HoloPose: Holistic 3D Human Reconstruction In-The-Wild

Rıza Alp Güler Iasonas Kokkinos
Ariel AI

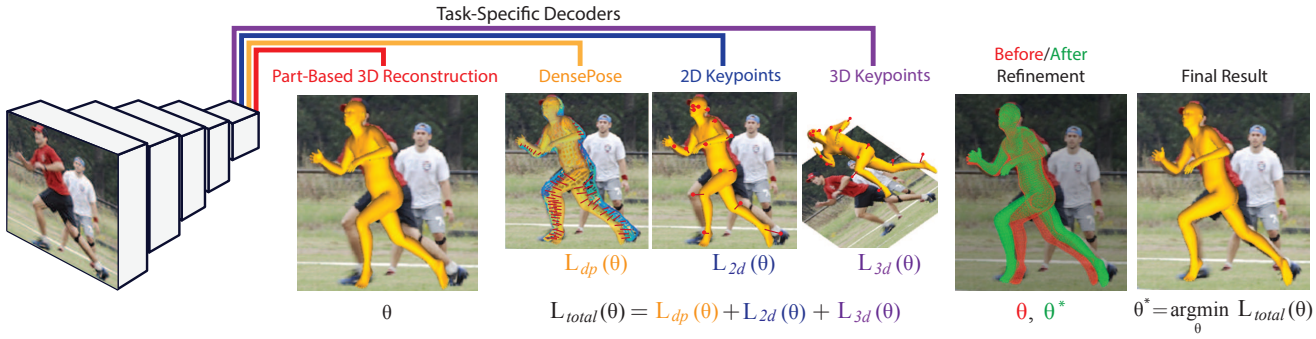

Figure 1: We introduce HoloPose, a method for holistic monocular 3D body reconstruction in-the-wild. We start with an accurate, part-based estimate of 3D model parameters θ, and decoupled, FCN-based estimates of DensePose, 2D and 3D joints. We then efficiently optimize a misalignment loss $L_{total}(\theta)$ between the top-down 3D model predictions to the bottom-up pose estimates, thereby largely improving alignment. The 3D model estimation and iterative fitting steps are efficiently implemented as network layers, facilitating multi-person 3D pose estimation in-the-wild at more than 10 frames per second.

Abstract

We introduce HoloPose, a method for holistic monocular 3D human body reconstruction. We first introduce a part-based model for 3D model parameter regression that allows our method to operate in-the-wild, gracefully handling severe occlusions and large pose variation. We further train a multi-task network comprising 2D, 3D and Dense Pose estimation to drive the 3D reconstruction task. For this we introduce an iterative refinement method that aligns the model-based 3D estimates of 2D/3D joint positions and DensePose with their image-based counterparts delivered by CNNs, achieving both model-based, global consistency and high spatial accuracy thanks to the bottom-up CNN processing. We validate our contributions on challenging benchmarks, showing that our method allows us to get both accurate joint and 3D surface estimates, while operating at more than 10fps in-the-wild. More information about our approach, including videos and demos is available at http://arielai.com/holopose.

1. Introduction

3D reconstruction from a single RGB image is a fundamentally ill-posed problem, but we perform it routinely when looking at a picture. Prior information about geometry can leverage on multiple cues such as object contours [32, 57, 4], surface-to-image correspondences [30, 38, 29] or shading [15, 12], but, maybe the largest contribution to monocular 3D reconstruction comes from semantics: the constrained variability of known object categories can easily resolve the ambiguities in the 3D reconstruction, for instance if the object's shape is bound to lie in a low-dimensional space [7, 21, 49].

This idea has been the basis of the seminal work of [56, 5] on morphable models for monocular 3D face reconstruction. Extending this to the more complicated, articulated structure of the human body, monocular human body reconstruction was studied extensively in the previous decade in conjunction with part-based representations, [40, 64], sampling-based inference [42, 40], spatio-temporal inference [44] and bottom-up/top-down computation [41]. Monocular 3D reconstrunction has witnessed a renaissance in the context of deep learning for both general categories, e.g. [50, 20, 23] and for humans in specific [6, 24, 55, 51, 52, 9, 34, 54, 31, 19, 60]. Most of the latter works rely on the efficient parameterisation of the human body in terms of skinned linear models [2] and in particular the SMPL model [26].

Even though 3D supervision is scarce, these works have exploited the fact that a parametric model provides a low-dimensional representation of the human body that can project to 2D images in a differentiable manner. Based

on this, these works have trained systems to regress model parameters by minimizing the reprojection error between model-based and annotator-based 2D joint positions [51, 52, 19], human segmentation masks and 3D volume projections [54, 51, 52, 31] or body parts [31].

In parallel with these works, 3D human joint estimation has seen a steady rise in accuracy [47, 63, 33, 35], most recently based on directly localizing 3D joints in a volumetric output space through hybrids of classification and regression [46, 27, 22].

Finally, recent work on Dense Pose estimation [10] has shown that one can estimate dense correspondences between RGB images and the body surface by training a generic, bottom-up detection system [13] to associate image pixels with surface-level UV coordinates. Even though DensePose establishes a direct link between images and surfaces, it does not uncover the 3D geometry of the particular scene, but rather gives a strong hint about it.

In this work we propose to link these separate research threads in a synergistic architecture that combines the powers of the different approaches. As in [51, 34, 54, 31, 19] we rely on a parameteric, differentiable human model of shape that allows us to describe the 3D human body surface in terms of a low-dimensional parameter vector, and incorporate it in a holistic system for monocular 3D pose estimation.

Our first contribution consists in introducing a part-based architecture for parameter regression. The present approaches to monocular 3D reconstruction estimate model parameters through a linear layer applied on top of CNN features extracted within the object's bounding box. As described in Sec. 3 our part-based regressor pools convolutional features around 2D joint locations estimated by an FCN-based 2D joint estimation head. This allows us to extract refined, localized features that are largely invariant to articulation, while at the same time keeping track of the presence/absence of parts.

We then exploit DensePose and 3D joint estimation to increase the accuracy of 3D reconstruction. This is done in two complementary ways. Firstly, we introduce additional reprojection-based losses that improve training in a standard multi-task learning setup. Secondly, we predict DensePose and 2D/3D joint positions using separate, FCN-based decoders and use their predictions to refine the top-down, model-based 3D reconstruction.

Our refinement process uses the CNN-based regressor estimates as an initialization to an iterative fitting procedure. We update to the model parameters so as to align the model-based and CNN-based pose estimates. The criterion driving the fitting is captured by a combination of a Dense Pose-based loss, detailed in Sec. 4 and the distances between the model-based and CNN-based estimates of the 3D joints. This allows us to update the model parameter estimates on-the-fly, so as to better match the CNN-based localization results. The iterative fitting is implemented as an efficient network layer for GPU-based Conjugate Gradients, allowing us to perform accurate real-time, multi-person 3D pose estimation in-the-wild.

Finally, in order to make a skinned model better compatible with generic CNN layers we also introduce two technical modifications that simplify modelling, described in Sec. 2. We first introduce a mixture-of-experts regression layer for the joint angle manifold which alleviates the need for the GAN-based training used in [19]. Secondly, we introduce a uniform, cartesian charting of the UV space within each part, effectively reparametrizing the model so as to efficiently implement mesh-level operations, resulting in a simple and fast GPU-based model refinement.

2. Shape Prior for 3D Human Reconstruction

Our monocular 3D reconstruction method heavily relies on a prior model of the target shape. We parameterize the human body using the Skinned Multi-Person Linear (SMPL) model [26], but other similar human shape models could be used instead. The model parameters capture pose and shape in terms of two separate quantities: θ comprises 3D-rotation matrices corresponding to each joint in a kinematic tree for the human pose, and β captures shape variability across subjects in terms of a 10-dimensional shape vector. The model determines a triangulated mesh of the human body through linear skinning and blend shapes as a differentiable function of θ, β, providing us with a strong prior on the 3D body reconstruction problem.

2.1. Mixture-of-Experts Rotation Prior

Apart from defining a prior on the shape given the model parameters, we propose here to enforce a prior on the model parameters themselves. In particular, the range of possible joint angle values is limited by the human body's mechanics, which is something we can exploit to increase the accuracy of our angle joint estimates. In [19] prior constraints were enforced implicitly through adversarial training, where a discriminator network was trained in tandem with an angle regression network and used to penalize statistically implausible joint angle estimates independently.

We argue that a simpler and potentially even tighter prior can be constructed by explicitly forcing the prediction to lie on the manifold of plausible shapes. Unlike earlier work that aimed at analytically modelling joint angles [43] we draw inspiration from recent works that use classification as a proxy to rotation estimation [49]: rather than predict Euler angles, the authors cast rotation estimation as a classification problem where the classes correspond to disjoint angular bins. This approach is aligned with the empirical observation that CNNs can improve their regression accuracy by exploiting classification within regression, as used

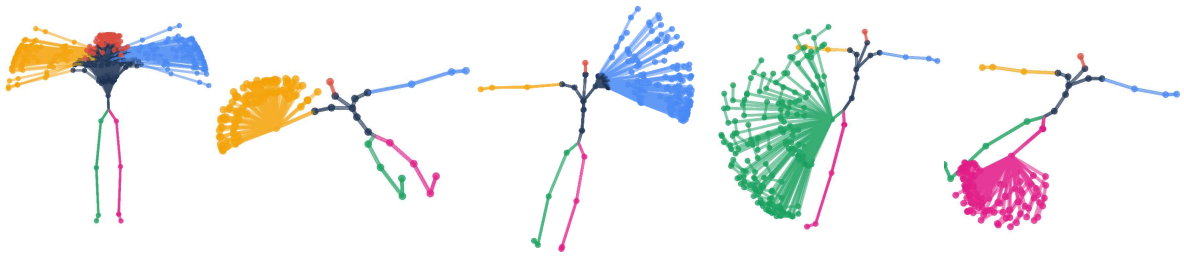

Figure 2: Visualization of Euler angle cluster centers, $\theta_{1,\dots,K}$, for several joints of the SMPL model. We limit the output space of our joint regressors to the convex hull of the centers, enforcing attainable joint rotations.

for instance in [11, 46, 27]

We propose a simple 'mixture-of-experts' angle regression layer that has a simple and effective prior on angles baked into its expression. We start by using the data collected by [1] of joint angle recordings as humans stretch. These are expected to cover sufficiently well the space of possible joint angles. For each body joint we represent rotations as Euler angles, θ and compute K rotation clusters $\theta_1, \dots, \theta_K$ via K-Means. These clusters provide us with a set of representative angle values. We allow our system to predict any rotation value within the convex hull of these clusters by using a softmax-weighted combination of the clusters. In particular, the Euler rotation Θ^i for the i'th body joint is computed as:

$$\theta^i = \frac{\sum_{k=1}^{K} \exp(w_k)\theta_K}{\sum_{k=1}^{K} \exp(w_k)} \quad (1)$$

where w_k are real-valued inputs to this layer. This forms a plausible approximation to the underlying angle distribution, as visualized in Fig. 2, while avoiding the need for the adversarial training used in [19], since by design the estimated angles will be coming from this prior distribution.

2.2. Cartesian surface parametrization

Even though we understand the body surface as a continuous structure, it is discretized using a triangulated mesh. This means that associating a pair of continuous UV coordinates with mesh attributes, e.g. 3D position, requires firstly identifying the facet that contains the UV coordinate, looking up the vertex values supporting the facet, and using the point's barycentric coordinates to interpolate these values. This can be inefficient in particular if it requires accessing disparate memory positions for different vertices.

We have found it advantageous to reparametrize the body surface with a locally cartesian coordinate system. This allows us to replace this tedious process with bilinear interpolation and use a Spatial Transformer Layer [17] to efficiently handle large numbers of points. In order to perform this reparametrization we first perform Multi-Dimensional Scaling to flatten parts of the model surface to two dimensions and then sample these parts uniformly on a grid.

In particular we use a 32×32 grid within each of the 24 body parts used in [10] which means that rather than the 6890 3D vertices of SMPL we now have 24 tensors of size $32 \times 32 \times 3$. We also sample the model eigenshapes on the same grid and express the shape synthesis equations in terms of the resulting tensors. We further identify UV-part combinations that do not correspond to any mesh vertex and ignore UV points that map there.

3. Part-Based 3D Body Reconstruction

Having outlined the parametric model used for 3D reconstruction, we now turn to our part-based model for parameter estimation. Existing model-based approaches to monocular 3D reconstruction estimate SMPL parameters through a single linear layer applied on top of CNN features extracted within the object's bounding box. We argue that such a monolithic system can be challenged by feature changes caused e.g. by occlusions, rotations, or global translations due to bounding box misalignments.

We handle this problem by extracting localized features around human joints, following the part-based modeling paradigm [8, 64, 61]. The position where we extract features co-varies with joint position. The features are therefore invariant to translation by design and can focus on local patterns that better reveal the underlying 3D geometry.

As shown in Fig. 3 we obtain features as a result of a deconvolution network and pool features at visible joint locations via bilinear interpolation. The joint locations are delivered by a separate network branch, trained for joint localization. Each feature extracted around a 2D joint can in principle be used to separately regress the full model parameters, but intuitively a 2D joint should have a stronger influence on model parameters that are more relevant to it. For instance a left wrist joint should be affecting the left arm parameters, but not those of kinetically independent parts such as the right arm, head, or feet. Furthermore, the fact that some joints can be missing from an image means that we cannot simply concatenate the features in a larger fea-

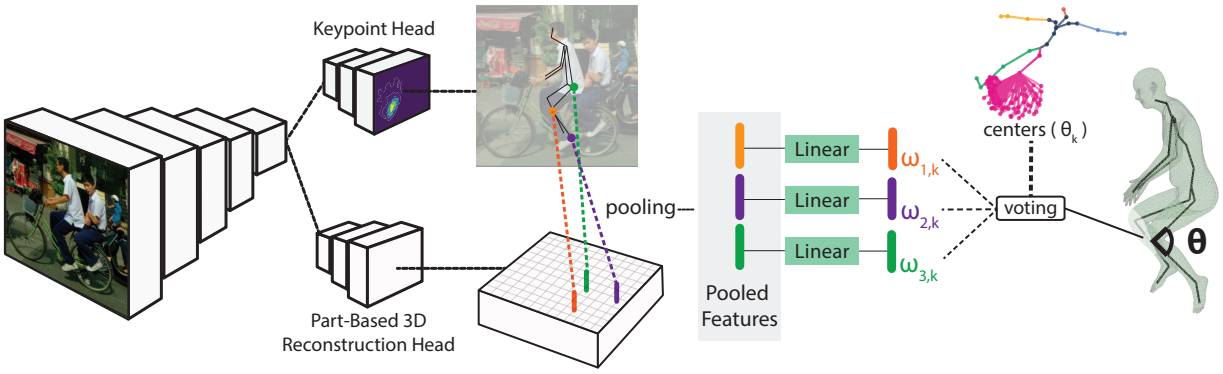

Figure 3: **Part-Based 3D Reconstruction.** A fully convolutional network for keypoint detection is used to localize 2D landmark positions of multiple human keypoints. We pool convolutional features around each keypoint, deriving a rich representation of local image structure that is largely invariant to global image deformations, and instead ellicits fine-grained, keypoint-specific variability. Each keypoint affects a subset of kinematically associated body model parameters, casting its own 'vote' for the putative joint angles. These votes are fused through a mixture-of-experts architecture that delivers a part-based estimate of body joint angles. In this figure for simplicity we show only pooling from the left-ankle, left-knee and left-hip local features which are relevant for the estimation of the left-knee angles.

ture vector, but need to use a form that accommodates the potential absence of parts.

We incorporate these requirements in a part-based variant of Eq. 1, where we pool information from $\mathcal{N}(i)$, the neighborhood of joint i corresponding to the angle θ^i:

$$\theta^i = \frac{\sum_{k=1}^{K} \sum_{j \in \mathcal{N}(i)} \exp(w_{k,j}^i)\theta_K}{\sum_{k=1}^{K} \sum_{j \in \mathcal{N}(i)} \exp(w_{k,j}^i)}. \quad (2)$$

As in Eq. 1 we perform an arg-soft-max operation over angle clusters, but fuse information from multiple 2D joints: $w_{k,j}^i$ indicates the score that 2D joint j assigns to cluster k for the i-th model parameter, θ^i. The neighborhood of i is constructed offline, by inspecting which model parameters directly influence human 2D joints, based on kinematic tree dependencies. Joints are found in the image by taking the maximum of a 2D joint detection module. If the maximum is below a threshold (0.4 in our implementation) we consider that a joint is not observed in the image. In that case, every summand corresponding to a missing joint is excluded, so that Eq. 2 remains valid. If all elements of $\mathcal{N}(i)$ are missing, we set θ^i to the resting pose.

4. Holistic 3D Body Reconstruction

The network described so far delivers a 'bottom-up' estimate of the body's 3D surface in a single-shot, i.e. through a forward pass in the network. In the same feedforward manner one can obtain 2D keypoints, 3D joint [46], or DensePose [10] estimates through fully-convolutional networks (FCNs). These provide complementary pieces of information about the human pose in the scene, with complementary merits. In particular, the model-based estimate of the body geometry is a compact, controllable, representation of a watertight mesh, that is bound to correspond to a plausible human pose. This is often not the case for the FCN-estimates, whose feedforward architecture makes it hard to impose lateral constraints between parts. At the same time the FCN-based estimates inspect and score exhaustively every image patch, allowing us to precisely localize human structures in images. By contrast, model-based estimates can be grossly off, e.g. due to some miscalculated angle in the beginning of the kinematic tree.

Motivated by this complementarity, we now turn to developing a holistic pose estimation system that allows us to have the best of both words. Our starting point is the fact that having a 3D surface estimate allows us to predict in a differentiable manner 3D joint positions, their 2D projections, alongside with dense surface-to-image correspondences. We can thus use any external pose information to construct a loss function that indicates the quality of our surface estimate in terms of geometric distances. Building on this, and as done also in [51, 52, 9, 34, 54, 31, 19, 60] we use multiple pose estimation cues to supervise the 3D reconstruction task, now bringing also DensePose [10] as a new supervision signal.

A more radical change with respect to prior practice is that we also introduce a refinement process that forces the model-based 3D geometry to agree with an FCN's predictions through an iterative scheme. This is effective also at test-time, where the FCN-based pose estimates drive the alignment of the model-based predictions to the image evidence through a minimization procedure.

In order to achieve both of these goals we exploit the geometric nature of the problem and construct a loss that penalizes deviations between the 3D model-based predictions and the pose information provided by complemen-

Figure 4: **DensePose refinement:** a bottom-up estimate of dense image-surface correspondence (row 1) is used to refine the 3D model estimate results (row 2), achieving a better alignment of the surface projection to the image (row 3).

tary cues. For example, Dense Pose associates an image position $\mathbf{x} = (x_1, x_2)$ with an intrinsic surface coordinate, $\mathbf{u} = (u_1, u_2)$. Given a set of model parameters $\phi = (\theta, \beta)$ we can associate every $\mathbf{u}$ vector with a 3D position $\mathbf{X}(\phi) = M(\phi, \mathbf{u})$, where M denotes the parametric model for the 3D body shape, e.g. [26]. This point in turn projects to a 2D position $\hat{\mathbf{x}}(\phi) = (\hat{x}_1, \hat{x}_2)$, which can be compared to $\mathbf{x}$ - ideally closing a cycle. Since this will not be the case in general, we penalize a geometric distance between $\hat{\mathbf{x}}(\phi)$ and $\mathbf{x} = (x_1, x_2)$, requiring that (ϕ) yields a shape that projects correctly in 2D. Summarizing, we have the following process and loss:

$$\mathbf{x} \overset{\text{DensePose}}{\rightarrow} \mathbf{u} \overset{M(\phi)}{\rightarrow} \mathbf{X} \overset{\Pi}{\rightarrow} \hat{\mathbf{x}} \tag{3}$$

$$\mathcal{L}_{\text{DensePose}}(\phi) = \sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2, \tag{4}$$

where $\hat{\mathbf{x}} = \Pi(M_\phi(\text{DensePose}(\mathbf{x})))$ is the model-based estimate of where $\mathbf{x}$ should be, Π is an orthographic projection matrix and i ranges over the image positions that become associated with a surface coordinate.

We can use Eq. 4 in two ways, as described above. Firstly, we can use it to supervise network training, where DensePose stands for Dense Pose ground-truth and ϕ is obtained by the part voting expression in Eq. 2. This will force the network predictions to comply with DensePose supervision, compensating for the lack of extensive 3D supervision.

Secondly, we can use Eq. 4 at test time to force the coupling of the FCN- and model- based estimates of human pose. We bring them in accord by forcing the model-based estimate of 3D structure to project correctly to the FCN-based DensePose/2D/3D joint predictions. For this we treat the CNN-based prediction as an initialization of an iterative fitting scheme driven by the sum of the geometric losses. We treat similarly the 2D and 3D joint predictions delivered by the FCN heads, and penalize the $L1$ distance of the model-based prediction to the CNN-based estimates. Furthermore, to cope with implausible shapes we use the following simple loss to bound the magnitude of the predicted β values: $\mathcal{L}_{beta} = \sum_i \max(0, b - |\beta_i|)$, where $b = 2$ is used in all experiments.

We use Conjugate Gradients (CG) to minimize a cost function formed by the sum of the above losses. We implement Conjugate Gradients as an efficient, recurrent network layer; our cartesian model parameterization outlined in Sec. 2.2 allows us to quickly evaluate and back-propagate through our losses using Spatial Transformer Networks. This gives us for free a GPU-based implementation of 3D model fitting to 2D images. If convergence is not achieved after a fixed number of 20 CG iterations we halt to keep the total computation time bounded. For the sparse, keypoint-based reprojection loss every iteration requires less than 20 msecs, while the single-shot feedforward surface reconstruction requires less than 10msecs. We anticipate that learning-based techniques, such as supervised descent [39, 59, 48] could be used to further accelerate convergence.

5. Experiments

We now describe our experimental setup and architectural choices, providing quantitative and qualitative results. We quantify performance in terms of two complementary

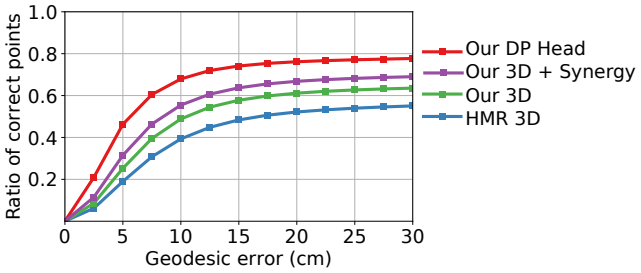

Figure 5: **Surface Correspondence Results:** Ratio of Correct points as a function of the the geodesic distance threshold. The proposed system uniformly outperforms the current state-of-the-art mesh-based results, while the refinement yields a further boost in surface alignment.

Method	$\mathbf{AUC}_{10}$	$\mathbf{AUC}_{30}$	$\mathbf{FR}_{10}$
HMR[19]	0.18	0.40	60.7
Ours	0.24	0.47	51.1
Ours+Synergy	0.29	0.53	44.6
Our DP Branch	0.40	0.63	32.0

Table 1: **DensePose Results:** Area Under the RCP curve (AUC) and Failure Rate (FR) results for DensePose estimation obtained by 3D body mesh reprojection to 2D.

problem aspects, namely mesh-based dense pose estimation and 3D object reconstruction. Our qualitative results demonstrate the performance of our system on challenging, "in-the-wild" images with heavy occlusion and clutter.

5.1. Experimental Setup

In all of our experiments we use as system backbone an ImageNet pre-trained ResNet-50 network [14]. For each dense prediction task we use a deconvolutional head following the architectural choices of [58], with three 4×4 deconvolution layers applied with batch-norm and ReLU, followed by a linear layer to obtain outputs of the desired dimensionality. Each deconvolution layer has a stride of 2, yielding a 64×64 tensor given a 256×256 image as input.

We train our system in two stages. We first train the 3D Keypoint and 2D Keypoint heads on the MPII[3], COCO[25] and H36m[16] datasets, following the practices of [46], including an integral loss for 3D keypoints. We then append the DensePose and part-based 3D reconstruction heads to the network and train the whole system end-to-end. The DensePose branch is trained using correspondences in the DensePose-COCO training set. The part-based 3D reconstruction head is trained using 2D and 3D keypoints and dense correspondences present in all datasets.

5.2. Surface Correspondence Performance

Dense correspondence measures 3D mesh alignment accuracy in challenging, "in the wild" scenarios. It complements 3D localization performance, because 3D localization can only be evaluated in constrained images where 3D pose information has been captured by appropriate setups, while dense correspondence can be established by manual annotators as in [10].

Method	*PA MPJPE*	*MPJPE*
3D Keypoint Localization		
Rogez *et al.* [36]	53.4	71.6
Pavlakos *et al.* [33]	51.9	71.9
Martinez *et al.* [28]	47.7	62.9
Sun *et al.* [45]	48.3	59.1
Sun *et al.* [46]	40.6	**49.6**
Multi-Task 3D Keypoints Branch		
BodyNet [54]	49.0	-
Our 3D Kps Branch	**36.82**	**50.42**
Keypoints on Reconstructed 3D Shape		
Zhou. *et al.* [62]	-	107
Reg. Forest (91 kps) [24]	93.9	-
SMPLify [6]	82.3	-
SMPLify (91 kps) [24]	80.7	-
Pavlakos *et al.* [34]	75.9	-
HMR unpaired	66.5	106.84
Omran *et al.* [31]	59.9	-
HMR	56.8	87.97
Ours	**50.56**	**64.28**
Ours+ Synergy	**46.52**	**60.27**

Table 2: **Results on Human3.6M Dataset.** MPJPE in mm. PA MPJPE means the estimated keypoints were rigidly aligned to ground truth prior to evaluation.

We measure the surface correspondence accuracy using the 'Ratio of Correct Points' (RCP) measure. A point correspondence is declared to be correct if the geodesic distance between the estimated and ground truth points is below a given threshold. On Fig. 5 we present RCP curves as a function of the geodesic error threshold. Following [10], we report RCP values between 0 and 30cm. The results show that the alignment accuracy of the proposed approach is clearly superior to the state-of-the-art approach of [20].

We also provide the results for the correspondence predicted by the discriminatively trained DensePose head of our system. The results clearly show that the synergistic refinement is successfully using the information provided by the DensePose head to improve the alignment accuracy, but there is still space for improvement, e.g. by using a more expressive 3D shape model.

In Table. 1 we complement these curves with three scalar values obtained from the RCP curve for quantitative comparison: the area under the curve(AUC) for 10 and 30 cm errors. We also provide the Failure Rate (FR) at 10cm, as

Evaluation Type	*Base perf.*	$\mathcal{L}_{Kps-2D}$	$\mathcal{L}_{Kps-3D}$	$\mathcal{L}_{DP}$	$\mathcal{L}_{Kps-2D} + \mathcal{L}_{DP}$	$\mathcal{L}_{Kps-3D} + \mathcal{L}_{DP}$
DensePose (FR_{10})	51.1	45.8	52.3	43.3	43.1	44.6
3D Keypoints (*PA MPJPE*)	50.56	52.32	46.78	56.87	51.20	46.52

Table 3: **Comparison of refinement strategies:** we examine the effects of different refinement loss choices on the performance of (i) DensePose estimation in-the-wild, (ii) 3D keypoint localization. We observe that a joint minimization is essential to attaining improvements in both 3D shape estimation and alignment of the body surface with the image domain.

the percentage of points that have above 10cm geodesic error. According to both measures we see that we are doing clearly better than HMR, while getting improvements through the refinement step.

5.3. 3D Keypoint Localization

We report the results of our system on the Human3.6M [16] benchmark. There are two commonly used evaluation protocols with different partitions of the dataset and different evaluation metrics; we report results on both.

We train our system using the frames obtained from subjects S1, S5, S6, S7 and S8 of [16]. In *Protocol 1*, we report mean per-joint position error after rigid alignment of the estimated key points to groundtruth using Procrustes analysis (PA MPJPE). We evaluate every 64-th frame of Subject 11s videos from the frontal camera (C2). In *Protocol 2*, we evaluate on all of the videos of S9 and S11 from all of the cameras, reporting MPJPE without alignment.

The results for both protocols are provided in Table 2. Firstly, the performance of our 3D keypoint branch is quite similar to Sun *et al*. [46] even though the same backbone is shared by a multitude of tasks. Our part-based 3D reconstruction system performs significantly better in both of the evaluation protocols with respect to all of the existing shape reconstruction methods. Specifically, our system improves over HMR, the state-of-the-art system by Kanazawa *et al*. [19] by 5.2 mm (PA MPJPE) in *Protocol 1* and 23.6 mm in *Protocol 2* (MPJPE). Furthermore, we show that the synergistic refinement leads to a further improvement of 4mm in both protocols.

We note that for all of the refinement results we are minimizing the reprojection error to the *estimated* 3D or DensePose estimates based on a CNN, while we assess accuracy in terms of the ground-truth for the respective tasks. This confirms that the CNN-based pose estimates guide the SMPL parameters to more accurate shape estimates.

5.4. Design Choices for Synergistic Refinement

Having validated the improvements attained thanks to the synergistic treatment of 3D shape reconstruction, we now turn to a more detailed ablation of the impact of different loss terms used during refinement. As can be seen in Table 3, using a single loss term results in higher performance for the task at hand, but results in a decrease for the remaining tasks.

For instance, as shown in the third column, minimizing 3D reprojection error improves 3D localization accuracy, but degrades dense pose estimation performance. Similarly, minimizing the DensePose-based loss improves the accuracy of dense correspondence, but results in a drop of 3D joint localization performance. This changes however when a combination of losses is used, where we observe a joint improvement in accuracy in both tasks.

5.5. Qualitative Results

Qualitative results of our system are provided in Fig. 6. We observe that HMR is often distracted by clutter while delivering a pose estimate, whereas our part-based estimate is visibly more accurate; the refinement step further aligns the surface with the image, correcting in particular limb estimates. Qualitative results of our system trained with SMPL+H model[37] is demonstrated in Fig. 7 for the multi-person case.

6. Conclusions and future work

In this work we have proposed HoloPose, a method that uses a tight prior model of human shape in tandem with a multitude of pose estimation methods to derive accurate monocular 3D human reconstruction. We have taken into account the articulated nature of the human body, showing that it substantially improves performance over a monolithic baseline, and have introduced a refinement procedure that allows to iteratively adapt the shape prediction results of a single-shot system so as to meet geometric constraints imposed by complementary, fully-convolutional networks.

In the future we intend to explore neural mesh synthesis models; the use of a distributed representation could more easily accommodate multi-modal distributions, encompassing male, female and child surfaces, which are currently treated by separate shape models. Furthermore, our approach could benefit from a more accurate modeling of geometry, for instance by incorporating perspective projection, surface normal and contour information [4], while we anticipate that the use of depth data, multiple views, [18] or temporal information [53] can help disambiguate 3D reconstruction errors.

Figure 6: **Qualitative results.** From left to right: (a) Input image, (b) HMR [20] results, (c) Our results, without refinement, (d) the visualization of the refinement, (e) our results, refined, (f) our results, refined, 3D rotated.

Figure 7: **Multi-person results.** We show the reconstructed 3D surfaces of multiple persons as colored surfaces (top), and surface normals (bottom). Videos of multi-person 3D reconstruction can be found in http://arielai.com/holopose

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[2] Brett Allen, Brian Curless, Zoran Popovic, and Aaron Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2006, Vienna, Austria, September 2-4, 2006*, pages 147–156, 2006. 1

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 6

[4] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 1, 7

[5] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, 2003. 1

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 1, 6

[7] João Carreira, Sara Vicente, Lourdes Agapito, and Jorge Batista. Lifting object detection datasets into 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1342–1355, 2016. 1

[8] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 3

[9] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 1, 4

[10] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4, 6

[11] Riza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, 2017. 3

[12] Bjoern Haefner, Yvain Quau, Thomas Mllenhoff, and Daniel Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] B. K.P. Horn. Shape from shading. Technical report, Cambridge, MA, USA, 1970. 1

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 6, 7

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*. 3

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 7

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4, 6, 7

[20] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 386–402, 2018. 1, 6, 8

[21] Abhishek Kar, Shubham Tulsiani, João Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1966–1974, 2015. 1

[22] Stefan Kinauer, Riza Alp Güler, Siddhartha Chandra, and Iasonas Kokkinos. Structured output prediction and learning for deep monocular 3d human pose estimation. In *EMMCVPR*, 2017. 2

[23] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3559–3568, 2018. 1

[24] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 3, 2017. 1, 6

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2, 5

[27] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *arXiv preprint arXiv:1710.02322*, 2017. 2, 3

[28] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision*, volume 1, page 5, 2017. 6

[29] Francesc Moreno-Noguer, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Capturing 3d stretchable surfaces from single images in closed form. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1842–1849, 2009. 1

[30] Dat Tien Ngo, Jonas Östlund, and Pascal Fua. Template-based monocular 3d shape recovery using laplacian meshes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):172–187, 2016. 1

[31] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 1, 2, 4, 6

[32] Martin R. Oswald, Eno Töppe, and Daniel Cremers. Fast and globally optimal single view reconstruction of curved objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 534–541, 2012. 1

[33] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1263–1272. IEEE, 2017. 2, 6

[34] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *arXiv preprint arXiv:1805.04092*, 2018. 1, 2, 4, 6

[35] A.I. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In *IEEE International Conference on Computer Vision and Pattern Recognition*, July 2017. 2

[36] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *CoRR*, abs/1803.00455v1, 2018. 6

[37] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245, 2017. 7

[38] Mathieu Salzmann and Pascal Fua. *Deformable Surface 3D Reconstruction from Monocular Images*. Synthesis Lectures on Computer Vision. Morgan & Claypool Publishers, 2010. 1

[39] Jason M. Saragih and Roland Göcke. A nonlinear discriminative approach to AAM fitting. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, 2007. 5

[40] Leonid Sigal, Michael Isard, Horst W. Haussecker, and Michael J. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98(1):15–48, 2012. 1

[41] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning Joint Top-down and Bottom-up Processes for 3D Visual Inference. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006. 1

[42] C. Sminchisescu and B. Triggs. Covariance-Scaled Sampling for Monocular 3D Body Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 447–454, Hawaii, 2001. 1

[43] Cristian Sminchisescu and Bill Triggs. Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics Research*, 22(6):371–391, 2003. 2

[44] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 69–76, Madison, 2003. 1

[45] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 7, 2017. 6

[46] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. *arXiv preprint arXiv:1711.08229*, 2017. 2, 3, 4, 6, 7

[47] Denis Tome, Christopher Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR 2017 Proceedings*, pages 2500–2509, 2017. 2

[48] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4177–4187, 2016. 5

[49] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 1, 2

[50] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 209–217, 2017. 1

[51] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 1, 2, 4

[52] Hsiao-Yu Fish Tung, Adam W. Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *ICCV*, 2017. 1, 2, 4

[53] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 238–245, 2006. 7

[54] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 1, 2, 4, 6

[55] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1

[56] Thomas Vetter, Michael J. Jones, and Tomaso A. Poggio. A bootstrapping algorithm for learning linear models of object classes. In *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico*, pages 40–46, 1997. 1

[57] Sara Vicente and Lourdes Agapito. Balloon shapes: Reconstructing and deforming objects with volume from images. In *2013 International Conference on 3D Vision, 3DV 2013, Seattle, Washington, USA, June 29 - July 1, 2013*, pages 223–230, 2013. 1

[58] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *arXiv preprint arXiv:1804.06208*, 2018. 6

[59] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2664–2673, 2015. 5

[60] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 4

[61] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014. 3

[62] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *European Conference on Computer Vision*, pages 186–201. Springer, 2016. 6

[63] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[64] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 3537–3546, June 2015. 1, 3