Reconstructing 3D Human Pose by Watching Humans in the Mirror

Qi Fang* Qing Shuai* Junting Dong Hujun Bao Xiaowei Zhou†
State Key Lab of CAD&CG, Zhejiang University

Abstract

In this paper, we introduce the new task of reconstructing 3D human pose from a single image in which we can see the person and the person's image through a mirror. Compared to general scenarios of 3D pose estimation from a single view, the mirror reflection provides an additional view for resolving the depth ambiguity. We develop an optimization-based approach that exploits mirror symmetry constraints for accurate 3D pose reconstruction. We also provide a method to estimate the surface normal of the mirror from vanishing points in the single image. To validate the proposed approach, we collect a large-scale dataset named Mirrored-Human, which covers a large variety of human subjects, poses and backgrounds. The experiments demonstrate that, when trained on Mirrored-Human with our reconstructed 3D poses as pseudo ground-truth, the accuracy and generalizability of existing single-view 3D pose estimators can be largely improved. The code and dataset are available at https://zju3dv.github.io/Mirrored-Human/.

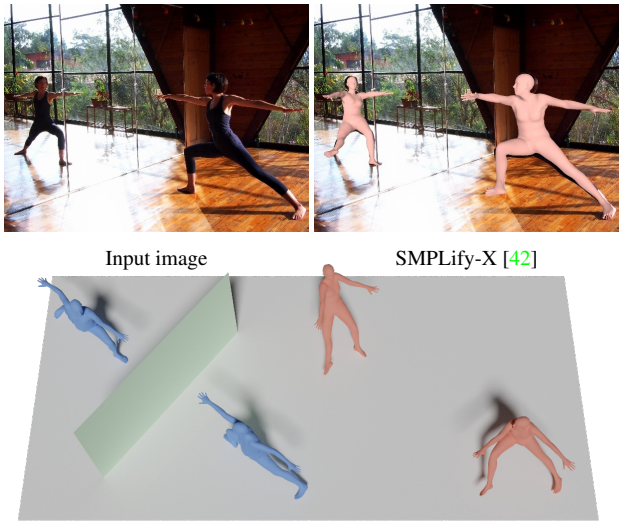
1. Introduction

3D human pose estimation has ubiquitous applications in sport analysis, human-computer interaction, and fitness and dance teaching. While there has been remarkable progress in 3D pose estimation from a monocular image or video [22, 37, 25, 24, 59], inevitable challenges such as the depth ambiguity and the self-occlusion are still unsolved.

In many scenes like dancing rooms and gyms, people are often in front of a mirror. In this case, we are able to see the person and his/her mirror image simultaneously. The mirror image actually provides an additional virtual view of the person, which can resolve the single-view depth ambiguity if the mirror is properly placed. Moreover, unseen part of the person can also be observed from the mirror image, so that the occlusion problem can be alleviated.

In this paper, we investigate the feasibility of leveraging such mirror images to improve the accuracy of 3D human pose estimation. We develop an optimization-based framework with mirror symmetry constraints that are applicable without knowing the mirror geometry and camera parameters. We also provide a method to utilize the properties of vanishing points to recover the mirror normal along with the camera parameters, so that an additional mirror normal constraint can be imposed to further improve the human pose estimation accuracy. The effectiveness of our framework is validated on a new dataset for this new task with 3D pose ground-truth provided by a multi-view camera system.

An important application of the proposed approach is to generate pseudo ground-truth annotations to train existing 3D pose estimators. To this end, we collect a large-scale set of Internet images that contain people and mirrors and generate 3D pose annotations with the proposed optimization method. The dataset is named Mirrored-Human. Compared with existing 3D human pose datasets [19, 33, 56] that are captured with very few subjects and background scenes, Mirrored-Human has a significantly larger diversity



3D visualization of our (left) and SMPLify-X (right) results

Figure 1: While the state-of-the-art single-view 3D pose estimator [42] yields a small reprojection error, the recovered 3D poses may be erroneous due to the depth ambiguity. We make use of the mirror in the image to resolve the ambiguity and reconstruct more accurate human pose as well as the mirror geometry.

*Equal Contribution. †Corresponding author.

in human poses, appearances and backgrounds, as shown in Fig. 7. The experiments show that, by combining Mirrored-Human with existing datasets as training data, both accuracy and generalizability of existing 3D pose estimation methods can be significantly improved for both single-person and multi-person cases.

In summary, we make the following contributions:

- We introduce a new task of reconstructing human pose from a single image in which we can see the person and the person's mirror image.
- We develop a novel optimization-based framework with mirror symmetry constraints to solve this new task, as well as a method to recover mirror geometry from a single image.
- We collect a large-scale dataset named Mirrored-Human from the Internet, provide our reconstructed 3D poses as pseudo ground-truth, and show that training on this new dataset can improve the performance of existing 3D human pose estimators.

2. Related work

2.1. 3D human pose

Benefiting from neural networks, the task of monocular 3D human pose estimation has made considerable progress. The skeleton-based approaches either lift the 2D pose to 3D [32, 7, 57], or adopt an end-to-end manner to regress the 3D pose directly from the image [51, 52, 67, 43]. The model-based approaches estimate the pose and shape simultaneously with parametric models [2, 31, 41, 60]. Approaches along this line can be divided into two categories. Optimization-based methods fit the model using 2D evidence and some human body priors [5, 28, 42, 11]. Regression-based methods directly regress the model parameters from the image [22, 40, 61, 23, 64]. Kolotouros *et al*. [25] incorporate their advantages and propose a self-improving framework. To relax the heavy reliance on the model's parameter space, model-free approaches directly regress the 3D locations of the mesh vertices [26, 37, 9]. For multi-person cases, previous works focus on how to extend single-person frameworks to multi-person ones [47, 34] or model the interaction between people [63, 14]. Some recent works [36, 66, 29, 30, 12] explore the representation of the absolute depth in the camera coordinate system.

In monocular settings, the depth ambiguity is inevitable. One solution is to utilize the additional supervision signal. Pavlakos *et al*. [43] use the ordinal depths of human joints to weakly supervise the network. Kanazawa *et al*. [22] train a discriminator network to judge if the estimated pose is reasonable. Some other works use temporal constraints [24, 55, 44] or geometric self-consistency [8]. Another line of work resolves the ambiguity with scene layouts. Hassan *et al*. [17] exploit static 3D scene structures with the inter-penetration and the contact constraints. Others [34, 63] integrate the ground plane information to recover the 3D location and pose. We use the additional view provided by the mirror reflection to resolve the depth ambiguity.

Learning-based methods are inseparable from the training data. Existing widely-used 3D datasets such as Human3.6M [19], Human Eva [48], MPI-INF-3DHP [33] and Panoptic Studio [21] are built with motion capture systems and thus have limited appearance and pose diversity. 3DPW [56] is a recent dataset consisting of ordinary activities. The combination of a camera and several IMUs attached to the human body provides accurate 3D poses, but the dataset still lacks diversity. Kanazawa *et al*. [23] contribute some Internet datasets which however lack 3D annotations. Arnab *et al*. [3] propose a bundle-adjustment-based algorithm, based upon which they generate 3D annotations for Internet data. However, this method still suffers from the depth ambiguity derived from monocular videos. To summarize, a *large-scale Internet* dataset with *3D annotations* is still missing. We collect numerous Internet images containing mirrors and people with a large diversity in appearances and poses, and build a dataset using our reconstructed 3D poses as pseudo ground-truth, to address this problem to some extent.

2.2. Reconstruction with mirrors

Earlier years have witnessed some researches on the mirror geometry of a catadioptric system (mirrors + lenses), especially for reconstruction and extrinsic camera calibration [54, 45]. For 3D reconstruction that is more relevant to our task, some works use the configuration of two planar mirrors to capture stereo images [38, 15, 53, 27] or produce four virtual views by assuming there is a one-time interreflection between two mirrors [62]. They calibrate the mirror based on the image correspondences or silhouettes. Nguyen *et al*. [39] use a depth sensor and at least two mirrors, trying to remove depth distortions. Akay *et al*. [1] reconstruct 3D object with a mirror and RGBD cameras. Hu *et al*. [18] only use one mirror and one RGB camera, but they need to label the object and mirror region. To the best of our knowledge, the task of exploiting mirror symmetry to recover 3D human pose and shape from an Internet image has not been discussed in literature.

Furthermore, not only does the symmetry property exist in scenes with mirrors, but it also appears on symmetrical objects, which has been explored to reconstruct faces, cars, etc. Some works [49, 20] explicitly detect the plane of symmetry with 2D correspondences and others [58] implicitly use the symmetry prior.

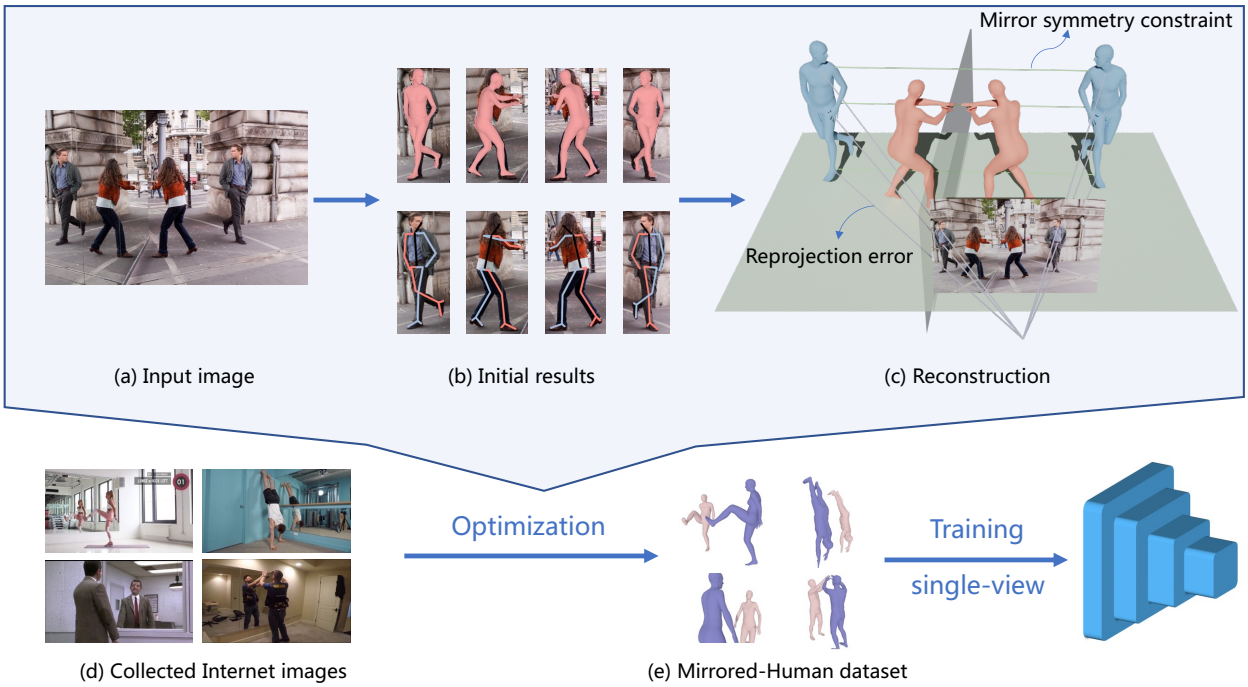


Figure 2: **Overview of our approach.** Given the input image (a), we first estimate the 2D keypoints and SMPL parameters as the initialization (b). Then, we minimize the reprojection error with mirror symmetry constraints for reconstruction (c). We collect a considerable number of Internet images (d) and build a dataset named Mirrored-Human (e) with pseudo ground-truth generated by our framework. The dataset can be used for the training of single-view methods.

3. Methods

Fig. 2 presents the pipeline of our framework. Given an image that contains a person and a mirror, our goal is to recover the human mesh considering the mirror geometry. The key insight is that the person and his/her mirror image can be treated as two people, and we reconstruct them together with the mirror symmetry constraints. This section will be organized as follows. First, the formulation of single-person mesh recovery is introduced (Sec. 3.1). Then the mirror symmetry constraints that relate the two people will be elaborated (Sec. 3.2). Finally, the objective functions and the whole optimization are described (Sec. 3.3).

3.1. Human mesh recovery with SMPL model

We adopt the SMPL model [31] as our human representation. The SMPL model is a differentiable function $\boldsymbol{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3\times N_v}$ mapping the pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ to a triangulated mesh with $N_v = 6890$ vertices. The 3D body joints $\boldsymbol{J}(\boldsymbol{\theta}, \boldsymbol{\beta})$ of the model can be defined as a linear combination of the mesh vertices. Hence for N_j joints, we defined the body joints $\boldsymbol{J}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3\times N_j} = \mathcal{J}(\boldsymbol{M}(\boldsymbol{\theta}, \boldsymbol{\beta}))$, where $\mathcal{J}$ is a pre-trained linear regressor. Let $\boldsymbol{R} \in SO(3)$ and $\boldsymbol{T} \in \mathbb{R}^3$ denote the global rotation and translation, respectively.

Given an image and the detected 2D bounding boxes, the 2D human keypoints $\boldsymbol{W}$ can be estimated with the cropped regions. The objective function for human mesh recovery generally consists of a reprojection term L_{2d} and a prior term L_p with respect to variables $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{R}$ and $\boldsymbol{T}$.

The reprojection term penalizes the weighted 2D distance between the estimated 2D keypoints $\boldsymbol{W}$ with the confidence c, and the corresponding projected SMPL joints:

$$L_{2d} = \sum_i c_i \rho(\boldsymbol{W}_i - \Pi_K(\boldsymbol{R}\boldsymbol{J}(\boldsymbol{\theta}, \boldsymbol{\beta})_i + \boldsymbol{T})), \quad (1)$$

where Π_K is the projection from 3D to 2D through the intrinsic parameter K. ρ denotes the Geman-McClure robust error function for suppressing noisy detections.

The human body priors are used to encourage realistic 3D human mesh results. Since the pose and shape parameters $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}})$ estimated by a neural network can be viewed as learned prior, the final results are supposed to be close to them:

$$L_p = ||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2^2 + \lambda_\beta ||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2^2, \quad (2)$$

where λ_β is a weight.

3.2. Mirror-induced constraints

If there is a mirror in the image, the relation between the person and the mirrored person can be used to enhance the

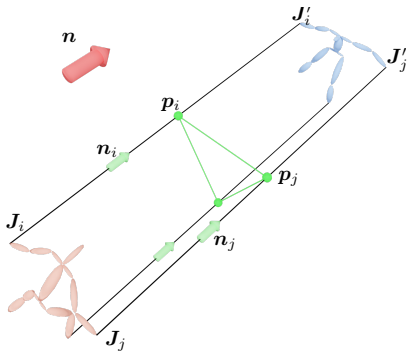


Figure 3: **An illustration of mirror-induced constraints.** The line segment connecting the joint $\boldsymbol{J}_i$ and its mirrored joint $\boldsymbol{J}_i'$ has the direction $\boldsymbol{n}_i$ and the middle point $\boldsymbol{p}_i$. Theoretically, $\boldsymbol{n}_i//\boldsymbol{n}_j$, and $\boldsymbol{n}_i \perp \overline{\boldsymbol{p}_i\boldsymbol{p}_j}$. If the mirror normal $\boldsymbol{n}$ (red arrow) is known, $\boldsymbol{n}//\boldsymbol{n}_i$ and $\boldsymbol{n}//\boldsymbol{n}_j$ should be satisfied as well.

reconstruction performance. This relation is a simple reflection transformation if the mirror geometry is known, which however is impracticable for an arbitrary image from the Internet. To tackle this problem and take advantage of the characteristic of the mirror, the following mirror-induced constraints are introduced, as illustrated in Fig. 3. Note that all symbols with the superscript prime refer to variables related to the mirrored person unless specifically mentioned.

Mirror symmetry constraints: Since the adopted human representation disentangles the orientation $\boldsymbol{R}$, pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$, $\boldsymbol{\beta}$ can be shared by the person and the mirrored person, and $\boldsymbol{\theta}$ is related to $\boldsymbol{\theta}'$ by a simple reflection operation as follows:

$$\boldsymbol{\beta}' = \boldsymbol{\beta},\ \boldsymbol{\theta}' = \mathcal{S}(\boldsymbol{\theta}), \tag{3}$$

where $\mathcal{S}(\cdot)$ denotes the reflection operation on axis angles.

As Eq. 3 does not take $\boldsymbol{R}$ and $\boldsymbol{T}$ into consideration, the constraint on 3D keypoints can be imposed to estimate the human orientation and position better. We abbreviate the global coordinates of the i-th joint $\boldsymbol{R}\boldsymbol{J}(\boldsymbol{\theta},\boldsymbol{\beta})_i + \boldsymbol{T}$ as $\boldsymbol{J}_i$. Given a pair of body joints i, j, we denote the direction of the line segment $\overline{\boldsymbol{J}_i\boldsymbol{J}_i'}$, $\overline{\boldsymbol{J}_j\boldsymbol{J}_j'}$ as $\boldsymbol{n}_i$, $\boldsymbol{n}_j$ and the middle point of them as $\boldsymbol{p}_i$, $\boldsymbol{p}_j$, respectively. Ideally, $\boldsymbol{n}_i$ should be parallel to $\boldsymbol{n}_j$ and $\boldsymbol{p}_i, \boldsymbol{p}_j$ are supposed to be on the mirror plane. Despite the fact that the mirror geometry is unknown, it needs to be satisfied that $\boldsymbol{n}_i$ is perpendicular to the line $\overline{\boldsymbol{p}_i\boldsymbol{p}_j}$. So for any pair of joints, we minimize the sum of the L2 norm of the cross product between $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$, and the inner product between $\boldsymbol{n}_i$ and $\boldsymbol{p}_j - \boldsymbol{p}_i$:

$$L_s = \sum_{(i,j)} (||\boldsymbol{n}_i \times \boldsymbol{n}_j||_2 + ||\boldsymbol{n}_i \cdot (\boldsymbol{p}_j - \boldsymbol{p}_i)||_2). \tag{4}$$

Mirror normal constraint: A mirror can be represented as a plane, parameterized as its normal and position. If its normal $\boldsymbol{n}$ is known, the geometric properties of the mirror can thus be utilized explicitly by constraining $\boldsymbol{n}_i$ and $\boldsymbol{n}$ to be parallel with the following loss function:

$$L_n = \sum_i ||\boldsymbol{n} \times \boldsymbol{n}_i||_2. \tag{5}$$

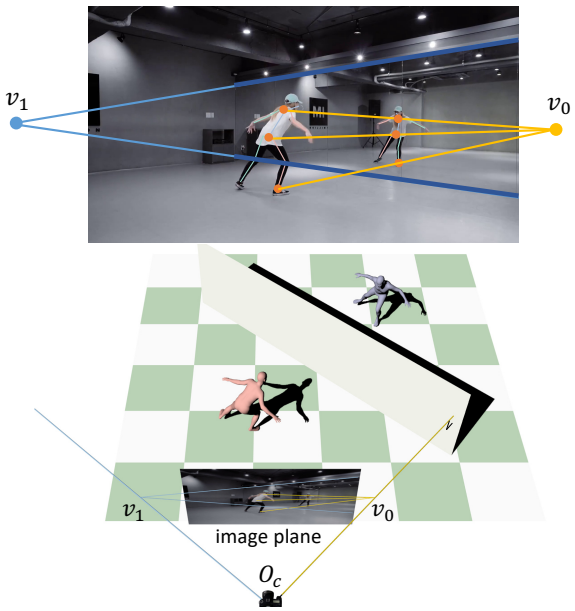


Figure 4: **Vanishing points in an image containing a person and a mirror.** In most cases at least two vanishing points can be found, where $\boldsymbol{v}_0$ comes from 2D human keypoints, and $\boldsymbol{v}_1$ comes from the annotated mirror edges. O_c denotes the camera center. Note that $\overline{O_c v_0}//\boldsymbol{n}$ and $\overline{O_c v_1} \perp \boldsymbol{n}$, where $\boldsymbol{n}$ is the mirror normal.

Mirror normal estimation: Though the mirror normal is not directly available, the vanishing points can be used to estimate it. The vanishing point of lines with direction $\boldsymbol{n}$ in 3D space is the intersection $\boldsymbol{v}$ of the image plane with a ray through the camera center with direction $\boldsymbol{n}$ [16]:

$$\boldsymbol{v} = K\boldsymbol{n}, \tag{6}$$

where the vanishing point $\boldsymbol{v} \in \mathbb{R}^3$ is in the form of homogeneous coordinates and K is the camera intrinsic matrix.

Eq. 6 reveals that obtaining the mirror normal $\boldsymbol{n}$ requires both K and $\boldsymbol{v}$. As the parallel lines connecting points on the real object and corresponding points on the mirrored object are perpendicular to the mirror, the vanishing point $\boldsymbol{v}$ with this direction can be estimated through their 2D positions. To get such correspondences, some previous works require additional inputs such as masks [18], which is infeasible for images from the Internet. Fortunately, since 2D human keypoints provide robust semantic correspondences, *e.g.* the left ankle of the real person and the right ankle of the mirrored person, this vanishing point can be acquired naturally and automatically in our setting ($\boldsymbol{v}_0$ in Fig. 4).

Note that if the intrinsic matrix K is provided, the mirror normal can thus be solved easily through $\boldsymbol{n} = K^{-1}\boldsymbol{v}_0$, otherwise K should be calibrated from a single image if possible. From the projective geometry [16], we know that it is possible to calibrate the camera intrinsic parameters from a single image. Suppose the camera has zero skew and square pixels. The intrinsic matrix K can be computed via three orthogonal vanishing points. Additionally, if the principal point is assumed to be in the image center (only the focal length is unknown), K can be computed via only two orthogonal vanishing points. Please refer to the supplementary material for more details.

As we have stated, one vanishing point $\boldsymbol{v}_0$ has been acquired based on reliable 2D human keypoints. Different from the general scene where finding orthogonal relations may be difficult, our setting contains richer information. Fig. 4 shows that if we annotate the mirror edges, at least one vanishing point $\boldsymbol{v}_1$ orthogonal to $\boldsymbol{v}_0$ can be obtained. With these vanishing points, the calibration can be performed. Note that images from the same video share the same intrinsic matrix K, thus the annotation process is not laborious.

The mirror normal constraint is optional, which depends on how easy it is to find mirror edges. In the experiment, we will show that our method can still achieve satisfactory performance without the mirror normal constraint.

3.3. Objective function and optimization

Combining all discussed above, the final objective function to optimize can be written as:

$$\min_{\Theta,\Theta'} \; L_{2d} + L'_{2d} + \lambda_p(L_p + L'_p) + \lambda_s L_s + \lambda_n L_n \\ s.t. \;\; \boldsymbol{\beta}' = \boldsymbol{\beta}, \; \boldsymbol{\theta}' = \mathcal{S}(\boldsymbol{\theta}), \tag{7}$$

where $\Theta = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{T}\}$ and $\Theta' = \{\boldsymbol{\theta}', \boldsymbol{\beta}', \boldsymbol{R}', \boldsymbol{T}'\}$. L'_{2d} and L'_p refer to the reprojection term and the prior term of the mirrored person, respectively. λ_p, λ_s and λ_n are weights. λ_n is set to zero whenever the mirror normal is unavailable. If there are two or more people, the optimization can be done for each subject separately.

We optimize Eq. 7 with respect to all parameters using L-BFGS and PyTorch. An off-the-shelf model [25] is adopted to generate the initial estimation. Given the 2D keypoints [50, 6], $\boldsymbol{R}$ and $\boldsymbol{T}$ are further optimized by aligning the initial SMPL model to the 2D keypoints. To improve the robustness of the initialization, we select the person with smaller reprojection error and apply the selected pose parameter to the other person after a reflection operation.

4. Evaluation

4.1. Dataset for evaluation

Since no dataset exists for our task which contains both mirrors and labeled 3D human keypoints, we collect a new dataset with six calibrated HD cameras, as shown in Fig. 5. All videos are recorded with a resolution of 1920×1080 pixels at 30 fps. Each person is performing various actions in front of a large mirror. A calibration board is placed on the mirror, thus the mirror geometry can be easily determined, which can also be used to evaluate the estimated mirror geometry. Note that the ground-truth of 3D keypoints is generated from all views. 2D bounding boxes, 2D keypoints, and the correspondences between multiple people are annotated manually.

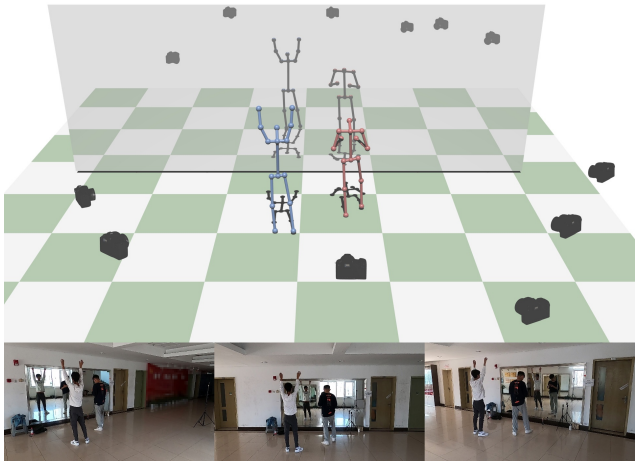


Figure 5: **The configuration of our evaluation set.** The poses of all subjects are captured with six cameras. The camera extrinsic paramters, camera intrinsic parameters, and mirror geometry are calibrated.

4.2. Reconstruction evaluation and ablation study

The 3D keypoints generated by this multi-view system are used as ground-truth to evaluate the reconstruction accuracy. Metrics include MPJPE, PA-MPJPE, and MRPE. MPJPE is the distance (mm) between predicted and ground-truth 3D keypoints after root alignment. PA-MPJPE is calculated with further Procrustes alignment. MRPE is defined as the distance (mm) between the predicted and the ground-truth root joint, which evaluates the absolute root position accuracy.

Pose Reconstruction: Two previous methods are compared here. SMPLify-X [42] is the state-of-the-art optimization-based method and SPIN [25] is the state-of-the-art regression-based method. 'Baseline' combines the two methods by using [25] as initialization and [42] as optimization. In Table 1, comparing 'Ours (full)' with first three rows, the result shows that our approach outperforms previous methods by a large margin.

The ablation study is performed to show the effect of our

Methods	MPJPE ↓	PA-MPJPE ↓	MRPE ↓
SMPLify-X [42]	143.73	90.57	368.00
SPIN [25]	109.79	67.42	167.62
Baseline ([25]+[42])	82.92	61.47	147.44
Ours (w/o L_s, L_n)	53.81	34.48	108.43
Ours (w/o L_n)	39.52	33.24	101.91
Ours (full)	**38.77**	**32.96**	**93.22**

Table 1: **Quantitative analysis.** 'Baseline' uses [25] as initialization and [42] as the optimization method to fit the body model to 2D keypoints for each person separately.

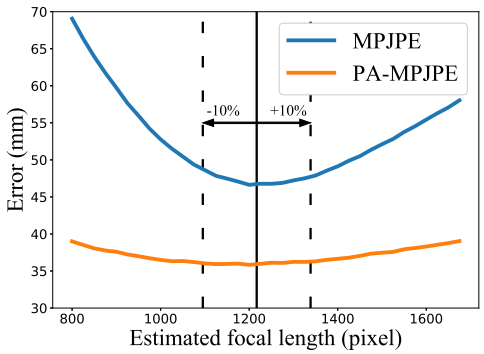


Figure 6: **Sensitivity of pose error to focal length estimation.** The vertical black line indicates the ground-truth focal length.

mirror-induced constraints. It can be observed in Table 1 that without the mirror normal constraint ('Ours (w/o L_n)'), our method can still perform well, indicating our applicability in the case where vanishing points are hard to acquire. If L_s is also discarded ('Ours (w/o L_s, L_n)'), MPJPE will degrade severely while PA-MPJPE has a relatively slight change, reflecting that the mirror symmetry constraint can adjust the human orientation effectively. 'Ours (w/o L_s, L_n)' is better than the baseline since $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are shared. For MRPE, more constraints will bring the better improvement and the position error of our full model is less than 10cm.

We also measure the accuracy of mirror normal estimation. As we have stated, the ground-truth mirror normal comes from calibration. We use the average angle between the ground-truth and the estimated normal. The angle is 1.9° with the ground-truth focal length, and 4.1° with the estimated focal length, both of which are quite small.

Focal length estimation: To evaluate the accuracy of focal length estimation, we simulate the case in which only the focal length is unknown and two orthogonal vanishing points are used. One vanishing point is computed by the lines parallel to the mirror and the other is from two ways: 1) if we label two lines that are perpendicular to the mirror from the scene, the error of focal length is 4.9% of the true value, and 2) if we make use of 2D human keypoints, the error reduces to 3.6%. It means that the 2D human keypoints provide reliable correspondences that are helpful to estimate the focal length.

To evaluate the influence of the predicted focal length on pose estimation, we modify the focal length to different values and calculate the reconstruction accuracy. As shown in Fig. 6, when the focal length deviates from the ground-truth by less than 10%, the change of the reconstruction error is less than 5%, showing the stability of our algorithm.

Figure 7: **Samples in our Mirrored-Human dataset.** From left to right, each column shows the variety of actions, appearances, viewpoints, poses and multiple mirrors/people, respectively.

Datasets	Frames(k)	Subjects	Annotations		
			3D	2D	Internet
InstaVariety [23]	2100	28272			✓
Penn Action [65]	77	2326		✓	✓
Human3.6M [19]	581	11	✓	✓	
3DPW [56]	51	18	✓	✓	
Mirrored-Human	1800	200+	✓*	✓	✓

Table 2: **Comparison of relevant datasets** in terms of the number of frames, subjects and annotation types. *Our Mirrored-Human dataset adopts our reconstructed 3D poses as pseudo ground-truth.

4.3. Qualitative comparison

Some representative results are visualized in Fig. 8, showing that our method can reconstruct accurate 3D human meshes as well as the mirror geometry. Compared with the monocular human mesh recovery algorithms [42, 25], our method produces much more consistent positions and orientations (Fig. 8) and more accurate poses (Fig. 9). More qualitative results can be found in the supplementary material.

5. Learning with the Mirrored-Human dataset

5.1. Mirrored-Human

Based on our framework, a large-scale Internet dataset can be built for the training of single-view tasks. The existing datasets [19, 33] lack the variety of both appear-

Methods	3DPW		Human3.6M	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
HMR [22]	-	81.3	88.0	56.8
HMMR [23]	-	72.6	-	56.9
Arnab. [3]	-	72.2	77.8	54.3
CMR [26]	-	70.2	-	50.1
SPIN [25]	98.2*	59.2	62.3*	41.1
MeshNet [37]	93.2	58.6	55.7	41.7
Baseline	90.0	57.5	54.7	41.7
[37]+MiHu	**85.1**	**54.8**	**53.6**	**41.0**

Table 3: Results on 3DPW and Human3.6M datasets. 'MiHu' is our Mirrored-Human dataset. 'Baseline' means training MeshNet [37] on Mirrored-Human with 3D annotations given by SMPLify-X [42]. *MPJPE of [25] is obtained by their released model.

ances and poses, making the training easy to overfit. For multi-person tasks, collecting data is more difficult . Therefore, previous methods exploit MuCo [35], a pseudo multi-person dataset composited from MPI-INF-3DHP [33] by masks, or JTA [13], a synthetic dataset. The gap between these datasets and the real scene may limit the performance of learning-based methods.

To alleviate the training data issue, we provide a large-scale Internet dataset named Mirrored-Human with our framework. Specifically, we collect a large number of videos from the Internet, in which we can see the person and the person's mirror image. Actions cover dancing, fitness, mirror installation, swing practice, etc. Fig. 7 demonstrates both the appearance and pose diversity of our dataset. Table 2 shows a thorough comparison between our dataset and relevant datasets. Please refer to our supplementary material for more details of our dataset.

5.2. Single-person mesh recovery

For this task, we choose MeshNet [37], a state-of-the-art method for single-view 3D pose estimation, for evaluation. Two datasets are used for evaluation. Human3.6M [19] is an indoor benchmark with 3D annotations. 3DPW [56] is an outdoor benchmark to test the generalization ability and only its defined test set is used. Following standard protocols, we report both MPJPE and PA-MPJPE. We also test the baseline method that uses the state-of-the-art optimization-based method SMPLify-X [42] to generate pseudo ground-truth to train the same network. Table 3 shows that with our dataset, the performance of MeshNet can be improved significantly, especially when tested on the 3DPW dataset without using training data from 3DPW. We also outperform the baseline, indicating that our framework is more accurate than [42].

5.3. Multi-person 3D pose estimation.

For this task, previous methods fall into two categories. Top-down methods detect human first and then estimate keypoints with a single-person pose estimator. Bottom-up methods localize all keypoints in the image first and then group them into people. We choose the top-down method [36] and the bottom-up method [66] for evaluation.

	Methods	AP^{25}_{root} ↑	PCK_{rel} ↑	PCK_{abs} ↑
TD	LCRNet [46]	-	53.8	-
	LCRNet++ [47]	-	70.6	-
	Dabral. [10]	-	71.3	-
	PandaNet [4]	-	72.0	-
	HMOR [29]	-	82.0	**43.8**
	Moon. [36]	31.0	81.8	31.5
	Moon. [36]+MiHu	**42.2**	**82.3**	43.0
BU	Mehta. [35]	-	65.0	-
	Xnect [34]	-	70.4	-
	SMAP [66]	37.3	73.5	35.4
	SMAP [66]+MiHu	**42.3**	**74.1**	**38.0**

Table 4: Results on the MuPoTS-3D dataset. The numbers are calculated for all people. 'MiHu' is our Mirrored-Human dataset. 'TD' and 'BU' mean 'top-down' and 'bottom-up', respectively.

The MuPoTS-3D [35] dataset is used. Following previous methods [36, 66], AP^{25}_{root}, PCK_{rel} and PCK_{abs} are measured. AP^{25}_{root} is the average precision of 3D human root location, which treats the prediction as correct if it lies within 25cm from the ground-truth. PCK_{rel} is the percentage of correct keypoints after root alignment. A keypoint is correct if the distance between the prediction and the ground-truth is smaller than 15cm. PCK_{abs} has almost the same definition as PCK_{rel}, but without the root alignment it measures the absolute pose accuracy. Note that AP is calculated only for the root and PCK is for all keypoints.

It can be observed from Table 4 that with our dataset, AP^{25}_{root} and PCK_{abs} are improved significantly compared with [36]. For bottom-up methods, we also improve the performance of [66] apparently.

6. Conclusion

In this paper, we present an optimization-based framework that leverages the mirror reflection to reconstruct accurate 3D human pose. We collect a large-scale Internet dataset named Mirrored-Human with our reconstructed 3D poses as pseudo ground-truth and show that training on this dataset can enhance the performance of existing 3D human pose estimators. Our work opens many new directions for future research. We plan to extend the method to handle multiple mirrors, multiple people, multiple frames and more detailed reconstruction of shape and appearance.

Acknowledgement: The authors would like to acknowledge the support from the National Key Research and Development Program of China (No. 2020AAA0108901) and NSFC (No. 61806176).

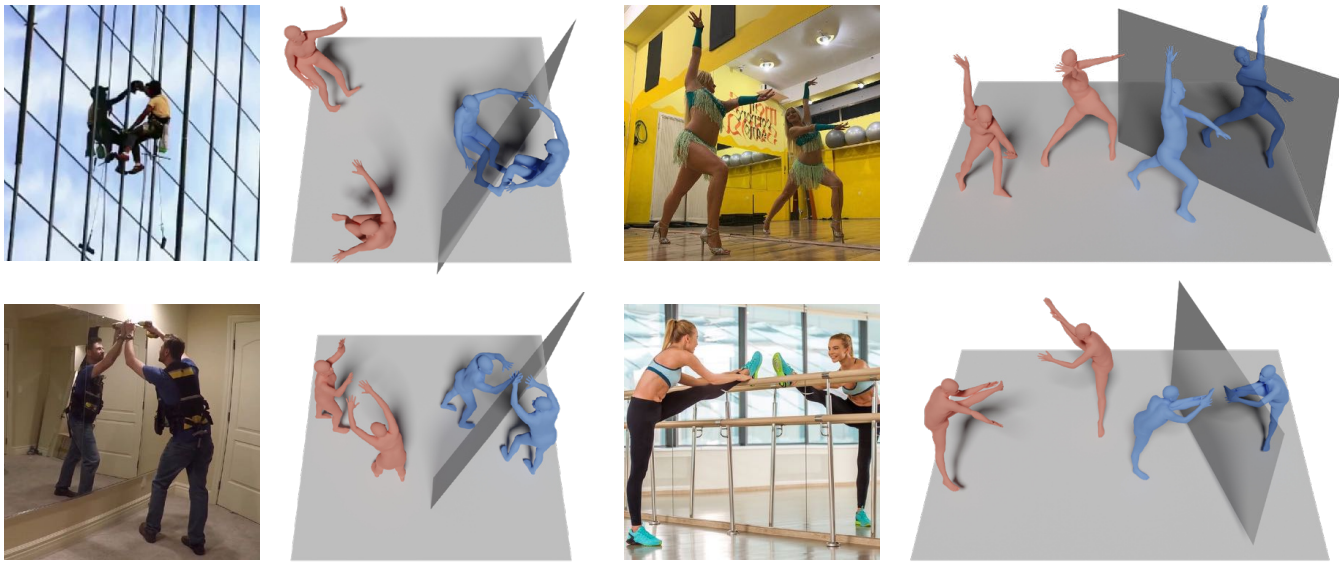


Figure 8: **Reconstruction results of Internet images.** Given the same 2D keypoints, we compare our method (blue) with SMPLify-X [42] (red). Our method simultaneously reconstructs precise poses, global locations and mirror geometry, while SMPLify-X [42] produces inconsistent results due to the depth ambiguity.

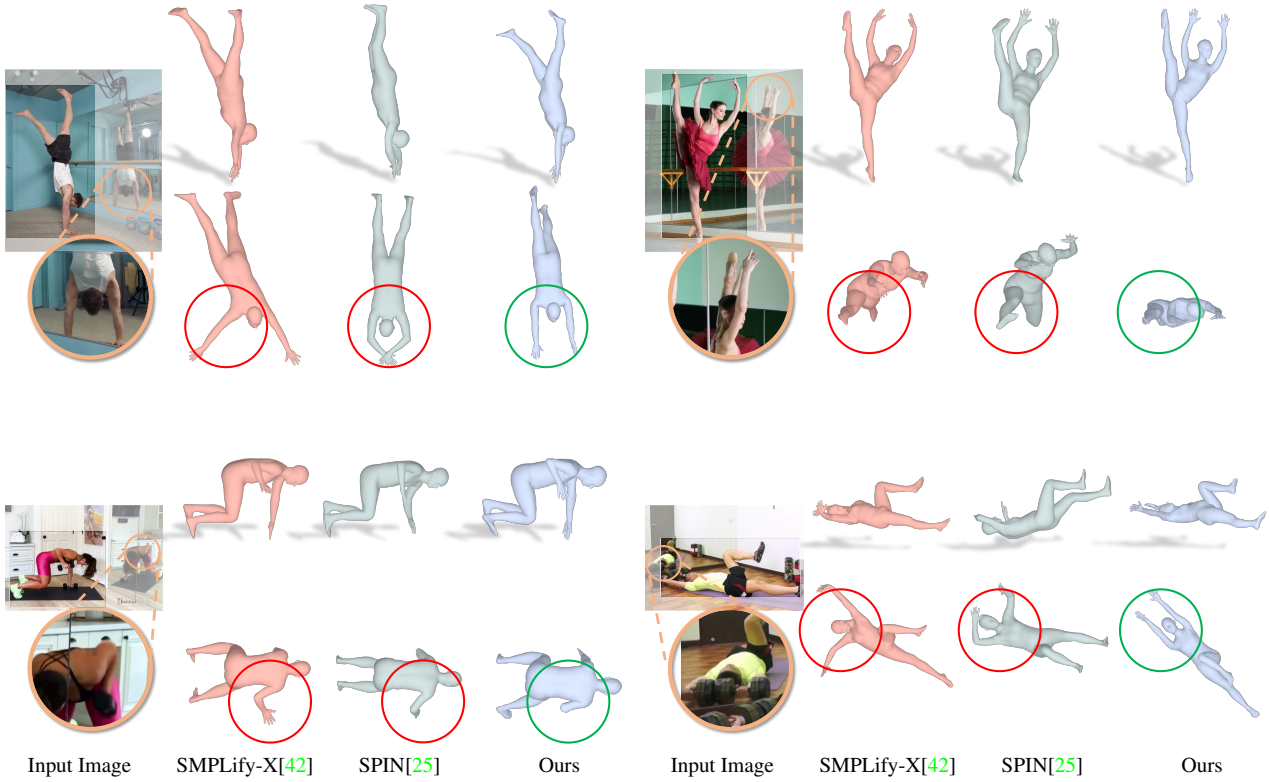


Figure 9: **Qualitative comparison of 3D pose estimation.** For each image, we show the results of the person estimated by the optimization-based method SMPLify-X [42] (red), the CNN-based method SPIN [25] (green) and our method (blue). For each block, top row shows the predicted mesh from the camera view and the bottom row shows another view. The circles emphasize some representative differences among three methods. The single-view optimization-based method produces inaccurate 3D poses despite smaller reprojection error. Our method produces more accurate 3D outputs given the same input.

References

[1] A. Akay and Y. S. Akgul. 3d reconstruction with mirrors and rgb-d cameras. In *VISAPP*, 2014. 2

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *SIGGRAPH*. 2005. 2

[3] Anurag* Arnab, Carl* Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019. 2, 7

[4] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *CVPR*, 2020. 7

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 5

[7] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017. 2

[8] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, 2019. 2

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 2

[10] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *3DV*, 2019. 7

[11] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *ECCV*, 2020. 2

[12] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *CVPR*, 2020. 2

[13] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *ECCV*, 2018. 7

[14] M. Fieraru, M. Zanfir, E. Oneata, A. I. Popa, V. Olaru, and C. Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020. 2

[15] Joshua Gluckman and Shree K Nayar. Catadioptric stereo using planar mirrors. *IJCV*, 2001. 2

[16] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4, 5

[17] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2019. 2

[18] B. Hu, C. M. Brown, and R. Nelson. Multiple-view 3-d reconstruction using a mirror. 2005. 2, 4

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2014. 1, 2, 6, 7

[20] Nianjuan Jiang, Ping Tan, and Loong-Fah Cheong. Symmetric architecture modeling with a single image. In *ACM SIGGRAPH Asia*. 2009. 2

[21] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE TPAMI*, 2017. 2

[22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 7

[23] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2, 6, 7

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2

[25] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1, 2, 5, 6, 7, 8

[26] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2, 7

[27] Douglas Lanman, Daniel Crispell, and Gabriel Taubin. Surround structured lighting: 3-d scanning with orthographic illumination. *CVIU*, 2009. 2

[28] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 2

[29] Jiefeng Li, Can Wang, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, 2020. 2, 7

[30] Jiahao Lin and Gim Hee Lee. Hdnet: Human depth estimation for multi-person camera-space localization. In *ECCV*, 2020. 2

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 2, 3

[32] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2

[33] 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. 1, 2, 6, 7

[34] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and

Christian Theobalt. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *ACM TOG*, 2020. 2, 7

[35] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 7

[36] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2019. 2, 7

[37] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 1, 2, 7

[38] S. A. Nene and S. K. Nayar. Stereo with mirrors. In *ICCV*, 1998. 2

[39] Trong-Nguyen Nguyen, Huu-Hung Huynh, and Jean Meunier. 3d reconstruction with time-of-flight depth camera and multiple mirrors. *IEEE Access*, 2018. 2

[40] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *3DV*, 2018. 2

[41] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A spare trained articulated human body regressor. In *ECCV*, 2020. 2

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 2, 5, 6, 7, 8

[43] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *CVPR*, 2018. 2

[44] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2

[45] Rui Rodrigues, Joao P Barreto, and Urbano Nunes. Camera pose estimation using images of planar mirror reflections. In *ECCV*, 2010. 2

[46] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 7

[47] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE TPAMI*, 2019. 2, 7

[48] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010. 2

[49] Sudipta N Sinha, Krishnan Ramnath, and Richard Szeliski. Detecting and reconstructing 3d mirror symmetric objects. In *ECCV*, 2012. 2

[50] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 5

[51] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 2

[52] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 2

[53] Tomu Tahara, Ryo Kawahara, Shohei Nobuhara, and Takashi Matsuyama. Interference-free epipole-centered structured light pattern for mirror-based multi-view active stereo. In *3DV*, 2015. 2

[54] K. Takahashi, S. Nobuhara, and T. Matsuyama. A new mirror-based extrinsic camera calibration using an orthogonality constraint. In *CVPR*, 2012. 2

[55] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In *3DV*, 2020. 2

[56] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 1, 2, 6, 7

[57] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *CVPR*, 2019. 2

[58] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 2

[59] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 1

[60] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020. 2

[61] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 2

[62] Xianghua Ying, Kun Peng, Yongbo Hou, Sheng Guan, Jing Kong, and Hongbin Zha. Self-calibration of catadioptric camera with two planar mirrors from silhouettes. *IEEE TPAMI*, 2012. 2

[63] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 2

[64] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *CVPR*, 2020. 2

[65] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 6

[66] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, 2020. 2, 7

[67] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 2