Human Pose and Shape Estimation From Single Polarization Images

Shihao Zou †, Member, IEEE, Xinxin Zuo, Sen Wang, Yiming Qian, Chuan Guo ‡, and Li Cheng †, Senior Member, IEEE

Abstract—This paper focuses on a new problem of estimating human pose and shape from single polarization images. Polarization camera is known to be able to capture the polarization of reflected lights that preserves rich geometric cues of an object surface. Inspired by the recent applications in surface normal reconstruction from polarization images, in this paper, we attempt to estimate human pose and shape from single polarization images by leveraging the polarization-induced geometric cues. A dedicated two-stage pipeline is proposed: given a single polarization image, stage one (Polar2Normal) focuses on the fine detailed human body surface normal estimation; stage two (Polar2Shape) then reconstructs clothed human shape from the polarization image and the estimated surface normal. To empirically validate our approach, a dedicated dataset (PHSPD) is constructed, consisting of over 500 K frames with accurate pose and parametric shape annotations. Empirical evaluations on this real-world dataset as well as a synthetic dataset, SURREAL, demonstrate the effectiveness of our approach. It suggests polarization camera as a promising alternative to the more conventional RGB camera for human pose and shape estimation.

Index Terms—Human pose and shape estimation, human shape reconstruction, shape from polarization.

I. INTRODUCTION

A CRITICAL computer vision problem is to predict 3D human poses, i.e., 3D body joint locations, from single images. In recent years, rapid progress is made in 3D human pose prediction from RGB images [1]–[10]. Moreover, fueled by the development in parametric human body shape modelling, such as SCAPE [11] and SMPL [12], it becomes feasible to estimate human body shapes from single RGB image, as is evidenced by a number of end-to-end deep learning methods [13]–[25]. On the other hand, the problems of 3D human pose and shape estimation from single RGB images are still far from being solved. This is mainly due to the inherently lack of 3D cues in an RGB image. Furthermore, as the human shape models (e.g., SMPL) are usually learned from large sets of scanned human naked bodies, they are often lacking in clothing details.

The above observation inspires us to investigate in this paper a new imaging modality, polarization images. That is, we consider the problem of estimating human pose and shape from a single polarization image. Polarization camera is realised upon a basic physics principle: a light ray reflected from an object is usually polarized. The polarized signal thus carries sufficient geometric cues of object surface details to reliably infer its surface normal [26], [27]. It is worth mentioning the biological fact that light polarization could be directly perceived by some species of bees, ants, and shrimps for purposes such as 3D navigation [28], [29]. Motivated by the physical and biological facts, we propose a dedicated two-stage approach for human pose and shape estimation by integrating the geometric cues from the input polarization images.

As shown in Fig. 1, our approach, also called HumanSfP, contains two main stages. Stage one, Polar2Normal, concentrates on predicting accurate surface normal maps from single polarization images by exploiting the associated physics laws as priors. It is then fed into stage two, Polar2Shape, in reconstructing the initial shape is parameterized by SMPL model, estimated from the raw polarization image and the surface normal.
clothed human shape guided by the obtained surface normal and an initial shape estimated by SMPL.

Different from previous efforts in estimating detailed depth map [25], [30], our stage one focuses on surface normal estimation from a polarization image. By explicitly integrating the underlying physical principles, it gives rise to more reliable estimation. To achieve this goal, two main challenges need to be addressed, namely $\pi$-ambiguity of the azimuth angle and the possible environmental noise in practical applications. Based on the physical laws of polarization, two ambiguous normal maps $m_1$ and $m_2$ (Section VIII-A) can be directly calculated from a polarization image. It is often reasonable to assume light reflected by human clothes is dominated by diffused reflection. Then, different from [27], we consider a two-branch strategy: one branch is employed to categorize each pixel into three categories: the two ambiguous normal maps and background; the second branch is to infer a coarse surface normal map. A fused normal map is produced by incorporating the classification results as well as the two ambiguous normal maps. Finally, as is revealed by [31], the fused normal map output may still be noisy, owing to environmental noise and the digital quantization of the polarization camera. We instead work with the normal residual calculated as the difference between the coarse and the fused normal map, which is used to refine the coarse normal map, and to produce our final normal estimation.

Based on the raw polarization image and final surface normal output of stage one, stage two concerns the estimation of 3D human pose and the reconstruction of its clothed shape. It starts from estimating an initial parametric shape, i.e. SMPL shape, as 3D pose, which is then deformed by leveraging the geometric details from the surface normal, to reconstruct the final clothed human shape. Different from previous works [20], [22], [31], geodesic distance is employed in SMPL shape estimation, since the pose representation of SMPL is naturally in a product space of SO(3), a classical example of Lie group. Our two-stage pipeline is empirically shown to be capable of inferring detailed surface normal faithfully, and estimating human pose and clothed body shape accurately.

To summarize, there are three main contributions in our work. (i) A new problem, namely human pose and shape estimation from single polarization images, is proposed. A dedicated deep learning approach, HumanSIP, is proposed, where the detail-preserving surface normal maps are obtained following the physical laws of light polarization, and are shown to estimate more accurate pose and body shape. (ii) In tackling this new problem, a dedicated Polarization Human Shape and Pose Dataset, PHSPD, has been created. It now consists of ~527 K frames and their corresponding pose and shape annotations. Overall there are 21 different subjects performing 31 unique actions, and ~9.5 hours of videos are recorded in total. The dataset and our code are publicly available. Improved performance over state-of-the-art [27] is demonstrated on the related benchmarks. Further, our proposed model for normal estimation also achieves the SOTA results for the task of Shape from Polarization (SfP). Together they demonstrate the effectiveness of our approach.

III. RELATED WORK

III. SHAPE FROM POLARIZATION (SfP)

It is a related task that focuses on inferring shape (normally represented as surface normal) from a polarization image, where each channel captures the polarimetric information of reflected light under a linear polarizer with a different angle. There are two main issues involved in SfP: angle ambiguity and the discrimination of specular and diffuse reflection. Previous efforts are mainly physics-based that rely on additional information or assumptions to elucidate the possible ambiguities, such as smooth object surfaces [34], coarse depth map [26], [35] and multi-view geometric constraint [36], [37]. The first deep learning-based method is conceived in [27] that integrates physical priors (ambiguous normal maps) with deep models in estimating normal maps. It has shown that deep model can learn to leverage the angle ambiguity and environmental noise. A following-up work [31] advocates to first classify each pixel into different types of ambiguous angle and obtain a fused normal map, which is shown to extract more explicit geometric cues from polarization images.
IV. HUMAN POSE ESTIMATION FROM SINGLE IMAGES

In the past few years, 3D human pose estimation from single images, mainly based on RGB or depth images, has been extensively studied. Many early efforts [38]–[40] utilize dictionary-based learning strategies to capture prior knowledge from large motion-capture dataset. Recent efforts focus on end-to-end deep learning based methods, including CNNs [4], [41] and Graph CNNs [42], [43] to estimate 3D human poses. In particular, a common framework has been adopted by a number of recent works [5], [9], [10], [44]–[49], which first infer 2D poses (either 2D joint positions or heatmaps), which are then lifted to 3D. Self-supervised learning [9], [10] and adversarial learning [47], [49] are also considered to exploit the benefits of additional re-projection or adversarial constraints.

V. HUMAN SHAPE ESTIMATION FROM SINGLE IMAGES

Going beyond pose estimation, the availability of parametric human shape models such as SMPL model [12], has fueled growing attentions in single image-based human shape estimation. SMPL is a statistical low-dimensional representation of human shape, realized by principal component analysis of empirical body shapes of naked and minimally dressed humans. Early efforts emphasize on optimization-based methods to fit SMPL model to point cloud or annotated pose & human silhouette [16]–[18], [50]. Recent deep learning based methods [20], [51], [52] instead learn to predict SMPL parameters under various constraints such as 2D/3D pose, silhouette, and adversarial examples. Human body pixel-to-surface correspondence map is considered in [53] for parametric shape estimation. In [54], optimization and regression are integrated to form a self-improvement loop. Temporal information is also investigated in [23] that estimates human shape in a video. Point cloud is also explored in parametric human shape estimation [55]. Rather than RGB images, our work works with single polarization images. Moreover, our aim is to estimate human shapes with clothing details, which goes beyond the capability of parametric SMPL human shapes for naked and minimally dressed humans.

In terms of human shape reconstruction, volume-based methods [24], [56]–[59] are popular in reconstructing detailed body shapes. They unfortunately suffer from the limitation of computation scalability and lack of reliable 3D cues. The most recent work, PaMIR [24], combines the parametric body model with the free-form deep implicit function to reconstruct human shape. Unlike surface normal, the implicit field cannot provide explicit cues of human body and leads to in-accurate reconstructed shapes. Saito et al. [58], [60] introduce a pixel-aligned implicit surface function to encode detailed body surface. However, these two works are not able to handle complex poses, partly owing to the lack of complex poses in their training set. A closely related work is [25] that considers a hierarchical framework to incorporate robust parametric shape estimation and flexible 3D shape deformation. It however employs a network trained on additional small dataset to infer shading information, which are inherently unreliable given the lack of ground-truth information of surface normal, albedo and environmental lighting. On the contrary, our method infers clothed shape with the reliable normal map estimated from the polarization image. Another related work is [30] that iteratively integrates rough depth map and estimated surface normal for improved clothes details.

VI. HUMAN POSE AND SHAPE DATASET

A number of human pose and shape datasets have been released in recent years, including MPII [39], MS COCO [61] and PoseTrack [62], which provide in-the-wild RGB images and 2D pose annotations. Human3.6M [32], MPI-INF-3DHP [33] and 3DPW [63] are three benchmark datasets with 3D annotations for human pose estimation from RGB images. There are also human shape datasets [30], [57], [58] that mainly consist of RGBD images and the corresponding human shape annotations. Our PHSPD dataset is the first such dataset based on polarization images.

VII. PRELIMINARY BACKGROUND

A. Polarization Image Formation

The reflected light from a surface mainly includes three components [37], the polarized specular reflection, the polarized diffuse reflection, and the unpolarized diffuse reflection. A polarization camera is equipped with an array of linear polarizers mounted right on top of its CMOS imager, in place of the RGB Bayer filters. During the imaging process, a pixel’s intensity typically varies sinusoidally with the angle of the polarizer [26]. In this work, we assume that the light reflected off human clothes is dominated by polarized diffuse reflection and unpolarized diffuse reflection. For a specific polarizer angle $\phi_{\text{pol}}$, the illumination intensity at a pixel with dominant diffuse reflection is

$$I(\phi_{\text{pol}}) = \frac{I_{\text{max}} + I_{\text{min}}}{2} + \frac{I_{\text{max}} - I_{\text{min}}}{2} \cos(2(\phi_{\text{pol}} - \varphi)).$$

(1)

Here $\varphi$ is the azimuth angle of surface normal, $I_{\text{max}}$ and $I_{\text{min}}$ are the upper and lower bounds of the illumination intensity. $I_{\text{max}}$ and $I_{\text{min}}$ are mainly determined by the unpolarized diffuse reflection, and the sinusoidal variation is mainly determined by the polarized diffuse reflection. If the intensity images $I(\phi_{\text{pol}})$ under three or more polarizer angles can be obtained, such as $I(0^\circ)$, $I(45^\circ)$ and $I(90^\circ)$, $\varphi$ can be solved in closed form. Note that there is $\pi$-ambiguity in the azimuth angle $\varphi$ in Eq. (1), which means that $\varphi$ and $\pi + \varphi$ will result in the same illumination intensity of the pixel. As for the zenith angle $\theta$, it is related to the degree of polarization $\rho$,

$$\rho = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}.$$

(2)

According to [34], when diffuse reflection dominates, the degree of polarization $\rho$ becomes a function of the zenith angle $\theta$ and the refractive index $n$,

$$\rho = \frac{(n - \frac{1}{n})^2 \sin^2 \theta}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2 \theta + 4 \cos \theta \sqrt{n^2 - \sin^2 \theta}}.$$

(3)

In this paper, we assume the refractive index $n = 1.5$ since the material of human clothes is mainly cotton or nylon. The solution of $\theta$ in Eq. (3) is thus a close-form expression of $n$ and $\rho$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
is the target normal map with \( \rightarrow (4) \)

\( \omega \) (3) is followed by body shape refinement from the surface normal estimation from a single polarization image; (2) Polar2Shape: initial human body shape estimation from the raw polarization image and the obtained surface normal, which is followed by body shape refinement from the surface normal. The classification loss is measured by the cross entropy between the pixel-wise probabilities, where Huber loss is employed to train the network, defined as

\[
H(x, \alpha) = \begin{cases} 
0.5 x^2, & |x| < \alpha \\
\alpha (|x| - 0.5 \alpha), & \text{otherwise}
\end{cases}
\]

The loss of the coarse surface normal becomes

\[
\mathcal{L}_{\text{coarse}} = \sum_{i,j} H(1 - \langle \mathbf{m}_3[i,j], \mathbf{m}_i[j] \rangle, \alpha),
\]

where \( \mathbf{m} \) is the target normal map with \((i, j)\) being the pixel coordinate, \( \langle \mathbf{m}_3[i,j], \mathbf{m}_i[j] \rangle \) is the cosine similarity between the two normal vectors, and \( \alpha \) controls the trade-off between the squared and the absolute losses. Now, the second decoder is to classify each pixel into three categories: background, ambiguous normal \( \mathbf{m}_1 \), and ambiguous normal \( \mathbf{m}_2 \), with the corresponding pixel-wise probabilities \( p_0, p_1, \) and \( p_2 \), respectively. The fused normal is thus obtained by

\[
\mathbf{m}_4 = (1 - p_0) \cdot \frac{p_1 \mathbf{m}_1 + p_2 \mathbf{m}_2}{\|p_1 \mathbf{m}_1 + p_2 \mathbf{m}_2\|_2},
\]

where \( 1 - p_0 \) acts as a soft mask for the foreground human body. The classification loss is measured by the cross entropy between

\[
\text{VIII. OUR HUMANSfP APPROACH}
\]

There are two main stages in our approach: (1) Polar2Normal: surface normal estimation from a single polarization image; (2) Polar2Shape: initial human body shape estimation from the raw polarization image and the obtained surface normal, which is followed by body shape refinement from the surface normal.
the predicted pixel-wise category and the target category,
\[
\mathcal{L}_{\text{category}} = \sum_{i,j} \sum_c y_{c[i,j]} \log p_c[i,j] + (1 - y_{c[i,j]} \log(1 - p_c[i,j]).
\]
(11)
Here \(c\) indexes among the three categories. \(y_{c[i,j]} \in \{0, 1\}\) is the multi-class label indicating which category the pixel \([i, j]\) belongs to. Note that the label \(y_{c[i,j]}\) is created by discriminating whether the pixel is background or which ambiguous normal has higher cosine similarity with its target normal.

Now let us look at the second step. Different from our previous work [31] that directly regresses the normal map given the polarization image and fused normal, we propose a residual update scheme to produce a more detailed and accurate surface normal estimation, as displayed in Fig. 2. Due to the environmental noise and the digital quantization of the polarization image formation process, the fused normal map \(m_3\) is often noisy and non-smooth. Given the coarse normal map \(m_3\) and fused normal map \(m_4\), the normal residual \(m_r\) is evaluated by
\[
m_{r[i,j]} = 1 - (m_3[i,j], m_4[i,j]).
\]
(12)
A denoising network is then trained to take both normal maps \(m_3, m_4\) and the normal residual \(m_r\) as input, to produce a smoothed and detailed normal \(\hat{m}\). The residual update scheme takes place as the coarse normal \(m_3\) is replaced by the final predicted normal \(\hat{m}\) to produce the new normal residual \(m_r\), which is then employed to regress an updated normal map \(\hat{m}\). The loss for \(\hat{m}\) is defined by
\[
\mathcal{L}_{\text{final}} = \sum_{i,j} \left| 1 - (\hat{m}[i,j], m[i,j]) \right|.
\]
(13)
Here \(|·|\) denotes the absolute value. Finally, our surface normal estimation model is learned by minimizing the following loss
\[
\mathcal{L}_{\text{polar2normal}} = \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{category}} + \mathcal{L}_{\text{final}}.
\]
(14)

B. Polar2Shape: Human Shape Reconstruction

Stage two of our approach, also referred to as Polar2Shape, focuses on the reconstruction of clothed body shape from a polarization image, which consists of two steps. In the first step, the naked body shape, represented by SMPL model [12], is estimated. The following step is to deform the initial human body shape with the estimated surface normal, and finally the clothed body shape is reconstructed. The process is shown in Fig. 3.

Initial Shape Estimation: The core of SMPL model [12] lies in a differentiable function \(M(\beta, \theta) \in \mathbb{R}^{6,890\times3}\) that outputs a triangular mesh with 6,890 vertices from 82 parameters \([\beta, \theta]\), \(\theta \in \mathbb{R}^{24}\) is the pose parameter vector to characterize pose articulations in axis-angle representation, consisting of one global rotation of the body and the relative rotations of its 23 joints. Human pose is therefore represented as \(\theta = (\theta_m)_{m=1}^{24}\), where \(\theta_m \in \mathbb{R}^3\), and \(m\) denotes the index of relative rotation in axis-angle. The shape parameter vector \(\beta \in \mathbb{R}^{10}\) is the linear coefficients of a PCA shape space that mainly determines individual body features such as height, weight and body proportions. The PCA shape space is learned from a large dataset of naked and minimal clothed human body scans. A specific SMPL shape is produced by first applying pose-dependent and shape-dependent deformations to the template pose, then using forward-kinematics to articulate the body to its current pose, and deforming the surface mesh by linear blend skinning. At the same time, the 3D joint positions, denoted by \(J_{3D} \in \mathbb{R}^{24\times3}\), are obtained by linear regression from the output mesh vertices. In addition to the SMPL model parameters, the global translation of the human body is also a necessary factor in aligning with the projection in the 2D image space, denoted by \(t \in \mathbb{R}^3\). This results in an 85-dimensional parameter space, \(\Theta = [\beta, \theta, t]\), that are used by SMPL to dictate a specific human shape.

Under mild assumption, the axis-angle representation of human pose in SMPL, \((\theta_m)_{m=1}^{24}\), has a bijective map to the corresponding 24-dim product space of \(\mathfrak{so}(3)\) manifold. Existing efforts such as [31] normally predict the pose in axis-angle representation, and measure the Euclidean distance between the predicted pose and target pose as \(\sum_{m=1}^{24} \|\theta_m - \hat{\theta}_m\|_2\). However, as the two poses corresponds to two points in the aforementioned curved space, their distance is better characterized by their geodesic distance, which is not necessarily the Euclidean distance. Thus we propose to represent the human pose as a set of rotation matrices \({R_m}_{m=1}^{24}\) in \(\mathbb{SO}(3)\) with \(R_m = \exp(\theta_m)\). The geodesic distance between the predicted pose and target pose becomes
\[
\mathcal{L}_{\text{pose}} = \sum_{m=1}^{24} D(R_m, \hat{R}_m) = \sum_{m=1}^{24} \cos^{-1}\left(\frac{\text{Tr}(R_m^T \hat{R}_m) - 1}{2}\right),
\]
(15)
which is also referred as the pose estimation loss. Moreover, the losses for SMPL shape, global translation and joint positions are defined as
\[
\mathcal{L}_{\text{shape}} = \|\beta - \hat{\beta}\|_2^2,
\]
(16)
\[ \mathcal{L}_{\text{trans}} = \| t - \hat{t} \|^2, \tag{17} \]
\[ \mathcal{L}_{\text{joints}} = \| \mathbf{J}_{3D} - \hat{\mathbf{J}}_{3D} \|^2. \tag{18} \]

At the moment, our initial shape estimation is learned by minimizing the following loss

\[ \mathcal{L}_{\text{polar2shape}} = \mathcal{L}_{\text{pose}} + \lambda_{\text{shape}} \mathcal{L}_{\text{shape}} + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}} + \lambda_{\text{joints}} \mathcal{L}_{\text{joints}}, \tag{19} \]

where \( \lambda_{\text{shape}}, \lambda_{\text{trans}} \) and \( \lambda_{\text{joints}} \) are the tuning parameters of the corresponding loss terms.

**Shape Reconstruction:** The initial human shape obtained by SMPL representation still lacks fine surface details. Therefore, the aim of this step is to refine the initial SMPL shape guided by our surface normal estimate, as follows. The SMPL body shape is rendered on the image plane to form an initial depth map. The technique of [65] is then engaged here to obtain an optimized depth map \( I_d \) from the predicted surface normal \( \hat{n} \) and the initial depth \( I_d \) by minimizing the objective function,

\[ E(I_d) = \lambda_n E_n(I_d) + \lambda_d E_d(I_d) + \lambda_s E_s(I_d), \tag{20} \]

which contains three energy terms. The first term, \( E_n(I_d) \), ensures the predicted normal to be perpendicular to the tangents of the optimized depth surface. The second term, \( E_d(I_d) \), encourages the optimized depth to be close to the initial depth. The third and final term preserves smoothness of nearby pixels over the optimized depth map. More details can be found in the supplementary material.

### C. Our In-House PHSPD Dataset

To facilitate empirical evaluation of our approach in real-world scenarios, a home-grown dataset is curated, which is referred as Polarization Human Pose and Shape Dataset or PHSPD. In what follows, we start by presenting the early version, PHSPDv1, as well as the more recent addition, PHSPDv2.

**PHSPDv1:** It is the early version used in our preliminary work [31]. During PHSPDv1 data acquisition, a system of 4 soft-synchronized cameras are used, consisting of a polarization camera and three RGBD cameras. 12 subjects are recruited in data collection, in which 9 are male and 3 are female. Each subject performs 3 different sets of actions (out of 18 distinct action types) 4 times, plus an additional period of free-form motion at the end of each session. Thus for each subject, there are 12 short videos (around 1,800 frames per video with 10-15 FPS). The total number of frames for each subject amounts to 22 K. Overall, PHSPDv1 dataset consists of 287 K frames, each frame here contains a synchronized set of images: one polarization image, three RGBD images.

**PHSPDv2:** It contains the newly acquired data, where our multi-camera acquisition system is extended to 11 cameras of different modality: one polarization camera and five RGBD cameras. 15 subjects are recruited for the data acquisition, where 11 are male and 4 are female. Each subject is required to perform 3 groups of actions (21 different actions in total) for 4 times, where each group includes actions of fast/medium/slow speed, respectively. Finally, 12 videos are collected for each subject; each video has around 1,300 frames of 15 FPS. In total, there are 180 videos, with each video lasting about 1.5 minutes. This amounts to 240 k frames with each frame including a polarization image and five RGBD images.

To summarize, our PHSPD dataset contains 334 videos of 21 different subjects performing 31 types of actions, amounts to around 527 K frames and 9.5 hours of recorded videos. Each frame contains a synchronized set of images including both a polarization image and RGBD images. Further details regarding our PHSPD dataset can be found in the supplementary material.

**Annotation:** The human full-body 3D pose and shape annotations, RGB images of higher resolution, more depth images, it is also has a limited amount of annotated images. In what follows, we start by presenting the early version, PHSPDv1, as well as the more recent addition, PHSPDv2.

**Compared with Existing Datasets:** Our PHSPD dataset is compared side-by-side with six widely-used human pose and shape datasets in Table I. The datasets of MPI [39], MS COCO [61] and PoseTrack [62] provide only RGB images with 2D pose annotations, where depth images and shape annotations are not considered; MPI-INF-3DHP [33] contains 1.3 M in-house frames with 3D pose annotations, but without depth channel images. 3DPW [63] has 51 K frames with extreme pose and 3D shape annotations, meanwhile it does not possess depth images, it also has a limited amount of annotated images.


to Human3.6M, our PHSPD dataset comes with 3D shape annotations, RGB images of higher resolution, more depth images and with higher resolution, more subjects, and more action types. More importantly, PHSPD is the only dataset that comes with polarization images.

### IX. Experiments

Our proposed approach is empirically examined in two major aspects, namely surface normal estimation, and pose and shape estimation from a polarization image. Section IX-A focuses on the empirical evaluations of surface normal estimation on the

| Dataset   | Sub | Act | MM | RGB | Depth | P | S |
|-----------|-----|-----|----|-----|-------|---|---|
| MS COCO   | -   | -   | X  | 330K|       | 2D|   |
| MPI [39]  | -   | -   | X  | 40K |       | 2D|   |
| PoseTrack [62] | - | - | X | 22K |       | 2D|   |
| MPI-3DHP [33] | 8 | -  | 1.3M| 3D  |       |   |   |
| 3DPW [63] | 7   | 8   | X  | 31K |       | 3D|   |
| Human3.6M [32] | 11 | 15 | 3.6M | 0.45M | 3D |   |   |
| PHSPD (ours) | 21 | 31 | 2.1M | 2.1M | 3D |   |   |
widely used SfP benchmark [27] in shape from polarization, as well as our PHSPD dataset. Section IX-B and IX-C evaluate 3D pose estimation on the synthetic SURREAL dataset [51] and the PHSPD dataset, and shape estimation on our PHSPD dataset. Ablation study is presented in Section IX-D to analyze the effect of individual components in our approach.

Evaluation Metrics: For surface normal estimation, we report mean angle error (MAE), which measures the angle error between the target and estimated normal map. For human pose and shape estimation, we report the mean per joint position error (MPJPE), pelvis-aligned MPJPE (PEL-MPJPE) where two pelvis joints are aligned between the predicted and target 3D pose, Procrustes-aligned MPJPE (PA-MPJPE) where the predicted pose is aligned with the target by rigid transformation, and percentage of correct key-points (PCK) where the joint position error less than 100 mm is considered as the correct prediction. As for the evaluation of human shape, 3D point to surface error (P2S) is employed, where the iterative closest point (ICP) alignment between predicted body mesh and the ground-truth human body point cloud is applied; for each vertex of the human body triangle mesh, its closest point in the point cloud is identified to form a pair, and the average distance of the pairs is computed.

Evaluation Datasets: The widely-used SfP dataset [27] is employed to evaluate the performance of our proposed Polar2Normal component. We follow the typical training (236 polarization images with \(1224 \times 1024\) resolution) and testing (27 images) scheme as in [27], where a \(256 \times 256\) patch on a image is randomly cropped for training, and 20 overlapped patches of one image are first cropped for testing and then fused to form the final predicted normal map. We also demonstrate the effectiveness of our approach on SURREAL [51], a synthetic dataset of color images rendered from motion-captured SMPL human shapes. Polarization images can be synthesized using color and depth images (details are in supplementary material). We choose subset “run1” and subsample over time to recruit one from every 20 consecutive frames. Finally, the train set comprises 123,860 samples and test set has 26,650 samples. SURREAL dataset is only used for the evaluation of normal and pose estimation due to the lack of ground truth point cloud of the human body. Our real-world PHSPD dataset is involved in the evaluations of all three tasks of normal estimation, pose estimation and clothed shape estimation. The target surface normal is obtained by fusing multi-view normal maps calculated from multiple depth images. Subject 4, 7, 11 in PHSPDv1 and subject 1, 2, 7 in PHSPDv2 are chosen to form the test set, resulting in 117,860 samples. The training set consists of the rest 400,785 samples.

Implementation Details: For Polar2Normal, each of the encoders and decoders contains 6 sequential blocks, with each block having one up/down-sampling and two convolutional layers. Further details of our Polar2Normal architecture can be found in supplementary material. The Polar2Normal model is trained for 600 epochs on SfP dataset, 30 epochs on SURREAL dataset and 8 epochs on PHSPD dataset with Adam optimizer [69]. The learning rate is 0.001, which decays to 0.0001 after training 200 epochs, 10 epochs, and 6 epochs, respectively. The batch size is 16 for training on three datasets. For Polar2Shape model, ResNet50 [70] is used as the backbone CNN model. The extract 1024-dimensional feature is directly regressed to the final outputs: \(\beta, \theta,\) and \(t.\) The pose estimation model is trained for 30 epochs on SURREAL dataset and 8 epochs on PHSPD dataset with Adam optimizer [69]. The learning rate is 0.001 and decays to 0.0001 after training 10 epochs and 4 epochs respectively. The batch size for training is 32. The trade-off parameter \(\alpha\) is set to 0.5. The tuning parameters of the loss terms are set to \(\lambda_{\text{shape}} = 0.1, \lambda_{\text{trans}} = 0.1\) and \(\lambda_{\text{joints}} = 10,\) respectively. The three weights used in the Polar2Shape stage, namely the normal term \(\lambda_{\text{n}},\) the depth data term \(\lambda_{\text{d}},\) and the smoothness term \(\lambda_{\text{s}}\) are empirically set to 1.0, 0.06, and 0.55, respectively.

A. Evaluation of Surface Normal Estimation

In this task, our approach is compared with three baselines: a conventional method HeightfP [71], a most recent work DeepSfP [27] and our preceding work HumanSfP1 [31].

From the quantitative results in Table II and the qualitative results in Fig. 4, our method consistently outperforms the state-of-the-art SfP methods HeightfP [71] and DeepSfP [27] for the task of shape from polarization. The poor performance of HeightfP [71] could be attributed to its noise-free assumption that may not hold in the captured images. Though DeepSfP [27] incorporates the ambiguous normal maps as physical priors, the ambiguous normal maps are directly concatenated with the polarization image to form its input, which may overlook the implicit geometric clues. As a result, it performs less well when comparing to our method, especially in complex scenes such as Christmas, Dragon and Horse, where the results of our method achieve \(\sim 3° - 5°\) improvement. It is also demonstrated in the visual results of Fig. 4. Comparing to our previous work HumanSfP1 [31], as illustrated in Table II, our new method, HumanSfP2, exceeds HumanSfP1 in most scenarios, especially Horse and Dragon, where we have \(\sim 2°\) improvement; Only for the two scenes of Box and Christmas, the results of both HumanSfP1 and HumanSfP2 are almost identical. The observation demonstrates the advantages of our newly proposed 2-step strategy for surface normal estimation.

We also evaluate the normal estimation of human body surface on both SURREAL dataset and our PHSPD dataset. Through both the quantitative results of Table III and the visual results of Fig. 5, it is observed that our method consistently

---

### Table II

| Scene      | HeightfP [71] | DeepSfP [27] | HumanSfP1 [31] | HumanSfP2 (ours) |
|------------|--------------|-------------|----------------|-----------------|
| Dragon     | 49.15        | 21.55       | 18.71          | 16.88           |
| Horse      | 55.87        | 22.27       | 21.27          | 19.64           |
| Christmas  | 39.68        | 13.50       | 8.56           | 8.57            |
| Box        | 31.00        | 23.31       | 21.94          | 22.04           |
| Fianingo   | 36.05        | 20.19       | 20.51          | 20.44           |
| Vase       | 36.88        | 10.32       | 9.20           | 9.18            |
| Whole Set  | 41.44        | 18.32       | 17.22          | 16.73           |

---

TABLE II

**Quantitative Evaluations of Surface Normal Estimation on SfP dataset [27].** MAE is reported in the case of six scenes and the whole set. HumanSfP2 (ours) outperforms two baseline methods [27], [71] by a large margin. It also demonstrates that our new method, HumanSfP2 (ours) outperforms the previous effort, HumanSfP1 [31], especially in the cases of complex shapes such as Horse and Dragon.
outperforms the state-of-the-art surface normal prediction methods, namely HeightSfP [71], DeepSfP [27] and HumanSfP1 [31]. Similar to the results presented in Table II, the conventional method HeightSfP shows least appealing performance, which possibly results from its assumptions of polarization images with high-precision pixel intensity and ideal noise-free environment. This is in contrast to the deep neural network based approaches that are often more robust to environmental noise and polarization images with standard pixel representation. Rather than directly concatenating and feeding the input polarization image and the corresponding ambiguous normal maps into a deep model, HumanSfP1 and HumanSfP2 categorize each pixel into one of the ambiguous normal maps, and obtain a fused normal that incorporates explicit geometric information for normal estimation. This could be the main factor that our methods exceed the state-of-the-art DeepSfP on both SURREAL and PHSPD datasets. Moreover, HumanSfP2 out-performs HumanSfP1 over both datasets, which may be attributed to the 2-step model proposed in HumanSfP2, where the first step focuses more on the coarse normal map, and the second step pays more attention to the fine normal details. Qualitative results in Fig. 5 also demonstrate the superior results of the proposed HumanSfP2.

### B. Evaluation of Pose Estimation

This section concerns the qualitative and quantitative evaluation of the estimated SMPL poses. Our comparison methods consist of HMR [20] and SPIN [54]. Since HMR and SPIN is trained on single RGB images, for PHSPD dataset, images from an RGB camera having similar angle of view with the polarization camera are used as the input for evaluation. In addition to HMR and SPIN, for fair comparison, HMR (polar) is included as another baseline, where HMR model is trained from scratch on the polarization images of either SURREAL dataset or our PHSPD dataset. HumanSfP1 [31] is also employed as another baseline that uses Euclidean distance to measure the distance between the predicted and target poses.

From Table IV, it is observed that our HumanSfP2 method produces the lowest errors in MPJPE, PEL-MPJPE and PA-MPJPE, and the highest PCK score among all competing methods. Comparing to HMR and SPIN that takes RGB images as the input source of data, and HMR (polar) that takes polarization images as the input source, our HumanSfP2 out-performs them consistently by a large margin on both SURREAL and PHSPD datasets. Such superior performance lies in two important factors. The first is the engagement of normal map as part of the input, which will be explained in Section IX-D in detail. It is of interest to point out that normal map as part of the input data source is capable of reducing the average joint error by about 10 ~ 20 mm in MPJPE, PEL-MPJPE and PA-MPJPE, while the three competing methods do not include the normal clues in pose estimation. The second factor is the introduction of geodesic loss to measure the predicted and target SMPL poses. Comparing to HumanSfP1, our approach improves the joint error by about 5 mm and PCK by 0.05 in PHSPD dataset, and 15 mm and 0.13 in SURREAL dataset. The quantitative results illustrate the effectiveness of leveraging geodesic distance in pose estimation. Visual results in Fig. 6 also demonstrate the effectiveness of our HumanSfP2 approach, where HumanSfP1 may predict unnatural poses as the Euclidean distance is not suitable in measuring joint displacements due to relative joint rotations.

### C. Evaluation of Shape Estimation

Considering there is no previous work in this new task of single polarization image based clothed human shape reconstruction, four RGB-based methods are recruited for comparison. They are DepthHuman [30], HMD [25], PIFu [58], PIFuHD [60] and PaMiR [24], where RGB image having closest view with polarization image is used as the input for evaluation. The estimated clothed shape is then compared with human point cloud to calculate 3D point to surface error (P2S).

Quantitative results are displayed in Table V. Here, DepthHuman performs the worst, which may be partly attribute to its consideration of the surface depth instead of the entire body shape. PIFu [58] and PIFuHD [60] show similar results, with PIFuHD having slightly larger error. The explanation is PIFu requires human mask as a prior, while PIFuHD do not have such assumption. However, both methods do not take human pose
Fig. 5. Exemplar results of normal map prediction on PHSPD dataset from four comparison methods: HeightSfP [71], DeepSfP [27], HumanSfP1 [31], and HumanSfP2 (ours).

### TABLE IV
**QUANTITATIVE EVALUATIONS OF HUMAN POSE ESTIMATION ON BOTH SURREAL AND PHSPD DATASETS. THE UNIT OF JOINT ERROR IS IN MILLIMETER.**

| Method          | Input | SURREAL          | PHSPD          |
|-----------------|-------|------------------|----------------|
|                 |       | PEL-MPIPE        | PA-MPIPE | PCK ↑ | PEL-MPIPE        | PA-MPIPE | PCK ↑ |
| HMR [20]        | C     | 133.95           | 100.66      | 0.53  | -                 | 106.27   | 68.58 | 0.37  |
| SPIN [54]       | C     | 95.08            | 80.67       | 0.58  | -                 | 91.92    | 49.79 | 0.64  |
| HMD (polar)     | P     | 113.82           | 86.85       | 0.55  | 113.74            | 88.24    | 59.05 | 0.68  |
| HumanSfP1 [31]  | P     | 84.78            | 60.82       | 0.72  | 66.97             | 63.12    | 42.08 | 0.83  |
| HumanSfP2 (ours)| P     | 59.17            | 46.58       | 0.85  | 62.04             | 54.98    | 36.88 | 0.88  |

### TABLE V
**QUANTITATIVE EVALUATIONS OF CLOTHED HUMAN SHAPE ESTIMATION.**

| Method          | Input | SURREAL          |
|-----------------|-------|------------------|
| DepthHuman [30] | C     | 83.86            |
| PIFuHD [60]     | C     | 67.05            |
| PIFu [58]       | C     | 62.13            |
| HMD [25]        | C     | 43.72            |
| PaMIR [24]      | C     | 47.24            |
| HumanSfP2 (initial) | P | 39.79          |
| HumanSfP2 (ours) | P     | 38.24            |

### TABLE VI
**ABLATION STUDY OF OUR NORMAL ESTIMATION COMPONENT ON SURREAL AND PHSPD DATASETS**

| Method          | SURREAL | PHSPD |
|-----------------|---------|-------|
| HumanSfP2 (color) | 13.44   | 24.79 |
| HumanSfP2 (no-prior) | 12.16 | 24.75 |
| HumanSfP2 (ours)   | 6.79    | 21.36 |

Exemplar visual results are presented in Fig. 7. It is observed that for DepthHuman, only a partial mesh with respect to the view in the input image is produced, as is also evidenced in Table V. HMD, on the other hand, does not work well, as evidenced by the often error-prone surface details. This may be attributed to the less reliable shading representation, given the new environmental lighting and texture ambiguities existed in these RGB images. PIFu [58] and PIFuHD [60] are able to predict human with clothing details, but both suffer from the relative inaccurate pose inference, making their results ill-aligned with the subject in the image space. PaMIR [24] estimates reasonable poses but reconstructs inaccurate human shapes, which could attribute to the in-correct information provided by their deep implicit function. Our HumanSfP approach is shown capable of producing reliable prediction of clothed body shapes, which once again demonstrates the applicability of polarization imaging in shape estimation, as well as the benefit of engaging the surface normal maps in our approach.

Qualitative results presented in Fig. 8 showcase the generalization ability of our approach. Note the polarization images are...
acquired at different physical locations with distinct background scenes that are very dissimilar to those in the training images.

D. Ablation Study

Polarization and RGB modalities for normal estimation: Here we want to compare the performance of normal map reconstruction from RGB images vs. polarization images. Let HumanSPF2 (color) denote the model that uses only RGB image, HumanSPF2 (no-prior) be the model without incorporating the ambiguous normal maps as the physical priors and with only the polarization image as input. The quantitative results are summarized in Table VI. We observe that HumanSPF2 (ours) exceeds HumanSPF2 (color) by $\sim 7^\circ$ in SURREAL dataset and $\sim 3.5^\circ$ in PHSPD dataset. The larger improvement of MAE in SURREAL dataset may due to the fact that SURREAL dataset is synthesized from naked and minimal dressed body shapes, such that the normal map of human body is smooth thus relatively easier to predict than the real-world PHSPD dataset. Moreover, similar MAE results are presented by HumanSPF2 (no-prior) and HumanSPF2 (color), which showcases the importance of ambiguous normal maps in our approach that carries the critical geometric clues for high performance in surface normal estimation from polarization images.

Normal maps for pose estimation: This section demonstrates the significant performance gain that normal maps provides in pose estimation. As in Table VII, the three methods of HumanSPF1 [31], HumanSPF2 (color) and HumanSPF2 (ours) are equipped with different input combinations: polar/RGB image, polar/RGB image with foreground mask, and polar/RGB image with predicted normal map. Within each method, the performance gain is particularly significant when normal map

| Method       | Input                | SURREAL          | PHSPD            |
|--------------|----------------------|------------------|------------------|
|              |                      | PEL-MPJPE $\downarrow$ | PA-MPJPE $\downarrow$ | PCK $\uparrow$ | MPJPE $\downarrow$ | PEL-MPJPE $\downarrow$ | PA-MPJPE $\downarrow$ | PCK $\uparrow$ |
| HumanSPF1    | Polar                | 95.28 (4.05)     | 67.84 (2.35)     | 0.67 (0.02)     | 75.76 (9.98)     | 73.04 (10.66)    | 0.77 (0.05)     |
|              | Polar+Mask           | 84.78 (14.55)    | 60.82 (9.37)     | 0.72 (0.07)     | 66.97 (18.77)    | 63.12 (20.58)    | 0.83 (0.11)     |
|              | Polar+Normal         | 96.27            | 74.88            | 0.69            | 88.35            | 77.97            | 0.74            |
| HumanSPF2    | Color                | 84.84 (11.45)    | 62.74 (12.14)    | 0.73 (0.04)     | 80.79 (7.76)     | 71.58 (6.39)     | 0.80 (0.06)     |
|              | Color+Mask           | 80.28 (15.99)    | 59.39 (15.49)    | 0.76 (0.07)     | 76.34 (12.21)    | 68.23 (9.74)     | 0.84 (0.10)     |
|              | Color+Normal         | 74.12            | 56.39            | 0.78            | 68.95 (1.78)     | 60.28 (3.54)     | 0.85 (0.02)     |
| HumanSPF2    | Polar                | 72.20 (1.92)     | 55.48 (0.91)     | 0.78 (0)        | 62.04 (8.69)     | 54.98 (8.84)     | 0.88 (0.05)     |
|              | Polar+Mask           | 59.17 (14.95)    | 46.58 (9.81)     | 0.85 (0.07)     | 60.73            | 63.82            | 0.84 (0.05)     |
|              | Polar+Normal         | 59.17 (14.95)    | 46.58 (9.81)     | 0.85 (0.07)     | 60.73            | 63.82            | 0.84 (0.05)     |
Fig. 7. Exemplar estimation results of clothed body shapes. The first and sixth column are RGB and polarization images, respectively. DepthHuman [30], HMD [25], PIFu [58], PIFuHD [60] and PuMIR [24] produce results based on RGB images. HumanSfP (ours) works with polarization input image.

Fig. 8. Exemplar estimation results of clothed body shapes, obtained on polarization images from new scene context.

is incorporated as input. This may due to the rich geometric information encoded in the normal map representation. Less significant gain is obtained when only mask is incorporated as input, which further demonstrates the effectiveness of geometric information in pose estimation. When comparing across polarization and RGB modalities in HumanSfP2 (ours) and HumanSfP2 (color), there is still noticeable improvement in HumanSfP2 (color), which combines RGB image with normal map as the input. However, the overall performance of HumanSfP2 (color) is worse than that of HumanSfP2 (ours). The explanation is that the normal maps estimated from RGB images are not as reliable as those obtained from the polarization image counterparts.

X. Conclusion

We tackle in this paper a new problem of estimating human shapes from single 2D polarization images. Our work exemplifies the applicability of engaging polarization cameras as a promising alternative to the existing imaging modalities for human pose and shape estimation. Moreover, by exploiting the rich geometric details in the surface normal of the input polarization images, our SfP approach is capable of reconstructing human body shapes of surface details. We expect this could be a useful tool in many downstream applications.

Acknowledgment

The authors would like to thank all the volunteers who contribute to the dataset, and W. Ji, C. Xu and M. Gong for their constructive advice.

References

[1] G. Ning, Z. Zhang, and Z. He, “Knowledge-guided deep fractal neural networks for human pose estimation,” IEEE Trans. Multimedia, vol. 20, pp. 1246–1259, 2017.
[2] A. Kamel, B. Sheng, P. Li, J. Kim, and D. D. Feng, “Hybrid refinement-correction heatmaps for human pose estimation,” IEEE Trans. Multimedia, vol. 23, pp. 1330–1342, 2021.

[3] Q. Zhang et al., “Single person dense pose estimation via geometric equilibrium consistency,” IEEE Trans. Multimedia, to be published, doi: 10.1109/TMM.2021.3129056.

[4] S. Park, J. Hwang, and N. Kwak, “3D human pose estimation using convolutional neural networks with 2D pose information,” in Proc. IEEE Eur. Conf. Comput. Vis., 2016, pp. 156–169.

[5] R. Zhao, Y. Wang, and A. M. Martinez, “A simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 12, pp. 2059–2066, Dec. 2018.

[6] B. X. Nie, P. Wei, and S.-C. Zhu, “Monocular 3D human pose estimation by predicting depth on joints,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 3447–3455.

[7] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, “DRPose3D: Depth ranking in 3D human pose estimation,” in Proc. Int. Joint Conf. Artif. Intell., 2018, pp. 978–984.

[8] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, “Learning pose grammar to encode human body configuration for 3D pose estimation,” in Proc. AAAI Conf. Artif. Intell., 2018, pp. 6821–6828.

[9] I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, and C. Theobalt, “In the wild human pose estimation using explicit 2D features and intermediate 3D representations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10905–10914.

[10] K. Wang, L. Lin, C. Jiang, C. Qian, and P. Wei, “3D human pose machines with self-supervised learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 5, pp. 1069–1082, May 2020.

[11] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, “SCAPE: shape completion and animation of people,” in ACM Trans. Graph., ACM, vol. 24, no. 3, pp. 408–416, 2005.

[12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” ACM Trans. Graph., vol. 34, no. 6, pp. 1–16, 2015.

[13] T. Zhao, S. Li, K. N. Ngan, and F. Wu, “3D reconstruction of human body shape from a single commodity depth camera,” IEEE Trans. Multimedia, vol. 21, no. 1, pp. 114–123, Jan. 2018.

[14] Y. Lu, J.-H. Cha, S.-K. Youn, and S.-W. Jung, “Parametric shape estimation of human body under wide clothing,” IEEE Trans. Multimedia, vol. 23, pp. 3657–3667, 2021.

[15] X. Zou et al., “SparseFusion: Dynamic human avatar modeling from sparse RGBD images,” IEEE Trans. Multimedia, vol. 23, pp. 1617–1629, 2021.

[16] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker, “Detected human shape and pose from images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.

[17] E. Díbá, H. Jain, C. Öztireli, R. Ziegler, and M. Gross, “HS-Net: Estimating human body shape from silhouettes with convolutional neural networks,” in Proc. Int. Conf. 3D Vis., 2016, pp. 108–117.

[18] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3857–3866.

[19] Y. Ba et al., “Deep shape from polarization,” in Proc. IEEE Eur. Conf. Comput. Vis., 2020, pp. 554–564.

[20] Z. Cui, J. Gu, B. Shi, P. Tan, and J. Kautz, “Polarimetric multi-view stereo,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1558–1567.

[21] X. Zhou, M. Zhu, S. Leonards, K. G. Derpanis, and K. Daniilidis, “Spareness meets depthness: 3D human pose estimation from monocular video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4966–4975.

[22] I. Akhter and M. J. Black, “Pose-conditioned joint angle limits for 3D human pose reconstruction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1446–1455.

[23] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, “Robust estimation of 3D human poses from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 2361–2368.

[24] S. Li, W. Zhang, and A. B. Chan, “Maximum-margin structured learning with deep networks for 3D human pose estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2015, pp. 2848–2856.

[25] H. Ci, C. Wang, X. Ma, and Y. Wang, “Optimizing network structure for 3D human pose estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 2262–2271.

[26] Y. Cai et al., “Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 2272–2281.

[27] D. Tome, C. Russell, and L. Agapito, “Lifting from the deep: convolutional neural networks for 3D pose estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2018, pp. 7782–7791.

[28] J. Martinez, R. Hosain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3D human pose estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 2640–2649.

[29] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, “Towards 3D human pose estimation in the wild: A weakly-supervised approach,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 398–407.

[30] W. Yang, X. Qiao, X. Wang, J. Ren, H. Li, and X. Wang, “3D human pose estimation in the wild by adversarial learning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2018, pp. 5255–5264.

[31] G. Pavlakos, X. Zhou, and K. Daniilidis, “Ordinal depth supervision for 3D human pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2500–2509.

[32] M. Kadambi, V. Taamazyan, B. Shi, and R. Raskar, “Depth sensing using geometrically constrained polarization normals,” Int. J. Comput. Vis., vol. 125, no. 1–3, pp. 34–51, 2017.

[33] L. Chen, Y. Zheng, A. Subapa-Asa, and I. Sato, “Polarimetric three-view geometry,” in Proc. IEEE Euro. Conf. Comput. Vis., 2018, pp. 20–36.

[34] R. Wehner and M. Müller, “The significance of direct sunlight and polarized light in natural environments,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[35] D. Mehta et al., “Monocular 3D human pose estimation in the wild using improved CNN supervision,” in Proc. Int. Conf. 3D Vis., 2017, pp. 506–516.

[36] G. A. Atkinson and E. R. Hancock, “Recovery of surface orientation from diffuse polarization,” IEEE Trans. Image Process., vol. 15, no. 6, pp. 1653–1664, Jun. 2006.

[37] A. Kazemi, G. Neekhoo, A. S. Mattausch, F. Bogo, V. Jetley, and R. Raskar, “Strengths and limitations of deep pose estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 4801–4809.

[38] D. Wehenkel and M. Müller, “A simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 7, pp. 682–695, Jul. 2014.

[39] J. Martinez, R. Hosain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3D human pose estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 506–516.
[52] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, “Neural body fitting: Unifying deep learning and model based human pose and shape estimation,” in Proc. Int. Conf. 3D Vis., 2018, pp. 484–494.

[53] Y. Xu, S.-C. Zhu, and T. Tung, “Denserac: Joint 3D pose and shape estimation by dense render-and-compare,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 7760–7770.

[54] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, “Learning to reconstruct 3D human pose and shape via model-fitting in the loop,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2252–2261.

[55] H. Jiang, J. Cai, and J. Zheng, “Skeleton-aware 3D human shape reconstruction from point clouds,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 5431–5441.

[56] G. Varol et al., “Bodynet: volumetric inference of 3D human body shapes,” in Proc. IEEE Eur. Conf. Comput. Vis., 2018, pp. 20–36.

[57] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, “DeepHuman: 3D human reconstruction from a single image,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 7739–7749.

[58] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 2304–2314.

[59] Z. Yang et al., “S3: Neural shape, skeleton, and skinning fields for 3D human modeling,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13284–13293.

[60] S. Saito, T. Simon, J. Saragih, and H. Joo, “PIFuHD: Multi-level pixel-aligned implicit function for high-resolution clothed human digitization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2020, pp. 84–93.

[61] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. IEEE Int. Conf. Comput. Vis., 2014, pp. 740–755.

[62] M. Andriluka et al., “PoseTrack: A benchmark for human pose estimation and tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5167–5176.

[63] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, “Recovering accurate 3D human pose in the wild using IMUs and a moving camera,” in Proc. IEEE Eur. Conf. Comput. Vis., 2018, pp. 601–617.

[64] R. M. Murray, Z. Li, S. S. Sastry, and Z. Li, A Mathematical Introduction to Robotic Manipulation. Boca Raton, FL, USA: CRC, 1994.

[65] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, “Efficiently combining positions and normals for precise 3D geometry,” ACM Trans. Graph., vol. 24, no. 3, pp. 536–543, 2005.

[66] Z. Cao, G. H. Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021.

[67] G. Pavlakos et al., “Expressive body capture: 3D hands, face, and body from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10975–10985.

[68] R. Bollapragada, J. Nocedal, D. Mudigere, H.-J. Shi, and P. T. P. Tang, “A progressive batching L-BFGS method for machine learning,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 620–629.

[69] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015, pp. 1–15.

[70] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[71] W. A. Smith, R. Ramamoorthi, and S. Tozza, “Height-from-polarisation with unknown lighting or albedo,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 12, pp. 2875–2888, Dec. 2019.