Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos

Yasamin Jafarian Hyun Soo Park
University of Minnesota
{yasamin, hspark}@umn.edu

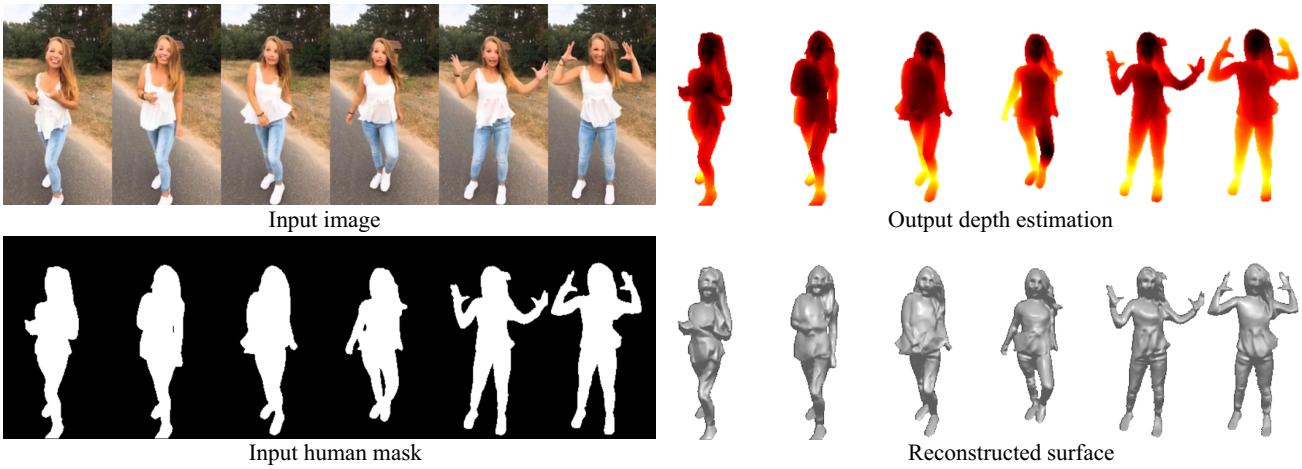


Figure 1: This paper presents a novel approach to estimate high fidelity depths of dressed humans from a single view image by leveraging a new data resource: a number of social media dance videos that span diverse appearance, clothing styles, performances, and identities. We show an example of a sequence of this data and the corresponding human mask along with the estimated depth (the darker, the closer) and the reconstructed surface.

Abstract

A key challenge of learning the geometry of dressed humans lies in the limited availability of the ground truth data (e.g., 3D scanned models), which results in the performance degradation of 3D human reconstruction when applying to real-world imagery. We address this challenge by leveraging a new data resource: a number of social media dance videos that span diverse appearance, clothing styles, performances, and identities. Each video depicts dynamic movements of the body and clothes of a single person while lacking the 3D ground truth geometry. To utilize these videos, we present a new method to use the local transformation that warps the predicted local geometry of the person from an image to that of another image at a different time instant. This allows self-supervision as enforcing a temporal coherence over the predictions. In addition, we jointly learn the depth along with the surface normals that are highly responsive to local texture, wrinkle, and shade by maximizing their geometric consistency. Our method is end-to-end trainable, resulting in high fidelity depth estimation that predicts fine geometry faithful to the input real image. We demonstrate that our method outperforms the state-of-the-art human depth estimation and human shape recovery approaches on both real and rendered images.

1. Introduction

Clothes are an integral part of our everyday life to function, express, and protect ourselves. With the increasing prevalence of VR and AR, the ability to precisely model the complex geometry of dressed humans is becoming the key to authentic social tele-presence. To capture the local geometry, e.g., wrinkle and fabric texture, photogrammetry based on massive camera infrastructure (e.g., 40-500 cameras to cover full body shape) [12, 25, 55] has been used, resulting in production-level rendering [9, 32] and 3D fabrication [3, 5]. Despite its promise, the practical deployment of such massive camera systems in our daily environment is still challenging because of its hardware requirements and computational complexity. Single view reconstruction is an immediate remedy to address this challenge where 3D representation of humans can be learned from the scanned human 3D models [1–3]. Nonetheless, the amount of these

data is limited (e.g., a few hundreds of static models), which do not span diverse poses, appearance, and complex cloth geometry resulting in the performance degradation of 3D human reconstruction when applying to real-world imagery. In this paper, we present a method to reconstruct high fidelity 3D geometry of dressed humans in the form of depths and surface normals from a single view image by exploiting hundreds of dance videos shared in social media (e.g., TikTok mobile application) as shown in Figure 1.

The main characteristics of these dance videos are that 1) each video depicts a sequence of diverse poses of a single person; and 2) 3D ground truth is not available, i.e., existing fully supervised methods are not applicable. We conjecture that since the geometry of dressed humans is an inherent semi-rigid structure, the local geometry of the same person approximately remains constant up to some transformations. For instance, the cloth movement on the left upper arm region undergoes, by large, a rigid transformation when its pose changes. Therefore, it is possible that the geometric consistency over different poses can be applied to learn from the real dance videos. We estimate a transformation for each body part that can warp its 3D geometry from one image to another image at a different time instant. This allows us to self-supervise the predicted geometry of the dressed humans without 3D supervision.

While modern learning based depth estimators are capable of recovering holistic scene geometry, it is shown [30] that it often fails to encode fine local geometry such as complex cloth wrinkles and face profile features, which constitutes the dominant factor of realism. On the other hand, surface normals are highly responsive to fine visual structures such as texture and wrinkles [54]. We exploit the geometric relationship of depths and surface normals to learn jointly (e.g. matching the surface normal to the curvature of the depth).

Our end-to-end trainable method takes as input an RGB image, corresponding human foreground, and human UV coordinates and outputs a high fidelity depth that captures fine wrinkles and shapes that is faithful to the input image. We design a network called *HDNet* that learns the spatial relationship between the image and UV coordinate to produce an intermediate surface normals. These predicted surface normals are, in turn, used to predict the high fidelity depths of dressed humans. We use a Siamese design of HDNet to measure the self-consistency of a pair of geometric predictions. To that end, our method is semi-supervised by leveraging both 3D scanned models and real dance videos. We demonstrate that our method outperforms the state-of-the-art human depth estimation approaches on both real and rendered images.

We present four core contributions: (1) a new dataset called *TikTok dataset* that consists of more than 300 sequences of dance videos shared in a social media mobile platform, TikTok, totaling more than 100K images along with the human mask and human UV coordinates; (2) a novel formulation that warps the 3D geometry of dressed humans from one image to the other image at a different time instant to measure self-consistency, which allows us to utilize the real dance videos; (3) HDNet design that learns to predict fine depths reflective of surface normal prediction by enforcing their geometric consistency; (4) strong qualitative and quantitative prediction on real world imagery.

2. Related Works

Our fundamental contribution lies at the intersection of human body reconstruction, single view depth estimation, and human 3D datasets.

Human Body Reconstruction There are two predominant representations in human body reconstruction: parametric and non-parametric. Similar to face modeling [16], parametric mesh models such as SCAPE [8] and SMPL [33] are an attractive choice of the human body representation, which can be used for single view human reconstruction [13, 26, 29, 39, 41, 44, 56] and synthetic data generation [51, 52]. While the parametric representation effectively limits the space of solutions where learning based approaches can be readily applicable and show remarkable performance, the reconstructed geometry has limited resolution, which prevents from expressing fine details of dressed humans. This has been addressed by refining parametric models with residual geometry [6, 7, 28, 35]. Depth [30, 50] or volumetric representation [23, 60] as a non-parametric representation can describe the geometry of dressed humans. Unlike parametric models, obtaining the ground truth data is challenging: Li et al. [30] addressed by exploiting a large community dataset of Mannequin Challenge, and Tang et al. [50] incorporated semantic labels (pose and segmentation) to regularize their depth estimator.

Single View Depth Estimation Single view depth estimation is a core task of scene understanding where sophisticated designs of convolutional neural networks (CNNs) enable predicting scene geometry [34, 49]. To capture fine details of depth reconstruction, additional cues such as surface normals have been incorporated [18, 36, 40, 42, 43, 50, 59, 61]. Iterative least squares [50] and kernel regression [42] have been used to fuse the surface normals and depths, and coarse-to-fine learning is used to densify LiDAR data for outdoor scenes or missing depth data [43] for indoor scenes [59]. Recently, integrating the surface normal into the depth prediction [54] (e.g. identifying whether a normal representation is realistic or not using GAN [21]) has shown to be effective in restoring local geometry such as cloth wrinkles and face profile features. Unlike previous work, we focus on recovering sub-centimeter detailed geometry tailored to dressed humans by jointly learning depths and surface normals and leveraging a large dataset of social media dance videos.

Human 3D Datasets While there are a number of RGBD

datasets for structural scene understanding [14, 17, 48, 57], a limited amount of data address the problem of the geometry prediction for dressed humans in the wild. A few RGBD datasets [11, 15, 31, 47] are designed for humans action recognition. However, these data lack the geometric details such as cloth wrinkles. For human geometry, the 3D scanned models [1–3] or multiview generated models [53, 58] can be used to generate photorealistic images from multiple views, which has been used for training a geometry predictor with full supervision [45, 46]. However, the amount of data is still limited to a few hundreds of static models, which prevents from learning a generalizable prediction model. In this paper, we introduce a new source of data: real dance videos from the Internet to generalize the human depth estimation to different view points, human appearance, clothing styles and poses.

3. Method

Given a single image of a dressed human $\mathbf{I}$, we reconstruct its high fidelity depth, i.e., $z = g(\mathbf{x}; \mathbf{I})$, where $\mathbf{x} \in \mathbb{R}^2$ is the xy-location in the image, and $z \in \mathbb{R}_+$ is the depth at the corresponding location.

Existing approaches learn g directly from the ground truth data, which shows two limitations in estimating depths of clothed humans. (1) While existing depth estimators are highly responsive to predict holistic scene geometry, it is shown [30] that its expressibility is limited at encoding fine local geometry such as irregular and complex wrinkles, that constitute the dominant factor of realism. (2) It requires a large amount of 3D ground truth data (e.g., ScanNet [17] and KITTI [19, 20]). Such large ground truth data for humans that span diverse appearance, cloth styles, and poses do not exist (e.g., a few hundreds of posed scanned models [1–3]).

3.1. Self-supervised Human Depths from Videos

We present a new method to address these limitations by leveraging large video data of real humans in motion. Albeit lacking of 3D ground truth, each video depicts the movement of a single person across time where her/his geometry approximately remains constant up to local transformations.

Consider a coordinate transform $h(\mathbf{u}) = \mathbf{x}$ that maps a canonical human body surface coordinate $\mathbf{u} \in \mathbb{R}^2$ (UV surface coordinate) to the corresponding point $\mathbf{x}$ in an image. A key feature of the UV surface coordinate is that it is invariant to poses, clothes, and appearance.

We parametrize a 3D point $\mathbf{p} \in \mathbb{R}^3$ reconstructed by the depth prediction using the UV coordinate, i.e.,

$$\mathbf{p}_i(\mathbf{u}) = z\mathbf{K}^{-1}\widetilde{\mathbf{x}} = g(h_i(\mathbf{u}); \mathbf{I}_i)\mathbf{K}^{-1}\widetilde{h}_i(\mathbf{u}), \qquad (1)$$

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the camera intrinsic parameter, $\widetilde{\cdot} \in \mathbb{P}^2$ is the homogeneous representation [24], and $\mathbf{x}$ is the pixel location in the image domain corresponding to $\mathbf{u}$ in the UV domain. The subscript i indicates the time instant.

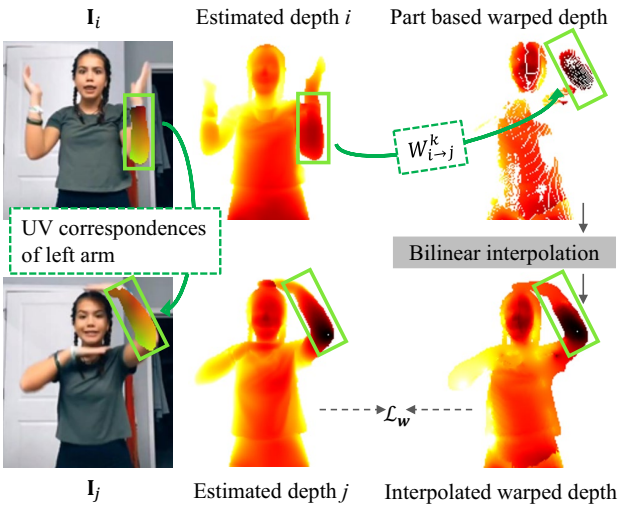


Figure 2: Given the depth estimate at the i^{th} time instant, we use a part based transformation that warps the 3D local geometry of the image to the image at the j^{th} time instant. The green boxes in two images show the UV correspondences of the left arm. The depths of the left arm are reconstructed in 3D and transformed to the j^{th} time to form the part based warped depths. We apply bilinear interpolation on the foreground range, resulting in the warped depths that can supervise the depth estimate at the j^{th} time instant through the warping loss $\mathcal{L}_w$.

We transform a set of points in the k^{th} body part at the i^{th} time instant to the j^{th} time instant:

$$\mathbf{p}_{i\to j}(\mathbf{u}) = \mathcal{W}^k_{i\to j}(\mathbf{p}_i(\mathbf{u})), \quad \mathbf{u} \in \mathcal{U}_k \qquad (2)$$

where $\mathcal{W}$ is a 3D part based warping function, and $\mathcal{U}_k$ is the set of UV coordinates associated with the k^{th} body part. The body part is defined as a region of the body where its local geometry approximately undergoes a rigid transformation, e.g., lower arm. An analogous warping is used for non-rigid tracking [37] without the part based representation, which allows mapping between consecutive frames. With the part based warping, we substantially extend the time horizon by parametrizing the 3D point using the UV coordinate, which does not require an offline iterative closest point method between the consecutive frames.

We use the Special Eucliean Transform ($SE3$) to model $\mathcal{W}^k_{i\to j}(\mathbf{p}_i) = \mathbf{R}^k_{i\to j}\mathbf{p}_i + \mathbf{t}^k_{i\to j}$ where $\mathbf{R}$ and $\mathbf{t}$ are the rotation and translation[1]. Among the correspondences $\mathbf{p}_i(\mathbf{u}) \leftrightarrow \mathbf{p}_j(\mathbf{u})$, we pre-define a subset of correspondences that represent the overall transformation for each part. With the pre-defined correspondences, we compute the transformation by minimizing the following error:

$$\underset{\mathbf{R},\mathbf{t}}{\text{minimize}} \sum_l \left\|\mathbf{p}_j(\mathbf{v}_l) - \mathcal{W}^k_{i\to j}(\mathbf{p}_i(\mathbf{v}_l))\right\|^2, \quad \mathbf{v}_l \in \mathcal{V}_k \subset \mathcal{U}_k,$$

[1] An affine transformation used in the non-rigid tracking [37] can be a complementary to the SE3.

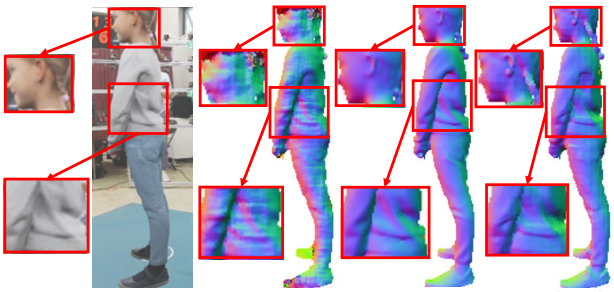


Figure 3: We compare our method with Tang et al. [50] on the surface normals derived from the depths. While two methods use the surface normals to enhance the depths, unlike Tang et al., our method jointly learns surface normals and depths by supervising them with each other, which produces more realistic and less noisy prediction that preserves the detailed geometry of wrinkles and face.

where $\mathcal{V}_k$ is the subset of the UV coordinates that represent the overall transformation. We minimize the objective using least squares [10]. In practice, we choose the sparse correspondences in the subset by discretizing the UV coordinates. This transformation is computed online, i.e., the transformation changes as the depth prediction is updated at each training iteration.

Figure 2 illustrates the self-supervision via warping the 3D geometry of humans between two frames of a video. By having the estimated depth, we use the UV coordinates to warp the local geometry for each part from the i^{th} time instant to the j^{th} time instant resulting in a sparse warped depth. We apply bilinear interpolation on the foreground range to get a dense warped depths that can supervise the depth estimate at the j^{th} time instant by minimizing warping loss $\mathcal{L}_w$.

We minimize the following loss to measure geometric discrepancy between two time instances:

$$\mathcal{L}_w = \sum_{l} \sum_{(i,j)\in\mathcal{V}_l} \sum_{k} \sum_{\mathbf{u}\in\mathcal{U}_k} \|\mathbf{p}_j(\mathbf{u}) - \mathbf{p}_{i\to j}(\mathbf{u})\|^2, \quad (3)$$

where $\mathcal{V}_l$ is the set of time instances within the l^{th} video.

Equation (3) allows us to utilize a large amount of real videos without the 3D ground truth via self-supervision, i.e., the estimated depth in one pose can be used to supervise the depth in the other pose. This makes the depth estimation responsive to real data of diverse human poses and appearances.

3.2. Joint Learning of Surface Normal and Depth

Surface normal estimation is highly responsive to the local texture, wrinkle, and shade [50, 54]. We jointly estimate surface normal and depth to benefit from each other. We estimate the surface normals of an image $\mathbf{I}$, i.e., $\mathbf{n} = f(\mathbf{x}; \mathbf{I})$ where $\mathbf{n} \in \mathbb{S}^2$ is the unit surface normal vector represented in the camera coordinate system.

Surface normal $\widehat{\mathbf{n}}(\mathbf{x})$ is the curvature that is perpendicular to the tangential plane of the corresponding 3D point $\mathbf{p}(\mathbf{x})$ (we override the notation $\mathbf{p}(\mathbf{u})$ in Equation (1)), i.e.,

$$\widehat{\mathbf{n}}(\mathbf{x}) = \frac{\partial \mathbf{p}(\mathbf{x})}{\partial x} \times \frac{\partial \mathbf{p}(\mathbf{x})}{\partial y} \Big/ \left\| \frac{\partial \mathbf{p}(\mathbf{x})}{\partial x} \right\| \left\| \frac{\partial \mathbf{p}(\mathbf{x})}{\partial y} \right\|, \quad (4)$$

where $\widehat{\mathbf{n}}$ denotes the surface normal estimate derived by the depth estimate.

We ensure geometric consistency between the predicted surface normals and the derived surface normals from the depth estimates by minimizing their geometric error:

$$\mathcal{L}_s = \sum_{\mathbf{I}_i\in\mathcal{D}} \sum_{\mathbf{x}\in\mathcal{R}(\mathbf{I}_i)} \cos^{-1}\left(\frac{\mathbf{n}^{\mathsf{T}}(\mathbf{x})\widehat{\mathbf{n}}(\mathbf{x})}{\|\mathbf{n}(\mathbf{x})\|\|\widehat{\mathbf{n}}(\mathbf{x})\|} \right), \quad (5)$$

where $\mathcal{R}(\mathbf{I})$ is the coordinate range of the image $\mathbf{I}$, and $\mathcal{D}$ is the image dataset including the dance videos and scanned 3D models.

Note that the relationship between surface normal and depth has been used to obtain the details of depth estimates. GeoNet [42] has leveraged the derived surface normals to refine the surface normal estimate for an indoor scene understanding. In the human domain, Tang et al. [50] uses the surface normal prediction to refine the human depth in a post-processing manner. Unlike these methods, we use the surface normal estimates to supervise the depths and the depth estimates to supervise the surface normals by enforcing their geometric consistency in the training phase. This end-to-end online pipeline enables learning the depths from the real videos without the ground truth depth. Figure 3 illustrates the comparison of the surface normal generated from the predicted depth of our method and Tang et al. [50]. Our result is realistic, which captures the wrinkles of the cloth fabric compared to Tang et al. [50].

3.3. Network Design

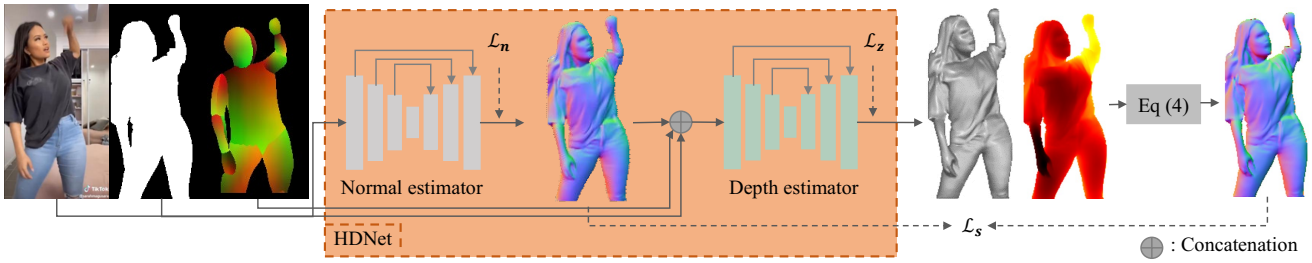
We minimize the following overall loss to learn the depth and surface normal estimators from real videos and 3D scanned models:

$$\mathcal{L} = \mathcal{L}_z + \lambda_n\mathcal{L}_n + \lambda_s\mathcal{L}_s + \lambda_w\mathcal{L}_w, \quad (6)$$

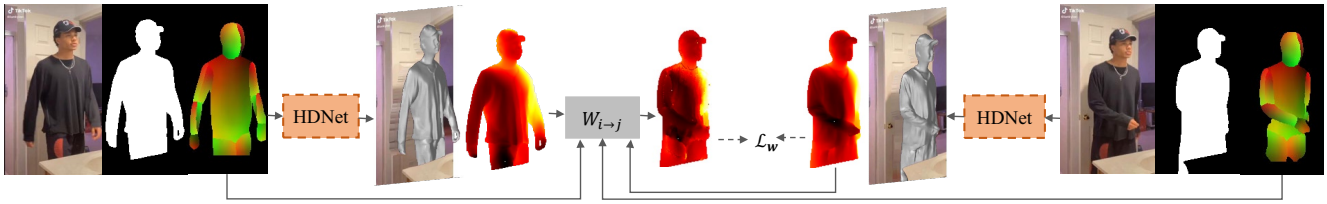
where λ_n, λ_s, and λ_w are relative weights between losses. In addition to self-consistency losses (λ_w and λ_s), we utilize the 3D ground truth data from the 3D scanned models [2]. This depth and surface normal can be learned by minimizing the following error between ground truth normal $\mathbf{N}(\mathbf{x})$ and the prediction.

$$\mathcal{L}_z = \sum_{\mathbf{I}_i\in\mathcal{D}_s} \sum_{\mathbf{x}\in\mathcal{R}(\mathbf{I}_i)} \|Z(\mathbf{x}) - g(\mathbf{x}; \mathbf{I})\|^2, \quad (7)$$

$$\mathcal{L}_n = \sum_{\mathbf{I}_i\in\mathcal{D}_s} \sum_{\mathbf{x}\in\mathcal{R}(\mathbf{I}_i)} \|\mathbf{N}(\mathbf{x}) - f(\mathbf{x}; \mathbf{I})\|^2, \quad (8)$$



(a) Network design of HDNet.



(b) HDNet self-supervision using two images from different time instances.

Figure 4: (a) Our network *HDNet* takes as input an image with the correspondending human foreground and UV coordinates and predicts the high fidelity depth of the human. The HDNet is composed of the depth and surface normal estimators. The surface normal estimator takes the input image and the foreground human mask and outputs the surface normal estimation. The surface normal estimation is, in turn, used as an input along with the image, foreground human mask, and part based UV coordinate to the depth estimator. We enforce the geometric consistency between the estimated depths and surface normals. (b) We build a Siamese design of HDNet to leverage real dance videos. The estimated depth of one image is warped to the other image at a different time instant using a part based transformation. We measure the geometric consistency between the predicted depth and warped depth through $\mathcal{L}_w$.

where $\mathcal{D}_s$ is the 3D scanned dataset with the ground truth depths $Z(\mathbf{x})$ and surface normals $\mathbf{N}(\mathbf{x})$.

Network Design and Details We design our neural network called *HDNet* (Human Depth Neural Network) that allows us to utilize both real videos and 3D scanned model data as shown in Figure 4(a). HDNet is composed of two estimators: surface normal and depth estimators. The surface normal estimator $f(\mathbf{x};\mathbf{I})$ takes as input an RGB image and its foreground mask, and outputs the surface normal estimates. The depth estimator, $g(\mathbf{x};\mathbf{I})$, in turn, takes as input a triplet of an RGB image, foreground mask, and UV coordinate, and outputs the depth estimates. The geometric consistency between the surface normal and depth is enforced by minimizing $\mathcal{L}_s$. For the 3D scanned model data, both estimators are supervised by the ground truth surface normal and depth ($\mathcal{L}_n$ and $\mathcal{L}_z$), respectively.

For the real videos, we build a Siamese network with HDNet where two triplets from two time instances within the same video are used for the depth estimates as shown in 4(b). The UV coordinates from both images are used to compute the special Euclidean transformation that is used to warp the depth from one image to the other image. At each time instant, we make five image pairs by randomly selecting time instances that have the UV coordinates of at least five common visible body parts while each contains more than 50 overlapping UV coordinates.

For the two estimators, we use the stacked hourglasses network [38] as a backbone network. The image and its foreground mask are cropped from the input image and resized to 256×256, and h is approximated by the inverse of the UV map obtained by DensePose [22]. We use Adam optimizer [27] with the following parameters for the training. Batch size: 10; learning rate: 0.001; the number of epochs: 380; $\lambda_\mathbf{n}$: 1; $\lambda_\mathbf{s}$: 0.5; and $\lambda_\mathbf{w}$: 5; GPU model: NVIDIA V100.

4. TikTok Dataset

We learn high fidelity human depths by leveraging a collection of social media dance videos scraped from the TikTok mobile social networking application. It is by far one of the most popular video sharing applications across generations, which include short videos (10-15 seconds) of diverse dance challenges as shown in Figure 5. We manually find more than 300 dance videos that capture a single person performing dance moves from TikTok dance challenge compilations for each month, variety, type of dances, which are moderate movements that do not generate excessive motion blur. For each video, we extract RGB images at 30 frame per second, resulting in more than 100K images. We segmented these images [4], and computed the UV coordinates. The dataset and code can be found in https://www.yasamin.page/hdnet_tiktok.

	Tang et al. dataset [50]			RenderPeople dataset [2]			Vlasic et al. dataset [53]		
Method	D. error	N. error	R. error	D. error	N. error	R. error	D. error	N. error	R. error
Li et al. [30]	1.59±1.11	0.68±0.13	0.10±0.05	3.54±3.78	0.48±0.13	0.10±0.06	5.55±5.98	0.76±0.13	0.27±0.10
Tang et al. [50]	1.21±1.61	0.54±0.12	0.08±0.12	3.66±3.29	0.59±0.10	0.12±0.05	2.29±2.03	0.73±0.09	0.23±0.07
PIFu [45]	1.52±1.07	0.57±0.09	0.10±0.06	2.28±1.86	0.43±0.09	0.09±0.04	3.04±2.97	0.69±0.09	0.22±0.06
PIFuHD [46]	1.45±0.86	0.60±0.09	0.09±0.05	2.33±1.92	0.47±0.09	0.09±0.04	3.50±3.43	0.80±0.09	0.19±0.05
Ours	**1.21±0.81**	**0.51±0.07**	**0.08±0.05**	**1.11±0.75**	**0.27±0.05**	**0.05±0.02**	**1.21±0.98**	**0.44±0.07**	**0.13±0.04**

Table 1: Quantitative Results. D. error (normalized error), N. error (rad) and R. error represent depth error, normal error, and reconstruction error respectively (mean±std).

Figure 5: TikTok Dataset. We present a new dataset called *TikTok dataset* that consists of more than 300 sequences of dance videos shared in a social media mobile platform, TikTok, totaling more than 100K images along with the human mask and human UV coordinates.

5. Experiments

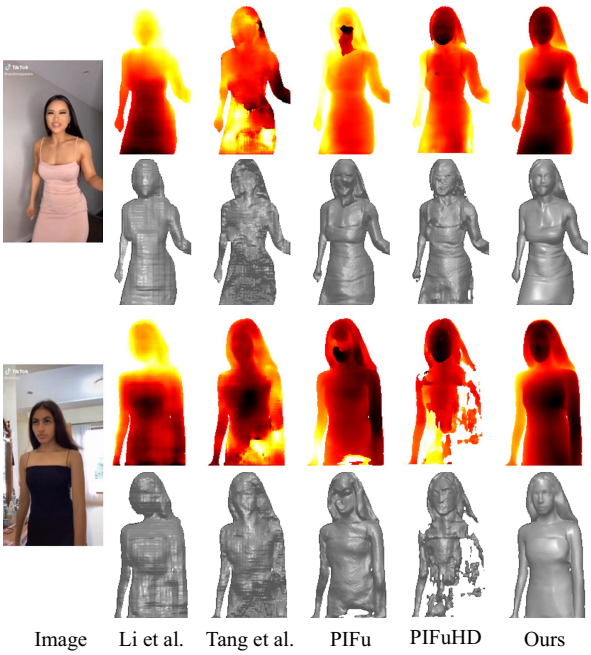
We evaluate our method both quantitatively and qualitatively compared with the state-of-the-art methods of human depth estimation and human shape recovery on real and synthetic data.

Training Datasets We use two datasets for training: 340 subjects from 3D scanned model (RenderPeople) [2] with 3D ground truth and our TikTok dataset without 3D ground truth (Section 4). We render the 3D scanned mesh models from approximately 100 viewpoints sampled uniformly across a camera rig (6m diameter) that encircles each subject with 16.5mm focal length. Total 34,000 and 100,000 images are used for training from RenderPeople and TikTok data, respectively.

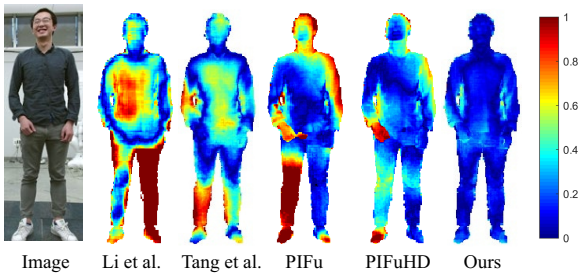
Evaluation Datasets We use three datasets to compare the performance of ours and baseline methods: Tang et al. [50], RenderPeople [2], and Vlasic et al. [53]. 1) *Training dataset of Tang et al.* This dataset is made of sequences of depth and RGB image pair for 25 subjects. We randomly choose around 70 frames for each subject, totaling 1300 images. 2) *RenderPeople dataset* This dataset is made of 3D scanned models with texture. We choose 6 subjects that are not part of our training data and render the images from 100 viewpoints, totaling 600 images. We use a ray tracing algorithm to compute the ground truth depth and render the human textured model. 3) *Vlasic et al. dataset* This dataset consists of 10 sequences of different people viewed from 8 views. Each video includes average of 200 frames of diverse activities such as swing dancing, samba dancing, jumping, squat, and marching. The dataset provides the RGB images and the meshes along with the camera parameters. We use a ray tracing algorithm to generate the ground truth depth from the meshes. We randomly choose total of 2000 images from this dataset. This dataset is in particular challenging because the viewpoints are substantially different from the existing datasets, i.e., a subject is viewed from an oblique view.

Evaluation metric We evaluate the performance in two aspects: (1) accuracy of depths, surface normals, 3D reconstruction, and (2) impact of joint training of surface normal and depth ($\mathcal{L}_s$) and integration of real dance videos ($\mathcal{L}_w$). We use mean squared error and mean absolute angular error as a metric for depth and surface normal, respectively. The surface normals are computed via Equation (4) and compared with the ground truth. In addition, we measure the 3D error by reconstructing the estimated depth. Since depths are reconstructed up to scale, we scale the reconstructed depth to match to the ground truth, i.e., the predicted depth is translated to the median of ground truth and scaled to match the minimum/maximum depths.

We followed the evaluation protocol of Li et al. [30], i.e., no retraining of the baseline models. We categorize the baseline methods into two: human depth estima-



(a) Qualitative comparison on TikTok dataset



(b) Error map

Figure 6: (a) We compare our method with the baseline methods (Li et al. [30], Tang et al. [50], PIFu [45], and PIFuHD [46]) on the TikTok dataset. (b) We measure the normalized error of estimated depths on the training data of Tang et al [50].

tion [30,50], and human shape recovery [45,46]. The quantitative comparison is summarized in Table 1.

i) *Human shape recovery* We compare our method with non-parametric human shape recovery designed for dressed humans (PIFu [45] and PIFuHD [46]) using an implicit function. Note that these methods predict not only the frontal body surface but also occluded body surface where we measure error only for the visible region. We apply a ray tracing method to identify the frontal surface where we measure the depth and surface normal.

ii) *Human depth estimation* We compare with depth estimation baselines that are tailored to dressed humans, which are most relevant to our work. Li et al. [30] used a large community dataset called MannequinChallenge dataset to train the stacked hourglasses [38], and Tang et al. [50] leveraged surface normals and depths to preserve detailed dressed human shapes. Since these two methods were designed for human depths, in particular, Li et al. [30] shows strong performance on both depths and surface normals.

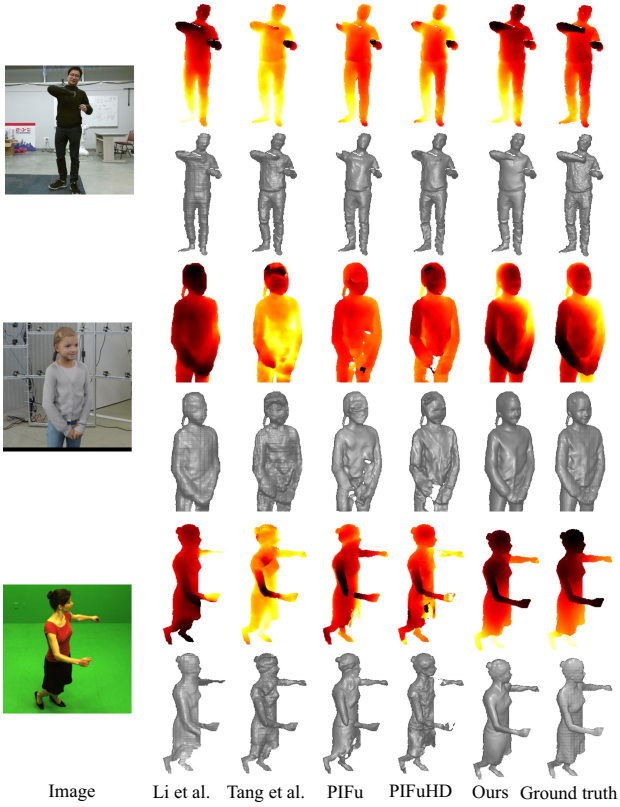


Figure 7: Qualitative results of Li et al. [30], Tang et al. [50], PIFu [45], PIFuHD [46] and ours. We show the results on the evaluation datasets: (1) Tang et al. training dataset [50] (first row), (2) RenderPeople dataset [2] (middle row) and Vlasic et al. dataset [53] (last row).

Figure 6(a) shows the evaluation of our method compared to the baseline methods on TikTok dataset. We can get the most representative depth estimation compared to other methods. Figure 6(b) visualizes the error map of our depth prediction and other baseline methods on Tang et. al. dataset [50]. Figure 7 shows the qualitative results of our method and the baseline methods on the evaluation datasets. We normalize the error $\widehat{e}$ with respect to the ground truth depth, i.e., $\widehat{e} = |z - Z|/Z$ where z and Z are the zero-mean predicted and ground truth depths, respectively.

Ablation study We conduct an ablation study to analyze the impact of the losses in training: $\mathcal{L}_{\mathbf{z}}$, $\mathcal{L}_{\mathbf{w}}$ and $\mathcal{L}_{\mathbf{s}}$. We consider three combinations: $\mathcal{L}_{\mathbf{z}}$, $\mathcal{L}_{\mathbf{z}}$+$\mathcal{L}_{\mathbf{s}}$, and $\mathcal{L}_{\mathbf{z}}$+$\mathcal{L}_{\mathbf{s}}$+$\mathcal{L}_{\mathbf{w}}$. We use the Tang et al. dataset [50] for the evaluation. We scale the predicted depth to match to the ground truth, i.e., the predicted depth is translated to the median of ground truth and scaled to match the minimum/maximum depths.

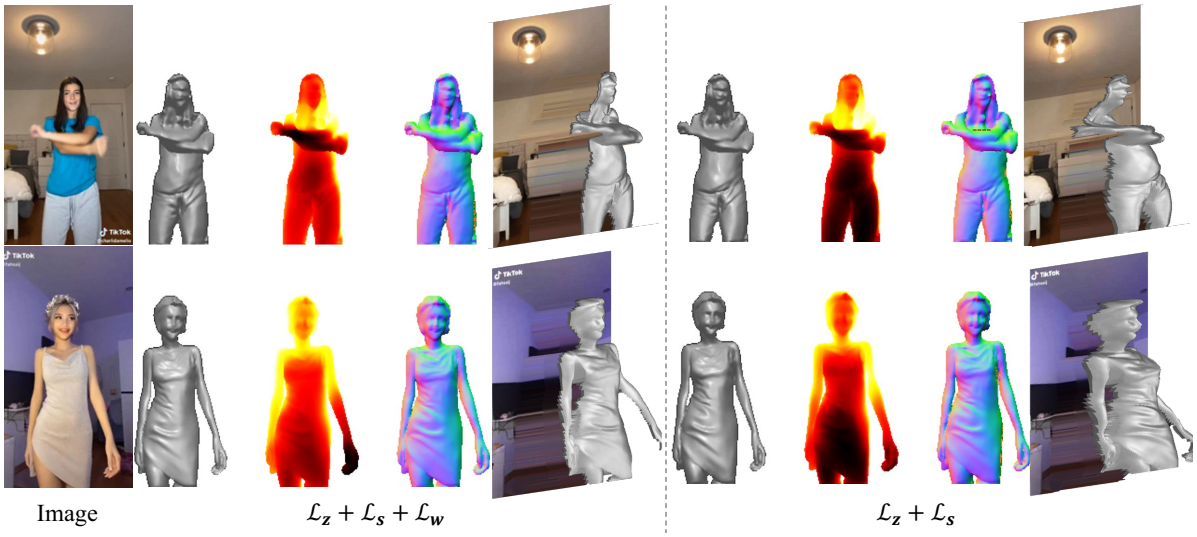


Figure 8: Ablation study on loss functions. from left to right: the image, the full method results, and the results without self-supervision.

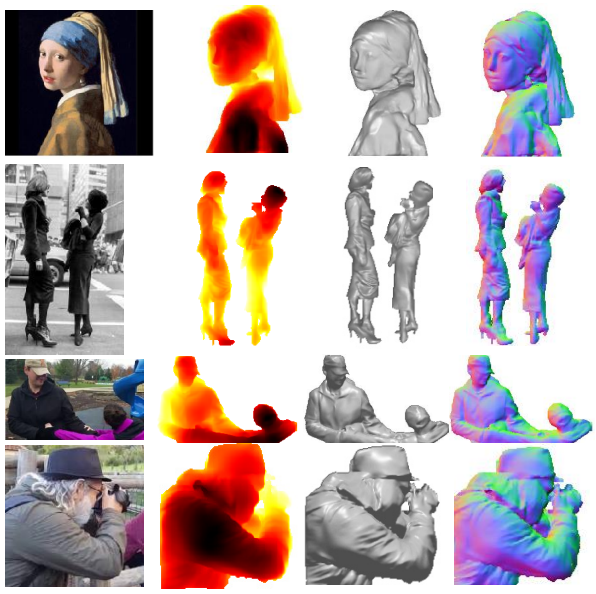


Figure 9: Qualitative results of our method on web images. From left to right: image, predicted depth, recosntructed surface and surface normal.

Table 2 summarizes the comparison of the combinations. On the one hand, $\mathcal{L}_\mathbf{w}$ enforces the network to learn the geometric consistency from the videos. This loss is highly effective, allows learning from a limited amount of 3D data. On the other hand, $\mathcal{L}_\mathbf{s}$ enforces to learn to recover the details, which can further reduce the depth and surface normal errors. Our method that leverages all three losses shows the most accurate prediction in reconstructing the depths and surface normals (last row of Table 2).

Figure 8 shows the qualitative results of our ablation study on two examples of TikTok data. From left to right we have the image, the final method results and the results without self supervision.

Losses	D. error	N. error	R. error
$\mathcal{L}_\mathbf{z}$	1.486±1.083	0.597±0.099	0.096±0.060
$\mathcal{L}_\mathbf{z} + \mathcal{L}_\mathbf{s}$	1.290±0.874	0.523±0.084	0.087±0.054
$\mathcal{L}_\mathbf{z} + \mathcal{L}_\mathbf{s} + \mathcal{L}_\mathbf{w}$	**1.212±0.812**	**0.512±0.076**	**0.083±0.051**

Table 2: Ablation study on Tang et al. dataset [50]. D. error (normalized error), N. error (rad) and R. error represent depth error, normal error, and reconstruction error respectively (mean±std).

We also visualize the performance of our method on a set of web images in Figure 9. Our method is generalizable to gray scale images, paintings, and images with multiple people.

6. Conclusion

This paper presents a new method to utilize large data of video data shared in social media to predict the depths of dressed humans. Our formulation allows self-supervision of depth prediction by leveraging local transformations to enforce geometric consistency across different poses. In addition, we jointly learn the surface normal and depth to generate high fidelity depth reconstruction. A new dataset called TikTok dataset is collected, consisting of more than 300 sequences of dance videos shared in a social media mobile platform, TikTok, totaling more than 100K images. Our method produces strong qualitative and quantitative prediction on real world imagery compared to the state-of-the-art human depth estimation and human shape recovery.

Acknowledgement This work was supported by a NSF NRI 2022894 and NSF CAREER 1846031.

References

[1] http://secure.axyz-design.com/. 1, 3

[2] https://renderpeople.com/3d-people. 1, 3, 4, 6, 7

[3] https://web.twindom.com/. 1, 3

[4] https://www.remove.bg/. 5

[5] https://www.shapify.me/. 1

[6] H. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3D people models. In *CVPR*, 2018. 2

[7] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 2

[8] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE:shape completion and animation of people. In *SIGGRAPH*, 2005. 2

[9] M. Armando, J.-S. Franco, and E. Boyer. Adaptive mesh texture for multi-view appearance modeling. In *3DV*, 2018. 1

[10] K. Arun, T. Huang, and S. Blostein. Least-squares fitting of two 3-d point sets. *TPAMI*, 1987. 4

[11] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *First International Workshop on Re-Identification*, 2012. 3

[12] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *SIGGRAPH*, 2010. 1

[13] 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 *ECCV*, 2016. 2

[14] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *3DV*, 2017. 3

[15] C. Chen, R. Jafari, and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *IEEE International Conference on Image Processing*, 2015. 3

[16] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 2001. 2

[17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 3

[18] X. Fei, A. Wong, and S. Soatto. Geo-supervised visual depth prediction. In *ICRA*. 2019. 2

[19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 3

[20] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. 2014. 2

[22] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense human pose estimation in the wild. In *CVPR*, 2018. 5

[23] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 2

[24] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Second edition, 2004. 3

[25] H. Joo, H. S. Park, and Y. Sheikh. Map visibility estimation for large-scale dynamic 3D reconstruction. In *CVPR*, 2014. 1

[26] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[27] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[28] Z. Laehner, D. Cremers, and T. Tung. DeepWrinkles: Accurate and realistic clothing modeling. In *ECCV*, 2020. 2

[29] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 2

[30] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 2, 3, 6, 7

[31] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 2019. 3

[32] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *SIGGRAPH*, 2018. 1

[33] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 2

[34] X. Luo, J. Huang, R. Szeliski, K. Matzen, and J. Kopf. Consistent video depth estimation. In *SIGGRAPH*, 2020. 2

[35] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. Black. Learning to dress 3D people in generative clothing. In *CVPR*, 2020. 2

[36] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. In *SIGGRAPH*, 2005. 2

[37] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 3

[38] A. Newell, K. Yang, and J. Deng. Based on stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 5, 7

[39] 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 *3DV*, 2018. 2

[40] R. Or-el, G. Rosman, A. Wetzler, R. Kimmel, and A. Bruckstein. RGBD-Fusion: Real-time high precision depth recovery. In *CVPR*, 2015. 2

[41] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2

[42] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia. GeoNet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018. 2, 4

[43] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys. DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *CVPR*, 2019. 2

[44] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, 2016. 2

[45] 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 *ICCV*, 2019. 3, 6, 7

[46] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 3, 6, 7

[47] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016. 3

[48] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3

[49] F. Tan, H. Zhu, Z. Cui, S. Zhu, M. Pollefeys, and P. Tan. Self-supervised human depth estimation from monocular videos. In *CVPR*, 2020. 2

[50] S. Tang, F. Tan, K. Cheng, Z. Li, S. Zhu, and P. Tan. A neural network for detailed human depth estimation from a single image. In *ICCV*, 2019. 2, 4, 6, 7, 8

[51] G. Varol, D. Ceylan, B. Russell, J. Yang, E. ¨Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 2

[52] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[53] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. 2008. 3, 6, 7

[54] L. Wang, X. Zhao, T. Yu, S. Wang, and Y. Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In *ECCV*, 2020. 2, 4

[55] A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, and P. Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *SIGGRAPH*, 2005. 1

[56] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 2

[57] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *ICCV*, 2013. 3

[58] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, 2017. 3

[59] Y. Zhang and T. Funkhouser. Deep depth completion of a single rgb-d image. In *CVPR*, 2018. 2

[60] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. DeepHuman: 3D human reconstruction from a single image. In *ICCV*, 2019. 2

[61] T. Zhu, P. Karlsson, and C. Bregler. Simpose: Effectively learning densepose and surface normal of people from simulated data. In *ECCV*, 2020. 2