3D Clothed Human Reconstruction in the Wild

Gyeongsik Moon[1*], Hyeongjin Nam[2*], Takaaki Shiratori[1], and Kyoung Mu Lee[2,3]

[1] Meta Reality Labs Research
[2] Dept. of ECE & ASRI, Seoul National University, Korea
[3] IPAI, Seoul National University, Korea
{mks0601,tshiratori}@fb.com, {namhjsnu28,kyoungmu}@snu.ac.kr

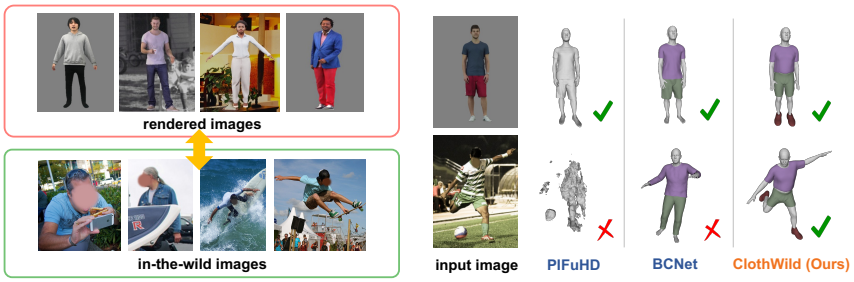
Abstract. Although much progress has been made in 3D clothed human reconstruction, most of the existing methods fail to produce robust results from in-the-wild images, which contain diverse human poses and appearances. This is mainly due to the large domain gap between training datasets and in-the-wild datasets. The training datasets are usually synthetic ones, which contain rendered images from GT 3D scans. However, such datasets contain simple human poses and less natural image appearances compared to those of real in-the-wild datasets, which makes generalization of it to in-the-wild images extremely challenging. To resolve this issue, in this work, we propose ClothWild, a 3D clothed human reconstruction framework that firstly addresses the robustness on in-the-wild images. First, for the robustness to the domain gap, we propose a weakly supervised pipeline that is trainable with 2D supervision targets of in-the-wild datasets. Second, we design a DensePose-based loss function to reduce ambiguities of the weak supervision. Extensive empirical tests on several public in-the-wild datasets demonstrate that our proposed ClothWild produces much more accurate and robust results than the state-of-the-art methods. The codes are available in here.

1 Introduction

3D clothed human reconstruction aims to reconstruct humans with various poses, body shapes, and clothes in 3D space from a single image. It is an essential task for various applications, such as 3D avatars and virtual try-on. Most of the recent 3D clothed human reconstruction methods [33,37,2,34,15,13,14,6,17,7,3] require 3D scans for the training; hence, they are trained on synthetic datasets, which consist of 3D scans [31,4] and rendered images from the scans. Although significant progress has been made by utilizing such synthetic datasets, all of them fail to produce robust results on in-the-wild images.

The in-the-wild images are taken in our daily environments, such as cluttered offices and concert halls, and have diverse human poses, appearances, and severe human occlusions. On the other hand, the 3D scans are captured from restricted

* equal contribution
This work was primarily done while Gyeongsik Moon was in SNU.



(a) Domain gap (pose, appearance, occlusion, etc) (b) Qualitative comparison

Fig. 1. (a) Domain gap between rendered images and in-the-wild images. (b) Due to the large domain gap, existing methods (*e.g.* PIFuHD [34] and BCNet [17]) fail on in-the-wild images, while our ClothWild successfully reconstructs. Colors of reconstructed 3D clothes are manually assigned to represent cloth types.

environments, such as motion capture studios. To ease 3D scan capturing, subjects are often instructed to take simple poses. Therefore, rendered images from the 3D scans have artificial appearances and simple human poses compared to those of in-the-wild images, as shown in Fig. 1(a). This large domain gap between in-the-wild and rendered images makes the methods trained on synthetic datasets fail to generalize to in-the-wild images, as shown in Fig. 1(b).

For the robustness to the domain gap, we present ClothWild, a 3D clothed human reconstruction framework that leverages a weak-supervision strategy to learn 3D clothed humans from in-the-wild datasets. Here, the weak supervision means that the supervision target is not full 3D data, but 2D data defined only in the single-camera viewpoint (*i.e.*, 2D cloth segmentations [9]). For the weak-supervision strategy, our reconstruction pipeline is divided into two networks: a cloth generative model and a regressor, ClothNet. The cloth generative model generates 3D cloth geometry around the T-posed 3D body of SMPL [23] from the cloth latent codes, where the 3D cloth geometry is animatable using a skinning function of SMPL. It is fully supervised on the synthetic datasets beforehand, and we freeze it during the weakly supervised training. Although the cloth generative model covers various cloth types and styles, it does not cover complicated human poses and diverse image appearances of in-the-wild datasets. For the robustness to such in-the-wild datasets, we design a regressor, called ClothNet, and weakly supervise it on in-the-wild datasets to predict latent codes of the cloth generative model. The final output can be obtained by passing the output of the ClothNet to the cloth generative model. As the cloth generative model is fixed during the weakly supervised training, we can protect the generative model from being hurt by imperfect 2D supervision targets of in-the-wild datasets, while making our ClothNet robust to diverse human poses and image appearances of in-the-wild images. Our weak-supervision strategy can recover full 3D geometry only from the evidence of the single-camera viewpoint by regularizing the predicted cloth latent codes to be in the cloth latent space.

Note that such regularization is not possible for previous approaches [33,34,36,3] as they do not model cloth latent space.

For the weak supervision, we can enforce the projection of 3D reconstruction results to be close to 2D supervision targets (*i.e.*, cloth segmentations [9]) of in-the-wild datasets. However, naively adopting such a strategy has three difficulties in learning 3D clothed humans. First, the 2D supervision targets provide information for only a single camera viewpoint. Therefore, there is no guarantee that the weak supervision can recover 3D geometry of other viewpoints. Second, severe depth ambiguity of the cloth segmentations hampers learning accurate 3D clothes. Since each pixel in cloth segmentations corresponds to innumerable points in 3D space, it is difficult to specify 3D points corresponding to clothes. Third, pixel-level misalignment can occur between projected 3D reconstruction results and 2D cloth segmentation. As camera parameters are not available for in-the-wild images, we should additionally predict the camera parameters to project the 3D clothed humans. This is a highly ill-posed problem as various combinations of camera parameters and 3D clothed humans correspond to the same 2D cloth segmentation. Due to the ill-posedness, camera prediction can often fail, which results in the wrong projection to the 2D image space.

In this regard, we design a DensePose-based loss function to resolve the issues which can occur when weakly supervising 3D clothed human reconstructions based on their projections. DensePose [10] informs where each human pixel corresponds to the 3D human body surface of SMPL. Using DensePose, our DensePose-based loss function obtains 3D points that correspond to the cloth segmentations around the SMPL surface. Then, it enforces such 3D points to be a part of the 3D cloth surface. As the DensePose effectively limits possible positions of 3D points around the SMPL surface, it significantly reduces the depth ambiguity. In addition, DensePose is already aligned with cloth segmentations in the image space; therefore, we do not have to predict camera parameters and project 3D reconstructions to the 2D image space. Hence, our DensePose-based loss function does not suffer from pixel-level misalignment.

We show that the proposed ClothWild produces far more robust and better results than the previous 3D clothed human reconstruction methods from in-the-wild images. As robustness on in-the-wild images has not been extensively studied in the 3D human reconstruction community, we believe ours can give useful insights to the following research.

Our contributions can be summarized as follows.

- We present ClothWild, which reconstructs robust 3D clothed humans from a single in-the-wild image. To the best of our knowledge, it is the first to explicitly address robustness on in-the-wild images.
- For the robustness to the domain gap between synthesized and in-the-wild datasets, we propose a weakly supervised pipeline that is trainable with 2D supervision targets of in-the-wild datasets.
- Our DensePose-based loss function resolves the ambiguities of weak supervision by effectively limiting possible positions of 3D points using DensePose.
- ClothWild largely outperforms previous methods on in-the-wild images.

2 Related works

3D clothed human reconstruction. Varol et al. [35] and Jackson et al. [16] proposed volumetric regression networks that directly predict a voxel representation of a 3D clothed human from a single image. Saito et al. [33,34] and He et al. [13] presented a pixel-aligned implicit function that predicts 3D occupancy fields of clothed humans. Alldieck et al. [1] utilized displacement vectors between a human body surface and clothes for reconstruction. Alldieck et al. [2] presented a method estimating normal and displacement maps on top of a human body model, such as SMPL [23]. Bhatnagar et al. [6] and Jiang et al. [17] presented 3D clothed human reconstruction systems that predict PCA coefficients of cloth generative model space. Huang et al. [15] and He et al. [14] handle implicit function representation of 3D clothed humans in arbitrary poses and produce animatable reconstruction results. Xiu et al. [36] presented ICON, which utilizes local features for robust 3D clothed human reconstruction. Alldieck et al. [3] proposed PHORHUM to photorealistically reconstruct the 3D geometry and appearance of a dressed person. As all of the above methods require 3D scans for training, they are trained on synthetic datasets, which consist of 3D scans and the rendered images from the scans. As the synthetic datasets mainly contain simple human poses with artificial appearances, the methods trained on such datasets fail to generalize to in-the-wild datasets. In contrast, our ClothWild is the first work that can be weakly supervised with 2D supervision targets of in-the-wild datasets, which results in robust outputs on in-the-wild images.

Recently, Corona et al. [7] presented a fitting framework that fits their cloth generative model to 2D cloth segmentations. Although their fitting framework handles reconstruction on in-the-wild images, there are two major differences from ours. First, their fitting framework is not a learning-based system; hence, it cannot utilize image features. As it only relies on 2D cloth segmentations for fitting, it produces wrong results when cloth segmentations are not available due to truncations or occlusions. On the other hand, ClothWild is a learning-based system that utilizes image features; hence, it is much more robust to occlusions and truncations by considering contextual information of image features, as shown in Fig. 5. Second, their fitting framework suffers from the depth ambiguity and pixel-misalignment as described in Section 1, since it projects 3D reconstruction results to the 2D image space and compares the projected one with 2D cloth segmentations. On the other hand, ClothWild leverages our newly proposed DensePose-based loss function. Hence, ClothWild suffers much less from the depth ambiguity and is free from the pixel-misalignment issue. The detailed description of it is provided in Section 6.4.

3D cloth generative model. 3D cloth generative models parameterize 3D clothes by embedding acquired 3D cloth scans in a latent space. Bhatnagar et al. [6] and Jiang et al. [17] used the PCA algorithm to represent clothes in a latent space. Ma et al. [25] used a graph convolutional neural network-based generative model to embed the clothes in a latent space. Bertiche et al. [5] embedded displacements from the human body surface to registered cloth meshes and cloth types as latent codes. Patel et al. [29] decomposed 3D clothes into low-

and high-frequency components and represented them as a function of human poses, shapes and cloth styles. Corona et al. [7] presented a cloth generative model, SMPLicit, which embeds 3D clothes as latent codes that represent cloth styles and cloth cuts. SMPLicit covers a wide variety of clothes, which differ in their geometric properties, such as sleeve length and looseness. In our framework, we employ SMPLicit as a cloth generative model.

3D human body reconstruction. 3D human body reconstruction methods predict parameters of the SMPL [23] body model from a single image. They perform well on in-the-wild images via weakly supervision with 2D GTs of in-the-wild images. Kanazawa et al. [18] proposed an end-to-end framework that utilizes adversarial loss to reconstruct plausible human bodies. Kolotouros et al. [20] combined a regressor and iterative fitting framework. Moon et al. [27] presented Pose2Pose utilizing local and global image features. We use Pose2Pose as an off-the-shelf 3D human body reconstruction method of our ClothWild due to its superior accuracy compared to other works.

3 ClothWild

Fig. 2 shows the overall pipeline of our ClothWild, which consists of ClothNet and BodyNet. We provide a detailed description of each module below.

3.1 ClothNet

Given an input image $\mathbf{I}$, ClothNet predicts cloth existence scores $\mathbf{c} = (c_1, \dots, c_{N_c})$, a set of cloth latent codes $\{\mathbf{z}_i\}_{i=1}^{N_c}$, and a gender $\mathbf{g}$. $N_c = 5$ denotes the number of clothes we consider, which include upper cloth, coat, pants, skirt, and shoes. The ith cloth existence score c_i represents the probability that a human is wearing ith cloth. The ith cloth latent code $\mathbf{z}_i \in \mathbb{R}^{d_i}$ represents a low-dimensional code of ith 3D cloth, embedded in latent space of the 3D cloth model, SMPLicit [7]. We set $d_i = 4$ for shoes and $d_i = 18$ for other clothes following SMPLicit [7]. The gender $\mathbf{g} \in \mathbb{R}^2$ is a one-hot encoded vector that represents a label for the appropriate human body model (*i.e.*, male and female). We use ResNet-50 [12] to extract an image feature vector $\mathbf{f} \in \mathbb{R}^{2048}$ from the input image after removing the fully-connected layer of the last part of the original ResNet. The image feature vector $\mathbf{f}$ is passed into a fully-connected layer, followed by a sigmoid activation function, to predict the cloth existence scores $\mathbf{c}$. In addition, we use N_c fully-connected layers to predict a set of cloth latent codes $\{\mathbf{z}_i\}_{i=1}^{N_c}$ from $\mathbf{f}$, where ith fully-connected layer predicts the latent codes of ith cloth, $\mathbf{z}_i$. Finally, another fully-connected layer, followed by a softmax activation function, predicts the gender from $\mathbf{f}$. The predicted set of cloth latent codes $\{\mathbf{z}_i\}_{i=1}^{N_c}$ and gender $\mathbf{g}$ are passed to SMPLicit.

3.2 BodyNet

The BodyNet takes the input image $\mathbf{I}$ and predicts a shape parameter $\beta \in \mathbb{R}^{10}$ and a pose parameter $\theta \in \mathbb{R}^{72}$ of the SMPL human body model [23]. The shape

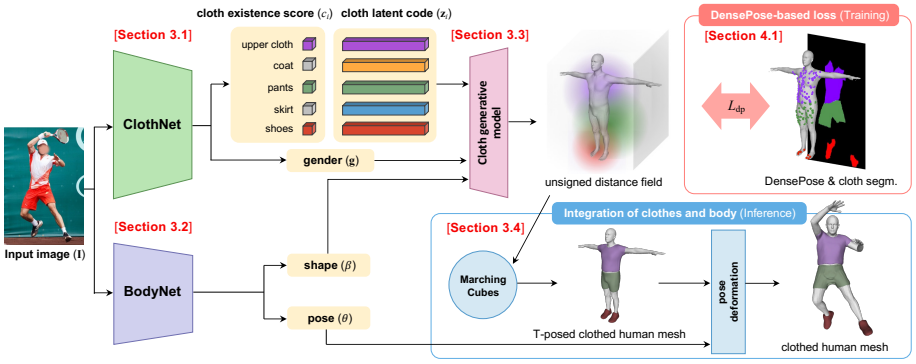


Fig. 2. The overall pipeline of ClothWild. ClothNet predicts cloth latent codes of the cloth generative model, SMPLicit, and produces an unsigned distance field by passing the codes to the SMPLicit. BodyNet predicts SMPL pose and shape parameters. At the training stage, the unsigned distance field is supervised with a combination of DensePose and cloth segmentations with our DensePose-based loss function. At the inference stage, the final 3D clothed human reconstruction is obtained through Marching Cubes and pose deformation steps.

parameter β represents PCA coefficients of T-posed human body shape space, and the pose parameter θ represents 3D rotations of human body joints. The shape parameter β is forwarded to SMPLicit that is described in Section 3.3. The pose parameter θ is used in the inference stage, of which detailed descriptions are in Section 3.4. We use Pose2Pose [27] as the BodyNet, which achieves state-of-the-art performance on in-the-wild benchmarks.

3.3 Cloth generative model

The cloth generative model embeds 3D clothes in the latent space. We use SMPLicit [7] as our cloth generative model, which produces a continuous unsigned distance field of a 3D cloth from a cloth latent code, gender, and human shape. We use the pre-trained SMPLicit and fix it while training ClothWild. Given the cloth latent codes $\{\mathbf{z}_i\}_{i=1}^{N_c}$, gender $\mathbf{g}$, and human shape β, SMPLicit outputs the unsigned distance field for the ith cloth, as follows:

$$C(\mathbf{x}, \mathbf{z}_i, \mathbf{g}, \beta) \longrightarrow \mathbb{R}^+, \quad i = 1, 2, \ldots, N_c, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^3$ is a 3D query point in a canonical 3D space where the human is in a T-pose. In the canonical 3D space, a T-posed naked human body mesh is founded by deriving from the gender $\mathbf{g}$ and human shape β. Around the T-posed naked human body, a set of 3D query points indicate the closest distance to the 3D cloth surface as the unsigned distance field. The above step only proceeds for the ith cloth when its cloth existence score c_i is larger than a threshold, which we empirically set to 0.25.

3.4 Integration of clothes and body

Marching Cubes. In the inference stage, we obtain cloth meshes using Marching Cubes [24], which extracts a mesh of an isosurface from a 3D discrete scalar field, such as the unsigned distance field. We calculate the unsigned distance field by densely sampling 3D query points in the 3D space and forwarding them to the SMPLicit model. Then, we extract cloth meshes on the T-pose from the unsigned distance field by Marching Cubes. At the end of this step, we obtain a T-posed clothed human mesh that comprises a T-posed naked human body mesh and cloth meshes.

Pose deformation. Finally, we apply the pose deformation to the naked human body mesh and cloth meshes of the T-posed clothed human mesh, respectively, following Corona et al. [7]. To deform the naked human body, we use the skinning deformation of SMPL with the pose parameter θ predicted by the BodyNet. To deform cloth meshes, we allocate each cloth vertex of the cloth meshes to its closest naked human body vertex and apply the same pose deformation of the human body vertex. Such the SMPL-driven pose deformation has the strength that the reconstruction results can be deformed with an arbitrary pose instead of the predicted pose, which enables an animation.

4 Learning from in-the-wild datasets

In this section, we describe loss functions to train our framework on in-the-wild datasets without any 3D scan GTs. ClothNet is the only trainable module in our framework, and all other modules, including BodyNet and SMPLicit, are fixed during the training. The overall loss function is

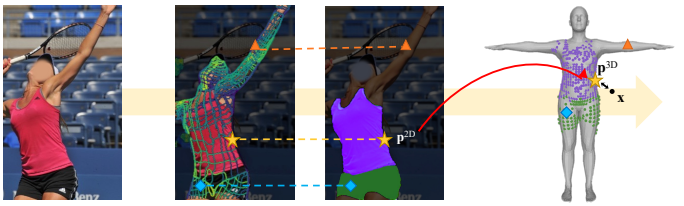
$$L_{\text{total}} = \lambda_{\text{dp}} L_{\text{dp}} + \lambda_{\text{reg}} L_{\text{reg}} + \lambda_{\text{exist}} L_{\text{exist}} + \lambda_{\text{gender}} L_{\text{gender}}, \tag{2}$$

where $\lambda_{\text{dp}} = 1$, $\lambda_{\text{reg}} = 0.1$, $\lambda_{\text{exist}} = 0.01$, and $\lambda_{\text{gender}} = 0.01$.

4.1 DensePose-based loss function

The DensePose-based loss function L_{dp}, which uses a combination of cloth segmentations and DensePose, makes reconstructed clothes close to cloth segmentations. The loss is computed in three steps: cloth-to-body mapping, query point selection, and loss calculation.

Cloth-to-body mapping. In the first step, we place pixels of 2D cloth segmentations in 3D space. Fig. 3 shows the procedure of the cloth-to-body mapping step. In the area where DensePose is defined, we sample 2D cloth points $\{\mathbf{p}_k^{\text{2D}}\}_{k=1}^{N_p}$ on cloth segmentations, where $N_p = 196$ is the number of the sampled 2D points. We utilize DensePose that informs where each human pixel corresponds to the 3D human body surface of SMPL to map the 2D cloth points $\{\mathbf{p}_k^{\text{2D}}\}_{k=1}^{N_p}$ to 3D cloth points $\{\mathbf{p}_k^{\text{3D}}\}_{k=1}^{N_p}$ on the T-posed human body surface. Each of the mapped 3D cloth points represents a cloth label (*e.g.*, upper cloth, pants, or non-cloth) elicited from the cloth segmentations.



(a) image (b) DensePose (c) cloth segm. (d) T-posed human body

Fig. 3. The procedure of the cloth-to-body mapping step. From (a) an in-the-wild image, we exploit 2D supervision targets: (b) DensePose and (c) cloth segmentations. (d) We sample a 2D cloth point $\mathbf{p}^{\text{2D}}$ of the cloth segmentations, and map the 2D cloth point $\mathbf{p}^{\text{2D}}$ to the 3D cloth point $\mathbf{p}^{\text{3D}}$ on the T-posed human body surface. We select the 3D query point $\mathbf{x}$ within constant distance from the 3D cloth points, and the DensePose-based loss function supervises the selected 3D query point.

Query point selection. In this step, we select 3D query points around the T-posed human body for loss calculation. For each cloth type, we uniformly sample 3D query points at a resolution of $21 \times 21 \times 21$ from a 3D bounding box, where the sampling strategy is described in the supplementary material. We select 3D query points $\{\mathbf{x}_j\}_{j=1}^{N_q}$ within a distance threshold τ from the 3D cloth points $\{\mathbf{p}_k^{\text{3D}}\}_{k=1}^{N_p}$ on the 3D human body surface, where N_q is the number of the sampled 3D query points. The distance threshold τ is set 10 cm for the coat and 3 cm for others.

Loss calculation. To the formal description of the loss function, we define a cloth binary function $S_i(\cdot)$. If the closest 3D cloth point from a 3D query point $\mathbf{x}_j$ belongs to the ith cloth, $S_i(\mathbf{x}_j)$ becomes 1 and 0 else. With the cloth binary function $S_i(\cdot)$, the DensePose-based loss function follows:

$$\begin{aligned} L_{\text{dp}} = \frac{1}{N_c N_q} \sum_{i=1}^{N_c} \sum_{j=1}^{N_q} & (S_i(\mathbf{x}_j) |C(\mathbf{x}_j, \mathbf{z}_i, \mathbf{g}, \beta)| \\ & + (1 - S_i(\mathbf{x}_j)) |C(\mathbf{x}_j, \mathbf{z}_i, \mathbf{g}, \beta) - d_{\max}|), \end{aligned} \tag{3}$$

where $d_{\max}$ is a maximum cut-off distance of the unsigned distance fields, set to 0.01 for shoes and 0.1 for others. The first loss term forces the 3D query point $\mathbf{x}$ to belong to ith cloth by making its unsigned distance close to zero when the 3D point matches cloth label i. The second loss term forces the 3D query point $\mathbf{x}$ to not belong to ith cloth when the 3D point does not match cloth label i. As a result, the DensePose-based loss function supervises the 3D query points established by the above steps without the depth ambiguity of the cloth segmentations.

4.2 Other loss functions

Regularization loss. The regularization loss function L_{reg} makes predicted latent codes close to the mean of the cloth latent space, which results in plau-

sible 3D clothes. The regularization loss function is defined as follows: $L_{\text{reg}} = \sum_{i=1}^{N_c} \alpha_i \|\mathbf{z}_i\|_2$, where α_i is set 0.1 for shoes and 1.0 for others. As the DensePose-based loss supervises 3D points that belong to partial areas of clothes, there is a possibility to learn implausible clothes (*e.g.*, overly thick cloth or torn cloth) that only rely on the areas. The regularization loss prevents it by constraining output clothes to be close to the mean.

Cloth existence loss. For the cloth existence prediction, we calculate a binary cross-entropy as follows: $L_{\text{exist}} = -\frac{1}{N_c} \sum_{i=1}^{N_c} (c_i^* \log c_i + (1 - c_i^*)(1 - \log c_i))$, where the asterisk denotes the groundtruth.

Gender classification loss. For the gender classification, we calculate a cross-entropy as follows: $L_{\text{gender}} = -(g_m^* \log g_m + g_f^* \log g_f)$, where g_m and $g_f = 1 - g_m$ is the probability of being male and female of the input human, respectively. The asterisk denotes the groundtruth.

5 Implementation details

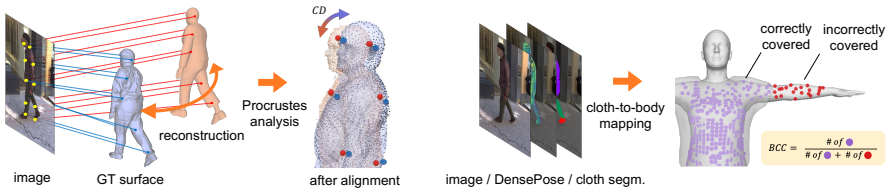
PyTorch [28] is used for implementation. The backbone part is initialized with the publicly released ResNet50 [12] pre-trained on ImageNet [32]. The weights are updated by Adam optimizer [19] with a mini-batch size of 8. The human body region is cropped using a GT box in both training and testing stages following previous works [33,34]. The cropped image is resized to 256×192. Data augmentations, including scaling, rotation, random horizontal flip, and color jittering, are performed in training. The initial learning rate is set to 10^{-4} and reduced by a factor of 10 after the 5th epoch. We train the model for 8 epochs with an NVIDIA GTX 2080 Ti GPU.

6 Experiment

6.1 Datasets

MSCOCO. MSCOCO [22] is a large-scale in-the-wild dataset, which provides in-the-wild images with diverse human poses and appearances. We train ours on the training split and evaluate on the validation split. We use outputs of the DensePose regression model [10] for the DensePose-based loss function as GT DensePose provides only sparse annotations. For the cloth segmentations, we use LIP [9], which contains cloth segmentation annotations of MSCOCO images. We acquire gender annotations by running Homogenus [30] on MSCOCO and use its predictions as supervision targets.

DeepFashion2. DeepFashion2 [8] is a comprehensive dataset that focuses on capturing clothed humans with a wide variety of cloth styles, and we use it as the additional train set along with the MSCOCO. We obtain DensePose and gender annotations by the same procedure as MSCOCO, described above. For cloth segmentation annotations, we use SCHP [21] to obtain cloth segmentations of the dataset.



(a) Chamfer distance (CD) (b) Body-cloth correspondence (BCC)

Fig. 4. Descriptions of our evaluation metrics. (a) Chamfer distance (CD) is a 3D distance between reconstruction and GT surface. It is measured after aligning them based on each 2D projection. (b) BCC is a proportion of 3D points that have correctly matched cloth types. The purple and red points on the right body surface are 3D points that are mapped from cloth segmentations using DensePose. The purple points represent 3D points that are covered with a correct type of reconstructed 3D clothes, while the red points represent others. It is measured after normalizing 3D poses.

3DPW. 3DPW [26] is an in-the-wild dataset, and we use it only for the evaluation purpose after sampling every 25th frame of test set videos. We use two items of 3DPW: T-posed clothed human meshes registered to each subject's 3D scan and SMPL pose parameters of humans in images. As the registered 3D clothed human meshes have the same mesh topology as that of SMPL body mesh, we deform the registered ones with SMPL pose parameters following the same skinning algorithm of SMPL. We use the posed registered meshes as GT 3D clothed humans of 3DPW.

6.2 Evaluation metrics

Fig. 4 shows our two evaluation metrics. Since there has been no work dealing with 3D clothed human reconstruction on the in-the-wild dataset, we propose two evaluation metrics in the following.

Chamfer distance (CD). Chamfer Distance (CD) is a 3D distance between the 3D clothed human reconstruction and the GT surface, widely used in previous works [33,34,17]. Before measurement, we rigidly align global rotation, scale, and translation of the reconstruction to the GT surface. The alignment is not trivial as there are no semantically matching pairs between the reconstruction and GT surface. For the alignment, we assume that the same human parts (*e.g.*, shoulders and knees) of both the reconstruction and the GT surface are projected to the same pixel in the image. We rasterize both the reconstruction and GT surface and pair two vertices (*i.e.*, one from the reconstruction and the other from the GT surface) that correspond to the same pixel. Based on the matching pairs, we align the reconstruction and measure the Chamfer distance (CD) in millimeter. More details of CD metric is described in the supplementary material.

Body-cloth correspondence (BCC). Body-cloth correspondence (BCC) is a proportion of 3D points that have correct cloth types on the T-posed naked human body surface. It only considers 3D cloth predictions, excluding 3D human

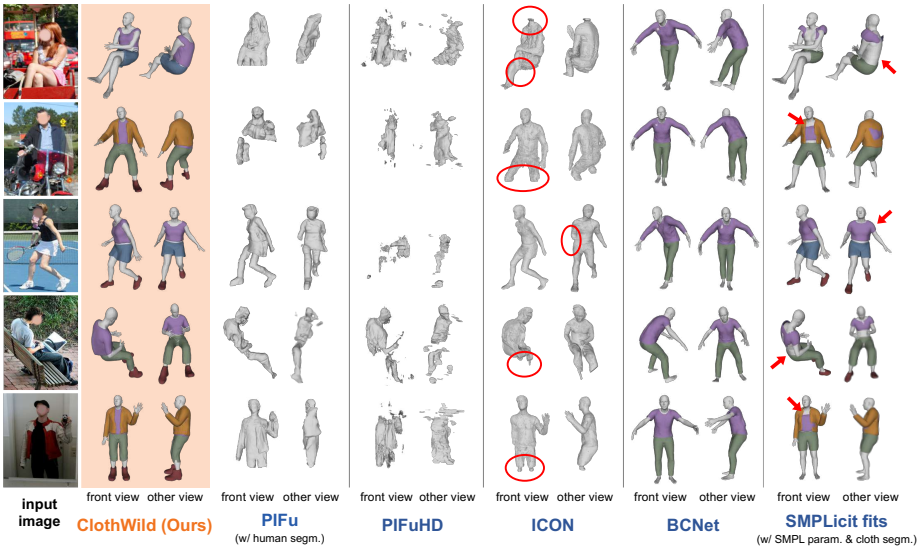


Fig. 5. Qualitative result comparisons on MSCOCO validation set. PIFu additionally uses human segmentation obtained from Mask R-CNN [11] for reconstruction. SMPLicit fits use SMPL parameter and cloth segmentations obtained from our BodyNet and SCHP [21], respectively. Colors of reconstructed 3D clothes are manually assigned to represent cloth types.

pose predictions. With cloth-to-body mapping described in Section 4.1, we map 2D points in a GT cloth segmentation for one cloth label to 3D points on the T-posed human body surface with the cloth-to-body mapping. We consider 3D points are correctly covered ones if the distance between them and reconstructed cloth is shorter than 3 cm. We calculate the proportion of the correctly covered points and average the proportions for all cloth labels (*e.g.*, upper cloth, pants, or non-cloth). This metric is used for methods that support cloth-to-body mapping based on the T-posed naked body of SMPL. For example, as PIFu [33], PIFuHD [34], and ICON [36] do not tell us cloth type which points belong, we could not evaluate them using the BCC metric.

6.3 Comparison with state-of-the-art methods

We compare our ClothWild with recent 3D clothed human reconstruction methods: PIFu [33], PIFuHD [34], ICON [36], BCNet [17], and the fitting framework of SMPLicit [7]. For the inference, PIFu requires human segmentation, and the fitting framework of SMPLicit requires SMPL parameter and cloth segmentations. We provide human segmentation, obtained by Mask R-CNN [11], to PIFu. The same SMPL parameters of our BodyNet and cloth segmentations from SCHP [21] are provided to the SMPLicit fitting framework. All of their results are obtained by using their officially released codes and pre-trained weights. Please note that

Table 1. CD comparison with state-of-the-art methods on 3DPW. Methods with $*$ and † additionally use human segmentation and cloth segmentations as an input for the inference, respectively.

Methods	CD ↓
PIFuHD [34]	137.50
BCNet [17]	118.75
ICON [36]	75.00
PIFu [33]$*$	67.25
SMPLicit fits [7]†	45.66
ClothWild (Ours)	**40.34**

Table 2. BCC comparison with state-of-the-art methods on MSCOCO. Methods with † additionally use cloth segmentations as an input for the inference.

Methods	upper body	lower body	non-cloth	average
BCNet [17]	0.415	0.729	0.800	0.648
SMPLicit fits [7]†	0.645	0.493	**0.961**	0.700
ClothWild (Ours)	**0.830**	**0.820**	0.887	**0.846**

the fitting framework of SMPLicit fits their cloth latent codes to the cloth segmentations. We call the fitting results of their framework as SMPLicit fits.

Qualitative results. Fig. 5 shows that our ClothWild produces much better reconstruction results than previous state-of-the-art 3D clothed human reconstruction methods on MSCOCO [22]. PIFu [33], PIFuHD [34], and ICON [36] suffer from undesirable results on in-the-wild images with diverse human poses and appearances, especially in the occluded human scene. ICON produces better results than PIFu and PIFuHD; however, it still suffers from missing human parts. On the other hand, our ClothWild successfully reconstructs invisible parts of clothed humans even when the input human is occluded or truncated. Like PIFu, PIFuHD, and ICON, BCNet [17] also requires 3D scans for the training. Accordingly, it is trained on synthetic datasets without in-the-wild images, the reason for weak results on in-the-wild images. SMPLicit fits [7] does not utilize image features, which provide contextual information of occluded and truncated human parts. Hence, it cannot reconstruct clothes at occluded or truncated human parts. In addition, the depth ambiguity and pixel-level misalignment of their silhouette-based loss function make the fitting attain the incorrect cloth styles. Our ClothWild resolves such issues by utilizing image features and the DensePose-based loss function.

Quantitative results. Table 1 shows that ClothWild achieves the best CD on 3DPW [22]. Also, Table 2 shows that ClothWild produces much better BCC than previous methods on MSCOCO [22]. Unlike ours, BCNet [17] handles only three cloth types (*i.e.*, upper cloth, pants, and skirt). Hence, we categorize upper

Table 3. Running time (seconds per image) comparisons.

PIFuHD [34] (CVPR 20)	SMPLicit fits [7] (CVPR 21)	ICON [36] (CVPR 22)	**ClothWild (Ours)**
20.43	105.43	87.88	**10.21**

cloth and coat into the upper body and pants and skirts into the lower body for a fair comparison. Then, we calculate BCC only for such clothes and exclude shoes. As the BCC evaluates only predicted 3D clothes without considering 3D human poses, the comparisons show that our ClothWild predicts the cloth styles of the image accurately, such as sleeveless and length of upper cloth and pants.

Running time. Table 3 shows that ours takes the shortest computational time to process a single image than recently presented works. The running times are measured in the same environment with Intel Xeon Gold 6248R CPU and NVIDIA RTX 2080 Ti GPU. We exclude pre-processing stages, such as 3D body pose estimation, human segmentation, and cloth segmentations. The fitting framework of SMPLicit takes a much longer time, although it is based on the same cloth generative model [7] as ours. This is because the fitting framework of SMPLicit forwards the cloth generative model about 200 iterations to fit its cloth latent codes to cloth segmentations. ICON suffers from a similar problem as it iteratively fits initial results based on normal maps and silhouettes. In contrast, our ClothWild performs the feed-forward only a single time. The above methods, including ours, require Marching Cubes to obtain final 3D geometry, one of the main bottlenecks of the running time. The running time of ClothWild's each component will be reported in the supplementary material.

6.4 Ablation study

Effectiveness of DensePose-based loss. The top block of Table 4 shows that the proposed DensePose-based loss has a vital role in properly learning 3D clothed humans. Fig. 6 (a) and (b) additionally shows the effectiveness of our DensePose-based loss compared to previous silhouette loss [7]. The silhouette loss enforces 3D cloth reconstruction results to be projected onto the 2D cloth segmentations. The top block of the table shows that using the DensePose-based loss achieves better BCC than using the silhouette loss. The silhouette loss suffers from depth ambiguity and pixel-misalignment, as described in Section 1. On the other hand, the proposed DensePose-based suffers much less from the depth ambiguity and is free from the misalignment issue as it designates accurate 3D points around the human body surface. Hence, it has significant benefits to learning precise 3D clothes.

Effectiveness of regularization loss. The bottom block of Table 4 shows the effectiveness of the regularization loss. Fig. 6 (b) and (c) additionally demonstrates the effectiveness. The regularization loss makes predicted clothes be in the latent space of the 3D cloth model. It is necessary because 2D supervision

Table 4. BCC comparison among different loss configurations.

DensePose	Silhouette	Regularization	BCC ↑
*** Effectiveness of DensePose-based loss**			
✗	✓	✓	0.644
✓	✓	✓	0.684
✓	✗	✓	**0.689 (Ours)**
*** Effectiveness of regularization loss**			
✓	✗	✗	0.381
✓	✗	✓	**0.689 (Ours)**

(a) DP. + Reg. (**Ours**) (b) Sil. + Reg. (c) Ours – Reg. (a) DP. + Reg. (**Ours**) (b) Sil. + Reg. (c) Ours – Reg.

Fig. 6. Comparison of three different loss configurations: (a) our proposed loss configuration, (b) replacing our DensePose-based loss with the silhouette loss, and (c) removing the regularization loss.

targets, such as cloth segmentations, only provide information of partial areas of clothes. Such limited information can lead our ClothWild to produce implausible clothes, such as overly thick cloth and torn cloth. The regularization loss prevents such improper learning and encourages reconstructing reasonable 3D clothes, despite only information of partial areas of clothes.

7 Conclusion

We propose ClothWild, a 3D clothed human reconstruction framework that produces significantly robust results from in-the-wild images. For the robustness to the domain gap between synthesized and in-the-wild datasets, we propose a weakly supervised pipeline and a DensePose-based loss function. As a result, our ClothWild outperforms previous 3D clothed human reconstruction methods on in-the-wild images.

Acknowledgements. This work was supported in part by IITP grant funded by the Korea government (MSIT) [No. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University), No.2022-0-00156], and in part by the Bio & Medical Technology Development Program of NRF funded by the Korean government (MSIT) [No. 2021M3A9E4080782].

References

1. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: CVPR (2019)
2. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: ICCV (2019)
3. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3D reconstruction of humans wearing clothing. In: CVPR (2022)
4. aXYZ: (2018), https://secure.axyz-design.com
5. Bertiche, H., Madadi, M., Escalera, S.: CLOTH3D: Clothed 3D humans. In: ECCV (2020)
6. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-Garment Net: Learning to dress 3D people from images. In: ICCV (2019)
7. Corona, E., Pumarola, A., Alenya, G., Pons-Moll, G., Moreno-Noguer, F.: SMPLicit: Topology-aware generative model for clothed people. In: CVPR (2021)
8. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: CVPR (2019)
9. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR (2017)
10. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: Dense human pose estimation in the wild. In: CVPR (2018)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-PIFu geometry and pixel aligned implicit functions for single-view human reconstruction. In: NeurIPS (2020)
14. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: ARCH++: Animation-ready clothed human reconstruction revisited. In: ICCV (2021)
15. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: ARCH: Animatable reconstruction of clothed humans. In: CVPR (2020)
16. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3D human body reconstruction from a single image via volumetric regression. In: ECCV (2018)
17. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. In: ECCV (2020)
18. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014)
20. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: ICCV (2019)
21. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE TPAMI (2020)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG (2015)
24. Lorensen, W.E., Cline, H.E.: Marching Cubes: A high resolution 3D surface construction algorithm. ACM siggraph computer graphics (1987)

25. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3D people in generative clothing. In: CVPR (2020)
26. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using imus and a moving camera. In: ECCV (2018)
27. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: CVPRW (2022)
28. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
29. Patel, C., Liao, Z., Pons-Moll, G.: TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In: CVPR (2020)
30. 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 (2019)
31. Renderpeople: (2018), https://renderpeople.com/3d-people
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge (2015)
33. 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 (2019)
34. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: CVPR (2020)
35. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3D human body shapes. In: ECCV (2018)
36. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: CVPR (2022)
37. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D human reconstruction from a single image. In: ICCV (2019)