iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57389

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Virtual Styling and Fitting using AI

Tarash Budhrani[1], Parth Jain[2], Neelanjaan De[3], Bharat Dedhia[4]
[1, 2, 3, 4]*Information Technology, K J Somaiya College of Engineering*

Abstract: ***The digitization of human forms holds significant relevance across domains like virtual reality, medical imaging, and robot navigation. While sophisticated multi-view systems excel at precise 3D body reconstruction, they remain largely inaccessible to the general consumer market. Seeking more accessible approaches, methods in human digitization explore simpler inputs, such as single images. Among these approaches, pixel-aligned implicit models [1] have garnered attention for their efficiency in capturing intricate geometric details like clothing wrinkles. These models, notably lightweight, employ parametric human body models without necessitating mappings between canonical and posed spaces. PIFu [1], a prime example, translates a parametric body model into pixel-aligned 2D feature maps, offering more comprehensive information than mere (x, y, z) coordinates. This research paper, rooted in PIFu [1], delves into the implementation intricacies of 3D human digitization within a specific domain – 3D digitization for virtual styling and fitting. In today's tech-driven world, online shopping has boomed, offering unparalleled convenience by letting users swiftly browse and buy items from home. However, despite its speed and ease, trying on clothes remains a hurdle. Without a physical store, customers struggle to gauge fit and size, leading to increased returns and order abandonment. To tackle this, a project aims to revolutionize the online clothing shopping experience. By offering a solution that allows virtual try-ons, users can visualize how clothes look and fit before buying, potentially reducing return rates and enhancing the overall shopping journey.***

I. OBJECTIVE OF THE RESEARCH

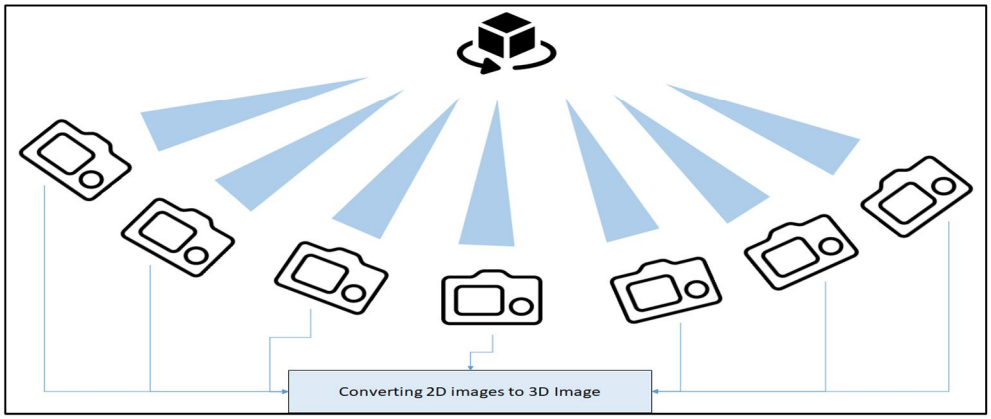
The central objective of this research is to demonstrate the capabilities of an artificial intelligence system designed to create a detailed three-dimensional representation of a person using an image supplied by the user. Subsequently, this AI-driven system enables users to virtually experiment with different clothing options on the generated 3D model. The aim of the 3D virtual fitting and styling model is to present an extensive visualization, offering a complete 360-degree perspective on how various products will look and fit on the individual. Users gain the ability to mix and match different clothing items, creating and visualizing a cohesive and personalized style.

The project attached to this research paper is aimed to achieve the following:

1) Development of a comprehensive 3D human body model tailored to individual customers based on their body measurements and provided images.
2) Incorporation of features within the application to personalize the body model, encompassing attributes such as skin color, hairstyle, hair color, and more.

II. BACKGROUND

Although it is now feasible to create detailed and exact replicas of people using multi-view systems, this capability has mostly been out of reach for most people because it depends on expensive professional capture setups that have strict requirements like a large number of cameras and controlled lighting. These setups are difficult and costly to set up. The below diagram explains this.



A. *Precursor to this Research Paper*

This work aims to accomplish a highly detailed 3D reconstruction of clothed individuals from just a single image, capturing fine elements like fingers, facial expressions, and intricate clothing folds. Current methods fall short in utilizing the full potential of high-resolution images. This is because they rely on a holistic approach to link the 2D appearance of a photographed person to their 3D form, overlooking crucial cues found in localized image patches that are pivotal for detailed 3D rendering.

Efforts to tackle this constraint fall into two main categories:

1) *Shape from Shading* [5]: This method overlays high frequency details onto low-resolution surfaces. It employs a lower image resolution to capture a general shape, subsequently supplementing it with fine details like surface normal or displacements through post-processing. This method's primary limitation lies in its reliance on low-resolution surfaces supplemented with high-frequency details. While it initially captures a general shape using lower image resolutions and enhances it with fine details during post-processing, this approach can struggle to accurately depict intricate features. Additionally, the reliance on post-processing for adding detailed information might result in inconsistencies or inaccuracies in the final reconstruction.
2) *Scape* [6]: This approach employs highly detailed human models to simulate realistic intricacies. While both methods yield reconstructions that seem detailed, they often fail to accurately replicate the true intricacies present in the original input image. Although SCAPE utilizes highly detailed human models to create realistic nuances, it encounters limitations in faithfully reproducing the exact intricacies present in the original input image. Despite generating reconstructions that appear detailed, there's a gap between the simulated details and the actual finer elements visible in the original image. This method might fail to capture subtle nuances and intricacies accurately, leading to a loss of fidelity in the reconstructed output.

B. *3D Modelling using Pixel-Aligned Implicit Function*

Pixel-Aligned Implicit Function (PIFu) [1] introduces an incredibly powerful implicit method that harmonizes the local pixel details of 2D images with the wider context of their corresponding 3D object. PIFu presents a comprehensive deep learning approach for digitizing highly intricate clothed human figures, capable of deducing both 3D surface structures and textures from just a single image, and if needed, multiple input images. This method unifies the digitization of complex shapes like hairstyles, clothing variations, and deformations into a seamless process. In contrast to existing 3D deep learning representations, PIFu excels in generating high-resolution surfaces, even in areas traditionally unseen, such as the back of a person. Notably, it maintains efficiency in memory usage, unlike voxel representations, accommodates diverse surface topologies, and ensures spatial alignment with the input image. Moreover, while prior techniques cater to either single-image processing or multiple views, PIFu naturally extends its capabilities to accommodate an arbitrary number of perspectives.

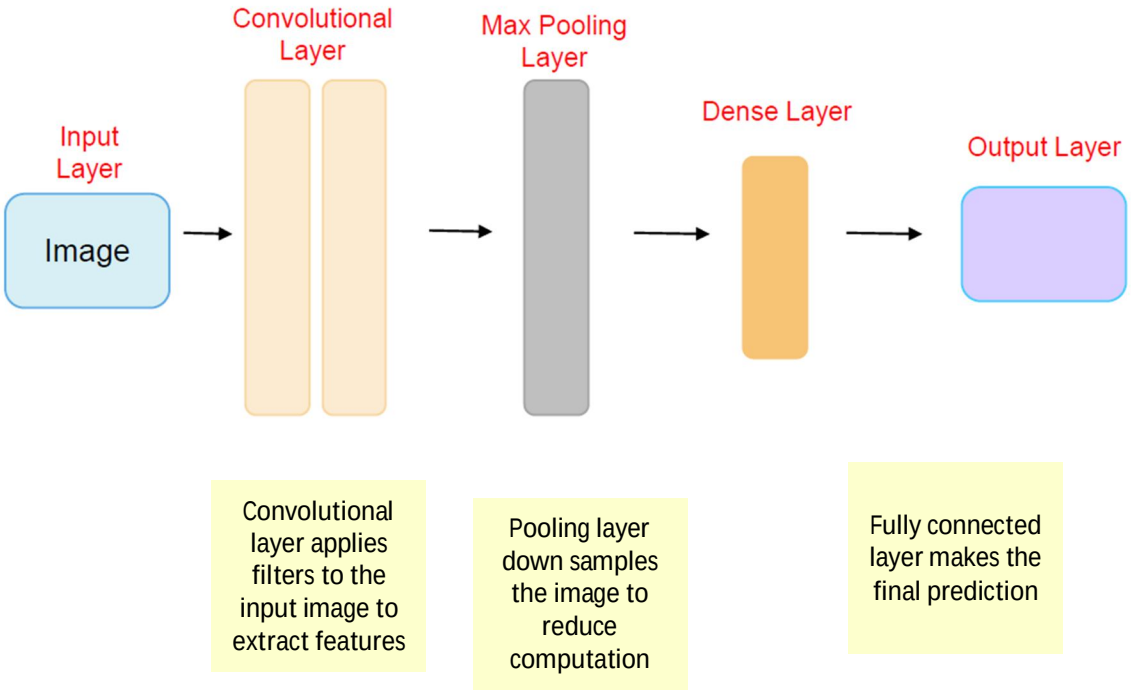
III. METHODOLOGY

The approach, whether given a solitary image or multiple viewpoints, revolves around reconstructing the intrinsic 3D structure and texture of a clothed human while upholding the image's intricate details. The formulation of an implicit function delineates a surface as the zero set of a function f, such as $f(X) = 0$ [50]. This approach yields a resource-efficient representation of a surface, obviating the necessity to explicitly store the embedding space of said surface. The novel pixel-aligned implicit function proposed comprises a fully convolutional image encoder g and a continuous implicit function f delineated by multi-layer perceptrons (MLPs). Here, the surface is defined as a level set of $f(F(x), z(X)) = s : s \in R$, wherein for a 3D point X, $x = \pi(X)$ represents its 2D projection, $z(X)$ signifies the depth value within the camera's coordinate space, and $F(x) = g(I(x))$ represents the image feature at x. To attain the pixel-aligned feature $F(x)$, we employ bilinear sampling, recognizing that the 2D projection of X is defined within a continuous space, not discretized like pixels. Learning an implicit function across the 3D realm utilizing pixel-aligned image features, rather than global features, empowers the learned functions to uphold the intricate local details discernible in the image. The continuous framework of PIFu enables the creation of detailed geometry with varied topologies in an efficient use of memory resources.

The crucial insight lies in our approach of acquiring an implicit function across 3D space using pixel-aligned image features instead of global ones. This strategy enables the learned functions to retain the specific local details inherent in the image. Leveraging the continuous nature of PIFu [1] enables the generation of intricate and varied geometries with efficiency in memory usage. Additionally, PIFu stands as a versatile framework capable of extension into diverse co-domains, including but not limited to RGB colors.

A. Design Principles

Yes, Pixel Aligned Implicit Function (PIFu) is based on convolutional neural networks (CNNs) [4] to some extent. PIFu leverages neural networks, particularly CNNs, as a fundamental component within its framework to capture and process image information effectively. In PIFu, CNNs play a pivotal role in encoding and extracting features from input images. For instance, when reconstructing a 3D shape from a single 2D image, the CNN is employed to encode the image features and generate a continuous representation of the 3D shape.

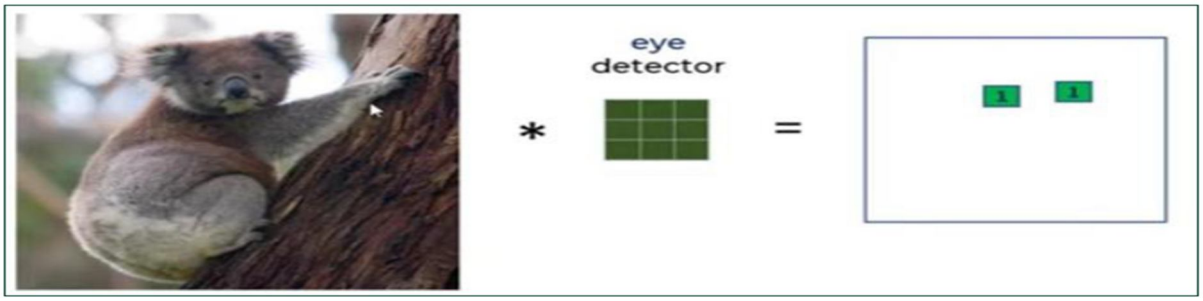


Here's a simplified breakdown of how a CNN might be used within the PIFu framework

1) *Feature Extraction:* The CNN takes in the 2D image as input and passes it through several convolutional layers. These layers are designed to detect and extract hierarchical features, learning patterns from the image data such as edges, textures, and shapes.
2) *Encoding Spatial Information:* The CNN learns to encode spatial information by understanding the relationships between different parts of the image. This encoded representation retains valuable information about the image content.
3) *Implicit Function Learning:* PIFu uses the output of the CNN as a basis for learning an implicit function that maps the 2D image to a 3D representation. This function estimates the continuous 3D shape, incorporating fine details like wrinkles in clothing or subtle variations in geometry.

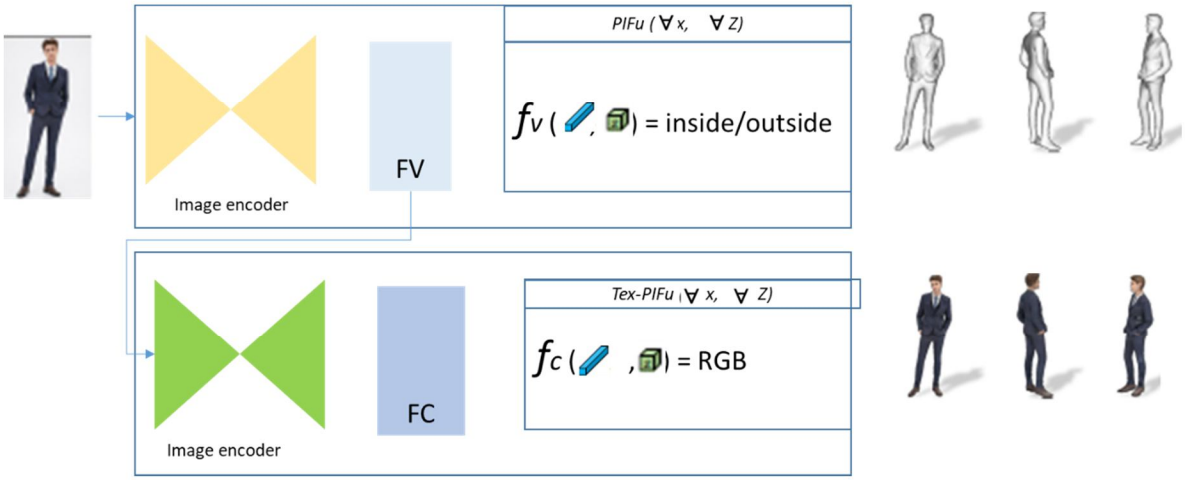
For example, imagine a scenario where you have a single 2D image of a person wearing a coat. The CNN within PIFu would process this image, extracting features like the outline of the person, details of the clothing, and other relevant visual cues. Through its learned representation, PIFu's implicit function then reconstructs a detailed 3D model of the person, considering the clothing's texture, fit, and overall shape based on the information extracted by the CNN.

PIFu's strength lies in its ability to leverage CNNs to understand and process complex visual information, enabling the reconstruction of detailed 3D shapes from limited 2D inputs.



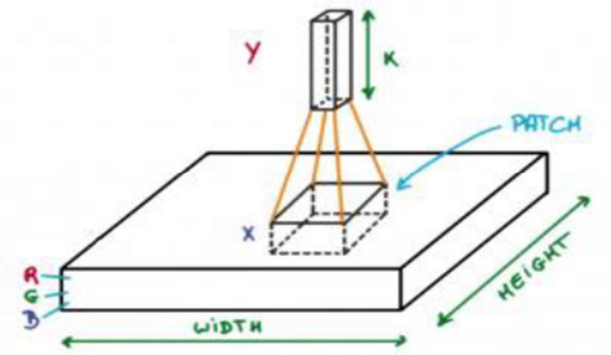
B. Digitization of an Image

When presented with an input image, the pixel-aligned implicit function (PIFu) operates by predicting the continuous inside/outside probability field concerning a clothed human. Likewise, in the case of texture inference (Tex-PIFu), it deduces the RGB value corresponding to specific 3D positions within the surface geometry, accommodating diverse topologies.



In surface reconstruction, PIFu utilizes the input image to forecast the continuous inside/outside probability field of a clothed human, facilitating the easy extraction of iso-surfaces (Sec. 3.1). Concurrently, in texture inference (Tex-PIFu), the model generates RGB values for 3D positions on the surface geometry, enabling texture inference even in self-occluded surface regions and shapes with varying topologies. Additionally, our approach seamlessly handles both single-view and multi-view inputs, showcasing enhanced results with increased fidelity when multiple views are available.

A Convolutional Neural Network (is like a super smart detective specializing in recognizing patterns in images Imagine you're looking at a huge picture full of details, trying to find specific shapes or objects like distinguishing between a cat and a dog. The CNN is built with layers that act like detectives working together Each layer focuses on certain aspects of the image, like edges or textures These detectives (called filters or kernels) examine small portions of the picture at a time kind of like looking through a magnifying glass at different parts of the image.



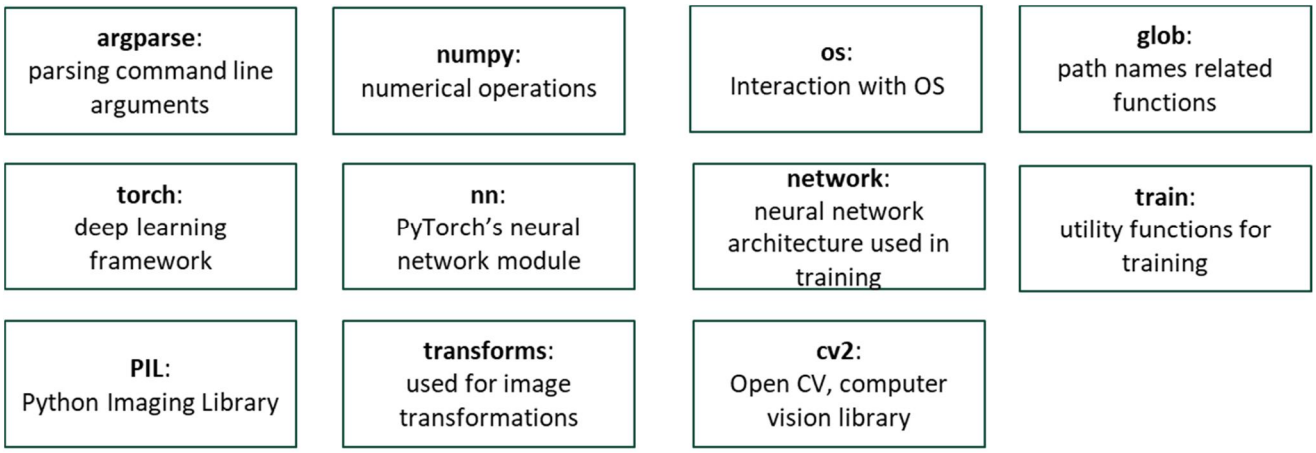
They do this by sliding across the image, picking up important features and passing on what they find to the next layer of detectives These layers gradually piece together the information to understand the bigger picture like recognizing that certain combinations of edges and textures form the shape of a cat's ear or a dog's tail. By working together in this way, CNNs can learn to recognize and classify images making them really good at tasks like identifying objects in photos, reading handwritten text, or even diagnosing medical conditions from scans. They're powerful tools that have revolutionized image recognition and analysis.

Let's delve into the mathematical intricacies intertwined within the convolution process. Convolution layers are comprised of a collection of adaptable filters, or kernels, characterized by petite widths, heights, and identical depth to that of the input volume. For instance, when applying convolution to an image measuring 34 x 34 x 3, the conceivable filter size could be axax 3, where 'a' can vary—say, 3, 5, or 7—yet remains smaller in comparison to the image's dimensions.

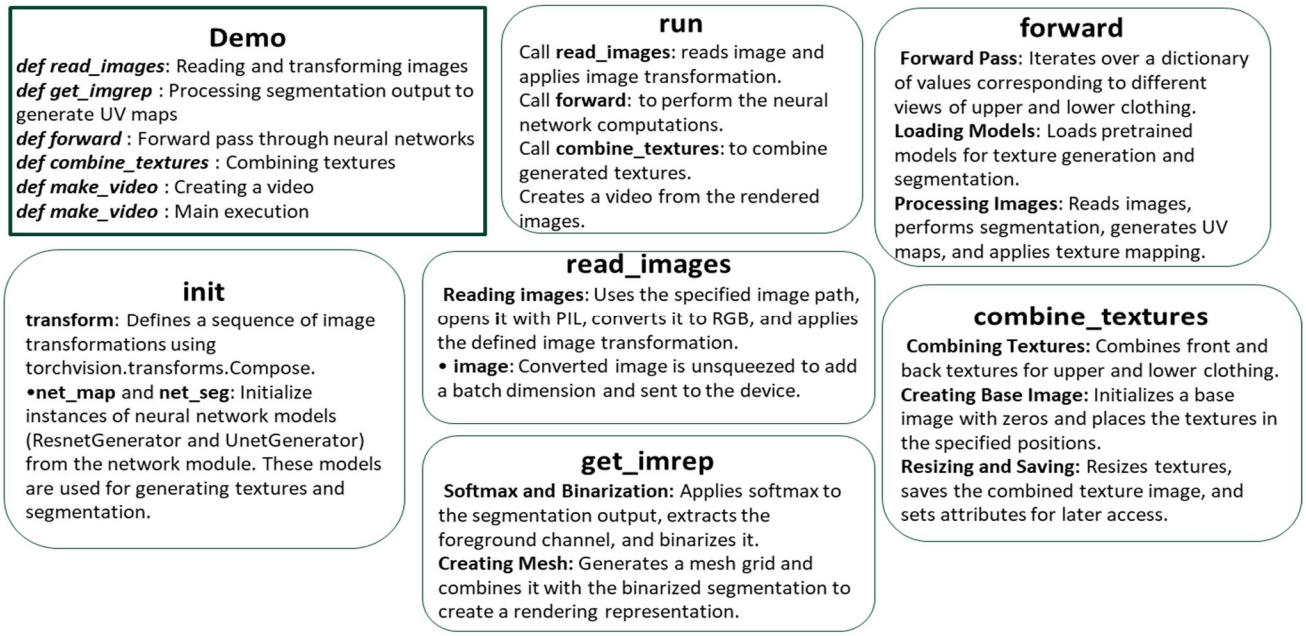
During the forward pass, each filter glides systematically across the entire input volume, maneuvering in increments known as strides, which can take on values such as 2, 3, or even 4 for images of heightened dimensions. At each step, the dot product is computed between the kernel weights and the corresponding patch extracted from the input volume. As these filters traverse, they generate a 2D output for each, subsequently stacked together, culminating in an output volume possessing a depth equivalent to the number of filters employed. This process enables the network to acquire the ability to learn the nuances encapsulated within these filters.

IV. SOLUTION DESIGN

The solution is designed and implemented in Python. The Convolutional Neural Network implementation is achieved by leveraging the libraries available in the Python stack.



Here is how the Class view of the solution look.



V. RESULTS

Figure 4 showcases our digitization outcomes utilizing real-world input images sourced from the DeepFashion dataset. Our PIFu method adeptly manages an array of garments, spanning skirts, jackets, and dresses. It excels in generating high-resolution intricate details while inferring credible 3D surfaces in unexplored areas. Our technique proficiently derives comprehensive textures from a single input image, enabling a comprehensive 360-degree view of our 3D models. For further insights, we refer to supplemental video2, featuring both static and dynamic outcomes. Specifically, we elucidate how 3D digitization from a solitary 2D input video captures dynamic clothed human performances and intricate deformations.

We subject our reconstruction accuracy to quantitative evaluation via three metrics. Within the model space, we gauge the average point-to-surface Euclidean distance (P2S) in cm, measuring from the vertices on the reconstructed surface to the ground truth. Additionally, we calculate the Chamfer distance between the reconstructed and ground truth surfaces. To delve deeper, we introduce the normal reprojection error, assessing the fidelity of reconstructed local details and the projection consistency from the input image viewpoint. Both reconstructed and ground truth surfaces undergo rendering into normal maps within the image space. Subsequently, we compute the L2 error between these two normal maps to ascertain alignment and accuracy.

VI. CONCLUSION

In conclusion, the Pixel Aligned Implicit Function (PIFu) [1] stands as a transformative approach, revolutionizing the realm of 3D reconstruction from 2D images. Its ability to extrapolate detailed 3D structures from single images while preserving high-resolution local details marks a significant leap in the field. PIFu's adaptability across diverse clothing types, its capacity to handle dynamic performances from videos, and its robust quantitative evaluation metrics underscore its efficacy and versatility.

The potential applications of PIFu span across various domains, including virtual try-on experiences in e-commerce, augmented reality, medical imaging, and entertainment industries. As highlighted throughout this paper, PIFu's accuracy in reconstructing 3D models, coupled with its capacity for fine detail preservation and consistency, positions it as a frontrunner in advancing digital reconstruction technologies.

Nevertheless, there remain avenues for further exploration and refinement. Future research could delve deeper into enhancing computational efficiency, refining accuracy in extreme poses or lighting conditions, and expanding its applicability to diverse object categories beyond apparel.

The strides made by PIFu in enabling realistic, immersive 3D reconstructions signify not just technological advancements but also the potential for redefining user experiences in numerous sectors. As the landscape of computer vision and 3D reconstruction continues to evolve, PIFu stands as a beacon, guiding further innovations and opening doors to uncharted possibilities.

REFERENCES

[1] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2020[1].

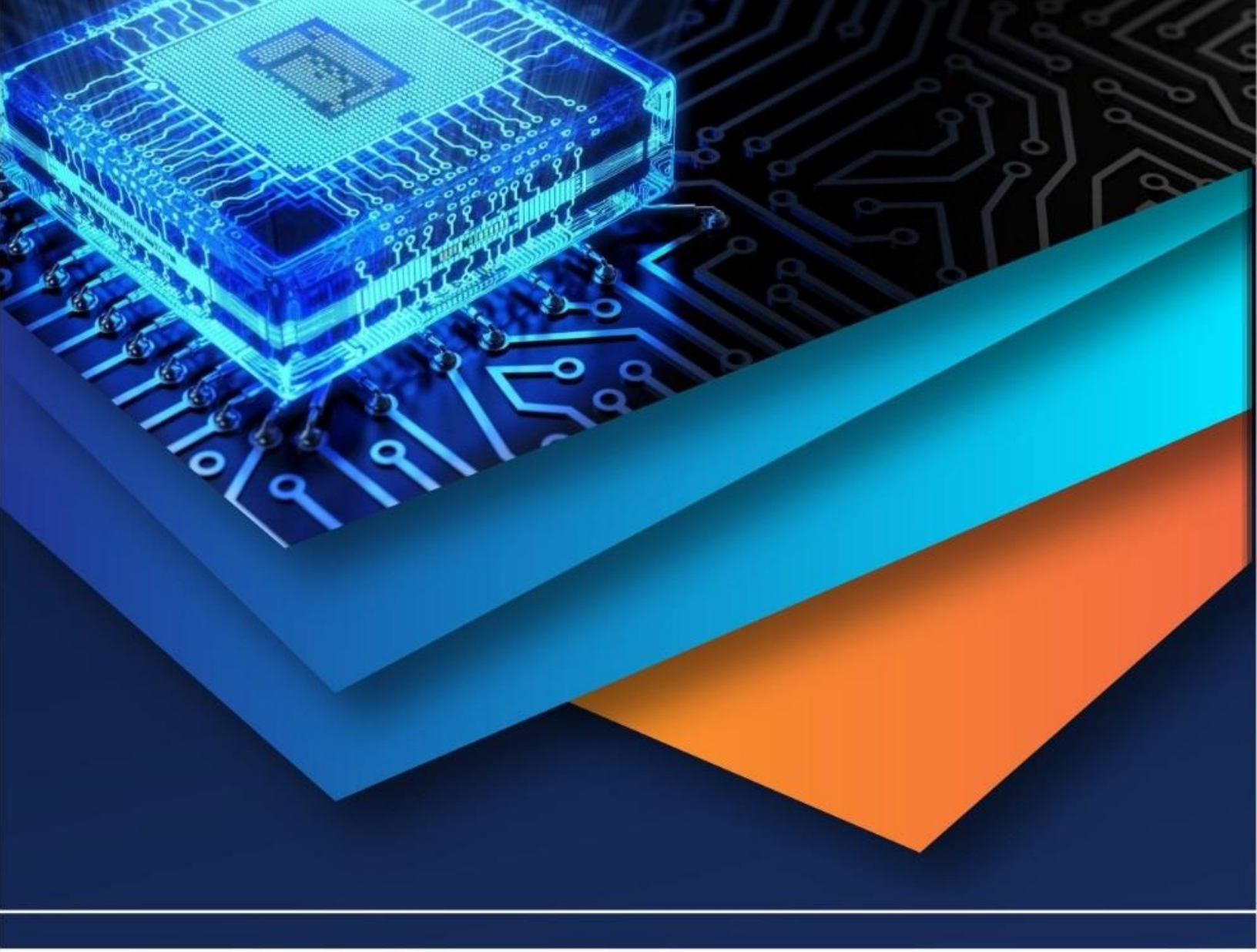
[2] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, October 2019. Human3.6M: This dataset provides 3D human pose and shape data captured in controlled environments, making it useful for evaluating pose and body shape estimation. MPII Human Pose Dataset: This dataset includes images of people in various poses, and it can be used for pose estimation tasks.

[3] Y. Wu and K. He. "Group Normalization." In European Conference on Computer Vision, pp. 3–19, 2019. This source is tapped for Image classification dataset.

[4] Convolutional Neural Networks, Explained. Published in Towards Data Science, Aug 27, 2020 https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

[5] Shape from Shading From: Visual Thinking for Information Design (Second Edition), 2022 https://www.sciencedirect.com/topics/computer-science/shape-from-shading

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. ACM Transactions on Graphics, 2005.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY