2025 IEEE International Joint Conference on Biometrics (IJCB) | 979-8-3315-0364-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/IJCB65343.2025.11410699

M3DHMR: Monocular 3D Hand Mesh Recovery

Yihong Lin
South China University of Technology
Guangzhou, China
jlxxlyh@163.com

Xianjia Wu
Huawei Cloud
Shenzhen, China
wuxianjia1996@gmail.com

Xilai Wang
South China University of Technology
Guangzhou, China
auwangxilai@mail.scut.edu.cn

Jianqiao Hu
South China University of Technology
Guangzhou, China
jianqiaohu2000@163.com

Songju Lei
Nanjing University
Nanjing, China
leisongju@smail.nju.edu.cn

Xiandong Li†
Huawei Cloud
Shenzhen, China
lxdphys@smail.nju.edu.cn

Wenxiong Kang†
South China University of Technology
Guangzhou, China
auwxkang@scut.edu.cn

Abstract

Monocular 3D hand mesh recovery is challenging due to high degrees of freedom of hands, 2D-to-3D ambiguity and self-occlusion. Most existing methods are either inefficient or less straightforward for predicting the position of 3D mesh vertices. Thus, we propose a new pipeline called Monocular 3D Hand Mesh Recovery (M3DHMR) to directly estimate the positions of hand mesh vertices. M3DHMR provides 2D cues for 3D tasks from a single image and uses a new spiral decoder consist of several Dynamic Spiral Convolution (DSC) Layers and a Region of Interest (ROI) Layer. On the one hand, DSC Layers adaptively adjust the weights based on the vertex positions and extract the vertex features in both spatial and channel dimensions. On the other hand, ROI Layer utilizes the physical information and refines mesh vertices in each predefined hand region separately. Extensive experiments on popular dataset FreiHAND demonstrate that M3DHMR significantly outperforms state-of-the-art real-time methods. The code is available at https://github.com/Jackson-coder/M3DHMR.

1. Introduction

Monocular 3D hand mesh recovery has attracted great attention in VR/AR, human-computer interaction, etc [30, 16]. This task aims to estimate the 3D spatial positions of hand mesh vertices from a single RGB image. It improves the accuracy and security of hand gesture authentication by restoring biometric features. On the one hand, it repairs low-quality samples caused by factors such as wear, contamination, or limitations of capture devices, which reduces false recognition rates. On the other hand, it reconstructs incomplete data, thereby boosting performance of multimodal hand gesture authentication system. However, high degrees of hand freedom, 2D to 3D ambiguity and self-occlusion make significant challenges for hand reconstruction [1, 9, 5].

To overcome these challenges, parametric model-based approaches [34, 6, 29, 27] usually use MANO [28] which obtains a hand mesh from coefficients of shape and pose to facilitately generate diverse hand poses for various applications with flexibility and high-precision. Though these methods are flexible, their 3D structure is not intuitive enough, especially the shape parameters are too difficult to optimize. Conversely, vertex-based methods can directly estimate the coordinates of the 3D mesh. Specifically, these

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

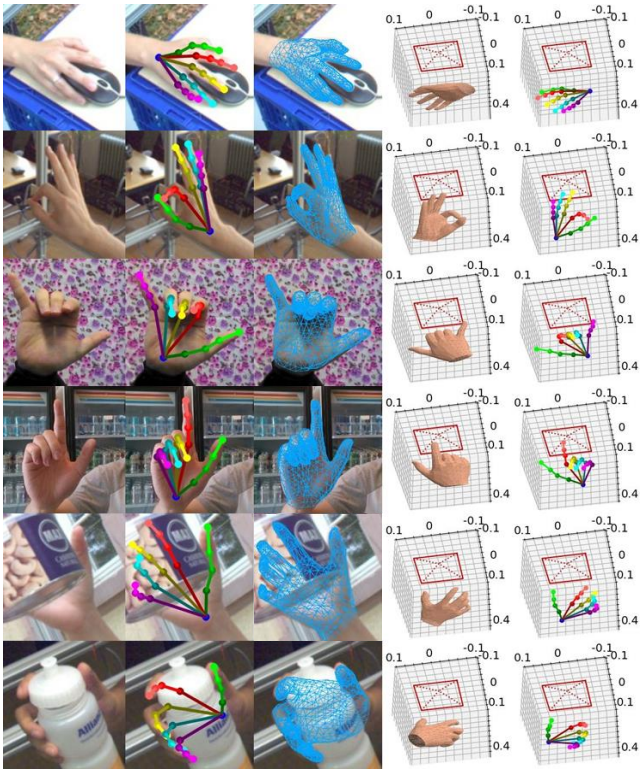

Figure 1. Qualitative results of our M3DHMR. We show the 2D pose, mesh projection, camera space mesh and pose in metres. The red rectangle indicates the camera.

models usually define the wrist of the hand as the root and calculate the root-relative 3D mesh coordinates. Some latest researches [22, 23, 18] are unable to implement in real-time due to their exclusive focus on improving the reconstruction accuracy, without considering the scale of models. Others [13, 20, 7, 33, 2, 30, 15, 31] can achieve real-time performance but are limited in reconstruction accuracy. Additionally, all methods previously mentioned have to provide the absolute root coordinate in camera space to get absolute camera coordinates of the 3D mesh, which limits the application to many high-level tasks. Though CMR [5] and MobRecon [3] provide pipelines to directly acquire the absolute 3D coordinates of the mesh in real-time, their reconstruction accuracy still needs to be improved.

Herein, we propose a Monocular 3D Hand Mesh Recovery (M3DHMR) pipeline to further improve the performance of real-time camera-space 3D hand mesh recovery. The pipeline reconstructs the 3D hand through multiple stages and facilitates the 3D reconstruction by regressing 2D joint landmarks as a subtask. Novel Dynamic Spiral Convolution (DSC) Layers and Region of Interest (ROI) Layer are used to decode the 3D vertex features to obtain mesh 3D coordinates. Instead of employing stacked spiral convolution layers or increasing the number of channels in these layers to enhance the model capacity, which is often accompanied with expensive computation, DSC Layers explore a more reasonable spiral convolution architecture to further enhance the model's representation capability with minimal increase in number of parameters and computational cost. More concretely, two convolution layers of DSC are used to adaptively adjust the normalized weights of the K spiral convolutions to extract the spatial-channel information. Furthermore, ROI Layer introduces the spatial information of the hand in the physical world by dividing the hand into multiple physical regions and refining the hand mesh vertices in each region separately. Extensive experiments indicate that M3DHMR surpasses current state-of-the-art real-time models in 3D hand mesh reconstruction. Some reconstruction results of M3DHMR in situations of complex pose, self-occlusion and object occlusion are shown in Fig. 1. Our contributions are summarized as follows:

- We design DSC Layers to adapt the weights of spiral convolutions based on the vertex positions and explore the mesh features in both space and channels.
- We propose a ROI Layer that divides the mesh vertices of the hand into several sets based on the physical region to separately refine the hand mesh features.
- Extensive experiments demonstrate that our method achieves superior performance in terms of reconstruction accuracy over existing state-of-the-art real-time methods.

2. Related Work

2.1. Hand Mesh Estimation

Hand mesh estimation has become an important task in computer vision and augmented reality, which aims to reconstruct detailed 3D hand structures from monocular RGB images. Existing mainstream methods can be categorized into model-based and vertex-based approaches.

Model-based approaches commonly rely on parametric hand models, such as MANO [28], which provides a low-dimensional representation of hand poses and shapes. Building upon this foundation, Zhou et al. [34] proposed the first real-time monocular hand capture framework, integrating multi-modal data (RGB, keypoints, and contours) to achieve highly accurate 3D reconstructions. Furthermore, Chen et al. [6] developed a self-supervised model-based 3D hand reconstruction method that eliminates the need for 3D annotation by leveraging geometric constraints and 2D-3D consistency for training. Meanwhile, RealisticHands [29] combined parametric mesh priors with non-parametric details, enabling high-quality hand pose and shape estimation from monocular RGB inputs. HaMeR [27] used a fully transformer-based architecture to analyze hands with significantly increased accuracy and robustness. Despite their

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

efficiency, these methods are inherently limited by the expressiveness of parametric models.

Vertex-based methods directly predict the 3D coordinates of mesh vertices, offering flexibility beyond parametric models. Among these, Mesh Graphormer [23] and PointHMR [18] pioneered the use of transformer architectures for monocular 3D hand mesh recovery, achieving notable improvements in accuracy and robustness. However, their computational complexity hinders real-time deployment. To address this limitation, several efficient alternatives have emerged. MobileHand [13] introduced a lightweight network design, while YoutubeHand [20] and Pose2Mesh [7] leveraged graph convolutions for efficient mesh reconstruction. Further explorations include HIU-DMTL [33], which employed deep multi-task learning, and I2UV-HandNet [2], which adopted an image-to-UV mapping strategy. Tang et al. [30] enhanced real-time performance and significantly improved pose and shape accuracy through a novel alignment optimization module. CMR [5] and MobRecon [3] proposed end-to-end pipelines to directly estimate absolute 3D coordinates. More recently, CLIP-Hand3D [15] integrated vision-language pretraining with context-aware prompting to improve generalization in challenging scenarios, and FastViT [31] introduced a hybrid vision transformer with structural reparameterization for optimal speed-accuracy trade-offs.

2.2. Spiral Convolution

Defferrard et al. [8] pioneered graph convolutional networks (GCNs) that operate on non-Euclidean graph structures while maintaining computational complexity comparable to that of classical CNNs. Masci et al. [24] developed spatial graph convolution operators that incorporate geodesic-aware sampling of graph signals, providing more intuitive handling of 3D mesh vertices in the spatial domain. This line of research was further advanced by Lim et al. [21], who introduced spiral convolution based on graph convolution as an effective approach for processing 3D mesh vertices.

Subsequent improvements in mesh spatial convolution techniques include SpiralNet++ [10], which employed truncated spiral lines to limit sampled vertices while expanding the receptive field through hollow spiral lines. Kolotouros et al. [20] directly learned the mapping relationships from 2D images to 3D meshes by implementing a straightforward encoder-decoder architecture with spiral convolution. Recent works like CMR and MobRecon designed ISM and DSConv to further enhance feature extraction.

Despite these advancements, the weights of current methods remain fixed when training is finished. This fixed nature restricts their ability to adapt to varying spatial locations of mesh vertices. To solve this problem, our work explores a more flexible convolution operator that can dynamically adjust to the spatial distribution of mesh vertices, potentially offering improved performance in 3D mesh processing tasks.

3. Method

3.1. Overview

Taking a single RGB image as input, our goal is to get the 3D coordinates of the hand mesh vertices. M3DHMR firstly uses a two-stack hourglass network [26] to generate encoding features F^e and predict 2D joint landmarks L^p. Note that the first stack and the first layer of the second stack are pretrained under guidance of heatmaps and landmarks. Moreover, M3DHMR adopts a spiral decoder to infer the root-relative coordinates of the 3D mesh vertices from the encoding features F^e. Specifically, pose features F^p are first obtained by grid sampling:

$$F^p = F^e(L^p). \tag{1}$$

Then a learnable matrix M lifts 2D pose features to 3D mesh features:

$$F^m = M \cdot F^p. \tag{2}$$

Subsequently, multiple graph-based DSC Layers produce multiple scale mesh vertex maps in a coarse-to-fine manner. Each layer upsamples output mesh features from the previous layer and explores spatial connectivity and channel dependencies between vertices to get fine-grained vertex features. Finally, M3DHMR adopts a ROI Layer to refine the result and uses the same method as CMR [5] to estimate the position of root in the camera space and thus recovers the coordinates of the 3D hand mesh vertices in the camera space. The overall network structure is shown in Fig. 2.

3.2. Dynamic Spiral Convolution Layer

Similar to SpiralConv [21] and SpiralConv++ [11], dynamic spiral convolution determines the convolution centre and produces a sequence of enumerated centre vertices based on adjacency, followed by the 1-ring vertices, the 2-ring vertices, and so on, until all vertices containing n rings are included:

$$\begin{aligned} 0{-}ring(v) &= v, \\ (n+1){-}ring(v) &= \mathcal{N}(n{-}ring(v)) \setminus n{-}disk(v), \\ n{-}disk(v) &= \cup_{i=0,\ldots,n} i{-}ring(v), \end{aligned} \tag{3}$$

where $\mathcal{N}(V)$ selects all vertices in the neighborhood of any vertex in set V. Different from previous works, dynamic spiral convolution comprises K convolution kernels that share the same kernel size and input/output dimensions. To enable adaptive adjustment of convolution weights for vertices at different positions, it employs two convolution layers and softmax to generate normalized attention weights

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

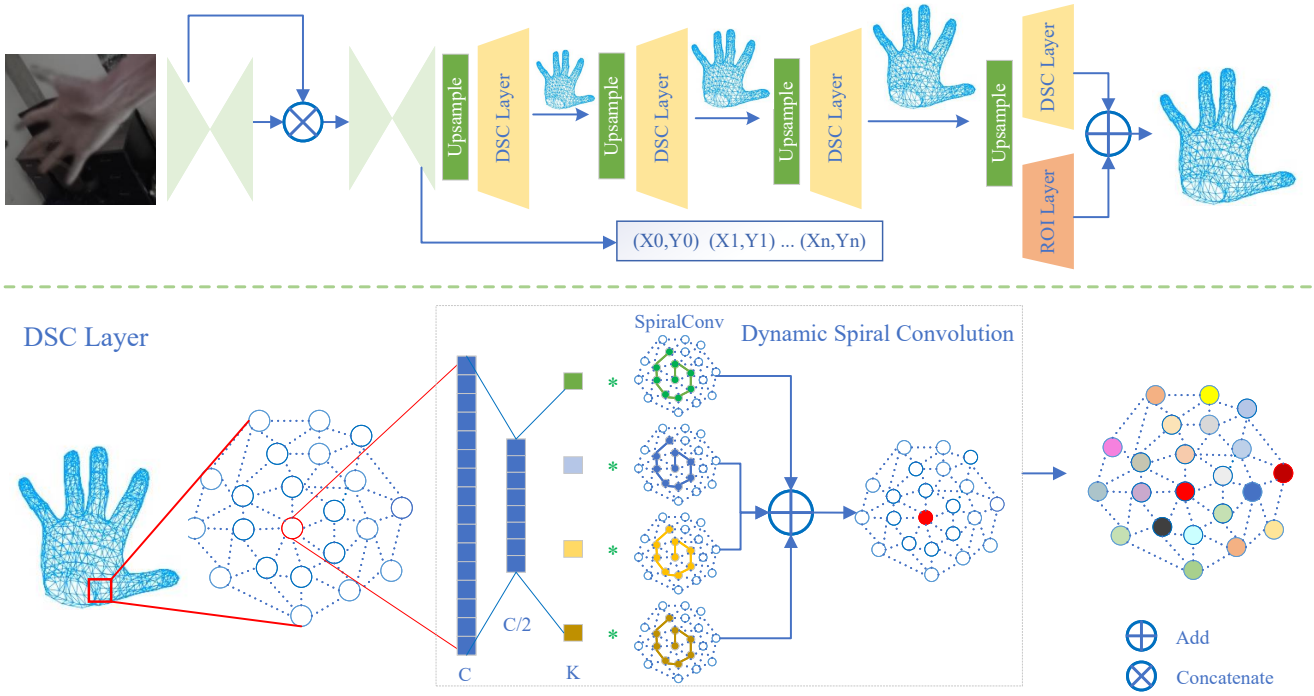

Figure 2. Overview of our M3DHMR framework. The architecture of ROI Layer is a dilated spiral convolution. Different colors in Dynamic Spiral Convolution indicate different weights corresponding to the spiral convolution kernels.

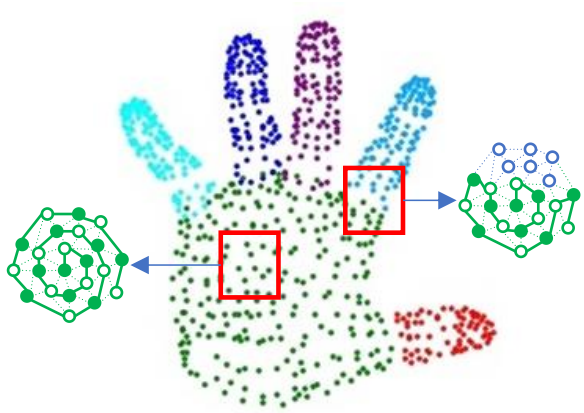

Figure 3. The preset regions for the ROI Layer. Different colors denote different regions. Dilated spiral convolution in ROI layer focuses only on the interior of the region.

for the K convolution kernels:

$$\pi_i(x) = f_2(f_1(x)),$$
$$s.t. 0 \leq \pi_i(x) \leq 1, \sum_{i=1}^{K} \pi_i(x) = 1, \tag{4}$$

where f_1 and f_2 are the two convolution layers, and π_i is the attention weight for the i-th convolutional kernel. The initial convolution layer reduces the dimension to half, while the latter computes the normalized attention weights for the K convolutional kernels, ensuring computational efficiency. Once the K attention weights are obtained, the final convolution weight parameters for that position can be calculated by weighted summation:

$$\hat{W}_1 = \sum_{i=1}^{K} W_1, \hat{b}_1 = \sum_{i=1}^{K} b_1,$$
$$\hat{W}_2 = \sum_{i=1}^{K} W_2, \hat{b}_2 = \sum_{i=1}^{K} b_2. \tag{5}$$

The final output can be denoted as:

$$DSC(v) = \hat{W}_2(\hat{W}_1(f(k{-}disk(v))) + \hat{b}_1) + \hat{b}_2. \tag{6}$$

3.3. Region of Interest Layer

The hand is artificially divided into six regions: palm, thumb, index finger, middle finger, ring finger, and little finger based on the shape and motion characteristics (See Fig. 3). We believe that the mesh vertices in the same region are highly correlated. The movement of the vertices strongly

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

affects others within the same region but has less effect on vertices in other regions. For example, movement at the joint of the index finger causes significant displacement at the tip of the finger, but has minimal impact on the palm and other fingers. Therefore, we assume that for each vertex, the finger or palm region where it is located will be the region of interest. Benefiting from the large receptive field of dilated spiral convolution [11], ROI Layer can sufficiently explore the global features in each hand region of convolution centre to refine the mesh in predefined hand regions.

3.4. Loss Function

We adopt L_1 loss to the 3D mesh loss L_{mesh} and 2D pose loss L_{pose2D}. Inspired by [9], we adopt the normal loss L_{norm} and the edge-length loss L_{edge} to preserve the surface normals and penalize flying vertices. Additionally, we conduct consistency supervision in both 3D and 2D space similar to [14] and [32]. Formally, we have

$$\begin{aligned}
\mathcal{L}_{mesh} &= ||V - V^*||_1, \\
\mathcal{L}_{pose2D} &= ||L^p - L^{p,*}||_1, \\
\mathcal{L}_{norm} &= \sum_{c \in C} \sum_{(i,j) \subset c} |\frac{V_i - V_j}{||V_i - V_j||_2} \cdot n_c^\star|, \\
\mathcal{L}_{edge} &= \sum_{c \in C} \sum_{(i,j) \subset c} |||V_i - V_j||_2 - ||V_i^\star - V_j^\star||_2|, \\
\mathcal{L}_{con3D} &= ||V_{view1} - V_{view2}||_1, \\
\mathcal{L}_{con2D} &= ||L^p_{view1} - L^p_{view2}||_1,
\end{aligned} \tag{7}$$

where C, V are face and vertex sets of a mesh; $n_c^\star$ indicates unit normal vector of face c; and $*$ denotes the ground truth.

Our overall loss is $\mathcal{L}_{total} = \mathcal{L}_{mesh} + \mathcal{L}_{pose2D} + \mathcal{L}_{norm} + \mathcal{L}_{edge} + \mathcal{L}_{con3D} + \mathcal{L}_{con2D}$.

4. Experiments

4.1. Datasets and Metric

FreiHAND is the current largest hand dataset for hand pose and shape estimation from single color images. It contains 130,240 training images and 3,960 evaluation samples. The data of FreiHAND is obtained from real-world scenarios in situations of complex pose, self-occlusion and object occlusion. Meanwhile, it is the only dataset that is annotated with both high-precision 3D keypoints and mesh models.

We use the following metrics in quantitative evaluations.

PA-MPJPE/MPVPE ignores global variations and measures the mean per joint/vertex position error in terms of Euclidean distance (mm) between the root-relative prediction and the ground truth coordinates based on Procrustes analysis [12]. For brevity, the metric is abbreviated as PJ/PV.

F-Score is the harmonic mean between precision and recall between two meshes *w.r.t.* a specific distance threshold, which enables a balanced assessment of the model's performance, particularly useful when the distribution is unbalanced. F@5/F@15 denotes to a threshold of 5mm/15mm.

FPS measures the number of frames an algorithm can process per second.

	PJ↓	PV↓	F@5↑	F@15 ↑
Baseline	6.8	7.0	0.720	0.967
w/o DSC Layer	6.8	6.9	0.723	0.977
w/o ROI Layer	6.7	6.8	0.728	0.978
Ours	**6.6**	**6.7**	**0.734**	**0.978**

Table 1. Ablation study of DSC Layer and ROI Layer with Kernel Size of 8. "Baseline" denotes the model with only Spiral Convolution Layers.

4.2. Experiment Setup

We set the batch size to 32 and train 38 epochs with Adam optimizer [19]. The learning rate is initialized to 0.001, which is divided by 10 at 30th epoch. Data augmentation includes random box scaling, box rotation, box shift and color jitter. We use PyTorch as the framework to conduct all experiments and use a single NVIDIA RTX 3090 for training and a single NVIDIA RTX 2080 Ti for inference.

4.3. Ablation Study

DSC Layer and ROI Layer Tab. 1 illustrates the influence of the two layers on the network performance. Compared to the DSC Layer, ROI Layer employs dilated spiral convolution with a larger convolution kernel, thereby enabling a larger receptive field. Furthermore, by limiting the receptive field to the hand region where the convolution centre is located, ROI Layer mitigates the impact of mesh vertices with low correlation across regions. As can be seen from the first and second rows of Tab. 1, the ROI Layer results in a decrease in PV by 0.1 mm, an increase in F@5 by 0.003, and an increase in F@15 by 0.01. The DSC Layer employs a mechanism of adjusting the weights, which allows the convolution to adapt to hand features in different positions. As shown in the first and third rows of Tab. 1, the performance of the dynamic spiral convolution is superior to that of the normal spiral convolution, resulting in a decrease in PJ and PV by 0.1mm and 0.2mm, respectively, and an increase in F@5 and F@15 by 0.008 and 0.011, respectively. The combination of both layers yields the best reconstruction results, as can be seen in the last row of the table.

Methods of 3D Decoding As illustrated in Tab. 2, Dynamic Spiral Convolution decreases PJ and PV of the 3D decoder and increases F@5 compared with SpiralConv++ and Depth-Separable SpiralConv. The results demonstrate

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

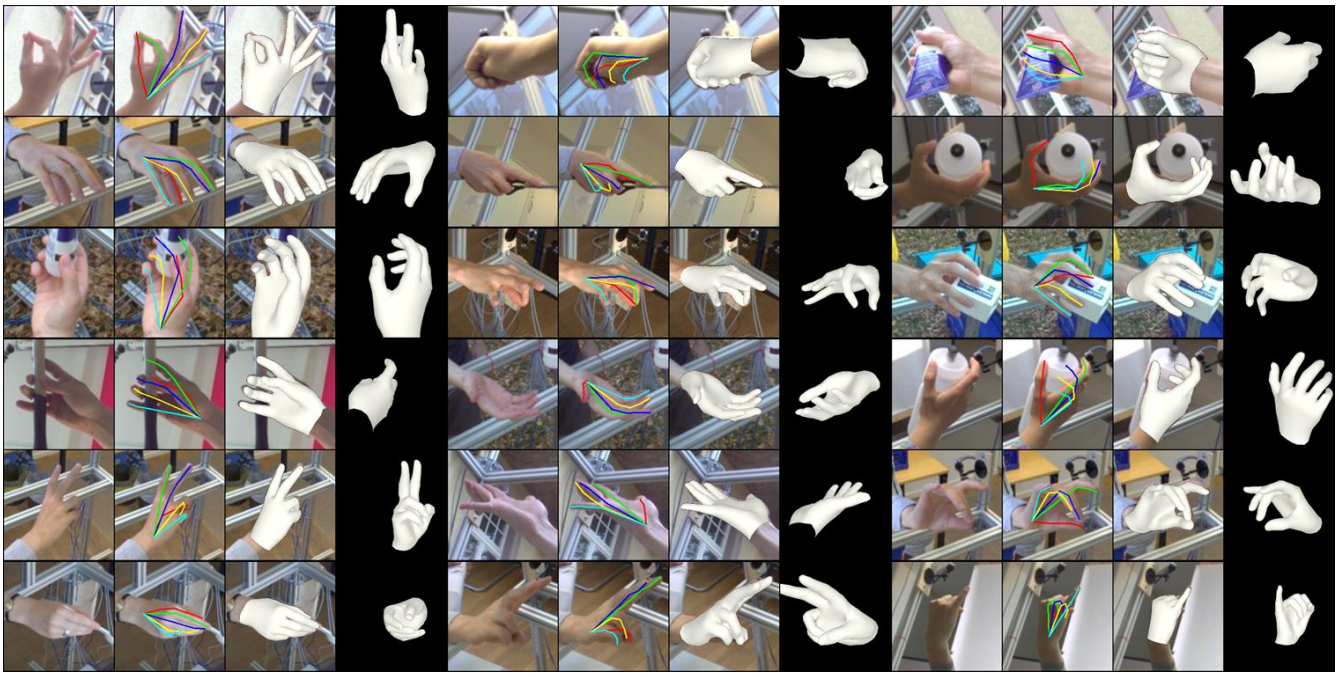

Figure 4. Several qualitative results of our predicted 2D pose, front view mesh and side view mesh on the test set of FreiHAND.

3D decoding	PJ↓	PV↓	F@5↑	F@15 ↑
SpiralConv++[5]	6.8	6.9	0.723	0.976
Depth-Separable SpiralConv[3]	6.8	6.9	0.724	0.977
Dynamic Spiral Convolution (K=8)	**6.6**	**6.7**	**0.734**	**0.978**

Table 2. Ablation study of 3D decoding.

K	PJ↓	PV↓	F@5↑	F@15 ↑
2	6.7	6.9	0.728	0.976
4	6.7	6.8	0.730	0.976
8	**6.6**	**6.7**	**0.734**	**0.978**
16	6.7	6.8	0.731	0.977

Table 3. Ablation study of Dynamic Spiral Convolution.

that adaptively adjusting the convolution weights for vertices at different positions significantly enhances the reconstruction outcomes.

Kernel Size of Dynamic Spiral Convolution Tab. 3 demonstrates the effect of DSC Layers at different values of K. It can be observed that the model exhibits optimal performance with K=8. A value that is too small may result in limited expression and insufficient adaptive ability. Conversely, a value that is too large may lead to an excessive focus on detailed features, which may result in the neglect of some main features and an adverse effect on the model's generalization ability.

4.4. Visualization results of our M3DHMR

As shown in Fig. 1, the qualitative evaluation results on the wild data demonstrate that our approach achieves high reconstruction accuracy in the situation of complex pose (first and second rows), self-occlusion (third and fourth rows) and object occlusion (fifth and sixth rows).

We randomly select some samples from the test set of FreiHAND and show the comprehensive qualitative results of our predicted 2D pose, front view mesh and side view mesh in Fig. 4. Overcoming challenges including complex poses, object occlusion, truncation and poor lighting, our method generates accurate 2D poses and 3D meshes. More results are shown in Fig. S1 in the Supplemental Material.

4.5. Comparisons with Contemporary Methods

In Tab. 4, we compare our method with several state-of-the-art approaches on hand mesh recovery task using the FreiHAND dataset. Since the implementation of Stacked-ResNet in MobRecon is not available, we implement Stack-ResNet18 and fully reproduce the original results of MobRecon. Considering that Stack-ResNet34 can be achieved by adjusting hyperparameters of Stack-ResNet18, we se-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

	Methods	Venue	Backbone	PJ↓	PV↓	F@5↑	F@15 ↑	FPS↑
Non Real-Time Methods	METRO [22]	CVPR 21	HRNet	6.7	6.8	0.717	0.981	-
	MeshGraphormer [23]	ICCV 21	HRNet	5.9	6.0	0.764	0.986	-
	PointHMR [18]	CVPR 23	HRNet	6.1	6.6	0.720	0.984	-
	HaMeR [27]	CVPR 24	ViTPose	6.0	5.7	0.785	0.990	-
	HaMeR-170k [27]	CVPR 24	ViTPose	6.1	5.8	0.782	0.990	-
Real-Time Methods	FreiHAND [35]	ICCV 19	ResNet50	11.0	10.9	0.516	0.934	-
	ObMan [17]	CVPR 19	ResNet18	-	13.2	0.436	0.908	20
	Boukhayma et al. [1]	CVPR 19	ResNet50	-	13.0	0.435	0.898	-
	MobileHand [13]	ICONIP 20	MobileNetV3	-	13.1	0.439	0.902	-
	YoutubeHand [20]	CVPR 20	ResNet50	8.4	8.6	0.614	0.966	-
	I2L-MeshNet [25]	ECCV 20	ResNet50	7.4	7.6	0.681	0.973	33
	Pose2Mesh [7]	ECCV 20	HRNet	7.7	7.8	0.674	0.969	22
	HIU-DMTL [33]	ICCV 21	Hourglass	7.1	7.3	0.699	0.974	-
	CMR [5]	CVPR 21	ResNet50	6.9	7.0	0.715	0.977	30
	I2UV-HandNet [2]	ICCV 21	ResNet50	6.7	6.9	0.707	0.977	-
	Tang et al. [30]	ICCV 21	ResNet50	6.7	6.7	0.724	**0.981**	39
	MobRecon [4]	CVPR 22	GhostStack	8.8	9.1	0.597	0.960	-
	MobRecon [4]	CVPR 22	DenseStack	6.9	7.2	0.694	0.979	-
	MobRecon [4]	CVPR 22	ResNet18	6.7	6.8	0.727	0.979	69
	MobRecon [4]	CVPR 22	ResNet34	6.6	6.7	0.730	0.979	43
	CLIP-Hand3D [15]	ACM MM 23	ResNet50	6.6	6.7	0.728	**0.981**	**77**
	FastViT [31]	ICCV 23	FastViT	6.6	6.7	0.722	**0.981**	-
	M3DHMR	**Ours**	**ResNet18**	**6.6**	**6.7**	**0.734**	0.978	42
	M3DHMR	**Ours**	**ResNet34**	**6.5**	**6.6**	**0.739**	0.980	28

Table 4. Results on the FreiHAND dataset.

lect Stack-ResNet18 and Stack-ResNet34 as backbones for fair comparison between MobRecon and M3DHMR. Our ResNet34-based M3DHMR achieves the best performance among real-time methods on camera-space mesh recovery. Specifically, M3DHMR reaches lowest 6.5mm PA-MPJPE, lowest 6.6mm PA-MPVPE and highest F@5 of 0.739. Additionally, its F@15 of 0.980 also outperforms most compared approaches, only slightly lower than CLIP-Hand3D and FastViT. Furthermore, a qualitative comparison of our M3DHMR with two other state-of-the-art real-time methods is presented in Fig. 5. It can be observed that our method demonstrates superior performance in terms of detail reconstruction. More results are shown in Fig. S2 in the Supplemental Material.

5. Conclusion and Discussion

In this paper, we provide a monocular 3D hand mesh recovery method named M3DHMR. Firstly, M3DHMR adopts a stacked hourglass structure to introduce 2D cues. Then, a spiral decoder with DSC Layers and a ROI Layer lifts the 2D cues to 3D mesh. By adaptively extracting local features at different positions and fully exploring the global features in each hand region, our model achieves the state-of-the-art performance on real-time 3D hand mesh recovery tasks when evaluated on the FreiHAND dataset.

However, there is still some room for improvement in depth ambiguity. In the future, we intend to explore the potential of incorporating data distribution priors to mitigate depth ambiguity by introducing generative models, such as VAE and diffusion models.

Ethics

This study utilizes the publicly available FreiHAND dataset for monocular 3D hand mesh recovery. The original dataset collection was conducted in accordance with ethical research guidelines, with informed consent obtained from all participants. No additional data collection was performed in this work. We affirm our commitment to responsible data usage practices and emphasize the importance of maintaining ethical vigilance in all applications of this research.

References

[1] A. Boukhayma, R. d. Bem, and P. H. Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019.

[2] P. Chen, Y. Chen, D. Yang, F. Wu, Q. Li, Q. Xia, and Y. Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *Proceedings*

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

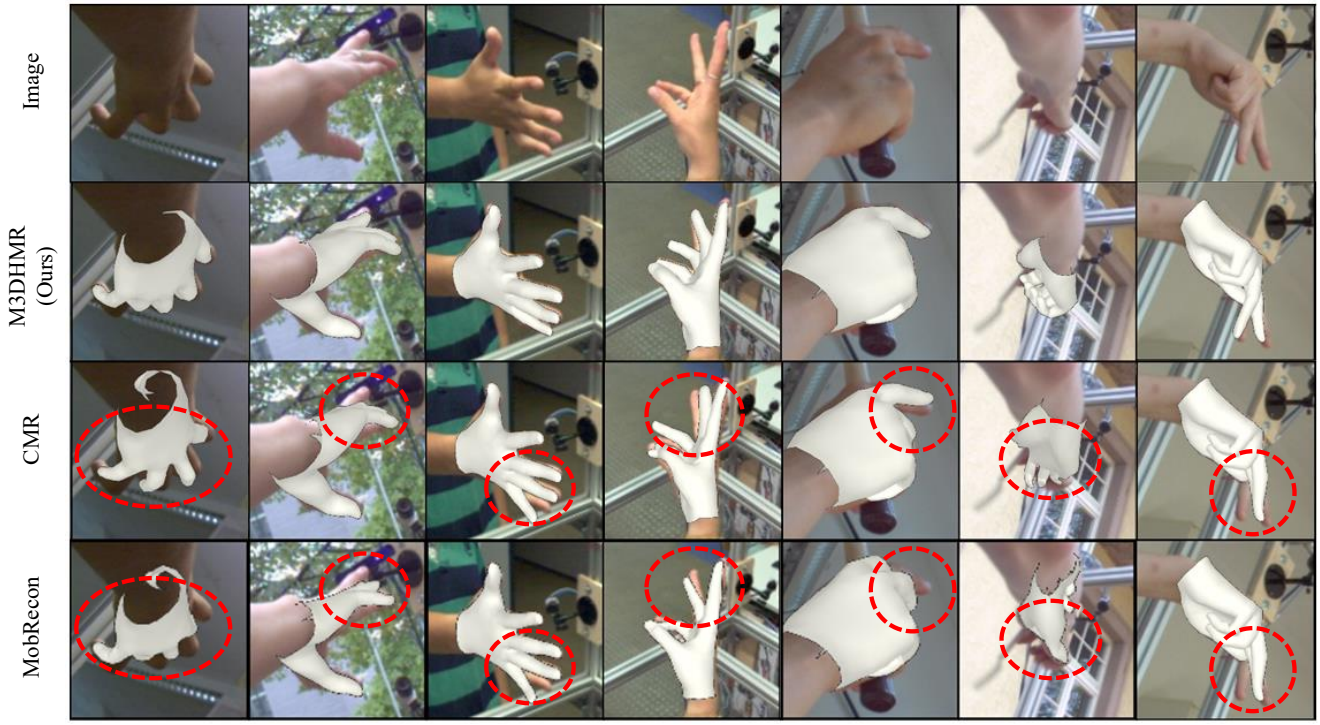

Figure 5. Qualitative comparisons of M3DHMR, CMR and MobRecon for predicting front view mesh on the test set of FreiHAND. The shortcomings of the latter two methods are marked with red circles.

of the IEEE/CVF International Conference on Computer Vision, pages 12929–12938, 2021.

[3] X. Chen, Y. Liu, Y. Dong, X. Zhang, C. Ma, Y. Xiong, Y. Zhang, and X. Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022.

[4] X. Chen, Y. Liu, Y. Dong, X. Zhang, C. Ma, Y. Xiong, Y. Zhang, and X. Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022.

[5] X. Chen, Y. Liu, C. Ma, J. Chang, H. Wang, T. Chen, X. Guo, P. Wan, and W. Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021.

[6] Y. Chen, Z. Tu, D. Kang, L. Bao, Y. Zhang, X. Zhe, R. Chen, and J. Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10451–10460, 2021.

[7] H. Choi, G. Moon, and K. M. Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, 2020.

[8] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[9] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[10] S. Gong, L. Chen, M. Bronstein, and S. Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[11] S. Gong, L. Chen, M. Bronstein, and S. Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Oct 2019.

[12] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[13] L. Guan Ming, J. Prayook, and A. Wei Tech. Mobilehand: Real-time 3d hand shape and pose estimation from color image. In *27th International Conference on Neural Information Processing (ICONIP)*, 2020.

[14] J. Guo, J. Deng, N. Xue, and S. Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. 2018.

[15] S. Guo, Q. Cai, L. Qi, and J. Dong. Clip-hand3d: Exploiting 3d hand pose estimation via context-aware prompting. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023.

[16] S. Han, B. Liu, R. Cabezas, C. D. Twigg, and R. Wang. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics*, 39(4), 2020.

[17] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black,

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.

I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[18] J. Kim, M.-G. Gwon, H. Park, H. Kwon, G.-M. Um, and W. Kim. Sampling is Matter: Point-guided 3d human mesh reconstruction. In *CVPR*, June 2023.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[20] D. Kulon, R. A. Guler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020.

[21] I. Lim, A. Dielen, M. Campen, and L. Kobbelt. A simple approach to intrinsic correspondence learning on unstructured 3d meshes. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[22] K. Lin, L. Wang, and Z. Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021.

[23] K. Lin, L. Wang, and Z. Liu. Mesh graphormer. In *ICCV*, 2021.

[24] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015.

[25] G. Moon and K. M. Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020.

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016.

[27] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[28] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, page 1–17, Dec 2017.

[29] M. Seeber, R. Poranne, M. Polleyfeys, and M. R. Oswald. Realistichands: A hybrid model for 3d hand reconstruction. In *2021 International conference on 3D vision (3DV)*, pages 22–31. IEEE, 2021.

[30] X. Tang, T. Wang, and C.-W. Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11698–11707, 2021.

[31] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[32] L. Yang, S. Chen, and A. Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11364–11373, 2021.

[33] X. Zhang, H. Huang, J. Tan, H. Xu, C. Yang, G. Peng, L. Wang, and J. Liu. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11281–11292, 2021.

[34] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5346–5355, 2020.

[35] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:02:12 UTC from IEEE Xplore. Restrictions apply.