2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

CAMInterHand: Cooperative Attention for Multi-View Interactive Hand Pose and Mesh Reconstruction

Guwen Han[2], Qi Ye[1,†], Anjun Chen[1], Jiming Chen[1]

Abstract— **Interactive hand mesh reconstruction from single-view images poses a significant challenge with the severe occlusion and depth ambiguity inherent in interactive hand gestures. Recent approaches that employ probabilistic models and token-pruned techniques have shown decent results in multi-view human body reconstruction. Nevertheless, these methods have not fully utilized multi-scale semantic information from multi-view images and are not applicable in scenarios involving severe occlusion during dual-hand interactions. Simultaneously, current single-view methods independently reconstruct the left and right hands, which are ineffective in enhancing the interaction between both hands. To address these challenges, we propose CAMInterHand, a cooperative attention-based method for multi-view interactive hand pose and mesh reconstruction. Specifically, CAMInterHand extracts local pyramid features and global vertex features from multi-scale feature maps of multi-view images, enabling the exploration of rich local semantic information and facilitating effective feature alignment. Furthermore, CAMInterHand employs the cooperative attention fusion module to fuse all features from multi-view images, enhancing interactions among vertices of dual hands within global and local contexts. We conduct extensive experiments on the large-scale multi-view dataset InterHand2.6M and CAMInterHand achieves a substantial performance improvement over existing methods for multi-view and single-view interactive hand reconstruction.**

I. Introduction

Interactive hand pose and mesh reconstruction is a fundamental task in many robotic applications, including human-robot interaction [1], [2], XR technologies, and robot grasping [3] and manipulation. With the development of deep learning [4], [5] and the proposing of large-scale datasets [6], [7], interacting hand pose and mesh reconstruction [8]–[11] has made significant advancements. Most prevailing approaches [11], [12] leverage single-view images as input. However, the performance of reconstruction using a single image is still limited by the challenges of substantial occlusions and depth ambiguity posed in interactive hand scenarios. Therefore, fusing multi-view images from different viewpoints is essential to achieve promising interactive hand reconstruction.

Recent work on multi-view human body reconstruction [13]–[17] has yielded exciting results. However, these methods have not fully leveraged multi-level semantic information from multi-view images in feature fusion. Additionally, due to the discrepancy between the human body and hand reconstruction, the direct application of these multi-view human body reconstruction methods to interactive hands will lead to undesirable results. Besides, adopting the strategy of extracting features separately for the right and left hands without interaction, current single-view methods struggle to accurately reconstruct the relative positions of both hands, resulting in the recovery of collisions and incorrect hand shapes.

To address these challenges, in this paper, we present CAMInterHand, a novel framework based on **C**ooperative **A**ttention for **M**ulti-view **Inter**active **Hand** pose and mesh reconstruction without calibration. In our framework, different from previous multi-view methods fusing features from a single layer [14], we resort to leveraging local pyramidal features extracted from multi-scale feature maps of multi-view images to fully explore the local semantic information from each perspective and to address the issue of interactive hand occlusion. In addition to local features, CAMInterHand also extracts global features from various viewpoint images to reinforce the feature alignment and the interaction between global and local contexts. Furthermore, to allow our fusion framework to effectively fuse these multi-scale features from different feature maps and different viewpoints, we formulate our fusion problem into the attention framework by integrating local pyramidal and global vertex-level features using a Transformer-based fusion module. Additionally, in contrast to existing single-view methods that individually reconstruct the left and right hand [12], our framework leverages cooperative attention to interactively fuse features from both hands, further enhancing the reconstruction performance.

We conduct extensive experiments on the large-scale multi-view dataset Interhand2.6M [7]. Our CAMInterHand exhibits favorable reconstruction results and outperforms traditional single-view interactive hand reconstruction methods by a large margin. Additionally, it also significantly outperforms multi-view fusion methods for human body reconstruction. The main contributions of this paper can be summarized as follows:

- We propose CAMInterHand, a novel fusion framework for multi-view interactive hand pose and mesh reconstruction based on the cooperative attention mechanism. The network can effectively utilize the image contextual features from each perspective to handle occlusions and depth ambiguity without camera calibration.
- Our novel notion of integrating local pyramid features

[1]College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China.

[2]College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China.

†Qi Ye (Corresponding author, qi.ye@zju.edu.cn) is with the College of Control Science and Engineeringthe State Key Laboratory of Industrial Control Technology, Zhejiang University, and the Key Key Lab of CS&AUS of Zhejiang Province.

This work was supported in part by the National Natural Science Foundation of China (Grant Number: 62233013, 62088101, 62103372).

and global vertex features extracted from multi-level feature maps fully leverages multi-scale semantic information while also facilitating an effective feature alignment.
- We propose a progressive interactive hand cooperative attention mechanism by employing fusion among vertex-level features of interacting hands, enabling extensive interactions between vertices of different hands.
- CAMInterHand demonstrates excellent performance on the multi-view interactive hand dataset and significantly outperforms state-of-the-art methods.

II. Related Works

A. Interactive Hand Pose and Mesh Reconstruction

Hand reconstruction is one of the key challenges in human motion reconstruction. Rapid progress has been made in interactive hand pose and shape reconstruction from monocular images, thanks to the development of 3D hand parametric models (e.g., MANO [8] and DeepHandMesh [9]) and the emergence of large hand pose and shape datasets (e.g., InterHand2.6M [7]). The simplest way to handle the reconstruction of both hands is to crop out each hand image and then use the single-hand reconstruction method to reconstruct each hand separately and then stitch them back to the human body model [18], [19]. Obviously, this cannot solve the problem of interactions with severe occlusions between interacting hands. In early monocular interactive mesh recovery methods mainly based on depth sensor fusion, Mueller et al. [20] proposed embedding parameterized hand pose and shape models and deep neural network-based dense correspondence predictors into a suitable energy minimization framework for shape fitting. Smith et al. [21] present the first algorithm capable of tracking high-fidelity hand deformations through highly self-contact and self-occlusion gestures.

Recently, due to the application of deep neural networks such as transformers [22], hand reconstruction based on a single RGB camera has made great progress. Zhang et al. [11] proposed to exploit the interactive context presented by the interacting hands to propose a context-aware cascaded refinement network. Li et al. [12] utilize pyramid image features and self-attention modules to directly regress the vertices of interacting hands. Yu et al. [23] used interactive hand context collaborative attention to predict MANO [8] parameter models. These methods do not utilize global attention between interacting hands, which leads to depth ambiguities between interacting hands and an inability to correctly reconstruct their relative positions, including collisions and incorrect shapes.

B. Multi-View Human Pose and Mesh Restoration

For camera calibration to be available, it is usually achieved by enforcing consistency in the estimated poses across multiple viewpoints to reconstruct the human body [24], [25]. For example, Iskakov et al. [26] proposed two multi-view 3D humans pose estimation solutions based on learnable triangulation methods, which combine 3D information from multiple 2D views. He et al. [27] proposed a differentiable core-polar transformer that enables 2D detectors to exploit 3D perceptual features to improve 2D pose estimation. These methods ignore the feature semantic information in multi-view images.

For camera calibration not available, the reconstruction of the human body is typically performed by fusing the features extracted from multiple viewpoints [15], [17]. For example, Xie et al. [28] proposed a pre-trained cross-view fusion model that can strictly adapt to unknown camera poses with only a small amount of labeled training data. Li et al [13] proposed a modified multi-view optimization method (MV-SMPLify) based on the SMPLify method to simultaneously fit SMPL models to multi-view images. Kolotouros et al. [14] proposed probabilistic modeling of 3D human mesh recovery. Ma et al. [29] proposed token-Pruned Pose Transformer. To block unimportant image tokens. Jia et al. [16] proposed pixel alignment feedback fusion for accurate and efficient human mesh recovery from multi-view images. However, these methods do not exploit the multi-level semantic features of multi-view images in the fusion of image features.

III. Methodology

A. Problem Statement and Method Overview

In this section, we present the technical details of our proposed method, CAMInterHand, for multi-view image interactive hand pose and mesh reconstruction. Fig. 1 illustrates the framework of CAMInterHand. Multi-view images are initially processed through encoders and decoders to extract both global and local features. Multi-view global features are concatenated into a single vector $G_F \in \mathbb{R}^{1536}$, while local features are fed into a multi-view image pyramid feature extractor, generating multi-view image pyramid features. These global features are then combined with local pyramid features and utilized in a progressive interactive hand cooperative attention fusion module to reconstruct both left-hand and right-hand meshes.

We use non-parametric methods to reconstruct the left-hand mesh $M_L \in \mathbb{R}^{778\times3}$ and the right-hand mesh $M_R \in \mathbb{R}^{778\times3}$, with 778 vertices in the left and right hands respectively. Subsequently, the 3D joint $J_{3D} = R(M)$ can be regressed from the output mesh using the trained MANO parameters, where R is the pre-trained linear regressor, $M \in \mathbb{R}^{778\times3}$.

B. Multi-View Image Pyramid Feature and Global Feature Extraction

To better leverage the multi-view image semantic information, we introduce the concept of multi-view image pyramid features and global feature extraction. This innovative strategy involves extracting local pyramid features and global vertex features from multi-level feature maps, enabling comprehensive utilization of multi-scale semantic information and facilitating effective feature alignment.

Multi-View Image Encoder-Decoder: We utilize the Swin Transformer [30] encoder to take multiple views as

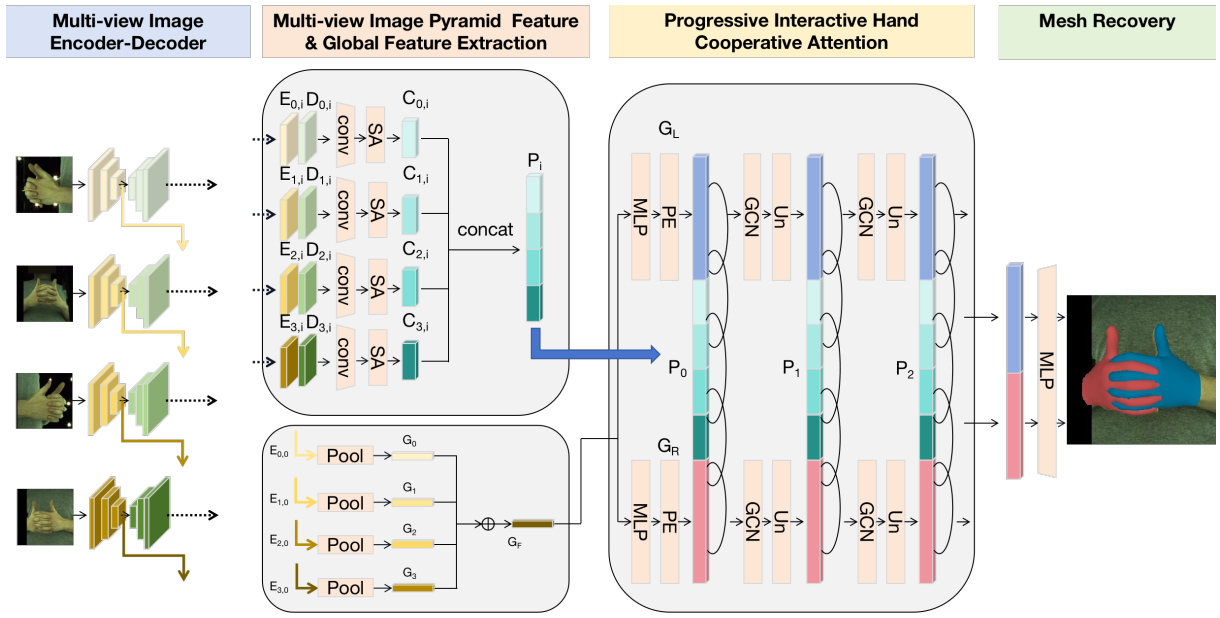

Fig. 1. Overview of our proposed CAMInterHand pipeline. Initially, multi-view images are independently input into encoders and decoders to facilitate the extraction of global and local image features. Subsequently, the global features originating from the multi-view images are merged into a global vector denoted as G_F. In parallel, the local features extracted from the multi-view images are directed into the multi-view image pyramid feature extractor, which captures the multi-view image pyramid feature. Finally, the global features are concatenated with the local pyramid features and input into the progressive interactive hand cooperative attention, facilitating the fusion of local and global features to reconstruct the meshes for the left and right hands.

input and encode them into features. These features are then fed to a simple decoder composed of several deconvolution layers to recover high-resolution feature maps. The final layer of the decoder is responsible for reconstructing dense maps, heatmaps, and 2D mask information, which are used for 2D supervised pretrain.

Multi-View Image Pyramid Features and Global Features Extraction: As shown in Fig. 1, we concatenate the feature maps of the same scale from the image encoder outputs E and decoder outputs D from four views. These feature maps are then separately input into convolution (Conv) layers and self-attention (SA) layers, further output the attention features $C_{v,i}$ for each view v, where v represents the view id totaling 4 perspectives, and i indicates the different scale features, ranging from 1 to 3. This process is formulated as shown in (1). Then, we concatenate the attention features $C_{v,i}$ of the same scale from the four views and output pyramid attention features P_i ($P_0 \in \mathbb{R}^{8\times8\times256}$, $P_1 \in \mathbb{R}^{16\times16\times256}$,$P_2 \in \mathbb{R}^{32\times32\times256}$) at three different scales, as shown in (2). As a result, the bottom-level pyramid attention features capture coarse-grained texture semantic information from multiple views, while the top-level pyramid attention features capture fine-grained semantic context from multiple views.

$$C_{v,i} = SA(Conv(concat(E_{v,i}, D_{v,i}))), \tag{1}$$

$$P_i = concat(C_{0,i}, C_{1,i}, C_{2,i}, C_{3,i}). \tag{2}$$

For global feature extraction, as depicted in Fig. 1, we extract multi-view features $E_{0,0}$, $E_{1,0}$, $E_{2,0}$, $E_{3,0}$ from the last layer of the encoder. These features are individually fed into pooling layers $fp()$ to generate global feature vectors for each perspective: G_0, G_1, G_2, and G_3. These perspective global feature vectors are then integrated to generate a single global feature G_F, as shown in (3).

$$G_F = sum(fp(E_{0,0}), fp(E_{1,0}), fp(E_{2,0}), fp(E_{3,0})). \tag{3}$$

C. Progressive Interactive Hand Cooperative Attention

Handling the relative depth ambiguity between interacting hands is a challenging task, which requires an efficient fusion of features between interacting hands. Excitingly, the multi-head self-attention (MHSA) module [22] is good at modeling relationships between information tokens and handling disordered, heterogeneous, and variable-length data structures effectively. Therefore, we propose a progressive interactive hand cooperative attention method to solve the above challenge. We ingeniously devise a fusion module to merge the vertex features of both hands with multi-view image features.

Firstly, the global feature G_F is input to a linear layer to reduce its dimensions. Subsequently, position encoding (PE) is applied separately using the left-hand and right-hand MANO templates, developing the left-hand feature $G_L \in \mathbb{R}^{98\times256}$ and right-hand feature $G_R \in \mathbb{R}^{98\times256}$. Next, the left-hand vertex features G_L, right-hand vertex features G_R, and multi-view image pyramid features P_0 are concatenated and fed into the multi-head self-attention module to output enhanced vertex features $G_L^{'}$ and $G_R^{'}$ using equation (4). The MHSA enables the fusion of vertex features with multi-level semantic information from local features at multi-view,

facilitating the fusion of vertex features from arbitrary points between both hands.

$$G_L^{'}, G_R^{'} = MHSA(G_L, Flatten(P_0), G_R). \tag{4}$$

We utilize graph convolution layers (GCN) [31] within each hand to enhance feature fusion among single-hand vertices. Each graph convolutional block consists of two layers of spiral convolution, and the feature dimension remains unchanged after the GCN. To accelerate convergence, we employ a progressive vertex regression approach to estimate vertex positions from coarse to fine. In specific, we employ the approach in Mesh Autoencoder [32] to upsample the number of vertices. The number of vertices of the three-layer interactive hand cooperative attention are V_0= 98, V_1=195, V_2=389 respectively. Through multiple dimension-reduction layers of MHSA and GCN, the last MLP layer directly regresses the coordinates of the interactive hand vertices.

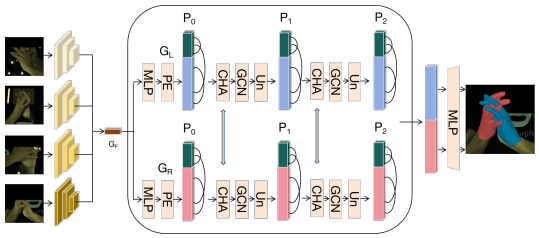

Fig. 2. Single-hand attention method framework: Unlike the interactive hand cooperative attention, the single-hand attention method connects the vertex features of the left and right hands with image features separately and then applies attention to each of them.

Compared to the IntagHand [12] approach, of which the framework is shown in Fig. 2, single-hand vertex features are independently fused with image features for attention. Despite incorporating cross-hand attention, it primarily focuses on local interactions within individual hands while overlooking interactive hand interactions. Our approach employs cooperative attention between interacting hands and multi-view pyramid features, which strengthens global interactions among different hand vertices, thereby implicitly improving the capacity to learn ambiguous relative depth between interacting hands.

D. Loss Functions

We train the multi-view method CAMInterhand with 2D loss function and 3D loss function. Concretely, to train the image encoder-decoder, we use L1 loss to supervise the 2D dense map and use the mean square error loss to supervise 2D heatmaps and masks. To train 3D regression vertices, we utilize vertex loss, joint loss, mesh normal loss, and edge length loss.

Vertex Loss. We use the L1 loss function to supervise the 3D coordinates of the interacting hand vertices.

$$L_V = \sum_{i}^{n} \|V_{h,i} - V_{h,i}^*\|_1, \tag{5}$$

where $V_{h,i}$ is the vertex i, $h \in \{left, right\}$ indicating the left or right hand. $*$ represents the ground truth.

Joint Loss. The predicted hand vertices V are multiplied by the pre-trained linear joint regressor J to obtain the regressed hand joints. Our joint loss function error is computed as follows:

$$L_J = \sum_{h \in \{left, right\}} \|JV_{h,i} - JV_{h,i}^*\|_2. \tag{6}$$

Normal Loss. To make the prediction mesh smoother, we use normal loss by aligning the predicted mesh surface normal with the ground truth mesh normal direction.

$$L_n = \sum_{f=1}^{F} \sum_{e=1}^{3} \|e_{f,i,h} \cdot n_{f,h}^*\|_1, \tag{7}$$

where f is the face of the reconstructed mesh, e represents the edge vector of each face, and each face is composed of three edge vectors. n^* denotes the normal vector of the ground truth mesh.

Edge Loss. We use edge length loss to keep the predicted edge lengths consistent with the ground truth edge lengths.

$$L_e = \sum_{f=1}^{F} \sum_{e=1}^{E} \|e_{f,i,h} - e_{f,i,h}^*\|_1, \tag{8}$$

where E=3 represents the number of edges on each face.

IV. Experiments

A. Experimental Settings

Dataset. InterHand2.6M [7] is a large-scale real-captured dataset with human(H) and machine(M) 3D pose and mesh annotations. The dataset is divided into two subsets, the interactive hands (IH) and the single hands (SH). Since we focus on multi-view interactive hand reconstruction, four camera views (400009, 400018, 400031, 400053) are selected from the 5 FPS IH subset with H+M annotations as the multi-view dataset. Invalid labels are discarded based on valid annotations of *hand_type_valid*. In total, 20,132 training samples and 1,356 testing samples from InterHand2.6M were used. **Evaluation Metrics.** To evaluate the pose and shape accuracy of the reconstructed hands, we compare the Mean Per Joint Position Error (MPJPE) and the Mean Per Vertex Position Error (MPVPE) in millimeters. We scaled the mid-metacarpal length of each hand to 9.5cm during training, aligned the root joint of each hand during evaluation, and rescaled back to true skeletal length.

Implementation Details. For the backbone network, we employ the Swin Transformer [30] pre-trained on ImageNet [5] as the image encoder. We use the Adam optimizer to train the model with an initial learning rate of 1e-4 and multiply the learning rate by 0.1 after every 200 epochs.

B. Experimental Results

Comparison to State-of-the-art Methods. We compare our CAMInterHand with state-of-the-art monocular and multi-view reconstruction methods, as shown in Table I. Interhand2.6m comprises over 70 camera viewpoints, while

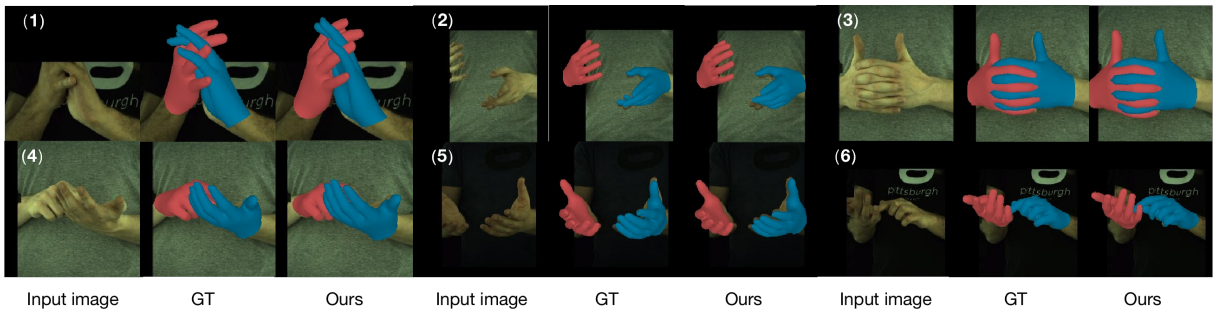

Fig. 3. Comparison with ground truth on InterHand 2.6M [7] test dataset. Our method produces good results in hand reconstruction in challenging situations such as truncation (1-2), severely occluded hands (3-4), and severely bent fingers (6).

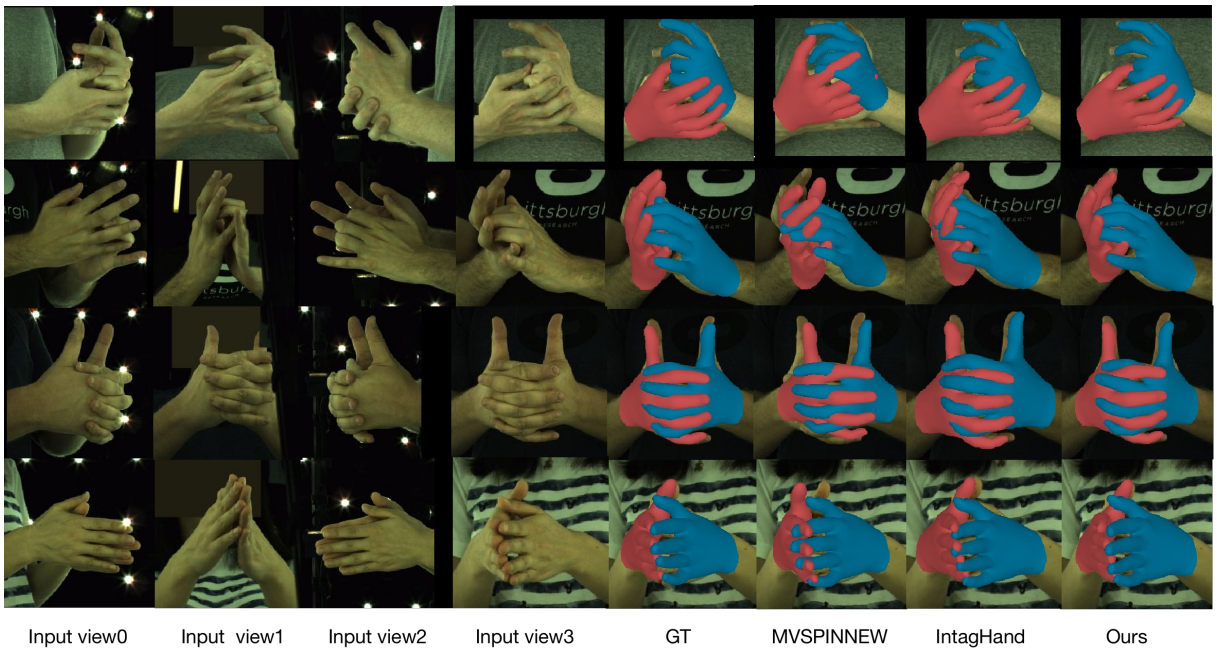

Fig. 4. Qualitative comparison with state-of-the-art monocular and multi-view reconstruction methods on the InterHand2.6M dataset. The four columns on the left are four input images for multi-view reconstruction and input view 3 for monocular reconstruction. We show the reconstruction results of view 3. Our method yields superior reconstruction results in various challenging interaction situations, while the MVSPINNEW [13] and IntagHand [12] methods exhibit severe collisions and significant deviations in hand shapes and the relative distance between the hands.

TABLE I. Comparison with state-of-the-art methods on interhand2.6m [7]. (-) indicates monocular reconstruction.

	MPJPE	MPVPE
(-) IntagHand [11]	8.76	9.04
(-) IntagHand [11] (all view)	6.63	6.86
MVSPINNEW [13]	11.22	11.49
ours(w/o pre-trained)	6.58	6.79
ours(w/ pre-trained)	**5.65**	**5.87**

we select a subset of four camera viewpoints, as detailed in the experimental settings. In Table 1, except for our CAMInterhand with pre-trained and IntagHand (all view) methods, which were trained on the all viewpoints interhand2.6m dataset, all other methods were trained on the subset of data from four camera viewpoints. All methods were tested on the subset dataset.

Compared to monocular reconstruction methods, our model achieves 6.58 MPJPE and 6.79 MPVPE when trained on the Interhand2.6M sub-dataset from four camera viewpoints. These results represent a significant improvement of 24.9% over the previous best method (IntagHand). Additionally, while compared to IntagHand trained using all camera viewpoints from the Interhand2.6M dataset, our CAMInterHand is pre-trained on all camera viewpoints and fine-tunes on a subset of four camera viewpoints, achieving 5.65 MPJPE and 5.87 MPVPE, which is significantly better than intaghand's 6.63 MPJPE and 6.86 MPVPE.

We further compare our method to the state-of-the-art multi-view reconstruction approach, MVSPINNEW [13], which is an improved multi-view optimization method based on the SMPLify framework, allowing the fitting of the SMPL model to multiple-view images simultaneously. To ensure a fair comparison, we run the code published by MVSPINNEW on the multi-view InterHand dataset to obtain test results. Our method achieves the lowest MPJPE of 5.65 and MPVPE of 5.87 on the Multi-View InterHand

dataset, significantly outperforming the recent multi-view reconstruction methods.

Qualitative Evaluation. Our qualitative results on InterHand2.6M are shown in Fig. 3 and Fig. 4. Notably, in challenging situations such as truncated hands, severe occlusions, and heavily bent fingers, our method consistently produces reconstructions that are close to the ground truth. Fig. 3 provides examples of such cases. Furthermore, compared to previous state-of-the-art methods, our approach accurately reconstructs various occlusions and interactive actions in the input images, avoiding collisions between the hands, as shown in Fig. 4.

C. Ablation Study

TABLE II. Ablation study on the selection of pyramid feature scales for image pyramid features (IPF).

	MPJPE	MPVPE
Ours(w/o IPF)	7.74	7.59
Ours + IPF-8	6.73	6.98
Ours + IPF-32	6.46	6.66
Ours + IPF	5.65	5.87

TABLE III. Ablation study on view number selection of image pyramid feature(IPF).IPF_v represents the number of viewpoints.

	MPJPE	MPVPE
Ours($IPF_v = 1$)	6.62	6.87
Ours($IPF_v = 2$)	6.13	6.31
Ours($IPF_v = 3$)	5.83	6.09
Ours($IPF_v = 4$)	5.65	5.87

TABLE IV. Analysis of interactive hand cooperative attention (interactive) and single hand attention method (single).

	MPJPE	MPVPE
single	5.97	6.19
interactive	5.65	5.87

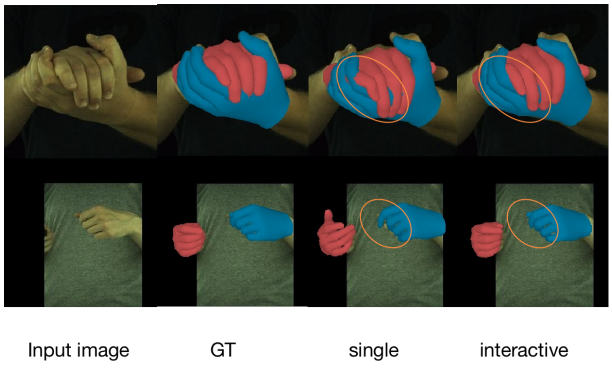

Fig. 5. Qualitative analysis of interactive hand cooperative attention (interactive) and single-hand attention methods (single), interactive hand cooperative attention can avoid collisions and align with images.

Effectiveness of the Multi-View Image Pyramid Feature Extraction: We compare four experimental settings of pyramid feature structures: removal of multi-view image pyramid features (w/o IPF), only using multi-view image pyramid features at one resolution (8*8 or 32*32), and using complete multi-scale Multi-view image pyramid features. The above four pyramid structures are independently input to the interactive hand cooperative attention fusion module, as shown in Table II. It is evident that the removal of the image pyramid feature structure leads to a significant decline in reconstruction accuracy, thereby underscoring the effectiveness of the multi-view image pyramid feature extraction. Additionally, the complete multi-scale multi-view image pyramid feature setting, results in the highest attainable reconstruction accuracy, providing compelling evidence for the efficacy of multi-scale pyramid feature extraction. The multi-view image pyramid feature extractor effectively captures low-level coarse-grained texture features as well as high-level fine-grained semantic attributes within the image.

Effectiveness of the Number of Views in the Image Pyramid Feature: We concatenate different numbers of image pyramid features for interactive hand cooperative attention fusion, as shown in Table III, it is evident that as the number of views in the image pyramid feature increases, the reconstruction error continually decreases, our interactive hand cooperative attention fusion module effectively integrates image semantic information from different numbers of perspectives.

Effectiveness of Progressive Interactive Hand Cooperative Attention: In the Methods section, we analyze the distinctions between the interactive hand cooperative attention and single-hand attention methodologies. The specific network structure is shown in Fig. 1 and Fig. 2. Quantitative experiments are shown in Table IV. The experimental results demonstrate that interactive hand cooperative attention leads to lower reconstruction errors compared to single-hand attention. This implies that progressive interactive hand cooperative attention can capture global interactions between arbitrary vertices of both hands, effectively improving the relative depth estimation between interacting hands. Qualitative analysis is shown in Fig. 5. Interactive hand cooperative attention can prevent collisions between both hands and align them with the image.

V. Discussion

In this paper, we propose CAMInterHand for reconstructing the interactive hand mesh from multi-view images. Specifically, we introduced a multi-view image pyramid feature extraction to fully exploit multi-level semantic features from multiple viewpoints. Additionally, we presented progressive interactive hand cooperative attention, allowing attention interaction between arbitrary two vertex features of both hands to capture global information. This method is dedicated to interactive hand reconstruction using calibrated fixed-position cameras. In future work, we will explore multi-view reconstruction using uncalibrated cameras placed at arbitrary positions.

REFERENCES

[1] Y. Shu, Z. Li, B. F. Karlsson, Y. Lin, T. Moscibroda, and K. Shin, "Incrementally-deployable Indoor Navigation with Automatic Trace Generation," in *IEEE Conference on Computer Communications (INFOCOM)*, 2019.

[2] Y. Shu, C. Bo, G. Shen, C. Zhao, L. Li, and F. Zhao, "Magicol: Indoor Localization Using Pervasive Magnetic Field and Opportunistic WiFi Sensing," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 7, pp. 1443–1457, Jul. 2015.

[3] Q. Liu, Y. Cui, Q. Ye, *et al.*, "Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 3153–3160.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[6] 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*, 2019, pp. 813–822.

[7] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, "Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, Springer, 2020, pp. 548–564.

[8] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.

[9] X. Chen, Y. Liu, C. Ma, *et al.*, "Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 274–13 283.

[10] H. Zhang, Y. Tian, X. Zhou, *et al.*, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 446–11 456.

[11] B. Zhang, Y. Wang, X. Deng, *et al.*, "Interacting two-hand 3d pose and shape reconstruction from single color image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 354–11 363.

[12] M. Li, L. An, H. Zhang, *et al.*, "Interacting attention graph for single image two-hand reconstruction," *arXiv preprint arXiv:2203.09364*, 2022.

[13] Z. Li, M. Oskarsson, and A. Heyden, "3d human pose and shape estimation through collaborative learning and multi-view model-fitting," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1888–1897.

[14] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 605–11 614.

[15] Z. Yu, L. Zhang, Y. Xu, *et al.*, "Multiview human body reconstruction from uncalibrated cameras," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7879–7891, 2022.

[16] K. Jia, H. Zhang, L. An, and Y. Liu, "Delving deep into pixel alignment feature for accurate multi-view human mesh recovery," *arXiv preprint arXiv:2301.06020*, 2023.

[17] A. Chen, X. Wang, K. Shi, *et al.*, "Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 2752–2758.

[18] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, Springer, 2020, pp. 20–40.

[19] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *2021 International Conference on 3D Vision (3DV)*, IEEE, 2021, pp. 792–804.

[20] F. Mueller, M. Davis, F. Bernard, *et al.*, "Real-time pose and shape reconstruction of two interacting hands with a single depth camera," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 4, pp. 1–13, 2019.

[21] B. Smith, C. Wu, H. Wen, *et al.*, "Constraining dense hand surface tracking with elasticity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] Z. Yu, S. Huang, C. Fang, T. P. Breckon, and J. Wang, "Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 955–12 964.

[24] H. Rhodin, J. Spörri, I. Katircioglu, *et al.*, "Learning monocular 3d human pose estimation from multi-view images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8437–8446.

[25] Y. Zheng, R. Shao, Y. Zhang, *et al.*, "Deepmulticap: Performance capture of multiple characters using sparse multiview cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6239–6249.

[26] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7718–7727.

[27] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu, "Epipolar transformers," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 7779–7788.

[28] R. Xie, C. Wang, and Y. Wang, "Metafuse: A pre-trained fusion model for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 686–13 695.

[29] H. Ma, Z. Wang, Y. Chen, *et al.*, "Ppt: Token-pruned pose transformer for monocular and multi-view human pose estimation," in *European Conference on Computer Vision*, Springer, 2022, pp. 424–442.

[30] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[31] 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*, 2019, pp. 0–0.

[32] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3d faces using convolutional mesh autoencoders," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 704–720.