Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4351951, 9 pages
https://doi.org/10.1155/2022/4351951

Research Article

Relative Pose Estimation for RGB-D Human Input Scans via Implicit Function Reconstruction

Pengpeng Liu,[1,2] Tao Yu,[3] Zhi Zeng,[2] Yebin Liu,[3] Guixuan Zhang,[2] and Zhen Song[4]

[1]*School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing, China*
[2]*Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China*
[3]*Department of Automation and BNRist, Tsinghua University, Beijing, China*
[4]*Advanced Research Center for Digitalization of Traditional Drama of the Central Academy of Drama, Beijing, China*

Correspondence should be addressed to Zhen Song; songzhen@zhongxi.cn

Received 15 October 2021; Revised 16 November 2021; Accepted 9 December 2021; Published 11 February 2022

Academic Editor: Ming Yan

To achieve a promising performance on relative pose estimation for RGB-D scans, a considerable overlap between two RGB-D inputs is often required for most existing methods. However, in many practical applications for human scans, we often have to estimate the relative poses under arbitrary overlaps, which is challenging for existing methods. To deal with this problem, this paper presents a novel end-to-end and coarse-to-fine optimization method. Our method is self-supervision which firstly combines implicit function reconstruction with differentiable render for RGB-D human input scans at arbitrary overlaps in relative pose estimation. The insight is to take advantage of the underlying human geometry prior as much as possible. First of all, for stable coarse poses, we utilize the implicit function reconstruction to dig out abundant hidden cues from unseen regions in the initialization module. To further refine the poses, the differentiable render is leveraged to establish a self-supervision mechanism in the optimization module, which is independent of standard pipelines for feature extracting and accurate correspondence matching. More importantly, our proposed method is flexible to be extended to multiview input scans. The results and evaluations demonstrate that our optimization module is robust for real-world noisy inputs, and our approach outperforms considerably than standard pipelines in non-overlapping setups.

1. Introduction

Relative pose estimation between two RGB-D scans is a fundamental problem in three-dimension (3D) vision and computer graphics. For previous multiview human motion capture [1] and human performance capture [2, 3] systems, camera extrinsic parameters (or camera poses) are required to be as accurate as possible to generate plausible results, which limit their usage. An automatic but robust autocalibration method for multiview human performance capture [4] is urgent. Examples include 3D human reconstruction from sparse views without pose parameters [5], self-calibration [6] for systems when there are disturbances to camera setups [7], and enhancing the reconstruction performance due to the inaccurate camera calibration [8].

Although there exist a lot of works focusing on relative pose estimation [9–11], they either only focus on static scene reconstruction or still require enough overlay between neighbouring RGB-D observations, which cannot be used in multiview human capture systems directly. In this paper, we are interested in relative pose estimation for human input scans, especially with nonoverlapping input scans. We present a novel end-to-end method, totally discarding the three-step paradigm [10] in the optimization module, handling the RGB-D human input scans at an arbitrary level of overlap.

Inspired from the intuition that human can estimate accurately the relative pose of two input scans, even non-overlapping, according to the underlying geometry prior. We hypothesize that the insight is to use the human prior as much as possible, especially the geometry knowledge. Therefore, this paper utilizes the human reconstruction model to enrich the underlying geometry cues of invisible regions. Taken performance, efficiency, and generalization capacity into consideration, we adopt the state-of-the-art

human reconstruction model RGB-D pixel-aligned implicit function (PIFu) [2] to obtain the detail-preserving human geometry. More importantly, we observe that multiview 3D human reconstruction is tightly related to external pose parameters, more accurate pose contributes to better reconstruction, and vice versa. It implies that we can implicitly optimize the pose parameters by supervising the human reconstruction process, free of feature extracting and matching like standard pipelines. To dig out multilevel semantic cues of human geometry prior, differentiable render [12] is used to establish a self-supervision mechanism. Thus, in the loop, we iteratively alternate between human reconstruction and relative pose estimation.

We observe that the above self-supervision optimization module is sensitive to initial result; thus, a good start is essential for faster convergence and better performance. To make our method more robust, the initialization module is added to serve the following optimization module for a stable coarse estimation. Taking inspiration from scan completion [11], firstly, we use 3D human reconstruction to complete the unseen regions and then sample point clouds with abundant geometry prior from human completion; finally, the off-the-shelf standard methods [9, 10] are used to get the coarse estimation. Based on the stable initialization, our optimization module can be more robust.

Furthermore, our coarse-to-fine framework can be easily extended to the scope of multiview inputs, owing to our special optimization mechanism via multiview implicit function reconstruction model [2]. For multiview inputs, our proposed framework can firstly optimize pairwise via the above 2-view relative pose estimation model, and then we can use multiview implicit function model and differentiable render to establish more semantic constraints, and based on the previous coarse results, we can align hierarchically the whole views to refine the pose parameters iteratively, while the most existing standard methods [9–11] can only optimize pairwise, which are error-prone for multiview extension [13].

Our contributions can be summarized as follows:

(i) The first end-to-end self-supervision relative pose estimation method for RGB-D human input scans at arbitrary overlaps, which combines the implicit function reconstruction model and differentiable render

(ii) Our proposed optimization module is robust, free of feature extraction, and correspondence matching and able to recovery good camera poses even using real-world noisy inputs. Our method also outperforms state-of-the-art standard optimization methods considerably, especially in small overlapping settings

(iii) Our method is flexible and can be easily extended to multiview input scans

2. Related Work

2.1. Relative Pose Estimation via Traditional Optimization. Traditional relative pose estimation approaches generally fall into two categories—global optimization and local optimization. Global optimization methods [7, 14] usually follow a three-step procedure [10]: extracting features, establishing correspondences of the features, and fitting a rigid transform to a subset of consistent feature correspondences. Local optimization methods [15] often focus on local pose refinement, relying on a good initialization. A popular method is geometric alignment [16], which minimizes pointwise distances between the input pairs. However, considerable overlapping regions between input scans are all required. Our approach allows input scans at any overlap and provides a stable initial result by our initialization module.

2.2. Learning-Based Relative Pose Estimation. Early works [17, 18] usually replace the modules of standard pipelines such as feature extracting or correspondence matching with deep learning networks. These methods still require massive overlaps between the input scans. Recent works [19, 20] design an end-to-end network to directly regress the relative pose. The problem of their discontinuous rotation representation such as quaternions or Euler angles is put forward in [17]. The method in [21] introduces a six-dimension (6D) continues rotation representation to alleviate the discontinuous issues. To generate probabilistic estimates, multiple rotation regressors are introduced by [22]. However, these data-driven approaches usually require large amounts of data to have a good generalization for real data. In this paper, our proposed approach is end-to-end, self-supervised, and effective for real data.

2.3. The Extreme Pose Estimation. Recently, several works addressed the extreme relative pose estimation between two input RGB-D scans, Caspi and Irani [23] focuses on image sequences and search for consistent temporal behaviour, supposing that two cameras are rigidly attached and move jointly, Yang et al. [11, 24] perform scan completion and then match the completed scans, essentially following the three-step paradigm [10]. Our initialization module is inspired from the scan completion, but our work is much different, one for that our work is well-designed for human input scans while they mainly focus on indoor scenes, not directly applied to our human input settings. Besides, our proposed optimization module discards the standard feature extracting and matching procedures and can be easily extended to multiview input settings, fully exploring the underlying human prior from the RGB-D input.

2.4. 3D Human Mesh Reconstruction via Implicit Function Representation. Recently, human reconstruction via implicit function shape representation [25, 26] has become a trend, which turns the reconstruction process more like classification task [27–29]. These methods [25, 26, 30, 31] learn a continuous implicit function representing shape using neural networks, to determine the query points inside or outside the surface. Compared with the voxel, point, and mesh representations for geometry, the implicit function representation is free of a fixed topology and not limited by the output resolution. PIFuHD [30] can produce very compelling and detailed results but struggles for complex poses, IPNet [31] focuses on 3D reconstruction from sparse and

dense point clouds and occupancy grids to produce global structure while retaining fine-scale detail, and recent work Function4D [2] achieves state-of-the-art results for 3D human reconstruction with RGB-D input scans, producing higher quality results and in real time. In this paper, we utilize [2] to obtain the detailed human geometry prior from a few RGB-D views, considering the performance and efficiency.

2.5. Differentiable Render. Differentiable render [12, 32–34] is a novel field which empowers the gradients of 3D objects to be calculated and propagated through images, which attracts increasing attention in academia and industry. The methods of differentiable render can be grouped into four categories, according to the underlying data representation: mesh, voxels, point clouds, and neural implicit functions. OpenDR [34] performs traditional rasterization in the forward pass and computes the approximate gradients in the backward pass. Recently, PyTorch3D [12] introduces a modular renderer by redesigning and exposing intermediates computed during rasterization. In this work, we use the available library PyTorch3D [12], considering its significant speed and memory improvements and convenience.

3. Approach

Given a pair of human RGB-D scans S_1 and S_2 with the camera intrinsic parameters as the input, the goal is to output the accurate rigid transformation T_{12}. To explore the hidden cues in the challenging nonoverlapping settings, we firstly exploit RGB-D PIFu [2] to recover the underlying human geometry prior from a sparse view and then propose a coarse-to-fine strategy for a robust optimization. As illustrated in Figure 1, our architecture consists of two modules; one is the initialization module producing a stable and fair good coarse pose; the other is the optimization module, refining the pose iteratively. Besides, we can easily extend our proposed method to multiview input scans.

3.1. Human Reconstruction with Implicit Function. Implicit function f is a spatially aligned and memory-efficient 3D shape representation, which consists of multilayer perceptrons [2, 30]. Instead of outputting the 3D volume, it learns a continuous function which determines the 3D voxel in space inside or outside the surface and then uses Marching Cube [35] to infer the surface based on the output Signed Distance Function (SDF) field. Recently, RGB-D PIFu proposed by [2] introduces extra depth observation to improve the reconstruction performance and accelerate the inference with filtering some exterior voxel using depth value. The network consists of an implicit function f and a fully convolutional encoder F. Following the RGB-D PIFu proposed by [2], the human surface can be defined as a level set of:

$$\begin{aligned} &f(F(\Pi(q)), q_{.z}, T(q)) = s : s \in \mathbb{R}, \\ &T(q) = \text{trunc}(q_{.z} - D(\Pi(q))), \end{aligned} \tag{1}$$

where f is a continuous implicit function represented by multilayer perceptrons. For a query point q, $\Pi(\cdot)$ is the perspective projection function, $F(\cdot)$ is the feature extracted by encoder networks, $D(\cdot)$ is a bilinear function sampling depth values on the depth image, the Probabilistic Signed Distance Function (PSDF) value of q is introduced by $q_{.z} - D(\Pi(q))$ to fully utilize the depth observation, and $T(\cdot)$ is used to truncate PSDF value to $[-\sigma_i, \sigma_i]$, eliminating the ambiguities of using global depth values. The sign of s represents whether the point is inside or outside the surface.

The RGB-D PIFu [2] meets the requirements of good generalization as discussed in [36], using view-centric coordinate systems, extracting geometry-aware feature maps, and exploiting multiview aggregation strategy for fusion. As demonstrated in Figures 2 and 3, the model shows the good generalization in real data collected by Kinect, which is trained only using 300 high-fidelity data from [2]. We observe that the reconstruction of visible parts is generally detailed and lifelike, while the unseen part completion is a little oversmooth but still human-like according to the learning prior from large data, which still empowers us abundant hidden cues. Considering the efficiency, compelling performance, and good generalization in real data, we adopt the RGB-D PIFu [2] to dig out the underlying geometry prior, which can produce a detail-preserving complete surface in real time.

3.2. Initialization Module. A good initial result can accelerate the convergence and contribute to better performance. As in Figure 1, the black arrows represent the flow of the initialization module which serves as a coarse relative pose estimation. The challenge is how to ensure a fair good result even in nonoverlapping input settings where almost all the standard pipelines [7, 9] cannot work well due to lack of enough correspondence. Enlightened by scan completion [11], human geometry prior through implicit function reconstruction is introduced to provide the hidden cues of the invisible region. Given the RGB-D scan S_i and camera intrinsic parameters K_i, for single-view input, the world coordinate can be assumed identical to the camera coordinate; thus, the completed human mesh M_i^1 can be reconstructed with single-view RGB-D PIFu without the relative pose. And then, using point clouds sampled from the mesh M_i^1, the initial relative pose can be estimated with the off-the-shelf standard method easily. In this paper, we leverage a popular registration method which combines the global registration [14] with local registration [15, 16]. Compared to the point clouds derived from depth images, the human completion mesh M_i^1 provides detail-preserving surface in visible observation and much richer hidden cues of human geometry prior in unseen parts. In addition, our initialization process is also efficient owing to the real-time property of the RGB-D PIFu [2].

3.3. Optimization Module. This module is served as a relative pose refinement. For multiview human reconstruction, adding a new view but with erroneous pose can be suicide, which may lead to artifacts, distortion, or even collapsed results. Fortunately, we can just grasp this property and establish a self-constraint end-to-end optimization mechanism via differentiable render. The insight for the proposed optimization process is that the inferred models (including

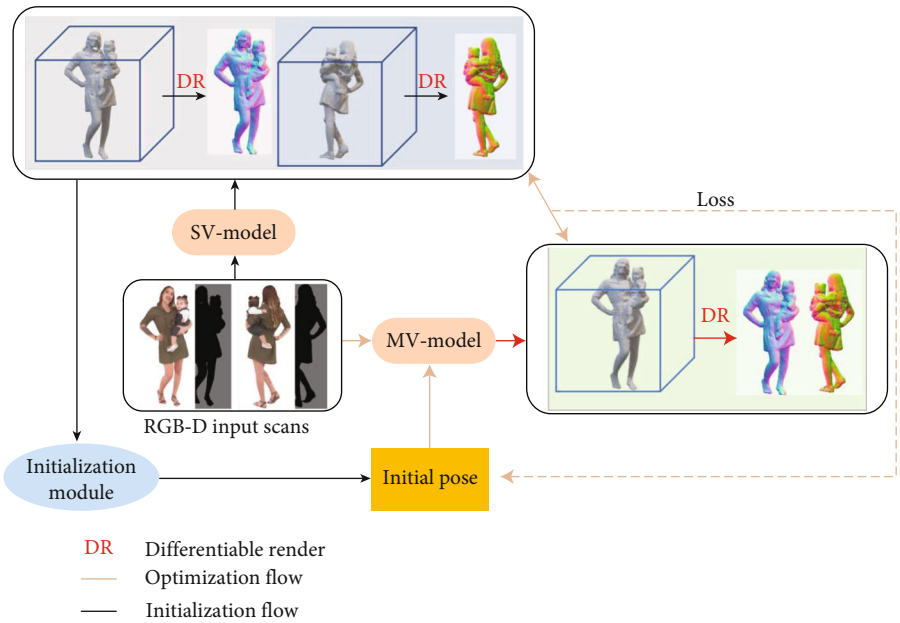

Figure 1: Methods pipeline. SV is the abbreviation of single-view, and SV model means our single-view human reconstruction model; similarly, MV means multiview and MV model means our multiview reconstruction model. The proposed framework combines the initialization module and optimization module. The black lines represent the initialization flow, and the red lines are the optimization flow. Given the RGB-D input scans, the initialization module produces a coarse initial pose; then, the optimization module refines the relative pose iteratively.

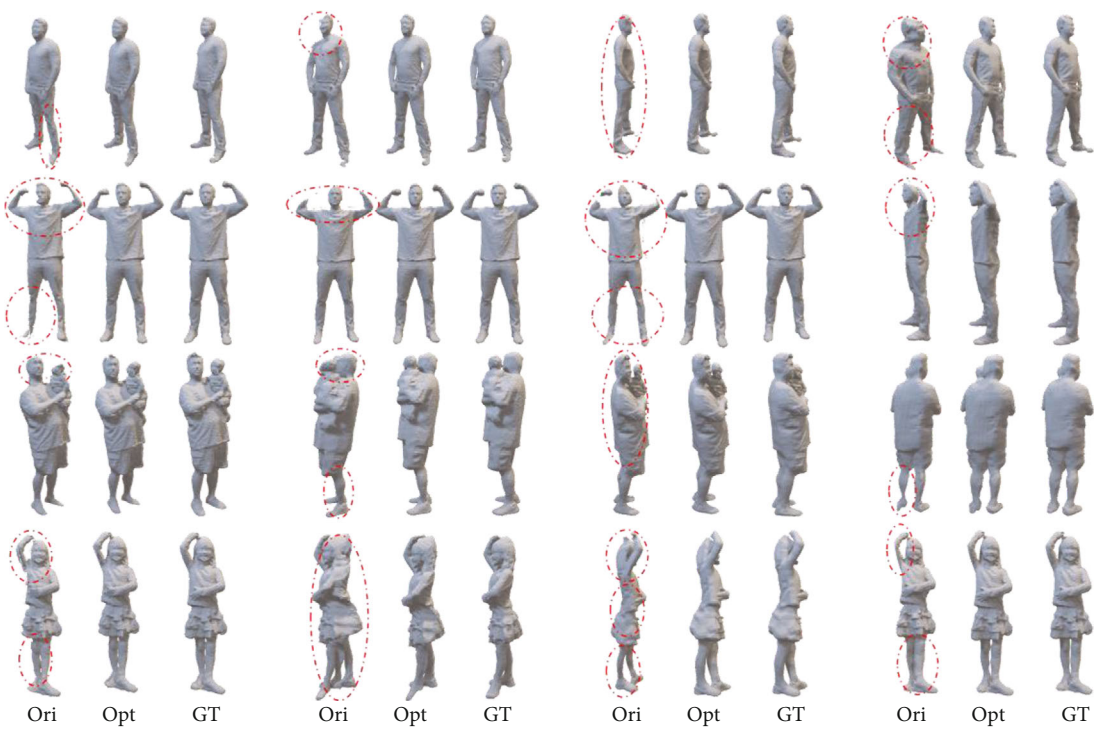

Figure 2: We qualitatively show the robustness of our optimization module. We add different small random Gauss noise to the groundtruth pose. The first column is the 2-view RGB-D PIFu reconstruction with original pose (with noise), the second column is the reconstruction after our optimization module, and the last column is the human reconstruction with groundtruth pose.

single-view inferred models from each view and also the multi-view inferred model) should be consistent with each other. If the multiview inferred model is quite different from the single-view inferred models, the relative pose must not be accurate. Moreover, the multiview inferred model should also match the real RGB-D observations from different

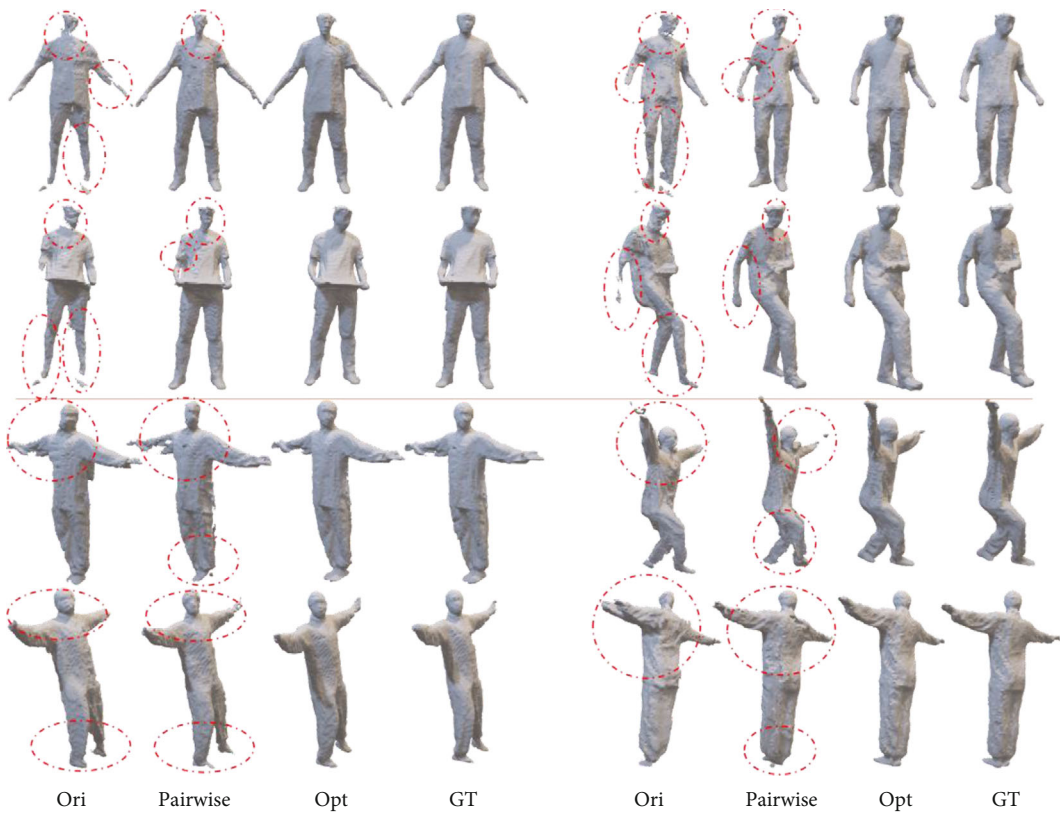

Figure 3: We show the performance of our multi-view extension. The first two rows show the 3-view input scans while the last two rows with 4-view input scans. We still add different small disturbances to the groundtruth poses. The first column is the reconstruction with original pose, the second column is the reconstruction with only pairwise optimization, the third column is the human reconstruction with our proposed optimization method, and the last column with the groundtruth pose.

views, which means that the multi-view inferred model should not violate the two-dimension (2D) semantic constraints from the real observations. Specifically, the differentiable render can be firstly used to render the human mesh M_i^1 from the input perspective to get the groundtruth semantic label such as mask $\hat{m}_i$ and normal image $\hat{n}_i$, for that the lifelike rendering results could be ensured from the visible perspective. Note that additional supervisions can also be added by rendering from novel views if the rendering results are enough realistic.

Given the initial relative pose parameters $T_{i0} = [R, t]$, the human mesh M_i^2 can be generated by our 2-view RGB-D PIFu. Similarly, the predicted semantic label such as mask $\hat{m}_i$ and normal image $\hat{n}_i$ can be rendered via differentiable render. Then, the loss terms can be proposed based on the following principles and be back propagated to update the relative pose parameters iteratively.

For loss terms, instead of penalizing the relative pose directly (which is also impossible), we implicitly supervise the multiview human reconstruction process. Specifically, we compute (R, t) by solving

$$\min_{R,t} a_1 L_{2D}(R, t) + a_2 L_{\mathrm{sdf}}(R, t) + a_3 L_{\mathrm{orth}}(R), \tag{2}$$

where a_i is the weight coefficient for each energy term. From Equation (2), our loss is composed of three parts, the 2D semantic loss $L_{2\mathrm{D}}(\bullet)$, the SDF field loss $L_{\mathrm{sdf}}(\bullet)$, and the orthogonal constraint $L_{\mathrm{orth}}(\bullet)$.

The first term $L_{2\mathrm{D}}(\bullet)$ is a 2D semantic supervision to keep the 2D semantic consistency. We mainly use mask and normal as 2D semantic priors to make sure that the rendered images (from the multiview RGBD-PIFu result) should be consistent with the input images. Thus, our 2D semantic loss can be described as $L_{2\mathrm{D}} = b_1 l_{\mathrm{mask}} + b_2 l_{\mathrm{normal}}$. We empirically set the weight coefficient $b_1 = b_2 = 1$. The normal loss l_{normal} is a $L2$ norm loss, described as $\|n_i - \hat{n}_i\|_2$, enhancing the normal consistency. The mask loss penalizes the small overlap of the two mask images, described as $\|m_i^{\mathrm{dist}} - \hat{m}_i^{\mathrm{dist}}\|_2$, the superscript dist representing distance transformation. Instead of calculating mean square error (MSE) for original mask images directly, we firstly using distance transform and then compute the MSE, which is more robust by experience.

The SDF field loss $L_{\mathrm{sdf}}(\bullet)$ is a 3D supervision, ensuring that the multiview reconstruction mesh is human-like. It can be described as $\|\mathrm{sdf}^2 - \mathrm{sdf}^1\|_2$, the superscript representing the view numbers, sdf^1 representing the predicted SDF of the single-view RGB-D PIFu. We observe that only supervised by the sparse view 2D projection is not enough, it is probably to be trapped into local minima that satisfy each perspective's constraint but looking like deformed and

weird. To solve the problem, we introduce the SDF field loss to make the consistency between the multiview SDF field and each single-view SDF field, to avoid constraining SDF filed into an abnormal field.

The last constraint $L_{\text{orth}}(\bullet)$ is an orthogonal constraint which is used to keep the orthogonality of rotation matrix R, described as $\|RR^T - I\|_2$, where I is identity matrix.

3.4. Multiview Extension. The above model is aimed at estimating the relative pose of a pair of images, while the goal of multiview extension is to output the relative pose of each view, given multi-view RGB-D input scans of the same person. However, most existing relative pose estimation approaches [11, 14, 24] are not suitable for multiview input, for that they mainly focus on pairwise input, which can only optimize pairwise for multi-view setting. As [13] has observed, only pairwise registration is error-prone, leading to odometry drift due to sensor noise and false pairwise alignments. Some multiway registrations [13] align multiple pieces of geometry in a global space via pose graph optimization [37], they combine local optimization such as point-to-plane ICP [9] with global optimization such as Levenberg-Marquardt algorithm [38], yet they still rely on large overlapping region to match considerable and accurate correspondences.

Our proposed approach can be flexibly extended to multiview setups owing to the good generalization of our multiview implicit function human reconstruction model RGB-D PIFu. For multiview settings, the n-view optimization can be based on the results of $n-1$ views, just like the pyramid structure; thus, we can optimize hierarchically in a coarse-to-fine strategy. As shown in Figure 4, the relative pose of the three views is agnostic, assuming the view_1 as the world coordinate system; we firstly optimize pairwisely, view_{12}, view_{13}, and view_{23}, note that the initial pose of view_{23} can be referred after view_{12} and view_{13} optimizations. Then, we align globally via 3-view pose optimization module to refine the pose. The architecture of 3-view pose optimization is analogous to 2-view optimization as Figure 1; using 3-view RGB-D PIFu and differentiable render to establish self-supervision mechanism, the loss term can be described as the following equation:

$$\min_{R,t} \sum_{i=1}^{n} a_{i1} L_{2\text{D}}^{i}(R,t) + a_{i2} L_{\text{sdf}}^{i}(R,t) + \text{a}_3 L_{\text{orth}}(R). \quad (3)$$

In practice, we can skip some middle-level optimization in practice, because of the diminishing marginal utility [39]; the more top-level optimization, the less effect it improves.

4. Results and Discussion

In this section, we present an experimental evaluation of the proposed approach. We firstly describe our evaluation dataset and evaluation metrics. Then, we compare our method with several baseline techniques, assessing the performance quantitively on different overlapping RGB-D input pairs. Finally, we present an ablation study to quantitively examine the impact of our initialization module and qualitatively demonstrate the robustness of our proposed optimization module.

4.1. Dataset. We perform experimental evaluation on two types of data. One is the synthesized data, rendering 200 high-quality scans from 60 views with rotation and random shifts. Note that to keep consistent with the real data, for the colour image, we use the PRT-based render as in [25]; for the depth image, we first render the groundtruth depth maps and then add the TOF sensor noise on top of them following [2]. Finally, we synthesize RGB-D data with resolution 512×512. The other is real data collected by multiview Kinects, which includes large poses, various clothes, and different people. We firstly segment the human mask provided by Kinect. Then, we align the colour image and depth image with the pose parameters of Kinect depth camera, getting the final real RGB-D data with resolution 640×576.

For a more comprehensive and detailed analysis, we classify all the data into four categories according to the overlap rate: overlap rate 0-5%, overlap rate 10%-30%, overlap rate 40%-70%, and overlap rate over 80%. Then, we select about 400 characteristic data pairs for each category as our final evaluation data, besides each category is mixed by half the real data and half the synthetic data. Note that for the RGB-D PIFu, we train it using 500 high-quality scans following [2].

4.2. Evaluation Metrics. We evaluate the rotation matrix R and translation part t of the relative pose $T = (R, t)$, respectively. We follow the standard protocol of reporting the rotation angle error $\arccos((tr(R^* R^T) - 1)/2)$ and translation error $\|t - t^*\|_2$, let (R^*, t^*) be the groundtruth relative pose and (R, t) be the predicted pose.

4.3. Quantitative Evaluation. We consider four baseline approaches: Super4PCS [7], Greg [10], ICP [9], and combine Greg with ICP. Super4PCS is a widely used global scan matching method between two 3D point clouds. Greg is another state-of-the-art global registration, combining cutting-edge features and reweighted least squares for rigid pose registration. ICP (iterative closet point), a local optimization algorithm, has been a mainstay of geometric registration in both research and industry. In this paper, we use point-to-plane ICP which has a faster convergence. We also combine global registration with local optimization as a baseline; the former provides an initial pose, and the latter refines the pose, just as our coarse-to-fine strategy.

Table 1 provides the quantitative results of our approach and baseline methods. We show the mean error for rotation and translation components for overlapping rate ($0-5\%$, $10\%-30\%$, $40\%-70\%$, and $\geq 80\%$) scan pairs, respectively. Overall, we observe that the less overlap rate, the greater advantage of our method. Our approach outperforms baseline approaches considerably in small-overlapping and almost overlapping settings, while performing slightly better in large-overlapping settings. In the four baselines, combining Greg with ICP is much better than others, especially in the significant overlap scenes, while in small-overlapping settings, all the baselines perform badly, making no difference.

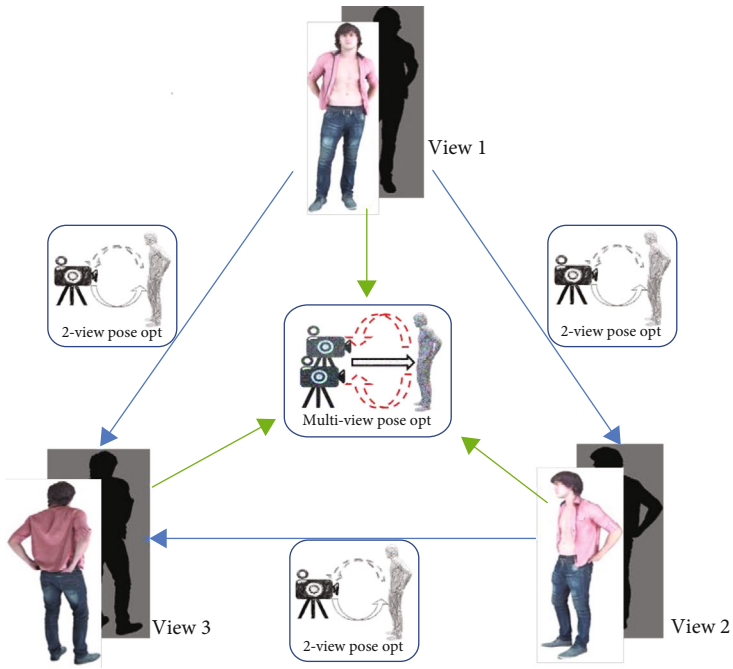

Figure 4: Multiview optimization. Opt is the abbreviation of optimization, 2-view pose Opt in picture means our 2-view relative pose optimization, whose structure is as described in Figure 1; similarly, multiview pose Opt means our multiview relative pose optimization. We optimize the pose parameters hierarchically, just as the pyramid structure. We firstly optimize pairwise to get a coarse pose and then hierarchically align the whole views based on the coarse results via our proposed multiview optimization module.

Table 1: The relative pose estimation evaluation on our dataset. We divide all the data into four categories according to the overlap rate: 0-5%, 10%-30%, 40%-70%, and over 80%, and report the mean error of rotation angle (Rot. (°)) and translation (Trans. (m)), respectively.

	Overlap (0-5%)		Overlap (10-30%)		Overlap (40-70%)		Overlap (≥80%)	
	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.
Super4PCs [7]	126.08	1.715	103.95	1.468	38.99	0.574	4.46	0.084
Greg [10]	124.41	1.679	92.50	1.538	64.47	1.109	18.44	0.325
ICP [9]	165.97	2.252	122.02	1.943	67.86	1.208	27.75	0.540
Greg+ICP	126.72	1.664	103.88	1.461	35.10	0.508	1.27	0.0262
Ours-initial	7.15	0.138	6.54	0.130	5.90	0.106	1.29	0.0259
Ours	6.16	0.127	5.88	0.119	4 .60	0.084	1.15	0.0209

In small overlap (overlapping rate 10%-30%) or almost no overlap (overlapping rate 0-5%), all the baselines perform very badly, with over 120 rotation errors and over 1.5 m translation errors. These methods rely on accurate correspondence in the overlap regions; thus, they cannot handle the small-overlapping or nonoverlapping settings. Even so, our approach performs much better, with mean errors in rotation/translation 6.16/0.127 m for almost no overlap and 5.88/0.119 m for small overlap. It fully demonstrates the effectiveness of our proposed approach for nonoverlapping settings, thanks to the coarse-to-fine optimization strategy.

In middle overlap (overlapping rate 40%-70%), although the Greg with ICP is much better than other baselines, with rotation/translation errors 35.10/0.508 m, our approach achieves much better results, with corresponding errors 4.60/0.084 m.

In significant overlap (overlapping rate over 80%), there is no significant difference between all the methods. All the baselines perform fairly good, especially, the Greg with ICP with small rotation/translation errors 1.27/0.0262 m. It further shows that the standard baselines require input scans possessing considerable overlapping regions for superior performance. However, our approach is still competitive, performing slightly better with mean rotation/translation errors 1.15/0.0209 m.

In this experiment, all the four baselines rely highly on the large overlap of input scans, not able to handle the extreme pose setting, while our approach performs stable and good in all settings.

4.4. Ablation Study. We conduct two experiments to evaluate the effectiveness of our proposed module. Firstly, we evaluate quantitatively our initialization module. From Table 1,

we find that our initialization module performs very stably. Even in the almost nonoverlapping settings, it can generally provide fair good initial results with small mean rotation/translation errors 7.15/0.138 m. In this module, we use RGB-D PIFu to reconstruct the full geometry, digging out the abundant hidden cues of the unseen regions. Therefore, our proposed initialization module can be free of overlap of the inputs.

Secondly, we evaluate qualitatively the robustness of our optimization module. This module serves as a refinement, conditioned on a not much bad pose. Thus, during the experiment, we add an arbitrary small Gauss random noise to the groundtruth pose parameter as a coarse initial pose. For a more intuitive distinction in the results, we qualitatively demonstrate the reconstruction performance, considering that the reconstruction procedure will amplify the impact of small deviation. Figure 2 shows the performance of different small noises. We can see that our proposed optimization can generally recovery a fair perfect human mesh in different small disturbances, which fully proves the robustness of our model.

Our approach can be easily extended to multiview input settings. As demonstrated in Figure 3, our proposed methods perform better than only pairwise optimization, closely to the reference reconstruction.

5. Conclusions

This paper proposes an end-to-end approach which consists of an initialization module and an optimization module, to estimate the relative pose between the RGB-D human input scans. We do not limit the input scans to have large overlap. Our initialization module can handle nonoverlapping settings effectively by incorporating the learned 3D human shape prior. Our optimization module refines our pose parameters by implicitly supervising the human reconstruction with multilevel constraints via implicit function reconstruction and differentiable render. The experiments demonstrate our optimization module is robust to small disturbances. Besides, our method can easily extend the scope to multiview input scans. Through evaluation on different overlap data, our method considerably outperforms the state-of-the-art baselines, especially for nonoverlapping scans. We believe the proposed method will stimulate the wide spread of multiview human reconstruction systems by eliminating the sophisticated camera pose calibration process.

Data Availability

We collect about 500 human models from Twindom dataset for training RGB-D PIFu. You can purchase the Twindom dataset from the webpage: https://web.twindom.com/. We provide part of our evaluation data, you can download from the hyperlink: https://drive.google.com/file/d/1SB3eTvbcG2b6CJbpF84jm_pApNMCTyvf/view?usp=sharing.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by "Beijing Outstanding Young Scientist Program" (BJJWZYJH01201910048035) and "the Fundamental Research Funds for the Central Universities" (YNZDA1805).

References

[1] W. Li, Y. Zeng, Q. Zhang, Y. Wu, and G. Chen, "Human motion capture based on incremental dimension reduction and projection position optimization," *Wireless Communications and Mobile Computing*, vol. 2021, 9 pages, 2021.

[2] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu, "Function 4d: real-time human volumetric capture from very sparse consumer rgbd sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5746–5756, 2021.

[3] T. Yu, Z. Zheng, K. Guo et al., "DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor," in *IEEE conference on computer vision and pattern recognition*, pp. 7287–7296, Salt Lake City, USA, 2018.

[4] Z. Li, T. Yu, Z. Zheng, K. Guo, and Y. Liu, "POSEFusion: pose-guided selective fusion for single-view hu-man volumetric capture," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14172, 2021.

[5] X. Li, S. Liu, K. Kim et al., "Self-supervised single-view 3d reconstruction via semantic consistency," in *European Conference on Computer Vision*, Springer, 2020.

[6] O. D. Faugeras, Q. T. Luong, and S. J. Maybank, "Camera self-calibration: theory and experiments," in *European Conference on Computer Vision*, pp. 321–334, Springer, 1992.

[7] N. Mellado, D. Aiger, and N. J. Mitra, "Super 4PCS fast global pointcloud registration via smart indexing," *Computer Graphics Forum*, vol. 33, no. 5, pp. 205–215, 2014.

[8] J. Chen, K. Benzeroual, and R. S. Allison, "Calibration for high-definition camera rigs with marker chess-board," in *IEEE Computer Society Conference on Coputer Vision and Pattern Recognition Workshops*, pp. 29–36, Providence, RI, USA, 2012.

[9] S. Bouaziz, A. Tagliasacchi, and M. Pauly, "Sparse iterative closest point," *Computer Graphics Forum*, vol. 32, no. 5, pp. 113–123, 2013.

[10] G. K. Tam, Z. Cheng, Y. Lai et al., "Registration of 3d point clouds and meshes: a survey from rigid to nonrigid," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 7, pp. 1199–1217, 2012.

[11] Z. Yang, J. Z. Pan, L. Luo, X. Zhou, K. Grauman, and Q. Huang, "Extreme relative pose estimation for RGB-D scans via scene completion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4531–4540, California, USA, 2019.

[12] N. Ravi, J. Reizenstein, D. Novotny et al., "Accelerating 3d deep learning with PyTorch3D," 2020, https://arxiv.org/abs/2007.08501.

[13] S. Choi, Q. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5556–5565, Boston, USA, 2015.

[14] D. Aiger, N. J. Mitra, and D. Cohen-Or, "4-points congruent sets for robust pairwise surface registration," *ACM SIGGRAPH*, vol. 27, no. 3, pp. 1–10, 2008.

[15] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[16] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and Vision Computing*, vol. 10, no. 3, pp. 145–155, 1992.

[17] S. Kim, S. Lin, S. R. Jeon, D. Min, and K. Sohn, "Recurrent transformer networks for semantic correspondence," *Advances in Neural Information Processing Systems*, vol. 3, pp. 6126–6136, 2018.

[18] R. Ranftl and V. Koltun, "Deep fundamental matrix estimation," in *Computer Vision – ECCV 2018: European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 284–299, Springer Link, 2018.

[19] E. Insafutdinov and A. Dosovitskiy, "Unsupervised learning of shape and pose with differentiable point clouds," 2018, https://arxiv.org/abs/1810.09381.

[20] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Relative camera pose estimation using convolutional neural networks," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 675–687, Springer, 2017.

[21] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, California, USA, 2019.

[22] V. Peretroukhin, B. Wagstaff, and J. Kelly, "Deep probabilistic regression of elements of so (3) using quaternion averaging and uncertainty injection," in *CVPR Workshops*, pp. 83–86, California, USA, 2019.

[23] Y. Caspi and M. Irani, "Aligning non-overlapping sequences," *International Journal of Computer Vision*, vol. 48, no. 1, pp. 39–51, 2002.

[24] Z. Yang, S. Yan, and Q. Huang, "Extreme relative pose network under hybrid representations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2455–2464, 2020.

[25] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: pixel-aligned implicit function for high-resolution clothed human digitization," in *IEEE/CVF International Conference on Computer Vision*, pp. 2304–2314, California, USA, 2019.

[26] Z. Li, T. Yu, C. Pan, Z. Zheng, and Y. Liu, "Robust 3d self-portraits in seconds," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1344–1353, 2020.

[27] M. Yan, S. Li, C. A. Chan, Y. Shen, and Y. Yu, "Mobility prediction using a weighted Markov model based on Mobile user classification," *Sensors*, vol. 21, no. 5, p. 1740, 2021.

[28] M. Yan, H. Yuan, Z. Li, Q. Lin, and J. Li, "Energy savings of wireless communication networks based on mobile user environmental prediction," *Journal of Environmental Protection and Ecology*, vol. 22, no. 1, pp. 206–217, 2021.

[29] M. Yan, X. Lou, and Y. Wang, "Channel noise optimization of polar codes decoding based on a convolutional neural network," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 1434347, 10 pages, 2021.

[30] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 84–93, 2020.

[31] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Combining implicit function learning and parametric models for 3d human reconstruction," 2020, https://arxiv.org/abs/2007.11432.

[32] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: learning dynamic renderable volumes from images," 2019, https://arxiv.org/abs/1906.07751.

[33] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui, "DIST: rendering deep implicit signed distance function with differentiable sphere tracing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2019–2028, 2020.

[34] W. Yifan, F. Serena, S. Wu, C. Öztireli, and O. Sorkine-Hornung, "Differentiable surface splatting for point-based geometry processing," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–14, 2019.

[35] W. E. Lorensen and H. E. Cline, "Marching cubes: a high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[36] M. A. Bautista, W. Talbott, S. Zhai, N. Srivastava, and J. M. Susskind, "On the generalization of learning-based 3d reconstruction," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2180–2189, 2021.

[37] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert, "Initialization techniques for 3d slam: a survey on rotation estimation and its use in pose graph optimization," in *IEEE international conference on robotics and automation (ICRA)*, pp. 4597–4604, Seattle, WA, USA, 2015.

[38] J. J. Mor'e, "The Levenberg-Marquardt algorithm: implementation and theory," in *Numerical Analysis*, pp. 105–116, Springer, Berlin, Heidelberg, 1978.

[39] T. Veblen, "The limitations of marginal utility," *Journal of Political Economy*, vol. 17, no. 9, pp. 620–636, 1909.