Computer Vision and Image Understanding 236 (2023) 103830

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

View consistency aware holistic triangulation for 3D human pose estimation

Xiaoyue Wan, Zhuo Chen, Xu Zhao *

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200000, China

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
41A05
41A10
65D05
65D17

Keywords:
3D pose estimation
View consistency
Pose coherence
Anatomy prior

ABSTRACT

The rapid development of multi-view 3D human pose estimation (HPE) is attributed to the maturation of monocular 2D HPE and the geometry of 3D reconstruction. However, 2D detection outliers in occluded views due to neglect of view consistency, and 3D implausible poses due to lack of pose coherence, remain challenges. To solve this, we introduce a Multi-View Fusion module to refine 2D results by establishing view correlations. Then, Holistic Triangulation is proposed to infer the whole pose as an entirety, and anatomy prior is injected to maintain the pose coherence and improve the plausibility. Anatomy prior is extracted by PCA whose input is skeletal structure features, which can factor out global context and joint-by-joint relationship from abstract to concrete. Benefiting from the closed-form solution, the whole framework is trained end-to-end. Our method outperforms the state of the art in both precision and plausibility which is assessed by a new metric.

1. Introduction

3D human pose estimation (HPE) is a significant computer vision problem with numerical applications such as human behavior analysis, X-reality, etc (Wang et al., 2021). To estimate 3D pose, there are two sensor setting streams: monocular (Martinez et al., 2017; Pavlakos et al., 2017a; Xu and Takano, 2021) and multi-view (Gavrila and Davis, 1996; Burenius et al., 2013). In this paper, we focus on multi-view 3D HPE, for its capability to estimate absolute 3D position without inherent depth ambiguities which monocular suffers.

One of the most common frameworks (Iskakov et al., 2019; Dong et al., 2019; Remelli et al., 2020; Kocabas et al., 2019) of multi-view methods follows a two-step procedure: (1) detect 2D keypoints of human skeleton at each view separately, (2) apply Linear Triangulation (LT) which utilizes epipolar geometry (Hartley and Zisserman, 2003) to reconstruct 3D pose. The framework is elegant because 2D detectors can be off-the-shelf and closed-form solution LT enables end-to-end training but without any learning cost. However, there are still two main drawbacks: (1) 2D keypoints detected in each view are independent of each other, and will be hampered by the occlusion and overlap due to lack of view consistency. (2) LT in step 2 calculates each 3D joint individually, neglecting the global context of whole pose. Hence, it is unable to identify the 3D outliers, which usually causes implausible poses.

To solve the first problem, Multi-View Fusion (MVF) module is proposed to refine the 2D keypoint by establishing view correlations. We argue that multiple image points projected from a 3D point share similar representations. In another word, two most similar points in different views are mostly intersected to one 3D point. According to this assumption, MVF utilizes keypoints detected in source views to generate pseudo heatmaps which represents the probability distribution the keypoint localized in reference view through feature matching. These pseudo heatmaps can guide the reference keypoints to perceive other views. There are also some works aimed to enhance view consistency through feature fusion: the fully-connected CrossView (Qiu et al., 2019) and the epipolar sample fusion in Epipolar Transformer (He et al., 2020). But, MVF primarily emphasizes heatmap generation and fusion which is more intuitive and the utilization of the detected keypoint location makes calculation more efficient.

Then to boost the plausibility of 3D poses, Holistic Triangulation (HT) with anatomy constraints is proposed, which enables all 3D keypoints to gain access to pose coherence through 2D–3D phase. Firstly, we modify the formulation of objective function so that all joints can be inferred as an entirety. Then, to model the joints linear dependence in the objective function, a PCA reconstruction term is injected. By doing so, joints are coupled in an abstract PCA subspace spanned by the principle components, which contains the global context of whole pose. Human anatomy prior therefore is implicitly introduced. Furthermore, to make the prior more explicit, PCA feature is extended from keypoint position to skeletal structure feature by applying kinematic chain space (KCS) (Wandt et al., 2018). Benefiting from the linear property of PCA, HT is still closed-optimized and differentiable, which maintains the elegance of LT.

* Corresponding author.
E-mail address: zhaoxu@sjtu.edu.cn (X. Zhao).

https://doi.org/10.1016/j.cviu.2023.103830
Received 28 February 2023; Received in revised form 31 July 2023; Accepted 1 September 2023
Available online 7 September 2023

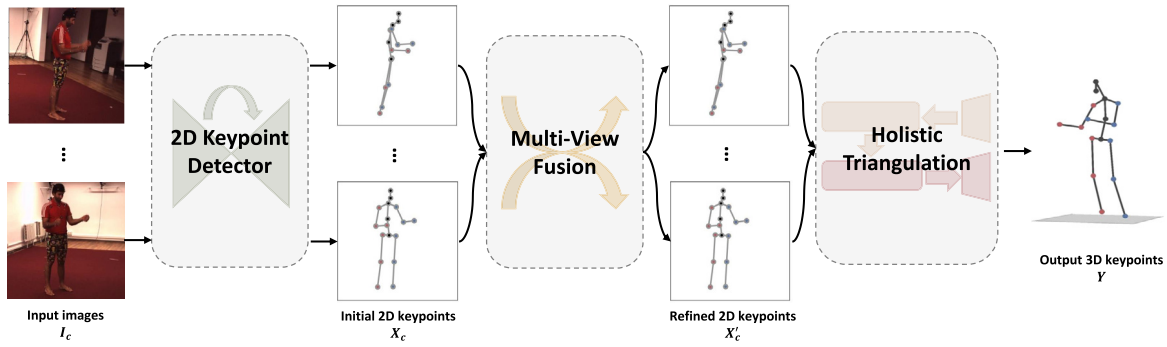

Fig. 1. The framework of our approach. 2D keypoint detector achieves 2D poses from multi-view RGB images. MVF refines the 2D results considering the views consistency. And HT generates the 3D pose under the constraints of the anatomy coherence.

Consequently, we integrate the 2D detector, MVF and HT into one end-to-end framework and introduce reprojected loss, bone length loss and joint angle loss to promote the view consistency and anatomy coherence during training procedure. In addition, a plausible-pose evaluation metric is proposed to fill in the gap of pose plausibility criterion.

Without bells and whistles, MVF-HT method exhibits competitive performance with state-of-the-art techniques, surpassing them in both precision, plausibility and generalization. Moreover, the anatomy prior extracted by PCA is explored through visualization. The main contributions are summarized below:

- We propose an efficient and intuitive MVF module to enhance the view consistency in 2D keypoint estimation. MVF refines 2D keypoint p through perceiving the possible position the same keypoints in other views may localize in the view of p.
- To the best of our knowledge, this is the first work that reconstructs the entire 3D pose at once using the triangulation framework. By integrating pose prior, 2D observations, and geometric constraints within the triangulation process and solving them in a closed form, the plausibility of the pose estimation is significantly improved.
- Our framework can be trained end-to-end but without any learning cost in 2D–3D phase because of the closed-form solution of HT. And a plausible-pose evaluation metric is proposed to fill in the gap of pose plausibility criterion.

2. Related work

Multi-View 3D HPE. The current multi-view 3D HPE methods can be divided into two categories according to the aforementioned two steps. The first category focuses on enhancing the 2D pose estimator. In Qiu et al. (2019), He et al. (2020) and Remelli et al. (2020), 2D detectors are enabled to perceive 3D information in the process of 2D detection, where (Qiu et al., 2019; He et al., 2020) use the epipolar constraints to fuse features of corresponding views while (Remelli et al., 2020) directly generate a canonical representation using convolution network. Our MVF is similar to Qiu et al. (2019) and He et al. (2020), but uses the results of 2D detector to sample joint feature and generate the corresponding pseudo heatmap to provide the assistant for the reference view.

The other category focuses on the second procedure which lifts 2D keypoints to 3D poses. The approach can be summarized as learning-based and optimization-based. In Iskakov et al. (2019), Dong et al. (2019), Remelli et al. (2020) and Kocabas et al. (2019), based on the camera projection geometry and multi-view 2D points, triangulation (Hartley and Zisserman, 2003) is used to obtain 3D results by SVD or Least-Square method. In Remelli et al. (2020), a lightweight DLT method is proposed and exceeds the SVD in time cost. In Kadkhodamohammadi and Padoy (2021), triangulation is replaced with a convolutional network to learn the lifting process. In Burenius et al. (2013), Pavlakos et al. (2017b) and Qiu et al. (2019), the human skeleton is modeled as 3D-PSM to establish the potential function combining the 2D observation and skeletal bone length constraints. 3D convolution is applied in Iskakov et al. (2019) and Tu et al. (2020) to make the inference directly from a volume which is aggregated by multi-view 2D features. PSM and learning-based methods have disadvantages in high computing and time consumption. Conventional triangulation methods (Hartley and Zisserman, 2003) only utilize the observation information and geometric constraints but ignore the skeletal prior. Our work not only inherits the cost advantage of triangulation but also injects anatomy prior to maintain the pose coherence.

Anatomy Prior Extraction. The prior extraction can be classified as model-based and learning-based. Model-based methods rely on a predefined model to interpret the body structure and utilize model constraints to represent the pose prior. Optimization fitting of the model generates a plausible pose. Zhou et al. (2016) uses the basis pose as the dictionary to represent the pose prior. Bogo et al. (2016) employs the SMPL model to limit the result. Although the model brings strong constraints, the iterative optimization used to solve the dictionary weights or SMPL parameters is time-consuming. In contrast, learning-based methods encode features to enhance joint correlations or generate distribution to represent prior. GCN (Cai et al., 2019; Liu et al., 2020; Zhao et al., 2019) and Attention (Guo et al., 2021) are used to capture the relationship between two joints. In Yang et al. (2022), Chen et al. (2019b) and Habibie et al. (2019), the encoder–decoder is applied to create the latent space which is used to mine the inter-dependencies between joints. GAN (Tian et al., 2021; Wandt and Rosenhahn, 2019; Chen et al., 2019a) and VAE (Pavlakos et al., 2019) are another kind of model to capture the distribution of poses. Learning-based methods leverage the power of deep network to capture more generic constraints, but at the cost of more computing resources and network complexity. In Malleson et al. (2020) and Romero et al. (2017), PCA is employed as a dimensionality-reduction method to acquire pose prior. The linear and network independent properties of PCA attract us. Hence, we use PCA with skeletal structure features as input to learn the relationships from near and distant joints.

3. Methodology

The overview of the proposed method is depicted in Fig. 1. There are three major modules: (1) 2D Keypoint Detector, to detect multi-view 2D joint locations respectively, where an off-the-shelf ResNet-152 backbone (Xiao et al., 2018) is directly applied. (2) Multi-View Fusion (MVF), to refine 2D poses considering the view consistency. (3) Holistic Triangulation (HT), to reconstruct the final 3D pose by closed-form optimization.

The input to the whole framework is a set of multi-view RGB images I_c, whose index is the number of the synchronized cameras and $c \in \{1, 2, \dots, C\}$. The output is 3D pose $Y = [\mathbf{y}_1^T, \mathbf{y}_2^T, \dots, \mathbf{y}_K^T]^T \in \mathbb{R}^{(3K,1)}$, where $\mathbf{y}_k = [x_k, y_k, z_k]^T$ and $K = 17$. Each image will be fed into the 2D

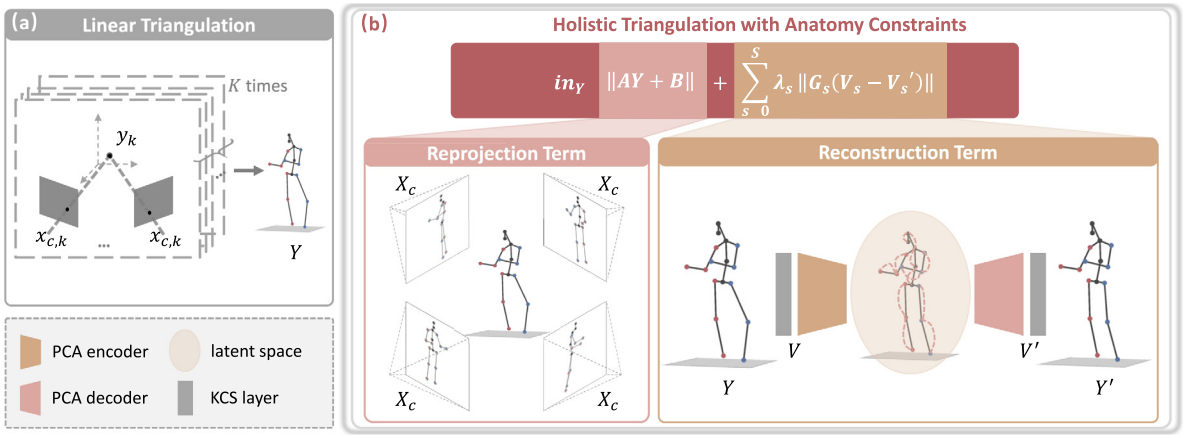

Fig. 2. The schematic diagram of (a) Linear Triangulation (LT) and (b) our Holistic Triangulation (HT) with anatomy constraints. There are two major differences between two methods: (1) LT reconstructs 3D keypoints separately and concatenates all keypoints to a pose, while HT reconstructs an entire 3D pose at once. (2) LT only consider about the geometric constraints, however, HT includes anatomy constraints extracted by PCA encoder–decoder.

detector to generate the initial 2D pose $X_c = [\mathbf{x}_{(c,1)}^T, \mathbf{x}_{(c,2)}^T, \dots, \mathbf{x}_{(c,K)}^T]^T \in \mathbb{R}^{(2K,1)}$, where $\mathbf{x}_{c,k} = [x_{(c,k)}, y_{(c,k)}]^T$ is the location of the kth joint in view c. Then, the MVF module obtains the refined 2D poses X_c' from the initial ones by fusing all heatmaps corresponding to different views. After that, HT reconstructs the 3D pose Y from the refined 2D poses through optimization. Finally, a loss function, takes multi-view consistency and whole pose coherence into account, supervises the network when end-to-end training.

In this section, we first introduce HT, since the goal of our task is 3D pose. Then, the MVF is introduced as an assistance to refine the 2D results. Finally, the overall loss function of the end-to-end framework will be present.

3.1. Holistic triangulation with anatomy constraints

LT (Hartley and Zisserman, 2003) is classic and elegant because of the closed-form solution, but is lack of joint-by-joint relation modeling. As depicted in Fig. 2(a), LT infers 3D position $\mathbf{y}_k$ of each keypoint separately for K times, and the keypoints are then concatenated to generate the 3D pose. To fix this issue, we propose Holistic Triangulation (HT), shown in Fig. 2(b), to reason the whole pose at once through reprojection and reconstruction term:

$$\min_Y \quad \|AY + B\| + \mathbb{H} \tag{1}$$

$$A = \begin{bmatrix} \mathbf{w}_1 \circ A_1 & O & \cdots & O \\ O & \mathbf{w}_2 \circ A_2 & \cdots & O \\ \vdots & \vdots & \vdots & \vdots \\ O & O & \cdots & \mathbf{w}_k \circ A_k \end{bmatrix}, \quad B = \begin{bmatrix} \mathbf{w}_1 \circ \mathbf{b}_1 \\ \mathbf{w}_2 \circ \mathbf{b}_2 \\ \vdots \\ \mathbf{w}_k \circ \mathbf{b}_k \end{bmatrix}$$

where $\circ$ is the Hadamard product. $A_k \in \mathbb{R}^{(2C,3)}$ is the first three columns of $\widetilde{A}_k$ and $\mathbf{b}_k \in \mathbb{R}^{(2C,1)}$ is the last column. $\widetilde{A}_k$ is same as LT (refer to supplementary). We draw the idea from Algebraic Triangulation (AT) (Iskakov et al., 2019) to add the learnable confidence $\mathbf{w}_k = [\omega_{1,k}, \omega_{1,k}, \dots, \omega_{C,k}, \omega_{C,k}]^T$ to mitigate the impact of 2D positions with low confidence.

However, owing to the blockwise linear independence in the reprojection term of Eq. (1), resulting vector of each block in Y has no difference from AT. To solve this problem, we introduce a reconstruction term $\mathbb{H}$.

Vanilla Reconstruction Term. The reconstruction term aims to enhance the blockwise linear dependence in A and inject anatomy coherence to Y. PCA (Hotelling, 1933), a simple but effective module is chosen to model the anatomy prior for three major reasons: (1) The PCA low-dimension latent space is capable to extract the correlations between different keypoints and factor out the generic pose global context. The generality of the pose context is guaranteed by the fact that training data contains various motions. And then we approach the estimated pose Y close to the PCA recovered pose Y' from the latent space, to inject the pose prior. (2) The linear property of PCA will not change the closed-form solution superiority of HT, which will not hinder end-to-end training. (3) PCA does not require additional network training, making it a more computationally efficient choice.

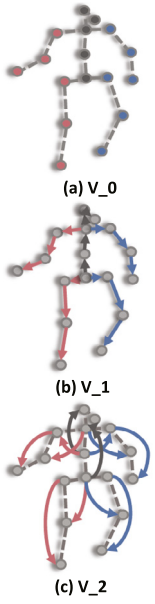

Fig. 3. Skeletal structure features.

Note that the training set of PCA is root-relative, $Y_{re} = Y - Y_{root} \in \mathbb{R}^{(3K,1)}$. And Y_{root} is the pelvis position which is estimated by LT. By adding a reconstruction term, the objective function is expressed as:

$$\min_Y \quad \|AY + B\| + \lambda \|Y_{re} - Y_{re}'\| \tag{2}$$

where $Y_{re}' = M^T M(Y_{re} - Y_{mean}) + Y_{mean}$ is the recovered pose; $M \in \mathbb{R}^{(D,3K)}$ is the feature extracting matrix of PCA encoder; $Y_{mean} \in \mathbb{R}^{(3K,1)}$ is the mean pose of PCA training set; and λ is a learnable weight of reconstruction term. Because of the convexity of Eq. (2), the 3D pose can be closed-form solved using Least-Square method (see supplementary for

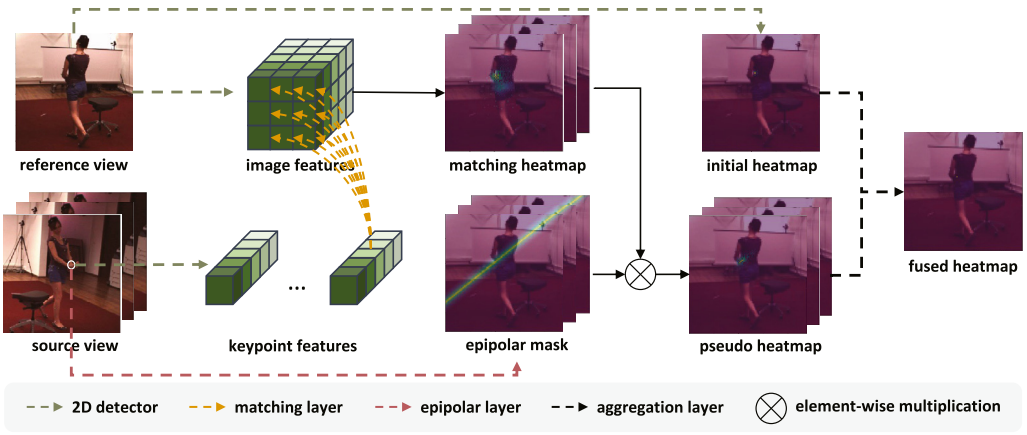

Fig. 4. The pipeline of MVF module. The reference view image provides the image features and initial heatmap while source views provide keypoint features. Matching heatmap is generated through a matching layer by comparing keypoint features with image features. And it is multiplied by an epipolar mask to avoid mismatching. Finally we aggregate initial heatmap of reference view and the pseudo heatmaps from source views to a fused heatmap.

proof):

$$(A^T A + \lambda N^T N)Y = \lambda N^T N(Y_{root} + Y_{mean}) - A^T B \tag{3}$$

where $N = I - M^T M \in \mathbb{R}^{(3K,3K)}$.

Skeletal Structure Feature Extraction Module. One inadequacy of the basic reconstruction term above is that the prior extracted is implicit. To address this, we transform the data from keypoint space to skeletal structure space, enhancing associations between joints and introducing explicit features.

To generate skeletal structure feature V, KCS (Wandt et al., 2018), a matrix multiplication algorithm to create vector between two selected points, is used to transform joints to joint-connected vectors by a mapping matrix $T \in \mathbb{R}^{(3J,3K)}$. And $G \in \mathbb{R}^{(3K,3J)}$ is applied to transform back.

The feature of connected joint with s hops is named as V_s. As shown in Fig. 3, both keypoints and bone vectors can be compatible with V_0 and V_1. Because longer distances will result in fewer extracted vectors with less information, only $hop = 0, 1, 2$ is defined. By fusing different V_s features, the objective function can be adapted to:

$$\min_{Y} \quad \|AY + B\| + \sum_{s=0}^{S} \lambda_s \|G_s(V_s - V_s')\| \tag{4}$$

where G_s remaps the reconstructed error from feature space back to keypoint space in order to keep two terms in the same dimension. Replacing $V\_s = T\_sY$, the solution is:

$$(A^T A + \sum_{s=0}^{S} \lambda_s H_s^T H_s)Y = \sum_{s=0}^{S} \lambda_s H_s^T H_s(Y_{root} + Y_{mean}) - A^T B \tag{5}$$

where $H_s = G_s N_s T_s \in \mathbb{R}^{(3K,3K)}$, G_s and T_s are mapping matrices of V_s feature, N_s is similar to N in Eq. (3) but is relative to different feature V_s.

3.2. Multi-view fusion 2D keypoint refinement

The initial 2D keypoints achieved by 2D backbone detector are independent from each view. To enhance the cross view correlations, we introduce the MVF module. The pipeline of MVF is illustrated in Fig. 4. To make the keypoints in the reference view consistent with other views, the pseudo heatmaps corresponding to the same keypoints in other views are generated. Concretely, the pseudo heatmap is the product of matching heatmap and epipolar mask, and represents the probability source view keypoint localizing in the reference view. After that, the initial heatmap are fused with pseudo heatmaps through an aggregation layer which is a $1 * 1$ convolution kernel to product the refined fused heatmap.

Matching Layer. The idea of cost volume in stereo matching methods (Kendall et al., 2017; Xu and Zhang, 2020) inspires us. The corresponding matching heatmaps H_{match}, which indicates the matching degree of the keypoints p' in the source view and all pixels $p(i, j)$ in the reference view, is generated. The pixel gets higher matching score as its features are better matched with p'. We also explore two types of matching strategy:

- inner dot: $\frac{1}{N}(F(i, j) \cdot F(p'))$
- fully connected layer: $\mathbf{w}^T \cdot cat(F(i, j), F(p'))$

where F represents the features generated by 2D backbone, $F(p')$ is the sampled feature of p' via bilinear interpolation, $cat()$ means concatenation and $\mathbf{w}$ is the learnable parameters.

Epipolar Layer. In stereo matching task, only the points in the horizontal direction will be compared because the given image pair are rectified. However, in matching layer, the p' is compared with all pixels in the reference view due to lack of rectification. The matching instability will be caused by the similar feature vectors of wrong pixels. To solve the problem, we generate the epipolar mask by the epipolar field (Ma et al., 2021) to limit the matching pixels locating near the epipolar line of p'. The epipolar field indicates the probability pixels $p(i, j)$ in the reference view lies in the epipolar line of p':

$$E(p, p') = (1 - |(\overrightarrow{c'p'} \times \overrightarrow{cc'}) \cdot \overrightarrow{cp(i,j)}|)^{\gamma} \tag{6}$$

where c, c' are camera centers and γ is the soft factor to control the epipolar field margin. The field gets narrower as γ gets bigger, $\gamma = 10$ is chosen empirically to generate epipolar mask.

3.3. Loss

The overall loss function consists of four parts: (1) Mean Square Error (MSE) between estimated 3D pose and groundtruth, (2) reprojected error: L2 loss of reprojected 2D pose and estimated 2D pose, (3) bone length loss: L1 loss of estimated bone vector and groundtruth and (4) joint angle loss of 3D poses:

$$L(Y) = L_{MSE}(Y, \hat{Y}) + \beta_{pj} L_{pj}(X', Y) + \beta_{bl} L_{bl}(Y, \hat{Y}) + \beta_{ja} L_{ja}(Y) \tag{7}$$

where $\hat{Y}$ is the groundtruth of the pose; β_{pj}, β_{bl} and β_{ja} are set to 0.1, 0.01, 0.01 separately by empirical results. Based on the common L_{MSE}, we subjoin the L_{pj} to enhance the multi-view consistency and L_{bl}, L_{ja} to promote the anatomy coherence.

Joint Angle Loss. The multivariate Gaussian Mixture Model (Reynolds et al., 2009) is used to model joint angle distribution $p(x_k)$, $x_k = [\sin\theta_k, \sin\varphi_k, \cos\varphi_k]^T$, and φ_k, θ_k are azimuth and polar angle of joint

in a local spherical coordinate system (Akhter and Black, 2015) (details in suppl.). And the joint angles with low probability are penalized:

$$L_{ja} = \frac{1}{K_{se}} \sum_{k=0}^{K_{se}} sigmoid\left(\left(p(x_k) - \frac{a}{2}\right) * (-\frac{10}{a})\right) \tag{8}$$

where K_{se} is the number of selected joints; a is probability border $p(x_k \pm 3\sigma)$, the angle with the probability $(0, a)$ should be penalized. So transformation $-\frac{a}{2}$, coefficient $\frac{10}{a}$ are used to offset $(0, a)$ to $(5, -5)$ to suit the variable domain of sigmoid.

4. Experiments

4.1. Datasets and evaluate metrics

Human3.6M Dataset. The Human3.6M (Ionescu et al., 2013) is one of the most universal 3D HPE dataset with 3.6 million annotations. The videos are acquired from 4 synchronized cameras in laboratory. We use Joint Detection Rate (JDR) to evaluate 2D pose, Mean Per Joint Position Error (MPJPE) to evaluate relative 3D pose and Percentage of Plausible Pose (PPP), elaborated in Section 4.2, to assess plausibility.

Total Capture Dataset. The Total Capture Dataset (Trumble et al., 2017) is a common dataset recorded by 8 cameras which are distributed over different pitch angles from top to bottom. The dataset contains various actions, including some challenging motions like crawling and yoga. Hence, cross-dataset experiments are executed on it to evaluate the generalization.

4.2. Plausible-pose evaluation metric

Plausible-Pose Protocol. A plausible pose should meet two requirements: all bones have appropriate length and all joints are flexed in a limited range. A suitable bone length should be as near as possible to the groundtruth, and a reasonable joint angle is located in an occupancy matrix $OC(\theta, \varphi)$ which indicates whether the angle pair (θ, φ) appears in the training set:

$$P_{bl}(BL) = \begin{cases} 1, & |\frac{BL}{\hat{BL}} - 1| < R \\ 0, & others \end{cases}, P_{ja}(\theta, \varphi) = \begin{cases} 1, & OC(\theta, \varphi) = 1 \\ 0, & others \end{cases} \tag{9}$$

where BL is predicted bone length, $\hat{BL}$ is bone length groundtruth and R is bone length proportion threshold. The morphology technique (Gonzales and Wintz, 1987) is used to smooth the occupancy matrix OC, so that the continuous feasibility space can be covered even though the (θ, φ) is discrete. Ultimately, a reasonable pose is:

$$P_p(Y) = \prod_{j=0}^{J} P_{bl}(BL_j) \prod_{k=0}^{K_{se}} P_{ja}(\theta_k, \varphi_k) \tag{10}$$

Plausible-Pose Metric. To evaluate the plausibility performance statistically in testing dataset, a new metric PPP is defined as:

$$PPP = \frac{1}{T} \sum_{t=1}^{T} P_p(Y_t) \tag{11}$$

where T is the number of testing samples, and the metric is divided according to the bone length threshold into PPP@R.

4.3. Comparison with state-of-the-art methods

We compare quantitative and qualitative performance with the state of the art using all views on Human3.6M, and conduct cross-dataset experiments on Total Capture.

Implementation Details. The 2D keypoint detection backbone is same as AT (Iskakov et al., 2019), to ensure that the performance is not influenced by differences in the 2D backbone. The low-dimension feature extraction matrix M and mean pose Y_{mean} of PCA are both generated by an augmented training set to cover a diverse range of pose distribution. The dataset consists of Human3.6M (Ionescu et al., 2013) and MPII-INF-3DHP (Mehta et al., 2017) which is a large dataset with over 1.3 million frames of samples. In order to avoid the influence of orientations diversity, the orientation normalization is applied in the training data. It should be clarified that hyperparameters and feature extraction strategies are determined by ablation study (refer to supplementary): the feature V_0, V_1, V_2 are fused and the corresponding PCA reserved dimension D are set to 25, 20, 15 respectively; and coefficient of reconstruction term λ are learnable with initial value 8000, 4000, 4000. To train the MVF module and the whole network, we only utilize Human3.6M or Total Capture (Trumble et al., 2017). Firstly, We train the MVF network with MSE loss of 2D keypoints for 2 epochs with a batch size of 12. The learning rate is initially set to 10^{-2} and decays every 25 000 iterations by a factor of 0.1. After that, the whole network which combines three modules are trained for 4 epochs with 10^{-4} learning rate under the supervision of loss function in Section 3.3. If not mentioned explicitly, the baseline is AT (Iskakov et al., 2019) method.

Quantitative Results on Human3.6M. We first evaluate the refined 2D results after MVF refinement module. Following convention, the threshold of JDR is set to the half of the head size. As shown in Table 1, MVF outperforms CrossView by at least 2.1% regardless of the matching strategy. And fcl MVF also surpass Epipolar Transformer. The improvement demonstrates that the initial keypoint location can be leveraged to generate reliable pseudo heatmap with fewer calculations.

To evaluate the 3D pose estimation, we first compare precision performance with state-of-the-art methods whose input of 2D–3D step is only 2D keypoint locations. In addition to the whole framework, two networks are trained separately: (1) only MVF, uses MSE loss and reprojected loss to supervise, (2) only HT, supervised by MSE loss, bone length loss and joint angle loss. The PCA matrices were trained using two strategies: (1) only Human3.6M, and (2) augmented training data that combines orientation-normalized Human3.6M and MPII-INF-3DHP. With data augmentation, the HT module performs better for all types of actions, demonstrating the importance of pose diversity. Both proposed modules achieves average MPJPE of 21.6 mm, surpassing AT by 1 mm (relative 4.4%). The improvement demonstrates that view consistency and anatomy coherence are both meaningful for pose estimation. As Table 2 shows, the method combined with two modules achieves the state-of-the-art results, with 21.1 mm MPJPE, 6.6% better than AT. The performance is improved on almost all actions.

We also compare our approach with other methods whose input of 2D–3D reconstruction is heatmap or intermediate feature which contains more infromation than keypoint location. As shown in Table 3, our method achieves a balance of implementation complexity (calculated by thop[1]), consuming time and accuracy performance. MVF-HT surpasses AT in both precision and plausibility with 6.6% and 2.4%. Even though Volumetric Triangulation (VT) (Iskakov et al., 2019) surpasses us with 0.3 mm MPJPE and 0.17% PPP@0.2, the MVF-HT almost outperforms it by 145 billion in the number of operations and 48 ms in time costing. In 3D reconstruction procedure, VT utilizes 2D features as input and 3D CNN as inference network, which considers more information and is more complicate.

Qualitative Results on Human3.6M. To evaluate the multi-view consistency performance of MVF, the estimated, reprojected and groundtruth 2D keypoints are compared. The estimated 2D keypoint will be close to the reprojected keypoint from 3D result if the keypoint is consistent with other views. As illustrated in Fig. 5, the blue (reprojected) and green (estimated) points are generally closer after MVF refinement, especially for some self-occlusion. The improvement suggests that the MVF module can make views perceive others and provide assistant for some unseen view from other seen views.

[1] https://github.com/Lyken17/pytorch-OpCounter.

Table 1
Comparison with state-of-the-art methods on Human3.6M in terms of 2D pose estimation accuracy metric JDR (%). "dot" means using inner dot matching strategy and "fcl" represents fully connected layer.

Method	shlder	elb	wri	hip	knee	ankle	root	belly	neck	nose	head	Avg.
CrossView (Qiu et al., 2019)	95.6	95.0	93.7	96.6	95.5	92.8	96.7	96.4	96.5	96.4	96.2	95.9
Epipolar (He et al., 2020)	97.7	97.3	94.9	**99.8**	98.3	97.6	**99.9**	**99.9**	**99.8**	99.7	**99.5**	98.3
Ours-dot	96.4	96.8	**99.8**	97.2	98.3	99.5	97.3	99.7	**99.8**	99.6	93.7	98.0
Ours-fcl	**97.8**	**97.5**	**99.8**	97.7	**98.7**	**99.6**	97.8	99.7	**99.8**	**99.8**	95.4	**98.5**

Table 2
Comparison with state-of-the-art methods on Human3.6M in terms of MPJPE, where the input of 2D–3D step is 2D locations. T. is short for triangulation, and DA is short for data augmentation.

MPJPE (mm)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Canonical (Remelli et al., 2020)	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	31.7	31.2	29.9	26.9	33.7	30.4	30.2
CrossView-T. (Qiu et al., 2019)	25.2	27.9	24.3	25.5	26.2	23.7	25.7	29.7	40.5	**28.6**	32.8	26.8	26.0	28.6	25.0	27.9
Epipolar-T. (He et al., 2020)	29.0	30.6	27.4	26.4	31.0	31.8	26.4	28.7	34.2	42.6	32.4	29.3	27.0	29.3	25.9	30.4
Algebraic-T. (Iskakov et al., 2019)	20.4	22.6	20.5	19.7	22.1	20.6	19.5	23.0	25.8	33.0	23.0	21.6	20.7	23.7	21.3	22.6
Ours-MVF	20.1	21.5	20.0	18.7	21.3	20.3	18.4	21.9	24.3	30.6	22.1	20.4	19.6	23.4	20.2	21.6
Ours-HT w/ DA	**19.4**	21.5	20.0	18.7	21.6	20.9	18.2	21.5	24.8	31.7	21.7	20.2	**18.9**	23.2	19.6	21.6
Ours-HT w/o DA	19.9	21.6	20.3	19.2	21.6	21.3	18.6	21.7	25.5	30.5	22.0	20.2	19.6	23.4	20.5	21.8
Ours	19.5	**20.9**	**19.5**	**18.3**	**21.1**	**20.0**	**17.9**	**21.3**	**23.9**	30.1	**21.6**	**19.9**	**18.9**	**22.8**	**19.5**	**21.1**

Table 3
Comparison of MPJPE, inference time, computation complexity and PPP@0.2 on Human3.6M. MACs and param are shorthand of the number of multiply-add operations and parameters.

Method	Input			Complexity		MPJPE (mm)	Time (ms)	PPP@0.2 (%)
	Feature	Heatmap	Keypoint	Param	MACs			
CrossView-RPSM (Qiu et al., 2019)		✓		570M	212B	26.2	1.88×10^3	–
Epipolar-RPSM (He et al., 2020)		✓		78M	205B	26.9	1.88×10^3	68.54
Algebraic-T. (Iskakov et al., 2019)			✓	79M	210B	22.6	**75**	79.36
Volumetric-T. (Iskakov et al., 2019)	✓			80M	359B	**20.8**	152	**81.41**
Ours			✓	79M	214B	21.1	104	81.24

Table 4
Comparison with the state of the art on Total Capture in terms of MPJPE.

MPJPE (mm)	Subject 1, 2, 3			Subject 4, 5			Avg.
	W2	FS3	A3	W2	FS3	A3	
IMUPVH[a] (Trumble et al., 2017)	30	91	49	36	112	**10**	70
AutoEnc[a] (Trumble et al., 2018)	**13**	49	24	**22**	71	40	35
CrossView[a] (Qiu et al., 2019)	19	28	21	32	54	33	29
GeoFuse[a] (Zhang et al., 2020)	14	26	18	24	49	28	25
Baseline	64	60	53	72	78	62	63
Ours	45	49	45	52	64	57	51
Ours[a]	**13**	**24**	**17**	23	**41**	29	**23**

[a] Methods are trained on the Total Capture.

Furthermore, qualitative experiments are used to evaluate the ability to amend the implausible pose of HT approach. As illustrated in Fig. 6, the extracted anatomy prior can amend some unreasonable errors. It is particularly noteworthy that HT has the capability to correct the pose to have normal joint angles and bone lengths in such a tough situation.

Generalization to the Total Capture Dataset. To substantiate generalization of our model, we first conduct cross-dataset experiments on Total Capture, the testing model is only trained by Human3.6M training set. As shown in Table 4, our method surpasses baseline by 12 mm (19%), which demonstrates the generalization of our method. Then we train our model with Total Capture training data under the same strategy clarified in Section 4.3. Our method achieves 23 mm MPJPE, which also exceeds GeoFuse (Zhang et al., 2020) by 8% and (Trumble et al., 2018) by 34%. It is worth noting that, the PCA training set does not contain the Total Capture, which demonstrates that our anatomy coherence has the ability to deal with unseen gestures.

4.4. Ablation study

All ablation studies are conducted on the Human3.6M. Both MPJPE and PPP@0.2 metrics are used for evaluation.

HT Module Design. To determine the hyperparameters, HT is only used as a post-processing step, replacing the baseline triangulation. We compare different choices of low-dimension D preserved by PCA. As the dimension decreases, less variance is kept, which means lower precision but higher abstraction. As shown in Table 5, the D of V_0 feature is changed from 35 to 10 with stride 5, which corresponds to preserving variance from 99.9% to 88.3%. As D decreases, both precision MPJPE and PPP get improved, and peaks at $D = 25$ (99.5% variances) where a balance is struck in prior extraction and precision preservation. Compared with baseline, the improvement demonstrates the importance of PCA reconstruction term, implying that the pose global context is extracted.

MVF Module Evaluation. Beside the matching strategy, we also evaluate the number of views when fusing, there are two pipelines: (1) all

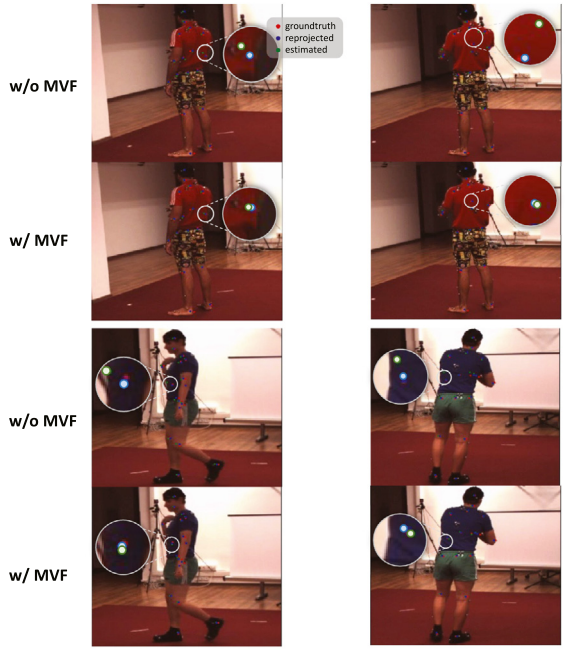

Fig. 5. Visualization of the 2D keypoints with and without MVF refinement. We use different colors to distinguish different types of 2D results, where red: groundtruth, green: estimates, blue: reprojected results from 3D reconstruction. As reprojected results get closer to estimates, the different view keypoints are more consistent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

view fusion, each view generates pseudo heatmap assist to other views, (2) most-conf fusion, only most-confident view is chosen to generate pseudo heatmap. The comparison is shown in Table 6, fully connected layer slightly outperforms the inner dot matching. Moreover, all view fusion performs better. We conjecture that since most initial results are reliable, as the number of fused views increases, more accurate auxiliary information is provided.

Effect of Loss. As shown in Table 7, PJ loss brings 2.1% relative MPJPE improvement, which is more than bone length loss and joint angle loss. We suppose it is because that view consistency enhance correspondence of multi-view 2D keypoints. But the plausibility raises little with PJ loss. BL loss and JA loss bring more improvement in PPP@0.2. It demonstrates that the kinematic skeleton structure can boost the plausibility of pose. Finally we retrain the network which combines all loss function, and obtain the results whose MPJPE is 21.89 mm and PPP@0.2 is 79.7%.

Effect of the Number of Views. Views are reduced from 4 to 2 during testing to explore influence of the number of views. As the number of views decreases, precision degrades. But MPJPE equals to 29.99 mm when there are two views as shown in Table 8, which is still excellent. By comparing ours with ours-HT or ours-MVF, the conclusion that both view consistency and anatomy coherence can improve pose estimation is proved.

Effect of Orientation Normalization. Whether and what the anatomy coherence PCA factors out still bother us. To observe it, we change each latent variable individually with a small step to generate the recovered 3D pose. The changes in 3D poses represents the physical meaning of the corresponding latent variable. Without orientation normalization, there are only 8 out of 25 latent variables describe joint correlations in motion, and the remaining 17 describe rotation invariant property. The results have been improved after orientation normalization, where 25 variables all describe the joint-coupled motion. And the transformation of the recovered 3D poses demonstrate the capability of PCA to restrict joints correlation through motion.

Table 5
Low dimension preserved design comparison. We refer to baseline as $D = 51$ whose dimension is not reduced.

Dimension D	51	35	30	**25**	20	15	10
MPJPE- (mm)	22.60	22.10	22.04	**22.04**	22.06	22.08	22.11
PPP@0.2 (%)	79.36	79.90	79.91	**80.97**	80.14	80.19	80.33

Table 6
Effect of matching strategy in MVF. First row is baseline.

Matching strategy		View		MPJPE (mm)
dot	fcl	most-conf	all	
				22.60
✓		✓		22.05
✓			✓	21.88
	✓	✓		21.92
	✓		✓	**21.61**

Table 7
Effect of loss strategy. PJ: reprojected loss, BL: bone length loss, JA: joint angle loss.

Strategy			MPJPE (mm)	PPP@0.2 (%)
PJ	BL	JA		
			22.60	79.36
✓			22.12	79.37
	✓		22.23	79.55
		✓	22.40	79.43
	✓	✓	22.10	79.68
✓	✓	✓	21.89	79.70

Table 8
Effect of the number of views during testing in terms of MPJPE. When combined with MVF, the case of using three views is not tested because it takes a long time to train.

#(views)	MPJPE (mm)			
	Baseline	Ours-HT	Ours-MVF	Ours
4	22.60	21.58	21.61	21.12
3	27.08	26.11	25.54	24.83
2	33.43	31.83	30.74	29.99

4.5. Limitations

Our method has some limitations that need improvement. Firstly, the low-dimensional subspace spanned by PCA aims to represent the anatomy pose prior across the entire motion space. However, despite combining Human 3.6M (Ionescu et al., 2013) and MPII-INF-3DHP (Mehta et al., 2017) datasets to enhance pose diversity, the prior obtained through PCA struggles to handle freestyle actions with unseen poses (FS, Table 4). Expanding the datasets alone may not be enough to overcome the generalization limitations for unseen motions. Secondly, our network structure is not adaptable to different view numbers, requiring retraining when the view number changes. In addition, the reduction in the number of view can lead to a lack of diverse observations, making the model more susceptible to errors caused by outliers or incorrect initial predictions. The interdependence between views amplifies the impact of inaccuracies in one view on the overall performance.

5. Conclusion

We propose view consistency aware holistic triangulation to improve the performance of both precision and plausibility in 3D HPE. The key contribution is that the geometric correspondences of multi-view 2D keypoints are enhanced and anatomy coherence is injected to 2D–3D process. Meanwhile, a PPP metric is raised to evaluate the pose plausibility. Experiments not only exhibit that our approach

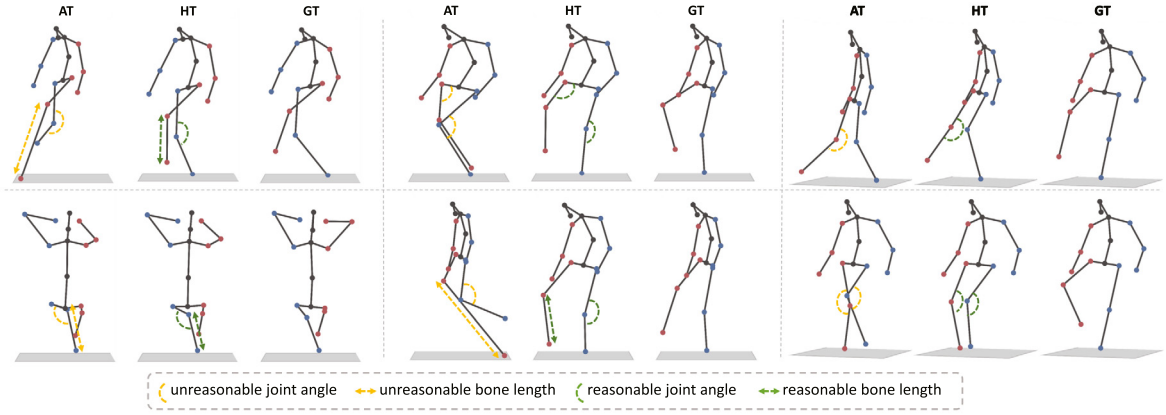

Fig. 6. Visualization of estimated 3D poses. Different colors are used to distinguish whether the pose is reasonable or not, where yellow represents unreasonable and green is reasonable. HT can amend the unreasonable poses. For example, in the first column of the first row, the pose generated by AT (baseline method, w/o reconstruction term) has a too long right leg (red side) and an unreasonable left knee angle (blue side), which is corrected by HT. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

outperforms state-of-the-art methods, but also demonstrate that the reconstruction term with extracted skeletal structure features can abstract the human anatomy prior.

CRediT authorship contribution statement

Xiaoyue Wan: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Zhuo Chen:** Validation, Writing – review & editing. **Xu Zhao:** Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work has been funded in part by the NSFC grant 62176156 and the Fundamental Research Funds for the Central Universities.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2023.103830.

References

Akhter, I., Black, M.J., 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1446–1455.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J., 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, pp. 561–578.

Burenius, M., Sullivan, J., Carlsson, S., 2013. 3D pictorial structures for multiple view articulated pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3618–3625.

Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., Thalmann, N.M., 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: ICCV. pp. 2272–2281.

Chen, X., Lin, K.-Y., Liu, W., Qian, C., Lin, L., 2019a. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10895–10904.

Chen, C.-H., Tyagi, A., Agrawal, A., Drover, D., Mv, R., Stojanov, S., Rehg, J.M., 2019b. Unsupervised 3d pose estimation with geometric self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5714–5724.

Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X., 2019. Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7792–7801.

Gavrila, D.M., Davis, L.S., 1996. 3-d model-based tracking of humans in action: a multi-view approach. In: Proceedings Cvpr Ieee Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 73–80.

Gonzales, R.C., Wintz, P., 1987. Digital Image Processing. Addison-Wesley Longman Publishing Co., Inc..

Guo, Y., Ma, L., Li, Z., Wang, X., Wang, F., 2021. Monocular 3D multi-person pose estimation via predicting factorized correction factors. Comput. Vis. Image Underst. 213, 103278.

Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C., 2019. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10905–10914.

Hartley, R., Zisserman, A., 2003. Multiple View Geometry in Computer Vision. Cambridge University Press.

He, Y., Yan, R., Fragkiadaki, K., Yu, S.-I., 2020. Epipolar transformers. In: Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition. pp. 7779–7788.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components.. J. Educ. Psychol. 24 (6), 417.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36 (7), 1325–1339.

Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y., 2019. Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7718–7727.

Kadkhodamohammadi, A., Padoy, N., 2021. A generalizable approach for multi-view 3d human pose regression. Mach. Vis. Appl. 32 (1), 6.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 66–75.

Kocabas, M., Karagoz, S., Akbas, E., 2019. Self-supervised learning of 3d human pose using multi-view geometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1077–1086.

Liu, K., Zou, Z., Tang, W., 2020. Learning global pose features in graph convolutional networks for 3d human pose estimation. In: Proceedings of the Asian Conference on Computer Vision.

Ma, H., Chen, L., Kong, D., Wang, Z., Liu, X., Tang, H., Yan, X., Xie, Y., Lin, S.-Y., Xie, X., 2021. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. arXiv preprint arXiv:2110.09554.

Malleson, C., Collomosse, J., Hilton, A., 2020. Real-time multi-person motion capture from multi-view video and IMUs. Int. J. Comput. Vis. 128, 1594–1611.

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision (3DV). IEEE, pp. 506–516.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J., 2019. Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10975–10985.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017a. Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7025–7034.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017b. Harvesting multiple views for marker-less 3d human pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6988–6997.

Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W., 2019. Cross view fusion for 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4342–4351.

Remelli, E., Han, S., Honari, S., Fua, P., Wang, R., 2020. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6040–6049.

Reynolds, D.A., et al., 2009. Gaussian mixture models. Encycl. Biom. 741 (659–663).

Romero, J., Tzionas, D., Black, M.J., 2017. Embodied hands: Modeling and capturing hands and bodies together. ACM Trans. Graph. 36 (6), http://dx.doi.org/10.1145/3130800.3130883.

Tian, L., Wang, P., Liang, G., Shen, C., 2021. An adversarial human pose estimation network injected with graph structure. Pattern Recognit. 115, 107863.

Trumble, M., Gilbert, A., Hilton, A., Collomosse, J., 2018. Deep autoencoder for combined human pose estimation and body model upscaling. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 784–800.

Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J., 2017. Total capture: 3d human pose estimation fusing video and inertial sensors. In: Proceedings of 28th British Machine Vision Conference. pp. 1–13.

Tu, H., Wang, C., Zeng, W., 2020. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer, pp. 197–212.

Wandt, B., Ackermann, H., Rosenhahn, B., 2018. A kinematic chain space for monocular motion capture. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

Wandt, B., Rosenhahn, B., 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7782–7791.

Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., Shao, L., 2021. Deep 3D human pose estimation: A review. Comput. Vis. Image Underst. 210, 103225.

Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 466–481.

Xu, T., Takano, W., 2021. Graph stacked hourglass networks for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16105–16114.

Xu, H., Zhang, J., 2020. Aanet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1959–1968.

Yang, J., Ma, Y., Zuo, X., Wang, S., Gong, M., Cheng, L., 2022. 3D pose estimation and future motion prediction from 2D images. Pattern Recognit. 124, 108439.

Zhang, Z., Wang, C., Qin, W., Zeng, W., 2020. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2200–2209.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3425–3435.

Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K., 2016. Sparseness meets deepness: 3d human pose estimation from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4966–4975.