Self-supervised 3D Representation Learning of Dressed Humans from Social Media Videos

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

Abstract—A key challenge of learning a visual representation for the 3D high fidelity 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 learn a visual representation from these videos, we present a new self-supervised learning 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 by enforcing a temporal coherence over the predictions. In addition, we jointly learn the depths 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 further provide a theoretical bound of self-supervised learning via an uncertainty analysis that characterizes the performance of the self-supervised learning without training. 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.

Index Terms—single view 3D reconstruction, depth estimation, normal estimation, high fidelity human reconstruction, self-supervised learning, dataset.

1 Introduction

Consider a historic photograph of Frida Kahlo wearing a beautiful shoulder scarf and ornaments as shown in Figure 1. We as humans can effortlessly perceive the fine-grained 3D geometry of her face, draped scarf, hair, and earrings from this 2D photograph. Can a machine be equipped with such perceptual capability such that we can travel back to the early 1930s to see her lively moment? With the increasing prevalence of VR and AR, this perceptual capability to precisely model the complex geometry of humans is becoming the key to authentic social tele-presence. Note that existing parametric body models for humans such as SMPL and its variants [13], [16], [17], [30], [33], [37], [41], [45], [47], [50], [62] have limited expressibility to model the complex 3D geometry of dressed humans.

To capture the fine-grained 3D geometry, e.g., wrinkle and fabric texture, photogrammetry based on massive camera infrastructure (e.g., 40-500 cameras to cover full body shape) [12], [29], [61] has been used, resulting in production-level rendering [9], [36], [64] 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 in a supervised fashion from the scanned human 3D models [1]–[3], [64]. Nonetheless, the number of these 3D data to train such model 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. This makes a sharp contrast with the existing datasets available for scene understanding (e.g., ScanNet [19]) that are made of millions of data instances to learn geometry and visual semantics, i.e., the size of data for humans is at least one or two orders of magnitude smaller than that for scenes.

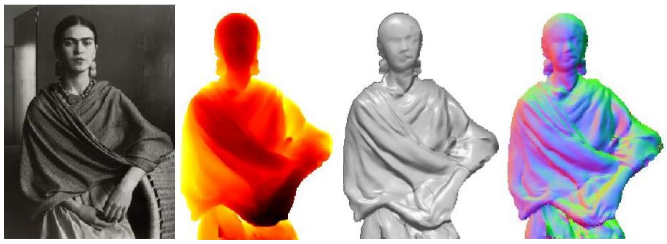


Fig. 1: We present a novel method that takes as input an image of dressed human(s) and outputs high fidelity depths and its surface normals. The estimated depths capture fine wrinkles of scarf, dress, and body shape. Photograph of Frida Kahlo by Imogen Cunningham.

In this paper, we fill this data gap by utilizing a new type of human visual data—hundreds of dance videos shared in social media (e.g., TikTok mobile application) that span diverse appearance, pose, shape, and identities. This enables us to reconstruct high fidelity 3D geometry of dressed humans in the form of depths and surface normals from a single view image as shown in Figure 2. 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 paradigm is not applicable.

To learn a visual representation of high fidelity 3D human geometry from social media dance videos, we make use of two geometric properties agnostic to 3D ground truth. (1) Local shape invariance: we conjecture that since the geometry of dressed

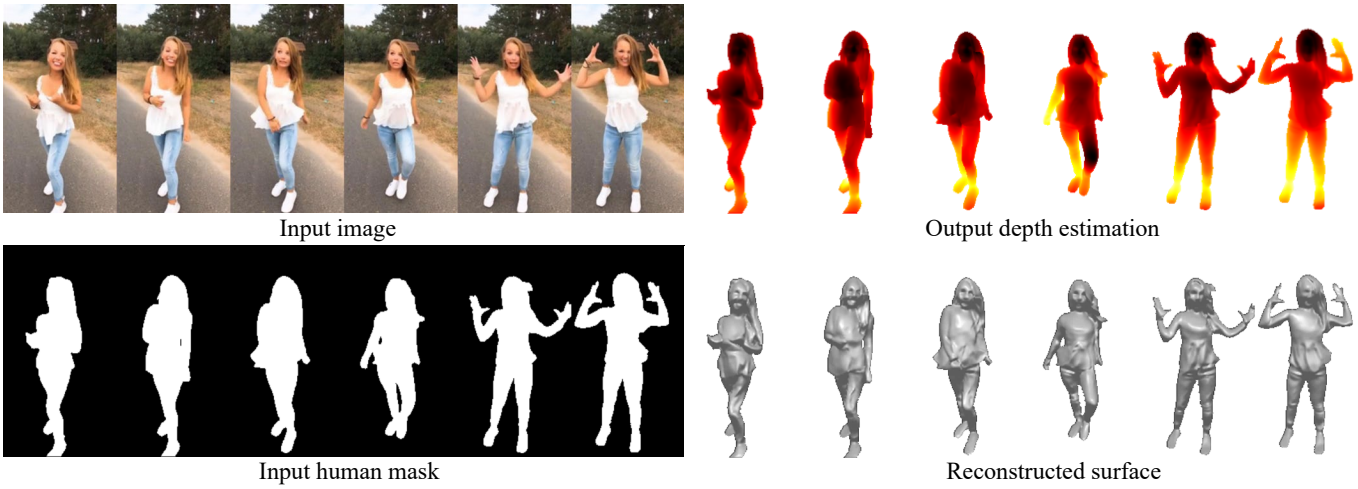


Fig. 2: This paper presents a novel approach to estimate high fidelity depths of dressed humans from a single view image by leveraging a new data resource: a number of social media dance videos that span diverse appearance, clothing styles, performances, and identities. We show an example sequence and the corresponding human mask along with the estimated depth (the darker, the closer) and the reconstructed surface.

humans is inherently semi-rigid so that the local geometry of the same person approximately remains constant up to some pose transformations. For instance, the cloth movement on the left upper arm region undergoes approximately 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. This geometry transformation allows further applying photometric consistency. (2) Geometric consistency: while modern learning based depth estimators are capable of recovering holistic scene geometry, it often fails to encode fine local geometry such as complex cloth wrinkles and face profile features [34], 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 [60]. We exploit the geometric relationship to jointly learn depths and surface normals (e.g. matching the surface normal to the curvature of the depth).

Our end-to-end trainable method takes as input an RGB image, the corresponding human foreground, and human UV coordinates and outputs high fidelity depths of fine wrinkles and shapes that are faithful to the input image. We design a network called *HDNet* that learns to predict the depths and surface normals, and the predicted surface normals are, in turn, used to ensure the geometric consistency with the predicted depths. We use a Siamese design of HDNet to measure the self-consistency across time by warping one prediction to another. To the 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.

Our core contributions include: (1) a new dataset called *TikTok dataset* that consists of more than 340 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.

Building upon the earlier version [28], we make the following additional contributions: (1) incorporating photometric consistency to learn the visual representation; (2) generalizing Euclidean transformation to affine transformation to handle large deformation (3) including a new baseline PaMIR [67] and a new evaluation on THuman2.0 [64] dataset.

2 Related Works

Our paper tackles a problem that 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 [18], parametric mesh models such as SCAPE [8] and SMPL [37] are an attractive choice of the human body representation, which can be used for single view human reconstruction [13], [16], [17], [30], [33], [41], [45], [47], [50], [62] and synthetic data generation [57], [58]. The number of parameters to model a 3D full body is relatively small (pose and shape), which makes regressing the parameters from a single view image possible. However, despite their remarkable performance, the reconstructed geometry has a limited resolution predefined by the mesh topology, which prevents them from expressing the fine details of dressed humans. For example, the fine-grained 3D geometry such as clothes and hair cannot be modeled. These challenges have been addressed by refining parametric models with residual geometry [6], [7], [32], [39]. Depth [34], [56] or volumetric representation [26], [68] as a non-parametric representation can describe the geometry of dressed humans. Tan et al. [55] combined the parametric and nonparametric representations in a semi-supervised manner by leveraging the videos of people in motion. They predicted the SMPL representation and then refined it to include more details on the surface by leveraging the photometric consistency in the temporal domain. However, the

complex clothing items such as skirts still cannot be captured in this method. A non-parametric representation is a viable solution to model such complex geometry of dressed humans. However, it requires a large amount of 3D ground truth data to predict a number of parameters (e.g., for depths, the number of prediction variables is comparable to the number of pixels). Li et al. [34] addressed this challenge by exploiting a large community dataset of Mannequin Challenge, and Tang et al. [56] 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 [38]. To capture fine details of depth reconstruction, additional cues such as surface normals have been incorporated [20], [42], [46], [48], [49], [56], [66], [69]. Iterative least squares [56] and kernel regression [48] 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 [49] for indoor scenes [66]. Recently, integrating the surface normal into the depth prediction [60] (e.g. identifying whether a normal representation is realistic or not using GAN [24]) 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. Unfortunately, to date, there exists no human visual data of which scale is comparable to the scene understanding datasets such as ScanNet [19] and KITTI [21]–[23], [40]. This presents a new challenge for learning a visual representation for human single view depth estimation.

Human 3D Datasets While there are a number of RGBD datasets for structural scene understanding [14], [19], [54], [63], a limited amount of data address the problem of the geometry prediction for dressed humans in the wild. A few RGBD datasets [11], [15], [35], [53] 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 [59], [65] can be used to generate photorealistic images from multiple views, which has been used for training a geometry predictor with full supervision [51], [52]. However, the amount of data is still limited to a few hundreds of static models, which prevents learning a model that can predict images of humans in the wild. In this paper, we introduce a new source of data: real dance videos from social media to generalize the human depth estimation to different viewpoints, 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 dressed humans. (1) While existing depth estimators are highly responsive to predict holistic scene geometry, it is shown [34] that its expressibility is limited at encoding fine local geometry such as irregular and complex wrinkles, which constitute the dominant factor of human geometry/rendering realism. (2) It requires a large amount of 3D ground truth data (e.g., ScanNet [19] and KITTI [22], [23]). 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]).

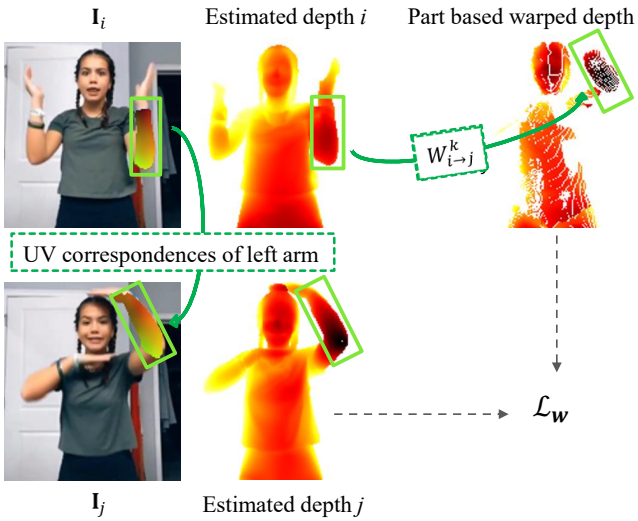


Fig. 3: Given the depth estimate at the i^{th} time instant, we use a part based transformation that warps the 3D local geometry of the image to the image at the j^{th} time instant. The green boxes in two images show the UV correspondences of the left arm. The depths of the left arm are reconstructed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

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.,

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

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the camera intrinsic parameter, $\widetilde{\cdot} \in \mathbb{P}^2$ is the homogeneous representation [27], 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.

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

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

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 parametric 3D transformation such as affine or rigid, e.g., lower arm. An analogous warping is used for non-rigid tracking [43] 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 an affine transformation, i.e., $\mathcal{W}^k_{i\to j}(\mathbf{p}_i) = \mathbf{A}^k_{i\to j}\mathbf{p}_i + \mathbf{t}^k_{i\to j}$ where $\mathbf{A}$ is a 3×3 nonsingular matrix, and $\mathbf{t}$ is a 3×1 translational vector. With the pre-defined correspondences, we compute the transformation by minimizing the following error:

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

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 3 illustrates the self-supervision via warping the 3D geometry of humans between two arbitrary frames of a video. We use the UV coordinates to warp the estimated depth for each body part from the i^{th} time instant to the j^{th} time instant, resulting in a sparse warped depth 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.

With the warped geometry, we can further enforce photometric consistency, i.e.,

$$\mathcal{L}_p = \sum_l \sum_{(i,j)\in\mathcal{V}_l} \sum_k \sum_{\mathbf{u}\in\mathcal{U}_k} \|\mathbf{I}_j(\mathbf{x}_{i\to j}(\mathbf{u})) - \mathbf{I}_i(h_i(\mathbf{u}))\|^2, \quad (4)$$

where $\widetilde{\mathbf{x}}_{i\to j}(\mathbf{u}) = \lambda\mathbf{K}\mathbf{p}_{i\to j}(\mathbf{u})$ is the 2D projection of $\mathbf{p}_{i\to j}(\mathbf{u})$ to the j^{th} time instant.

Equation (3) and (4) 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 normals are known to be highly correlated with the local texture, wrinkle, and shade [56], [60] because of its first order nature of pixel intensity, i.e., under Lambertian lighting model, the pixel intensity is linear in the surface normal. We jointly estimate surface normals and depths 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.,

$$\widehat{\mathbf{n}}(\mathbf{x}) = \frac{\partial\mathbf{p}(\mathbf{x})}{\partial x} \times \frac{\partial\mathbf{p}(\mathbf{x})}{\partial y} \Big/ \left\| \frac{\partial\mathbf{p}(\mathbf{x})}{\partial x} \times \frac{\partial\mathbf{p}(\mathbf{x})}{\partial y} \right\|, \quad (5)$$

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

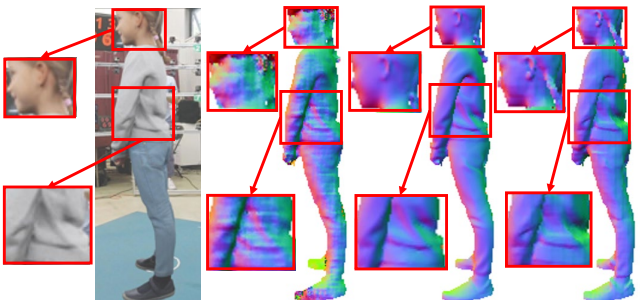


Fig. 4: We compare our method with Tang et al. [56] on the surface normals derived from the depths. While two methods use the surface normals to enhance the depths, unlike Tang et al., our method jointly learns surface normals and depths by supervising them with each other, which produces more realistic and less noisy prediction that preserves the detailed geometry of wrinkles and face.

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

$$\mathcal{L}_s = \sum_{\mathbf{I}_i\in\mathcal{D}} \sum_{\mathbf{x}\in\mathcal{R}(\mathbf{I}_i)} \cos^{-1}\left(\frac{\mathbf{n}^{\mathsf{T}}(\mathbf{x})\widehat{\mathbf{n}}(\mathbf{x})}{\|\mathbf{n}(\mathbf{x})\|\|\widehat{\mathbf{n}}(\mathbf{x})\|} \right), \quad (6)$$

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 [48] has leveraged the derived surface normals from the predicted depths to refine the predicted surface normals for an indoor scene understanding. In human domain, Tang et al. [56] uses the surface normal prediction to refine the human depth prediction 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 4 illustrates the comparison of the surface normal generated from the predicted depth of our method and Tang et al. [56]. Our result is realistic, which captures the wrinkles of the cloth fabric compared to Tang et al. [56].

3.3 Network Design

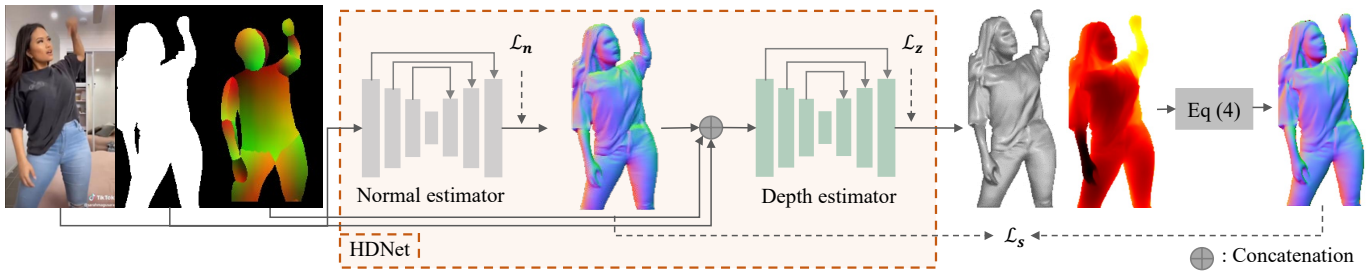
We minimize the following overall loss to learn the depth and surface normal estimators from real videos and 3D scanned models:

$$\mathcal{L} = \mathcal{L}_z + \lambda_n\mathcal{L}_n + \lambda_s\mathcal{L}_s + \lambda_w\mathcal{L}_w + \lambda_p\mathcal{L}_p, \quad (7)$$

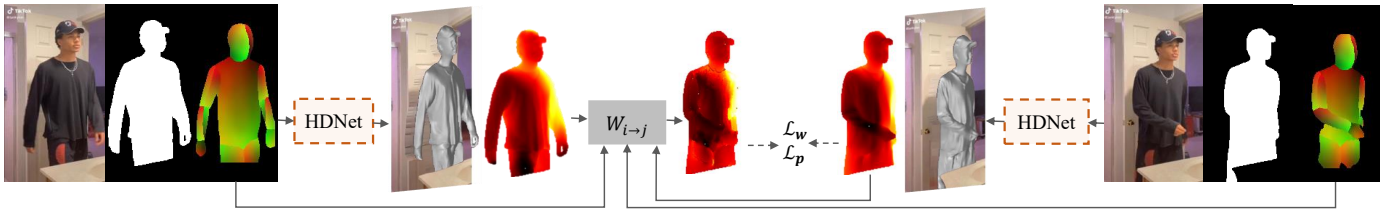
where λ_n, λ_s, λ_w, and λ_p are relative weights between losses. In addition to self-consistency losses (λ_w and λ_s), we utilize the 3D ground truth data from the 3D scanned models [2]. This depth and surface normal can be learned by minimizing the following error between ground truth normal $\mathbf{N}(\mathbf{x})$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

$$\mathcal{L}_z = \sum_{\mathbf{I}\in\mathcal{D}_s} \sum_{\mathbf{x}\in\mathcal{R}(\mathbf{I})} \|Z(\mathbf{x}) - g(\mathbf{x};\mathbf{I})\|^2, \quad (8)$$

$$\mathcal{L}_n = \sum_{\mathbf{I}\in\mathcal{D}_s} \sum_{\mathbf{x}\in\mathcal{R}(\mathbf{I})} \cos^{-1}\left(\frac{\mathbf{N}^{\mathsf{T}}(\mathbf{x})f(\mathbf{x};\mathbf{I})}{\|\mathbf{N}(\mathbf{x})\|\|f(\mathbf{x};\mathbf{I})\|} \right), \quad (9)$$



(a) Network design of HDNet.



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

Fig. 5: (a) Our network *HDNet* takes as input an image with the correspondending human foreground and UV coordinates and predicts the high fidelity depths of the dressed human. The HDNet is composed of the depth and surface normal estimators. The surface normal estimator takes as input, an image and its foreground human mask and outputs the surface normals. The estimated surface normals are, 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 and photometric consistency between the predicted depths and warped depths through $\mathcal{L}_w$ and $\mathcal{L}_p$ respectively.

Fig. 6: TikTok Dataset. We present a new dataset called *TikTok dataset* that consists of 340 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.

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 5(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 Figure 5(b). The UV coordinates from both images are used to compute the affine 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 the time instances that are at least 5 frames apart.

For the two estimators, we use the stacked hourglasses network [44] 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 [25]. We use Adam optimizer [31] 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; $\lambda_{\mathbf{w}}$: 5, and $\lambda_{\mathbf{p}}$: 5; GPU model: NVIDIA V100.

Noisy UV Filtering For every pair of images with the DensePose correspondence, we evaluate the validity of correspondences based on two criteria. (1) The pair must share at least five visible body

Fig. 7: The theoretical bound of depth prediction with respect to uncertainty on Tang et al. dataset [56]. From left to right, we show (1) the depth prediction error as a function of reconstruction uncertainty on torso, head, arms, and legs area, which shows that the performance of the depth prediction by self-supervision is bounded by the uncertainty, (2) the first, middle, and last frames of the sequence, (3) depth error heatmap for Ex1 (small motion) and Ex6 (large motion), and (4) the uncertainty heatmap for Ex1 and Ex6.

parts where each common body part is defined by the one with more than 50 overlapping UV correspondences. (2) We choose the pairs to be at least 5 frames apart to ensure sufficient motion between frames by assuming that the noise in the prediction is not coherent across time. For minor spurious correspondences, (i.e. some noisy pixels in a specific body part) it is handled by the least squares warping solving.

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 6. 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.

5 Theoretical Analysis on Self-supervised Learning via Reconstruction Uncertainty

We formulate our self-supervised learning based on the assumption that the depth predictions in different time instances can be complementary to each other, i.e., a depth prediction in one frame can provide a new information to that of another frame. This assumption applies a majority of our dance videos while there are a few trivial cases where the assumption does not apply effectively. One extreme case would be a sequence of static pose (no motion). Since the depth prediction from one frame does not provide any new information to other frames, the self-supervised learning must be not effective. We characterize the impact of the body motion with respect to the depth prediction using an uncertainty analysis. This analysis allows us to anticipate the impact of the self-supervised learning without training as it provides theoretical performance bound of the self-supervised learning.

Uncertainty Modeling: Consider a 3D point $\mathbf{X}_i \in \mathbb{R}^3$ reconstructed by the i^{th} time instant. The uncertainty of the point can be modeled by a covariance matrix $\mathbf{C}_\mathbf{X} \in \mathbb{R}^{3\times3}$ where the covariance specifies the direction of uncertainty, i.e., the singular vectors and values. The 3D point is visible in the j^{th} time instant, which can be mapped to the i^{th} time instance to form $\mathbf{X}_{j\to i}$. This allows us to find the expected value and its covariance:

$$\mathbf{C}_\mathbf{X} = \mathbb{E}[(\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i])(\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i])^\mathsf{T}], \tag{10}$$

$$\text{where} \quad \mathbb{E}[\mathbf{X}_i] = \frac{1}{T}\sum_{j=1}^{T} \mathbf{X}_{j\to i}, \tag{11}$$

where $\mathbb{E}[\mathbf{X}]$ is the expected value of $\mathbf{X}$. The covariance of the 3D point can be measured by:

$$u(\mathbf{X}) = \sum_k \sigma_k, \tag{12}$$

$$\text{where} \quad \mathbf{C}_\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{U}\begin{bmatrix}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\\ 0 & 0 & \sigma_3\end{bmatrix}\mathbf{V}^\mathsf{T}, \tag{13}$$

where $u(\mathbf{X})$ is the uncertainty of $\mathbf{X}$ that is a sum of singular values of the covariance matrix, and $\mathbf{C}_\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$ is its singular value decomposition. In practice, we initialize our covariance matrix with small covariance on x-y direction (5 pixel) while with large variance on z-direction (e.g., 1E+5).

The major implication of this covariance analysis is a theoretical characterization of self-supervised learning. The uncertainty is high if there is small body motion as the z directional uncertainty is high. The uncertainty can be only reduced as the body move significantly where the body parts must be seen from different viewing angles. Since our self-supervised learning leverages the expectation of the 3D point $\mathbb{E}[\mathbf{X}_i]$ transformed by different time instances, the uncertainty provides a theoretical performance bound (lower bound) of self-supervised learning. This further indicates the effectiveness of self-supervised learning can be *predicted* without training as the uncertainty analysis does not requires training. For instance, a sequence of a static pose would be nearly infinite along the z direction and therefore, it can be predicted that there will be no improvement with self-supervised learning.

Experimental Validation: We compare the reconstruction uncertainty in relation with the depth error using short sequences of the Tang et. al. dataset [56]: Ex1 made of frame 1-5, Ex2 made of 1-10, Ex3 made of 1-15, Ex4 made of 1-20, Ex5 made of 1-25, and Ex6 made of 1-30. Ex1 is a short sequence with small body displacement that is expected to be high uncertainty while Ex6 is a long sequence with full body rotation that is expected to be low

	Tang et al. dataset [56]				RenderPeople dataset [2]				Vlasic et al. dataset [59]				THuman2.0 dataset [64]			
Method	D. error	3cm	4cm	5cm	D. error	3cm	4cm	5cm	D. error	14	18	22	D. error	3cm	4cm	5cm
Li et al. [34]	6.1±3.2	4%	23%	48%	6.4±4.1	8%	28%	46%	37.6±13.7	1%	4%	8%	8.2±4.4	3%	11%	22%
Tang et al. [56]	**4.9±7.1**	**41%**	**65%**	**80%**	6.9±2.8	2%	11%	28%	27.1±7.9	1%	9%	27%	9.0±4.3	1%	6%	15%
PIFu [51]	6.3±3.4	5%	22%	46%	5.3±2.6	17%	35%	54%	30.3±6.6	0%	1%	8%	7.3±3.3	3%	11%	26%
PIFuHD [52]	5.5±3.0	11%	37%	57%	5.6±2.4	10%	29%	48%	27.3±6.6	1%	7%	23%	7.8±3.9	3%	11%	24%
PaMIR [67]	5.4±3.0	12%	37%	58%	5.6±2.0	5%	21%	42%	22.1±6.3	9%	27%	53%	7.2±3.2	3%	13%	27%
Ours [28] (affine)	**4.8±2.9**	**26%**	**52%**	**66%**	**3.2±1.1**	**46%**	**80%**	**95%**	**16.8±5.2**	32%	**63%**	**84%**	**5.1±2.5**	**18%**	**38%**	**57%**
Ours [28] (rigid)	5.0±3.0	25%	49%	65%	**3.2±1.1**	**46%**	**80%**	**95%**	16.9±5.1	**33%**	61%	**84%**	**5.1±2.5**	**17%**	37%	**57%**
Ours [28] + $\mathcal{L}_p$	5.1±2.9	20%	47%	64%	**2.9±1.0**	**56%**	**88%**	**97%**	**15.9±5.4**	**42%**	**69%**	**87%**	**5.2±2.6**	**17%**	**39%**	**57%**

TABLE 1: Quantitative results on the depth prediction. We report the depth error and the percentage of test samples having an error less than three error tolerances (3cm, 4cm, and 5cm) except for Vlasic et al. [59] . All the errors are reported in centimeter (cm), except for Vlasic et al. dataset [59] for which the conversion to metric scale is not known and the reported numbers are in their scale. The best and the second best methods are marked as **red bold** and **blue bold**, respectively.

	Tang et al. dataset [56]				RenderPeople dataset [2]				Vlasic et al. dataset [59]				THuman2.0 dataset [64]			
Method	N. error	25°	30°	35°	N. error	25°	30°	35°	N. error	25°	30°	35°	N. error	25°	30°	35°
Li et al. [34]	33±4	0%	19%	72%	28±7	40%	66%	84%	43±8	2%	7%	20%	32±7	17%	**42%**	67%
Tang et al. [56]	31±7	**16%**	54%	78%	35±5	2%	19%	53%	40±7	0%	4%	22%	39±7	1%	9%	28%
PIFu [51]	33±5	1%	35%	68%	25±5	**51%**	**79%**	**95%**	38±7	1%	10%	34%	**32±6**	12%	40%	69%
PIFuHD [52]	34±5	0%	21%	58%	27±6	34%	65%	86%	47±6	0%	0%	2%	35±8	8%	25%	49%
PaMIR [67]	34±5	0%	15%	58%	29±4	18%	60%	87%	33±5	4%	29%	62%	**32±6**	9%	40%	71%
Ours [28] (affine)	**29±4**	8%	**67%**	**89%**	**16±2**	**100%**	**100%**	**100%**	**24±4**	**59%**	**87%**	**97%**	**22±5**	**75%**	**92%**	**97%**
Ours [28] (rigid)	**29±4**	**9%**	**65%**	**86%**	**16±2**	**100%**	**100%**	**100%**	**25±4**	**52%**	**82%**	**96%**	**22±5**	**74%**	**92%**	**98%**
Ours [28] + $\mathcal{L}_p$	**30±4**	5%	60%	84%	**15±2**	**100%**	**100%**	**100%**	25±5	49%	80%	95%	**22±5**	**74%**	**92%**	**98%**

TABLE 2: Quantitative results on surface normal estimated from the depth prediction. We report the normal error and the percentage of test samples having an error less than three error tolerances (25°, 30°, and 35°). All the errors are reported in degree (°). The best and the second best methods are marked as **red bold** and **blue bold**, respectively.

	Tang et al. dataset [56]				RenderPeople dataset [2]				Vlasic et al. dataset [59]				THuman2.0 dataset [64]			
Method	R. error	3cm	4cm	5cm	R. error	3cm	4cm	5cm	R. error	14	18	22	R. error	3cm	4cm	5cm
Li et al. [34]	5.4±2.8	5%	35%	59%	5.1±3.0	16%	42%	65%	27.1±10.5	4%	15%	34%	6.9±3.6	7%	18%	34%
Tang et al. [56]	4.6±6.8	**47%**	**71%**	**83%**	5.8±2.3	6%	19%	42%	22.9±7.2	6%	25%	54%	7.6±3.6	2%	10%	24%
PIFu [51]	5.6±3.3	11%	41%	58%	4.2±2.0	29%	53%	72%	22.3±6.9	8%	30%	53%	6.2±2.9	7%	23%	41%
PIFuHD [52]	4.9±2.8	21%	49%	65%	4.3±1.7	24%	51%	73%	21.5±5.8	7%	29%	59%	6.5±3.3	6%	22%	38%
PaMIR [67]	4.9±2.9	21%	48%	67%	4.7±1.5	10%	33%	62%	17.0±5.5	32%	64%	83%	6.0±2.8	7%	23%	41%
Ours [28] (affine)	**4.4±2.6**	31%	**59%**	72%	**2.8±0.9**	61%	89%	96%	**12.8±4.4**	**64%**	**88%**	**96%**	**4.4±2.2**	**29%**	**53%**	**71%**
Ours [28] (rigid)	**4.5±2.7**	**33%**	58%	72%	**2.8±0.9**	**63%**	**90%**	**97%**	**13.1±4.6**	62%	**86%**	**95%**	**4.4±2.2**	28%	**52%**	**70%**
Ours [28] + $\mathcal{L}_p$	**4.4±2.6**	32%	**59%**	**73%**	**2.5±0.8**	**71%**	**93%**	**99%**	**12.8±4.7**	**66%**	**86%**	**95%**	**4.4±2.3**	**30%**	**53%**	**70%**

TABLE 3: Quantitative results on surface reconstruction. We report the reconstruction error and the percentage of test samples having an error less than three error tolerances (3cm, 4cm, and 5cm) except for Vlasic et al. [59]. All the errors are reported in centimeter (cm), except for Vlasic et al. dataset [59] for which the conversion to metric scale is not known and the reported numbers are in their scale. The best and the second best methods are marked as **red bold** and **blue bold**, respectively.

uncertainty. In Figure 7, we illustrates the depth error produced by self-supervised learning as a function of the uncertainty for torso, head, arms, and legs. As expected, the depth error after self-supervised learning increases as the uncertainty increases.

6 Experiments

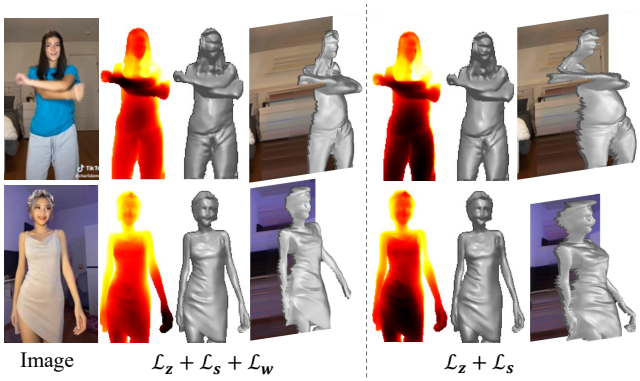
We evaluate our method both quantitatively and qualitatively compared with the state-of-the-art methods of human depth estimation and human shape recovery on real and synthetic data.

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.

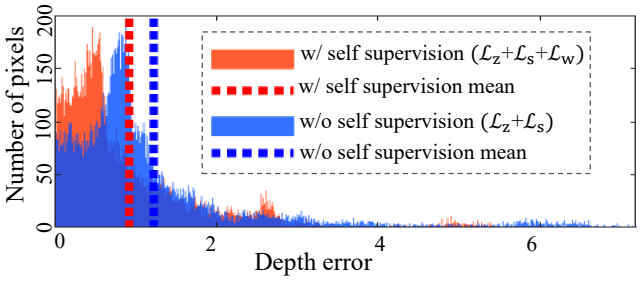
Evaluation Datasets We use four datasets to compare the performance of ours and baseline methods: Tang et al. [56], RenderPeople [2], Vlasic et al. [59], and THuman2.0 [64]. 1) *Training dataset of Tang et al.* This dataset is made of sequences of depth and RGB image pair for 25 subjects. We randomly 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

Losses	D. error	N. error	R. error
$\mathcal{L}_{\mathbf{z}}$ (trained on RenderPeople dataset [2])	5.66±3.85	34.24±5.72	5.17±3.09
$\mathcal{L}_{\mathbf{z}} + \mathcal{L}_{\mathbf{s}}$ (trained on RenderPeople dataset [2])	5.11±3.20	29.99±4.85	4.66±2.82
$\mathcal{L}_{\mathbf{z}} + \mathcal{L}_{\mathbf{s}} + \mathcal{L}_{\mathbf{w}}$ (trained on RenderPeople [2] and TikTok dataset)	**4.89±2.93**	**29.36±4.40**	**4.46±2.65**

TABLE 4: Ablation study on Tang et al. dataset [56]. Here we report the depth error (cm), normal error (°) and reconstruction error (cm) (mean±std).



(a) Impact on self-supervision



(b) Depth error histogram.

Fig. 8: (a) Ablation study on loss functions. From left to right: the image, the full method results, and the results without self-supervision. (b) We show the depth error histogram that illustrates the long tail error distribution of the models without self-supervision.

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. 4) *THuman2.0 dataset* This dataset consists of 526 3D scanned models with texture. We use a ray tracing algorithm to compute the ground truth depth and render the human textured model. We randomly chose 140 3D models and rendered them on 15 cameras around them, to generate totaling 2100 test images.

Evaluation Metric We evaluate the performance in two aspects: (1) accuracy of depths, surface normals, and 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 (Table 1) and surface normal (Table 2), respectively. The surface normals are computed via Equation (5) and compared with the ground truth. In addition, we measure the 3D error by reconstructing 3D point cloud from the estimated depths. To handle unknown scale and focal length, we estimate a relative transformation between the estimated 3D point cloud and the ground truth to measure the shape error, i.e., the estimated point cloud is translated to the median of ground truth and scaled to match the minimum/maximum point cloud distance. The reconstruction error is computed using mean square error (Table 3).

6.1 Quantitative Evaluation

We followed the evaluation protocol of Li et al. [34], i.e., no retraining of the baseline models. We categorize the baseline methods into two: human depth estimation [34], [56], and human shape recovery [51], [52], [67]. The quantitative comparison is summarized in Table 1, 2, 3. We report the performance of our method for rigid transformation warping (first row), affine transformation warping (second row), and the effect of photometric consistency ($\mathcal{L}_{\mathbf{p}}$) (last row).

i) *Human shape recovery* We compare our method with non-parametric human shape recovery designed for dressed humans (PIFu [51], PIFuHD [52], and PaMIR [67]) 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. [34] used a large community dataset called MannequinChallenge dataset to train the stacked hourglasses [44], and Tang et al. [56] leveraged surface normals and depths to preserve detailed dressed human shapes. Note that Tang et al. [56] is both trained and tested on the Tang et al. dataset (no testing data are provided). This results in strong performance of Tang et al., which forms an upper bound performance on Tang et al. dataset. Nonetheless, our method without any adaptation to the dataset performs competitively and generalizes well.

As shown in Table 1, 2, 3, affine transformation is more expressive, in general, compared to rigid body transformation. This is due to the ability to model nonrigid transformation to some extent where most body parts undergo nonrigid motion. Therefore, as expected results of the affine transformation generally surpass the rigid transformation. For future research, the warping solving in our framework can be further extended to perspective transformation or even nonlinear fittings for more degree of freedom and a better warping representation.

As reported in Table 1, 2, 3, except in the RenderPeople dataset, the photometric loss ($\mathcal{L}p$) is not effective as other losses. This stems from the fact that human appearance in RenderPeople is well textured compared to other datasets, which can benefit from the photometric coherence.

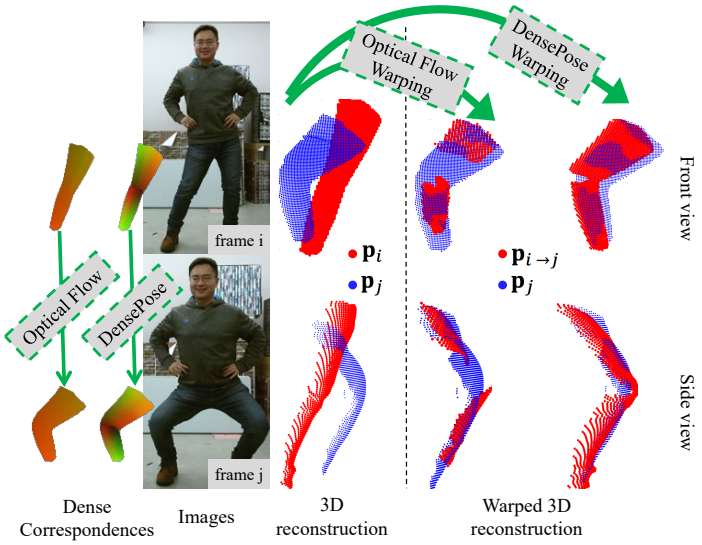


Fig. 9: Comparison between DensePose and optical flow correspondences and their estimated warping on two frames (3 frames apart) of a sequence. From left to right we show (1) the correspondences from frame i to j using optical flow, (2) the correspondences from frame i to j using DensePose, (3) images in frame i and j, (4) the point cloud reconstruction from the predicted depth, in frame i and j, from the front and side view, (5) the warped point cloud from i to j using optical flow correspondences, and point cloud in frame j, and (6) the warped point cloud from i to j using DensePose correspondences, and point cloud in frame j.

6.2 Ablation Study

We conduct an ablation study to analyze the impact of the losses and the usage of TikTok videos 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 [56] without an adaptation for the evaluation. We scale the predicted depths to match to the ground truth, i.e., the predicted depths are translated to the median of ground truth and scaled to match the minimum/maximum depths. Table 4 summarizes the comparison of the combinations. The first two rows in Table 4 is trained on only the RenderPeople dataset as we have the ground truth depth and surface normal. The last row is trained on RenderPeople and the TikTok dataset together. Note that $\mathcal{L}_w$ and $\mathcal{L}_p$ can be only applied to TikTok data to leverage motion. The RenderPeople dataset we had access to was 3D posed data which only captures the human mesh in one frame so we do not have access to any other pose of that mesh; thus, we cannot apply $\mathcal{L}_w$ and $\mathcal{L}_p$ on this dataset. On the one hand, $\mathcal{L}_\mathbf{w}$ enforces the network to learn the geometric consistency from the videos. This loss is highly effective and 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 4).

Our self-supervision makes a positive impact on the plausibility of reconstruction. Without it, the trained model is highly overfitted to the scanned data, which produces unrealistic reconstruction as shown in Figure 8(a). From left to right we have the image, the final method results and the results without self supervision. Without the self supervision, The head is reconstructed far behind the torso mainly due to the small size of the head. As the mean and median errors are not the best descriptive metrics to capture such qualitative plausibility, we further analyze the error by computing its distribution using error histogram shown in Figure 8(b). The self-supervision results in majority of pixels remaining in the lower error regions and a smaller number of pixels in outlier regions (shorter tail error distribution).

Comparison with Optical Flow: To examine the choice of DensePose as the underlying dense correspondences in our framework, we conduct a new comparison of body part warping using DensePose and optical flow. Optical flow that is designed for small pixel displacement fails to make correspondences for far distant frames (e.g., more than 10 frames) while DensePose can be applied regardless of frame distance. Nonetheless, Figure 9 shows the 3D warping based on DensePose and optical flow for the right leg in the Tang et al. dataset (3 frames apart). It illustrates the warping from the DensePose produces accurate alignment of point cloud that can cover the entire leg.

6.3 Qualitative Evaluation

Figure 10 shows the evaluation of our method compared to the baseline methods on TikTok dataset [28]. We get the most representative depth estimation compared to other methods.

We visualize the performance of our method on a set of web images in Figure 11(a). Our method is generalizable to gray scale images, paintings, and images with multiple people.

We also evaluate our method compared to the baselines qualitatively on the evaluation datasets (Figure 12). Figure 12(a) shows the performance of our method and the baselines compared to the ground truth on Tang et al. dataset [56]. Note that Tang et al. [56] was trained on this data. Our method has the most plausible results compared to the baselines on this dataset. Figure 12(b) shows the performance of our method and the baselines compared against the ground truth on RenderPeople dataset [2]. Our prediction is the closest to the ground truth compared to the baseline methods. Figure 12(c) and 12(d) also show that our method outperforms the baselines on Vlasic et al. [59] data and THuman2.0 data [64] respectively.

Handling Large Pose Variation Our method can handle moving body parts, such as arms and legs, that induce significant depth variation across time. Specifically, the 3D translation in $\mathcal{W}^k_{i\to j}$ is designed to account for such changes in depth. Figure 11(b) shows a large depth and pose change of the left leg between frames where the 3D points ($\mathbf{p}_i$) can be correctly transformed to the other frame ($\mathbf{p}_{i\to j} = \mathcal{W}^k_{i\to j}(\mathbf{p}_i)$).

6.4 Failure Cases

As our method takes advantage of in-the-wild internet data, it is robust to the majority of different viewpoints and appearances. However, in some extreme and rare cases, our method fails to predict a realistic human depth. The failure scenarios (shown in Figure 13) are caused by one of these factors: 1) **uncommon camera view:** first row (The left leg is predicted way further from the body and the bottom of the coat is predicted closer to the camera because of the unusual point of view), 2) **extreme lighting and unnatural coloration:** second row (the left lower part of the leg is predicted curved because of the lighting of the room) and third row (the shadow of the hat over the body made the depth prediction of the head and neck unrealistic), 3) **highly crowded texture:** third row and fourth row (the texture on both of the dresses can be miss-identified as shadows in normal estimator and lead to uneven reconstruction), and 4) **occluding accessories:**

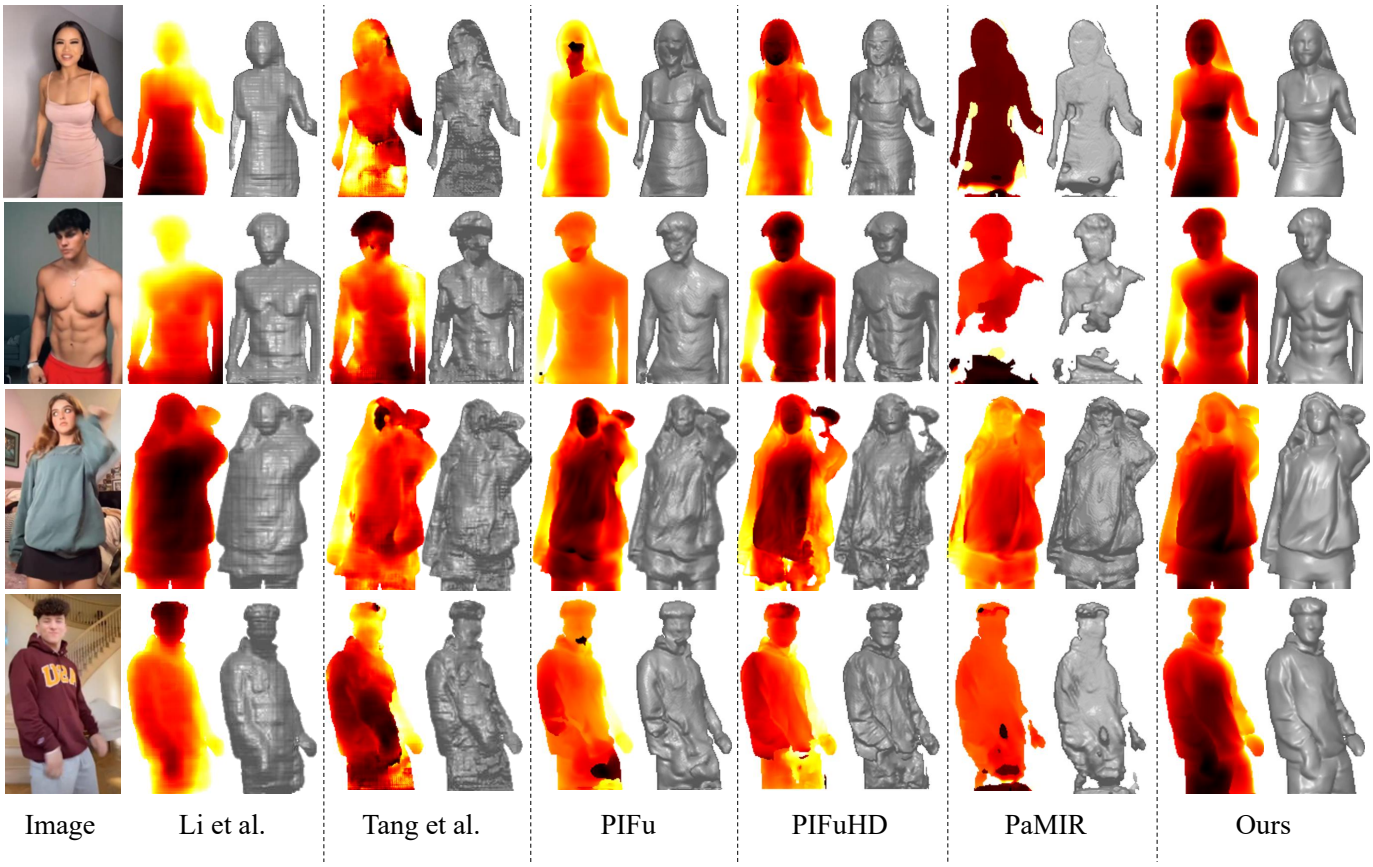
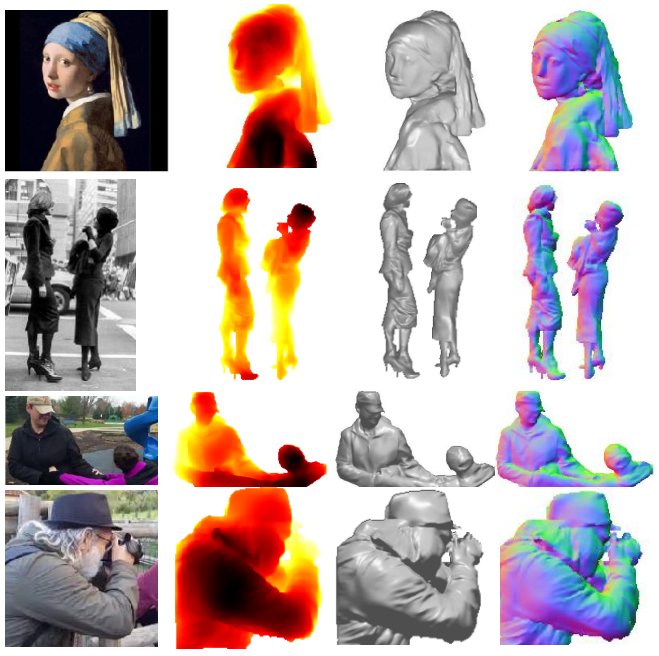
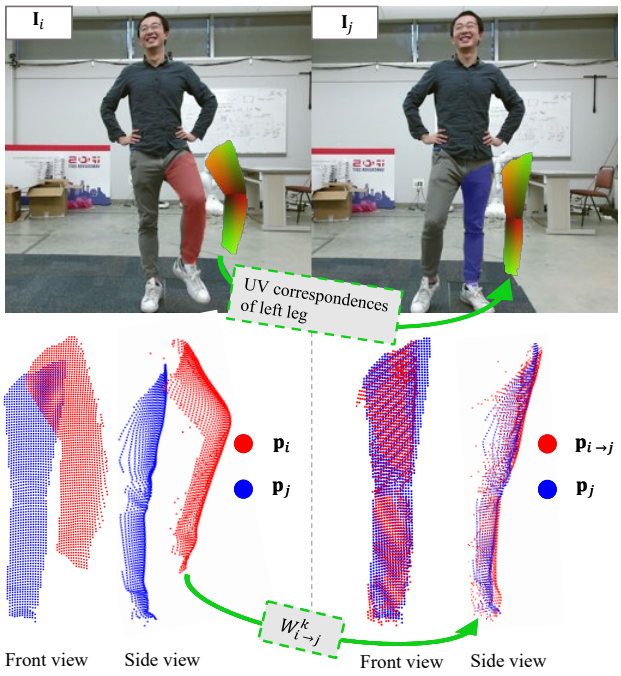


Fig. 10: Qualitative comparison on TikTok dataset [28].

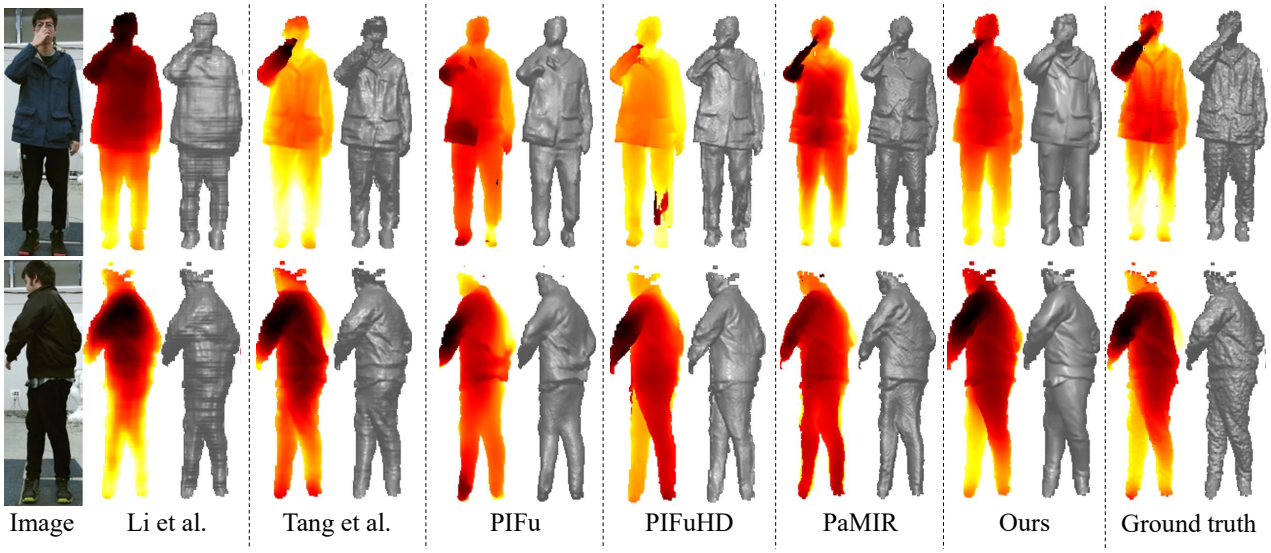


(a) Generalization to internet images

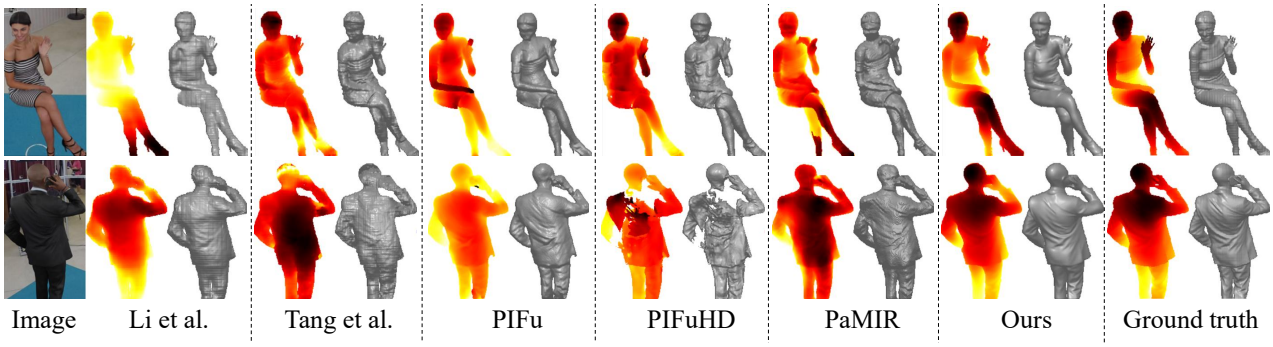


(b) Large pose transformation

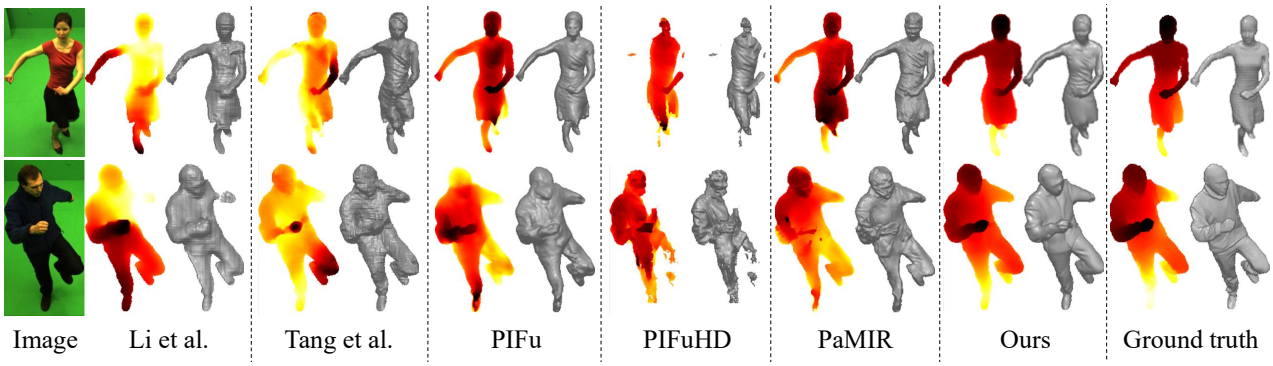
Fig. 11: (a) Qualitative results of our method on web images. From left to right: image, predicted depth, reconstructed surface and surface normal. (b) Our method can handle significant depth change.



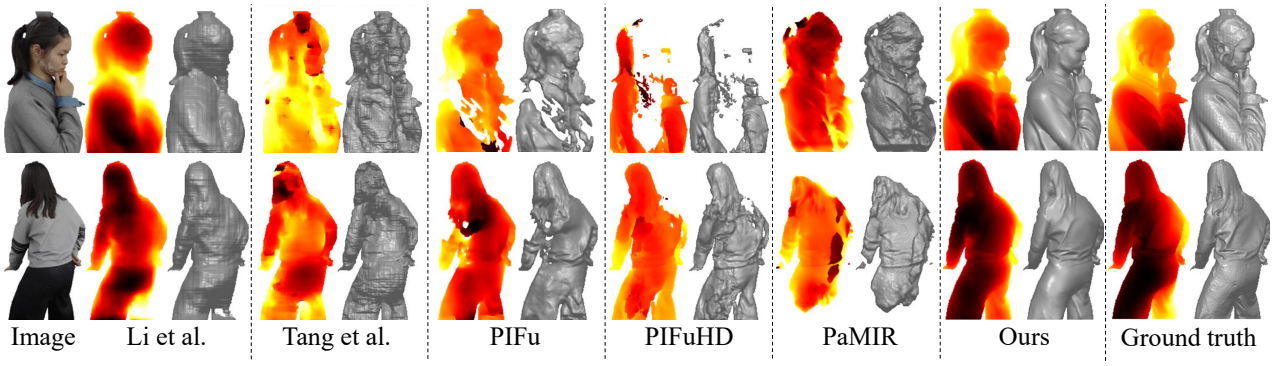
(a) Qualitative evaluation on Tang et al. dataset [56].



(b) Qualitative evaluation on RenderPeople dataset [2].



(c) Qualitative evaluation on Vlasic et al. dataset [59].



(d) Qualitative evaluation on THuman2.0 dataset [64].

Fig. 12: Qualitative results of our method and baselines on 4 different evaluation dataset.

fifth row (the net accessories around the dress and head can cause confusion for the normal estimator in identifying the surface and lead to uneven reconstruction).

7 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 340 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, 7, 8, 9, 11
[3] https://web.twindom.com/. 1, 3
[4] https://www.remove.bg/. 6
[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 *ECCV workshop*, 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. 1, 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 *ICIP*, 2015.
[16] H. Choi, G. Moon, and K. M. Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 1, 2
[17] H. Choi, G. Moon, and K. M. Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021. 1, 2
[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 2001. 2
[19] 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. 1, 3
[20] X. Fei, A. Wong, and S. Soatto. Geo-supervised visual depth prediction. In *ICRA*. 2019. 3
[21] J. Fritsch, T. Kuehnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *ITSC*, 2013. 3
[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 3
[23] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3
[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3
[25] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense human pose estimation in the wild. In *CVPR*, 2018. 5
[26] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 2
[27] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Second edition, 2004. 3
[28] Y. Jafarian and H. S. Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, 2021. 2, 7, 9, 10
[29] H. Joo, H. S. Park, and Y. Sheikh. Map visibility estimation for large-scale dynamic 3D reconstruction. In *CVPR*, 2014. 1
[30] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2
[31] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5
[32] Z. Laehner, D. Cremers, and T. Tung. DeepWrinkles: Accurate and realistic clothing modeling. In *ECCV*, 2020. 2
[33] 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. 1, 2
[34] 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, 7, 8
[35] 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.
[36] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *SIGGRAPH*, 2018. 1
[37] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 1, 2
[38] X. Luo, J. Huang, R. Szeliski, K. Matzen, and J. Kopf. Consistent video depth estimation. In *SIGGRAPH*, 2020. 3
[39] 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
[40] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 3
[41] G. Moon and K. M. Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 1, 2
[42] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. In *SIGGRAPH*, 2005. 3
[43] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 3
[44] A. Newell, K. Yang, and J. Deng. Based on stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 5, 8
[45] 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. 1, 2
[46] R. Or-el, G. Rosman, A. Wetzler, R. Kimmel, and A. Bruckstein. RGBD-Fusion: Real-time high precision depth recovery. In *CVPR*, 2015. 3
[47] 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. 1, 2
[48] 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. 3, 4
[49] 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. 3
[50] 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. 1, 2
[51] 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, 7, 8
[52] 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, 7, 8
[53] 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.
[54] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3
[55] 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
[56] 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, 3, 4, 6, 7, 8, 9, 11
[57] 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
[58] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev,

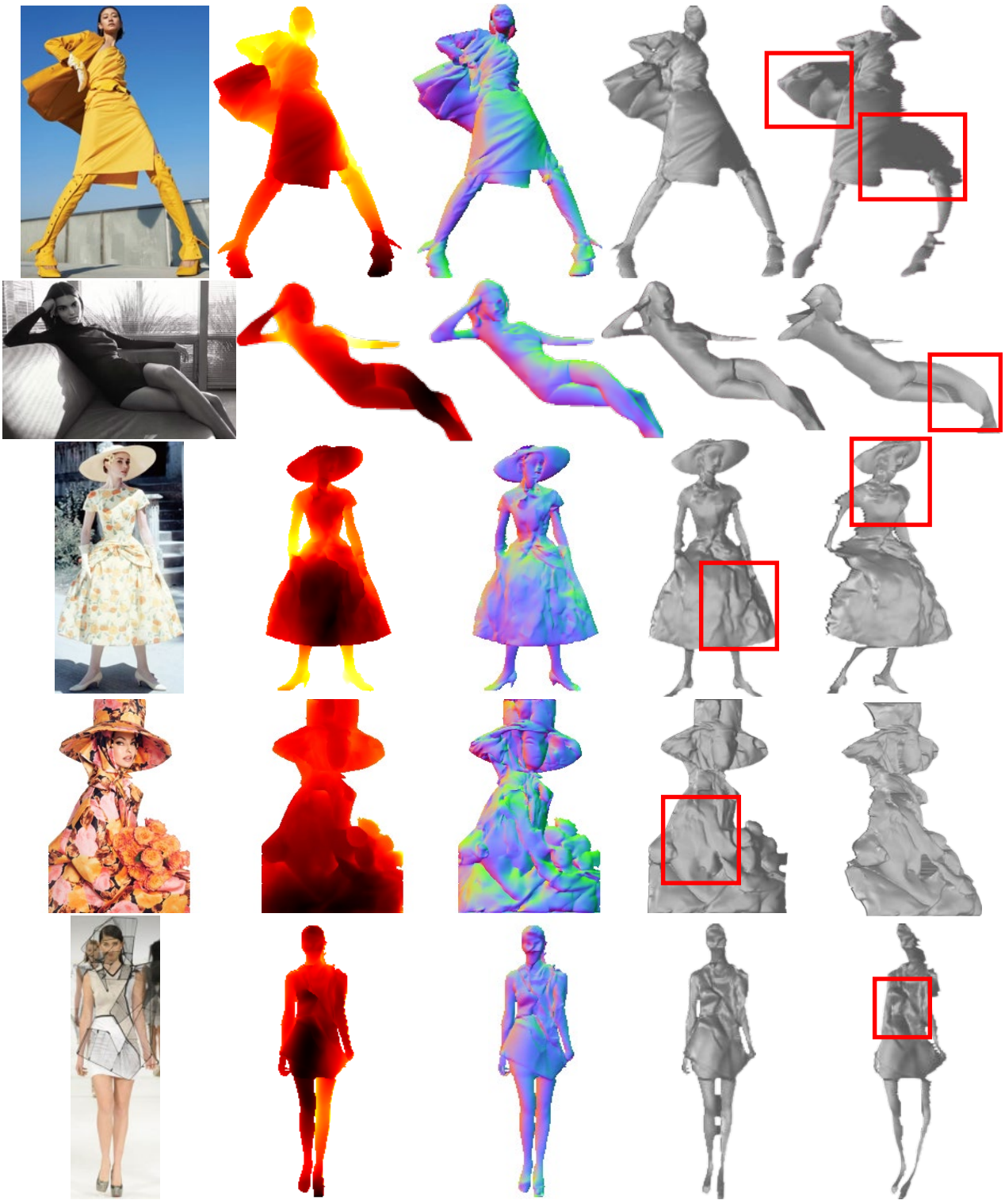
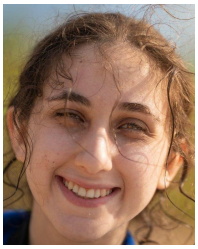
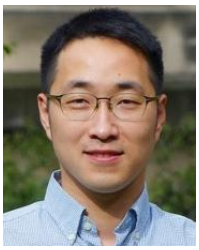


Fig. 13: Failure cases. Here we show the input image, predicted depth, surface normal, and the reconstructed 3D surface from the front and side view. Despite robustness in most of the images in the wild, in some extreme cases our method can fail to predict a realistic depth. These failure scenarios can be categorized as: 1) **uncommon camera view:** first row (*credit: Max Mara Clothing on Instagram*), 2) **extreme lighting and unnatural coloration:** second row (*credit: Kendall Jenner on Instagram*) and third row (*credit: Audrey Hepburn in FUNNY FACE 1956 Paramount film*), 3) **highly crowded texture:** third row and fourth row (*credit: Linda Evangelista's British Vogue Interview: The Rebirth Of The Indomitable Super*), and 4) **occluding accessories:** fifth row (*credit: Collection by Richard Sun of UCA Rochester. Graduate Fashion Week 2012 at London's Earl's Court.*)

and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2
[59] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. 2008. 3, 7, 9, 11
[60] 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, 3, 4
[61] 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
[62] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 1, 2
[63] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *ICCV*, 2013. 3
[64] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 1, 2, 7, 9, 11
[65] 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
[66] Y. Zhang and T. Funkhouser. Deep depth completion of a single rgb-d image. In *CVPR*, 2018. 3
[67] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2021. 2, 7, 8
[68] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. DeepHuman: 3D human reconstruction from a single image. In *ICCV*, 2019. 2
[69] T. Zhu, P. Karlsson, and C. Bregler. Simpose: Effectively learning densepose and surface normal of people from simulated data. In *ECCV*, 2020. 3



Yasamin Jafarian Yasamin Jafarian is a Ph.D. candidate in the Department of Computer Science and Engineering at the University of Minnesota. She is interested in understanding the high fidelity 3D human geometry from single view images using self-supervised learning. She received the CVPR 2021 Best Paper Honorable Mention Award.



Hyun Soo Park Hyun Soo Park is an assistant professor in the Department of Computer Science and Engineering at the University of Minnesota. He is interested in modeling human and animal behaviors. Prior to joining the UMN, he was a postdoctoral fellow in the GRASP Lab at the University of Pennsylvania, and earned his Ph.D. from Carnegie Mellon University. He received NSF CAREER Award (2019) and CVPR 2021 Best Paper Honorable Mention.