DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras

Yang Zheng*, Ruizhi Shao*, Yuxiang Zhang, Tao Yu, Zeron Zheng, Qionghai Dai, Yebin Liu
Department of Automation and BNRist, Tsinghua University

Figure 1: Given only sparse multi-view RGB videos (6 views for the left and middle, 8 views for the right), our method is able to reconstruct various kinds of 3D shapes with temporal-varying surface details even under challenging occlusions for multi-person interactive scenarios.

Abstract

We propose DeepMultiCap, a novel method for multi-person performance capture using sparse multi-view cameras. Our method can capture time varying surface details without the need of using pre-scanned template models. To tackle with the serious occlusion challenge for close interacting scenes, we combine a recently proposed pixel-aligned implicit function with parametric model for robust reconstruction of the invisible surface areas. An effective attention-aware module is designed to obtain the fine-grained geometry details from multi-view images, where high-fidelity results can be generated. In addition to the spatial attention method, for video inputs, we further propose a novel temporal fusion method to alleviate the noise and temporal inconsistencies for moving character reconstruction. For quantitative evaluation, we contribute a high quality multi-person dataset, MultiHuman, which consists of 150 static scenes with different levels of occlusions and ground truth 3D human models. Experimental results demonstrate the state-of-the-art performance of our method and the well generalization to real multiview video data, which outperforms the prior works by a large margin.

1. Introduction

Recent years have witnessed great progress in vision-based human performance capture, which is promising to enable various applications (e.g., tele-presence, sportscast, gaming and mixed reality) with enhanced interactive and immersive experiences. To achieve surprisingly detailed geometry and texture reconstruction, dense camera rigs even equipped with sophisticated lighting systems are introduced [56, 8, 29, 28, 20, 2]. However, the extremely expensive and professional setups limited their popularity. Although other light-weight multi-view human performance capture systems have achieved impressive results even in real-time, they still relies on pre-scanned templates [36, 35], custom-designed RGBD [15, 13] or commercial RGBD [61, 62] sensors, or limited to single-person reconstruction [51, 18, 25, 19].

Benefiting from the fast improvement of deep implicit functions for 3D representations, recent methods [47, 48, 32] are able to recover the 3D body shape only from a single RGB image. Compared with the voxel-based [53, 69] or mesh-based [40, 1] representations, an implicit function guides the deep learning models to notice geometric details in a more efficient way. Specifically, PIFu [47, 32] achieves plausible single human reconstruction using only RGB images, and PIFuHD [48] further utilizes normal maps and high resolution images to generate more detailed results.

Despite the prominent performance in digitizing 3D human body, both PIFu [47] and PIFuHD [48] suffer from several drawbacks when extending the frameworks to multi-person scenarios and multi-view setups. Firstly, the average-pooling-based multi-view feature fusion strategy in PIFu will lead to over-smoothed outputs when high frequency details (e.g., normal maps) are included in multi-
view features. More importantly, in the two approaches, reconstruction results are only promised with ideal input images without severe occlusions in multi-person performance capture scenarios. The reconstruction performance of [47, 48] will be significantly deteriorated due to the lack of observations caused by severe occlusions.

To address the aforementioned problems, we propose a novel framework to perform multi-person reconstruction from multi-view images. First of all, inspired by [54], we design an spatial attention-aware module to adaptively aggregate information from multi-view inputs. The module is effective to capture and merge the geometric details from different view points, and finally contributing to the significant improvement of results under multi-view setups. Moreover, for multi-person reconstruction, we further combine the attention module with parametric models, i.e., SMPL, to enhance the robustness while maintaining the fine-grained details. The SMPL model serves as a 3D geometry proxy which compensates for the missing information where occlusions take place. With the semantic information provided by SMPL, the network is capable of reconstructing complete human bodies even under close interactive scenarios. Finally, when dealing with moving characters from video, we propose a temporal fusion method by weighting the signed distance field (SDF) across the time domain, which further enhance the temporal consistency of the reconstructed dynamic 3D sequences.

Another urgent problem is that the lack of high-quality scans of multi-person interactive scenarios in the community makes it difficult for accurately evaluating multi-person performance capture systems like ours. To fill this gap and better evaluate the performance of our system, we contribute a novel dataset, MultiHuman, which consists of 150 high-quality scans with each containing from 1 to 3 multi-person interactive actions (including both natural and close interactions). The dataset is further divided into several categories according to the level of occlusions and number of persons in the scene, where a detailed evaluation can be conducted. Experimental results demonstrate the state-of-the-art performance and well generalization capacity of our approach. In general, the main contribution in this work can be summarized as follows:

- We propose a novel framework for high-fidelity multi-view reconstruction for multi-person interactive scenarios. By leveraging the human shape and pose prior for resolving the ambiguities introduced by severe occlusions, we achieve the state-of-the-art performance even with partial observations in each view.
- We design an efficient spatial attention-aware module to obtain fine-grained details for multi-view setups, and introduce a novel temporal fusion method to reduce the reconstruction inconsistencies for moving characters from video inputs.

- We contribute an extremely high quality 3D model dataset containing of 150 multi-person interacting scenes. The dataset can be used for training and evaluation of related topics in future research.

2. Related Work

**Single-view performance capture** Many methods have been proposed to reconstruct detailed geometry from single-view inputs. Typical techniques include silhouette estimation [40], depth estimation [17, 50] and template-based deformation [1, 71, 21]. Moreover, SMPL [37] regression or optimization can be incorporated to generate more reliable and robust outputs as shown in [69, 68, 4]. Real-time methods can be implemented with the aid of a single depth sensor [60, 62] or by innovating computation and rendering algorithms [32]. Regarding to the 3D representations used in these methods, we can split them into two categories: explicit [53, 40, 69, 22] and implicit [47, 48, 25, 6, 7, 11, 5, 58, 49] reconstruction methods. Compared with traditional explicit 3D representations, implicit representations show certain advantages in domain-specific shape learning and detail preservation. For example, PIFu [47] define the surface as a level set of function \( f \). Similarly, [25] defines a probability field of surface points, and ARCH [26] predicts a 3D occupancy map. However, all of the methods above are mainly focusing on single-person reconstruction, and it remains difficult for them to achieve accurate reconstruction under multi-person scenarios.

**Multi-view performance capture** Motion capture has been developed to make accurate motion predictions in multi-person interaction scenes [3, 36, 35, 28, 46, 45, 67, 65, 30, 41, 52, 34, 63]. Some of them even achieve real-time performance [6, 12, 66, 39]. However, these works only capture skeleton motions instead of the detailed geometries. Regarding to multi-view geometry reconstruction, previous studies use template-based deforming methods [10, 55, 18], skeleton tracks [55, 18] or fusion-based techniques [14]. Aside from the long computation time, these methods often show deficiency in mapping textures, handling topology changes or dealing with drastic frame-to-frame motion. Moreover, the aforementioned methods also show limited adaptability for multi-person capture as they cannot effectively deal with occlusions. Robust quality reconstruction methods often come at prohibitive dependencies and constraints. Some methods depend on dense viewpoints [8, 29] and even controlled lighting [56, 20] to reconstruct detailed geometry. Another branch of multi-view RGBD systems [15, 13, 61, 70, 42], achieve impressive real-time performance capture results even for multi-person scenarios benefiting from the strong depth observations. Note that Huang et.al. [25] also presents a volumetric capture ap-
Figure 2: Pipeline of our system. With estimated SMPL models and segmented multi-view, we design a spatial attention-aware network and temporal fusion method to reconstruct each character separately. An overview of our approach is illustrated in Fig 2. With input as the segmented multi-view single person images and the corresponding SMPL, the system outputs the reconstructed 3D persons. The results are combined together directly with no need for modifying the relative position, since the multi-view setting ensures the 3D spatial relationship between different individuals.

To obtain the inputs, we firstly fit SMPL-X [43] models through 3D keypoints estimated from multi-view video by a light-weight total capture method [67]. For multi-person segmentation, we refer to a self-correction method [31] and use SMPL projection maps to track different characters in multi-view scenes. Finally, the 3D human can be generated through the spatial attention-aware network based on the pixel-aligned implicit function, and further polished by the temporal fusion method when the time information is available in the video inputs, which will be described detailedly in Section 4.

3. Overview

Our method is based on the implicit function. An implicit function represents the surface of a 3D model as a level set of an occupancy field function F, e.g. F(X) = 0.5. Specifically, PIFu [47] combines 3D points with conditional variables to formulate a pixel-aligned implicit function:

\[ F(\Phi(x, I), z(X)) = s : s \in [0, 1] \] (1)

where for an image I and a given 3D point X, x = \Pi(X) is the 2D projection coordinate on the image plane, z(X) is the depth value in the camera space, and \Phi(x, I) is the image embedding feature at location x. In PIFu, a multi-layer perceptron (MLP) is trained to fit F.

In order to improve the quality of the reconstruction results, PIFuHD [48] maintains the origin PIFu framework as a coarse level prediction while adding high resolution images to a fine level network:

\[ F^H(\Phi(x, I_H, N_F, N_B), \Omega(X)) = s : s \in [0, 1] \] (2)

where \( I_H, N_F, N_B \) are the high resolution image, the predicted frontal and back normal map, and \( \Omega(X) \) is the 3D embeddings extracted from the intermediate features in the coarse level. More detailed human models can be reconstructed with additional information brought by the increasing resolution and high frequent details in the normal maps.

For multi-view images, a naive strategy is proposed in PIFu [47] to extract multi-view features, i.e., performing mean pooling on the embeddings from the intermediate layer of the MLP. However, this simple method may lead to loss of details and even collapse in real world cases, especially when the multi-view features are not consistent due to the various depth in different views and occlusions.

4. Single Person Reconstruction

Reconstruction of a single person from multi-view is a challenging problem. The main concern is to extract the meta information of the observations from different views. For this end, we propose a novel feature fusion module based on self-attention mechanism, which is effective to help the network aware of geometry details shown in the
Figure 3: Architecture of our attention-aware network. We leverage a two-level coarse to fine framework (left) with a multi-view feature fusion module based on self-attention (right). Human body prior SMPL is used in the coarse level to ensure the robustness of reconstruction, and a specially designed SMPL global normal map helps the fine level network better capture the details. To merge multi-view features efficiently, we leverage the self-attention mechanism to extract meta information from different observations, which significantly improve the reconstruction quality.

4.1. Attention-aware Multi-view Feature Fusion

In PIFu [47], the simple strategy for multi-view reconstruction is averaging the multi-view feature embeddings from the intermediate layer of MLP. We argue that the method is not efficient enough to merge the geometry details from multi-view scenes, which could lead to losing information. As shown in Figure 4, when the strategy is applied to PIFuHD [48], we obtain a smoother output. The geometry features may not remain consistent since the visible regions changes from view to view. The mean pooling method cannot handle these cases effectively.

To capture correlations between different views, inspired by [54], we propose a multi-view feature fusion method based on self-attention mechanism. The detailed architecture of the module is illustrated in Figure 3. In practice, given n observations, the multi-view features are stacked together as $\phi_m$, which is then embedded with three different linear layers and self-attention mechanism is applied:

$$
\text{attention}(\phi_q, \phi_s, \phi_t) = \text{softmax}\left(\frac{\phi_q^T \phi_s}{\sqrt{d_k}}\right) \phi_t \quad (3)
$$

where $\phi_q = \phi_m W_q$, $\phi_s = \phi_m W_s$, $\phi_t = \phi_m W_t$ denote the query, source and target feature embedded by learnable weights $W$, and $\phi_m \in \mathbb{R}^{n \times d_k}$ with $d_k$ as the embedding size. The dot-product result is divided by $\sqrt{d_k}$ to prevent the gradient vanishing problem. For linear weights $W$, we use a multi-head attention strategy [54], i.e., $W_q, W_s, W_t \in \mathbb{R}^{n_{\text{head}} \times d_k}$ which encode the multi-view features into $n_{\text{head}}$ different embedding subspaces, allowing the model to better notice different geometry patterns jointly. As the result, the weights of $n$ observations are obtained through $\text{softmax}$ function by calculating the similarity between views in the query feature $\phi_q$ and the source feature $\phi_s$. The salient covisible details tend to have large weights and will be maintained, while the invisible regions which lead to small weights have little influence on the outputs.

Finally, we stack the linear and attention layers to form a self-attention encoder as proposed in [54]. The meta-view prediction is then generated as:

$$
F^T(X) = g^T(T(\phi_m)) \quad (4)
$$

where $T(\phi_m)$ is the feature output of the self-attention encoder, and the implicit function $g^T$ predicts the occupancy field. The output meta-view feature is expected to contain the global spatial information. As demonstrated in Figure 4, when combining the attention module with PIFuHD [48], we are able to capture and preserve details with increasing observations.
4.2. Embedded with Parametric Body Model

Although the attention-aware feature fusion module is effective to mine details from multi-view, without auxiliary 3D information, the network struggles to make a reasonable prediction when information is lost due to occlusions. To address the limitation, we combine the strength of attention mechanism and parametric models.

A parametric body model, e.g. SMPL, contains the pose and shape information of human bodies. The semantic feature of SMPL is extracted by 3D convolution network for geometry inference. To improve the efficiency of attention module, inspired by the position encoding introduced in [54], we further design an informative view representation by rendering SMPL global normal maps. The global maps offer guidance for the network to identify the particular visible body parts in multi-view observations, and the corresponding geometry features can be easily extracted. Specially, to render the global normal maps, SMPL is transformed to the canonical model space, where RGB color is obtained from the normal vector and standard rendering procedure can be applied. In multi-person scenes, though images of single person can be fragmentary due to occlusions, the extra information provided by SMPL compensates for the missing part and remain consistent under different views, which significantly improves the quality and robustness of reconstruction results.

With SMPL, we rewrite the two-level pixel-aligned function Eqn. 1 and Eqn. 2. The coarse level has the formulation:

$$F^L(X) = g^L(\Phi^L(x, I), \Psi(X, \mathbf{V}_M))$$

and the fine level:

$$F^H(X) = g^H(\Phi^H(x, I, \mathbf{N}_F, \mathbf{N}_{V_M}), \Omega^L(X))$$

where $\mathbf{V}_M$ denotes the volumetric representation of SMPL, $\Psi(X, \mathbf{V}_M)$ is the SMPL semantic features, $\mathbf{N}_F, \mathbf{N}_{V_M}$ refer to the predicted frontal normal map and the rendered SMPL global normal map, and $\Omega^L(X)$ is the 3D embeddings from the coarse level. Note that here we perform the reconstruction under the SMPL model space where SMPL is normalized to a unit cube, allowing for the same setups during training and inference. The reconstruction results are then transformed to the world space and aligned together.

4.3. Temporal Fusion

For moving characters in video inputs, inconsistencies could raise between continuous frames due to the change of visible parts. To address the limitation, we propose a simple temporal fusion method. Suppose $p_{t,i}$ is a vertex of the reconstructed mesh at time $t$, we first calculate the blending weight by:

$$w_{t,j} = \sum_{j \in N_{t,i}} \frac{w_{t,j \rightarrow i}}{w_{t,i}} W_j$$

where $N_{t,i}$ is the nearest SMPL vertex set of $p_{t,i}$, $W_j$ is the blending weight of SMPL vertex $v_{j,t}$, and

$$w_{t,j \rightarrow i} = \exp\left(-\frac{\|p_{t,i} - v_{j,t}\|}{\sigma^2}\right), w_{t,i} = \sum_{j \in N_{t,i}} w_{t,j \rightarrow i}$$

is the weight of vertex $v_{j,t}$. Given the estimated SMPL models at time $t$ and $t'$, the reconstructed vertices $V_t$ can be warped to time $t'$ through the standard blend skinning:

$$V_{t',t} = W(W^{-1}(V_t, J(\beta_t, \theta_t, \mathbf{W}_t), J(\beta_{t'}, \theta_{t'}, \mathbf{W}_{t'}))$$

where $W$ refers to the skinning procedure, and $J, \beta, \theta$ are the SMPL parameters. With the warped mesh, we calculate the signed distance field (SDF) and perform mean pooling to generate continuous reconstructions:

$$S_{fusion}(t) = \frac{1}{h} \sum_{t' \in \mathcal{F}_t} S(t' \leftarrow t)$$

where $S$ denotes the SDF, and $\mathcal{F}_t$ is a sliding time window with size of $h$. In our approach, $h$ is set to 3 for consistent results while maintaining details.

5. Extend to Multi-person Reconstruction

Multi-person reconstruction is implemented by reconstructing each individual separately. The key challenge is to train the network to maintain robust against occlusions in interactive scenes. For this end, we add handcrafted occlusions during training. We firstly collect 1700 single human models from Twindom and THuman2.0 [61] to construct a large scale dataset. To simulate multi-person cases, we render images via taichi [23] and randomly project other persons to the masks, where various situations can be generated from non-occlusion to heavy occlusions. The high quality 3d models in the dataset enable us to render photo-realistic pictures, and thus the network can be trained with well generalization ability to real world data.

https://web.twindom.com/
6. Dataset and Experiment

In this section we explain our experimental settings and results. We highly recommend readers to refer to the supplementary document and video for more implementation details and a better visualization of our dataset and results.

6.1. MultiHuman Dataset

Current human dataset only provides single person scans (Twindom and THuman2.0 [61]), or 3D skeletons and deformable parametric models [29, 27]. The lack of high-quality scans of multi-person scenes limit the development of the community. For this end, we propose MultiHuman dataset, which is collected using a dense camera-rig equipped with 128 DLSRs and a commercial photogrammetry software. This system has also been used in constructing THuman2.0 [61]. Our dataset contains 150 multi-person static scenes. In total there are 278 characters with mostly university students wearing casual clothes. Each scene contains 1 to 3 persons, where each model consists of about 300,000 triangles with photorealistic texture.

To evaluate our method, we divide the dataset into different categories by the level of occlusions and number of persons, i.e., 30 single human scenes, 18 occluded single human scenes (by different objects), 46 natural interactive two person scenes, 30 closely interactive two person scenes, and 26 scenes with three persons.

6.2. Evaluation

Performance on MultiHuman We compare our method with current state-of-the-art approaches, i.e., PIFu [47], PIFuHD [48] and PaMIR [68] (PIFu + SMPL). All the methods are trained with the same setting as described in Sec. 5. For PIFuHD, the backside normal maps are not used in our implementation, and multi-view features are fused by mean pooling as [47, 68]. During test, the ground truth models are normalized to 180 centimeters height and we render 6 view images as the input. The point-to-surface distance and chamfer distance between the reconstruction and ground truth geometry are used as evaluation matrix. Quantitative results are shown in Table 1. When occlusions intensify with increasing persons and interacting elements, the loss of prior methods exacerbate while ours remains competitive. Qualitative examples in Figure 4 further demonstrate how the module can help the baseline model maintain geometry details with increasing views.

Variant 1: Self-attention Module We design a self-attention module to better capture the details from different observations. To figure out the strength of our multi-view feature fusion method, we combine the attention module with PIFu [47] (PIFu + Att) and PIFuHD [48] (PIFuHD + Att), and further evaluate our method’s performance without the module (replaced by mean pooling). Quantitative results in Table 1 shows that the module benefits baseline models under non occluded and occluded scenes. PIFuHD with attention module even outperforms ours on single human reconstruction, since the limitations brought by SMPL (Section 6.4) can lead to a lower accuracy for our method. For PIFu the improvement is marginal, indicating that the module is more effective to merge multi-view features with the detailed geometry information offered by image normal maps. For our method, we lose the competitive performance without the module. Qualitative examples in Figure 4 further demonstrate how the module can help the baseline model maintain geometry details with increasing views.

Variant 2: Use of SMPL SMPL is used in our method as a 3D proxy for the network to generate a reasonable output, and we further design a SMPL global normal map.
Figure 6: Reconstruction results on MultiHuman dataset of single person, occluded single person, two natural-interactive person, two closely-interactive person, three person scenes (top to bottom). Our method (e) generates robust and highly detailed humans, significantly narrowing the gap between ground truth (f) and performance of current state-of-the-art methods.

Figure 7: Our method generates robust results even when parts of human are not visible in closely interactive scenes.

(50x121) to improve the robustness of reconstruction against occlusions and preserving details. The huge gap between PaMIR [68] and ours indicates SMPL is not only the factor contributing to our advantages. Table 1 shows the performance of our method without the designed global maps (Ours w/o SN). The results demonstrate lower accuracy of reconstruction, which implies the efficiency of the global maps as a visual reference to guide the attention network merge multi-view information.

**Variant 3: Temporal Fusion** Figure 8 illustrates the results of our method with and without temporal fusion on real world image sequence. The temporal fusion method further enhance the reconstruction consistency, which can
Table 1: Quantitative evaluation on MultiHuman dataset. We compare our method with PIFu [47], PIFuHD [48], PaMIR [68] with mean pooling feature fusion method [47], and several variants including PIFu + Att (attention module), PIFuHD + Att, our method without attention (w/o att) and our method without the SMPL global normal maps (SN).

Figure 8: Performance on real world image sequence. (b) shows the original result of each frame, (c) demonstrates the results polished by temporal fusion.

Figure 9: Reconstruction of challenging clothes.

Figure 10: Failure case with inaccurate SMPL input. Our method is misled by the incorrect SMPL information.

7. Discussion and Future Works

Our method is capable of reconstructing high-fidelity multi-person with the suppose that the preprocessing is well conditioned, i.e., calibrated cameras, accurate segmentation, well fitted SMPL. Though in real world cases the system remains robust against small noise, large errors in preprocessing could lead to failure cases. Future works can focus on a more simplified pipeline, e.g., autonomous calibration, reconstruction using implicit human template, and lightweight networks to achieve real-time inference, which will surely make the system more applicable.

Acknowledgements This work is supported by the National Key Research and Development Program of China No.2018YFB2100500, the NSFC No.61827805, the NSFC No.62125107, the NSFC No.62171255 and Shuimu Tsinghua Scholarship.
References


[31] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. IEEE TPAMI, 2020. 3
[47] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In ICCV, October 2019. 1, 2, 3, 4, 6, 8


