CGI2023 manuscript No.
(will be inserted by the editor)

Monocular Human Depth Estimation with 3D Motion Flow and Surface Normals

Yuanzhen Li · Fei Luo* · Chunxia Xiao*

Abstract We propose a novel monocular human depth estimation method using video sequences as training data. We jointly train the depth and 3D motion flow networks with photometric and 3D geometric consistency constraints. Instead of depth ground truth, we take the surface normal as the pseudo label to supervise the depth network learning. The estimated depth may exist texture copy artifact when the clothes on the human body have patterns and text marks (non-dominant color). Thus, we also propose an approach to alleviate the texture copy problem by estimating and adjusting the color of non-dominant color areas. Extensive experiments on public datasets and the Internet have been conducted. The comparison results prove that our method can produce competitive human depth estimation and has better generalization ability than the-state-of-art methods.

1 Introduction

Estimating human depth from one or a few 2D images is very useful for many applications, including AR/VR, teleconference, and virtual digital human creation in the recently emerging metaverse. The human body can be represented with parametric or non-parametric models. Parametric-based methods usually use the SCAPE [5] and SMPL [32] to represent the naked body shape, but it is hard for them to create 3D geometric details, like hair and clothes wrinkles. Non-parametric models based on depth [19], voxel [14], and implicit function representation [41] can produce more human details.

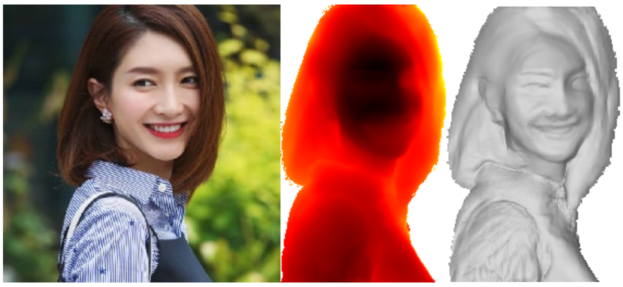

Fig. 1 An example of human depth estimation and corresponding mesh from a single image by our method. From left to right: input image, estimated depth, and the 3D mesh corresponding to the depth.

Compared to other non-parametric models, depth information is more convenient to acquire and store. Some methods [45, 19] use neural networks to estimate human depth from a single image by using depth ground truth to supervise training. However, high-precision depth scanners [54] are still not widely available in practice. Collecting a large amount of depth ground truth for humans with diverse appearances, clothes, and poses is challenging. Until now, only a few hundred human-scanned models have been freely made available in academia and industries. Thus, the supervised depth estimation methods are difficult to generalize. Furthermore, the predicted depth is relative, and it has an unknown scale and shift with regard to the depth ground truth.

To solve the problem of hard-to-obtain ground truth data, some methods [8, 58, 39] use pseudo labels instead of ground truth data. The surface normal is perpendicular to the surface at a given 3D point, which can

Yuanzhen Li
yuanzhen@whu.edu.cn

Fei Luo
luofei@whu.edu.cn

Chunxia Xiao
cxxiao@whu.edu.cn

*Corresponding author: Chunxia Xiao and Fei Luo.
School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China.

present more fine geometric details. Using the surface normal to supervise the depth estimation can avoid the impact of the unknown scale and shift. Therefore, we use surface normal as the pseudo label to supervise the depth learning indirectly.

As video data is easier to obtain and more information can be learned from consecutive frames, we use video sequences as the training data. Due to the non-rigid transformation of human motion, we train one more 3D motion flow network to enforce the accuracy of the depth network. Based on the estimated depth and 3D motion flow, we establish the photometric and 3D geometric consistencies between adjacent frames.

When human clothes exhibit patterns and text marks, the estimated depth may suffer from the texture copy problem. We design a linear transformation and image inpainting approach to alter the pattern and text color close to the domain color, which could alleviate the texture copy problem.

In this work, we propose a novel monocular human depth estimation method by exploiting the surface normal and 3D motion flow to supervise depth estimation. Meanwhile, we propose an approach to alleviate the texture copy problem. We perform quantitative and qualitative experimental evaluations and ablation study experiments on various datasets to verify the effectiveness of our method. In summary, we make the following contributions:

- We propose a novel human depth estimation method by jointly learning the depth and 3D motion flow.
- Instead of depth ground truth, we propose to use the surface normal as the pseudo label to supervise the depth estimation model.
- To alleviate the texture copy artifact, we develop a color component analysis and color transformation approach to deal with the colors on clothes.

2 Related Work

2.1 Human Body Reconstruction

In computer graphics, the depth representation is referred to as the 2.5D model [57]. Depth information of human is usually used to reconstruct the geometric surface of the human body [34]. The 3D human body can be represented by the parametric model or the non-parametric model. Parametric methods such as SCAPE [5], SMPL [32], and SMPL-X [37] treat the human body reconstruction as the determination of the pose and shape parameters. Although parametric human body shapes are readily applicable [38,20], the reconstructed geometry is difficult to represent the fine details of dressed humans with hair and loose clothes. Some methods [4,24,25] add residual geometry to parametric models. However, those methods are not powerful enough for non-rigid clothes reconstruction.

Non-parametric representations can describe the geometric surface details of dressed humans, such as depth [45,19], voxel grids representation [14], and implicit function representation [41,52]. Zheng et al. [61] and Habermann et al. [14] used voxel to reconstruct human shapes. The voxel grid processing requires intensive memory, and the results have limited resolution. Saito et al. [41] used a pixel-aligned implicit function to reconstruct the 3D human from a single image. Afterward, they used normal maps to improve the 3D geometric details [42] (PIFuHD). Zheng et al. [60] proposed a method called PaMIR to improve the robustness of the 3D model with the SMPL mesh. Xiu et al. [52] proposed a local-feature-based implicit 3D reconstruction method called ICON to improve the robustness of human pose estimation. They used the SMPL to guide normal prediction and regressed the occupancy field. The above methods based on implicit function representation are computationally heavy and time-consuming. Feng et al. [9] proposed an effective and flexible 3D geometry representation of the Fourier Occupancy Field (FOF) for the monocular real-time human reconstruction.

Compared to other non-parametric models, depth is more flexible to integrate other information to reconstruct the geometric surface. Li et al. [28] utilized the motion parallax from static scenes to guide the human depth estimation. Tang et al. [45] proposed a supervised human depth estimation method by incorporating pose and semantic labels. To improve the generalization ability of the human depth estimation, Jafarian and Park [19] proposed a semi-supervised human depth and normal estimation method. They assumed human motion to be a local rigid transformation and based on the DensePose maps [13] to calculate it. DensePose represents dense correspondence from 2D images to 3D surface-based representations of the human body, which regresses part-specific UV coordinates. The DensePose map has three channels.

Actually, human motion is a non-rigid transformation when the body is dressed in loose clothes. Thus, we propose to learn the 3D motion flow of the human between adjacent frames to establish the photometric and 3D geometry consistency losses. The difference between our method and the method of Jafarian and Park [19] are as follows. (1) instead of depth ground truth, we use the surface normals as the pseudo labels to supervise the depth learning. (2) we jointly learn the human depth and 3D motion flow to establish the photometric and 3D geometry consistency between adjacent frames.

2.2 Self-supervised Monocular Depth Estimation

Without needing depth ground truth, self-supervised monocular depth estimation has extracted a lot of attention. Stereo image pairs and video frames can train self-supervised depth estimation. Stereo image pair training utilizes stereo images with a known baseline distance between the left and right cameras [11,23,26]. Monocular video sequences training jointly estimates the depth and the camera pose on adjacent frames [62,12,27,33]. In our human depth estimation, a non-rigid human moves in front of a fixed camera. We jointly estimate the human depth and 3D motion flow on adjacent frames.

2.3 Scene Flow

Optical flow represents the motion direction and quantity of a 2D image object movement between two consecutive frames [3,56,44]. 3D Scene flow is a three-dimensional motion field of each point in the scene [48]. The 3D scene flow is widely used in video tracking and monitoring. According to the type of input data, various scene flow estimation methods have been proposed, such as stereo images [17,43], 3D point clouds [51,50,31], or a sequence of RGB-D images [46,18,53]. Huguet and Devernay [17] proposed a traditional method that used standard variational formulations and energy minimization to estimate scene flow from stereo images. Wei et al. [51] proposed a point-voxel recurrent field transform method to estimate scene flow from point clouds. Hur and Roth [18] proposed a monocular scene flow and depth estimation method with a single neural network to estimate the depth and 3D motion flow and adopted self-supervised learning with 3D loss functions.

Li et al. [29] estimated a scene flow field to design a dynamic scenes neural radiance field framework [35]. Zhang et al. [59] jointly trained a scene flow network and a pre-trained depth network in each video fragment to generate temporally consistent depth for arbitrarily moving objects.

3 Method

As shown in Fig. 1, given a single image of dressed human I, our goal is to estimate the human depth D. The overview of our method is shown in Fig. 2, which consists of two sub-modules: depth estimation and 3D motion flow estimation. The depth estimation network takes a single image of a dressed human as input and predicts the human depth. The 3D motion flow estimation network takes color images $\{I_i, I_j\}$ and DensePose maps $\{S_i, S_j\}$ of two adjacent frames as input and predicts the human 3D motion flow $F_{i\to j}$. Based on the estimated depth and 3D motion flow, we establish the photometric and 3D geometric consistency constraints. At the same time, we use the surface normal as the pseudo label to supervise the depth estimation network learning. In the inference stage, the two networks are applied independently.

When the clothes on the body have patterns and text marks, it could lead to the texture copy problem on depth estimation. Based on the connected domain size of the marks, we design a linear transformation and image inpainting method to reduce the color difference, and alleviate the texture copy problem (see Fig. 5). In the following subsections, we will describe the human depth estimation method and the approach to alleviate the texture copy problem.

3.1 Human Depth Estimation

The overview of our human depth estimation method is shown in Fig. 2. The depth estimation network G_{d} takes the color image I_i as input and predicts the human depth D_i. The 3D motion flow estimation network G_{f} takes the two frames' color image $\{I_i, I_j\}$ and the DensePose $\{S_i, S_j\}$ as input and predicts the 3D motion flow $F_{i\to j}$ of the human. Based on the estimated 3D motion flow and depth, we establish the photometric and 3D geometric consistency constraints. Meanwhile, we use the normal as the pseudo label instead of depth information to supervise the human depth estimation network learning.

Our objective loss for optimizing the depth estimation network G_{d} and the 3D motion flow estimation network G_{f} consisting of five terms: normal depth consistency loss $\mathcal{L}_{\mathrm{n}}$, photometric reprojection loss $\mathcal{L}_{\mathrm{c}}$, 3D geometric consistency loss $\mathcal{L}_{\mathrm{g}}$, depth smoothness loss $\mathcal{L}_{\mathrm{s}}$, and 3D flow loss $\mathcal{L}_{\mathrm{f}}$. Subsequently, we will detailedly describe each loss term.

Normal depth consistency loss Surface normal $N(x)$ of a 2D pixel coordinate x is the curvature that is perpendicular to the tangential plane of the corresponding 3D point $P(x)$:

$$N(x) = \left(\frac{\partial P(x)}{\partial \mathrm{x}} \times \frac{\partial P(x)}{\partial \mathrm{y}}\right) / \left(\left\|\frac{\partial P(x)}{\partial \mathrm{x}}\right\| \times \left\|\frac{\partial P(x)}{\partial \mathrm{y}}\right\|\right), \tag{1}$$

where $P(x) = D(x)K^{-1}\hat{x}$, $\hat{x}$ is the 2D homogeneous coordinate of pixel x. The intrinsic parameters matrix of the camera $K \in R^{3\times 3}$ is calibrated in our experiment. We constrain the geometric consistency between the pseudo labeled surface normal N and the derived surface normal $\widetilde{N}$ from the predicted depth D in Eq.

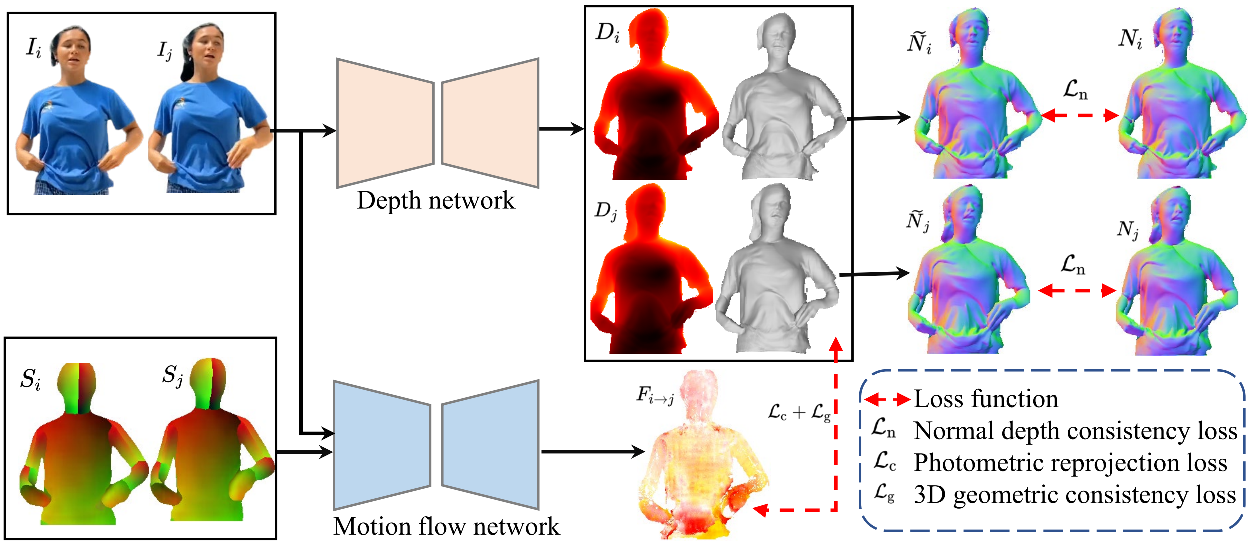
I_i
I_j
Depth network
D_i
D_j
$\widetilde{N}_i$
N_i
$\mathcal{L}_n$
$\widetilde{N}_j$
N_j
$\mathcal{L}_n$
S_i
S_j
Motion flow network
$F_{i\to j}$
$\mathcal{L}_c + \mathcal{L}_g$
Loss function
$\mathcal{L}_n$ Normal depth consistency loss
$\mathcal{L}_c$ Photometric reprojection loss
$\mathcal{L}_g$ 3D geometric consistency loss

Fig. 2 Overview of our training framework. It consists of two major modules. The depth estimation network takes a color frame I_i as input and predicts the depth D_i. The 3D motion flow network takes two adjacent frames $\{I_i, I_j\}$ and the corresponding DensePose $\{S_i, S_j\}$ as input and predicts the human 3D motion flow $F_{i\to j}$. Based on the estimated depth $\{D_i, D_j\}$ and 3D motion flow $F_{i\to j}$, we establish the photometric and 3D geometric consistency constraints to supervise the two networks. Meanwhile, the precomputed surface normal is used as the pseudo label to supervise the depth network learning.

(1):

$$\mathcal{L}_\text{n} = \frac{1}{k}\sum_{x\in I}\cos^{-1}\left(\frac{N^\text{T}(x)\widetilde{N}(x)}{\|N(x)\|\times\|\widetilde{N}(x)\|}\right), \tag{2}$$

where k is the number of pixels. In our experiment, we use the estimated surface normal from method [19] as the pseudo label N.

Photometric reprojection loss We assume that the illumination information is constant in all trained video data. The photometric reprojection loss $\mathcal{L}_\text{c}$ penalizes the photometric difference between the target frame I_i and the reconstructed target frame $I_{j\to i}$. The reconstructed image $I_{j\to i}$ is synthesized from the reference frame I_j, the predicted depth D_i, and the 3D motion flow $F_{i\to j}$. We back-project each pixel x on the target frame I_i into 3D space point $P_i(x)$, and translate it to the 3D point $P_{i\to j}(x)$ corresponding to the reference frame I_j with the estimated 3D motion flow:

$$P_{i\to j}(x) = P_i(x) + F_{i\to j}(x). \tag{3}$$

The 3D point $P_{i\to j}(x)$ is projected into the image plane: $x' = KP_{i\to j}(x)$. If the depth D_i and 3D motion flow $F_{i\to j}$ are accurate, the pixels x and x' have the same color. We use the bilinear interpolation to sample the reference frame I_j and synthesize the reconstructed image $I_{j\to i}$. Inspired by [11], we use a combination of the $L1$ and single scale SSIM [49] term as our photometric reprojection loss $\mathcal{L}_\text{c}$:

$$\begin{aligned}\mathcal{L}_\text{c} = \frac{1}{k}\sum_{x\in I}\frac{\mu}{2}(1 - \text{SSIM}(I_i(x), I_{j\to i}(x)))+ \\ (1-\mu)\left\|I_i(x) - I_{j\to i}(x)\right\|_1,\end{aligned} \tag{4}$$

where $\|.\|_1$ denotes the $L1$ norm, μ is set to 0.85, and SSIM() denotes the structure similarity index measure which is computed over a 3×3 block filter.

3D geometric consistency loss The 3D geometric consistency loss $\mathcal{L}_\text{g}$ penalizes the difference between the translated 3D point $P_{i\to j}(x)$ from the target frame I_i and the 3D point $P_j(x')$ corresponding to the matched pixel point x' in the reference frame I_j:

$$\mathcal{L}_\text{g} = \frac{1}{k}\sum_{x\in I}\|P_{i\to j}(x) - P_j(x')\|_2^2, \tag{5}$$

where the matched 3D point $P_j(x') = D_j(x')K^{-1}\widehat{x'}$. Fig. 3 shows the photometric reprojection loss $\mathcal{L}_\text{c}$ and 3D geometric consistency loss $\mathcal{L}_\text{g}$.

Depth smoothness loss As in [11], we use edge-aware depth smoothness loss weighted by image gradients to encourage locally smooth constraint of depth D:

$$\mathcal{L}_\text{s} = \frac{1}{k}\sum_{x\in I}\left|\frac{\partial D(x)}{\partial \text{x}}\right|e^{-\|\frac{\partial I(x)}{\partial \text{x}}\|} + \left|\frac{\partial D(x)}{\partial \text{y}}\right|e^{-\|\frac{\partial I(x)}{\partial \text{y}}\|}. \tag{6}$$

3D flow loss We penalize the difference between the 3D flow $F_{i\to j}$ and the 2D flow from the corresponding DensePose maps. We first calculate the 2D flow

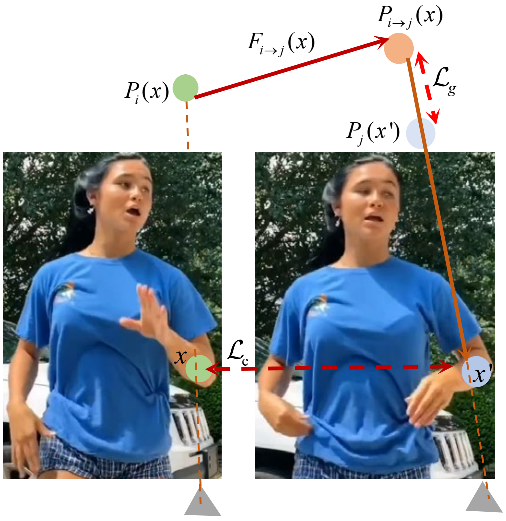

Fig. 3 The photometric reprojection loss $\mathcal{L}_{\mathrm{c}}$ and 3D geometric consistency loss $\mathcal{L}_{\mathrm{g}}$. The pixel x in image I_i is back-projected into a 3D point $P_i(x)$ using the predicted depth $D_i(x)$. The 3D points $P_i(x)$ is translated to point $P_{i\to j}(x)$ at time j using the predicted 3D motion flow $F_{i\to j}(x)$. In image space, the translated point $P_{i\to j}(x)$ is projected to 2D point x'. The pixel x' is back-projected into the 3D point $P_j(x')$ using the predicted depth $D_j(x)$. The photometric consistency $\mathcal{L}_{\mathrm{c}}$ penalizes the color difference between $I_i(x)$ and $I_j(x')$. The 3D geometric consistency $\mathcal{L}_{\mathrm{g}}$ penalizes the Euclidean distance between the translated 3D point $P_{i\to j}(x)$ and the matched 3D point $P_j(x')$. The black triangle represents the camera position.

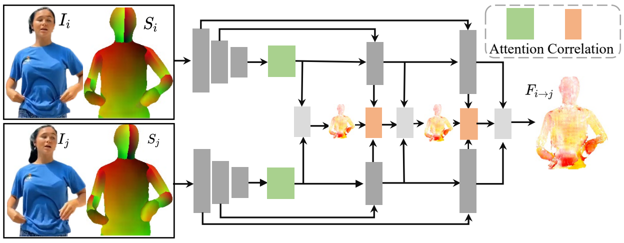

Fig. 4 The 3D motion flow network. It takes a variant of U-Net using the residual blocks in the encoder and decoder with skip connections as the backbone framework of the 3D motion flow G_{f}.

$F^{2D}_{i\to j}$ using the DensePose maps S_i and S_j. Second, we project the estimated 3D motion flow $F_{i\to j}$ into 2D flow $\widetilde{F^{2D}_{i\to j}}$ in 2D image space. Finally, we use the Euclidean distance between the 2D flow $F^{2D}_{i\to j}$ and the projected 2D flow $\widetilde{F^{2D}_{i\to j}}$ as the 3D flow loss:

$$\mathcal{L}_{\mathrm{f}} = \frac{1}{k}\sum_{x\in I}\left\| F^{2D}_{i\to j}(x) - \widetilde{F^{2D}_{i\to j}}(x)\right\|_2^2. \tag{7}$$

To sum up, the total loss function is as follows:

$$\mathcal{L}_{\mathrm{loss}} = \mathcal{L}_{\mathrm{n}} + \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}} + \lambda_{\mathrm{g}}\mathcal{L}_{\mathrm{g}} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{s}} + \lambda_{\mathrm{f}}\mathcal{L}_{\mathrm{f}}. \tag{8}$$

The hyper-parameters λ_{c}, λ_{g}, λ_{s}, and λ_{f} control the relative weights of the different terms. In our experiment, we set $\lambda_{\mathrm{c}} = 0.3$, $\lambda_{\mathrm{g}} = 0.5$, $\lambda_{\mathrm{s}} = 0.1$, and $\lambda_{\mathrm{f}} = 0.1$.

3.2 Network Details

Depth network We use the stacked hourglass network [36] as the backbone of the depth estimation network G_{d}. Jafarian and Park [19] also used it as their depth estimation network. The difference between ours and theirs is that our network does not directly output a continuous depth. The output of our depth network is pixel-wise disparity probability, called discrete disparity volume [27]. Concretely, the depth network outputs an M channel disparity probability volume $\{V_1, ..., V_M\}$ with M disparity layers. The corresponding disparity value d_m in the disparity layer m $(m = 1, 2, ..., M)$ is:

$$d_m = d_{\min} + \Delta_d \times (m-1), \tag{9}$$

where $d_{\min}$ and Δ_d are the minimum disparity value and disparity interval. In our experiment $M = 90$, $d_{\min} = 0.001$, and $\Delta_d = 0.01$. A depth-wise softmax operation processes probability volume V_m to produce an actual probability map for each disparity layer $V^d_m = softmax(V_m)$. The disparity by weighting the sum of the disparity probability volume:

$$\sigma = \sum_{m=1}^{M} d_m V^d_m. \tag{10}$$

We set the minimum depth as 2.0 and maximum depth as 4.0, and convert the disparity σ to the depth [12].

Motion flow network As shown in Fig. 4, we take a variant of U-Net [40] using the residual blocks [16] in the encoder and decoder with skip connections as the backbone framework of the 3D motion flow G_{f}. We input the color image and DensePose $\{I_i, S_i\}$, $\{I_j, S_j\}$ into the ResNet50 network and it outputs three feature layers. We use a self-attention module [47] to learn correlation features of the global context in the last features layer. We concatenate the two self-attention features and yield the low-resolution 3D motion flow by a convolutional layer with a filter. The low-resolution 3D motion flow is the first step of the multi-scale decoder. We adopt the feature correlation layer for the two additional stages to make the network stronger regulation [18]. The feature correlation is based on a matching score to quantize the feature similarity between images. We use a bilinear up-sampling operator to up-sample the low-resolution features, concatenate with the encoded features and correlation features as the next layers input, and output the high-resolution 3D motion flow.

3.3 Texture Copy

When the clothes exhibit patterns and text marks, it is possible to lead to texture copy phenomenon on depth estimation or 3D human reconstruction (see Fig. 6 and Fig. 7). We define the above areas as non-dominant color areas. We propose a method to transfer the non-dominant color to be similar to the dominant color, aiming to alleviate the texture copy. It was noted that we need to retain the wrinkle details of clothes when the non-dominant color is transferred.

As shown in Fig. 5, our method is based on a linear color transformation operator and image inpainting algorithm. First, we use the body parsing algorithm [30] to segment the clothes. Second, we divide the colors of the clothes into H groups and utilize the k-means clustering algorithm [22] to extract H areas with their central color values $\bar{C}_h^t$, where $t = \{1, 2, 3\}$ and $h = \{0, 1, ..., H-1\}$. The central color values $\bar{C}_h^t$ are normalization (0,1). The maximum area is the dominant color area C_0 with the central color value $\bar{C}_0$. For one of the three color channels t, when the difference between the central value $\bar{C}_h^t$ of the area C_h and the central values $\bar{C}_0^t$ of the dominant color is more than a threshold ϑ:

$$|\bar{C}_0^t - \bar{C}_h^t| > \vartheta, \tag{11}$$

the area C_h is the non-dominant color area, and its color needs to be transferred. In our experiment, we set $\vartheta = 0.1$. Based on the constraint, we extract non-dominant color area masks. To avoid boundary traces in color transformation, we use the guided filter [15] to dilate the masks.

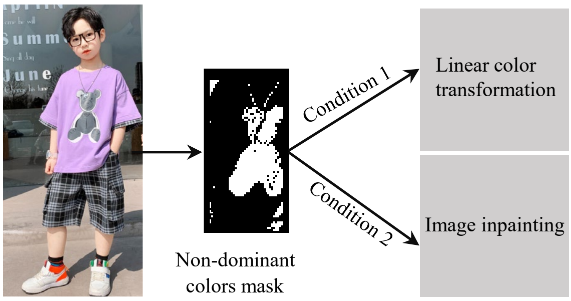

Fig. 5 Texture copy alleviation. First, we use the k-means clustering algorithm to extract non-dominant color areas. Then, we design a linear transformation and image inpainting algorithm to alter their color close to the dominant color based on the connected domain size of masks, respectively.

Based on the connected domain size of the non-dominant color mask, we design a linear color transformation operator and image inpainting to transfer the non-dominant color, respectively. When the connected domain of the non-dominant color mask is less than a threshold, we use the image inpainting algorithm [7] to fill the non-dominant color area (see Fig. 6).

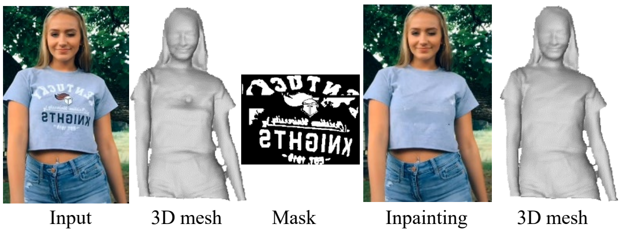

Fig. 6 Texture copy alleviation by the image inpainting. From left to right are the source image, the 3D mesh of the source image, the non-dominant color mask, the inpainted image, and the 3D mesh of the inpainted image.

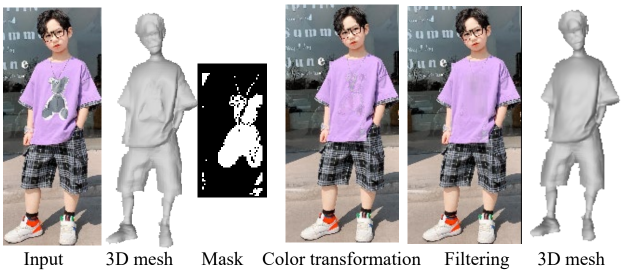

Fig. 7 Texture copy alleviation by the color transformation. From left to right are the source image, the 3D mesh of the source image,the non-dominant color mask, the color transformed image, the filtered image, and the 3D mesh of the filtered image.

When the connected domain of the non-dominant color mask is greater than a threshold, the wrinkle details of the clothes could be damaged when we use the image inpainting algorithm. Thus, we design a linear operator to transfer the color of the non-dominant color area to be close to the dominant color (see Fig. 7). First, we compute the mean between the dominant color central value and the non-dominant color central value:

$$\bar{C}_{0h} = (\bar{C}_0 + \bar{C}_h)/2. \tag{12}$$

Then, the color C_h of the non-dominant color area subtracts the mean value $\bar{C}_{0h}$ as its new color:

$$C_h = |C_h - \bar{C}_{0h}|. \tag{13}$$

The new color C_h can be close to the dominant color value and retain the wrinkle details of the clothes. Finally, we use the image filter algorithm [10] to smooth the edge area of the mask. Fig. 6 and Fig. 7 demonstrate that the images pre-processed by our method have less texture copy problem compared to the source results.

4 Experiments

To validate the effectiveness of our human depth estimation method, we compare it with existing methods and conduct ablation study experiments. We implement our method in the TensorFlow framework with a single NVIDIA GeForce RTX 3090Ti GPU. The batch size is set to 8, and the number of epochs is 200. We use the Adam Optimizer [21] with $\alpha = 0.0001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ to train.

4.1 Training Datasets and Evaluation Metrics

Training dataset Our training dataset is taken from Jafarian and Park [19], which consists of more than 300 sequences of dance videos shared on the TikTok social media mobile platform. We remove images with blurred human motion, crop and resize our training images to 256×256, utilize the method [1] to extract the human mask, and utilize the method [13] to obtain the corresponding DensePose map. Jafarian and Park [19] have calibrated the intrinsic parameters matrix of the camera K corresponding to the training dataset.

Evaluation metrics Similar to Jafarian and Park [19], we evaluate the performance of our human depth estimation method by measuring the accuracy of the predicted depth and the corresponding 3D point cloud. We use the mean squared error as a metric for the depth error. We reconstruct the 3D point cloud from the estimated depth and compute the mean square error as a reconstruction error. Due to the influence of depth scale and viewpoint, the two sets of corresponding 3D point cloud need to be aligned. Therefore, we estimate a relative rotation and translation between two sets of corresponding 3D point clouds based on least-squares [6]. The estimated point cloud is translated to the median of ground truth and scaled to match the minimum-maximum point cloud distance [2]. At the same time, we also compare the accuracy at different error thresholds, i.e., the percentage of pixels with an error smaller than some thresholds.

4.2 Quantitative and Qualitative Evaluation

Quantitative evaluation We quantitatively evaluate our human depth estimation method on Tang et al. [45] and THuman2.0 [55] datasets. Tang et al. provided an RGBD dataset with 25 subjects, and we randomly choose around 3000 frames as the test data. THuman2.0 dataset consists of 526 3D human models with texture, and we use a ray-tracing algorithm to render RGBD images on ten camera poses as our test data.

We compare our method with the state-of-the-art methods [19,45,42,60,52,9], and the above models are not retrained. The compared methods can be categorized into human depth estimation [45,19] and non-parametric human shape recovery [42,60,52,9]. The human shape recovery methods PIFuHD [42], PaMIR [60], ICON [52], and FOF [9] predict implicit function representation containing the back side 3D surface of a human body from a single image. We only compare the error of the input image region. We use a ray-tracing algorithm to identify the front surface and render the corresponding depth.

Table 1 and Table 2 report the quantitative comparison results. The quantitative results of our method are stable on two test datasets. Even though the compared methods use depth ground truth as the supervised information, our method produces a competitive result without depth ground truth. The quantitative results of our method and method [19] are quite close. However, the network input of our method is a single image, and we do not need ground truth depth to supervise. From the two tables, our method is effective.

Qualitative evaluation We qualitatively evaluate our depth estimation method on the following datasets: the above two quantitative evaluation datasets, the TikTok dataset (not in our training dataset), captured by a smartphone and from the Internet. We present the qualitatively compared results in Fig. 8, 9, 10, and 11. We only pre-process the second input image of Fig. 10, with the proposed linear color transformation to alter the non-dominant color.

As shown in Fig. 8, Tang et al. [45] can estimate a good depth on Tang et al. dataset. In contrast, the qualitative results are poor on other datasets (see Fig. 9-Fig. 11). In Fig. 9, the above non-parametric human shape recovery methods can produce good results because the dataset belongs to their training dataset. As shown in Fig. 10 and Fig. 11, when the human is not complete in the input image, the predicted human shape easily produces a distortion by PaMIR [60] and FOF [9]. As show in Fig. 8-Fig. 11, methods [42,60,9] are sensitive to pose. ICON [52] has good robustness in various postures, but the predicted human shape still needs to improve the detailed shape.

As shown in Fig. 10 and Fig. 11, compared with method [19], our reconstructed surface is more continuous, especially for the human arm. Jafarian and Park [19] predicted depth appears depth drifting. For our capturing image and Internet image shown in Fig. 11, our method has produced better results, which proves our better generalization ability than the others. Our method demonstrates a competitive performance in the quantitative and qualitative evaluations.

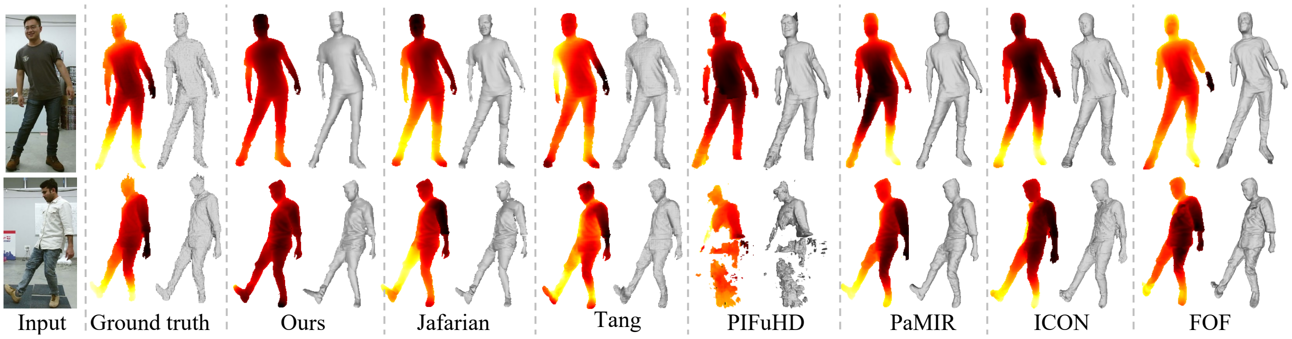

Fig. 8 Qualitative comparison with existing methods on Tang et al. dataset [45]. From left to right, these images are the input image, ground truth, results of our method, Jafarian [19], Tang et al. [45], PIFuHD [42], PaMIR [60], ICON [52], and FOF [9].

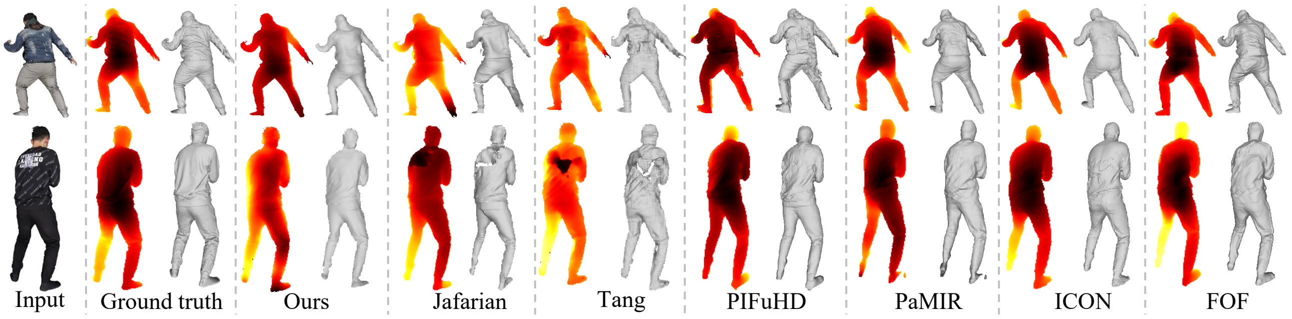

Fig. 9 Qualitative comparison with existing methods on THuman2.0 dataset [55]. From left to right, these images are the input image, ground truth, results of our method, Jafarian [19], Tang et al. [45], PIFuHD [42], PaMIR [60], ICON [52], and FOF [9].

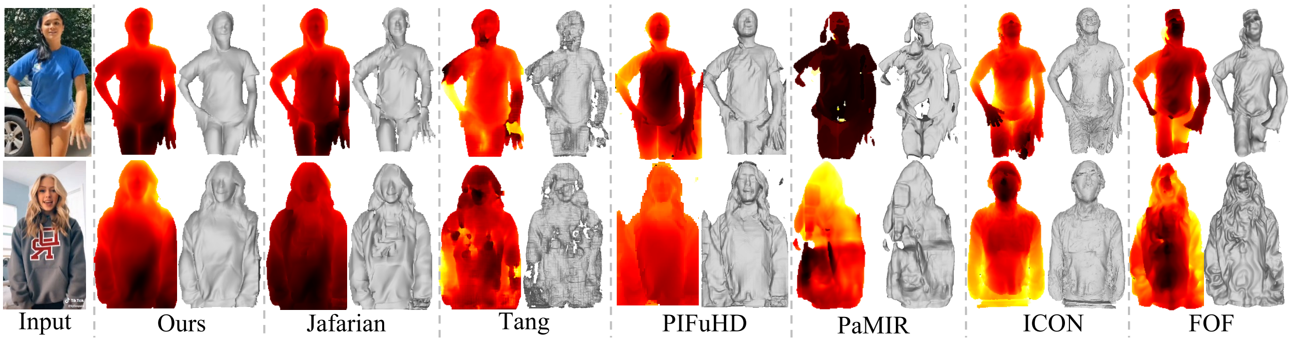

Fig. 10 Qualitative comparison with existing methods on TikTok dataset [19]. From left to right, these images are the input image, results of our method, Jafarian [19], Tang et al. [45], PIFuHD [42], PaMIR [60], ICON [52], and FOF [9].

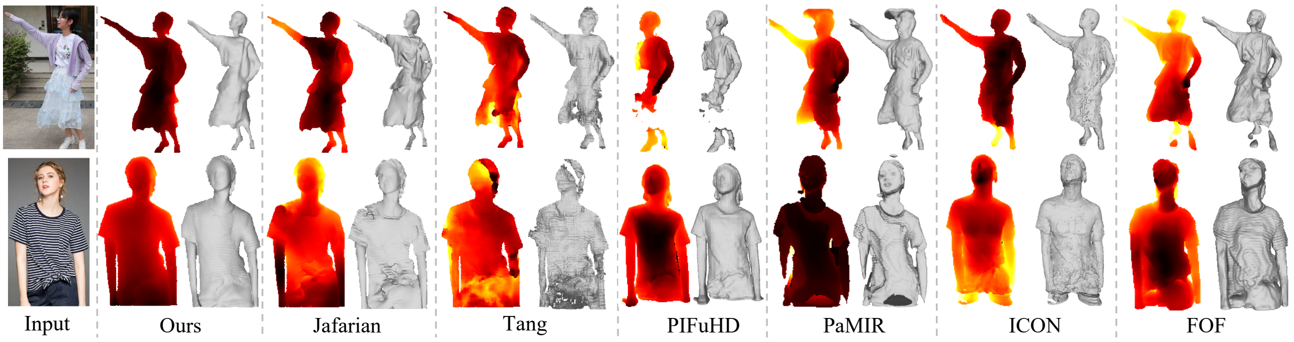

Fig. 11 Qualitative comparison with existing methods on our data and the Internet. From left to right, these images are the input image, results of our method, Jafarian [19], Tang et al. [45], PIFuHD [42], PaMIR [60], ICON [52], and FOF [9].

Table 1 Quantitative result on the depth estimation. The depth error and the percentage of test samples having an error of less than three error tolerances (3.0 cm, 4.0 cm, and 5.0 cm) on Tang et al. [45] and THuman2.0 [55] datasets. All the errors are reported in centimeter (cm). D. error represents the depth error. The best and second-best results are marked as **bold** and underline on each dataset, respectively.

	Tang et al. dataset				THuman2.0 dataset			
Method	D. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑	D. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑
Tang et al. [45]	5.1±4.2	**38%**	**63%**	**77%**	10.8±5.2	1%	5%	14%
Jafarian and Park [19]	**5.1±3.4**	23%	48%	61%	**5.4±2.2**	**15%**	34%	53%
Ours	5.2±1.4	25%	51%	65%	5.5±1.6	**15%**	**36%**	**55%**
PIFuHD [42]	6.0 ± 1.4	8%	30%	50%	7.3±1.6	3%	12%	22%
PaMIR [60]	6.1 ± 2.6	7%	28%	47%	7.3±2.5	2%	12%	24%
ICON [52]	5.7±4.2	17%	38%	53%	5.8±3.2	12%	23%	47%
FOF [9]	5.8±2.8	16%	36%	54%	5.7±2.5	14%	25%	49%

Table 2 Quantitative evaluation on surface reconstruction. The reconstruction error and the percentage of test samples having an error of less than three error tolerances (3.0 cm, 4.0 cm, and 5.0 cm) on Tang et al.[45] and THuman2.0 [55] datasets. R. error represents the reconstruction error.

	Tang et al. dataset				THuman2.0 dataset			
Method	R. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑	R. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑
Tang et al. [45]	4.9±5.4	**42%**	**65%**	**76%**	8.2±4.1	1%	8%	20%
Jafarian and Park [19]	**4.6±3.9**	25%	54%	65%	5.0±2.8	26%	**53%**	69%
Ours	4.8±2.1	27%	57%	68%	**4.8±1.0**	**28%**	50%	**72%**
PIFuHD [42]	5.2±3.4	18%	44%	61%	6.7±2.8	7%	21%	35%
PaMIR [60]	5.3±4.2	17%	40%	60%	6.4±3.6	5%	20%	37%
ICON [52]	4.9±2.2	21%	48%	58%	5.2±2.4	19%	40%	59%
FOF [9]	5.1±2.7	20%	47%	60%	5.1±1.3	18%	42%	62%

Table 3 Ablation study on Tang et al. [45] and THuman2.0 [55] datasets, respectively.

(a) The depth error and the percentage of test samples having an error of less than three error tolerances (3.0 cm, 4.0 cm, and 5.0 cm).

	Tang et al. dataset				THuman2.0 dataset			
Method	D. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑	D. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑
w/o 3D motion flow	6.0±2.1	19%	40%	53%	6.9±3.2	12%	30%	49%
with 3D motion flow	**5.2±1.4**	**25%**	**51%**	**65%**	**5.5±1.6**	**15%**	**36%**	**55%**

(b) The reconstruction error and the percentage of test samples having an error of less than three error tolerances (3.0 cm, 4.0 cm, and 5.0 cm).

	Tang et al. dataset				THuman2.0 dataset			
Method	R. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑	R. error ↓	3.0 cm ↑	4.0 cm ↑	5.0 cm ↑
w/o 3D motion flow	5.4±2.6	23%	50%	54%	5.4±2.6	23%	47%	68%
with 3D motion flow	**4.8±2.1**	**27%**	**57%**	**68%**	**4.8±1.0**	**28%**	**50%**	**72%**

4.3 Ablation Study

To validate the effectiveness of the 3D motion flow, we make an ablation study to omit the 3D motion flow from our objective ("w/o 3D motion flow"). We evaluate the above combinations on the Tang et al. [45] and THuman2.0 [55] datasets, respectively. The quantitative comparison result is summarized in Table 3. The 3D motion flow can effectively improve the accuracy of our depth estimation based on the pseudo-labeled surface normal. Various qualitative results show that "w/o 3D motion flow" easily produces distortion. When we jointly train the 3D motion flow, the photometric consistency $\mathcal{L}_\text{c}$ and 3D geometric consistency $\mathcal{L}_\text{g}$ can improve the accuracy of the network.

In Fig. 12, we present two qualitative results of the ablation study. Full supervision results are superior to only using the surface normal as the supervising information. With the cooperative contribution of each component in our method, we have improved the human body 3D structure completeness and decreased the depth distortion on non-wrinkle areas caused by texture copy.

In Fig. 13, we show an example with the 3D mesh from different viewing angles. From the results, we can see that the proposed method is effective. In Fig. 14,

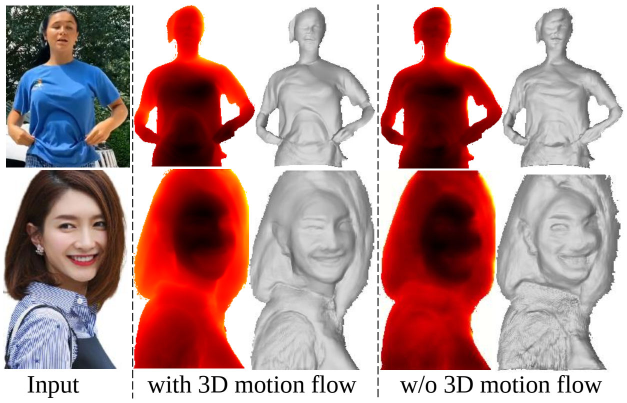

Fig. 12 Ablation study on w/o 3D motion flow. From left to right are the input image, the full method results, and the results without the 3D motion flow to supervise. The two input images are from the TikTok dataset and the Internet.

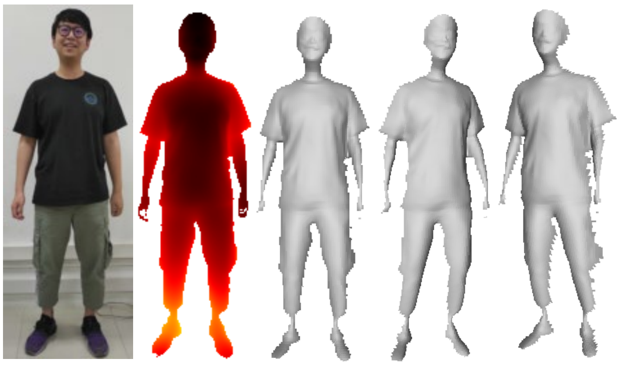

Fig. 13 The 3D meshes from different viewing angles. From left to right are the input image, predicted depth, 3D mesh on the current pose, and 3D meshes on the other two poses. The input image is captured by a smartphone.

we visualize the 2D projection image of the 3D motion flow.

5 Limitations

Our texture copy alleviation method cannot fully solve the texture copy problem. When the non-dominant and dominant color areas have depth differences, the image inpainting algorithm may produce a wrong modification. Our method may be invalid in a case with multiple dominant colors, such as striped shirts.

6 Conclusion

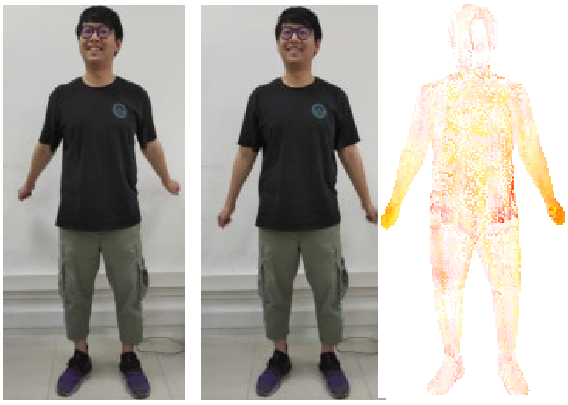

Fig. 14 The 2D projection of the 3D motion flow. From left to right are the target image I_i, the reference image I_j, and the 2D projection of the 3D motion flow $F_{i\rightarrow j}$.

We have proposed a monocular human depth estimation method via jointly learning 3D motion flow. Instead of depth information, we use the surface normal as the pseudo label to supervise the depth network learning. Based on the estimated depth and 3D motion flow, we design photometric consistency and 3D geometric consistency to enforce the accuracy of the depth estimation model. Moreover, to alleviate the texture copy artifact of the pattern and text areas, we design an approach of color component analysis and color transformation. Experiments demonstrate that our method can produce competitive results and has a good generalization ability compared with state-of-the-art methods. We have experimentally validated that 3D motion flow can improve the accuracy of depth estimation.

Acknowledgments

This work is partially supported by NSFC (No.61972298), Bingtuan Science and Technology Program (No.2019BC008), and CAAI-Huawei MindSpore Open Fund.

Data Availability Statement

The data that support the findings of this study are openly available in the public data repository at:
TikTok [19]: https://www.yasamin.page/hdnet_tiktok;
Tan [45]: https://github.com/sfu-gruvi-3dv/deep_human;
THuman2.0 [55]: https://github.com/ytrock/THuman2.0-Dataset.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. `https://www.remove.bg/upload`
2. `http://nghiaho.com/?page_id=671`.
3. Aleotti, F., Poggi, M., Mattoccia, S.: Learning optical flow from still images. In: CVPR, pp. 15,196–15,206 (2021)
4. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: ICCV, pp. 2293–2303 (2019)

5. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. ACM Transactions on Graphics (TOG) **24**(3), 408–416 (2005)
6. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-d point sets. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) (5), 698–700 (1987)
7. Bian, X., Wang, C., Quan, W., Ye, J., Zhang, X., Yan, D.M.: Scene text removal via cascaded text stroke detection and erasing. Computational Visual Media **8**, 273–287 (2022)
8. Chen, Z., Lu, X., Zhang, L., Xiao, C.: Semi-supervised video shadow detection via image-assisted pseudo-label generation. In: ACM MM, pp. 2700–2708 (2022)
9. Feng, Q., Liu, Y., Lai, Y.K., Yang, J., Li, K.: Fof: Learning fourier occupancy field for monocular real-time human reconstruction. In: NeurIPS (2022)
10. Gastal, E.S.L., Oliveira, M.M.: Domain transform for edge-aware image and video processing. ACM Transactions on Graphics (TOG) **30**(4), 1–12 (2011)
11. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR, pp. 270–279 (2017)
12. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV, pp. 3828–3838 (2019)
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR, pp. 7297–7306 (2018)
14. Habermann, M., Xu, W., Zollhofer, M., Pons-Moll, G., Theobalt, C.: Deepcap: Monocular human performance capture using weak supervision. In: CVPR, pp. 5052–5063 (2020)
15. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) **35**(6), 1397–1409 (2012)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
17. Huguet, F., Devernay, F.: A variational method for scene flow estimation from stereo sequences. In: ICCV, pp. 1–7 (2007)
18. Hur, J., Roth, S.: Self-supervised monocular scene flow estimation. In: CVPR, pp. 7396–7405 (2020)
19. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: CVPR, pp. 12,753–12,762 (2021)
20. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR, pp. 7122–7131 (2018)
21. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: ICLR (2015)
22. Krishna, K., Murty, M.N.: Genetic k-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **29**(3), 433–439 (1999)
23. Kuznietsov, Y., Stuckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: CVPR, pp. 6647–6655 (2017)
24. Lahner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: ECCV, pp. 667–684 (2018)
25. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: 3DV, pp. 643–653 (2019)
26. Li, Y., Luo, F., Li, W., Zheng, S., Wu, H.h., Xiao, C.: Self-supervised monocular depth estimation based on image texture detail enhancement. The Visual Computer (TVC) **37**(9), 2567–2580 (2021)
27. Li, Y., Luo, F., Xiao, C.: Self-supervised coarse-to-fine monocular depth estimation using a lightweight attention module. Computational Visual Media **8**(4), 631–647 (2022)
28. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: CVPR, pp. 4521–4530 (2019)
29. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: CVPR, pp. 6498–6508 (2021)
30. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) **41**(4), 871–885 (2018)
31. Liu, X., Qi, C.R., Guibas, L.J.: Flownet3d: Learning scene flow in 3d point clouds. In: CVPR, pp. 529–537 (2019)
32. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015)
33. Luo, F., Wei, L., Xiao, C.: Stable depth estimation within consecutive video frames. In: CGI, pp. 54–66 (2021)
34. Luo, F., Zhu, Y., Fu, Y., Zhou, H., Chen, Z., Xiao, C.: Sparse rgb-d images create a real thing: A flexible voxel based 3d reconstruction pipeline for single object. Visual Informatics **7**(1), 66–76 (2023)
35. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV, pp. 405–421 (2020)
36. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV, pp. 483–499 (2016)
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR, pp. 10,975–10,985 (2019)
38. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR, pp. 10,975–10,985 (2019)
39. Petrovai, A., Nedevschi, S.: Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation. In: CVPR, pp. 1578–1588 (2022)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241 (2015)
41. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV, pp. 2304–2314 (2019)
42. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: CVPR, pp. 84–93 (2020)
43. Schuster, R., Wasenmuller, O., Kuschk, G., Bailer, C., Stricker, D.: Sceneflowfields: Dense interpolation of sparse scene flow correspondences. In: WACV, pp. 1056–1065 (2018)
44. She, D., Xu, K.: An image-to-video model for real-time video enhancement. In: ACM MM, pp. 1837–1846 (2022)
45. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: ICCV, pp. 7750–7759 (2019)
46. Teed, Z., Deng, J.: Raft-3d: Scene flow using rigid-motion embeddings. In: CVPR, pp. 8375–8384 (2021)

47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
48. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. In: ICCV, pp. 722–729 (1999)
49. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing (IEEE TIP) **13**(4), 600–612 (2004)
50. Wang, Z., Li, S., Howard-Jenkins, H., Prisacariu, V., Chen, M.: Flownet3d++: Geometric losses for deep scene flow estimation. In: WACV, pp. 91–98 (2020)
51. Wei, Y., Wang, Z., Rao, Y., Lu, J., Zhou, J.: Pv-raft: point-voxel correlation fields for scene flow estimation of point clouds. In: CVPR, pp. 6954–6963 (2021)
52. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: implicit clothed humans obtained from normals. In: CVPR, pp. 13,286–13,296 (2022)
53. Yang, G., Ramanan, D.: Learning to segment rigid motions from two frames. In: CVPR, pp. 1266–1275 (2021)
54. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: CVPR, pp. 5746–5756 (2021)
55. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: CVPR (2021)
56. Zhang, F., Li, Y., You, S., Fu, Y.: Learning temporal consistency for low light video enhancement from single images. In: CVPR, pp. 4967–4976 (2021)
57. Zhang, W., Yan, Q., Xiao, C.: Detail preserved point cloud completion via separated feature aggregation. In: ECCV, pp. 512–528 (2020)
58. Zhang, X., Ge, Y., Qiao, Y., Li, H.: Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: CVPR, pp. 3436–3445 (2021)
59. Zhang, Z., Cole, F., Tucker, R., Freeman, W.T., Dekel, T.: Consistent depth of moving objects in video. ACM Transactions on Graphics (TOG) **40**(4), 1–12 (2021)
60. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) **44**(6), 3170–3184 (2022)
61. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: CVPR, pp. 7739–7749 (2019)
62. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR, pp. 1851–1858 (2017)