


Article

Enhanced Multi-Scale Attention-Driven 3D Human Reconstruction from Single Image

Yong Ren ^{1,2}, Mingquan Zhou ^{1,2,*}, Pengbo Zhou ³, Shibo Wang ^{1,2}, Yangyang Liu ^{1,2}, Guohua Geng ^{1,2}, Kang Li ^{1,2,*}  and Xin Cao ^{1,2,*}

¹ School of Information Science and Technology, Northwest University, Xi'an 710127, China; ryrory@stumail.nwu.edu.cn (Y.R.); wangshibo@stumail.nwu.edu.cn (S.W.); yylliu@nwu.edu.cn (Y.L.); ghgeng@nwu.edu.cn (G.G.);

² National and Local Joint Engineering Research Center for Cultural Heritage Digitization, Xi'an 710127, China

³ School of Art and Media, Beijing Normal University, Beijing 100875, China; zhoupengbo@bnu.edu.cn

* Correspondence: mqzhou@nwu.edu.cn (M.Z.); likang@nwu.edu.cn (K.L.); caoxin@nwu.edu.cn (X.C.)

Abstract: Due to the inherent limitations of a single viewpoint, reconstructing 3D human meshes from a single image has long been a challenging task. While deep learning networks enable us to approximate the shape of unseen sides, capturing the texture details of the non-visible side remains difficult with just one image. Traditional methods utilize Generative Adversarial Networks (GANs) to predict the normal maps of the non-visible side, thereby inferring detailed textures and wrinkles on the model's surface. However, we have identified challenges with existing normal prediction networks when dealing with complex scenes, such as a lack of focus on local features and insufficient modeling of spatial relationships. To address these challenges, we introduce EMAR—Enhanced Multi-scale Attention-Driven Single-Image 3D Human Reconstruction. This approach incorporates a novel Enhanced Multi-Scale Attention (EMSA) mechanism, which excels at capturing intricate features and global relationships in complex scenes. EMSA surpasses traditional single-scale attention mechanisms by adaptively adjusting the weights between features, enabling the network to more effectively leverage information across various scales. Furthermore, we have improved the feature fusion method to better integrate representations from different scales. This enhanced feature fusion allows the network to more comprehensively understand both fine details and global structures within the image. Finally, we have designed a hybrid loss function tailored to the introduced attention mechanism and feature fusion method, optimizing the network's training process and enhancing the quality of reconstruction results. Our network demonstrates significant improvements in performance for 3D human model reconstruction. Experimental results show that our method exhibits greater robustness to challenging poses compared to traditional single-scale approaches.

Keywords: enhancing multi-scale attention; single image; normal map; human reconstruction



Citation: Ren, Y.; Zhou, M.; Zhou, P.; Wang, S.; Liu, Y.; Geng, G.; Li, K.; Cao, X. Enhanced Multi-Scale Attention-Driven 3D Human Reconstruction from Single Image. *Electronics* **2024**, *13*, 4264. <https://doi.org/10.3390/electronics13214264>

Academic Editors: Kibum Kim and Huawei Tu

Received: 3 September 2024

Revised: 23 October 2024

Accepted: 24 October 2024

Published: 30 October 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the fields of computer vision and graphics, reconstructing 3D human models from a single image is both a challenging and significant task [1–5]. Compared to multi-view reconstruction, single-image reconstruction offers advantages such as ease of data acquisition, lower costs, and wide applicability. These benefits make single-image reconstruction particularly promising for various applications, providing rich possibilities for augmented reality, virtual try-on, human–computer interaction, and animation production [6–10].

Recent methods leverage implicit functions to generate the final human mesh [11,12], thus avoiding explicit mesh partitioning and more naturally capturing shapes and details. Implicit functions offer greater flexibility by producing outputs at arbitrary resolutions and densities [13]. Another key advantage is their ability to significantly reduce memory consumption while maintaining the accuracy and quality of the models. As a result, implicit functions are gaining increasing attention from researchers in the field of 3D

human reconstruction. The flexibility of implicit functions allows for the smooth handling of complex geometries and fine details, which is particularly beneficial for applications requiring high precision, such as medical imaging and detailed character modeling in video games.

Despite the development of robust representation functions, learning the detailed textures of the non-visible side from a single image remains a major challenge. Existing methods often employ generative adversarial networks to predict normal maps for the non-visible side, thereby providing rich details and clothing the bare parameterized models with intricate garments. Pifu [11] utilizes a pixel-aligned implicit function for high-resolution clothed human digitization, effectively capturing local details; however, its reliance on local information may lead to an oversight of the global structure. Similarly, while Pifuhd [13] improves upon this approach with a multi-level pixel-aligned implicit function, it still faces challenges in comprehensively capturing detailed features of the entire body. This limitation can result in the loss of important details in complex scenarios, negatively impacting the final output. Econ [14] aims to optimize the representation of clothed human bodies through normal integrals, but it mainly focuses on local feature extraction, which may lead to a lack of overall structure coherence. This limitation is particularly evident in complex scenes. Additionally, generating normal maps requires the accurate modeling of different parts of the input image. Traditional CNN networks, constrained by fixed convolution kernel sizes, struggle to capture spatial relationships across various scales and sizes, resulting in insufficient modeling capacity for spatial relationships between different scale features.

Inspired by the powerful attention mechanism [15], we propose an Enhanced Multi-Scale Attention module to improve normal map prediction for clothed human bodies. This mechanism enables the network to learn discriminative features at different scales, improving modeling for complex scenes and allowing adaptive weight adjustment based on spatial relationships. Traditional CNNs, with limited feature fusion, struggle to capture complex scene characteristics. Our approach enhances both detail and global structure capture, boosting accuracy and robustness in normal prediction. Additionally, a hybrid loss function is introduced to reduce noise and discontinuities, producing smoother and more coherent normal maps. Our key contributions are as follows:

- We propose a novel network model, Enhanced Multi-Scale Attention-Driven 3D Human Reconstruction from Single Image (EMAR), for robustly reconstructing 3D human meshes from a single image.
- To effectively capture and integrate global and local features, we introduce an enhanced multi-scale attention module that helps the network learn more discriminative feature representations at different scales, thereby improving its ability to model complex scenes.
- For effective multi-scale information fusion and interaction, we design a novel feature fusion module based on the enhanced multi-scale attention module, improving the accuracy and robustness of normal prediction.
- We introduce a hybrid loss function to supervise network training and ensure fast convergence, enhancing the network's robustness in handling complex structures and resulting in high-fidelity 3D human models.

2. Related Works

Single-image 3D human reconstruction. In the field of 3D human reconstruction from a single image, research has shown clear progression, evolving from early volumetric methods to fine-grained implicit representations. Early works, such as BodyNet [8], pioneered the use of deep learning to predict 3D human volumes directly from RGB images, though they struggled with handling occlusions. Mid-stage research focused on improving both reconstruction accuracy and robustness. For instance, Lasor [16] used synthetic occlusion-aware data and neural mesh rendering to enhance the accuracy of body pose and shape, while Cliff [17] incorporated scene information to improve performance in

complex environments. Recent studies have shifted toward ultra-fast and highly detailed reconstructions. Ultraman [18] achieved fast, fine-grained 3D reconstruction suitable for real-time applications, and Human as Points [19] introduced a method that generates precise point clouds directly from single-view images, enabling detailed geometric representation. Additionally, implicit methods like Pamir [20] and IUVD [21] explored efficient and compact implicit representations.

The latest advancements, such as Implicit Clothed Human Reconstruction [22], combine self-attention mechanisms with Signed Distance Functions (SDFs) to achieve precise and robust reconstruction of clothed human bodies, offering more realistic models that enhance virtual and augmented reality applications.

However, most existing methods primarily capture local features while overlooking global information. This imbalance makes it difficult to maintain accuracy and coherence in complex scenes, particularly when dealing with intricate clothing details and dynamic poses. Addressing these deficiencies is crucial for improving the robustness of reconstruction techniques.

Attention mechanism. In the field of 3D human reconstruction, attention mechanisms have become a pivotal tool, enhancing accuracy, optimizing processes, and improving model generalization. Wei et al. [23] pioneered temporal attention mechanisms to capture 3D human poses and shapes from monocular video, which significantly improved reconstruction accuracy and handling of temporal information. Building on this, Cho et al. [24] introduced a cross-modal attention-based method using Transformers for 3D human mesh recovery, effectively fusing image, depth map, and keypoint data and leading to greater robustness. Xue et al. [25] further refined this approach by incorporating adaptive token sampling into Transformers, enhancing both the speed and precision of mesh reconstruction. Lin et al. [26] streamlined human pose and mesh reconstruction through an end-to-end Transformer model, capturing intrinsic relationships between human pose and shape. For multi-person 3D reconstruction, Qiu et al. [27] applied a progressive video Transformer method to accurately reconstruct multiple humans in dynamic sequences.

Attention mechanisms have also been applied to clothed human reconstruction, as demonstrated by Zhang et al. [28] with the Side-View Conditioned Implicit Function (SIFU), and Li et al. [29] with the R3D-SWIN model, which uses shifted window attention for single-view 3D reconstruction. These advancements showcase the significant potential of attention mechanisms in enhancing 3D human reconstruction across various scenarios.

However, despite these advancements, many methods still face challenges in effectively integrating multi-scale features, which limits their ability to accurately model complex human shapes and movements.

3. Methodology

In this section, we present the details of our proposed EMAR. Section 3.1 describes the network overview of our EMAR. Section 3.2 explains the EMSA module. Section 3.3 shows the feature fusion module. Section 3.4 introduces our constructed hybrid loss function.

3.1. Network Overview

Our proposed EMAR, as shown in Figure 1, takes an RGB image I_{in} as input and obtains the SMPL(-X) mesh M through human pose estimation. Either the SMPL or SMPL-X mesh can be used interchangeably depending on the requirements. Based on our experiments, the choice between these two parameterized models has minimal impact on the final results. However, since this is not the primary focus of our research, we will not delve into further details on this aspect. Using Pytorch3D 0.7.1 rendering, we generate the normalized maps $N(V/IN)$ for both the visible and invisible sides of the mesh. These maps, along with the input image, are concatenated and fed into the clothing-normal prediction network. The network employs an enhanced multiscale attention module to extract features at multiple scales and a feature fusion module to combine these features, resulting in more accurate front and back clothing-normal maps $NCB(V/N)$. Finally, by combining SDF

features extracted from M , we predict the final 3D human reconstruction result. Below, we focus on detailing our innovative contributions.

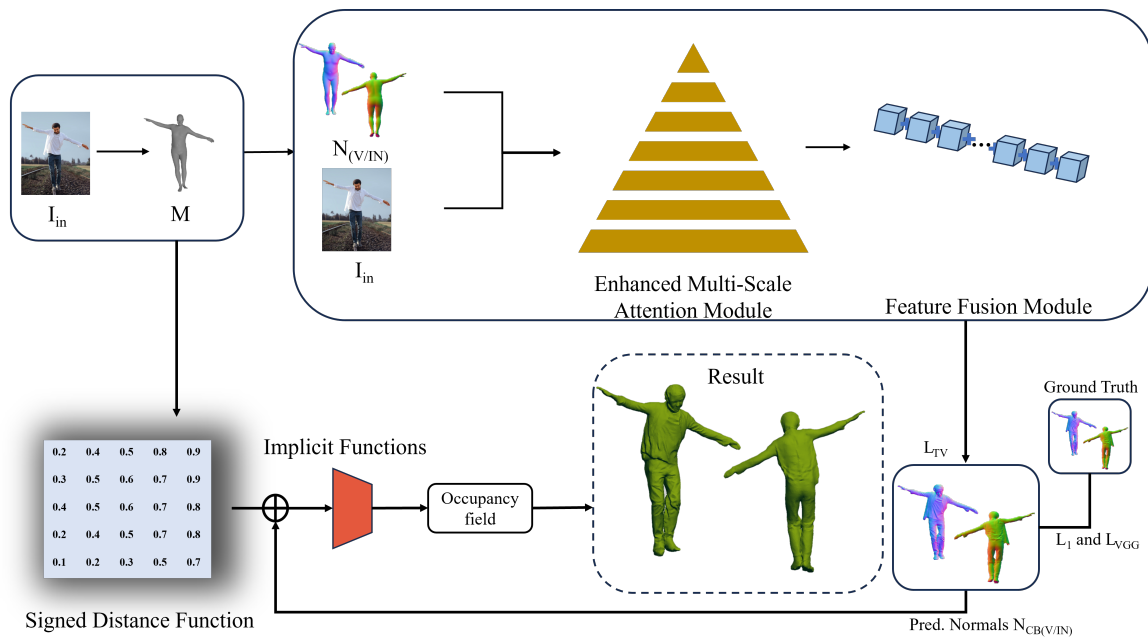


Figure 1. Overview of our proposed EMAR. Given a single-view image I_{in} and the corresponding SMPL-X mesh M , we first prepare the normal maps $N_{(V/IN)}$ and SDF features for M . Our enhanced multi-scale attention module aids the network in learning more discriminative feature representations across different scales. The proposed feature fusion module further enhances the feature representation, producing smoother and more continuous normal maps $N_{CB(V/N)}$. Finally, an Implicit Function (IF) is utilized to infer the occupancy field \hat{O} of the clothed human body.

3.2. Enhanced Multi-Scale Attention Module

In the following sections, we will provide a detailed description of the Enhanced Multi-Scale Attention Module (EMSA). As illustrated in Figure 2, the input features are denoted as $X \in \mathbb{R}^{(c \times h \times w)}$, where c refers to the number of channels, and h and w represent the spatial dimensions of the input features. To enhance the learning and representation capabilities of features, we first divide the features into g groups, with each group having c/g channels. In this paper, g is set to 3.

We map the input features of each group into four different branches. The second and third branches undergo horizontal and vertical pooling, respectively, extracting global information x_h and x_w along the corresponding directions. These features are then concatenated and processed through a 1×1 convolution to further fuse the global information in the height and width directions, forming a new feature map. Subsequently, the fused feature map is split along the height and width directions to obtain new nx_h and nx_w values. To constrain the resulting values within the range $[0, 1]$, nx_h and nx_w are transformed non-linearly through sigmoid activation functions. They are then combined with the original feature map of the first branch to inject global information along the height and width directions for feature enhancement. Normalization is performed to ensure that the mean feature value within each group is 0 and the variance is 1, further stabilizing the feature distribution. The process is outlined as follows:

$$f_1 = \text{GN}(\sigma(nx_h) \times \sigma(nx_w) \times \text{group}_x)$$

where f_1 represents the output feature map, GN represents the group normalization function, σ represents the sigmoid function, and group_x represents the grouped data. After applying the sigmoid function to nx_h and nx_w , element-wise multiplication is performed

with $group_x$, and the resulting feature map is then processed through the group normalization function. nx_h and nx_w are obtained through the following functions:

$$(nx_h, nx_w) = split(conv1 \times 1(Wavgpool(group_x) + Havgpool(group_x))) \tag{1}$$

where $split$ denotes the splitting function. $Wavgpool$ and $Havgpool$ respectively denote average pooling along the horizontal and vertical directions.

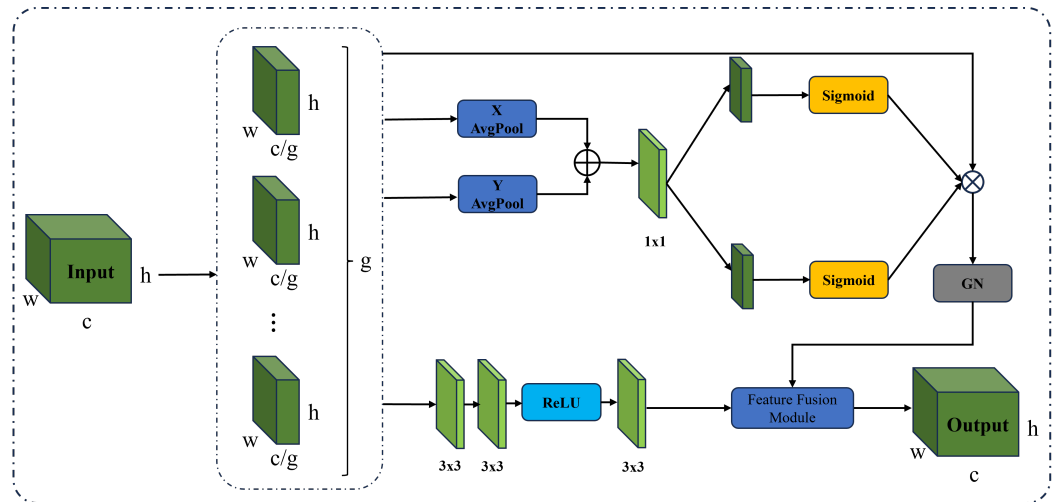


Figure 2. Enhanced Multi-Scale Attention Module.

For the fourth branch, we directly apply two 3×3 convolutions to increase the receptive field and enhance feature representation. Subsequently, the ReLU activation function is utilized to enhance the non-linear representation capability of features. Then, another 3×3 convolution is performed to further extract and enhance features. Finally, the result of the fourth branch is combined with the results of the first three branches, which have undergone group normalization, and inputted into the feature fusion module (Section 3.3) for integration, ultimately yielding the final feature map. The data processing procedure for the fourth branch is outlined as follows:

$$f_2 = conv3 \times 3(ReLU(conv3 \times 3(conv3 \times 3))) \tag{2}$$

Here, f_2 represents the result of processing in the fourth branch. The subsequent results from these two processes are fed into the feature fusion module, resulting in the final outcome:

$$f_{out} = ff(f1, f2) \tag{3}$$

where f_{out} represents the final output result, ff denotes the feature fusion module, and further details will be elaborated in Section 3.3.

3.3. Feature Fusion Module

In 3D human reconstruction tasks, models need to handle feature information from different scales and resolutions. In complex scenes, features from different scales may contain diverse crucial information. However, conventional Convolutional Neural Networks (CNNs) often rely on simple convolution and pooling operations to fuse these feature representations. This approach fails to adequately capture the complex relationships and interactions among features at different scales, leading to insufficient modeling capabilities for complex scenes. To address these limitations, we propose a Feature Fusion Module (FFM) aimed at enhancing the model’s integration capability of features at different scales. This module achieves better capture of details and global structures in complex scenes through fusion of multi-scale features and adaptive weight adjustments.

Figure 3 illustrates our proposed Feature Fusion Module, which efficiently integrates multi-scale features through a series of refined operations, thereby enhancing the model’s adaptability in handling complex scenes. Specifically, this module receives feature information from two input feature maps (Input1 and Input2) and performs a series of processing operations on these two feature maps. Firstly, adaptive average pooling (AvgPool) is applied to each input feature map separately, compressing the global information of the feature maps into 1×1 feature maps to extract global features. Then, the pooled feature maps are reshaped to match subsequent matrix operations.

$$\text{Reshape}(\text{AvgPool}(\text{Input})) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \text{Input}(i, j) \tag{4}$$

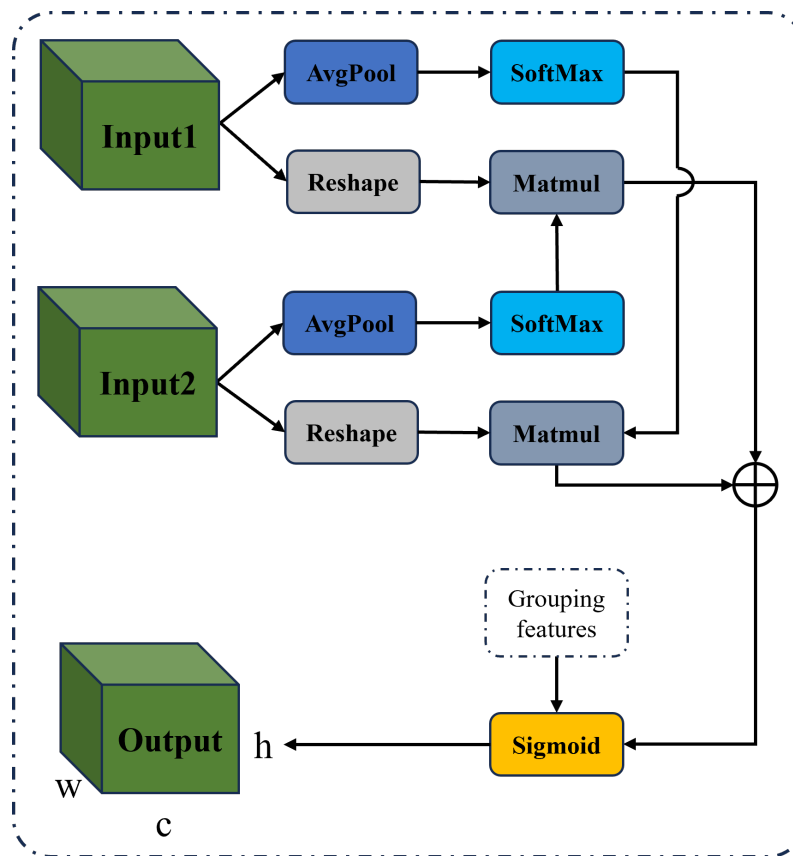


Figure 3. Feature Fusion Module.

Then, the reshaped feature maps are subjected to the Softmax function to generate normalized weight matrices. These weight matrices reflect the importance of the feature maps at a global scale. Subsequently, through matrix multiplication (Matmul), these weight matrices are multiplied by the corresponding feature matrices to compute the fused weights. This step ensures that information from the feature maps at a local scale is preserved, and by weighting the fusion of global and local features, the expressive power of the features is enhanced.

$$f = \text{Softmax}(\text{AvgPool}(\text{Input})) \tag{5}$$

$$\text{Weights} = \text{Matmul}((f_1), \text{Reshape}(\text{Input2})) + \text{Matmul}((f_2), \text{Reshape}(\text{Input1})) \tag{6}$$

Here, Input denotes two inputs, Input1 and Input2, where f_1 represents the processed value of Input1, and f_2 represents the processed value of Input2.

After generating the fusion weights, the initial feature maps are combined with the fused weights through element-wise multiplication. This step enables the feature maps to

adaptively adjust the weights of different features, thereby better integrating multi-scale feature information. Finally, after passing through the Sigmoid activation function, the final output feature map (Output) is generated. This feature fusion module significantly enhances the model's capability to capture both fine details and global structures, improving its adaptability in complex scenes. It also boosts the accuracy and robustness of normal prediction, effectively addressing the issue of information loss commonly seen in traditional convolutional neural networks when handling multi-scale features.

$$\text{Output} = \sigma(\text{sum}(\text{Weights}_1, \text{Weights}_2) + \text{group}_x) \quad (7)$$

3.4. Hybrid Loss Function

In normal map prediction, high-frequency noise and irregular edges can degrade the results. To mitigate this, we introduce a mixed loss function consisting of L1 loss, VGG perceptual loss, and Total Variation (TV) regularization. Each loss function has its unique strengths and weaknesses. The L1 loss is simple and effective, directly measuring the absolute error between the predicted and target values, thereby improving geometric accuracy. However, it fails to capture high-level semantic features and may overlook fine details. The VGG perceptual loss, by comparing feature space representations using a pre-trained VGG network, enhances the visual quality of the reconstruction, especially in terms of details and textures, but it comes with higher computational cost and emphasizes visual perception rather than geometric accuracy. The Total Variation (TV) regularization helps to suppress noise and maintain the smoothness of the normal map, reducing artifacts, but it may lead to excessive smoothing, potentially diminishing fine details in the image. By combining these loss functions, our hybrid loss achieves a balance between geometric accuracy, visual quality, and noise reduction. In the actual implementation, these losses are calculated separately for the visible and invisible parts of the normal map and then combined into the total loss. Through extensive experiments, we have validated that this hybrid loss function design achieves a good balance between accuracy and visual quality. The specific experimental results show that, compared to individual loss functions, our hybrid loss significantly improves reconstruction quality (see Section 4.2 for the results and analysis of the loss function ablation study).

$$L_1 = \|\text{pred} - \text{target}\|_1 \quad (8)$$

$$L_{VGG} = \|\phi(\text{pred}) - \phi(\text{target})\|_2 \quad (9)$$

$$L_{TV} = \sum_{(i,j)} ((\text{pred}_{i+1,j} - \text{pred}_{i,j})^2 + (\text{pred}_{i,j+1} - \text{pred}_{i,j})^2) \quad (10)$$

Here, pred represents the predicted normal map. Additionally, in TV regularization, there is usually a regularization weight λ to control the influence of the regularization term on the total loss, which typically ranges from positive real numbers, commonly between 0.01 and 10. In this paper, $\lambda = 0.1$. Next, we calculate the losses separately for the visible and invisible sides of the normal map:

$$L_{\text{total_V}} = 5.0 \times L_{L1_V} + L_{VGG_V} + \lambda_{\text{reg}} \times L_{TV_V} \quad (11)$$

$$L_{\text{total_IN}} = 5.0 \times L_{L1_IN} + L_{VGG_IN} + \lambda_{\text{reg}} \times L_{TV_IN} \quad (12)$$

Finally, the total losses of the frontal and back normals are combined to form the final total loss:

$$L_{\text{final}} = [L_{\text{total_V}}, L_{\text{total_IN}}] \quad (13)$$

4. Experiments

In this section, we verify the feasibility of the algorithm through extensive experiments. Section 4.1 introduces the dataset used, the experimental settings, and the three metrics adopted. Section 4.2 outlines the comparison results with state-of-the-art methods.

Section 4.3 provides an ablation study and related discussions. Section 4.4 introduces the failure cases and analysis.

4.1. Datasets and Implementation

The Thuman2.0 [30], CAPE [31] and RenderPeople [32] datasets are widely used in computer vision, each offering distinct features. Thuman2.0 focuses on human pose estimation, providing diverse samples that capture how different poses affect body morphology, essential for accurate 3D human reconstruction. CAPE, in contrast, includes images and models of clothed humans, where clothing occlusion and shape greatly influence body appearance. These factors are critical in training models, making both datasets ideal for our study. The RenderPeople dataset complements these by offering high-quality 3D scans of clothed humans in a variety of poses and clothing styles, further enriching the diversity of samples for testing model generalization. These factors make all three datasets ideal for our study.

Our network was trained separately on the Thuman2.0 and CAPE datasets. For the Thuman2.0 dataset, we split it into a training set consisting of 500 models, a validation set containing 21 models, and a test set with 5 models. For the CAPE dataset, we used the entire dataset for training. We then evaluated our method on the CAPE, Thuman2.0, and RenderPeople datasets.

To better analyze the network's generalization ability on complex poses, we further divided the CAPE dataset into "CAPE-Hard" and "CAPE-Easy" categories based on the difficulty of the poses at test time. The "CAPE-Hard" category contains 100 models, which include more challenging poses, while the "CAPE-Easy" category consists of 50 models with relatively simpler poses. This categorization allowed us to test the generalization ability of the network across poses of varying difficulty.

Additionally, the RenderPeople dataset includes highly detailed scans of 10 clothed human subjects in natural poses. The diversity of clothing and poses in this dataset allowed us to further evaluate the network's performance under real-world conditions. This additional evaluation helps provide a more comprehensive analysis of the method's generalization ability.

The proposed method was implemented on a desktop computer with an Intel(R) Core(TM) i9-10980XE CPU @ 3.00 GHz (Intel Corporation, Santa Clara, CA, USA), NVIDIA RTX A6000 GPU (NVIDIA Corporation, Santa Clara, CA, USA), and 128 GB of memory (Kingston Technology Corporation, Fountain Valley, CA, USA). The EMAR model, as proposed in this paper, supports end-to-end training and utilized the Adam optimizer with an initial learning rate of 1×10^{-4} , a batch size of 4, and was trained for 20 epochs.

Considering the characteristics of point clouds, we opted to use three specific loss functions: the point-to-surface distance (p2s), Chamfer Distance, and Normal Difference. These loss functions are particularly well suited for our task, as they directly measure the geometric accuracy and surface smoothness of the 3D models. The p2s loss ensures that the generated point cloud closely aligns with the target surface, Chamfer Distance minimizes the overall discrepancy between two point clouds, and Normal Difference helps maintain consistent and accurate surface normals, thereby enhancing the fidelity of the reconstructed 3D geometry.

4.2. Comparison Experiments

Qualitative Experiments. As shown in Figure 4, the qualitative experimental results demonstrate that our method outperformed PIFu [11], PaMIR [20], and 2K2K [33]. PIFu leverages pixel-aligned features to infer 3D geometry from 2D images, but its feature representation is often insufficient for capturing fine details in complex regions. PaMIR improves upon PIFu by introducing poseawareness, yet it struggles to generalize well to unseen poses and complex geometric variations. 2K2K, while utilizing a hierarchical approach, faces challenges in accurately estimating depth information from low-resolution images, leading to less precise surface reconstructions.

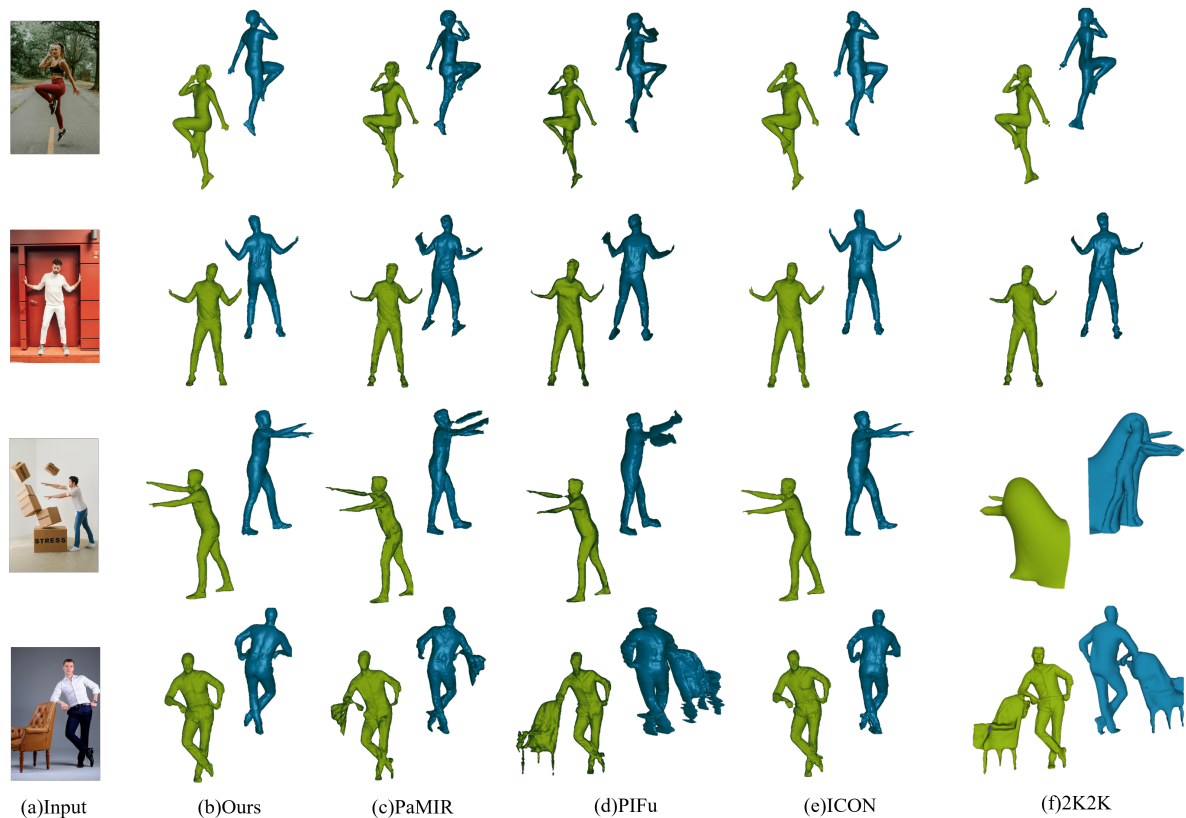


Figure 4. Qualitative experiments on in-the-wild photos, where green is the front of the model and blue is the back of the model: Column (a) shows the input images, column (b) presents the results of EMAR, column (c) shows the results of PaMIR, column (d) presents the results of PIFu, column (e) shows the results of ICON, and column (f) shows the results of 2K2K.

In contrast, our improvement is attributed to the incorporation of the Enhanced Multi-Scale Attention Module, which enables our model to focus on more discriminative features across multiple scales. This helps in capturing finer details and better handling complex geometric structures that PIFu, PaMIR, and 2K2K struggle with. For example, the 2K2K method encounters difficulties in depth estimation due to the lower resolution of input images, which results in less clear representations of the human form.

Compared to ICON [5], our method shows significant improvements in detail accuracy, as illustrated in Figure 5. ICON primarily relies on implicit surface reconstruction and can generate reasonably detailed 3D human models. However, its performance is limited by the lack of explicit surface normal prediction, which can result in less accurate surface details. In contrast, our method predicts more accurate normal maps, enabling the model to capture finer surface details, leading to superior performance in both qualitative and quantitative measures. This richer surface detail information enhances our model's ability to reconstruct more complex human forms.

Table 1 compares the time complexity of various 3D human reconstruction methods. PIFu exhibits a quadratic complexity of $O(n^2)$, making it computationally more demanding as input size increases. In contrast, Pamir, ICON, 2K2K, and EMAR all have a linear complexities of $O(n)$, indicating that they are more efficient and scalable for larger datasets. Overall, newer methods showed improved computational efficiency while maintaining performance.

Quantitative Experiments. In Table 2, we present the quantitative results on the CAPE, Thuman2.0, and RenderPeople datasets, with the CAPE column being the weighted average of the first two columns. The results of our method are shown in the last two rows: the second-to-last row (denoted as Ours (T)) reflects the performance of our model trained

on the Thuman2.0 dataset, while the last row (denoted as Ours (C)) shows the performance when trained on the CAPE dataset.

Table 1. Comparison of time complexity of various methods.

Methods	Time Complexity
PIFu	$O(n^2)$
PaMIR	$O(n)$
ICON	$O(n)$
2K2K	$O(n)$
EMAR	$O(n)$

Table 2. Quantitative results on the CAPE, Thuman2.0, and RenderPeople datasets. ↓ indicates lower values are better. Bold values represent the best performance.

Dataset	CAPE-Easy			CAPE-Hard			CAPE			Thuman2.0			RenderPeople		
	Chamfer↓	P2S↓	Normals↓	Chamfer↓	P2S↓	Normals↓	Chamfer↓	P2S↓	Normals↓	Chamfer↓	P2S↓	Normals↓	Chamfer↓	P2S↓	Normals↓
PIFu (2019)	2.823	2.796	0.100	4.029	4.195	0.124	3.627	3.729	0.116	3.024	2.297	0.201	2.103	1.452	0.191
PaMIR (2021)	1.936	1.263	0.078	2.216	1.611	0.093	2.123	1.495	0.088	1.730	1.330	0.118	1.196	0.984	0.124
ICON (2022)	1.233	1.170	0.072	1.096	1.013	0.063	1.142	1.065	0.066	1.013	1.050	0.082	0.735	0.831	0.080
2K2K (2023)	1.264	1.213	0.070	1.437	1.385	0.060	1.379	1.328	0.063	1.651	1.247	0.079	1.032	0.932	0.077
Ours (T)	0.802	0.768	0.050	0.884	0.861	0.053	0.857	0.830	0.052	0.986	1.024	0.076	0.735	0.797	0.051
Ours (C)	-	-	-	-	-	-	-	-	-	1.097	1.192	0.081	0.906	0.839	0.069

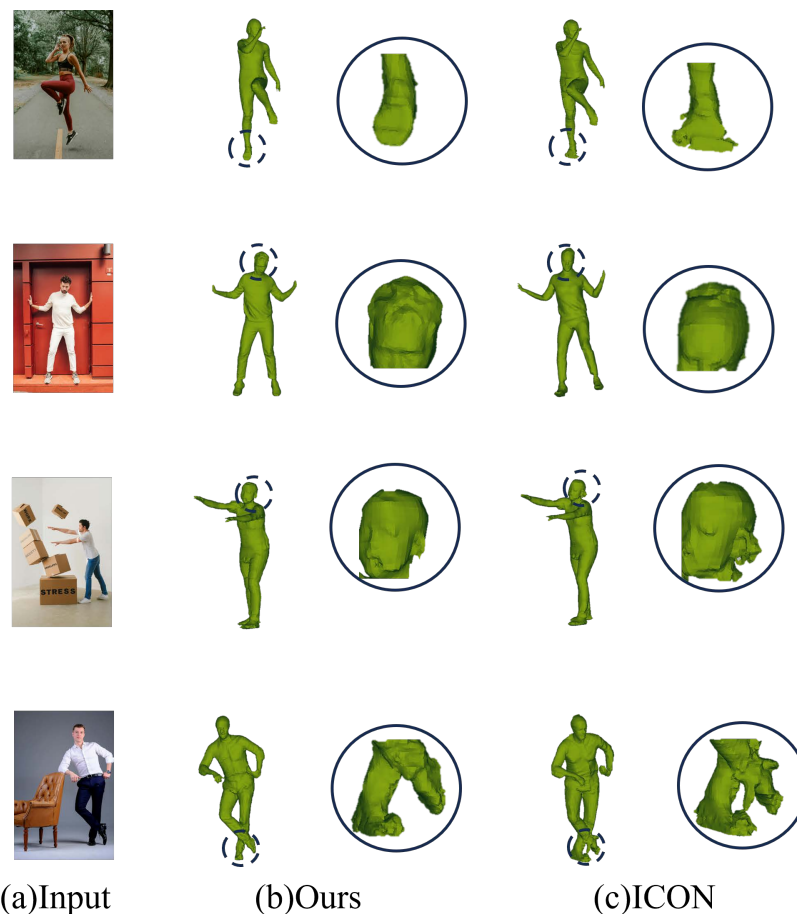


Figure 5. Comparison of Details with ICON.

When trained on Thuman2.0, our method achieved the best performance across almost all datasets and metrics, particularly excelling in the more challenging “CAPE-Hard” subset and demonstrating its strong generalization ability to various difficult poses.

However, when trained on the CAPE dataset, while the performance remained strong, there was a slight decrease in the Thuman2.0 and RenderPeople evaluations. This suggests that although the CAPE dataset includes a variety of clothing and poses, it may not fully capture the variability required for optimal generalization to other datasets. Nevertheless, our method still achieved competitive results, reaffirming its robustness across different datasets and conditions.

4.3. Ablation Experiments

In this section, we compare and discuss the roles of the proposed modules, including the EMSA module and the Feature Fusion module.

To evaluate the effects of each component, we conducted ablation experiments using ICON as the baseline model. We implemented the Feature Fusion module through channel concatenation, which is the most common method. Subsequently, we separately added the EMSA module and the Feature Fusion module to complete the experimental settings. Table 3 presents the numerical results of different combinations of these modules. With the addition of various key components, the performance of the network gradually improved, with our model achieving the best performance. Following this was the baseline model with the EMSA module, while the baseline model performed the worst among all indicators. This further confirms the necessity and effectiveness of our proposed approach.

Table 3. Ablation experiments on the CAPE dataset. ↓ indicates lower values are better.

Baseline	EMSA	FFM	Chamfer↓	P2S↓	Normals↓
✓			1.142	1.065	0.066
✓	✓		0.998	0.943	0.055
✓		✓	1.041	0.973	0.060
✓	✓	✓	0.857	0.830	0.052

The ablation study shows how different loss functions impact model performance, as shown in Table 4. Using only L1 loss improved geometric accuracy but failed in detail recovery, resulting in higher Normals values. Adding VGG loss unexpectedly worsened all metrics, indicating that it did not effectively capture necessary features in this context. Similarly, using only Total Variation loss negatively affected both geometric accuracy and detail preservation across all metrics. Combining L1 and VGG losses achieved a balance between shape fidelity and detail, resulting in the lowest values for Chamfer and P2S, with Normals values being only 0.003 higher than the optimal performance. However, it is noteworthy that changes in the Chamfer and P2S metrics, compared to those obtained with all three loss functions, were 2.02% and 1.22%, respectively, while the change in the Normals metric reached 5.45%. This discrepancy arises because the Normals values are inherently lower than those of the other two metrics, causing what may seem like a minor change to have a significantly different impact. After comprehensive consideration, we decided to use three loss functions together.

Table 4. Ablation experiment of loss function. ↓ indicates lower values are better.

Baseline	L ₁	L _{VGG}	L _{TV}	Chamfer↓	P2S↓	Normals↓
✓	✓			0.855	0.835	0.065
✓		✓		0.860	0.840	0.080
✓			✓	0.865	0.845	0.090
✓	✓	✓		0.840	0.820	0.055
✓	✓		✓	0.845	0.825	0.060
✓		✓	✓	0.855	0.830	0.075
✓	✓	✓	✓	0.857	0.830	0.052

4.4. Failure Cases and Analysis

As shown above, our proposed EMAR can effectively reconstruct the 3D human mesh from a single image. However, our model still faces some challenging issues. As shown in Figure 6, when the body undergoes self-intersection, the predicted human mesh by the network often exhibits various faults. In Figure 6A, the target person's hands are clasped together, but the predicted human mesh shows the hands crossing fingers. In Figure 6B, the target person is standing with one leg bent and the other leg stretched forward while trying to reach forward to grab the heel with both hands, but the predicted human mesh is in a squatting position. This is due to inaccurate human pose estimation, which affects the final reconstruction results. Precise pose estimation will be the direction of our next work.



Figure 6. Failure cases.

5. Conclusions

This paper presents an enhanced single-image 3D human body reconstruction method driven by multi-scale attention. To address the issue of inability to utilize feature correlations across different scales, we propose an Enhanced Multi-Scale Attention (EMSA) module, which helps the network learn more distinctive feature representations at various scales, thereby improving robustness to various human poses. To tackle the problem of ineffective integration of information at different scales, leading to insufficient modeling capabilities for complex scenes, we designed a Feature Fusion Module (FFM) to enhance the model's integration capability of features at different scales. Additionally, we introduced a loss function more suitable for normal map prediction, enabling the network to generate smoother normal maps. Finally, we conducted comparative experiments and ablation studies, and the results demonstrate that our method surpasses most state-of-the-art approaches. Code has been published on github (<https://github.com/R33333/EMAR> (accessed on 2 September 2024)).

Author Contributions: Conceptualization, Y.R. and M.Z.; methodology, Y.R., M.Z. and P.Z.; validation, Y.R., P.Z. and S.W.; writing—original draft, Y.R., S.W. and Y.L.; writing—review and editing, Y.L. and G.G.; funding acquisition, K.L. and X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China (62271393), the Key Laboratory Project of the Ministry of Culture and Tourism (1222000812, crrt2021K01), the Science and Technology Plan Project of Xi'an City (2024JH-CXSF-0014), the Key Research and Development Program of Shaanxi Province (2024SF-YBXM-681, 2021ZDLSF06-04).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable. The study used publicly available datasets, where informed consent was previously obtained by the dataset providers.

Data Availability Statement: The data used in this article are all public data sets, and there are no innovative data. Public datasets can be downloaded from relevant references.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, J.; Yoon, J.S.; Wang, T.Y.; Singh, K.K.; Neumann, U. Complete 3D Human Reconstruction from a Single Incomplete Image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 8748–8758.
2. Lochner, J.; Gain, J.; Perche, S.; Peytavie, A.; Galin, E.; Guérin, E. Interactive Authoring of Terrain using Diffusion Models. *Comput. Graph. Forum* **2023**, *42*, e14941. [[CrossRef](#)]
3. Zhu, H.; Cao, Y.; Jin, H.; Chen, W.; Du, D.; Wang, Z.; Cui, S.; Han, X. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16; Springer: Cham, Switzerland, 2020; pp. 512–530.
4. Chen, H.; Huang, Y.; Huang, H.; Ge, X.; Shao, D. GaussianVTON: 3D Human Virtual Try-ON via Multi-Stage Gaussian Splatting Editing with Image Prompting. *arXiv* **2024**, arXiv:2405.07472.
5. Xiu, Y.; Yang, J.; Tzionas, D.; Black, M.J. Icon: Implicit clothed humans obtained from normals. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 13286–13296.
6. Ma, X.; Zhao, J.; Teng, Y.; Yao, L. Multi-Level Implicit Function for Detailed Human Reconstruction by Relaxing SMPL Constraints. *Comput. Graph. Forum* **2023**, *42*, e14951. [[CrossRef](#)]
7. Ren, Y.; Zhou, M.; Wang, Y.; Feng, L.; Zhu, Q.; Li, K.; Geng, G. Implicit 3D Human Reconstruction Guided by Parametric Models and Normal Maps. *J. Imaging* **2024**, *10*, 133. [[CrossRef](#)] [[PubMed](#)]
8. Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; Schmid, C. BodyNet: Volumetric inference of 3D human body shapes. In Proceedings of the ECCV 2018, Munich, Germany, 8–14 September 2018.
9. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10975–10985.
10. Xu, W.; Chatterjee, A.; Zollhöfer, M.; Rhodin, H.; Mehta, D.; Seidel, H.P.; Theobalt, C. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph. (ToG)* **2018**, *37*, 1–15. [[CrossRef](#)]
11. Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; Li, H. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2304–2314.
12. Muttagi, S.I.; Patil, V.; Babar, P.P.; Chunamari, R.; Kulkarni, U.; Chikkamath, S.; Meena, S. 3D Avatar Reconstruction Using Multi-level Pixel-Aligned Implicit Function. In Proceedings of the International Conference on Recent Trends in Machine Learning, IOT, Smart Cities & Applications, Hyderabad, India, 16–17 September 2023; Springer: Cham, Switzerland, 2023; pp. 221–231.
13. Saito, S.; Simon, T.; Saragih, J.; Joo, H. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 84–93.
14. Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; Black, M.J. Econ: Explicit clothed humans optimized via normal integration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 512–523.
15. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
16. Yang, K.; Gu, R.; Wang, M.; Toyoura, M.; Xu, G. Lasor: Learning accurate 3d human pose and shape via synthetic occlusion-aware data and neural mesh rendering. *IEEE Trans. Image Process.* **2022**, *31*, 1938–1948. [[CrossRef](#)] [[PubMed](#)]
17. Li, Z.; Liu, J.; Zhang, Z.; Xu, S.; Yan, Y. Cliff: Carrying location information in full frames into human pose and shape estimation. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part V; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13695, pp. 590–606.
18. Chen, M.; Chen, J.; Ye, X.; Gao, H.a.; Chen, X.; Fan, Z.; Zhao, H. Ultraman: Single Image 3D Human Reconstruction with Ultra Speed and Detail. *arXiv* **2024**, arXiv:2403.12028.
19. Tang, Y.; Zhang, Q.; Hou, J.; Liu, Y. Human as Points: Explicit Point-based 3D Human Reconstruction from Single-view RGB Images. *arXiv* **2023**, arXiv:2311.02892.
20. Zheng, Z.; Yu, T.; Liu, Y.; Dai, Q. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3170–3184. [[CrossRef](#)] [[PubMed](#)]
21. Li, B.; Deng, Y.; Yang, Y.; Zhao, X. An Embeddable Implicit IUVD Representation for Part-Based 3D Human Surface Reconstruction. *IEEE Trans. Image Process.* **2024**, *33*, 4334–4347. [[CrossRef](#)] [[PubMed](#)]
22. Yao, L.; Gao, A.; Wan, Y. Implicit Clothed Human Reconstruction Based on Self-attention and SDF. In Proceedings of the International Conference on Neural Information Processing, Changsha, China, 20–23 November 2023; Springer: Cham, Switzerland, 2023; pp. 313–324.
23. Wei, W.L.; Lin, J.C.; Liu, T.L.; Liao, H.Y.M. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13211–13220.

24. Cho, J.; Kim, Y.; Oh, T.H. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part I; Lecture Notes in Computer Science Springer: Cham, Switzerland, 2022; Volume 13684, pp. 342–359.
25. Xue, Y.; Chen, J.; Zhang, Y.; Yu, C.; Ma, H.; Ma, H. 3d human mesh reconstruction by learning to sample joint adaptive tokens for transformers. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 6765–6773.
26. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1954–1963.
27. Qiu, Z.; Yang, Q.; Wang, J.; Feng, H.; Han, J.; Ding, E.; Xu, C.; Fu, D.; Wang, J. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21254–21263.
28. Zhang, Z.; Yang, Z.; Yang, Y. SIFU: Side-view Conditioned Implicit Function for Real-world Usable Clothed Human Reconstruction. *arXiv* **2023**, arXiv:2312.06704.
29. Li, C.; Xiao, M.; Gao, M. R3D-SWIN: Use Shifted Window Attention for Single-View 3D Reconstruction. *arXiv* **2023**, arXiv:2312.02725.
30. Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; Liu, Y. Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5746–5756.
31. Ma, Q.; Yang, J.; Ranjan, A.; Pujades, S.; Pons-Moll, G.; Tang, S.; Black, M.J. Learning to dress 3D people in generative clothing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6469–6478.
32. Renderpeople. Available online: <https://renderpeople.com/> (accessed on 19 October 2024) .
33. Han, S.H.; Park, M.G.; Yoon, J.H.; Kang, J.M.; Park, Y.J.; Jeon, H.G. High-fidelity 3d human digitization from single 2k resolution images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12869–12879.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.