materials: The datasets used in this research, are publicly available.

References

1. Shen, Y.; Yang, C.; Tang, X.; Zhou, B. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *44*, 2004–2018.
2. Tamoor, M.; Naseer, A.; Khan, A.; Zafar, K. Skin lesion segmentation using an ensemble of different image processing methods. *Diagnostics* **2023**, *13*, 2684.
3. Tamoor, M.; Gul, H.; Qaiser, H.; Ali, A. An optimal formulation of feature weight allocation for CBR using machine learning techniques. 2015 SAI Intelligent Systems Conference (IntelliSys). IEEE, 2015, pp. 61–67.
4. Saito, S.; Simon, T.; Saragih, J.; Joo, H. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 84–93.
5. Gecer, B.; Ploumpis, S.; Kotsia, I.; Zafeiriou, S. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*, 4879–4893.
6. Khalid, A.; Farhan, A.A.; Zafar, K.; Tamoor, M. Automated Cobb's Angle Measurement for Scoliosis Diagnosis Using Deep Learning Technique. *Preprints* **2024**.
7. Saleem, M.A.; Tamoor, M.; Asif, S. An efficient method for gender classification using hybrid CBR. 2016 Future Technologies Conference (FTC). IEEE, 2016, pp. 116–120.
8. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* **2017**.
9. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

 doi:10.20944/preprints202408.1391.v1

10. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
11. Naseer, A.; Tamoor, M.; Azhar, A. Computer-aided COVID-19 diagnosis and a comparison of deep learners using augmented CXRs. *Journal of X-ray Science and Technology* **2022**, *30*, 89–109.
12. Naseer, A.; Hussain, S.; Zafar, K.; Khan, A. A novel normal to tangent line (NTL) algorithm for scale invariant feature extraction for Urdu OCR. *International Journal on Document Analysis and Recognition (IJDAR)* **2022**, *25*, 51–66.
13. Naseer, A.; Tamoor, M.; Khan, A.; Akram, D.; Javaid, Z. Occupancy detection via thermal sensors for energy consumption reduction. *Multimedia Tools and Applications* **2024**, *83*, 4915–4928.
14. Nawaz, M.; Adnan, A.; Tariq, U.; Salman, F.; Asjad, R.; Tamoor, M. Automated career counseling system for students using cbr and j48. *Journal of Applied Environmental and Biological Sciences* **2015**, *4*, 113–120.
15. Khan, M.; Naseer, A.; Wali, A.; Tamoor, M. A Roman Urdu Corpus for sentiment analysis. *The Computer Journal* **2024**, p. bxae052.
16. Raza, N.; Naseer, A.; Tamoor, M.; Zafar, K. Alzheimer Disease Classification through Transfer Learning Approach. *Diagnostics* **2023**, *13*, 801.
17. Wali, A.; Naseer, A.; Tamoor, M.; Gilani, S. Recent progress in digital image restoration techniques: a review. *Digital Signal Processing* **2023**, p. 104187.
18. Nguyen-Phuoc, T.; Li, C.; Theis, L.; Richardt, C.; Yang, Y.L. Hologan: Unsupervised learning of 3d representations from natural images. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7588–7597.
19. Schwarz, K.; Liao, Y.; Niemeyer, M.; Geiger, A. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems* **2020**, *33*, 20154–20166.
20. Henzler, P.; Mitra, N.J.; Ritschel, T. Escaping plato's cave: 3d shape from adversarial rendering. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9984–9993.
21. Nasreen, G.; Haneef, K.; Tamoor, M.; Irshad, A. A comparative study of state-of-the-art skin image segmentation techniques with CNN. *Multimedia Tools and Applications* **2023**, *82*, 10921–10942.
22. Wali, A.; Ahmad, M.; Naseer, A.; Tamoor, M.; Gilani, S. Stynmedgan: medical images augmentation using a new GAN model for improved diagnosis of diseases. *Journal of Intelligent & Fuzzy Systems* **2023**, *44*, 10027–10044.
23. Malik, Y.S.; Tamoor, M.; Naseer, A.; Wali, A.; Khan, A. Applying an adaptive Otsu-based initialization algorithm to optimize active contour models for skin lesion segmentation. *Journal of X-Ray Science and Technology* **2022**, *30*, 1169–1184.
24. Chughtai, I.T.; Naseer, A.; Tamoor, M.; Asif, S.; Jabbar, M.; Shahid, R. Content-based image retrieval via transfer learning. *Journal of Intelligent & Fuzzy Systems* **2023**, pp. 1–26.
25. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. *CoRR* **2016**, *abs/1603.06937*, [1603.06937].
26. Nasir, S.; Taimoor, M.; Gul, H.; Ali, A.; Khan, M.J. Optimization of decision making in cbr based self-healing systems. 2012 10th International Conference on Frontiers of Information Technology. IEEE, 2012, pp. 68–72.
27. Gupta, A. Human Pose Estimation Using Machine Learning in Python. https://www.analyticsvidhya.com/blog/2021/10/human-pose-estimation-using-machine-learning-in-python/, 2008. [Online; accessed 19-July-2023].
28. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
29. Bashkirova, D.; Usman, B.; Saenko, K. Unsupervised video-to-video translation. *arXiv preprint arXiv:1806.03698* **2018**.
30. Cao, C.; Simon, T.; Kim, J.K.; Schwartz, G.; Zollhoefer, M.; Saito, S.S.; Lombardi, S.; Wei, S.E.; Belko, D.; Yu, S.I.; others. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* **2022**, *41*, 1–19.
31. Lauria, G.; Sineo, L.; Ficarra, S. A detailed method for creating digital 3D models of human crania: an example of close-range photogrammetry based on the use of structure-from-motion (SfM) in virtual anthropology. *Archaeological and Anthropological Sciences* **2022**, *14*, 42.

 doi:10.20944/preprints202408.1391.v1

32. Song, S.; Truong, K.G.; Kim, D.; Jo, S. Prior depth-based multi-view stereo network for online 3D model reconstruction. *Pattern Recognition* **2023**, *136*, 109198.
33. Nagano, K.; Seo, J.; Xing, J.; Wei, L.; Li, Z.; Saito, S.; Agarwal, A.; Fursund, J.; Li, H.; Roberts, R.; others. paGAN: real-time avatars using dynamic textures. *ACM Trans. Graph.* **2018**, *37*, 258–1.
34. Luo, H.; Nagano, K.; Kung, H.W.; Xu, Q.; Wang, Z.; Wei, L.; Hu, L.; Li, H. Normalized avatar synthesis using stylegan and perceptual refinement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11662–11672.
35. Henzler, P.; Mitra, N.; Ritschel, T. Escaping Plato's Cave: 3D Shape From Adversarial Rendering, 2021, [1811.11606].
36. Nguyen-Phuoc, T.H.; Richardt, C.; Mai, L.; Yang, Y.; Mitra, N. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems* **2020**, *33*, 6767–6778.
37. Lorensen, W.E.; Cline, H.E. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics* **1987**, *21*, 163–169.
38. Renderpeople. https://renderpeople.com/, 2018. [Online; accessed 19-July-2023].
39. Zafar, I.; Wali, A.; Kunwar, M.A.; Afzal, N.; Raza, M. A pipeline for medical literature search and its evaluation. *Journal of Information Science* **2023**, p. 01655515231161557.
40. Luo, H.; Nagano, K.; Kung, H.W.; Xu, Q.; Wang, Z.; Wei, L.; Hu, L.; Li, H. Normalized avatar synthesis using stylegan and perceptual refinement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11662–11672.
41. Saito, S.; Simon, T.; Saragih, J.; Joo, H. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 84–93.
42. Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; Cao, X. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 601–610.