Multiview Human Body Reconstruction from Uncalibrated Cameras

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a new method to reconstruct 3D human body pose and shape by fusing visual features from multiview images captured by uncalibrated cameras. Existing multiview approaches often use spatial camera calibration (intrinsic and extrinsic parameters) to geometrically align and fuse visual features. Despite remarkable performances, the requirement of camera calibration restricted their applicability to real-world scenarios, *e.g.*, reconstruction from social videos with wide-baseline cameras. We address this challenge by leveraging the commonly observed human body as a semantic calibration target, which eliminates the requirement of camera calibration. Specifically, we map per-pixel image features to a canonical body surface coordinate system agnostic to views and poses using dense keypoints (correspondences). This feature mapping allows us to semantically, instead of geometrically, align and fuse visual features from multiview images. We learn a self-attention mechanism to reason about the confidence of visual features across and within views. With fused visual features, a regressor is learned to predict the parameters of a body model. We demonstrate that our calibration-free multiview fusion method reliably reconstructs 3D body pose and shape, outperforming state-of-the-art single view methods with post-hoc multiview fusion, particularly in the presence of non-trivial occlusion, and showing comparable accuracy to multiview methods that require calibration.

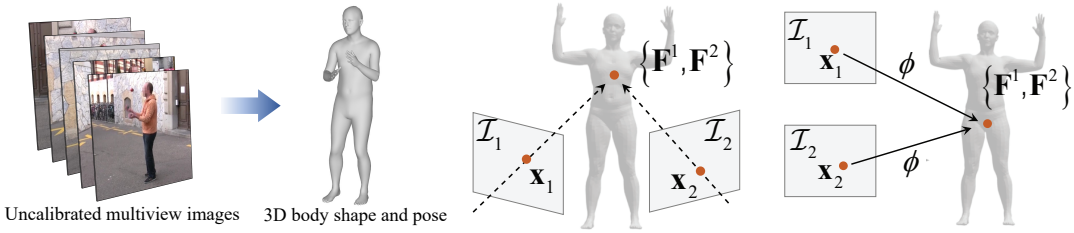
1 Introduction

Cameras are an integral part of our lives for us to capture and share priceless moments. In particular, social videos voluntarily captured by multiple viewers watching the same scene, *e.g.*, friends recording a street busker simultaneously, provide a new form of popular content for visual communication in social media.

These videos possess two properties. (1) The videos are, by nature, multiview from social members, which provide redundant yet distinctive visual information to model the 3D body geometry of a target subject. For instance, the busker's right elbow is occluded by his torso in one view but visible from another viewpoint as shown in Figure 1(a) [2]. This self-occlusion can be reasoned by consolidating multiview information. (2) These videos are, in general, not spatially calibrated (intrinsic and extrinsic parameters). There exists neither a calibration pattern nor common visual features to register 3D camera poses (*e.g.*, structure from motion) because of wide-baseline cameras. Existing computer vision solutions leveraging multiview calibration [13, 15, 12, 32] are therefore not applicable. On the other hand, single view approaches [19, 20, 39, 17] for 3D body reconstruction can be combined in a post-hoc way, *e.g.*, taking the average of global visual features, which shows limited performance.

In this paper, we propose a novel method to reconstruct 3D human pose and shape from uncalibrated multiview images. We use the commonly observed human body as a *semantic calibration target* to align visual features across views: the dense keypoints (correspondences) [10] detected in an image provide an injective mapping from image coordinates to the canonical body surface coordinates

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Input and output (b) Geometric fusion (c) Semantic fusion

Figure 1: (a) We present a new method to reconstruct the 3D human body shape and pose by fusing visual features from uncalibrated multiview images. All results in this manuscript are generated before blurring the face of the subject. (b) Fusing visual features $\{\mathbf{F}^1, \mathbf{F}^2\}$ from multiview images require camera calibration to align features. (c) Instead, we use body semantics, e.g., dense keypoints, to align features, eliminating the requirement of camera calibration.

regardless of views and poses, which allows semantically aligning and then, fusing visual features from multiview images.

We integrate the following desired properties to learn visual features for reconstructing 3D body pose and shape. (1) View independence: Without camera calibration, in order to learn and fuse visual features independent of views, we need to transport them to a shared space. To achieve this, we semantically, instead of geometrically, transfer and align visual features based on dense keypoint mapping to the body surface. Such a unified coordinate system makes visual feature extraction agnostic to views and poses. (2) Locality: We hypothesize that a local visual feature for each pixel can describe the fine-grained relationship between human body parts around the pixel. In conjunction with global features that describe a holistic pose, we learn local features from multiview images, jointly. (3) Confidence: Views are not equally informative and can complement each other in inferring different body parts. We design an attention mechanism to reason about the confidence of visual features, both globally across multiple views and locally within each view. Across multiple input images, a higher global confidence score is assigned to a view generating more accurate reconstruction, *e.g.*, a less occluded one. Within each view, higher local confidence scores are assigned for body parts that are easier to infer, *e.g.*, a clearly visible limb.

We design an end-to-end network that learns to reconstruct 3D body geometry from multiview images. It takes as input, a set of multiview images and extracts per-pixel features and their confidences. We use dense keypoint mapping to semantically align the features in the canonical coordinate system defined by the human body surface. These aligned features from multiview images are fused by using the predicted confidences, which form unified per-vertex features. We concatenate those fused per-vertex features together with image feature for learning body model parameters.

We apply our method to reconstruct the 3D human body from uncalibrated multiview images. Our method can effectively fuse visual features from multiview images, outperforming single view approaches with post-hoc fusions, in particular, for self-occlusion, and showing comparable performance compared to multiview approaches that use calibrated images. We demonstrate our work in realworld scenarios such as social videos where camera calibration is challenging.

Contributions This paper makes three major technical contributions. (1) We propose a novel multiview method for human pose and shape reconstruction that scales up to an arbitrary number of uncalibrated camera views (including single view), guided by dense keypoints. (2) We introduce a novel self-attention based multiview feature fusion method that takes uncertainties across and within different views into consideration. (3) We achieve state-of-the-art results among calibration-free 3D human body pose and shape reconstruction approaches.

2 Related Work

To handle the challenge due to occlusion, single-view-based methods infer the occluded parts in a data-driven manner, while multiview approaches can fuse visual features from different views.

Single View 3D Human Reconstruction Reconstructing occluded 3D content from a single view image is a geometrically ill-posed problem. With some assumptions about scenes, it is possible to reason about the occluded geometry. For instance, for humans, a low dimensional parametric model such as SMPL [27] can be used to reconstruct shape and pose by matching 2D sparse landmarks [8, 3]

and semantic segmentation masks [22]. On the other hand, learning-based approaches directly regress the model parameters from an image [19], or semantic representations such as 2D keypoint heatmaps and silhouettes [30], part segmentation [28], and dense correspondences [39]. Such learning-based approaches rely on a dataset with 3D ground truth such as Human3.6M [14] to learn the relationship between an image and the shape and pose, which is prone to overfitting when applying to a novel scene image. To address the limited 3D ground truth data, self-supervised learning can be used by generating photorealistic data [20] and by enforcing consistency for predictions from multiple views [29, 40]. To enhance the expressibility of the model, the parametric representation is relaxed by incorporating a volumetric representation [37, 16], a point cloud [7], or vertex locations of a pre-defined mesh [21, 42]. Compared with single view reconstruction, our approach is tailored for multiview feature fusion that significantly improves the performance.

Multiview Fusion for 3D Human Reconstruction When camera calibration (intrinsics and extrinsics) is available, single view reconstruction methods can be extend to multiview settings. For instance, SMPLify [3] is extended to reconstruct a 3D body geometry in an unified coordinate system where its validity can be measured by projecting onto multiview images, *i.e.*, the reprojection error of the 2D keypoints and silhouette is minimized to learn a geometrically coherent model [24]. Multiview geometry offers a geometric way to fuse multiview features for self-occlusion reasoning. For instance, visual features [12] or predictions [45] can be transferred to other views through epipolar lines, inverse projection [15, 36], or a latent view-invariant space [32].

When the camera calibration is not available, the learning-based approaches can be used to fuse multiview features. For example, a fusion module can be learned to warp a feature map from one view to align with that of another view [31, 38], which does not rely on calibration but requires learning a large number of parameters to model the cross view geometry. Besides, uncalibrated multiview images can be used in a sequential manner. For example, HMR [19] uses in a recurrent framework to progressively refine an initial single view prediction of camera and human model parameters, view by view and stage by stage [25]. The requirement of synchronization between cameras can be relaxed by learning a probabilistic distribution of the model parameters [34]. Unlike existing work, we fuse multiview images without calibration using the observed human body as a semantic calibration target. The features are transported to the canonical body surface where a self-attention module is used to combine multiview features effectively.

Dense Keypoints for 3D Human Reconstruction Dense keypoint estimation, *e.g.* DensePose [10], establishes correspondences between image pixels and the canonical body surface coordinate system. This provides a strong cue to infer 3D body pose and shape from a single image that alleviates the self-occlusion problem [33]. Due to such a desirable property, single view approaches employ dense correspondence for supervision [9, 42] or as an additional cue [43, 39]. Further, it can be used to reconstruct 3D body pose and shape from multiview images [23, 41]. Inspired by DecoMR [42], we leverage dense correspondences to re-arrange visual features from different images into a canonical body surface, which allows us to effectively fuse multiview features. Note that DecoMR [42] is a single view method that transfers features to a 2D canonical body surface coordinate and regress a discrete view-specific vertex location map; while ours transfer features from an arbitrary number of views to a continuous 3D canonical body mesh for fusion and regressing a view-indepdent body model.

3 Method

We present a novel method that reconstructs 3D human body pose and shape from an arbitrary number of uncalibrated multiview images. We leverage dense keypoints to map visual features from multiview images to a common canonical coordinate system, which allows effectively fusing them for 3D reconstruction.

3.1 Human Body as Semantic Calibration Target

Given an image $\mathcal{I} \in [0,1]^{W \times H \times 3}$ containing a person with width W and height H, we aim to recover the person's 3D body pose and shape represented as mesh vertices $\mathbf{Z} = \{\mathbf{z}_i \in \mathbb{R}^3\}_{i=1}^{N_v}$ by estimating SMPL model [27] parameters Θ.

This problem in general can be formulated as a composition of an image encoder and a mesh decoder:

$$\mathbf{F} = f_{\mathrm{E}}(\mathcal{I}; \theta_E), \quad \mathbf{Z} = f_{\mathrm{D}}(\mathbf{F}; \theta_D), \tag{1}$$

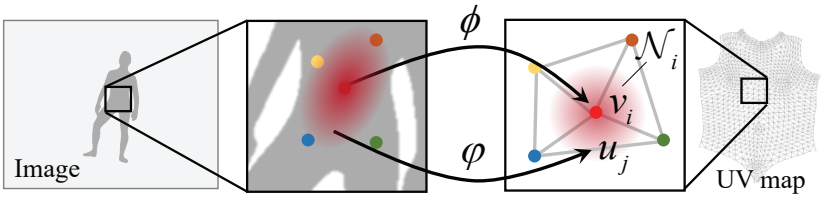


Figure 2: We approximate the discrete feature mapping ϕ using a continuous mapping ψ, i.e., the feature in the i-th vertex is approximated by the weighted average of the nearby features, which enables differentiable feature learning.

where $f_{\mathrm{E}}(\mathcal{I};\theta_E)$ denotes an image feature encoder that extracts feature maps $\mathbf{F}$ from the input image $\mathcal{I}$, and $f_{\mathrm{D}}(\cdot;\theta_D)$ denotes the mesh decoder that takes a set of the image features to reconstruct the 3D human pose and shape.

Existing multiview body pose and shape reconstruction methods extend single view methods by leveraging camera calibration to fuse multiview image information [12, 15]. In contrast, we leverage the observed human pose as a semantic cue to fuse visual features from multiview images. Consider a discrete mapping $\phi(\mathbf{x},\mathcal{I})$ that maps a pixel in an image $\mathcal{I}(\mathbf{x})$ to a vertex on the 3D body surface, $\mathbf{z}_i$. This offers a new way to align visual features in a common vertex coordinate system across views:

$$\mathbf{F}_i = f_{\mathrm{E}}(\mathbf{x},\mathcal{I};\theta_E), \quad \text{where} \quad i = \phi(\mathbf{x},\mathcal{I}), \tag{2}$$

and the encoder $f_{\mathrm{E}}(\cdot)$ learns the corresponding image features $\mathbf{F}_i$ for a certain body vertex $\mathbf{z}_i$. We assume the 3D body vertex index i are consistent across people, views, and poses.

Figure 1(b,c) illustrates the comparison between geometric fusion that relies on camera calibration to align visual features from multiview images via epipolar line [12] or triangulation [15]. In contrast, our approach uses mappings from image points to 3D body vertices, $\phi(\mathbf{x},\mathcal{I})$, to semantically align features without camera calibration.

3.2 Local Feature Registration

A key challenge of learning visual features from Equation (2) lies in the non-differentiability of the index map ϕ. The domain (pixel coordinates $\mathbf{x}$) is not continuous, and any $\mathbf{x}$ that does not map to an integer index is undefined. We address this challenge by relaxing the discrete domain, *i.e.*, representing the feature $\mathbf{F}_i$ using a set of continuous vicinity features from a continuous mapping.

Using a dense (continuous) keypoint map $\psi(\mathbf{x},\mathcal{I})$ that maps pixel coordinates $\mathbf{x} \in \mathbb{R}^2$ to continuous UV coordinates $\mathbf{u} \in \mathbb{R}^2$, we define the feature of the i-th body vertex $\bar{\mathbf{F}}_i$ as a weighted sum of all features nearby in the UV space (Figure 2):

$$\bar{\mathbf{F}}_i = \frac{\sum_{\mathbf{u}_j \in \mathcal{N}_i} w_j f_{\mathrm{E}}(\psi^{-1}(\mathbf{u}_j),\mathcal{I};\theta_E)}{\sum_{\mathbf{u}_j \in \mathcal{N}_i} w_j}. \tag{3}$$

Here, $\mathcal{N}_i$ is a set of neighboring locations of the i-th vertex for which the inverse are well-defined pixel coordinates; w_j is the weight for the visual feature corresponding to UV coordinates $\mathbf{u}_j$. Given the location of the i-th vertex in the UV coordinate system $\mathbf{v}_i$, its neighborhood is defined as:

$$\mathcal{N}_i = \{\mathbf{u}_j : d(\mathbf{u}_j,\mathbf{v}_i) < \tau,\ \psi^{-1}(\mathbf{u}_j) \in \Lambda_{\mathcal{I}}\}, \tag{4}$$

where $d(\cdot,\cdot)$ measures the geodesic distance between two points in the 3D body surface and $\Lambda_{\mathcal{I}}$ is the lattice of the input image. We further define the weight by applying an RBF kernel on the geodesic distance, *i.e.*,

$$w_j = \exp\left(-\frac{d(\mathbf{u}_j,\mathbf{v}_i)^2}{2\sigma^2}\right), \tag{5}$$

where w_j is the weight for the j-th neighbor $\mathbf{u}_j$, and σ controls the effective range of the neighbors. To avoid numerical instability, in practice, we assign zero for features when $\sum_{\mathbf{u}_j \in \mathcal{N}_i} w_j < 10^{-6}$.

3.3 Multiview Feature Fusion via Self-attention

Depending on the camera configuration and the human pose, some views can be more informative for reconstructing a certain body area than the others which suffer from occlusion or depth ambiguity.

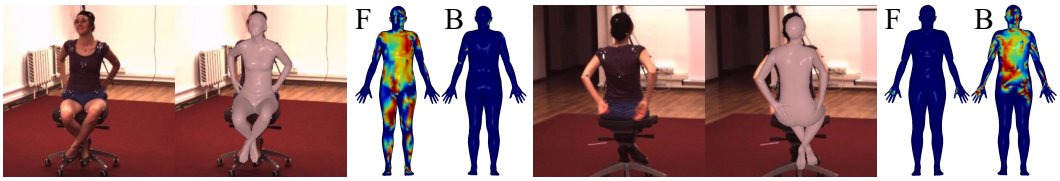


Figure 3: We use a self-attention mechanism to fuse the local and global features across views. The intensity in the heatmaps (right) indicates the attention predicted by the images (left). "F" denotes the front view and "B" denotes the back view. The attention is expected to highly correlated with the visible regions of the body because the features can be confidently predicted from the encoder. The reconstructed 3D mesh overlays the input image.

We address this by a self-attention mechanism that learns to assign a weight for each view at different body areas.

With Equation (2), visual features from multiview images can be aligned with respect to the 3D body surface indices, which allows us to fuse the image features from multiview images:

$$\bar{\mathbf{F}}_i^* = \sum_{c=1}^{N_c} \alpha_i^c \, \bar{\mathbf{F}}_i^c, \tag{6}$$

where $\bar{\mathbf{F}}_i^*$ is the i-th fused image feature and α_i^c is the weight indicating the feature confidence for the c-th camera view, respectively.

We represent the feature confidence α_i^c as a self attention, and it can be learned in an unsupervised way:

$$a_c(\mathbf{u}) = f_{\mathrm{A}}(f_{\mathrm{E}}(\psi^{-1}(\mathbf{u}), \mathcal{I}_c; \theta_E); \theta_A), \tag{7}$$

where $a_c(\mathbf{u})$ is the attention score at the UV coordinates $\mathbf{u}$ from the c-th view, and $f_{\mathrm{A}}(\cdot; \theta_A)$ is the learned function that predicts the attention given the set of features from the c-th view, respectively.

Similar to Equation (3), we represent the feature confidence as a weighted sum of attention scores from neighboring locations:

$$\alpha_i^c = \frac{\exp(\beta_i^c)}{\sum_{k=1}^{N_c} \exp(\beta_i^k)}, \quad \text{where} \quad \beta_i^c = \frac{\sum_{\mathbf{u}_j \in \mathcal{N}_i} w_j^c a_c(\mathbf{u}_j)}{\sum_{\mathbf{u}_j \in \mathcal{N}_i} w_j^c}, \tag{8}$$

and w_j^c is the weight for the j-th neighbor, $\mathbf{u}_j$, in the UV-coordinate system, as defined in Equation (4-5). We apply the softmax operation to normalize feature confidence across views. Figure 3 illustrates the transferred learned attention values on the body mesh. Overall, the learned attention scores are consistent with our intuition about contributions to body regions from different camera views.

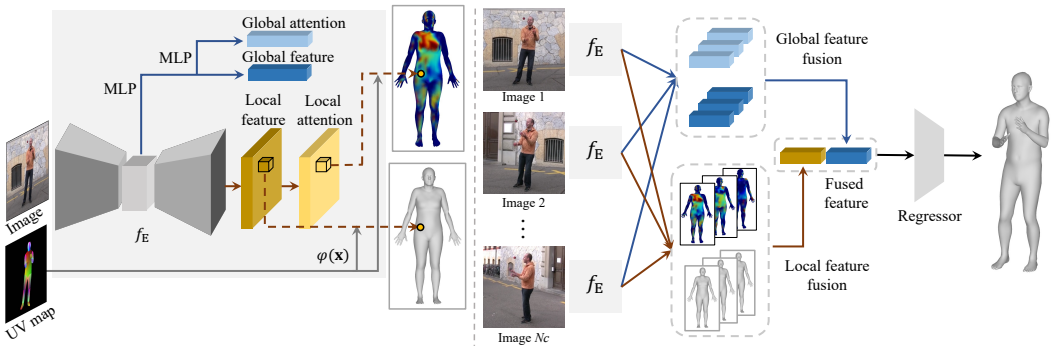
While per-pixel features are useful for capturing fine-grained geometry, the holistic shape such as pose can be better represented by a global feature. In conjunction with per-pixel features, we incorporate global features. To fuse global features across views, we utilize a self-attention mechanism similar to local feature fusion with minor modifications. Instead of learning a single score for each view, we learn an attention score for each channel of the global feature from each view and then normalize it to be used for combining global features across views.

3.4 Loss and Network Design

Given the neural network output $\mathbf{\Theta}$ and subsequent $\mathbf{Z}$, we learn our model in an end-to-end manner by minimizing the following loss:

$$\mathcal{L} = \sum_{\widehat{\mathbf{\Theta}} \in D_{\mathrm{V}}} \|\mathbf{\Theta} - \hat{\mathbf{\Theta}}\|_2 + \lambda_{\mathrm{P}} \sum_{\widehat{\mathbf{P}}_{\mathrm{3D}} \in D_{\mathrm{P}}} \|J\mathbf{Z} - \hat{\mathbf{P}}_{\mathrm{3D}}\|_2^2 + \lambda_{\mathrm{K}} \sum_{\widehat{\mathbf{P}}_{\mathrm{2D}} \in D_{\mathrm{K}}} \|\Pi(J\mathbf{Z}) - \hat{\mathbf{P}}_{\mathrm{2D}}\|_2^2, , \tag{9}$$

where three terms are the losses for model parameters, 3D joint error, and 2D joint error, respectively, and λ_{P} and λ_{K} balance the loss magnitudes. D_V, D_P, and D_K are their datasets, *i.e.*, we use multiple datasets to improve generalizability. $\widehat{\mathbf{\Theta}} \in \mathbb{R}^{85}$, $\widehat{\mathbf{P}}_{\mathrm{3D}} \in \mathbb{R}^{3M}$, and $\hat{\mathbf{P}}_{\mathrm{2D}} \in \mathbb{R}^{2M}$ are the ground truth model parameters, 3D vertices, 3D joint positions, and 2D joint positions, with N and M denoting the number of vertices and joints, respectively. $J \in \mathbb{R}^{3M \times 3N}$ is the pre-defined 3D joint regression



(a) Encoder (b) Multiview fusion decoder

Figure 4: (a) We design a new neural architecture for the feature encoder f_{E} that outputs visual features for 3D reconstruction. The global feature is computed from the bottleneck feature, and its global attention is obtained by transforming the global feature through fully connected layers. The encoder also output per-pixel (local) features and their attention that are mapped to body vertex coordinates via dense keypoints (UV map). (b) We combine the global and local features from the multiview images using their attention. The fused local features and global features are concatenated and feed into a MLP regressor for predicting 3D shape and pose.

		Multiview reconstruction			Single view reconstruction	
Method	Type	Calibration-free	MPJPE-PA	MPVPE-PA	MPJPE-PA	MPVPE-PA
HMR [19]	Mono	✓	57.8 ± 10.7	67.7 ± 15.4	56.8	65.5
GraphCMR [21]	Mono	✓	50.9 ± 9.1	59.1 ± 13.4	50.1	56.9
SPIN [20]	Mono	✓	44.5 ± 7.9	51.5 ± 11.8	41.1	49.3
DecoMR [42]	Mono	✓	42.0 ± 8.8	50.5 ± 14.1	**39.3**	47.6
PyMAF [44]	Mono	✓	N/A	N/A	40.5	N/A
MV-SPIN [35]	Multi	✗	35.4	N/A	N/A	N/A
LVS [35]	Multi	✗	**32.5**	N/A	N/A	N/A
Liang [25]	Multi	✓	48.5	57.5	59.1	69.2
Ours	Multi	✓	33.0	**34.4**	41.6	**46.4**

Table 1: Comparison between the proposed method and existing single view and multiview methods on Human3.6M dataset. For each method, we specify the supported type of input and whether camera calibration is required. We evaluate the performance of each method on both multiview and single view reconstruction. Since multiview reconstruction is not supported by single view methods, we report the mean and standard deviation of the error after evaluating the method on each of the input views. In addition, we report the multiview methods (MV-SPIN [35] and LVS [35]) that use camera calibration for fusion as a reference, i.e., it is expected to provide the performance upper bound of our calibration-free method. MPJPE-PA and MPVPE-PA are reported in millimeter.

matrix that linearly maps the 3D vertices to the 3D joints [27], and $\Pi(\cdot)$ is the operation of camera projection that projects the 3D joints to 2D joints in the image plane. $\mathcal{L}_{\Theta}$ measures the Euclidean error between the ground truth model parameters and predicted ones, $\mathcal{L}_{\mathrm{P}}$ measures the 3D error of the joints, i.e., $J\mathbf{Z}$ is the M joint locations, and $\mathcal{L}_{\mathrm{K}}$ measures the 2D projection of the joint locations.

Note that neither the local feature registration nor the multiview fusion assumes a fixed camera configuration or a fixed number of views. We fully leverage this advantage in our training pipeline: at each step, we randomly choose a subset of views from each data sample to train the model. Figure 4 shows the overview of our training pipeline. As a result, we are able to mix multiview datasets and single view datasets to train a model that can handle the variable number of views.

3.5 Implementation Details

We design the encoder $f_{\mathrm{E}}(\cdot;\theta_E)$ as a ResNet-50 backbone [11], that takes a $224\times224\times3$ image as an input and outputs a global feature vector with 256 dimensions and a 56×56 local feature map with 256 dimensions. The local attention function $f_{\mathrm{A}}(\cdot;\theta_A)$ is modeled by a sequence of convolutional layers, while the per-channel attention for global feature is modeled by a fully connected layer. We set $\tau = 0.05$ and $\sigma = 2.33 \times 10^{-2}$. We use a pre-trained DensePose model for ψ, which partitions the whole body surface to 24 different parts where UV map is defined for each part. In practice, $d(\mathbf{u}, \mathbf{v})$ in Equation (4) and Equation (5) is set to infinity if $\mathbf{u}$ and $\mathbf{v}$ are from different parts. Another important detail is that for a vertex that corresponds to multiple 2D UV coordinates, we take the

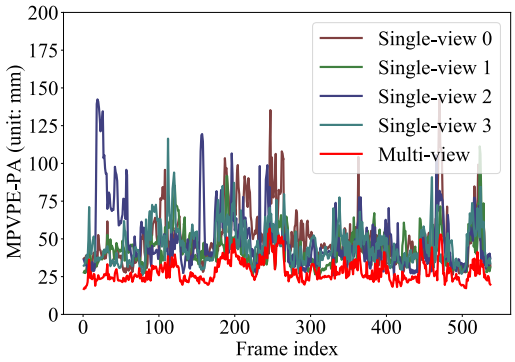


Figure 5: We compare ours that infer jointly from multiple views with SPIN [20] infer on each view independently on a sequence from Human3.6M [14]. Our multi-view results are equivalent to or better than single-view result from the best view for almost all frames.

Variant	MPJPE-PA	MPVPE-PA
W/o local feature	37.7	42.0
Average fusion	35.4	37.8
Max fusion	35.2	37.5
Att. fusion (ours)	**33.0**	**34.4**
Att. fusion (3 views)	34.2	36.6
Att. fusion (2 views)	37.3	40.2
Att. fusion (1 view)	44.1	50.1

Table 2: We compare different variants of the proposed method to verify the design decisions. All the variants are trained on the union of Human3.6M and UP-3D and evaluated on Human3.6M (use all four views unless specified). MPJPE-PA and MPVPE-PA are reported in millimeter.

minimum when computing the geodesic distance, *i.e.*, replacing $d(\mathbf{u}_j, \mathbf{v}_i)$ with $\min(\{d(\mathbf{u}_j, \mathbf{v}_i^k)\})$ in Equation (4) and Equation (5), where $\mathbf{v}_i^k$ is the k-th 2D UV coordinates that the same i-th vertex corresponds. In fact, this serves as an approximation of ϕ using ψ. For training, we use a batch size of 16 and a learning rate of 3×10^{-5}. We set $\lambda_{\mathrm{P}} = \lambda_{\mathrm{K}} = 5.0$.

4 Experiments

We validate our calibration-free multiview fusion approach on multiple datasets varying from indoor to outdoor, controlled to in-the-wild environments. We refer the reader to check additional results, experiments and implementation details in the *supplementary material*.

4.1 Datasets

Human3.6M [14] is a large-scale multiview dataset with ground truth 3D human pose annotation. We follow the standard training/testing split: using subject S1, S5, S6, S7 and S8 for training, and subject S9 and S11 for testing. We reconstruct the ground truth 3D human mesh in the format of the SMPL [27] model by applying MoSh [26] to the sparse 3D MoCap marker following previous works [19, 21, 42, 41].

UP-3D [22] is an in-the-wild single view dataset containing 8,515 images annotated with ground truth 2D keypoints and 3D body mesh. The 3D body mesh is obtained by fitting the SMPL [27] model to images from the Leeds Sports Pose [18], MPII Human Pose [1], and FashionPose [4] datasets. We use the standard training split [21, 42].

MARCOnI [5] is a multiview dataset including both indoor and outdoor images for evaluating marker-less motion capture methods. This dataset contains 12 real-world sequences captured by 3-16 cameras with varying modalities from 1-2 subjects each in different scenes, *e.g.*, Soccer, Kickbox, Juggling, etc. We use this dataset to qualitatively evaluate our method and test the generalizability of the model.

VBR [2] is a multiview dataset that include challenging social videos captured by multiple members in a scene. The baseline between cameras are wide. It contains three synchronized multiview sequences including Juggler, Magician, and Rothman. We use this dataset to qualitatively evaluate our method and test the generalizability of the model.

Social Videos are a new multiview dataset we collected for qualitative study. It features multiple persons performing dynamic group activities, such as playing basketball and name tag ripping, which introduces frequent occlusion and disocclusion. Up to 9 subjects are captured from 4 hand-held cameras in 6 sequences. We run an off-the-shelf 2D human pose detection and tracking tool [6] with manual correction and multiview association to obtain ground truth per-person multiview tracking data.

4.2 Metrics

Since the proposed method reconstructs the 3D body shape and pose (mesh) and uses the local features for recovering fine details, we evaluate of the method using Mean Per-Vertex Position Error (MPVPE) and Mean Per-Joint Position Error (MPJPE), both after Procrustes Analysis (PA). MPJPE-PA primarily evaluates the pose estimation accuracy of the model, whereas MPVPE-PA evaluates the accuracy of the reconstructed body shape.

4.3 Comparison with State-of-the-art Methods

Baselines HMR [19] reconstructs the 3D body shape and pose from a monocular image by estimating the SMPL model parameters in an iterative manner. GraphCMR [21] leverages a graph convolutional network to first reconstruct the non-parametric mesh, which is then refined by fitting SMPL model parameters. SPIN [20] further improves the accuracy by combining iterative fitting and regression into a self-improving loop. PyMAF [44] is a more recent monocular method that incorporate mesh alignment feedback loop for better estimation. Multiview SPIN and LVS [35] are two methods that leverage multiview calibration for estimating SMPL model parameters as a reference. To our best knowledge, Liang [25] is the only multiview method that does not require known camera calibration, and it utilizes a multi-stage network that is scalable to any number of views.

Multiview Reconstruction We reconstruct 3D body shape and pose using multiview images from Human3.6M dataset as summarized in Table 1, middle three columns. For all the single view methods, we observe large variations in accuracy due to the varying viewing angle. Multiview methods overcome this limitation by fusing information from multiple views and producing a unified output. Figure 5 shows such a comparison between ours and SPIN [20] at continuous frames of a Human3.6M [14] sequence under MPVPE-PA. Compared to the one proposed by Liang [25], our method outperforms in both MPJPE-PA and MPVPE-PA. We also report the performance of the multiview methods (MV-SPIN [35] and LVS [35]) that rely on the camera calibration as a reference, i.e., it is expected to provide the performance *upper bound* of our calibration-free method.

Single View Reconstruction Our method can be applied to an arbitrary number of cameras, including a monocular camera, *without* retraining the model. Applying our method on single view input, however, does not benefit from the attention-based multiview feature fusion. Evaluation results can be found in Table 1, last two columns. Even without specifically targeting for the single view use case, the results produced by our method is comparable with existing ones on MPJPE-PA while outperform others on MPVPE-PA. Further, our fusion method is complementary to the single view reconstruction approaches where we can extract the global and local features with a minor modification. We expect that a stronger performance can be achieved in the multiview reconstruction if a stronger baseline can be used.

4.4 Ablation study

For all the following experiments, we train the model using the Human3.6M and UP-3D datasets, and evaluate the model on the test split of Human3.6M.

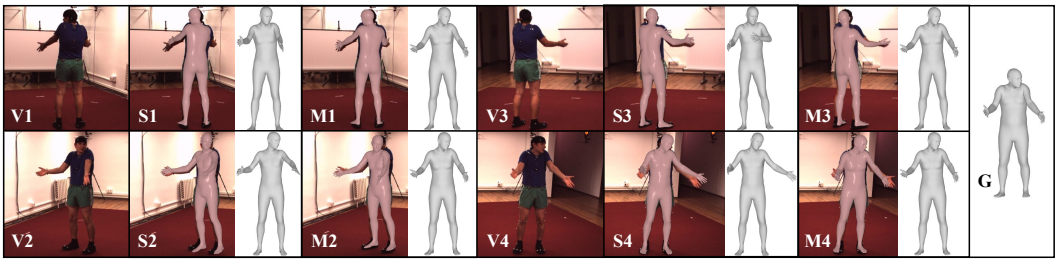
Effectiveness of Local Features We compare our model with the variant that is trained with the global features only. The features across views are fused with the self-attention mechanism. Table 2 shows that adding local features effectively improves 3D reconstruction accuracy.

Effectiveness of using Self-attention To demonstrate the effectiveness of our attention based fusion, we compare ours with other alternative fusion methods: average pooling and max pooling. Comparison in Table 2 demonstrates that the model trained with our self-attention fusion outperforms the other fusion methods by a large margin.

Benefit of using More Views To investigate the benefit of using more views for 3D reconstruction, we evaluate the performance of our model with varying number of views. Note that re-training the model is not necessary since our approach allows taking an arbitrary number of views as input. The results are summarized in Table 2. Using more views helps improve the performance.

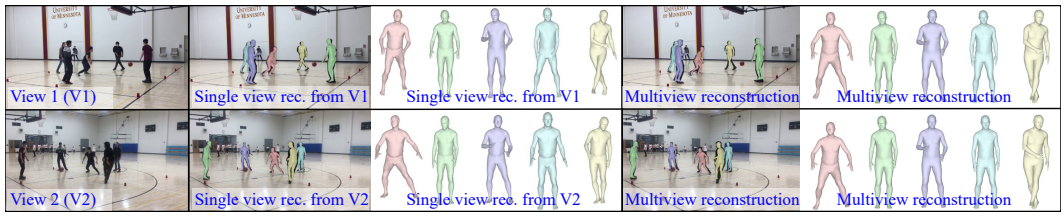
4.5 Qualitative Results

We apply our method to reconstruct 3D body shape and pose using multiview images and show the qualitative comparisons on images from Human3.6M [14] and self-collected Social Videos dataset against SPIN [20] in Figure 6.



(a) Qualitative comparison with SPIN [20] on Human3.6M [14].

(b) Qualitative results on Unstructured VBR (A, B, C), and MARCOnI (D).



(c) Qualitative comparison with SPIN [20] on our Social Videos dataset.

Figure 6: (a) While reconstruction results of both may have similar level of alignment with image after reprojection, 3D reconstructions of single-view method are not consistent across views due to bias of viewpoint. "V" stands for "view", "S" stands for "Single-view reconstruction [20]", "M" stands for "Multiview reconstruction" of ours, "G" stands for "Ground truth". (b, c) We reconstruct 3D body shape and pose from uncalibrated multiview images of characteristic social videos in diverse scenes, including Unstructured VBR, MARCOnI, and our Social Videos. While single view reconstruction produces inconsistent poses depending on views, ours produces a consistent pose invariant to views as leveraging multiview information jointly.

5 Summary and Limitation

We present a novel 3D human body reconstruction method that uses multiview images without calibration. Instead of relying on geometric camera calibration, our method leverages the observed human subject as a semantic calibration target that can align the visual features from multiview images. The visual features are learned in the 3D body surface coordinate, established by the dense keypoint mapping, that is agnostic to views and poses. We reason about self-occlusion via the confidence of the aligned features, which allows us to effectively fuse the features from multiview images. The fused features are fed into a MLP regressor to regress model parameters. Experiments show that our calibration-free method is able to effectively leverage multiview information in a principled way, outperforming state-of-the-art single view approaches with a post-hoc multiview fusion. Our approach is readily applicable to real world scenarios, including 3D reconstruction from social videos.

Our reliance on the dense keypoint estimation can be a double-edged sword. While it establishes the dense correspondences without camera calibration, the erroneous estimation of the dense keypoints may lead to misalignment of the local features, resulting in suboptimal 3D reconstruction. This is more prominent when the body is severely occluded or the appearance is ambiguous, which often leads to left-right or front-back confusion in the prediction. Our attention mechanism can mitigate fusing erroneous features if the majority of dense keypoint estimation from multiview images are correct. In practice, we found that the dense keypoint estimation [10] is highly reliable, which makes our method applicable to in-the-wild social videos.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014.

[2] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. In *ACM SIGGRAPH*, 2010.

[3] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, 2016.

[4] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2013.

[5] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[7] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[8] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE/CVF International Conference on Computer Vision*, 2009.

[9] R. A. Guler and I. Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[10] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[12] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[13] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision*, 2017.

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[15] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[16] A. S. Jackson, C. Manafas, and G. Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *European Conference on Computer Vision*, 2018.

[17] Y. Jafarian and H. S. Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[18] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

[19] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[20] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[21] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[22] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[23] X. Li, Y. Liu, H. Joo, Q. Dai, and Y. Sheikh. Capture dense: Markerless motion capture meets dense pose estimation. *arXiv*, 2018.

[24] Z. Li, M. Oskarsson, and A. Heyden. 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[25] J. Liang and M. C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[26] M. Loper, N. Mahmood, and M. J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics*, 2014.

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 2015.

[28] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision*, 2018.

[29] G. Pavlakos, N. Kolotouros, and K. Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[30] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[31] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng. Cross view fusion for 3d human pose estimation. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[32] E. Remelli, S. Han, S. Honari, P. Fua, and R. Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[33] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[34] A. Sengupta, I. Budvytis, and R. Cipolla. Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[35] S. Shin and E. Halilaj. Multi-view human pose and shape estimation using learnable volumetric aggregation. *arXiv*, 2020.

[36] H. Tu, C. Wang, and W. Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, 2020.

[37] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *European Conference on Computer Vision*, 2018.

[38] R. Xie, C. Wang, and Y. Wang. Metafuse: A pre-trained fusion model for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[39] Y. Xu, S.-C. Zhu, and T. Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[40] Y. Yao, Y. Jafarian, and H. S. Park. Monet: Multiview semi-supervised keypoint detection via epipolar divergence. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[41] Z. Yu, H. Yu, L. Sha, S. Ganguly, and H. S. Park. Semi-supervised dense keypoints using unlabeled multiview images. In *Annual Conference on Neural Information Processing Systems*, 2021.

[42] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang. 3d human mesh regression with dense correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[43] H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun. Learning 3d human shape and pose from dense body parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[44] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[45] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Supplementary Material.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [No]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Supplementary Material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]