Generalizable Human Pose Triangulation

Kristijan Bartol   David Bojanić   Tomislav Petković   Tomislav Pribanić

University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia

name.surname@fer.hr

Abstract

We address the problem of generalizability for multi-view 3D human pose estimation. The standard approach is to first detect 2D keypoints in images and then apply triangulation from multiple views. Even though the existing methods achieve remarkably accurate 3D pose estimation on public benchmarks, most of them are limited to a single spatial camera arrangement and their number. Several methods address this limitation but demonstrate significantly degraded performance on novel views. We propose a stochastic framework for human pose triangulation and demonstrate a superior generalization across different camera arrangements on two public datasets. In addition, we apply the same approach to the fundamental matrix estimation problem, showing that the proposed method can successfully apply to other computer vision problems. The stochastic framework achieves more than 8.8% improvement on the 3D pose estimation task, compared to the state-of-the-art, and more than 30% improvement for fundamental matrix estimation, compared to a standard algorithm.

1. Introduction

Human pose estimation is a vision task of detecting the keypoints that represent a standard set of human joints. The area is extremely competitive, especially due to the advances in deep learning. Pose estimation is particularly important for applications such as medicine, fashion industry, anthropometry, and entertainment [1]. In this work, we focus on 3D human pose estimation from multiple views in a single time frame.

The common approach to multi-view pose estimation is to (1) detect correspondent 2D keypoints in each view using pretrained pose detector [37, 8, 35], and then (2) triangulate [15, 25, 13, 26, 18, 32]. A naive approach takes 2D detections as they are and applies triangulation from all available views. Due to the variety of poses and self-occlusions, some views contain erroneous detections, which should be ignored or their influence mitigated in the triangulation process. One way to ignore the erroneous detections is to apply
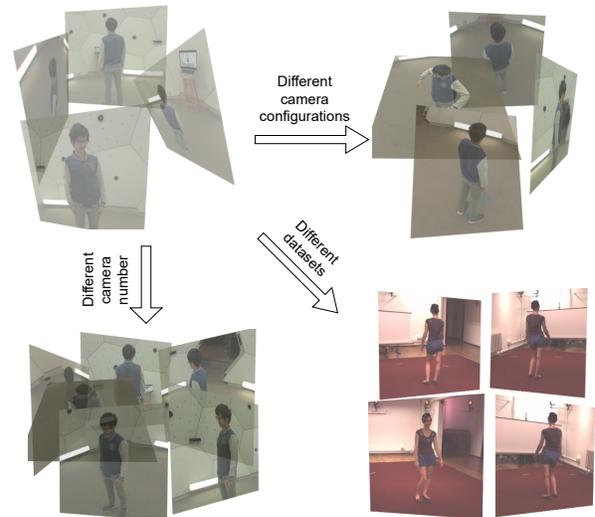

Figure 1: We propose a stochastic framework for human pose triangulation from multiple views and demonstrate its successful generalization across different camera arrangements, their number, and different public datasets. The upper two and the lower left image shows different camera arrangements and their number on CMU Panoptic Studio dataset [17]. The lower right part shows the Human3.6M's 4-camera arrangement [14].

RANSAC [10], marking the keypoints whose reprojection errors are above a threshold as outliers [30, 13]. The problem with vanilla RANSAC is that it is non-differentiable, so the gradients are not back-propagated, which disables end-to-end learning. Most of the state-of-the-art 3D pose estimation approaches extract 2D image features, such as heatmaps, from multiple views, and combine them for 3D elevation in an end-to-end fashion [15, 26, 25]; we refer to those approaches as the *learnable triangulation approaches*.

Due to a mostly-fixed set of cameras during training, the learnable triangulation approaches are often limited to a single camera arrangement and their number. Several works attempt to generalize outside the training data

[26, 15, 18, 13, 31, 29, 34], but the demonstrated performance on novel views is significantly lower than using the original (base) views.

Inspired by stochastic learning [27] and its applications in computer vision [4, 5, 6], we propose *generalizable triangulation* of human pose. First, we generate a pool of random hypotheses. A hypothesis is a 3D pose where the points are obtained by triangulating a random subset of views for each joint separately. Each generated hypothesis pass through a scoring neural network. The loss function is an expectation of the triangulation error, i.e. $\mathbb{E}(h_i) = \sum_i e_i s_i$, where e_i is the error of the hypothesis h_i and s_i is the hypothesis score. By minimizing the error expectation, the model learns the distribution of hypotheses. The key idea is to learn to evaluate 3D pose hypotheses without considering the spatial camera arrangement used for triangulation.

The proposed approach has several practical advantages over the previous methods. First, we demonstrate its consistent generalization performance across different camera arrangement on two public datasets - Human3.6M [14] and Panoptic Studio [17] (see Fig. 1). Second, we show that the proposed model learns human pose prior and define a novel metric for pose prior evaluation. Finally, we apply the same stochastic approach to the problem of fundamental matrix estimation from noisy 2D detections and compare it to the standard 8-point algorithm, showing that the proposed framework successfully applies to computer vision problems other than human pose triangulation.

2. Related Work

We distinguish two types of related work. First, we focus on triangulation-based 3D pose estimation methods and methods that attempt to generalize between the different camera arrangements and datasets. Second, we relate to keypoint correspondence methods and point out how our problem differs from the standard correspondence problem.

Triangulation. Most of the single-person image-based approaches either use robust triangulation (RANSAC) or apply learnable triangulation. Several methods [19, 30, 24] based on robust triangulation use RANSAC on many (more than four) views to apply triangulation only on inlier detection candidates to produce pseudo ground truth data. He et al. [13] exploit epipolar constraints to find the keypoint matches between multiple images and then apply robust triangulation.

The standard approach for learnable triangulation using deep learning models [15, 25, 32, 7, 3] is to first extract 2D pose heatmaps, where each heatmap represents the probability of a keypoint location. Cross-view fusion [25] builds upon the pictorial structures model [2] to combine 2D keypoint features from multiple views to estimate a 3D pose. An algebraic triangulation [15] estimates the confidence

for each keypoint detection and applies weighted triangulation. Their volumetric approach combines the multi-view features and builds the volumetric grid, obtaining the current state-of-the-art for single-frame 3D pose. Finally, [26] fuses the features into a unified latent representation that is less memory intensive than the volumetric grids. Similar to us, they also attempt to disentangle from the specific spatial camera arrangement.

Keypoint correspondence. The standard keypoint-based computer vision approaches, such as structure-from-motion [28], rely on sparse keypoint detections to establish initial 3D geometry. The core problem is to determine the correspondences between the extracted keypoint detections across images, under various illumination changes, texture-less surfaces, and repetitive structures [11]. The usual approach is to apply keypoint descriptor such as SIFT [21] and find inlier correspondences using RANSAC [10]. Even though this paradigm is successful in practice, it is not differentiable and, therefore, cannot be used in an end-to-end learning fashion.

Several works have proposed soft and differentiable versions of RANSAC (DSAC) [4, 5, 6, 39]. The successful soft RANSAC alternative [39] learns to extract both local features of each data point, as well as retain the global information of the 3D scene. Similar to us, they also demonstrate convincing generalization capabilities to unseen 3D scenes. On the other hand, DSAC and its variants [4, 5, 6] propose a probabilistic learning scheme, i.e. minimizing the error expectation. We follow their approach but also discover that different strategies work better for our problem (see Sec. 3 and 4).

In contrast to the standard keypoint matching approaches, we extract keypoints with already known human joint correspondences between the views. However, our correspondent keypoints are noisy, oscillating around the centers of the joints, which potentially leads to erroneous triangulation. Our model demonstrates robustness to erroneous keypoint detections.

3. Method

We first describe the generic stochastic framework, and then describe it more specifically for generalizable pose triangulation and fundamental matrix estimation. The framework consists of several steps, shown in Fig. 2:

1. **Pre-training.** Prior to stochastic learning, the 2D poses (keypoints) are extracted for all images in the dataset. In all our experiments, we use the keypoints extracted using a baseline model [37] pretrained on Human3.6M dataset. The input to stochastic model, therefore, consists only of keypoint detections, $\mathbf{y}$. In each frame, JxK keypoints are detected, where J is the number of joints, and K is the number of views.
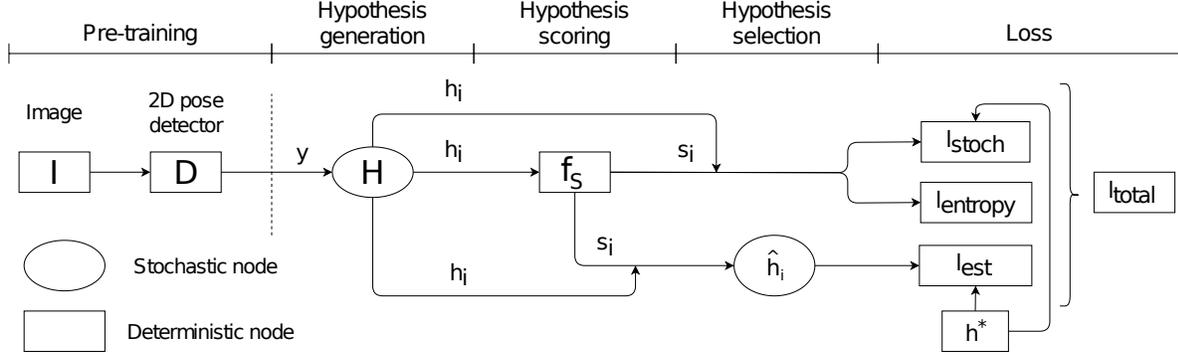
Figure 2: An overview of our method. Before stochastic learning, 2D keypoints, y, are extracted. In each frame, the hypothesis pool, $h_i \in \mathbf{H}$, is generated, and the poses are passed through the scoring network, f_S. The hypothesis $\hat{h}_i$ is selected based on the estimated scores s_i. Finally, the total loss, l_{total}, consists of three components (l_{stoch}, $l_{entropy}$, l_{est}), and is calculated with respect to the ground truth, h^*.

2. **Hypothesis generation, H**. As it is possible to generate an extremely large number of hypotheses, only a subset of random hypotheses is created. Following [27] and [4], we model the hypothesis generation step as a *stochastic* node.

3. **Hypothesis scoring, f_S**. Each generated hypothesis $h_i \in \mathbf{H}$ is scored using a scoring function, $f_S(h_i|\mathbf{y}) = s_i$. The scoring function is a neural network, i.e. a multi-layer perceptron. The network architectures for 3D pose triangulation and fundamental matrix estimation differ and are specified at the end of the Sec. 4. The network is the only learnable part of our model. The estimated scores s_i, passed through the Gumbel-Softmax, $\sigma_{GS}(s_i)$ (Eq. 3), represent the estimated probability distribution of the hypotheses $\mathbf{H}$, $\theta_{\mathbf{H}}$.

4. **Hypothesis selection, $\hat{h}_i$**. We experiment with several hypothesis selection strategies. The one that works the best for us is the weighted average of all hypotheses:

$$\hat{h}_{weight} = \sum_i s_i h_i, \quad \sum_i s_i = 1, \quad h_i \in \mathbf{H}, \quad (1)$$

where the scores s_i are used as weights. We also try other strategies, such as the stochastic selection:

$$\hat{h}_{stoch} = h_i, \quad \text{with} \quad i \sim \theta_{\mathbf{H}}, \quad (2)$$

where hypothesis h_i is selected based on the estimated distribution $\theta_{\mathbf{H}}$. As shown in Sec. 4, the stochastic selection performs worse than the weighted, in contrast to [4].

5. **Loss calculation, l_{total}**. The loss function consists of several components:

(a) Stochastic loss. Following [4], we calculate our stochastic loss as an expectation of error for all hypotheses, $l_{stoch} = \mathbb{E}(e_{\mathbf{H}}) = \sum_i e(h_i, h^*)s_i$, where e_i is the error of the estimated hypothesis with respect to the ground truth, h^*, and s_i represent the probability that the error is minimal.

(b) Entropy loss. Score estimations s_i tend to quickly converge to zero. To stabilize the estimation values, we follow [5] and minimize an entropy function, $l_{entropy} = - \sum_i s_i \log(s_i)$.

(c) Estimation loss. We define it as the error of the selected hypothesis with respect to the ground 3D pose, $l_{est} = e_i(\hat{h}_i, h^*)$. The estimation loss, in the case of generalizable pose triangulation, is most similar to the standard 3D pose estimation loss, used by the competing approaches [15, 26, 25, 32, 18].

Finally, the total loss is a sum of the three components, $l_{total} = \alpha\, l_{stoch} + \beta\, l_{entropy} + \gamma\, l_{est}$, where α, β, and γ are fixed hyperparameters that regulate relative values between the components.

In order for the estimated scores s_i to represent the probabilities, their values need to be normalized into $[0, 1]$ range. The standard way to normalize the output values is to apply the softmax function, $\sigma(s_i) = \frac{\exp s_i}{\sum_j \exp s_j}$. To avoid early convergence, we use the Gumbel-Softmax function [16, 22]:

$$\sigma_{GS}(s_i) = \frac{\exp((\log s_i + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log s_j + g_j)/\tau)}, \quad (3)$$

where τ is a temperature parameter, and g_i represent samples drawn from *Gumbel*(0, 1) [23] distribution. The temperature τ regulates the broadness of the distribution.

For lower temperatures ($\tau < 1$), the influence of lower-score hypotheses is limited compared to higher-score hypotheses, and vice versa. The purpose of *Gumbel*(0, 1) is to add noise to each sample while retaining the original distribution(s), which allows the model to be more flexible with the hypothesis selection.

3.1. Generalizable Pose Triangulation

We now describe the stochastic framework specifically for learning human pose triangulation.

Pose generation. The 3D human pose hypothesis, $h_i \in \mathbf{H}$, is generated in the following way. For each joint k, a subset of views, $\mathbf{v}_k$, is randomly selected. The detections from the selected views are triangulated to produce a 3D joint.

Pose normalization. The input to the pose scoring network, $f_{S,pose}$ are 3D pose coordinates, $\mathbf{p}$, normalized in the following way — we select three points: left and right shoulder and the pelvis (between the hips), calculate the rotation between the normal of the plane given by the three points, and the normal of the xy-plane, and apply that rotation to all coordinates. Other than the 3D pose coordinates, we also extract 16 body part lengths, given by all adjacent joints, e.g. left lower arm, left upper arm, left shoulder, etc. Finally, we concatenate both normalized 3D coordinates and the body part lengths into a 1D vector and pass it through the network. The output is a scalar, s_i, representing the score of the hypothesis h_i.

Pose estimation error. The pose estimation error, $e_i(\hat{h}_i, h^*)$, is a mean per-joint precision error (MPJPE) [14] between the estimated 3D pose, $\hat{\mathbf{p}}_i$, and the ground truth, $\mathbf{p}^*$:

$$e_i(\hat{h}_i, h^*) = e_i(\hat{\mathbf{p}}_i, \mathbf{p}^*) = \frac{1}{J} \sum_k^J ||\hat{p}_{ik} - p_k^*||_2, \quad (4)$$

where p_{ik} is the k-th keypoint of the i-th pose.

3.2. Fundamental Matrix Estimation

We describe how to learn fundamental matrix estimation between the pairs of cameras using the proposed stochastic framework. The fundamental matrix describes the relationship between the two views via $x_2^\top F x_1 = 0$, where x_1 and x_2 are the corresponding 2D points in the first (target) and the second (reference) view. From the fundamental matrix, relative rotation and translation (the relative camera pose) between the views can be obtained [12].

Hypothesis generation. The relative camera pose hypothesis, h_i, is generated in a slightly different way than the 3D pose hypothesis. The required number of points to determine the fundamental matrix is 8 when an 8-point algorithm is used [20]. However, with the presence of noise, the required number of points is usually much higher. Instead of

using a single time frame as in pose triangulation, we select the keypoints from M frames, having a total of $M * J$ individual point correspondences. The camera hypothesis h_i is obtained using a subset of $T < M * J$ correspondences, passed through an 8-point algorithm. The result of an 8-point algorithm are four possible rotations and translations; we select the correct one in a standard way [12].

Input preparation. The input to the camera pose scoring network, $f_{S,cam}$, are the distances between the corresponding projected rays. The rays are obtained using the reference camera parameters, (R_{ref}, t_{ref}), and the estimated relative camera pose, $(R_{rel,i}, t_{rel,i})$. To achieve the permutation invariance between the line distances on the input, we simply sort the values before passing it through the network.

Hypothesis selection. The camera pose hypothesis, $\hat{h}_{weight}$, is selected as the weighted average of the rotation[1], i.e. a weighted average of the translation of all hypotheses.

Estimation error. The hypothesis estimation error, e_i, is calculated as:

$$e_i(\hat{h}_i, h^*) = e_i(\hat{\mathbf{X}}_i, \mathbf{X}^*) = ||\hat{\mathbf{X}}_i - \mathbf{X}^*||_2 \quad (5)$$

where $\mathbf{X}^*$ are random 3D points (used as ground truth), and $\hat{\mathbf{X}}_i$ are 3D points obtained by projecting the points $\mathbf{X}^*$ to 2D planes, using the estimated parameters, $(\hat{R}_i, \hat{t}_i)$, and then projected back to 3D. More specifically, using the estimated, target projection matrix, $\hat{P}_i = K_i[\hat{R}_i|\hat{t}_i]$ and the reference projection matrix, $P_{ref} = K_{ref}[R_{ref}|t_{ref}]$, the points $\mathbf{X}^*$ are first projected to 2D, $\hat{\mathbf{x}}_i = \hat{P}_i \mathbf{X}^*$, and then triangulated using P_{ref} and $\hat{P}_i$. The intrinsic matrices K_i are assumed to be known for all cameras.

4. Experiments

The stochastic framework is evaluated on Human3.6M [14] and Panoptic Studio [17] datasets. As most of the previous 3D pose estimation approaches presented their results on Human3.6M, we use it for the quantitative comparison to state-of-the-art. Panoptic Studio contains a relatively large number of cameras (31) with useful data annotations (camera parameters, 3D and 2D poses). We use the Panoptic Studio dataset to evaluate the generalization performance between different camera arrangements and their number. We also evaluate the generalization between the Panoptic Studio and Human3.6M datasets. As experiments are based on a single-person pose estimation, we use Panoptic Studio sequences that contain single person in the scene, following [36].

Other than the evaluation of our best result ($\hat{h}_{weight}$), we also compare between different hypotheses:

- Weighted average hypothesis, $\hat{h}_{weight}$,

---

[1] The rotations are converted to quaternions, for simplicity.

Table 1: The demonstration of the generalization performance (MPJPE in mm) on five data sets, featuring different spatial camera placements, different number of cameras, and different datasets (CMU Panoptic Studio and Human3.6M). Each row shows the performance on five test sets when the specified train set is used. The maximal difference between the scores for particular test sets is shown in the last column. The last row demonstrates inter-dataset generalization performance, while other rows show intra-dataset performance.

Train	CMU1		CMU2		CMU3		CMU4		H36M		Max diff. ↓
	CMU1	25.8	CMU1	25.8	CMU1	25.6	CMU1	25.2	CMU1	25.6	2.3%
	CMU2	25.4	CMU2	26.0	CMU2	25.5	CMU2	25.6	CMU2	25.9	2.4%
Test	CMU3	24.9	CMU3	26.0	CMU3	25.0	CMU3	25.0	CMU3	25.7	4.4%
	CMU4	25.1	CMU4	25.6	CMU4	25.3	CMU4	25.1	CMU4	25.5	**2.0%**
	H36M	**33.5**	H36M	33.4	**H36M**	**31.0**	H36M	32.5	**H36M**	**29.1**	15.1%

Table 2: The evaluation of generalization performance from CMU Panoptic Studio [17] to Human3.6M dataset [14], compared to the volumetric approach of Iskakov et al. [15]. The proposed approach achieves 8.8% better performance on H3.6M compared to [15], when trained on a 4-camera CMU3 dataset (see Table 1).

CMU → H3.6M		
Ours	Iskakov et al. [15]	Improvement
31.0 mm	34.0 mm	**8.8%**

Table 3: The evaluation of generalization performance compared to Remelli et al. [26] (lower is better). We measure the performance drop between the base test set and the novel test set for intra-dataset and inter-dataset configurations. Note that we do not compare on the same datasets, so we only measure the relative drop in percentages. Still, our approach demonstrates a significantly smaller performance drop compared to the competing method in all setups. The † presents the canonical fusion, and the ‡ presents the baseline approach in [26].

Intra-dataset			
Method (train dataset)	Base test	Novel test	Diff. ↓
Remelli et al. [26] (TC1)[†]	27.5 mm	38.2 mm	38.9%
Remelli et al. [26] (TC1)[‡]	39.3 mm	48.2 mm	22.6%
Ours (CMU1)	24.9 mm	25.8 mm	3.6%
Ours (CMU3)	25.0 mm	25.6 mm	2.4%
Ours (CMU4)	25.0 mm	25.6 mm	2.4%
Ours (CMU2)	25.6 mm	26.0 mm	**1.6%**
Inter-dataset			
Method (train dataset)	H36M	CMU1	Diff. ↓
Ours (H36M)	29.1 mm	33.5 mm	**15.1%**

- Average hypothesis, $\hat{h}_{avg}$, obtained as an average of all hypotheses,

- Most and least probable hypotheses, $\hat{h}_{most}$ and $\hat{h}_{least}$, the hypotheses with maximal and minimal estimated score, s_{max} and s_{min},

- Stochastic hypothesis, $\hat{h}_{stoch}$, selected randomly, based on the estimated distribution $\theta_{\mathbf{H}}$,

- Random hypothesis, $\hat{h}_{random}$, selected randomly from

Table 4: The comparison to RANSAC, algebraic triangulation [15], and VoxelPose [34] on Panoptic Studio (intra-dataset) [mm]. The numbers show the performance on novel camera views. Our number is obtained as an average over 12 non-diagonal values of Table 1.

Intra-dataset (CMU Panoptic Studio)			
RANSAC	Algebraic	VoxelPose	Ours
39.5	33.4	25.5	**25.4**

an uniform distribution,

- Best and worst hypotheses[2], h_{best} and h_{worst}, with the lowest and the highest errors, e_{min} and e_{max}.

Additionally, we also compare ourselves with RANSAC as reported in [15] (see Subsec. 4.2).

4.1. Generalization Performance

One of the most important properties of the proposed model is that it generalizes well to different spatial arrangements and number of cameras, and different datasets, which is a major limitation of the previous models. To evaluate the generalization performance across data sets, we select five different camera arrangements:

1. Cameras $1, 2, 3, 4, 6, 7, 10$ (CMU1),

2. Cameras $12, 16, 18, 19, 22, 23, 30$ (CMU2),

3. Cameras $10, 12, 16, 18$ (CMU3),

4. Cameras $6, 7, 10, 12, 16, 18, 19, 22, 23, 30$ (CMU4), and

5. Cameras $0, 1, 2, 3$ (H36M).

The setup is as follows. Each of the five camera arrangements is first used for training, and then the generalization performance is tested on the remaining four arrangements.

---

[2]Note that the best and the worst hypotheses are not available in inference (missing ˆ sign), because they are determined using ground truth.

Table 5: No additional training data setup. Overall comparison to the state-of-the-art on Human3.6M dataset. The proposed method outperforms most of the state-of-the-art methods. All values are showing MPJPE scores (mm).

Protocol 1, abs. positions	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg ↓
Tome et al. [32]	43.3	49.6	42.0	48.8	51.1	64.3	40.3	43.3	66.0	95.2	50.2	52.2	51.1	43.9	45.3	52.8
Kadkhodamohammadi et al. [18]	39.4	46.9	41.0	42.7	53.6	54.8	41.4	50.0	59.9	78.8	49.8	46.2	51.1	40.5	41.0	49.1
Cross-view fusion [25]	28.9	32.5	26.6	28.1	28.3	29.3	28.0	36.8	41.0	30.5	35.6	30.0	28.3	30.0		31.2
Remelli et al. [26]	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
Epipolar transformers [13]	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	28.8	31.7	28.2	26.4	23.6	28.3	23.5	**26.9**
Ours (h_{weight})	27.5	28.4	29.3	27.5	30.1	28.1	27.9	30.8	32.9	32.5	30.8	29.4	28.5	30.5	30.1	29.1

The five selected sets differ with respect to the spatial camera arrangement and their number. Additionally, the fifth camera set (H36M) is used to test the transfer learning capabilities between the datasets. All the results in this subsection are obtained using h_{weight} hypothesis.

Our Generalization Performance. Table 1 shows consistent performance on each of the five test datasets, regardless of the selected training dataset. In particular, the performance between different test sets on the Panoptic Studio dataset is within 5% difference, which demonstrates robustness to various camera arrangements and their number (intra-dataset). The inter-dataset generalization is also successful, which we further evaluate against the competing methods [15, 26]. Note that the demonstrated generalization can be exploited both in training time and in inference.

Volumetric Triangulation. Table 2 compares our proposed method to the state-of-the-art 3D pose estimation approach [15]. Iskakov et al. reported an average 34.0 mm error on Human3.6M test set when they trained on CMU Panoptic Studio (4-camera arrangement). Compared to them, we achieve 31.0 mm on our 4-camera arrangement (CMU3), demonstrating an improvement in inter-dataset generalization (see Table 1 for the comprehensive results).

Remelli et al. Table 3 compares our method to Remelli et al. [26]. Similar to us, they explicitly address the generalization to novel views. They demonstrate their intra-dataset generalization performance on Total Capture [33], by comparing the test performances on cameras (1, 3, 5, 7) as a base arrangement (TC1) and testing it on cameras (2, 4, 6, 8) as a novel arrangement (TC2). We do not evaluate our model on Total Capture. Instead, to compare with Remelli et al., we measure the performance between the CMU camera test sets and evaluate relative score differences. The performance of our model is consistent across different camera arrangements and their number for intra-dataset configuration. Moreover, our inter-dataset performance from CMU Panoptic Studio to Human3.6M is 15.1%, which is still better than the best result by Remelli et al. Note that the inter-dataset experiment is the most difficult as it also includes the changes in camera arrangement.

RANSAC. We outperform RANSAC on Panoptic Studio by a large margin. We can explain this by the fact that CMU does not have a full view of a person in most cameras, leading to strong occlusions and missing parts. As RANSAC takes only reprojection errors of individual 3D joints as an inlier selection criterion, it is unable to evaluate the estimated 3D pose as a whole, in contrast to our model that learns human pose prior (see Sec. 4.3).

Algebraic Triangulation. The algebraic triangulation [15] is originally proposed as an improvement over RANSAC, where the weight is estimated for each joint location. The weight-based model indeed outperforms RANSAC both on Human3.6M and Panoptic Studio. However, as the authors point out in [15], it has several drawbacks. First, it processes each view independently, and second, it separately triangulates each joint. Therefore, the weight-based algebraic model suffers from the same problem as RANSAC by not taking the whole pose into account. Our model, on the other hand, successfully learns human pose prior, which allows it to select more feasible poses, making it more robust to occlusions and missing body parts. Note that [15] does not test their weighted model on unseen views. Therefore, Table 1 shows the result of the model w/o weights, as this model is consistent across different camera sets. The actual result of the weighted model might differ, but it is hard to estimate by how much.

VoxelPose. VoxelPose [34] reports 25.51mm MPJPE score on their intra-dataset experiment, compared to our 25.42mm. Even though we achieve comparable performances, the significant difference is that we did not pretrain our 2D backbone on the Panoptic Studio dataset, which would most likely further improve our 2D keypoint estimation and, consequently, final 3D pose estimations (Supp.).

4.2. Base Dataset Performance

The comparison to state-of-the-art is shown in Table 5. Note that the Table only shows the methods that use Human3.6M for training and testing, with no additional training data (therefore, excluding Iskakov et al. [15]). Compared to the best-performing single-frame method, Epipolar Transformers [13], we obtain 2.2 mm worse MPJPE, but outperform most of the other recent methods.

Table 6 shows the MPJPE scores of all previously described pose hypotheses on the two datasets, compared to the RANSAC result, as reported in [15]. Even though our weighted average hypothesis, $\hat{h}_{weight}$, is outperformed by the RANSAC approach on Human3.6M, we show a significant improvement on Panoptic Studio. Also, note that RANSAC is competitive against most of the state-of-the-art approaches on Human3.6M that do not use additional train-
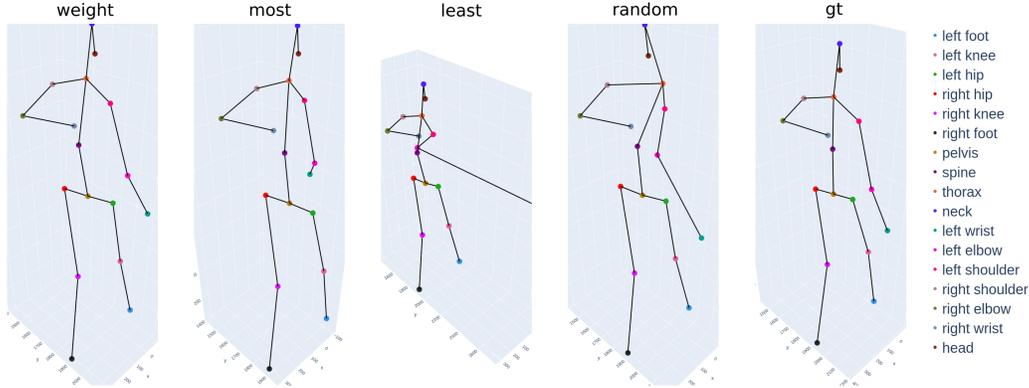
Figure 3: Qualitative comparison between four 3D pose hypotheses compared to ground truth (*gt*), on Human3.6M.

Table 6: Overall quantitative comparison between the hypotheses. The values are showing MPJPE scores in mm (the lower is better).

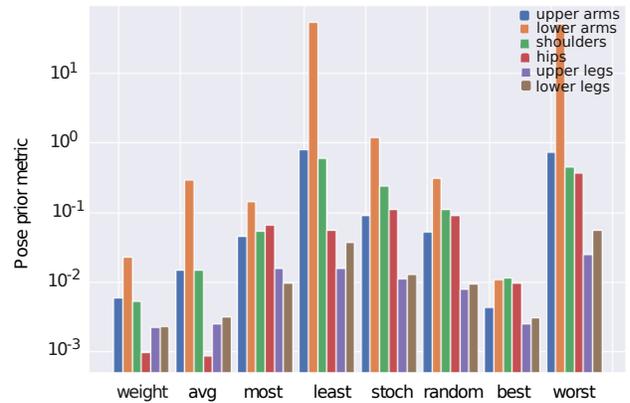| Hypothesis	Human3.6M ↓	Panoptic Studio ↓
$\hat{h}_{weight}$	29.1	**24.9**
$\hat{h}_{avg}$	31.2 +2.1	25.9 +1.0
$\hat{h}_{most}$	41.3 +12.2	25.0 +0.1
$\hat{h}_{least}$	74.5 +45.4	29.8 +3.9
$\hat{h}_{stoch}$	41.3 +12.2	26.5 +1.6
$\hat{h}_{random}$	45.0 +15.9	26.1 +1.2
h_{best}	22.3 -6.8	24.4 -0.5
h_{worst}	98.9 +69.8	31.0 +6.1
RANSAC	**27.4 -1.7**	39.5 +14.6

Figure 4: Evaluation of the human pose prior metric for different hypotheses, and six body part pairs (lower is better).

ing data.

Regarding other results, the average hypothesis, $\hat{h}_{avg}$ performs better than the stochastic, $\hat{h}_{stoch}$. The stochastic performs even worse than the random hypothesis on Panoptic Studio. The most probable hypothesis, $\hat{h}_{most}$, outperforms the average on Panoptic Studio. Note that the difference between best and worst hypothesis (h_{best}, h_{worst}) is significantly different on the two datasets. This suggests that the hypotheses generated on Panoptic Studio are more similar to each other and the distribution is less broad. The difference between the most and the least probable hypotheses ($\hat{h}_{most}$, $\hat{h}_{least}$) is reasonable on both datasets, which confirms that our model learned to differentiate between the poses.

Fig. 3 shows the qualitative performance comparison between several hypotheses. The least probable hypothesis, h_{least}, does not have visually plausible pose reconstruction, while the random hypothesis, $\hat{h}_{random}$, has some obvious errors in the upper body. The most probable hypothesis, $\hat{h}_{most}$ has minor reconstruction errors on the right arm and shoulder. The weighted hypothesis, $\hat{h}_{weight}$, is visually comparable to ground truth.

4.3. Human Pose Prior

We demonstrate the successful pose prior learning of the pose scoring network, $f_{S,pose}$. There are previous works that attempt learning human pose prior [9, 38, 2], but they do not quantitatively evaluate their methods. The idea of learning pose prior is to differentiate between the 3D poses that are more plausible and the poses that are less plausible with respect to several human body properties. The properties that can be extracted from the 3D pose are based on the body part lengths and between-joint angles. In this work, we focus on body part lengths, i.e. *left-right body symmetry*.

The body symmetry is measured for six different body left-right part pairs: upper arms, lower arms, shoulders, hips, upper legs, and lower legs. For each pair, l, we calculate the ratio, r_{il} between the left and right part, in each time frame, i. The final pose prior metric is a variance of the ratios over time:

$$S^2 = \frac{\sum_i (r_{il} - \bar{r}_l)^2}{T - 1},\qquad(6)$$

Table 7: Evaluation of fundamental matrix estimation for all pairs of views on Human3.6M, based on four error metrics. Note that the camera pairs (1, 3), and (2, 4) are diagonal, while other pairs are adjacent.

Camera pair	$E_R \downarrow$	$E_t \downarrow$	$E_{2D} \downarrow$	$E_{3D} \downarrow$
(1, 3)	1.7e-2	2.3e+1	3.7e+0	2.3e+1
(2, 4)	3.2e-2	4.9e+0	1.3e+0	2.3e+1
(1, 4)	9.8e-3	2.7e+1	2.3e+0	1.3e+1
(2, 3)	**2.1e-3**	9.9e+0	**7.2e-1**	1.3e+1
(3, 4)	4.7e-3	8.5e+0	1.2e+0	5.5e+0
(1, 2)	4.8e-3	**4.8e+0**	8.1e-1	**4.9e+0**

where $\overline{r}_l$ is the mean ratio for the pair l, and T is the number of frames. The reason for using ratios instead of the differences between the body parts is that some people are naturally asymmetric, so the idea is only to measure the consistency over time.

Fig. 4 shows the pose prior metrics for the subject 9 of the Human3.6M dataset, for different hypotheses. As expected, the values are generally the lowest for our best performing hypothesis, $\hat{h}_{weight}$, followed by the average hypothesis, $\hat{h}_{avg}$. The difference between the most probable and the least probable hypothesis ($\hat{h}_{most}$, $\hat{h}_{least}$) suggests that we successfully learned body pose prior, i.e. differentiate between the plausible poses with respect to the body symmetry consistency over time. Note that the best hypothesis, h_{best}, is comparable to $\hat{h}_{weight}$.

4.4. Fundamental Matrix Estimation

Table 7 shows the fundamental matrix estimation results on all 4-view combinations on Human3.6M. The four metrics are used for the evaluation:

- Rotation error, $E_R = ||\text{quat}(\hat{R}_{rel}) - \text{quat}(R^*_{rel})||_2$, where quat() represents the conversion to quaternions,

- Translation error (mm), $E_t = ||\hat{t}_{rel} - t^*_{rel}||_2$,

- 2D error (pixels), $E_{2D} = ||\hat{\mathbf{x}}_i - \mathbf{x}^*||_2$, where $\mathbf{x}^*$ represents random 3D points, $\mathbf{X}^*$, projected using the ground truth relative projection matrix, P^*, and

- 3D error (mm), $E_{3D} = e_i(\hat{h}_i, h^*)$, from Eq. 5.

The obtained results show that the model achieves sub-pixel error (E_{2D}) for two pairs of views ((1, 2) and (2, 3)), and only few pixels in the worst case, which corresponds to several millimeters when reprojected back to 3D (E_{3D}). Note that the adjacent pairs of views have lower errors than the opposite pairs, as expected.

In Fig. 5, we compare our 3D errors (E_{3D}) to the vanilla 8-point algorithm, on the (2, 3) camera pair, using different
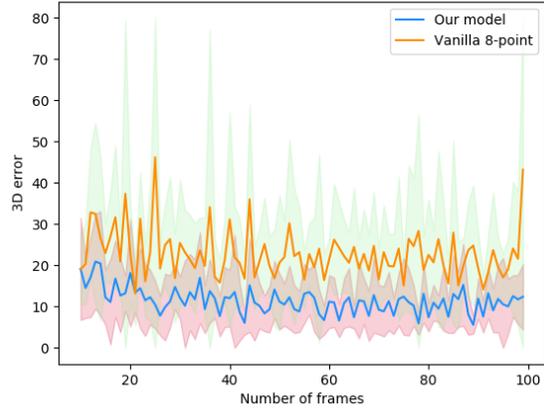

Figure 5: The comparison of the E_{3D} errors between the stochastic model and the 8-point algorithm, for different number of input frames (between 10 and 100), using the camera pair (2, 3) of Human3.6M. For every number of frames, the experiments is done 10 times. The lines show mean values, and the fill parts show standard deviations. The values are clipped to 80 mm.

number of input frames. Our model consistently outperforms the 8-point algorithm, showing robustness to noise and increased confidence due to lower variance.

5. Conclusion

The proposed generalizable approach is a promising novel direction for 3D human pose estimation, as well as other related computer vision problems, such as the camera pose estimation. The demonstrated results show convincing generalization capabilities between different camera arrangements and datasets, outperforming previous methods. The model requires relatively little training data, which makes training faster and more convenient for smaller datasets, as further discussed in Supplementary.

The overall performance is competitive in both human pose triangulation and camera pose estimation tasks. By combining these two steps, it is possible to transfer the performance of the base dataset to any novel multi-camera dataset, in inference. The next reasonable step is to exploit image features in an end-to-end learning fashion, which should further improve the performance and possibly outperform the state-of-the-art even on the base dataset. The current model supports only a single-person pose triangulation. To extend to multi-person, we need to solve the keypoint correspondence problem between the people.

Acknowledgement

References

[1] K. Bartol, D. Bojanić, T. Petković, and T. Pribanić. A review of body measurement using 3d scanning. *IEEE Access*, 9:67281–67301, 2021. 1

[2] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, 2014. 2, 7

[3] A. Bouazizi, J. Wiederer, U. Kressel, and V. Belagiannis. Self-supervised 3d human pose estimation with multiple-view geometry. *ArXiv*, abs/2108.07777, 2021. 2

[4] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. Dsac — differentiable ransac for camera localization. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2500, 2017. 2, 3

[5] E. Brachmann and C. Rother. Learning less is more - 6d camera localization via 3d surface regression. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018. 2, 3

[6] E. Brachmann and C. Rother. Neural-guided ransac: Learning where to sample model hypotheses. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4321–4330, 2019. 2

[7] S. Bultmann and S. Behnke. Real-time multi-view 3d human pose estimation using semantic feedback to smart edge sensors. In D. A. Shell, M. Toussaint, and M. A. Hsieh, editors, *Robotics: Science and Systems XVII, Virtual Event, July 12-16, 2021*, 2021. 2

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017. 1

[9] D. Drover, M. Rohith, C.-H. Chen, A. Agrawal, A. Tyagi, and C. P. Huynh. Can 3d pose be learned from 2d projections alone? In *ECCV Workshops*, 2018. 7

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. 1, 2

[11] Y. Furukawa and C. Hernández. *Multi-View Stereo: A Tutorial*. 2015. 2

[12] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 4

[13] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7776–7785, 2020. 1, 2, 6

[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*, 36(7):1325–1339, jul 2014. 1, 2, 4, 5

[15] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7717–7726, 2019. 1, 2, 3, 5, 6

[16] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *ArXiv*, abs/1611.01144, 2017. 3

[17] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 2, 4, 5

[18] A. Kadkhodamohammadi and N. Padoy. A generalizable approach for multi-view 3d human pose regression. *Mach. Vis. Appl.*, 32:6, 2021. 1, 2, 3, 6

[19] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1077–1086, 2019. 2

[20] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. In M. A. Fischler and O. Firschein, editors, *Readings in Computer Vision*, pages 61–62. Morgan Kaufmann, San Francisco (CA), 1987. 4

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 2

[22] C. J. Maddison, A. Mnih, and Y. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *ArXiv*, abs/1611.00712, 2017. 3

[23] C. J. Maddison, D. Tarlow, and T. Minka. A* sampling. In *NIPS*, 2014. 3

[24] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. *ArXiv*, abs/2008.09309, 2020. 2

[25] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng. Cross view fusion for 3d human pose estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4341–4350, 2019. 1, 2, 3, 6

[26] E. Remelli, S. Han, S. Honari, P. Fua, and R. Y. Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6039–6048, 2020. 1, 2, 3, 5, 6

[27] J. Schulman, N. Heess, T. Weber, and P. Abbeel. Gradient estimation using stochastic computation graphs. In *NIPS*, 2015. 2, 3

[28] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2

[29] H. Shuai, L. Wu, and Q. Liu. Adaptively multi-view and temporal fusing transformer for 3d human pose estimation. *ArXiv*, abs/2110.05092, 2021. 2

[30] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, 2017. 1, 2

[31] J. J. Sun, J. Zhao, L.-C. Chen, F. Schroff, H. Adam, and T. Liu. View-invariant probabilistic embedding for human

pose. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 53–70, Cham, 2020. Springer International Publishing. 2

[32] D. Tomè, M. Toso, L. Agapito, and C. Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. *2018 International Conference on 3D Vision (3DV)*, pages 474–483, 2018. 1, 2, 3, 6

[33] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017. 6

[34] H. Tu, C. Wang, and W. Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6

[35] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 1

[36] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966, 2019. 4

[37] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2

[38] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[39] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua. Learning to find good correspondences. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2666–2674, 2018. 2