Texturized Multi-level Implicit Modelling for High-Resolution 3D Human Digitization: The PIFuHD approach

1Shreya Junagade, 2Parth Gorde, 3Mihir Harne, 4Roma Thakur
[1,2,3,4]*Department of Information Technology MET BKC* Institute of Engineering Nashik, Maharashtra, India

Abstract: Our 3D human shape estimation network stands out for integrating volumetric feature transformation, merging diverse image features into 3D space to precisely recover surface geometry. Complemented by a rich dataset of 7000 real-world human models, our method, empowered by unique architecture, excels in single-image 3D human model estimation. Addressing challenges in estimating human pose and body shape from 3D scans over time, we introduce PIFuHD Pixel- aligned Implicit Function. PIFuHD enables end-to-end deep learning for digitizing detailed clothed humans from a single image, surpassing prior work with high-resolution reconstructions on the Render people dataset. Moreover, our innovative approach recovers fine details, even on occluded parts, by transforming shape regression into an aligned image-to-image translation problem. Using a partial texture map as input, our method estimates detailed normal and vector displacement maps, enhancing clothing representation on a low-resolution smooth body model. In the landscape of 3D human shape estimation, our multi-level architecture, balancing broad context and high resolution, significantly outperforms existing techniques, leveraging 1kresolution input images for enhanced single-image reconstructions.

1. INTRODUCTION

In recent years Texturized Multi-Level Implicit Modelling is introduced address the limitations of existing 3D human digitization methods. The proposed technique, PIFuHD, stands out for its utilization of deep neural networks and a combination of techniques to achieve precise reconstruction of high-resolution human models from 2D images. The core of PIFuHD lies in the blending of fully-connected techniques and deformable convolutional networks to accurately capture pose variations, surface details, and textures. This approach surpasses previous methods, providing a significant improvement in the authenticity of 3D human models. PIFuHD excels not only in its complex architecture but also in its ability to comprehend spatial relationships effectively through a careful interplay of feature extraction and fusion, resulting in three-dimensional representations that closely resemble the intricacies of the human form. The research delves into PIFuHD's capabilities, examining its architecture, training methods, and applications. The goal is to establish PIFuHD as a key player in the evolution of 3D human modelling. By demystifying its intricacies, this work contributes to the field's knowledge, offering insights that can impact industries such as virtual reality and human-computer interaction, promising a transformative future for 2D-to-3D human model synthesis. Pixel-aligned Implicit Function (PIFu), is novel approach for 3D deep learning in the context of textured surface inference for clothed 3D humans. The key innovation of PIFuHD lies in aligning individual local features at the pixel level to the global context of the entire object in a fully convolutional manner, enabling it to reason about 3D shapes accurately, even from a single view. The method employs an encoder to learn per-pixel feature vectors that consider the global context, preserving local details while inferring plausible ones in unseen regions. PIFuHD's end-to-end and unified digitization approach can predict high-resolution 3D shapes of individuals wearing complex clothing and hairstyles, handling a variety of clothing types and capturing high-frequency details like wrinkles at the pixel level. Additionally, PIFuHD can naturally extend to infer per-vertex colours, generating a complete texture of the surface. Comprehensive evaluations against ground truth 3D scan datasets highlight PIFuHD's

state-of-the-art performance in digitizing clothed humans.

2. LITERATURE SURVEY

Several papers focused on reconstructing 3D human models from images or video using methods like skinned multi-person linear models, convolutional neural networks, and volumetric discretization. However, they had limitations in capturing details like facial expressions, hands, and clothing. 2020, Saito etal[1]:They introduced ”PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization.” Their approach utilized a multi-level pixel-aligned implicit function to achieve high fidelity 3D human models. However, a notable limitation of this method is its reconstruction of 3D humans without incorporating textures or colours, potentially hiding the realism of the generated models. 2019, Zheng, Yu[2]: Zheng, Yu and their collaborators from Beihang University and Orbbec Company presented ”Deep-Human: 3D Human Reconstruction from a Single Image.” They implemented a 3D representation using the Skinned Multi-Person Linear (SMPL) model through volumetric discretization. While successful in generating 3D representations, this method faced challenges in re- constructing hands and facial expressions, limiting the comprehensiveness of the reconstructed human models 2019, Alldieck, Magnor[3]:In the same year, Alldieck, Magnor, and Bhatnagar proposed ”Learning to Reconstruct People in Clothing from a Single RGB Camera.” They trained a Convolutional Neural Network (CNN) to infer a 3D mesh model while simultaneously reconstructing the subject’s 3D model. How- ever, key limitation was the reliance on synthetic data, raising concerns about the generalization of the model to real-world scenarios. 2018, Joo, Simon and Sheikh[4]:They introduced ”Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies.” This method utilized the Frank model, integrating body part models to reconstruct a complete 3D human model. Despite its capability to capture overall body movements, the approach had limitations, such as providing limited surface detail and complexity in model integration. 2020, Aymen Mir1 ,Thiemo Alldieck[5]:They provide effective method to automatically transfer textures of clothing images (front and back) to 3D garments worn on top SMPL [42], in real time. Their model opens the door for applications such as virtual try-on and allows for generation of 3D humans with varied textures which is necessary for learning.

3. PIFUHD

Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization is an extension of the original PIFu method designed to achieve higher-resolution 3D human digitization. While the original PIFu focuses on reconstructing detailed 3D geometry and texture from a single image, PIFuHD aims to capture even finer details and higher resolutions.

In PIFuHD, the architecture is enhanced to handle high-resolution images and produce more detailed 3D reconstructions. This is particularly important for applications where capturing fine details, such as wrinkles in clothing or subtle facial features, is crucial.

PIFuHD is designed to handle high-resolution images, allowing for more detailed and accurate 3D reconstructions. The implicit function used in PIFuHD is refined to better capture fine details, resulting in more accurate surface representations.

PIFuHD often employs multi-scale processing to capture details at different resolutions, enabling it to handle both global and local features effectively. To maintain the quality of the reconstruction at higher resolutions, PIFuHD may employ more sophisticated sampling strategies or adaptive sampling techniques PIFuHD may incorporate additional regularization techniques to prevent overfitting and ensure smooth and realistic reconstructions. Training PIFuHD requires carefully curated datasets and may involve fine-tuning on high-resolution images to achieve the desired level of detail and accuracy.

4. METHODOLOGY

The method leverages the Pixel-aligned Implicit Function (PIFuHD) framework to advance the resolution of 3D human digitization. Initially, PIFuHD processes images at 512 × 512 resolution to generate low-resolution feature embeddings at 128 × 128. For enhanced resolution, an additional pixel-aligned prediction module is introduced, which processes 1024 × 1024 resolution images to encode high-resolution features at 512 × 512. This module utilizes high-resolution feature embeddings along with 3D embeddings from the first module to predict an

occupancy probability field. To further refine the reconstruction quality, the method predicts normal maps for both front and back sides in image space, incorporating them as additional input. The foundational concept of PIFuHD is briefly outlined, emphasizing its objective of 3D human digitization by estimating occupancy within a dense 3D volume. PIFuHD models a function f(X), predicting binary occupancy values for any 3D position in continuous camera space X=(Xx,Xy,Xz)ˆaR, outputting 1 if X is inside the mesh surface and 0 otherwise, based on a single RGB image.

Coarse Level: Within the PIFuHD architecture, the coarse level takes low-resolution images to encompass a broader spatial context and facilitates holistic reasoning. It furnishes context to the fine level for estimating highly detailed geometry. This level integrates global geometric information by processing a down sampled 512×512 image, generating backbone image features at 128×128 resolution. Conversely, the fine level introduces more nuanced details using the original 1024 × 1024 resolution image, resulting in backbone image features at 512×512 resolution. Notably, the fine level adopts 3D embedding features from the coarse level rather than absolute depth values. This module is structured similarly to PIFuHD but incorporates predicted frontside and backside normal maps.

Fine Level: The fine level within the PIFuHD framework aims to enhance the 3D human digitization process by adding subtle details. It utilizes the original high-resolution (1024x1024) image as input and generates backbone image features at 512 × 512 resolution, which is quadruple the resolution of previous implementations. This module employs high-resolution input and integrates a 3D embedding extracted from the coarse level network. Although the receptive field of the fine level doesn’t encompass the entire image, its fully convolutional architecture enables training with a random sliding window while inferring at the original image resolution, i.e., 1024 × 1024. These normal maps steer the 3D reconstruction to yield sharper geometry. The prediction of normal maps for both back and front utilizes a pix2pixHD network.

Texturized Fine Level: This stage enhances 3D human models by adding vibrant textures. It utilizes the original 1024x1024 image for texture mapping and generates detailed features at 256 × 256 resolution. Leveraging the 3D information from the Fine level, it applies these textures to the human model, enriching its visual appearance. In summary, the Texturized Fine Level method completes our multi-level 3D modelling approach by infusing the model with detailed and vibrant textures sourced from high-resolution images. This method not only enhances the model's visual realism but also ensures that the final 3D model is a harmonious blend of structural accuracy, intricate details, and visual appeal.

5. ALGORITHM

The marching cubes algorithm is a popular technique used in the creation of 3D human models.

• Data Acquisition: The first step is to obtain 3D data of the human body, often through the use of scanning technologies such as laser scanning, structured light scanning, or photogrammetry. These techniques can capture the surface geometry of the human body with high accuracy and detail.

• Volume Data Representation: The 3D scan data is typically represented as a volumetric grid, where each grid cell (or voxel) contains a value that represents the density or material properties of the underlying object. This volume data can be obtained by converting the 3D scan data into a 3D array or grid.

• Marching Cubes Algorithm: The marching cubes algorithm is then applied to the volume data to extract a polygonal mesh representation of the human model. The algorithm works by iterating through the volume data, examining each cube-shaped cell (or "marching cube") and determining how to triangulate the surface that passes through that cell based on the density values at the cube's vertices.

• Surface Reconstruction: The marching cubes algorithm generates a set of triangles that represent the surface of the human model. These triangles can then be further processed, smoothed, and optimized to create a high-quality 3D mesh representation of the human body.

Overall, the marching cubes algorithm is a powerful and widely-used technique for generating 3D human models from volumetric data, and it has been instrumental in the development of many applications in fields such as computer graphics, virtual reality, and medical imaging.

6. ARCHITECTURE

The methodology for creating a 3D human model begins with the Loading of a 2D Image, where users can upload a 2D image as the primary input for the modelling process. Following this, the Image Processing stage is initiated to remove the background and any non-human elements from the image. This step ensures that the subsequent model focuses solely on the human subject, enhancing the accuracy and realism of the final 3D model.

Once the image is processed, the Generate 3D Model phase utilizes the Pixel-aligned Implicit Function High-Definition (PIFuHD) technique to transform the 2D image into a 3D mesh. PIFuHD's advanced algorithms and neural network architectures enable the generation of a detailed and accurate 3D representation of the human subject from a single 2D image. This approach eliminates the need for multi-view setups or complex scanning equipment, simplifying the modelling process while maintaining high-quality results.

After generating the 3D mesh, the Apply Textures stage enriches the model by applying textures derived from the original 2D image. This step ensures that the 3D model retains the visual details, colours, and textures of the human subject, further enhancing its realism and visual fidelity.

SYSTEM PROPOSED ARCHITECTURE

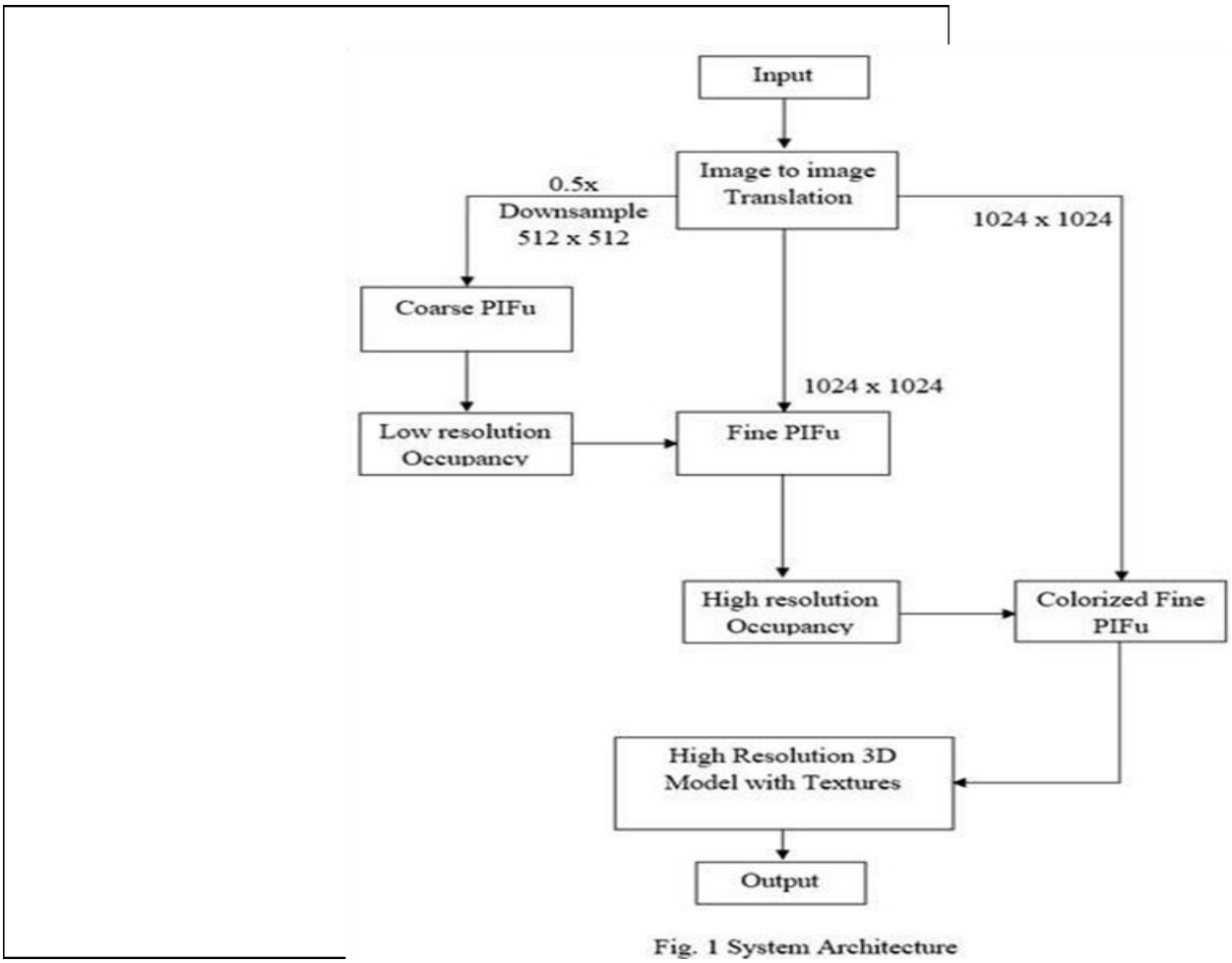


Figure 1: the system model of the proposed design

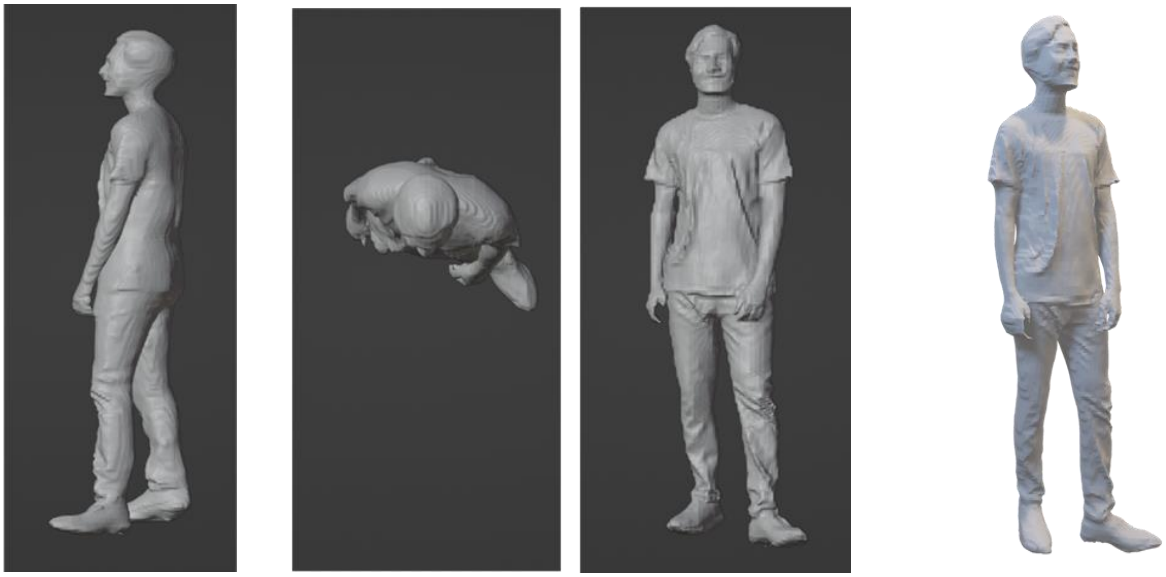
Finally, in the Render 3D Model phase, the system renders the textured 3D model for visualization. Using advanced rendering techniques and visualization tools, the system presents the 3D model in a format that can be viewed, analyzed, or integrated into various applications. This rendered 3D model serves as the final output, ready for deployment in applications such as virtual try-ons, simulations, or virtual environments, demonstrating the effectiveness and versatility of the proposed methodology.

7. RESULT

Our model showcases the outcomes of our digitization process using authentic images sourced from the Deep Fashion dataset [31]. Our PIFu framework exhibits proficiency in handling a diverse array of clothing items, encompassing skirts, jackets, and dresses. Notably, our approach achieves high-resolution local detailing and accurately infers 3D surfaces even in regions not directly visible in the input image. Furthermore, our method successfully deduces comprehensive textures from just a single input image, facilitating a full 360-degree view of our generated 3D models. For a more comprehensive view of our results, including both static and dynamic outputs, readers are directed to the supplementary video. This video demonstrates the capability of our approach to capture dynamic human movements in clothed scenarios and intricate deformations, all from a single 2D video input.

Figure 2: Image to Image processing of a 2D image

3D Human Model Presentation in different Views



(a) side View (b) Top view (c) front View (d) more refined textures of the 3D model

8. CONCLUSION

In conclusion our project, which focuses on generating 3D human models from high-resolution images, showcases the boundless creativity and innovation of human capabilities. This advancement stands poised to reshape numerous fields, with entertainment being just the tip of the iceberg. As we tackle the challenges and seize the opportunities that this venture offers, our dedication to pushing the boundaries in computer graphics and computer vision remains unwavering. Through teamwork, commitment, and a shared vision, we're unlocking the vast potential of lifelike 3D human models. We're steering towards a future where digital realism is limitless, ushering in fresh creative avenues and revolutionizing various industries.

In essence, the PIFuHD methodology marks a significant advancement in 3D modelling technology, seamlessly transforming 2D images into intricate and lifelike 3D representations. As research continues and technology advances, we can look forward to further enhancements in precision, efficiency, and versatility, cementing its pivotal role in the fields of computer graphics and computer vision.

9. FUTURE WORK

We introduce a multi-level framework designed to perform joint reasoning over holistic and local details, facilitating high-resolution 3D reconstructions of clothed humans from a single image without the need for additional post-processing or auxiliary information. Our multi-level Pixel-Aligned Implicit Function accomplishes this by progressively integrating global context through a scale pyramid as an implicit 3D embedding. This approach circumvents premature decisions about explicit geometry, a limitation observed in prior methods. Our experiments underscore the significance of incorporating 3D-aware context for achieving accurate and precise reconstructions. Additionally, by addressing ambiguity in the image-domain, we significantly enhance the consistency of 3D reconstruction details in occluded regions.

Given that the effectiveness of the multi-level approach is contingent on the preceding stages' success in extracting 3D embeddings, improving the baseline model's robustness is crucial for enhancing overall reconstruction accuracy. Future research directions might involve integrating human-specific priors, such as semantic segmentations, pose information, and parametric 3D face models. Incorporating 2D supervision of implicit surfaces could further bolster the methodology's performance and capabilities.

REFERENCE

List all the material used from various sources for making this project proposals:

[1] Shunsuke Saito, University of Southern California, Facebook Reality Labs, Facebook AI Research, PIFuHD: Multi-Level Pixel-Aligned Implicit Function For High-Resolution 3D Human Digitization 2020.

[2] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. PonsMoll. Learning to reconstruct people in clothing from a single RGB camera. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1175ˆa1186, 2019.

[3] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In IEEE Conference on Computer Vision and Pattern Recognition, pages 8387aˆ 8397, 2018.

[4] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor.Tex2shape: Detailed full human body geometry from a single image. In The IEEE International Conference on Computer Vision (ICCV), October 2019 .

[5] R.Alp GAˆ zˇ uler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7297aˆ 7306, 2018 .

[6] Shunsuke Saito, University of Southern California, Facebook Reality Labs, Facebook AI Research, PIFuHD: Multi-Level Pixel-Aligned Implicit Function For High-Resolution 3D Human Digitization 2020.

[7] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. PonsMoll. Learning to reconstruct people in clothing from a single RGB camera. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1175ˆa1186, 2019.

[8] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of

3d people models. In IEEE Conference on Computer Vision and Pattern Recognition, pages 8387aˆ 8397, 2018.

[9] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor.Tex2shape: Detailed full human body geometry from a single image. In The IEEE International Conference on Computer Vision (ICCV), October 2019 .

[10] R.Alp GAˆ z˙ uler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7297aˆ 7306, 2018 .