Neural-code PIFu: High-fidelity Single Image 3D Human Reconstruction via Neural Code Integration

Ruizhi Liu[1], Paolo Remagnino[1], and Hubert P. H. Shum[1]

Department of Computer Science, Durham University, Durham, UK
{ruizhi.liu,paolo.remagnino,hubert.shum}@durham.ac.uk

Abstract. We introduce neural-code PIFu, a novel implicit function for 3D human reconstruction, leveraging neural codebooks, our approach learns recurrent patterns in the feature space and reuses them to improve current features. Many existing methods predict normal maps from image feature space which easily overlook non-trivial features. Moreover, neglecting global geometric correlations restricted the use of repetitive features to improve the expressive power of current features. In this work, we propose neural-code PIFu, a novel framework that enhances initial features by fusing them with neural codes that are learned from the input features and geometric prior. It also models the global geometric correlation to facilitate the use of neural codes. Extensive experiments demonstrate that our method outperforms state-of-the-art (SoTA) PIFu-based approaches by a large margin, and achieves comparable results to parametric-models-based methods without the use of auxiliary data.

Keywords: 3D Human Reconstruction · Deep Learning · Neural Code Integration.

1 Introduction

The growing demand for realistic 3D human reconstruction has driven the development of diverse methodologies, serving as a crucial foundation for the metaverse, and AR/VR industries. The main objective of 3D human reconstruction is to transform 2D features onto 3D surfaces that accurately represent the human in the RGB images. Early techniques [1] relied on dense view reconstruction to model intricate 3D human surfaces, but their reliance on sophisticated camera arrays made large-scale applications impractical. Recently, deep learning has revolutionized the field. Explicit representation is commonly used with deep learning to model human surfaces, early methods [1,2] based on explicit surfaces cannot generate details for human surfaces. To address the issue, [3,4] predicts 3D geometric offsets as clothing details. Despite promising results, explicit surface representations suffer from the inflexibility of modeling shape and struggle to replicate intricate garments such as dresses due to the significant divergence in shape from the human body.

Unlike explicit surfaces such as meshes, implicit surfaces can model arbitrary shapes and are not limited by the resolution of input data. Pixel-aligned implicit function first proposed in [5] has emerged as a promising approach in the field. PIFu [5] and PIFuHD [6] represent pioneering methods that employ implicit functions to reconstruct a human surface from a single RGB image directly. To reconstruct more detailed human surfaces, some methods [6,8,9,10] attempt to predict normal maps from the image feature and use them as additional inputs to inform the models.

The main problems of many PIFu-based methods are twofold: (1) Predicting normal maps from image feature space has limited improvements on non-trivial details. Many current methods easily overlook subtle details in image features which are also underrepresented in the predicted normal maps. This limits the improvements provided by normal maps. (2) Neglecting global geometric correlations among query points hinders the exploitation of repetitive patterns. In this work, our proposed alternative method addresses these issues without relying on complex architectures or additional data assistance.

To address these challenges, we propose Neural-Code PIFu, an end-to-end trainable approach for 3D human reconstruction from a single image. Inspired by [11] which learns quality-dependent features using vector quantization. Our method effectively learns repetitive patterns via neural codebook learning modules and models the overall global geometric correlations via self-attention with positional encoding to facilitate the use of neural codes. We improve pixel-aligned features by fusing them with relevant neural codes locally and globally via context-aware latent fusion. Finally, We fully integrate local and global features by facilitating query points to sufficiently interact via neural code integration.

Our method outperforms SoTA quantitatively and qualitatively. We evaluate neural-code PIFu on Thuman2.0 [12] and BUFF dataset [7] as well as out-of-distribution images to show the generalization of the proposed method. Our method demonstrates promising results, outperforming PIFu-based SoTA by a noticeable margin, and achieving comparable results with parametric-model-informed methods (i.e. ICON [13] and ECON [14]). The out-of-distribution evaluation demonstrates that our method generalises well to unseen garments and poses with minimum artifacts.

Our main contributions are summarised as follows:

- We propose an end-to-end trainable approach named **Neural-Code PIFu** for 3D human reconstruction from a single image, which learns reoccurring patterns and stores them as neural codes. It also models the global geometric correlation among query points.
- We propose **Context-Aware Latent Fusion** to reuse learned neural codes to improve the expressiveness of the feature. This allows more geometric details even if they are blurry in the given latent space.
- We propose **Neural Code Integration** to facilitate the interaction between query points,and also encourage local and global features to be adequately integrated.

2 Related Works

In this section, we briefly review the development and the relevant domains of single-view 3D human reconstruction.

Explicit Reconstruction An explicit surface can be described as a prescription of the precise location of the surface. The early methods represent a surface via voxels which discrete a 3D surface into a grid. This allows the explicit surface reconstruction to align with modern learning-based image processing methods [22,23,24], which can properly transfer 2D features to 3D surfaces without sacrificing a massive amount of consistency between 2D and 3D feature space. However, the aforementioned methods are highly sensitive to the resolution of input data, the computational consumption non-linearly increases with resolutions, which makes large-scale applications unfeasible. The point clouds, on the other hand, are computationally friendly in comparison to voxel representations [26,27,29]. Taking advantage of the properties of point clouds, recently proposed learning-based methods [13,14,28,27] can encode a sophisticated surface into a compact and sparse latent space with the cost of a small amount of computational resources, but loss of information is inevitable when mapping from a dense to a sparse latent space, point clouds normally lack abundant geometric information. This results in a loss of details and over-smoothed surfaces. Our method proposes to reuse repetitive patterns in the learned image feature space to enhance the surface details without additional inputs.

Implicit Human Surface Reconstruction Implicit representation could be considered as a function of the level set of the function [5]. This representation can be implemented as a multi-layer perceptron predicting occupancy field or SDF values, which indicate the probability of whether query points lie within the surface [5,6]. To convey more useful information from 2D input data to 3D surface, recent methods predict occupancy field conditioned on pixel-aligned features [5,6,16]. These methods have achieved promising results. However, most of the methods suffer from over-smoothed reconstructed surfaces.

To address this challenge, recent methods either introduce auxiliary data as prior or strong constraints, such as normal maps and parametric models (e.g. SMPL [2] and SMPL-X [25]) or add more 3D supervisions to the models. However, these methods fail to fully explore the valuable 2D space, and useful information such as detailed features lost during the transition from 2D feature space to 3D space.

3 Methodology

Our objective is to extract a highly detailed 3D human surface from a single-view image using a novel implicit function. This function employs neural codebooks to capture repetitive patterns and preserve them as neural codes, leveraging them

to enhance the expressiveness of features. We argue that when image features are blurred or over- smoothed, normal maps struggle to capture details, consequently restricting the improvement offered by normal maps. Additionally, these methods fail to emphasize global geometric correlation. Some methods like [7] incorporate a global feature map derived from image feature space. However, it has limited improvement in global awareness of models, as features of each query point are still isolated. To alleviate these challenges, we propose a novel framework to improve initial features by fusing them with neural codes that are learned from the input features and geometric prior. As shown in figure 1, we propose a selective learning neural codebook that specifically preserves representative and reoccurring features as neural codes. We purposefully utilize these neural codes to enhance the expressiveness of the current features, achieving the addition of human surface details without the need for additional data assistance. Moreover, we introduce an extra branch for modeling global geometric correlations which facilitates the use of neural codes.

Preliminary We start by detailing the background of the implicit function representation. An implicit function parameterizes a 3D surface as a level set of functions. Given a query point in the 3D space, an implicit function classifies the point as either inside or outside the surface. This is denoted as:

$$f(X) = \begin{cases} 1, & \text{if X is inside the surface,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Pixel-aligned implicit function captures detailed features from RGB images. It predicts the occupancy field which represents the probability distribution of whether a query point is inside or outside the surface. The pixel-aligned implicit function is mathematically defined as:

$$f(F_c(x), \phi(X)) = s : s \in R, \tag{2}$$

where $F_c(x)$ is 2D image feature at position x which is the 2D projection of query point X, and $\phi(\cdot)$ is the depth value of point X in the relative camera coordinates. For more details, we refer readers to [5].

3.1 Neural-Code PIFu Representation.

The inferior performance of current methods [6,8,9,10] is attributed to the prediction of normal maps from image features and the absence of global geometric correlation. These methods reconstruct detailed human surfaces dependent on normal maps derived from image features. Although introducing normal maps has been proven useful in adding details, it does not address the core issues. Firstly, the improvement provided by normal maps is limited if the initial features are non-trivial in the image feature space. Secondly, the majority of PIFu-based methods [5,6,8,9] fail to consider the global geometric correlation within query

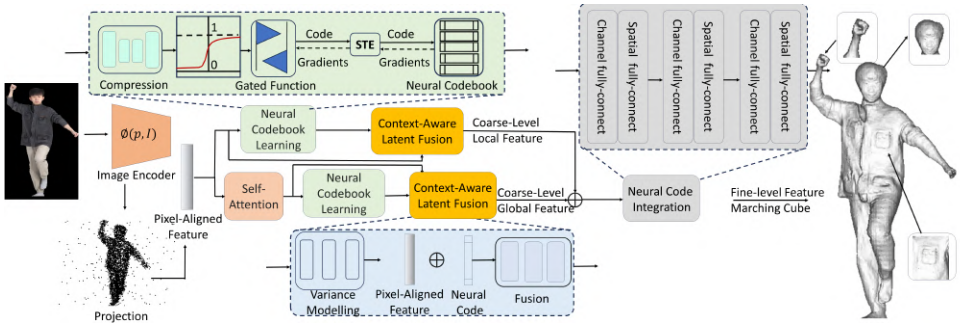

Fig. 1. The overview of our proposed method. The single view RGB image is fed into the image encoder, and the query points are projected to the image plane in order to obtain the pixel-aligned feature which is then used to obtain coarse-level local features and global features. The fine-level features are produced via neural codebook integration, which is used to predict the occupancy field for query points.

points. The global geometric correlation within the query points is essential for artifact reduction and completed human mesh.

To address the aforementioned challenges, we propose Neural-Code PIFu representation for human reconstruction, which possesses the ability to improve details based on input features and model global geometric correlations between query points. We adapt neural codebooks to learn representative and reoccurring features within the given latent space, selectively preserving them as neural codes. These neural codes are used as a complement for feature improvement, which allows the model not only to rely on image feature space but also on neural codebooks to acquire details. Moreover, modeling global geometric correlation informs the model with a global context, this allows a noticeable reduction of artifacts and efficient use of neural codes.

Our proposed model is mathematically represented as:

$$F_Q(x_c, F_g(f_l, f_g), \phi(X)) = s : s \in R, \tag{3}$$

where x_c is the input feature, F_g is the neural code integration which takes global and local features, denoted as f_g and f_l, as inputs. This module allows local and global features to be fully combined. Additionally, coarse global feature f_g and coarse local feature f_l are generated via the context-aware latent fusion described in section 2.3. We apply self-attention with positional encoding to model the overall global correlation within all query points. This is denoted as follows:

$$SA(x_c) = softmax(\frac{Pos(Q)Pos(K)^T}{\sqrt{d_k}}) \cdot Pos(V). \tag{4}$$

Each query point not only contains its features but is also weighted based on all the other query points after this operation.

Neural Codebook Learning. We use neural codebooks to effectively capture and reuse representative features for reconstructing detailed human surfaces. The goal of the neural codebook learning module is to learn the latent distribution representing a shared collection of appearance and geometry within the given features. Given a pixel-aligned feature x_c of query points, we first extract the n most representative features in x_c via a softmax relaxation of nearest-neighbor:

$$z_i \leftarrow \frac{e^{||z_i - x_c||_2}}{\sum_{n=1}^{k} e^{-||z_n - x_c||_2}}. \tag{5}$$

We adapt a straight-through-estimator (STE) to enable backpropagation through the neural codebook, which is vital for a learnable codebook. All neural codes are initialized with a standard Gaussian distribution $\mathcal{N} \sim (\mu, \sigma)$.

Gated Function. This function selectively preserves the neural codes of interest while discarding less relevant ones, ensuring the retention of the most distinct features captured in the input latent space. This step is crucial for reconstructing intricate surface details without introducing artifacts. The gated function is denoted as:

$$z_i \leftarrow \omega(\varphi(v_s - i^2/\lambda), T) \cdot z_i. \tag{6}$$

The gated function provides a hard decision boundary for neural code selection. The ω is a binarization function. T is a manually defined threshold, v_s is a scoring function that weights the inputs, and λ is a scaling hyperparameter based on the size of the neural codebook.

Discussions. Our method possesses better generalisation and flexibility in selecting features in comparison to existing methods like SuRS [15]. SuRS learns a prior difference between high- and low- resolution surfaces. This benefits reconstruction when the details in the image are non-trivial. Nevertheless, it is significantly constrained by the limitations imposed by the training distribution, demanding additional data and supervision. Additionally, it lacks the flexibility of applying learned prior knowledge to inform the model, which introduces a lot of artifacts. In contrast, our approach can selectively employ neural codes to enhance those blurred features. This contributes to artifact reduction and better generalisation.

3.2 Context-Aware Latent Fusion

Intuitively, details of the clothed human body, such as clothing wrinkles and facial contours, exhibit significant similarities. Existing works [13,14,16,6,15] fail to take advantage of these similarities and reuse them to enhance non-trivial details.

Therefore, we propose context-aware latent fusion leveraging neural codebook learning modules for improvements of both local and global features. This module generates coarse-level local and global features by combining the input

features with learned neural codes. This allows a better representative power to improve the non-trivial details in the initial image space. This also enables the model to process out-of-distribution images. The module has two primary steps, variance modeling and latent fusion between neural codes and input features.

Variance Modelling . To ensure the neural codes are distinctive within the codebook, we follow [17] to further maximize the distance between learned neural codes by modeling the intra-variance between each code. The intra-variance is modeled using a convolutional neural network V which takes both neural code z_i and input feature x_c and outputs the variance perturbation:

$$z_i = z_i + \epsilon \cdot \frac{V(z_i, x_c)}{||V(z_i, x_c)||_2}. \tag{7}$$

It draws a clearer boundary within different neural codes and benefits the reduction of artifacts on the reconstructed surface, as ambiguity within the features deteriorates the uncertainty of points near the surface [18]. Introducing variance perturbation to neural codes eases the uncertainty.

Latent Fusion. It aims to generate local and global coarse-level features by merging the input latent with its relevant neural codes. There are two branches to separately process local and global features. We concatenate the input feature and its neural code and feed it into the local fusion module which is modeled as a residual MLP to obtain both coarse-level features.

3.3 Neural Code Integration

Deficiency in communication between query points is one of the weaknesses of previous PIFu-based models. Existing approaches [5,6,8] fail to facilitate sufficient interaction among the query points, and the local and global features are not adequately integrated.

Hence, we propose neural code integration to integrate coarse-level local and global features into fine-level features. The purpose of this module is to enable spatial-wise and channel-wise communication between both features.

We adapt MLP-mixer architecture [19] over the commonly used vision transformer for not only its simplicity but also for its comparable performance with a lighter computational burden. We modify the original architecture and directly apply it to the latent space. There are two steps within the neural code integration module: channel-wise mixing and location-wise mixing. In our case, the former enables communication within each feature of query points, the latter allows interaction within different query points. The neural code integration module is defined as:

$$f_{channel} = x_{g,l} + MLP_{channel}(x_{g,l}), \tag{8}$$

$$f_{spatial} = f_{channel} + MLP_{spatial}(f_{channel}), \tag{9}$$

where $MLP_{channel}$ and $MLP_{spatial}$ are responsible for channel-wise mixing and spatial mixing respectively, the x_g, l is the concatenation of coarse-level local feature and global feature.

We use fine-level features to predict the occupancy field with an MLP surface classifier, and the reconstructed mesh is extracted following [20].

4 Experiments

To evaluate the performance of the proposed method, we conducted extensive experiments on two publicly accessible datasets that are widely accepted by the community, including Thuman2.0 [12], BUFF [7].

Datasets. Thuman2.0 [12] constitutes 524 high-resolution human meshes with rich details on the surface. We follow the split ratio mentioned in [15] to split the dataset into training and testing sets, which contain 402 and 122 meshes respectively. To evaluate the generalization of our proposed model, we conduct further experimentation on 143 human scans of both BUFF which no methods use for training.

Evaluation Metric. In our experiments, we leverage Chamfer Distance (CF) to measure the distance between the reconstructed surfaces and the ground truth surfaces. Average point-to-surface Euclidean distance (P2S) is applied to measure the distance from the vertices of the reconstructed surfaces to the ground truth surfaces. Lastly, we harness normal reprojection error to evaluate the projection consistency from input image. All metrics are measured in centimeters (cm).

Implementation Details . Our proposed model is trained with RGB image with the size of $(N_I \times N_I, N_I = 512)$. We follow the same rendering process used in PIFU [5] to generate images at every degree along the yaw axis for each human scan. The ground truth 3D points are sampled following the spatial sampling procedure mentioned in PIFu [5] The input image is first encoded via a 2D convolutional neural network containing a stacked hourglass network which has been proven to possess better generalization for human-related estimation. The encoded continuous image features, which have the shape of (W, H, C, W= 128, H=128, C=321). Pixel-aligned features then are obtained by projecting the query points to the image feature space. Pixel-aligned features are then passed to the neural codebook learning module to be decomposed and extract the most representative neural code. The coarse-level features are learnt through context-aware latent fusion which are learned via a 4-layer Multi-Layer Perceptron (MLP). A fine-level feature is produced via neural code integration which takes both global and local coarse-level features as input. Regarding the final occupancy prediction, we adapt a surface classifier formulated as a residual MLP to classify the fine-level features. Once the occupancy values are obtained, we visualize it using [20].

Our model is trained on TITAN X GPU with 8 batches, and a learning rate of 0.0001 with decay. The model is optimized via Adam.

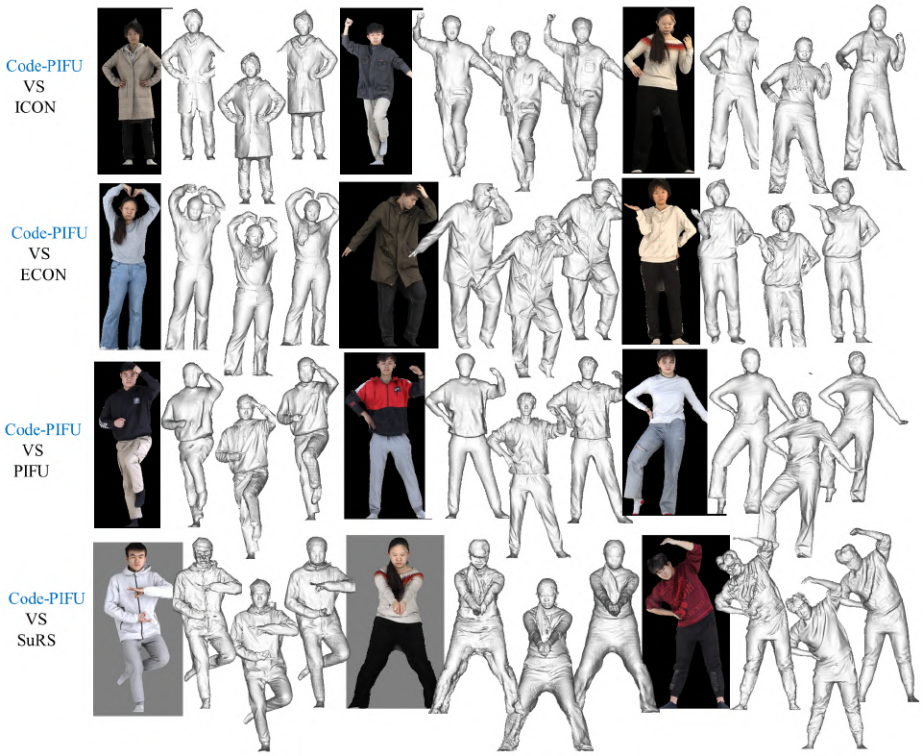

Fig. 2. The comparisons between our method and SoTA. From left to right are the results of the selected SoTA, the ground truth and our results.

4.1 Comparisons on SoTA.

We compare our methods with state-of-the-arts methods: PIFu [5], PIFuHD [6], PaMIR [16], SuRS [15], ICON [13], ECON [14]. We follow the training protocol of SuRS [15],and retrain all methods using their released source codes for fair comparisons.

Table 1 shows quantitative results on Thuman2.0 [12], BUFF [7] dataset. Our method outperforms all PIFu-based SoTA with noticeable margins on the Thuman2.0 test dataset and BUFF dataset. As shown in figure 2, our proposed method produces more plausible meshes with minimum artifacts.

Our method outperforms parametric model-based methods ICON [13] and ECON [14] on Thuman2.0 dataset [12],and achieves comparable results on

BUFF dataset [7]. This is largely attributed to the utilization of parametric models in these methods for rendering human normal vector maps, which are subsequently employed to predict normal vector maps with clothing. The normal vectors obtained through parametric model rendering demonstrate greater stability and accuracy compared to those predicted directly from image features. The discrepancy in performance is particularly noticeable on the BUFF dataset.

	THuman 2.0 Dataset			BUFF Dataset		
Models	Chamfer(↓)	P2S(↓)	Normal(↓)	Chamfer(↓)	P2S(↓)	Normal(↓)
PIFu	1.501 (↓50%)	1.523 (↓51%)	0.122 (↓37%)	1.781 (↓57%)	1.754 (↓55%)	0.142 (↓39%)
PIFuHD	1.372 (↓46%)	1.432 (↓49%)	0.124 (↓38%)	1.634 (↓54%)	1.671 (↓53%)	0.133 (↓35%)
PaMIR	1.713 (↓56%)	1.818 (↓60%)	0.134 (↓43%)	1.752 (↓57%)	1.872 (↓58%)	0.148 (↓42%)
SuRS	0.931 (↓20%)	1.151 (↓36%)	0.107 (↓28%)	1.532 (↓50%)	1.622 (↓51%)	0.136 (↓37%)
ICON	0.747 (↓0.5%)	0.735 (↓0.5%)	0.086 (↓10%)	0.832 (↓9%)	0.854 (↓9%)	0.087 (-)
ECON	0.748 (↓0.5%)	0.737 (↓0.5%)	0.079 (↓3%)	0.762 (↓0.5 %)	**0.732 (↑6%)**	**0.082 (↑5%)**
Ours	**0.745**	**0.733**	**0.077**	**0.759**	0.781	**0.086**

Table 1. The Quantitative results on Thuman 2.0 and BUFF dataset. The percentage shows the improvements of the proposed method in comparison to SoTA. Chamfer, P2S and Normal consistency evaluation: the smaller the better.

	Thuman 2.0 Dataset			BUFF Dataset		
Modules	Chamfer	P2S	Normal	Chamfer	P2S	Normal
w/ codebooks w/o fusion	0.774(↓4%)	0.764(↓4%)	0.082(↓6%)	0.787(↓4%)	0.791(↓1%)	0.090(↓4%)
w/o global codebook	0.762(↓2%)	0.754(↓3%)	0.081(↓5%)	0.797(↓5%)	0.798(↓2%)	0.089(↓3%)
w/o local codebook	0.780(↓5%)	0.773(↓5%)	0.084(↓8%)	0.815(↓7%)	0.824(↓5%)	0.092(↓7%)
w/o fusion	0.767(↓3%)	0.769(↓5%)	0.086(↓10%)	0.782(↓3%)	0.813(↓4%)	0.090(↓4%)
w/o Integration	0.782(↓5%)	0.791(↓7%)	0.088(↓13%)	0.812(↓7%)	0.833(↓6%)	0.092(↓7%)
w/ all modules	**0.745**	**0.733**	**0.077**	**0.759**	**0.781**	**0.086**

Table 2. The ablation results on Thuman 2.0 and BUFF dataset. The percentage shows the performance improvement with or without key components. Chamfer, P2S, and Normal consistency evaluation: the smaller the better.

4.2 Ablation Study.

We evaluate our methods with a series of ablation studies to assess the key components contributing to the overall performance. Table 2 illustrates the performance with and without some significant modules of the proposed method. First, we evaluate the importance of two neural codebook learning modules. It is obvious that the performance dramatically deteriorates without the two neural codebook learning modules. Moreover, deployment of either neural codebook learning module boosts the performance, but the local codebook learning module has a large impact on overall performance in comparison to its counterpart. Lastly, it is noticeable that without neural code integration results in the worst performance. Figure 3 shows the results of the qualitative ablation study. The local neural codebook contributes to the diversity of surface details such as fingers and facial details. It is noticeable that surfaces tend to suffer from local surface detail insufficiency without the local neural codebook. Similarly, neural

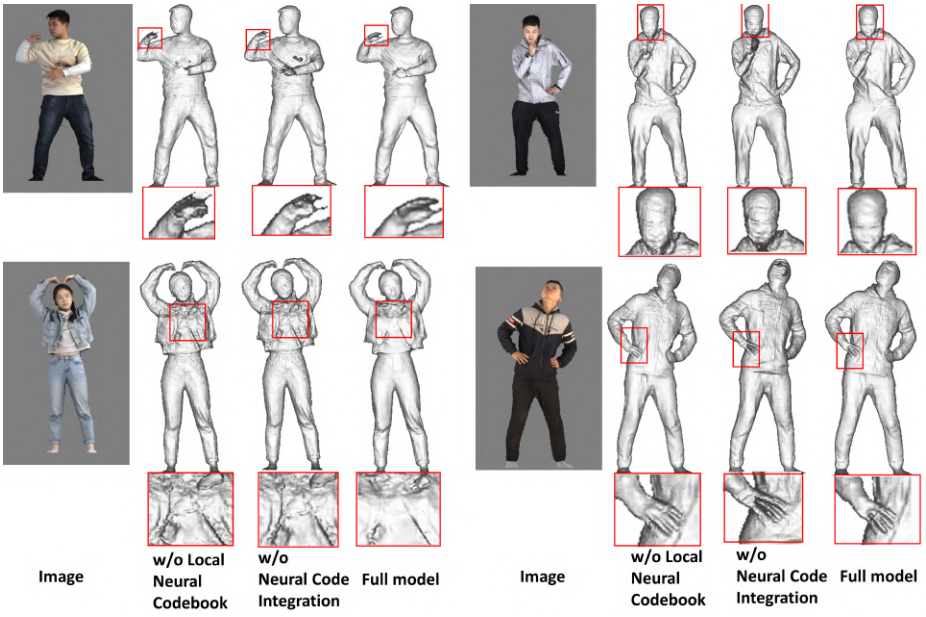

Fig. 3. Qualitative ablation study on Thuman 2.0 test dataset.

code integration encourages local and global neural codes to be fully integrated, this is beneficial for surface details preservation.

4.3 Out-Of-Distribution Image Evaluation.

We involve out-of-distribution image evaluation to further demonstrate the generality of our proposed models. As shown in figure 4, our model generalizes well on unseen images that are beyond the distribution of the training dataset. Our learned neural code book can generalize well to various unseen garment details and fashion poses without further training. Unlike SuRS [15] which are highly constrained by the distribution of training data, our method captures the most frequently appeared patterns in the training data, and utilizes them to improve the expressiveness of input features beyond training distribution. We also capture more details than ICON which also predicts normal maps from image feature space.

5 Conclusion and Discussions

In conclusion, we propose a novel framework for 3D human reconstruction from a single image named neural-code PIFu which bridges the pixel-aligned features and its neural codes for better expressiveness. Our method predicts the coarse-level feature for both local and global contexts and applies two neural code

Fig. 4. Out-of-distribution evaluation of the proposed method.

books to learn the distinctive neural codes. The fine-level feature is produced via a neural code integration which considers the global geometric correlation of each feature, resulting in much detailed human surfaces.

Although our method surpasses SoTA in terms of generalisation, details capturing, and preservation for unseen clothing, our method shows weaknesses in reconstructing unseen poses which may result in broken meshes. Additionally, our method tends to recognize hair as details of garments, this frequently occurs when reconstructing females in fashion poses. In future research, we will investigate combining uncertainty modeling, domain adaption, and diffusion models to alleviate the mentioned challenges.

References

1. HanbyulJoo, TomasSimon,andYaserSheikh, "Total capture: A3dde formation model for tracking faces, hands, and bodies," in Proceedings of the IEEE conference on computer vision and pattern recognition.

2. Loper et al., "Smpl: A skinned multi-person linear model," in Seminal Graphics Papers: Pushing the Boundaries.
3. Alldieck et al., "Learning to reconstruct people in clothing from a sin gle rgb camera," in Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition.
4. Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll, "360-degree textures of people in clothing from a single image," in 2019 Interna tional Conference on 3D Vision (3DV). IEEE, 2019, pp. 643–653.
5. Saito et al., "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in Proceedings of the IEEE/CVF interna tional conference on computer vision, 2019, pp. 2304–2314.
6. Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo, "Pi fuhd: 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, 2020, pp. 84–93.
7. Detailed, accurate, human shape estimation from clothed 3d scan se quences," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.
8. Lifang Chen, Jianghu Su, and Shiyong Luo, "Transpifu: Combining transformer and pixel-aligned implicit function for single-view clothed human reconstruction," Computers Graphics, vol. 111, pp. 1–13, 2023.
9. KennardYantingChan, GuoshengLin, HaiyuZhao, andWeisiLin, "In tegratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction," in European conference on computer vision. Springer, 2022, pp. 328–344.
10. Kennard Chan, Guosheng Lin, Haiyu Zhao, and Weisi Lin, "S-pifu: Integrating parametric human models with pifu for single-view clothed human reconstruction," Advances in Neural Information Processing Systems, vol. 35, pp. 17373–17385, 2022.
11. Zhou Yang, Weisheng Dong, Xin Li, Mengluan Huang, Yulin Sun, and Guangming Shi, "Vector quantization with self-attention for quality independent representation learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24438–24448.
12. Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu, "Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021), June 2021.
13. Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black, "Icon: Implicit clothed humans obtained from normals," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 13286–13296.
14. Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black, "Econ: Explicit clothed humans optimized via normal integration," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 512–523.
15. Marco Pesavento, Marco Volino, and Adrian Hilton, "Super-resolution 3d human shape from a single low-resolution image," in European Conference on Computer Vision. Springer, 2022, pp. 447–464.
16. ZerongZheng, TaoYu, YebinLiu, and Qionghai Dai, "Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 6, pp. 3170–3184, 2021.

17. Matthew Wallingford, Aditya Kusupati, Alex Fang, Vivek Ramanujan, Aniruddha Kembhavi, Roozbeh Mottaghi, and Ali Farhadi, “Neural radiance field codebooks,” arXiv preprint arXiv:2301.04101, 2023.
18. Xueting Yang, “D-if: Uncertainty-aware human digitization via implicit distribution field,” International Conference on Computer Vision, 2023.
19. Tolstikhin et al., “Mlp-mixer: An all-mlp architecture for vision,” Ad vances in neural information processing systems, vol. 34, pp. 24261 24272, 2021.
20. William E Lorensen and Harvey E Cline, “Marching cubes: A high-resolution 3d surface construction algorithm,” in Seminal graphics: pioneering efforts that shaped the field, pp. 347–353. 1998.
21. Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black, “Econ: Explicit clothed humans optimized via normal integration,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 512–523.
22. Ji, Mengqi and Gall, Juergen and Zheng, Haitian and Liu, Yebin and Fang, Lu,"SurfaceNet: An End-To-End 3D Neural Network for Multiview Stereopsis", Proceedings of the IEEE International Conference on Computer Vision (ICCV)
23. Jimenez Rezende, D., Eslami, S. M., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N. (2016). Unsupervised learning of 3d structure from images. Advances in neural information processing systems, 29.
24. Kar, A., Häne, C., Malik, J. (2017). Learning a multi-view stereo machine. Advances in neural information processing systems, 30.
25. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., Black, M. J. (2019). 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 (pp. 10975-10985).
26. Jiang, H., Cai, J., Zheng, J. (2019). Skeleton-aware 3d human shape reconstruction from point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5431-5441).
27. Wang, Junying, et al. "Complete 3D Human Reconstruction from a Single Incomplete Image." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
28. Lu, Yang, et al. "3d real-time human reconstruction with a single rgbd camera." Applied Intelligence 53.8 (2023): 8735-8745.
29. Zhao, Xiaoming, et al. "Occupancy planes for single-view rgb-d human reconstruction." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 3. 2023.