3D Avatar Reconstruction using Multi-Level Pixel-Aligned Implicit Function

Shreedhar I Muttagi, Vaishnavi Patil, Pooja P Babar, Ritvik Chunamari
Uday Kulkarni, Satish Chikkamath, and Meena S. M

[1] KLE Technological University, Hubli, Karnataka, India
[2] KLE Technological University, Hubli, Karnataka, India
[3] KLE Technological University, Hubli, Karnataka, India
[4] KLE Technological University, Hubli, Karnataka, India
[5] KLE Technological University, Hubli, Karnataka, India
[6] KLE Technological University, Hubli, Karnataka, India
[7] KLE Technological University, Hubli, Karnataka, India
uday_kulkarni@kletech.ac.in, vaishnavipatil0248@gmail.com

Abstract. In recent years, there has been significant interest in 3D avatar reconstruction from images, driven by its applications in gaming, entertainment, and augmented reality. This paper introduces a novel approach that utilizes the FSRCNN (Fast Super-Resolution Convolutional Neural Network) and ML-PIFu (Multi-level pixel-aligned implicit function) framework to reconstruct high-fidelity 3D avatars from a single input image. Our method integrates the FSRCNN with a two-module ML-PIFu pipeline, combining global geometric information and local fine details. The FSRCNN enhances the input image quality before it is processed by the ML-PIFu framework. The coarse-level module of ML-PIFu improves the image quality, capturing the overall structure and shape of the human form, while the fine-level module incorporates intricate details for enhanced output. Through the use of neural networks and an end-to-end training strategy, our approach achieves accurate and precise reconstructions without requiring multiple images or complex camera setups. The potential applications of our approach are extensive, including the creation of personalized avatars for telepresence, Virtual Reality (VR), Augmented Reality (AR), anthropometry studies, and virtual try-on experiences. By leveraging our approach, these applications can benefit from enhanced accuracy, realism, and versatility, thereby opening up new possibilities in the realm of digital human representation and interaction.

Keywords: 3D avatar reconstruction · Multi level pixel aligned implicit function (ML-PIFu) · Neural networks, · End-to-end training,· Single image reconstruction, · Coarse-level module, · Fine-level module,Augmented Reality (AR), · Virtual Reality (VR).

1 Introduction

The detailed digitization of the human body has significant implications in various fields, including medical imaging and virtual reality. However, its widespread

use has been limited due to the high cost and specialized requirements of professional capture systems, such as multiple cameras and controlled lighting conditions. Fortunately, there has been an increasing interest in leveraging advanced deep-learning models to overcome these limitations and enable reconstructions from just a single image. While these approaches have demonstrated promising results, there is still a need to enhance their performance further. Our objective is to achieve an exceptional level of fidelity in the 3D reconstruction of clothed individuals using just a single image, surpassing the performance of existing techniques and specialized capture devices. We strive to capture intricate details such as finger positions, facial expressions, and clothing folds with utmost accuracy, ensuring a resolution that allows for precise restoration. However, current methods have not fully harnessed the potential of high-resolution imaging and struggle to handle low-quality photographs typically taken with standard sensors on mobile devices. As a result, we aim to develop innovative approaches that take advantage of modern imaging technology while remaining effective even when dealing with sub-optimal image quality. By doing so, we aim to revolutionize the field of 3D reconstruction from a single image.

Prior methods often use down-sampled images due to memory limitations, relying on holistic reasoning to connect a human's 2D appearance to its 3D shape. Local image patches provide valuable details for precise 3D reconstruction, but their usage in high-resolution inputs is limited due to graphics technology constraints. Our approach aims to overcome these limitations, incorporating local patches for comprehensive 3D representation, despite memory constraints and graphics limitations.

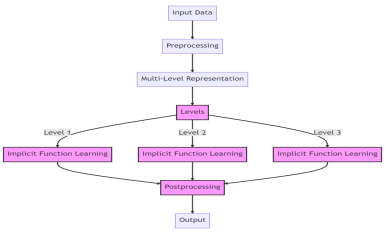

Figure 1. Using a high-resolution image of an individual, we are able to reconstruct remarkably detailed 3D representations of clothed individuals.

Local image patches are valuable for capturing crucial details in 3D reconstruction, yet their utilization in high-resolution inputs remains limited due to current graphics technology limitations. However, advancements in this field can potentially overcome these constraints, paving the way for comprehensive integration of local image patches in future high-resolution implementations. By harnessing the power of advanced graphics capabilities, the potential for leveraging local image patches in their entirety can be unlocked, enriching the accuracy and fidelity of 3D reconstructions.

The constraints that hinder the use of local image patches can be resolved in two ways. In the first category, the issue is solved by enhancing low-fidelity surfaces with high-frequency details in a coarse-to-fine fashion. With this method, a low-resolution image is used to first get a crude form. Subsequently, through post-processing methods like Shape From Shading[9] or neural network-based composition [16], precise features in the form of surface normals [10] or displacements [12] are added. The second category generates convincing details using high-fidelity human models like SCAPE. While both methods produce reconstructions that appear to be detailed, they frequently fall short of accurately recreating the true details found in the raw photos. By incorporating state-of-the-art algorithms and deep learning models, we strive to overcome the limitations of existing methods and achieve accurate and detailed 3D reconstructions. The goal is to enable the creation of realistic and expressive 3D avatars from minimal input, revolutionizing applications such as virtual reality, gaming, and entertainment. In this paper, we present the following sections. Section 2 offers an overview of the background and context of our work on existing 3D human avatar digitization models. In Section 4, we examine the limitations of these models and propose an architecture that effectively overcomes these limitations. Section 4 presents the results of our proposed architecture, comparing it to other existing models. Finally, Section 5 summarizes the key findings and concludes our proposed architecture.

2 Related Works

1. 3D Human Avatar Digitization from a Single Image. Research on 3D human reconstruction [13] encompasses various areas, including estimating body pose, single-image-based 3D human body reconstruction, video-based human body reconstruction, and inferring occluded or invisible texture color. The focus is on creating a 3D Human Avatar from a single image. A single image is used to reconstruct the human body. To reconstruct the final geometry, the proposed method utilizes 2D non-rigid registration. Infer GAN (Generative Adversarial Networks) [3] is used to extract texture based on foreground texture. In this method, a single image is used to reconstruct the entire geometry and intricate texture of a human body, and an animated model is generated based on the weights and joints of the parametric motion model. To demonstrate the capabilities of the reconstruction pipeline, a mobile application has been developed, showcasing the process and hosted on a server. However, it is imperative to note some limitations of this approach.

2.Self-Recon: Self Reconstruction Your Digital Avatar from the Monocular Video In this paper, the authors highlight the importance and challenges of reconstructing clothed bodies, particularly in industries such as film and gaming. Traditional methods for achieving high-quality human reconstructions often involve pre-captured templates, multi-camera setups, controlled environments, and manual labor by skilled artists. However, these requirements

are impractical for general consumers who seek personalized avatars for applications like telepresence, AR/VR, anthropometry, and virtual try-on. This necessitates the development of direct high-fidelity digital avatar reconstruction methods from monocular videos to meet the growing demand. Additionally, a non-rigid ray casting algorithm is introduced, combining explicit representation for differentiable intersections with the deformed implicit surface. As a result, Self Recon achieves high-fidelity reconstructions of detailed clothed body shapes without relying on pre-computed templates. The authors demonstrate the potential applications of Self Recon through the generation of high-quality avatars based on tracking results.

3. Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization The purpose of this paper is to construct high-resolution 3D reconstructions of clothed individuals that utilize the Pixel-Aligned Implicit Function (PIFu) to facilitate the generation of 3D reconstructions of clothed individuals. This technique involves populating a dense 3D volume to determine the presence of human body points in 3D space. The performance of recent 3D human reconstruction methods was compared using the People Snapshot dataset [5] , including multi-scale voxel representation (Deep Human), pixel-aligned implicit function (PIFu), and a model-based approach with texture mapping using displacements and surface normal (Tex2shape) [2].The results suggest that Tex2shape and Deep Human methods show limited refinement effects, primarily because their basic forms have limited representational power. Voxel representation introduces spatial resolution constraints, while model-based approaches face challenges in handling varying topologies and significant deformations.

3 Proposed methodology

In order to improve the accuracy and resolution of 3D reconstructions from 2D images, we propose an enhancement to the Pixel Aligned Implicit Function (PIFu) [13] framework. It utilizes 512x512 resolution images to generate low-resolution feature embeddings of size 128x128, which are then used to produce higher-resolution outputs. Two additional modules are introduced in this architecture to augment the PIFu framework. A higher-resolution image (1024x1024) is taken as input and encoded into 512x512 features. This module focuses on capturing fine details and enhancing the resolution of the reconstructed 3D object. The second module takes the high-resolution feature embedding from the first module, along with the 3D embedding from the PIFu framework, and utilizes them to estimate an occupancy probability field. The occupancy probability field represents the likelihood that a given point in 3D space is occupied by the object being reconstructed. By incorporating these additional modules, the proposed architecture extends the capabilities of the PIFu framework and improves the quality and resolution of 3D reconstructions from 2D images.

Overall the proposed architecture is designed to improve the accuracy and resolution of 3d reconstructions from 2d images by leveraging higher-resolution

inputs and fine details. This approach has several potential applications in fields such as computer graphics virtual reality and robotics where accurate and detailed 3d reconstructions are essential for creating realistic and immersive environments. The proposed architecture may also have practical applications in fields such as medicine and architecture where detailed 3d models are essential for planning and analysis. In our proposed work we have improved the existing state of art PIFu framework.To enhance the quality of the reconstructed avatars, we integrate the PIFu framework with some advanced image processing techniques present in the open-source computer vision (OpenCV (cv2)) library. These advanced image processing techniques preprocess the input images and improve the quality of the reconstructed avatars.Super Resolution model, Fast Super-Resolution Convolutional Neural Network (FSRCNN) has been used to improve the quality of the input image.By recovering a high-resolution (HR) image from a given low-resolution (LR) one, single image super-resolution (SR) attempts to improve image quality.

FSRCNN: **FSRCNN**:There are five fundamental steps in the development of the FSRCNN (Fast Super-Resolution Convolutional Neural Networks): removing the feature, shrinking the feature, mapping the feature, expanding the feature, and deconvolution the feature. As part of the network's architecture, convolutional and deconvolutional layers are used. A Super-Resolution Convolutional Neural Network (FSRCNN) is a development of the Super-Resolution Convolutional Neural Network (SRCNN). It eliminates the need for interpolation by directly learning the transformation between the original low-resolution image and the high-resolution image during the mapping process of FSRCNN. In addition, the output is further refined by adding a deconvolution layer. Secondly, FSRCNN incorporates a shrink-expand approach, where the input feature dimension is initially reduced (shrunk) before the mapping step and then expanded again afterward. This technique captures and preserves essential information throughout the network. Thirdly, FSRCNN employs multiple mapping layers with smaller filter sizes. This design choice enables the network to learn more detailed and localized features. The output of FSRCNN is passed through the MobileNet model, which is responsible for keypoint detection and annotation for human pose estimation. These modifications enhance the performance and efficiency of the FSRCNN model, allowing it to achieve fast and accurate super-resolution while maintaining significant image details.

As depicted in Figure 2, the network utilizes preprocessed images as inputs for its image encoders. Two encoders are employed: a low-resolution and a high-resolution encoder. The coarse image encoder generates feature maps sized at 128x128x256, which are then passed through a Multi-Layer Perceptron (MLP). The coarse-level MLP consists of six layers with neuron counts of (257, 1024, 512, 256, 128, 1). To facilitate information flow, skip connections are implemented at the third, fourth, and fifth layers. These connections improve feature preservation and enhance the overall performance of the network

The MLP layer produces 3D embedding as the final output. The pixel-based image encoder generates feature maps of 512x512x216. The coarse-level encoder

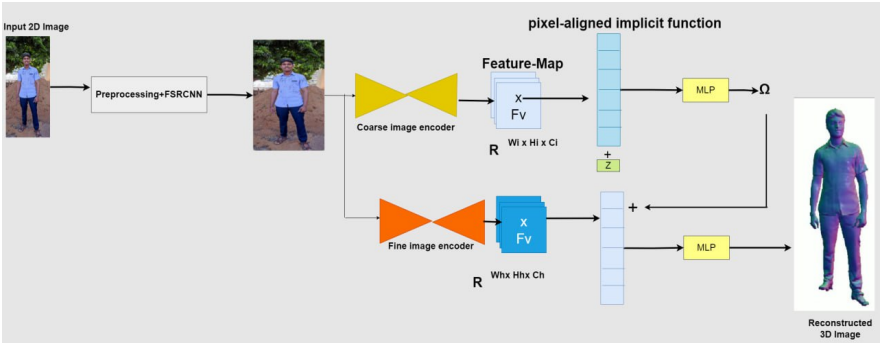

Figure 2. Multi-Level Pixel Implicit Function Architecture.

uses the feature maps along with the 3D embedding to create the Pixel-Aligned implicit function, which is then used to create the fine-level encoder. The output generated by the Pixel-Aligned implicit function is then passed through the fine-level encoder MLP. This process ultimately produces a 3D avatar of the human depicted in the input image as the final output.

Multi-Level Pixel-Aligned Implicit Function The Multi-Level pixel-aligned implicit Function [**?**] takes a 1024x1024 image as input and processes it through two levels of encoders.

The coarse level module
The multi-level approach in 3D human digitization is designed to follow the structure of the PIFu method described in previous research. It takes a down-sampled image of size 512x512 as input and generates backbone image features with a resolution of 128x128. The objective of this module is to incorporate global geometric information and establish a foundation for the subsequent fine-level module to add finer details to the 3D human digitization process. By processing the input image at a lower resolution, the module focuses on capturing larger-scale geometric features that are crucial for accurately representing the human form in 3D space.

The fine-level modules fine-level module:
In the proposed approach, it takes the original 1024x1024 resolution image as input and makes a backbone image feature with a resolution of 512x512. This resolution is four times higher than the resolution used in the previous PIFu research project. The fine-level module incorporates more subtle details into the 3D human digitization process that were not captured by the coarse-level module. By using 3D embedded features rather than absolute depth values, the fine-level module avoids relying on absolute depth values. This allows the module to focus on refining the geometric features provided by the coarse level and incorporating finer details into the final output. By utilizing these 3D embedding features, our approach achieves enhanced fidelity and an increased level of detail in the recon-

structed output. In our approach, we use a coarse-level module similar to that used in PIFu, but we modified it to enhance its performance. As a result of the modifications, the generated results should be more accurate and higher quality. The coarse-level module has been fine-tuned to better suit our approach by incorporating these adjustments, resulting in more precise and detailed results.

Multilayer Perceptron (MLP):

It is a type of artificial neural network with many layers of interconnected neurons. Each neuron in an MLP takes inputs, applies weights to those inputs, and then passes the weighted sum through an activation function to produce an output. In 3D avatar reconstruction, the MLP can estimate the avatar model parameters based on the extracted features. These parameters could include shape parameters (e.g., head size, body proportions), appearance parameters (e.g., skin color, hairstyle), or pose parameters (e.g., joint angles).

Predicting the exact shape and structure of the human back solely from images poses a challenge since direct observation is not possible. The reconstruction of the trunk using the MLP network adds another layer of difficulty to the task. Inherently ambiguous and multimodal problems result in back reconstructions that lack features. The occupancy loss function, designed to handle uncertainty, favors average reconstructions, further diminishing detail. The final layers of the MLP network face the challenge of learning a complex prediction function, adding to the overall complexity of the reconstruction process.

4 Experimental Results

4.1 Dataset Description.

In this study, we obtained high-quality 3D geometry and images from the Render People dataset . In total, 500 photogrammetry scans were analyzed, divided into 450 training scans and 50 test scans.Each subject's renderings were generated from various view points along the yaw axis while maintaining a fixed elevation of 0 degrees.To render the meshes, we employed pre-computed radiance transfer, utilizing 163 second-order spherical harmonics derived from HDRI Haven1. This technique allowed us to accurately capture and simulate the lighting conditions, leading to improved realism and visual quality in the rendered images. The combination of the Render People dataset and the radiance transfer approach provided a strong foundation for our research, enabling us to effectively evaluate and enhance our proposed methods. This integration played a crucial role in facilitating our research objectives and ensuring the validity of our findings. As discussed earlier our proposed architecture is built upon the Pixel-aligned Implicit Function (PIFu) framework [4]. The loss values obtained after comparing our model with the baseline model are shown in Table I.As compared to the baseline model, our model gives the less mean square error, a higher peak signal-to-noise ratio (PSNR), a lower mean square error, and a lower structural similarity index measure (SSIM).

In order to determine the similarity of two images, the Structural Similarity Index (SSIM) is used. It extracts three key features: luminance, contrast, and structure. The comparison between the two images is conducted based on these three features, enabling a comprehensive assessment of their similarity. The SSIM metric considers not only pixel values but also the structural information present in the images, providing a more robust evaluation of their resemblance.

SSIM The Structural Similarity Index (SSIM) is a metric utilized to measure the similarity between two provided images. It extracts three key features: luminance, contrast, and structure. The comparison between the two images is conducted based on these three features, enabling a comprehensive assessment of their similarity. The Structural Similarity Index (SSIM) is calculated as follows:

The Structural Similarity Index (SSIM) is calculated as follows:

$$SSIM(a,b) = \frac{(2\mu_a\mu_b + Z_1)(2\sigma_{ab} + Z_2)}{(\mu_a^2 + \mu_b^2 + Z_1)(\sigma_a^2 + \sigma_b^2 + Z_2)} \tag{1}$$

μ_a and μ_b represent the mean intensities of image patches a and b, respectively. σ_a^2 and σ_b^2 correspond to the variances of the intensities in image patches a and b, respectively. σ_{ab} indicates the covariance of intensities between the image patches a and b. Z_1 and Z_2 are constants introduced to stabilize the division and prevent division by zero. Typically, they are defined as $Z_1 = (k_1 L)^2$ and $Z_2 = (k_2 L)^2$, where L represents the dynamic range of the pixel values (e.g., $L = 255$ for 8-bit images). The constants k_1 and k_2 are small values, often set as 0.01 and 0.03 respectively.

MSE The Mean Square Error (MSE) is a metric used to quantify the cumulative squared error between a compressed image and its original counterpart. It provides a numerical measure of the overall discrepancy between the two images, with a lower MSE value indicating a higher level of fidelity in the compression process.

$$loss_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2 \tag{2}$$

PSNR An image's Peak Signal-to-Noise Ratio (PSNR) measures how well it compares with its original image after compression or reconstruction. It is typically expressed in decibels (dB), and a higher PSNR value indicates better image quality. The PSNR formula is as follows:

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right) \tag{3}$$

Table 1. comparison Table

	PIfu	Improved PIfu(our model)
MSE /s	108.81	37.9
PSNR /ms	9.594	11.244
SSIM /ms	0.156	0.58

The comparison table reveals that our Improved PIfu model outperforms the original PIfu model in three crucial metrics: MSE, PSNR, and SSIM. The substantially lower MSE of 37.9 signifies its superior accuracy in reconstructing and preserving image details. Additionally, the higher PSNR of 11.244 dB and superior SSIM of 0.58 demonstrate the improved signal-to-noise ratio, visual quality, and maintenance of structural information compared to PIfu's performance (MSE: 108.81, PSNR: 9.594 dB, SSIM: 0.156). These results confirm the effectiveness of our model for applications requiring high-quality images

Figure 3. 3D representation of a man with realistic anatomical proportions, facial features, and customizable clothing.

5 Conclusion

This research paper introduces a groundbreaking approach for 3D avatar reconstruction called the multi-level pixel-aligned implicit function (ML-PIFU) framework. The method effectively addresses the challenges of reconstructing high-fidelity 3D avatars from a single input image by employing a two-module pipeline that integrates global geometric information and local fine details. Extensive experiments validate the effectiveness and innovation of the approach, showcasing its superiority over existing state-of-the-art techniques. The comprehensive nature of the experiments solidifies the credibility and reliability of the findings, establishing ML-PIFU as a cutting-edge solution in the field. In summary, the method presents an accurate and precise approach for 3D avatar reconstruction.

References

1. Andrew Howard, M.Z.: pp. 4510–4520 (6 2018)
2. Angjoo Kanazawa, Michael J. Black, D.W.J.J.M.: End-to-end recovery of human shape and pose pp. 7122, 7131 (6 2018)
3. Boyi Jiang, Yang Hong, H.B.J.Z.: Selfrecon: Self reconstruction your digital avatar from monocular video (4 2022). https://doi.org/10.48550/arXiv.2201.12792
4. Garvita, B.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization pp. 1–8 (8 2019). https://doi.org/10.1109/ICCV.2019.00552
5. Jason Saragih, H.J.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization (6 2020). https://doi.org/10.1109/CVPR42600.2020.00016
6. Justin Johnson, Alexandre Alahi, L.F.F.: Perceptual losses for real-time style transfer and super-resolution (3 2016). https://doi.org/10.48550/arXiv.1603.08155
7. Kulkarni U, S M M, G.S.B.G.: Quantization friendly mobilenet (qf-mobilenet) architecture for vision based applications on embedded platforms (12 2020). https://doi.org/10.1016/j.neunet.2020.12.022.
8. L.-C. Chen, G. Papandreou, I.K.K.M.: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs pp. 1,10 (4 2017). https://doi.org/10.1109/TPAMI.2017.2699184
9. Lida Hu1, Q.G.: Automatic facial expression recognition based on mobilenetv2 in real-time (6 2020). https://doi.org/10.1088/1742-6596/1549/2/022136
10. Neverova, R.A.G.N.: ensepose: Dense human pose estimation in the wild (6 2018). https://doi.org/10.1109/CVPR.2018.00762
11. S. Saito, Z. Huang, R.N.S.M.A.K.: Pixel-aligned implicit function for highresolution clothed human digitization (12 2019). https://doi.org/10.48550/arXiv.1905.05172
12. Shunsuke Saito, Z.H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization pp. 1–8 (12 2019). https://doi.org/10.48550/arXiv.1905.05172
13. Shunsuke Saito1, .Z.H.: Pixel-aligned implicit function for high-resolution clothed human digitization (6 2018)
14. Thiemo Alldieck, Gerard Pons-Moll, C.T.M.M.: Tex2shape: Detailed full human body geometry from a single image (9 2019). https://doi.org/10.48550/arXiv.1904.08645
15. XIUMING ZHANG SEAN FANELLO, Y.T.T.: Neural light transport for relighting and view synthesis (1 2021). https://doi.org/10.1145/3446328
16. Zhong Li, L.C.: Neural network approximations of compositional functions with applications to dynamical systems (12 2020). https://doi.org/10.1109/CVPR.2018.00762
17. Zhong Li, Lele Chen, Y.X.: 3d human avatar digitization from single image pp. 1–8 (12 2019). https://doi.org/10.1145/3359997.3365707