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

Yong Ren , Mingquan Zhou * , Pengbo Zhou , Shibo Wang , Yangyang Liu , Guohua Geng , Kang Li * , Xin Cao *

Posted Date: 11 September 2024

doi: 10.20944/preprints202409.0871.v1

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

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
[2] National and Local Joint Engineering Research Center for Cultural Heritage Digitization, Xi'an 710127, China
[2] School of Art and Media, Beijing Normal University, Beijing 100875, China
* Correspondence: mquanzhou@126.com (M.Z.); likang@nwu.edu.cn (K.L.); xin_cao@163.com (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

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 [1,2]. 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 [3,4]. For instance, in augmented reality, accurate 3D human models enable more immersive experiences by seamlessly integrating virtual objects with the real world [5]. Virtual try-on applications benefit from single-image reconstruction by allowing users to visualize how clothes fit without physically trying them on, enhancing online shopping experiences [6]. In human-computer interaction, realistic 3D models improve gesture recognition and enhance interactive applications [7,8], while in animation production, single-image reconstruction simplifies the creation of lifelike characters, reducing the need for extensive motion capture setups [9,10].

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

 doi:10.20944/preprints202409.0871.v1

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 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 [11,14]. However, these methods typically capture only local features, neglecting global and multi-scale information. This limitation becomes evident in complex scenes, particularly when dealing with intricate human details and structures [15]. Additionally, generating normal maps requires 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[16], we introduce an enhanced multi-scale attention module, starting from predicting the normal map of clothed human bodies. This mechanism aids the network in learning more discriminative feature representations at different scales, thereby enhancing the network's modeling capability for complex scenes. Moreover, it allows the network to adaptively adjust weights based on feature information from different positions, facilitating better capture of spatial relationships between different positions in the image. This is beneficial for improving the accuracy of normal prediction and enhancing robustness to human body poses.

Furthermore, traditional CNN networks often rely on simple convolution and pooling operations to fuse features across different scales, limiting effective information fusion and interaction. This limitation hampers the network's ability to fully utilize the correlation between different scale features when learning complex scene characteristics. To address this, we enhance the model's capability to capture details and global structures in complex scenes by leveraging features at different scales and employing adaptive weights. This enhancement leads to improved accuracy and robustness in normal prediction.

To address issues of discontinuity and noise in the normal maps generated during reconstruction, we developed a hybrid loss function. This function reduces differences between adjacent pixels to suppress noise and discontinuities, resulting in more coherent and smooth normal maps. Our 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. With the progression of deep learning, novel methodologies for reconstructing 3D human bodies from single images have made significant strides in precision and efficiency. [8] leverages supervised learning to train convolutional neural networks for direct

 doi:10.20944/preprints202409.0871.v1

volumetric prediction of the human body from a single image. [17] improve model robustness against complex real-world scenes using synthetic data and neural mesh rendering techniques, ensuring accurate 3D human posture and shape reconstruction. Lastly, [18] includes scene location information within complete frames to better comprehend the spatial position and posture of the human body, thereby enhancing the accuracy and robustness of single-view human body reconstruction. [19] combines conditional implicit functions and efficient voxel networks to accurately reconstruct 3D human models from a single image in complex scenes. [20] introduces an efficient network architecture that significantly enhances computational speed while maintaining detail fidelity, making it suitable for real-time applications. [21] proposes an explicit point-based 3D human reconstruction method that uses a novel network to generate high-resolution 3D point clouds from single-view RGB images, improving precision by accurately capturing surface details. [22] presents a new implicit representation method (IUVD) that decomposes the human body into multiple parts for separate processing, thereby enhancing reconstruction accuracy and efficiency. [23] proposes a tri-directional implicit function for high-fidelity 3D character reconstruction. By modeling from multiple directions, it improves detail level and accuracy in the reconstruction process. [24] introduces an implicit reconstruction method based on self-attention mechanisms and Signed Distance Functions (SDF). The self-attention mechanism enhances the model's ability to capture complex geometric details, while SDF provides high-precision surface representation. This combination is particularly effective in reconstructing clothed human models, significantly improving reconstruction accuracy.

Attention mechanism. Attention mechanisms enhance the focus of neural networks on specific regions. To address partial occlusion issues, [25] employs motion continuity attention to capture temporal coherence, integrating features from temporally adjacent representations more effectively through a hierarchical attention feature mechanism. [26] uses a self-attention mechanism based on non-adjacent vertices of the body's triangular mesh topology to minimize memory overhead and accelerate inference speed. [27] proposed a learnable sampling module for human mesh recovery to reduce inherent depth ambiguity, aggregating global information through joint adaptive markers and leveraging non-local information in the input image. [28] is a mesh transformer for end-to-end human mesh recovery, where the encoder captures interactions between vertices and joints, and the decoder outputs 3D joint coordinates and mesh structure. [29] introduced an end-to-end Transformer-based method for multi-person human mesh recovery in videos, extracting global features with a spatio-temporal encoder and predicting human posture and mesh using spatio-temporal pose and shape decoders. [30] utilizes the cross-attention mechanism in the transformer architecture to effectively decouple side-view features in the 2D-to-3D feature mapping process, significantly improving the accuracy and robustness of 3D models, especially in cases where SMPL-X estimation is imperfect. [31] applies the shifted window attention mechanism to better capture spatial relationships and details, enhancing single-view 3D reconstruction and demonstrating significant performance improvements. [32] employs attention mechanisms to enhance the generalization capability of Neural Radiance Fields, allowing high-quality 3D human models to be generated from sparse inputs, including single images.

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, which can be replaced as needed. Using Pytorch3D rendering, the normalized maps $N_{(V/IN)}$ of the visible and invisible sides of the parameterized mesh are obtained. These two normal maps, along with the input image, are concatenated as the input to the clothing-normal prediction network. The network utilizes an enhanced multiscale attention module

 doi:10.20944/preprints202409.0871.v1

to extract multiscale features and a feature fusion module to merge features of different scales, thus predicting more accurate front and back clothing-normal maps $N_{CB(V/N)}$. Finally, combining SDF features extracted from M predicts the final 3D human reconstruction result. Below, we will focus on describing our innovative part.

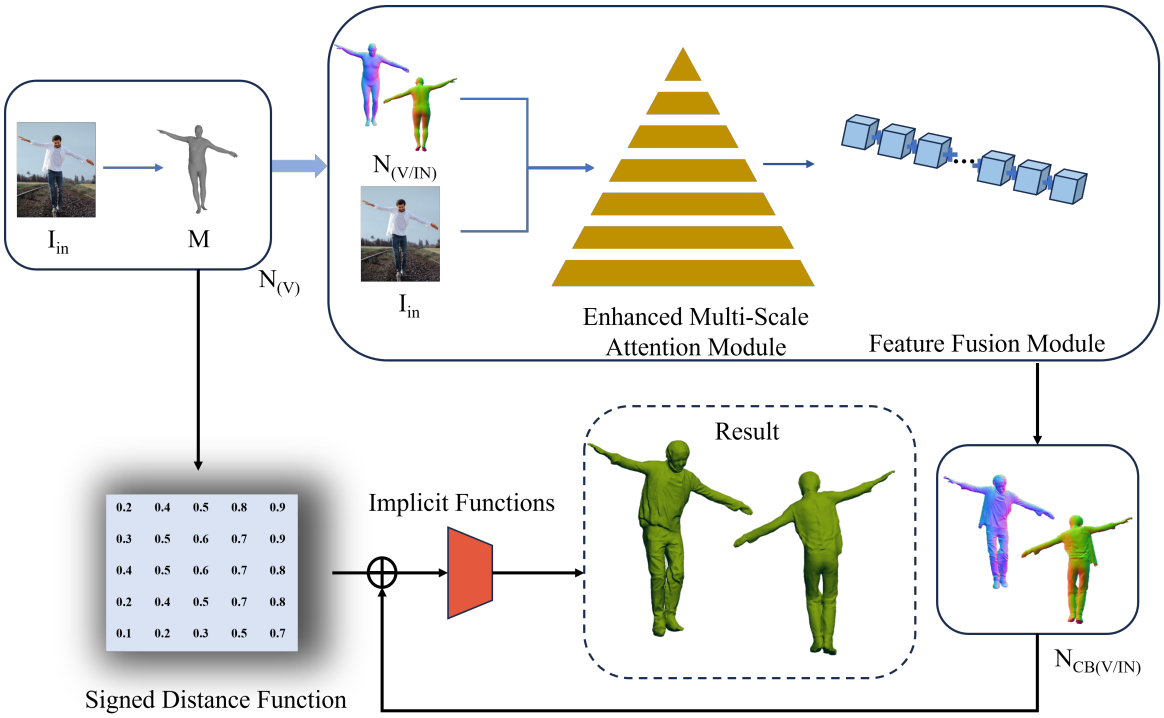

Figure 1. Overview of our proposed EMAR. Given a single-view image I_{in} and the corresponding SMPL-X mesh M, we first prepare the normal maps $N_{(V/IN)}$ and SDF features for M. Our enhanced multi-scale attention module aids the network in learning more discriminative feature representations across 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. 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:

$$f_1 = GN(œ(nx_h) \times œ(nx_w) \times group_x)$$

 doi:10.20944/preprints202409.0871.v1

Where f_1 represents the output feature map, GN represents the group normalization function, œ 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. Where nx_h and nx_w are obtained through the following functions:

$$(nx_h, nx_w) = \text{split}(conv1 \times 1(Xavgpool(group_x) + Yavgpool(group_x))) \tag{1}$$

Where split denotes the splitting function, which splits the feature map after convolution into nx_h and nx_w,Conv1 $\times$ 1 represents a convolutional neural network with a 1 $\times$ 1 kernel size. Xavgpool and Yavgpool respectively denote average pooling along the horizontal and vertical directions.

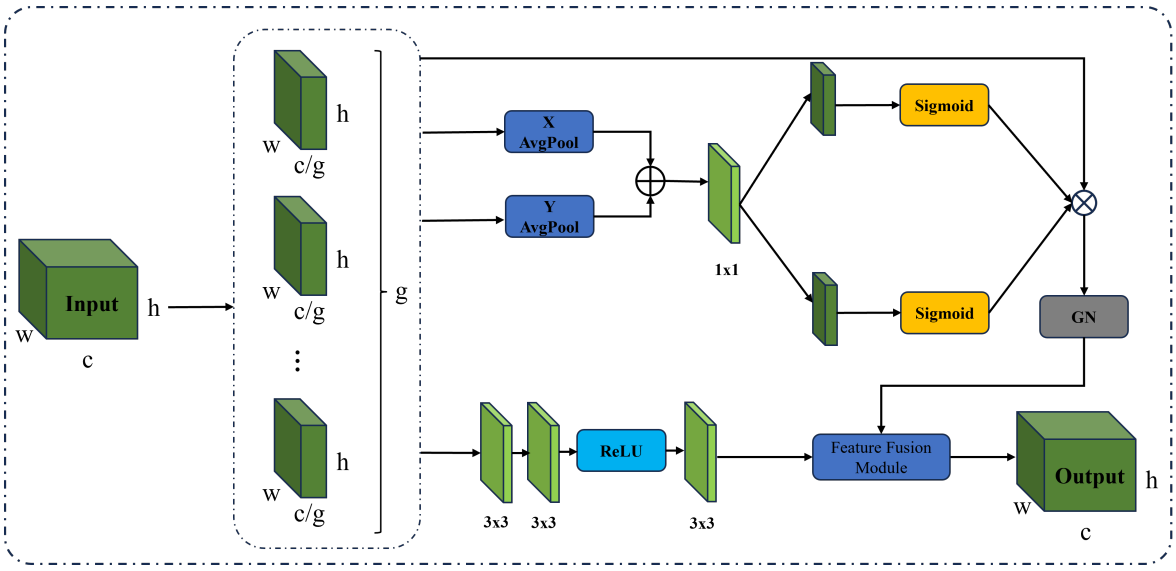

Figure 2. Enhanced multi-scale attention module.

For the fourth branch, we directly apply two 3 $\times$ 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 $\times$ 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 (Sec. 3.3) for integration, ultimately yielding the final feature map. The data processing procedure for the fourth branch is outlined as follows:

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

Here, f_2 represents the result of processing in the fourth branch, ReLU denotes the ReLU activation function, and conv3 $\times$ 3 represents a convolutional neural network with a 3 $\times$ 3 kernel size. 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 3C.

 doi:10.20944/preprints202409.0871.v1

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 1x1 feature maps to extract global features. Then, the pooled feature maps are reshaped to match subsequent matrix operations.

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

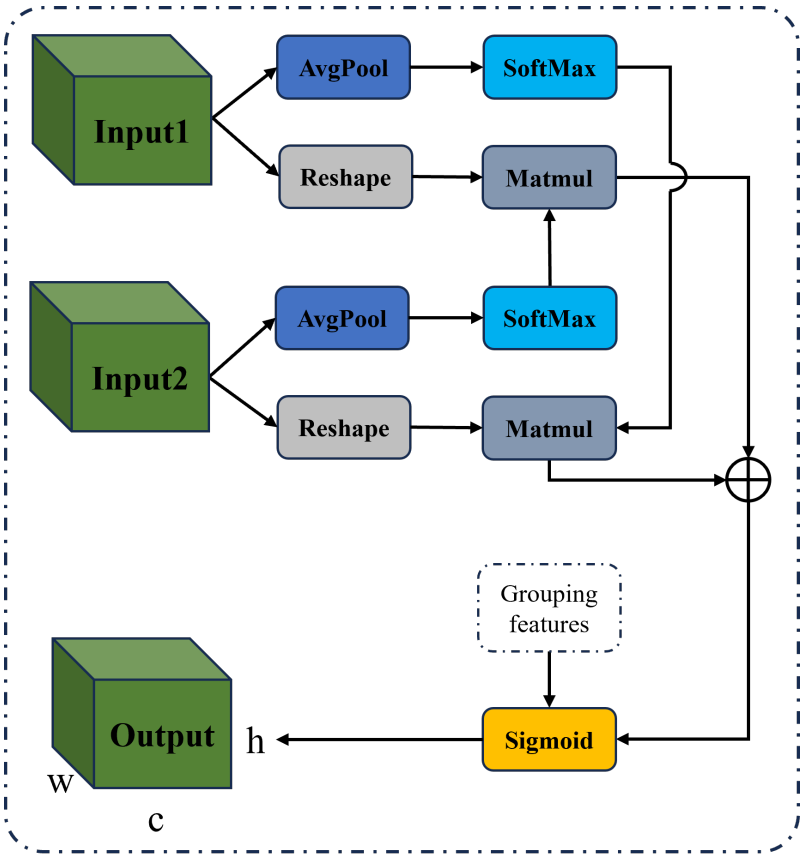

Figure 3. Enhanced multi-scale attention module.

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

 doi:10.20944/preprints202409.0871.v1

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.

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

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

Here, Input denotes two inputs, Input_1 and Input_2, where f_1 represents the processed value of Input_1 and f_2 represents the processed value of Input_2.

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 effectively enhances the model's ability to capture details and global structures, thereby improving its adaptability in complex scenes and enhancing the accuracy and robustness of normal prediction.

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

Through this multi-step feature fusion process, the module dynamically adjusts and integrates feature information from different scales, fully exploiting the complementarity between global and local features, thereby significantly enhancing the model's performance in handling complex scenes. This approach not only enhances the expressive power of features but also effectively addresses the issue of information loss in traditional convolutional neural networks when dealing with multi-scale features, greatly boosting the model's performance.

3.4. Hybrid Loss Function

In normal map prediction, high-frequency noise and irregular edges may affect the quality of the results. To address this issue, we propose a mixed loss function consisting of three parts: L1 loss, VGG perceptual loss, and Total Variation regularization. The L1 loss calculates the absolute error between the predicted normal and the target normal, aiming to improve prediction accuracy, which can be computed by the following formula:

$$L_1 = \|\text{pred} - \text{target}\|_1 \tag{8}$$

The VGG perceptual loss measures the difference between the predicted and target normals in the feature space using a pre-trained VGG network, enhancing visual quality and perceptual similarity, which can be calculated by the following formula:

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

Here, ϕ denotes the feature extraction function of the VGG network. Total Variation regularization is employed to reduce high-frequency noise in the normal map, maintaining smooth edges, which can be calculated by the following formula:

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

 doi:10.20944/preprints202409.0871.v1

Here, pred represents the predicted normal map, i and j are the pixel coordinates of the image, and $\text{pred}_{i+1,j}$ and $\text{pred}_{i,j+1}$ represent the pixel values of the image at positions $i+1, j$ and $i, j+1$, respectively. 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:

$$L_{\text{total\_V}} = 5.0 \times L_{\text{L1\_V}} + L_{\text{VGG\_V}} + \lambda_{\text{reg}} \times L_{\text{TV\_V}} \tag{11}$$

$$L_{\text{total\_IN}} = 5.0 \times L_{\text{L1\_IN}} + L_{\text{VGG\_IN}} + \lambda_{\text{reg}} \times L_{\text{TV\_IN}} \tag{12}$$

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

$$L_{\text{final}} = [L_{\text{total\_V}}, L_{\text{total\_IN}}] \tag{13}$$

4. Experiments

In this section, we validate the feasibility of our algorithm through extensive experiments. Section 4.1 introduces the datasets and experimental settings used. Section 4.2 presents the three evaluation metrics used in the experiments. Section 4.3 outlines the comparison results with state-of-the-art methods. Section 4.4 provides ablation studies and related discussions, while Section 4.5 presents failure cases and analysis.

4.1. Datasets and Implementation

The Thuman2.0 [33] and Cape [34] datasets are widely used in the computer vision domain, each with its own unique characteristics. Thuman2.0 is a dataset for human pose estimation, providing rich samples of human poses where different poses have varying effects on the morphology of different body parts, which is crucial for reconstructing 3D human models. The Cape dataset, on the other hand, provides images and models of humans wearing different clothing. Clothing occlusion and shape significantly impact the appearance of the human body, necessitating the consideration of these factors in the training data. Therefore, we chose these two datasets.

We evaluated our approach on the CAPE dataset, which consists of 150 models, to test the generalization of EMAR without any module trained on the Cape dataset. We divided the Cape dataset into "CAPE-Hard" and "CAPE-Easy" categories based on the presence of "difficult poses" and "non-difficult poses", with "CAPE-Hard" containing 100 models and "CAPE-Easy" containing 50 models, to better analyze the generalization ability on complex poses.

The proposed method was implemented on a desktop computer with an Intel(R) Core(TM) i9-10980XE CPU @ 3.00GHz, NVIDIA RTX A6000 GPU, and 128GB of memory. The EMAR model, as proposed in this paper, supports end-to-end training and utilizes the Adam optimizer with an initial learning rate of 1e-4, a batch size of 4, and trained for 20 epochs.

Considering the characteristics of point clouds, we opted to use three specific loss functions: 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 outperforms PIFu, PaMIR, and 2K2K. This improvement is attributed to our incorporation of the Enhanced Multi-Scale Attention Module, which enables the model to focus on more

 doi:10.20944/preprints202409.0871.v1

discriminative features at different scales. In contrast, PIFu and PaMIR are less effective in capturing details and handling complex features. The 2K2K method struggles with accurately estimating depth information due to the limited resolution of the input images, resulting in an unclear representation of the human form. Compared to ICON, our method shows significant improvements in detail, as illustrated in Figure 5. This is because we predict more accurate normal maps, providing the model with more reliable surface details.

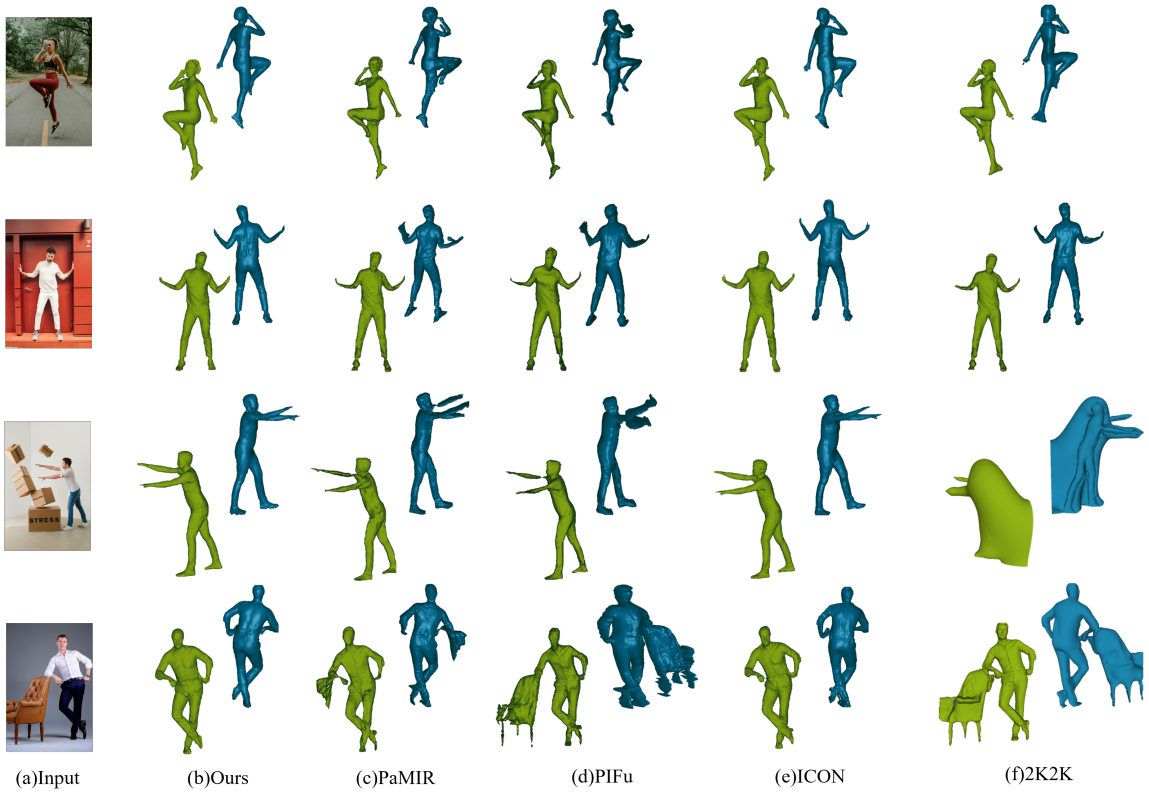

Figure 4. Qualitative experiments on in-the-wild photos: Column (a) shows the input images, column (b) presents the results of EMAR, column (c) shows the results of PaMIR, column (d) presents the results of PIFu, and column (e) shows the results of ICON.

Quantitative Experiments. We present the quantitative results in Table 1, where we tested all methods under two pose difficulty levels, "CAPE-Easy" and "CAPE-Hard." The metrics on the rightmost side under CAPE are the averaged results of "CAPE-Easy" and "CAPE-Hard" for the same row. It is evident from the table that our method outperforms the others.

Table 1. Quantitative results on the CAPE dataset. ↓ indicates lower values are better. Bold values represent the best performance.

Dataset	CAPE-Easy			CAPE-Hard			CAPE		
Method	Chamfer↓	P2S↓	Normals↓	Chamfer↓	P2S↓	Normals↓	Chamfer↓	P2S↓	Normals↓
PIFu	2.823	2.796	0.100	4.029	4.195	0.124	3.426	3.496	0.112
PaMIR	1.936	1.263	0.078	2.216	1.611	0.093	2.076	1.437	0.086
ICON	1.233	1.170	0.072	1.096	1.013	0.063	1.164	1.092	0.068
2K2K	1.264	1.213	0.070	1.437	1.385	0.060	1.351	1.299	0.065
Ours	**0.802**	**0.768**	**0.050**	**0.884**	**0.861**	**0.053**	**0.843**	**0.815**	**0.052**

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2024 doi:10.20944/preprints202409.0871.v1

Figure 5. Comparison of Details with ICON.

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 conduct ablation experiments using ICON as the baseline model. We implement the Feature Fusion module through channel concatenation, which is the most common method. Subsequently, we separately add the EMSA module and the Feature Fusion module to complete the experimental settings. Table 2 presents the numerical results of different combinations of these modules. With the addition of various key components, the performance of the network gradually improves, with our model achieving the best performance. Following is the baseline model with the EMSA module, while the baseline model performs the worst among all indicators. This further confirms the necessity and effectiveness of our proposed approach.

Table 2. Ablation experiments on the CAPE dataset.

Baseline	EMSA	FFM	Chamfer↓	P2S↓	Normals↓
✓			1.164	1.092	0.068
✓	✓		0.973	0.952	0.059
✓		✓	1.097	1.034	0.063
✓	✓	✓	0.843	0.815	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

 doi:10.20944/preprints202409.0871.v1

faults. In Figure 6-A, the target person's hands are clasped together, but the predicted human mesh shows the hands crossing fingers. In Figure 6-B, 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.

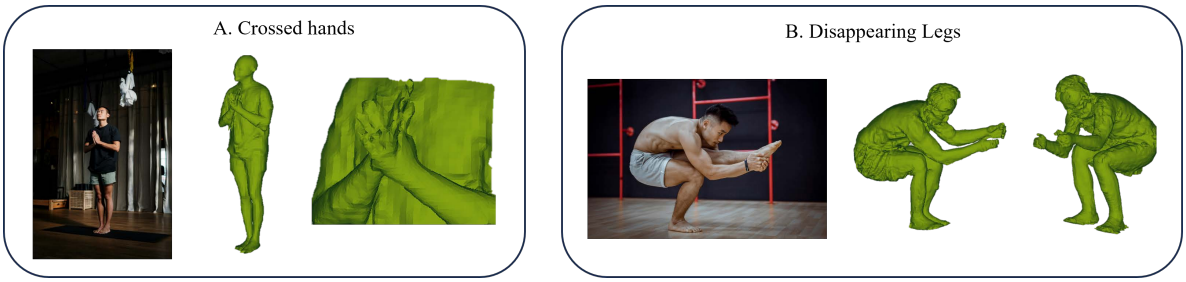

Figure 6. Comparison of Details with ICON.

5. Conclusion

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 design a Feature Fusion Module (FFM) to enhance the model's integration capability of features at different scales. Additionally, we introduce a loss function more suitable for normal map prediction, enabling the network to generate smoother normal maps. Finally, we conduct comparative experiments and ablation studies, and the results demonstrate that our method surpasses most state-of-the-art approaches.

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., B.S., 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, 2019GY215, 2021ZDLSF06-04), and the China Postdoctoral Science Foundation (2018M643719)

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 is no innovative data. Public data sets 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8748–8758.
2. Lochner, J.; Gain, J.; Perche, S.; Peytavie, A.; Galin, E.; Guérin, E. Interactive Authoring of Terrain using Diffusion Models. In Proceedings of the Computer Graphics Forum. Wiley Online Library, 2023, Vol. 42, p. e14941.

 doi:10.20944/preprints202409.0871.v1

3. Ma, X.; Zhao, J.; Teng, Y.; Yao, L. Multi-Level Implicit Function for Detailed Human Reconstruction by Relaxing SMPL Constraints. In Proceedings of the Computer Graphics Forum. Wiley Online Library, 2023, Vol. 42, p. e14951.
4. 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. *Journal of Imaging* **2024**, *10*, 133.
5. 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 preprint arXiv:2405.07472* **2024**.
6. 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, August 23–28, 2020, Proceedings, Part I 16. Springer, 2020, pp. 512–530.
7. 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). IEEE, 2022, pp. 13286–13296.
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.
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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 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 Transactions on Graphics (ToG)* **2018**, *37*, 1–15.
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 Proceedings of the IEEE/CVF international conference on computer vision, 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. Springer, 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 84–93.
14. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 84–93.
15. Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; Black, M.J. Econ: Explicit clothed humans optimized via normal integration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 512–523.
16. 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). IEEE, 2023, pp. 1–5.
17. 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 Transactions on Image Processing* **2022**, *31*, 1938–1948.
18. 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, October 23–27, 2022, Proceedings, Part V. Springer, 2022, Vol. 13695, *Lecture Notes in Computer Science*, pp. 590–606.
19. Zheng, Z.; Yu, T.; Liu, Y.; Dai, Q. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *44*, 3170–3184.
20. 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 preprint arXiv:2403.12028* **2024**.
21. Tang, Y.; Zhang, Q.; Hou, J.; Liu, Y. Human as Points: Explicit Point-based 3D Human Reconstruction from Single-view RGB Images. *arXiv preprint arXiv:2311.02892* **2023**.

22. Li, B.; Deng, Y.; Yang, Y.; Zhao, X. An Embeddable Implicit IUVD Representation for Part-based 3D Human Surface Reconstruction. *arXiv preprint arXiv:2401.16810* **2024**.
23. Lim, B.; Lee, S.W. TIFu: Tri-directional Implicit Function for High-Fidelity 3D Character Reconstruction. *arXiv preprint arXiv:2401.14565* **2024**.
24. 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. Springer, 2023, pp. 313–324.
25. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13211–13220.
26. 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, October 23–27, 2022, Proceedings, Part I. Springer, 2022, Vol. 13684, *Lecture Notes in Computer Science*, pp. 342–359.
27. 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 Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6765–6773.
28. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1954–1963.
29. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21254–21263.
30. Zhang, Z.; Yang, Z.; Yang, Y. SIFU: Side-view Conditioned Implicit Function for Real-world Usable Clothed Human Reconstruction. *arXiv preprint arXiv:2312.06704* **2023**.
31. Li, C.; Xiao, M.; Gao, M.; et al. R3D-SWIN: Use Shifted Window Attention for Single-View 3D Reconstruction. *arXiv preprint arXiv:2312.02725* **2023**.
32. Zhao, F.; Yang, W.; Zhang, J.; Lin, P.; Zhang, Y.; Yu, J.; Xu, L. Humannerf: Generalizable neural human radiance field from sparse inputs. *arXiv preprint arXiv:2112.02789* **2021**, *3*, 1.
33. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021; pp. 5746–5756.
34. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020; pp. 6469–6478.