Learning 3-D Human Pose Estimation from Catadioptric Videos

Chenchen Liu, **Yongzhi Li**, **Kangqi Ma**, **Duo Zhang**, **Peijun Bao**, **Yadong Mu**[*]
Peking University
{liuchenchen, yongzhili, makq, zhduodyx, peijunbao, myd}@pku.edu.cn

Abstract

3-D human pose estimation is a crucial step for understanding human actions. However, reliably capturing precise 3-D position of human joints is non-trivial and tedious. Current models often suffer from the scarcity of high-quality 3-D annotated training data. In this work, we explore a novel way of obtaining gigantic 3-D human pose data without manual annotations. In catedioptric videos (*e.g.*, people dance before a mirror), the camera records both the original and mirrored human poses, which provides cues for estimating 3-D positions of human joints. Following this idea, we crawl a large-scale Dance-before-Mirror (DBM) video dataset, which is about 24 times larger than existing Human3.6M benchmark. Our technical insight is that, by jointly harnessing the epipolar geometry and human skeleton priors, 3-D joint estimation can boil down to an optimization problem over two sets of 2-D estimations. To our best knowledge, this represents the first work that collects high-quality 3-D human data via catadioptric systems. We have conducted comprehensive experiments on cross-scenario pose estimation and visualization analysis. The results strongly demonstrate the usefulness of our proposed DBM human poses.

1 Introduction

Accurately reconstructing the human pose in 3-D from real images in a variety of indoor and outdoor scenarios, has a wide range of interesting applications in emerging fields such as virtual and augmented reality, human computer interaction, humanoid robotics and monitoring mobility. In the past 10 years, many 3-D pose-related datasets have emerged, which has greatly promoted the development of this field.

Existing 3-D pose datasets can be roughly divided into three categories. The first type collects 3-D pose data using a marker-based motion capture system synchronized with video, such as HumanEva [Sigal *et al.*, 2010], Human3.6M [Ionescu *et al.*, 2014] and GPA [Wang *et al.*, 2019]. This type of data is mostly collected indoors, and restricts recording to skin-tight clothing. To overcome the limitations of marker-based data collection, several marker-less approaches have also been used. 3DPW [von Marcard *et al.*, 2018] and MPI-INF-3DHP [Mehta *et al.*, 2017] propose marker-less capturing system (based on inertial measurement sensors and recording in green screen studio with multiple cameras in a dome system). CMU Panoptic Studio [Joo *et al.*, 2019] creates a panoptic studio and capture poses with 10 pre-calibrated RGBD cameras. This marker-less system enables diverse clothing but requires an expensive studio setup. The third is synthetic data which can be generated by retargeting MoCap sequences to 3-D avatars [Chen *et al.*, 2016]. However the results lack realism. Trained using the synthetic data, learning based methods often suffer from the peculiarities of the rendering, leading to poor generalization to real images.

In acquiring above data, either high-accuracy sensors or multiple-camera system are costly and have limited capturing scenarios. In this paper, we propose to exploit catadioptric videos for collecting 3-D pose. In conventional catadioptric systems (an optical system that combines refraction-based lenses and reflection-based mirrors), one can treat each mirror as a virtual camera. For example, [Gluckman and Nayar, 2001] designed and implemented a real-time catadioptric stereo system which uses only a single camera and two planar mirrors. There are two major challenges for the utilization of catadioptric videos. First, as pointed out by [Gluckman and Nayar, 2001], at least two mirrors are required for camera calibration and 3-D estimation, which is infeasible in socially-shared catadioptric videos (*e.g.*, YouTube videos themed dancing tutorial with a mirror). Secondly, catadioptric videos are often captured in realistic scenarios, which significantly complicates reliable estimation of human keypoints. Unlike the marked-based systems, the initial pose estimation is often noisy and demands further refinement.

To tackle above challenges, we harness the following insight: different from generic objects, human body has many strong structure priors, such as the symmetry of the left and right bone length, the degree of freedom of keypoints, etc. We prove that, for one-mirror system that dominates online catadioptric videos, using the human priors can significantly reduce the ambiguity of camera's interinsic parameter calibration, and serves as a guiding objective in refining noisy human keypoints.

[*]Corresponding author

Our main contributions are three-folds: 1) We propose a large-scale Dance-before-Mirror (DBM) video dataset, which is more than 24 times larger than existing largest Human3.6M [Ionescu *et al.*, 2014]. Overall 6,922 dance-before-mirror videos are crawled from video-sharing websites, and 175,652 20-second clips are extracted. The DBM data can serve as a new benchmark for 3-D human pose estimation; 2) We devise a novel algorithm for intelligently estimating high-quality 3-D human keypoints from online catadioptric videos. Briefly, we boil it down to an optimization problem between two sets of mirrored 2-D human keypoints. The key trait of the proposed algorithm is the joint use of epipolar geometric formulas and human skeleton priors. In particular, the symmetry prior of left / right human body is utilized for inferring camera intrinsics and recovering occluded human keypoints; 3) Comprehensive cross-data transfer experiments are conducted together with other three datasets (Human3.6M, MPI-INF-3DHP, 3DPW). The experimental results clearly demonstrate the usefulness of our data.

2 Related Works

Human Pose Estimation For 2-D human pose estimation [Fang *et al.*, 2017; Chen *et al.*, 2018; Cao *et al.*, 2017], most prevailing methods adopt an encoder-decoder architecture to predict a heatmap for each keypoint, from which the keypoint position is further inferred. Exemplar methods include Simple baseline [Xiao *et al.*, 2018] and CPN [Chen *et al.*, 2018]. More recently, HR-Net [Sun *et al.*, 2019] uses the repeated multi-scale fusions across high-to-low and low-to-high sub-networks to exchange the multi-resolution information. For 3-D human pose estimation, some traditional methods like [Wei and Chai, 2010] propose a generative method using physics priors to get the 3-D pose in a monocular view. Others [Guan *et al.*, 2009; Jain *et al.*, 2010] introduce semi-automatic analysis-by-synthesis fitting of parametric body models. Modern deep network based methods [Pavlakos *et al.*, 2017; Sun *et al.*, 2018] can broadly be classified into two classes: *direct regression* and *'lifting' based approaches*. The former requires lots of 3-D-pose labelled images and predict the 3-D location straightly from the image. But such datasets are either captured in studio scenarios with limited pose and appearance diversity [Ionescu *et al.*, 2014] or contains lots of synthetic imagery [Chen *et al.*, 2016]. While 'Lifting' based approaches predict the 3-D pose from a separately detected 2-D pose [Martinez *et al.*, 2017]. More recently, some literature [Qiu *et al.*, 2019; Remelli *et al.*, 2020] leverage the multi-view informantion, which further improve the performance.

Learning from Catedioptric Systems [Mariottini *et al.*, 2012] proposes a new image-based camera localization and 3-D scene reconstruction algorithm by observing a scene being reflected on two (or more) planar mirrors. [Zhou *et al.*, 2016] proposes an omnidirectional stereo vision sensor based on one single camera and catadioptric system. However, existing works are mostly geometry-oriented, not fully harnessing the learning approach and human skeleton priors, which inspires our work.

3 Dance-before-Mirror (DBM) Video Dataset

This section elaborates on the crawling / processing of large-scale Dance-before-Mirror (DBM) video dataset, particularly a novel algorithm that extracts 3-D human keypoints from videos without human annotations.

3.1 Video Crawling and Clip Generation

Motivating the construction of DBM videos we consider several desiderata, including the comprehensive coverage of human poses, clothes, and background visual appearance. Furthermore, camera motion and shot change should be maximally excluded for easing the estimation of camera intrinsics. We observed in the past few years tremendous dancing-teaching videos have been shared at social media websites. In particular, we choose Bilibili[1], a popular video-sharing website in Asian countries themed around animation, comic, and games, and YouTube[2] as two main resources. On both sites we search videos with queries "dance tutorial mirror" and its several variants. The top-ranked videos in the search results are mostly records of dancing tutorials. Mirrors are generally positioned in front of the dancing tutors during the video recording in order for a non-occlusion viewing of the dancing actions. We manually filter out videos without mirrors from the returned results. This totals 6,922 videos (5,840 from Bilibili and 1,082 from Youtube) with durations from half a minute to longer than an hour, all of which we suppose are amenable to further processing. Figure 1 illustrates a few randomly-drawn frames from these videos. As we can see, DBM videos span a wide spectrum of key factors, *e.g.*, human poses, races, genders, etc.

For data cleaning, a human detector (we use pretrained FPN [Lin *et al.*, 2017] with ResNet-101 [He *et al.*, 2016] on MS-COCO [Lin *et al.*, 2014]) first scans video frames at 5 fps (frames per second). Only frames with two persons (tutor and the correspondence in mirror) are kept. Continuous 20 seconds (100 key frames) are trimmed out as clips. For diversity, at most 100 clips are reserved for each video.

Next, we use HR-Net [Sun *et al.*, 2019] pretrained on MS-COCO and MPII [Andriluka *et al.*, 2014] for detecting keypoints from each person. For each keypoint $p_i = (x_i, y_i, s_i)$, where (x_i, y_i) is the pixel coordinates, and s_i is the confidence score. In DBM videos, self-occlusion is commonly observed. We simply use the confidence scores of keypoints for filtering purpose. When $s_i < \tau$, a keypoint is marked as occluded. When the number of occluded points in the entire clip is greater than N, the video is judged to be severely occluded and can not be used in subsequent algorithm. In practice, τ and N are empirically set to 0.7 and 1500, respectively.

3.2 Geometric Model

The geometry of our proposed model is depicted in Figure 2. Let O be the optical center and π be the single mirror in the catadioptric video. Conventionally, it is a dominating practice to add a virtual camera for each mirror (*e.g.*, the mirrored optical center O' in Figure 2).

[1] https://www.bilibili.com/

[2] https://www.youtube.com/

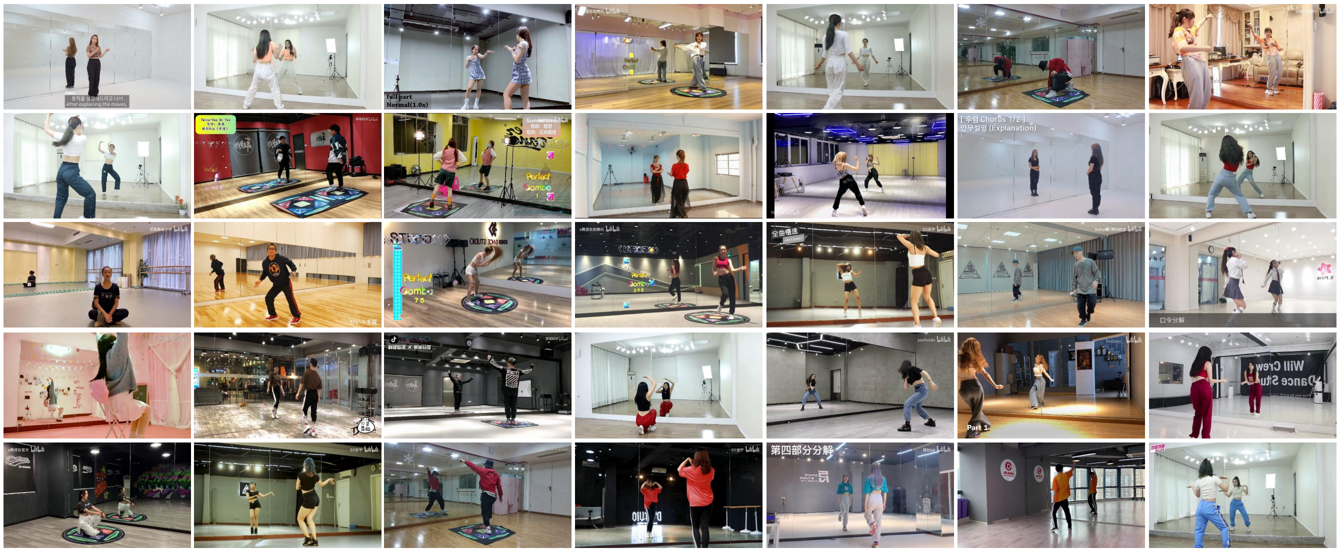

Figure 1: Sampled frames from our Dance-before-Mirror (DBM) videos. Each frame is randomly drawn from one of 20-second video clips. Main themes of these clips are teaching to dance. See main text for more description.

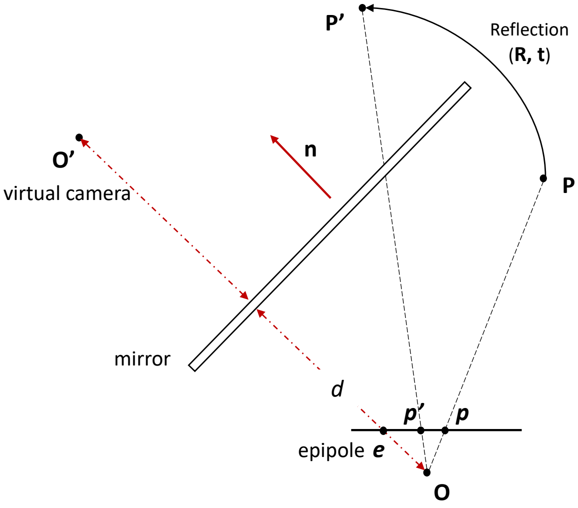

Figure 2: Geometric model of single-mirror catadioptric system. See text for more explanation.

With rare exception a dancing mirror can be modeled as a planar. Suppose the world coordinate system is aligned with the camera's. A planar mirror π can be uniquely determined by a normal vector n ($n \in \mathbb{R}^3$ and $\|n\|_2 = 1$) and a non-negative scalar $d \geq 0$ for representing the orthogonal distance between mirror π and the optical center O. Let $P \in \mathbb{R}^3$ be an arbitrary space point in the scene which is not on the mirror. P' is its reflected point. The relationship between P, P' is specified by:

$$\begin{aligned} P' &= P + 2(d - \mathbf{n}^T P)\mathbf{n} \\ &= (\mathbf{E} - 2\mathbf{n}\mathbf{n}^T)P + 2d\mathbf{n} \\ &= \mathbf{R}P + \mathbf{t}, \end{aligned} \tag{1}$$

with two key elements calculated as:

$$\mathbf{R} = \mathbf{E} - 2\mathbf{n}\mathbf{n}^T, \;\; \mathbf{t} = 2d\mathbf{n}, \tag{2}$$

which can be trivially verified by connecting P, P' in Figure 2 and integrating the connection [Gluckman and Nayar, 2001]. $\mathbf{E}$ is an identity matrix whose size can be inferred from its context. The matrix $\mathbf{S} = (\mathbf{R}\ \mathbf{t}; 0\ 1)$ is called the reflection matrix in mirror systems.

To obtain a Euclidean-sense estimation of human keypoints, it is necessary to determine the camera intrinsics. Generally, it is challenging to perform self-calibration on DBM videos due to the lack of informative clues. We simplify it through the following approximation as the work [Gluckman and Nayar, 2001] did: it is often the case that the skew is zero, the optical center is roughly above the image center. The remaining key variable to be learned is the focal length. In specific, our approximation leads to the following camera intrinsic matrix:

$$\mathbf{K} = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

Use p, p' to denote the pixels that correspond to P, P', respectively. For a pinhole camera model, their relationship can be represented as

$$Z \cdot \begin{pmatrix} p \\ 1 \end{pmatrix} = \mathbf{K}P, \;\; Z' \cdot \begin{pmatrix} p' \\ 1 \end{pmatrix} = \mathbf{K}(\mathbf{R}P + \mathbf{t}), \tag{4}$$

where Z, Z' are the values on the z-axis for P, P' respectively.

3.3 3-D Keypoint Estimation

The geometric model is parameterized by the mirror normal $\mathbf{n}$ and focal length f. For right now we assume f is known and defer its optimization in Section 3.5. To learn n, we have the following observation (detailed derivation is found in the supplemental material):

Theorem 1 (Fundamental matrix for 1-mirror system)**.** *For an arbitrary space point $P \in \mathbb{R}^3$, the following holds*

$$\begin{pmatrix} p' \\ 1 \end{pmatrix}^T (\mathbf{K}^{-1})^T [\mathbf{n}]_\times \mathbf{R}\mathbf{K}^{-1} \begin{pmatrix} p \\ 1 \end{pmatrix} = 0, \tag{5}$$

where $[\mathbf{n}]_\times = \begin{pmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & -n_1 \\ -n_2 & n_1 & 0 \end{pmatrix}$ *for* $\mathbf{n} = (n_1, n_2, n_3)$.

Eq. 5 paves the way of computing $\mathbf{n}$. From all matched points (p, p') in a clip, we use the RANSAC [Fischler and Bolles, 1981] algorithm to estimate an optimal $\mathbf{n}$. $\mathbf{R}, \mathbf{t}$ follow by Eq. 2. Next, to estimate the depth of human keypoints, we apply the standard triangulation technique [Hartley and Zisserman, 2003; Ma *et al.*, 2010] to Eq. 4, returning two depth values Z, Z'.

3.4 3-D Keypoint Refinement

Since our algorithm solves the 3-D keypoint estimation from matched 2-D points, the position accuracy of the 2-D keypoints becomes crucial. Unfortunately, in most DBM video, the self-occlusion inevitably appears and causes unreliable 2-D estimation.

To recover the occluded 2-D keypoints, we first divide the human keypoints into two categories: torso keypoints and the limbs keypoints. The keypoints of the torso mainly include head, neck, shoulder and hip, and the keypoints of the limbs include elbow, wrist, knee and ankle. When human moves, the motion of the torso is typically slow, yet the limbs often swing quickly. Based on this observation, we propose different strategies for refining the two sets of keypoints, respectively.

For slow-moving torso keypoints, we use *temporal filtering* to refine the occluded keypoints. As mentioned before, when the confidence of a keypoint estimation is below τ, it is regarded as an occluded point. For a torso keypoint, represent its point-sequence along time as $\{.., h_{i-1}, h_i, h_{i+1}, ...\}$. When h_i is occluded (*i.e.*, with low confidence) yet neither of h_{i-1}, h_{i+1} is, we simply use $0.5 \cdot (h_{i-1} + h_{i+1})$ to replace the original h_i. A sequential scan of all such sequences completes the refinement.

For the limbs keypoints, we refine them by jointly using epipolar geometry and human skeleton priors. In the single-mirror catadioptric system, point $\mathbf{e}$ in Figure 2 is known to be an epipole. According to the nature of the catadioptric system and epipoles, all lines that connect two matched points (*i.e.*, any pair (p, p')) will intersect at the epipole $\mathbf{e}$. If the camera that captures the video is visuable, the epipole will be right on where the camera locates. An example is shown in Figure 3.

To refine occluded points, it is important to first estimate the pixel position of the epipole. For a clip captured by a fixed camera, the epipole stays unchanged in all frames. In practice, we choose a few matched points of highest confidence scores. The line connecting each pair is estimated. We calculate the intersection between any two lines, obtaining a candidate point of the epipole. Finally, the median of all candidate points is calculated to serve as a robust estimation of the epipole's position. See Figure 3 for an illustration. Let $\mathbf{e} = (x_e, y_e)$ be the pixel coordinate of the epipole point. As shown in Figure 3, let $(a_{mr}, a_{ml}, a_r, a_l)$ be the elbow points and $(w_{mr}, w_{ml}, w_r, w_l)$ be the wrist points. Without loss of generality, assume w_r is occluded and suffers from inaccurate 2-D estimation. Since w_{mr} is not occluded, and w_r is located on the epipolar line $\overrightarrow{ew_{mr}}$. We have the epipolar constraint

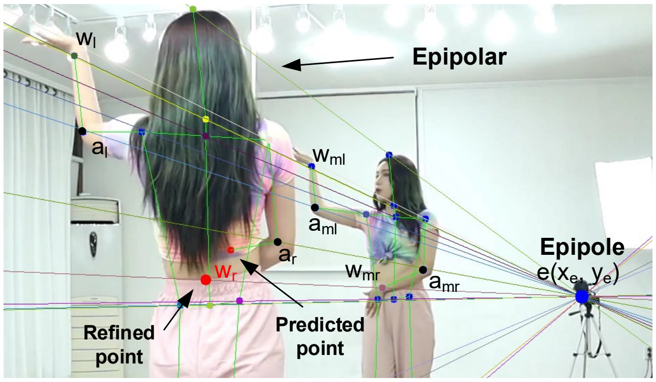

Figure 3: Illustration of the epipole in catadioptric videos.

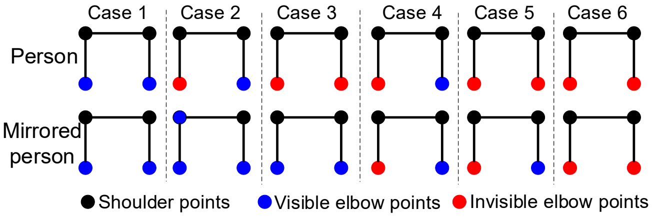

Figure 4: Six occlusion cases for shoulder-elbow point refinement.

$$(x_{w_r} - x_{w_{mr}})/(x_{\mathbf{e}} - x_{w_{mr}}) = (y_{w_r} - y_{w_{mr}})/(y_{\mathbf{e}} - y_{w_{mr}}), \tag{6}$$

where (x_{w_r}, y_{w_r}) is the candidate pixel coordinate of w_r.

For point pairs (w_l, w_{ml}), (a_l, a_{ml}), and (a_r, a_{mr}) that are not occluded, we can use the triangulation method in Section 3.3 to find their corresponding 3-D coordinates W_l, A_l and A_r. As a skeleton prior, the length of the left and right arms of a human should be equal. This prior leads to the following equation:

$$\|W_l - A_l\|_2 = \|W_r - A_r\|_2, \tag{7}$$

where $\| \cdot \|_2$ is the Euclidean distance. W_r is the 3-D coordinate of w_r, which can be inferred via triangulation after knowing (x_{w_r}, y_{w_r}). In practice, we can sample a few candidates for (x_{w_r}, y_{w_r}), and keep the one maximally satisfying Equations 6 and 7 as the best guess.

Figure 4 shows the all six possible occlusion cases of the shoulder-eblow pair. The joint optimization of Equations 6 and 7 solves Case 2 in Figure 4. Indeed, we claim that Cases 3 and 4 are also solvable by reducing them to a conditional version of Case 2. The details are deferred to the supplementary materials. To tackle other human keypoints, we associate

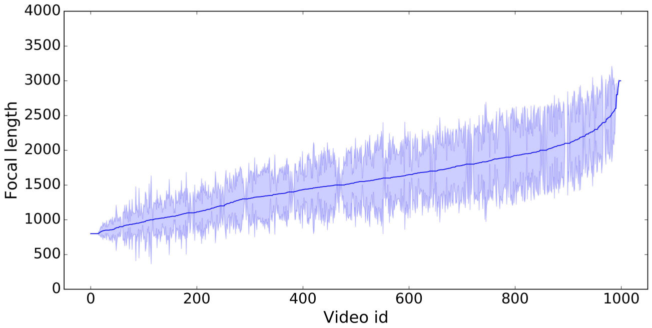

Figure 5: Mean / standard variation of clip-based estimation of the camera focal length.

Dataset	H36M	GPA	3DPW	3DHP	DBM
Year	2014	2019	2018	2017	2021
No. of Joints	32	34	24	28 or 17	16
No. of Cameras	4	5	1	14	1
No. of Subjects	11	13	18	8	-
GT Source	VICON	VICON	IMU	The Capture	Mirror
No. Images	3.6M	0.7M	68K	1.8M	87M

Table 1: Comparison of existing datasets commonly used for training and evaluating 3-D human pose estimation methods.

the human body keypoints into shoulder-eblow, elbow-wrist, hip-knee and knee-ankle. The occlusion in other configurations can be restored likewise.

3.5 Determination of Optimal Focal Length

All above computations assume a known focal length. Since f is usually in a moderate value range, we adopt an exhaustive search scheme for finding the optimal focal length. In specific, for each candidate f, we estimate all 3-D keypoints as aforementioned. Next, for each f a skeleton symmetry score is computed over all left-right limb pairs, such as $\min(\|W_l - A_l\|_2, \|W_r - A_r\|_2)/\max(\|W_l - A_l\|_2, \|W_r - A_r\|_2)$. The f that attains best skeleton symmetry will be kept. Figure 5 shows the mean / std of estimated focal lengths over all clips in a video. In practice we apply the mean to all clips.

4 Evaluations

4.1 Investigation of DBM Data

After all above data filtering operatioins, we divide the train/val/test set according to the ratio of 7:1:2, and get 124243/18114/33295 clips respectively. During the division process, we ensure that the clips of the same video do not appear in different sets. Table 1 contrasts the key statistics of mainstream 3-D human pose benchmarks. Our intelligently-collected DBM surpasses all others in term of data scale.

We next evaluate the quality of epipole estimation, which is crucial for 3-D keypoint refinement. Geometrically, in the mirrored image, the position of the camera is the position of the epipole point, as the case in Figure 3. We manually pick 100 videos that have the camera visible, and then mark the position of the camera as the groundtruth of epipole point. Table 2 summarizes the Euclidean distance between the estimated pole coordinates and the true pole coordinates by varying point pairs involved in the computation. More point pairs are observed to generate more robust estimation.

NumPointPair	100	200	300	400	500	600
L2 distance	42.0	27.9	25.1	23.8	21.4	19.7

Table 2: Accuracy of epipole point estimation with respect to the number of involved keypoint pairs (unit of distance: pixel). The image size is 1920x1080 or 1280x720.

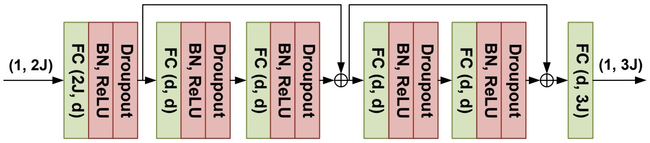

Figure 6: An instantiation of our fully-connected 3-D pose estimation architecture. The input consists of 2-D keypoints for $J = 14$ joints. Fully-connected layers are in green where $2J$, d denotes $2 \times J$ inputs dimensions and $d = 1024$ output dimensions.

Others about the distribution of the quality of 3-D keypoints in human skeleton prior (such as human skeleton symmetry), and the influence of refinement on the quality of 3-D keypoints are also important. We present more details in the supplemental materials.

4.2 3-D Pose Estimation Model

Following [Pavllo *et al.*, 2019], we use a fully-connected network with residual connections for 3-D pose estimation. The architecture is shown in Figure 6. The input layer takes the concatenated (x, y) coordinates of J joints for each frame and outputs a 1024-dimensional feature. This is followed by two ResNet-style blocks which are surrounded by a skip-connection. Each block performs two fully-connected layers and they are followed by batch normalization, rectified linear units, and dropout. Finally, the last layer outputs a prediction of the 3-D poses for the input frame.

4.3 Datasets for Comparison

Human3.6M (H36M) [Ionescu *et al.*, 2014] is a popular motion capture dataset containing 3.6 million video frames for 11 subjects, of which 7 are annotated with 3-D poses. Following [Pavllo *et al.*, 2019], we train on 5 subjects (S1, S5, S6, S7, S8) and test on 2 subjects (S9, S11). 3DPW [von Marcard *et al.*, 2018] is a wild video dataset taken from a moving phone camera. It contains 60 videos and about 68k frames covering multiple scenarios and actions, with the annotated 3-D positions of 24 keypoints. 3DHP [Mehta *et al.*, 2017] is a human pose dataset captured in a multi-camera studio, covering a wide range of viewpoints and actions. It contains about 1.8 million frames recorded by 14 cameras, including 8 actors and 8 activity sets. Each frame has the annotation of 3-D keypoint positions (28 for train and 17 for test). We train and test our model on the original train/test split for 3DPW and 3DHP.

4.4 Data Distribution Analysis and Visualization

To study the human pose coverage of the proposed DBM data, we adopted the t-SNE [Maaten and Hinton, 2008] to visualize

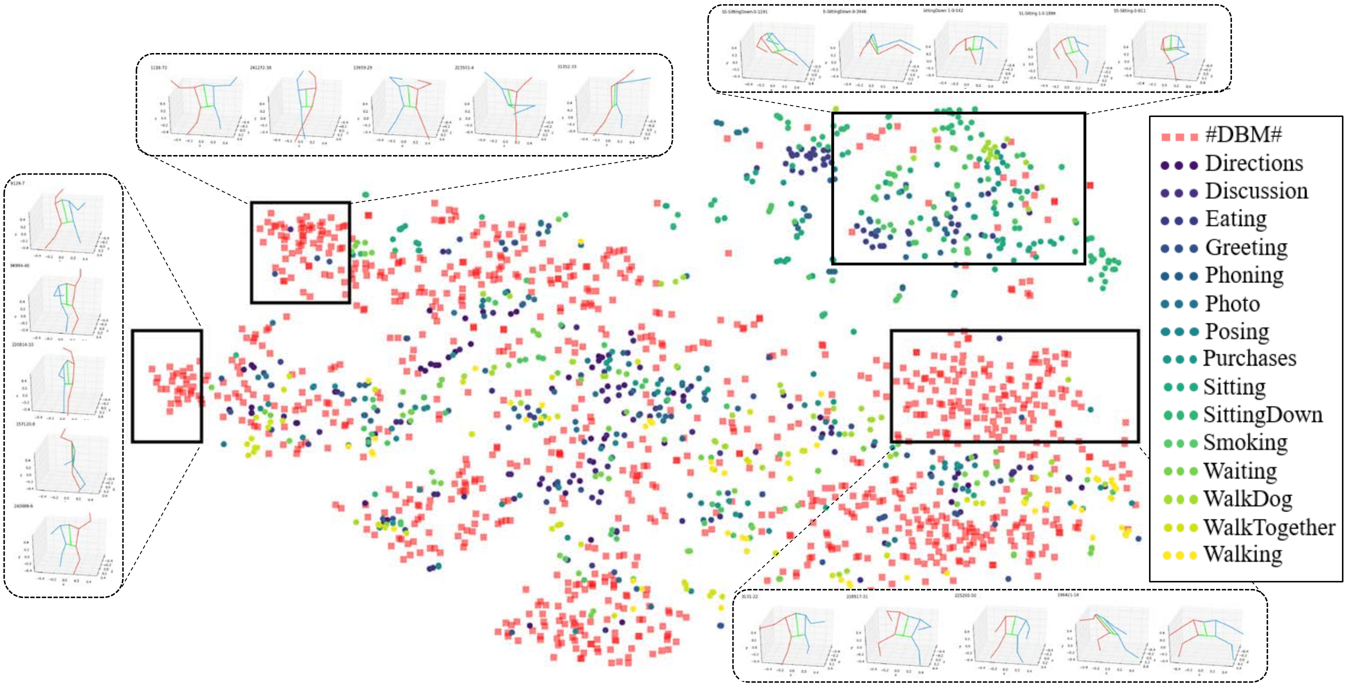

Figure 7: t-SNE visualization of the sampled poses in DBM and H36M. Different colors represent different action categories. Some representative poses for each selected area are shown in the dash boxes.

Test/Training	MPJPE					P-MPJPE				
	H36M	DBM	DBM+	3DPW	3DHP	H36M	DBM	DBM+	3DPW	3DHP
H36M	44.1	83.9	69.7	157.7	81.9	32.7	56.9	50.9	89.8	57.6
DBM	44.7	27.4	27.5	45.2	41.3	22.7	9.4	9.4	20.7	18.4
3DPW	128.6	125.6	119.5	72.9	114.8	80.6	70.2	60.1	50.7	73.7
3DHP	86.3	109.8	73.3	158.1	44.5	61.6	71.8	54.3	93.9	35.7

Table 3: Cross-data test error in MPJPE (left) and P-MPJPE (right). Red color indicates training and testing are on the same dataset.

the pose distribution in the dataset. Specifically, we randomly select 1000 poses from the training set of DBM and H36M respectively, and project them into the 2-dimensional diagram, as shown in Figure 7. Critically, since most DBM videos contain standing-like poses, poses like “sitting” and “sitting down” are inadequate in DBM in comparison with H36M, as seen in the top-right zone. Otherwise, DBM covers larger spectrum of human poses, such as the pose of raising both hands, and the pose of separating the hands and legs together. We highlight some zones with exemplar poses in Figure 7. This shows the DBM dataset can enrich existing data, thereby promoting the development of the entire community.

4.5 Cross-Dataset Generalization

A good 3-D pose dataset shall generalize well on other cross-scenario datasets. To validate the generalization capability of different 3-D pose datasets, we train our 3-D pose estimation model separately on four 3-D pose estimation datasets, DBM, H36M, 3DPW and 3DHP, and conduct cross-dataset evaluation by testing these models on the rest of the datasets.

We consider two metrics in our cross-dataset evaluation, following common practice. The first is mean per-joint position error (MPJPE) in millimeters, calculating the mean error between the predicted joint positions and the ground truth. The second is procrustes analysis MPJPE (P-MPJPE) in millimeters, calculating the MPJPE after alignment with the ground truth in rigid transformation, rotation and scale. For the whole four datasets, we train the same 3-D pose estimation model as specified in Figure 6. Hyperparameters such as learning rate and dropout ratio are the same as those in [Pavllo *et al.*, 2019]. In order to make a fair comparison on different datasets, the input of the 3-D pose estimation model is the ground truth 2-D keypoints. Following [Wang *et al.*, 2020], we choose the mutual 14 joints of DBM, H36M, SURREAL, 3DPW and 3DHP. Because the bone lengths of the human in each dataset are different, we use the average bone length of the target data to normalize the predicted 3-D coordinates before performing cross-data tests.

The cross-dataset evaluation results are shown in Table 3. In most settings, DBM and 3DHP are two best performers. DBM shows some more extraordinary results in terms of P-MPJPE. For instance, model trained on DBM achieves the optimal P-MPJPE when transferred to H36M and 3DPW compared to other datasets. We would emphasize that data in

Action		Average	Pose	Walk	WalkT.	Photo	SitD.	Phone	Purch.	Eat	Sit	Dir.	Greet	WalkD.	Disc.	Wait	Smoke
MPJPE	3DHP	81.9	97.2	75.1	77.4	84.7	105.2	70.7	91.7	64.0	63.6	88.0	92.4	76.1	78.9	92.0	70.8
	DBM	83.9	76.2	72.3	77.4	87.0	121.2	84.0	79.2	83.3	114.9	71.9	78.6	74.1	73.0	80.0	84.2
	DBM+	69.7	66.8	60.8	65.5	72.8	111.8	63.1	69.0	62.7	69.3	65.4	68.5	66.6	63.7	74.5	63.8
	H36M	44.1	48.1	37.1	38.1	48.0	49.0	43.5	42.1	38.6	49.1	42.7	44.8	43.8	48.2	44.6	42.7
	H36M+DBM	39.4	41.4	32.3	34.1	46.0	46.1	39.0	36.6	34.3	44.8	36.8	38.6	41.3	40.7	39.5	38.9
P-MPJPE	3DHP	57.6	61.9	55.7	59.2	56.9	81.3	52.1	57.3	46.6	53.9	55.5	60.0	56.8	51.8	61.7	52.4
	DBM	56.9	43.9	51.7	55.1	58.1	97.2	56.4	52.8	58.5	79.5	42.8	49.2	54.7	46.6	51.7	56.3
	DBM+	50.9	43.4	47.5	49.9	52.4	89.0	46.1	51.0	43.6	54.2	42.7	46.8	53.6	44.9	51.0	47.1
	H36M	32.7	34.6	27.5	29.8	35.7	38.9	30.8	29.6	29.2	36.6	30.9	33.1	33.1	34.8	32.4	32.5
	H36M+DBM	30.7	30.6	25.4	26.9	35.6	38.8	29.3	27.3	27.0	35.3	27.7	30.2	33.6	30.8	29.9	31.6

Table 4: Test errors of fine-grained activities in MPJPE (top rows) and P-MPJPE (bottom rows). In all evaluations, the target dataset for cross-scenario testing is H36M. Due to limited space, each action is indicated by its abbreviation, full names are given in legend of Figure 7.

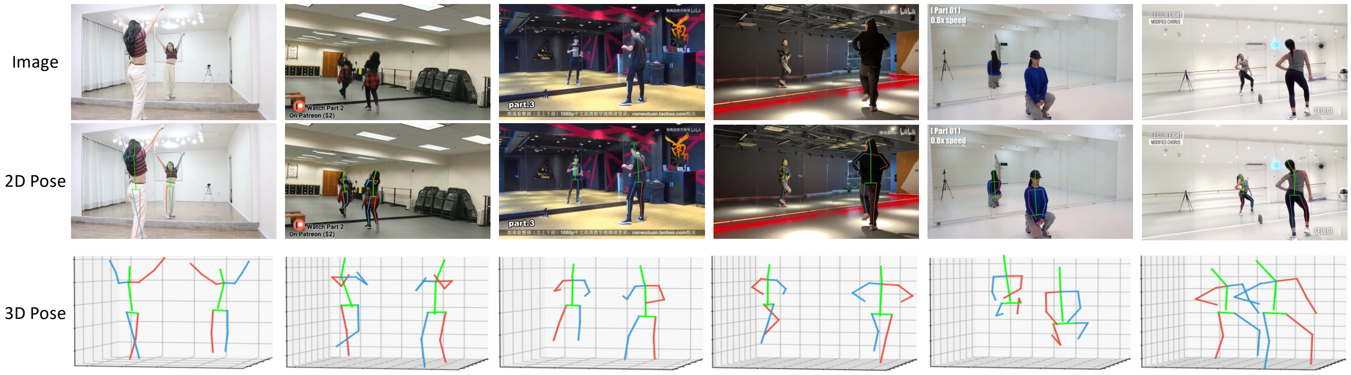

Figure 8: Visualization of 2-D pose and 3-D pose in our DBM dataset. The first row is the original video frames, and the second and third rows are the corresponding 2-D and 3-D poses of the persons, respectively. Better viewing if enlarging the images.

H36M and 3DHP are obtained by sensors, and the way we obtain DBM data is much cheaper in contrast. Nonetheless, the cheaply obtained DBM shows a generalization ability comparable or better to H36M and 3DHP. When compared to 3DPW, whose data is also obtained in a cheap way using IMU, model trained on DBM performs much better generalization ability on H36M (83.9 v.s. 157.7) and 3DHP (109.8 v.s. 158.1) test sets.

To further conduct fine-grained study, we test our DBM model on different kind of actions in H36M and compare them to the results trained on 3DHP. The results are shown in Table 4. Our main observation is that our performance in many actions is better than 3DHP except for sitting-related actions. DBM is a dataset related to dancing movements, thus the sitting scenario is relatively rare. Therefore, we conduct another experiment to complement 280K sitting poses from 3DHP to DBM, forming a new data DBM+. With the supplied sitting data, DBM almost dominates all action categories. In addition, we compare the model trained on H36M or H36M+DBM. The observed improvement (*e.g.*, 32.7 v.s. 30.7 in P-MPJPE) proves the DBM is strongly complementary to existing datasets.

4.6 Some Examples

In Figure 8, we illustrate some examples of the DBM dataset. The first row is the original video frames, and the second and third rows are the 2-D and 3-D pose estimation results of the persons obtained by our proposed method.

5 Concluding Remarks

This work constructs a new large-scale DBM benchmark for 3-D human pose estimation. We explore a new method of collecting 3-D pose from webly-available catedioptric videos, and propose a novel algorithm that combines geometry and human skeleton priors. In future we will include active annotation for further improving the data quality.

Acknowledgments

This work is supported by National Key R&D Program of China (2018AAA0100702), National Natural Science Foundation of China (61772037) and Beijing Natural Science Foundation (Z190001).

References

[Andriluka *et al.*, 2014] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[Cao *et al.*, 2017] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[Chen *et al.*, 2016] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016.

[Chen *et al.*, 2018] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018.

[Fang *et al.*, 2017] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

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

[Gluckman and Nayar, 2001] Joshua Gluckman and Shree K. Nayar. Catadioptric stereo using planar mirrors. *IJCV*, 44(1):65–79, 2001.

[Guan *et al.*, 2009] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, 2009.

[Hartley and Zisserman, 2003] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2003.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Ionescu *et al.*, 2014] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014.

[Jain *et al.*, 2010] Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. Moviereshape: Tracking and reshaping of humans in videos. *TOG*, 29(6):1–10, 2010.

[Joo *et al.*, 2019] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart C. Nabbe, Iain A. Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 41(1):190–204, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[Ma *et al.*, 2010] Yi Ma, Stefano Soatto, Jana Koseck, and S. Shankar Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, 2010.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Mariottini *et al.*, 2012] Gian Luca Mariottini, Stefano Scheggi, Fabio Morbidi, and Domenico Prattichizzo. Planar mirrors for image-based robot localization and 3-d reconstruction. *Mechatronics*, 22(4):398–409, 2012.

[Martinez *et al.*, 2017] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[Mehta *et al.*, 2017] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017.

[Pavlakos *et al.*, 2017] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017.

[Pavllo *et al.*, 2019] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

[Qiu *et al.*, 2019] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *ICCV*, 2019.

[Remelli *et al.*, 2020] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *CVPR*, 2020.

[Sigal *et al.*, 2010] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.*, 87(1-2):4–27, 2010.

[Sun *et al.*, 2018] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018.

[Sun *et al.*, 2019] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[von Marcard *et al.*, 2018] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018.

[Wang *et al.*, 2019] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless C. Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *CoRR*, abs/1905.07718, 2019.

[Wang *et al.*, 2020] Zhe Wang, Daeyun Shin, and Charless C. Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In Adrien Bartoli and Andrea Fusiello, editors, *ECCV*, 2020.

[Wei and Chai, 2010] Xiaolin Wei and Jinxiang Chai. Videomocap: Modeling physically realistic human motion from monocular video sequences. In *SIGGRAPH*. 2010.

[Xiao *et al.*, 2018] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.

[Zhou *et al.*, 2016] Fuqiang Zhou, Xinghua Chai, Xin Chen, and Ya Song. Omnidirectional stereo vision sensor based on single camera and catoptric system. *Applied optics*, 55(25):6813–6820, 2016.