Reconstructing NBA Players

Luyang Zhu, Konstantinos Rematas, Brian Curless,
Steven M. Seitz, and Ira Kemelmacher-Shlizerman

University of Washington

Abstract. Great progress has been made in 3D body pose and shape estimation from a single photo. Yet, state-of-the-art results still suffer from errors due to challenging body poses, modeling clothing, and self occlusions. The domain of basketball games is particularly challenging, as it exhibits all of these challenges. In this paper, we introduce a new approach for reconstruction of basketball players that outperforms the state-of-the-art. Key to our approach is a new method for creating poseable, skinned models of NBA players, and a large database of meshes (derived from the NBA2K19 video game) that we are releasing to the research community. Based on these models, we introduce a new method that takes as input a single photo of a clothed player in any basketball pose and outputs a high resolution mesh and 3D pose for that player. We demonstrate substantial improvement over state-of-the-art, single-image methods for body shape reconstruction. Code and dataset are available at http://grail.cs.washington.edu/projects/nba_players/.
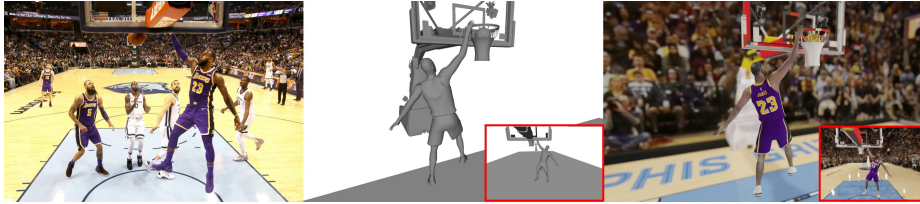
Keywords: 3D Human Reconstruction

1 Introduction

Given regular, broadcast video of an NBA basketball game, we seek a complete 3D reconstruction of the players, viewable from any camera viewpoint. This reconstruction problem is challenging for many reasons, including the need to infer hidden and back-facing surfaces, and the complexity of basketball poses, e.g., reconstructing jumps, dunks, and dribbles.

Human body modeling from images has advanced dramatically in recent years, due in large part to availability of 3D human scan datasets, e.g., CAESAR [56]. Based on this data, researchers have developed powerful tools that enable recreating realistic humans in a wide variety of poses and body shapes [41], and estimating 3D body shape from single images [58, 64]. These models, however, are largely limited to the domains of the source data – people in underwear [56], or clothed models of people in static, staged poses [4]. Adapting this data to a domain such as basketball is extremely challenging, as we must not only match the physique of an NBA player, but also their unique basketball poses.

Sports video games, on the other hand, have become extremely realistic, with renderings that are increasingly difficult to distinguish from reality. The player models in games like NBA2K [6] are meticulously crafted to capture each player's

Fig. 1. Single input photo (left), estimated 3D posed model that is viewed from **a new** camera position (middle), same model with video game texture for visualization purposes. The insets show the estimated shape from the input camera viewpoint. (Court and basketball meshes are extracted from the video game) *Photo Credit: [5]*

physique and appearance (Fig. 3). Such models are ideally suited as a training set for 3D reconstruction and visualization of real basketball games.

In this paper, we present a novel dataset and neural networks that reconstruct high quality meshes of basketball players and retarget these meshes to fit frames of real NBA games. Given an image of a player, we are able to reconstruct the action in 3D, and apply new camera effects such as close-ups, replays, and bullet-time effects (Fig. 1).
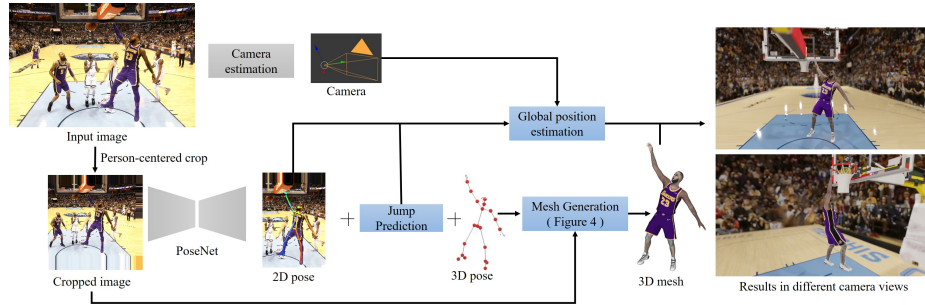
Our new dataset is derived from the video game NBA2K (with approval from the creator, Visual Concepts), by playing the game for hours and intercepting rendering instructions to capture thousands of meshes in diverse poses. Each mesh provides detailed shape and texture, down to the level of wrinkles in clothing, and captures all sides of the player, not just those visible to the camera. Since the intercepted meshes are not rigged, we learn a mapping from pose parameters to mesh geometry with a novel *deep skinning* approach. The result of our skinning method is a detailed deep net basketball body model that can be retargeted to any desired player and basketball pose.

We also introduce a system to fit our retargetable player models to real NBA game footage by solving for 3D player pose and camera parameters for each frame. We demonstrate the effectiveness of this approach on synthetic and real NBA input images, and compare with the state of the art in 3D pose and human body model fitting. Our method outperforms the state-of-the-art methods when reconstructing basketball poses and players even when these methods, to the extent possible, are retrained on our new dataset. This paper focuses on basketball shape estimation, and leaves texture estimation as future work.

Our biggest contributions are, first, a deep skinning approach that produces high quality, pose-dependent models of NBA players. A key differentiator is that we leverage *thousands of poses* and capture detailed geometric variations as a function of pose (e.g., folds in clothing), rather than a small number of poses which is the norm for datasets like CAESAR (1-3 poses/person) and modeling methods like SMPL (trained on CAESAR and $\sim$45 poses/person). While our approach is applicable to any source of registered 3D scan data, we apply it to reconstruct models of NBA players from NBA2K19 game play screen captures. As such, a second key contribution is pose-dependent models of different bas-

ketball players, and raw capture data for the research community. Finally, we present a system that fits these player models to images, enabling 3D reconstructions from photos of NBA players in real games. Both our skinning and pose networks are evaluated quantitatively and qualitatively, and outperform the current state of the art.

One might ask, why spend so much effort reconstructing mesh models that already exist (within the game)? NBA2K's rigged models and in-house animation tools are proprietary IP. By reconstructing a posable model from intercepted meshes (eliminating requirement of proprietary animation and simulation tools), we can provide these best-in-the-world models of basketball players to researchers for the first time (with the company's support). These models provide a number of advantages beyond existing body models such as SMPL. In particular, they capture not just static poses, but human body dynamics for running, walking, and many other challenging activities. Furthermore, the plentiful pose-dependent data enables robust reconstruction even in the presence of heavy occlusions. In addition to producing the first high quality reconstructions of basketball from regular photos, our models can facilitate synthetic data collection for ML algorithms. Just as simulation provides a critical source of data for many ML tasks in robotics, self-driving cars, depth estimation, etc., our derived models can generate much more simulated content under any desired conditions (we can render any pose, viewpoint, combination of players, against any background, etc.)

Fig. 2. Overview: Given a single basketball image (top left), we begin by detecting the target player using [15, 59], and create a person-centered crop (bottom left). From this crop, our PoseNet predicts 2D pose, 3D pose, and jump information. The estimated 3D pose and the cropped image are then passed to mesh generation networks to predict the full, clothed 3D mesh of the target player. Finally, to globally position the player on the 3D court (right), we estimate camera parameters by solving the PnP problem on known court lines and predict global player position by combining camera, 2D pose, and jump information. Blue boxes represent novel components of our method.

2   Related Work

Video Game Training Data. Recent works [55, 54, 38, 53] have shown that, for some domains, data derived from video games can significantly reduce manual labor and labeling, since ground-truth labels can be extracted automatically while playing the game. E.g., [14, 53] collected depth maps of soccer players by playing the FIFA soccer video game, showing generalization to images of real games. Those works, however, focused on low level vision data, e.g., optical flow and depth maps rather than full high quality meshes. In contrast, we collect data that includes 3D triangle meshes, texture maps, and detailed 3D body pose, which requires more sophisticated modeling of human body pose and shape.

Sports 3D reconstruction. Reconstructing 3D models of athletes playing various sports from images has been explored in both academic research and industrial products. Most previous methods use multiple camera inputs rather than a single view. Grau *et al.* [21, 20] and Guillemaut *et al.* [25, 24] used multi-view stereo methods for free viewpoint navigation. Germann *et al.* [18] proposed an articulated billboard presentation for novel view interpolation. Intel demonstrated 360 degree viewing experiences[1], with their True View [2] technology by installing 38 synchronized 5k cameras around the venue and using this multi-view input to build a volumetric reconstruction of each player. This paper aims to achieve similar reconstruction quality but from a *single* image.

Remptas *et al.* [53] reconstructed soccer games from monocular YouTube videos. However, they predicted only depth maps, thus can not handle occluded body parts and player visualization from all angles. Additionally, they estimated players' global position by assuming all players are standing on the ground, which is not a suitable assumption for basketball, where players are often airborne. The detail of the depth maps is also low. We address all of these challenges by building a basketball specific player reconstruction algorithm that is trained on meshes and accounts for complex airborne basketball poses. Our result is a detailed mesh of the player from a single view, but comparable to multi-view reconstructions. Our reconstructed mesh can be viewed from any camera position.

3D human pose estimation. Large scale body pose estimation datasets [31, 44, 42] enabled great progress in 3D human pose estimation from single images [45, 43, 62, 28, 46]. We build on [45] but train on our new basketball pose data, use a more detailed skeleton (35 joints including fingers and face keypoints), and an explicit model of jumping and camera to predict global position. Accounting for jumping is an important step that allows our method outperform state of the art pose.

3D human body shape reconstruction. Parametric human body models [10, 41, 51, 57, 33, 48] are commonly fit to images to derive a body skeleton, and provide a framework to optimize for shape parameters [12, 33, 48, 65, 39, 30, 69]. [64] further 2D warped the optimized parametric model to approximately account for clothing and create a rigged animated mesh from a single photo. [34, 50, 35, 37, 49, 26, 70, 36] trained a neural network to directly regress body shape param-

---

[1] https://www.intel.com/content/www/us/en/sports/technology/true-view.html

eters from images. Most parametric model based methods reconstruct undressed humans, since clothing is not part of the parametric model.

Clothing can be modeled to some extent by warping SMPL [41] models, e.g., to silhouettes: Weng *et al.* [64] demonstrated 2D warping of depth and normal maps from a single photo silhouette, and Alldeick *et al.* [8, 7, 9] addressed multi-image fitting. Alternatively, given predefined garment models [11] estimated a clothing mesh layer on top of SMPL.

Non-parametric methods [63, 47, 58, 52] proposed voxel [63] or implicit function [58] representations to model clothed humans by training on representative synthetic data. Xu *et al.* [67, 68] and Habermann *et al.* [27] assumed a pre-captured multi-view model of the clothed human, retargeted based on new poses.

We focus on single-view reconstruction of players in NBA basketball games, producing a complete 3D model of the player pose and shape, viewable from any camera viewpoint. This reconstruction problem is challenging for many reasons, including the need to infer hidden and back-facing surfaces, and the complexity of basketball poses, e.g., reconstructing jumps, dunks, and dribbles. Unlike prior methods modeling undressed people in various poses or dressed people in a frontal pose, we focus on modeling clothed people under challenging basketball poses and provide a rigorous comparison with the state of the art.

3   The NBA2K Dataset

Fig. 3. Our novel NBA2K dataset examples, extracted from the NBA2K19 video game. Our NBA2K dataset captures 27,144 basketball poses spanning 27 subjects, extracted from the NBA2K19 video game.

Imagine having thousands of 3D body scans of NBA players, in every conceivable pose during a basketball game. Suppose that these models were extremely detailed and realistic, down to the level of wrinkles in clothing. Such a dataset would be instrumental for sports reconstruction, visualization, and analysis. This

section describes such a dataset, which we call *NBA2K*, after the video game from which these models derive. These models of course are not literally player scans, but are produced by professional modelers for use in the NBA2K19 video game, based on a variety of data including high resolution player photos, scanned models and mocap data of some players. While they do not exactly match each player, they are among the most accurate 3D renditions in existence (Fig. 3).

Our NBA2K dataset consists of body mesh and texture data for several NBA players, each in around 1000 widely varying poses. For each mesh (vertices, faces and texture) we also provide its 3D pose (35 keypoints including face and hand fingers points) and the corresponding RGB image with its camera parameters. While we used meshes of 27 real famous players to create many of figures in this paper, we do not have permission to release models of current NBA players. Instead, we additionally collected the same kind of data for 28 synthetic players and retrained our pipeline on this data. The synthetic player's have the same geometric and visual quality as the NBA models and their data along with trained models will be shared with the research community upon publication of this paper. Our released meshes, textures, and models will have the same quality as what's in the paper, and span a similar variety of player types, but not be named individuals. Visual Concepts [6] has approved our collection and sharing of the data.

The data was collected by playing the NBA2K19 game and intercepting calls between the game engine and the graphics card using RenderDoc [3]. The program captures all drawing events per frame, where we locate player rendering events by analyzing the hashing code of both vertex and pixel shaders. Next, triangle meshes and textures are extracted by reverse-engineering the compiled code of the vertex shader. The game engine renders players by body parts, so we perform a nearest neighbor clustering to decide which body part belongs to which player. Since the game engine optimizes the mesh for real-time rendering, the extracted meshes have different mesh topologies, making them harder to use in a learning framework. We register the meshes by resampling vertices in texture space based on a template mesh. After registration, the processed mesh has 6036 vertices and 11576 faces with fixed topology across poses and players (point-to-point correspondence), has multiple connected components (not a watertight manifold), and comes with no skinning information. We also extract the rest-pose skeleton and per-bone transformation matrix, from which we can compute forward kinematics to get full 3D pose.

4   From Single Images to Meshes

Figure 2 shows our full reconstruction system, starting from a single image of a basketball game, and ending with output of a complete, high quality mesh of the target player with pose and shape matching the image. Next, we describe the individual steps to achieve the final results.

4.1   3D Pose in World Coordinates

2D pose, jump, and 3D pose estimation Since our input meshes are not rigged (no skeletal information or blending weights), we propose a neural network called *PoseNet* to estimate the 3D pose and other attributes of a player from a single image. This 3D pose information will be used later to facilitate shape reconstruction. PoseNet takes a single image as input and is trained to output 2D body pose, 3D body pose, a binary jump classification (is the person airborne or not), and the jump height (vertical height of the feet from ground). The two jump-related outputs are key for global position estimation and are our novel addition to existing generic body pose estimation.
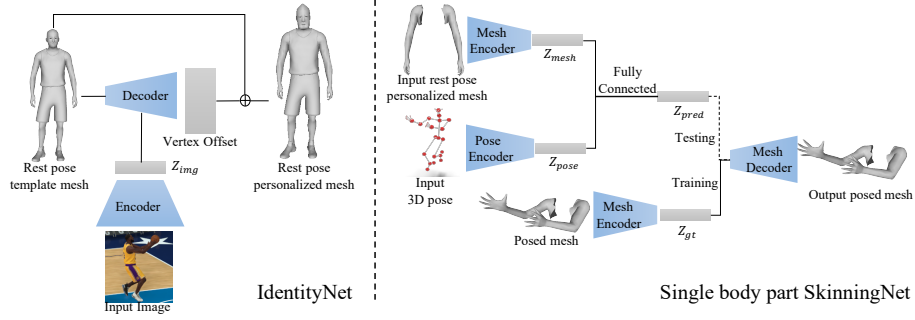
From the input image, we first extract ResNet [66] features (from layer 4) and supply them to four separate network branches. The output of the 2D pose branch is a set of 2D heatmaps (one for each 2D keypoint) indicating where the particular keypoint is located. The output of the 3D pose branch is a set of XYZ location maps (one for each keypoint) [45]. The location map indicates the possible 3D location for every pixel. The 2D and 3D pose branches use the same architecture as [66]. The *jump branch* estimates a class label, and the *jump height branch* regresses the height of the jump. Both networks use a fully connected layer followed by two linear residual blocks [43] to get the final output.

The PoseNet model is trained using the following loss:

$$\mathcal{L}_{pose} = \omega_{2d}\mathcal{L}_{2d} + \omega_{3d}\mathcal{L}_{3d} + \omega_{bl}\mathcal{L}_{bl} + \omega_{jht}\mathcal{L}_{jht} + \omega_{jcls}\mathcal{L}_{jcls} \tag{1}$$

where $\mathcal{L}_{2d} = \|H - \hat{H}\|_1$ is the loss between predicted (H) and ground truth ($\hat{H}$) heatmaps, $\mathcal{L}_{3d} = \|L - \hat{L}\|_1$ is the loss between predicted (L) and ground truth ($\hat{L}$) 3D location maps, $\mathcal{L}_{bl} = \|B - \hat{B}\|_1$ is the loss between predicted (B) and ground truth ($\hat{B}$) bone lengths to penalize unnatural 3D poses (we pre-computed the ground truth bone length over the training data), $\mathcal{L}_{jht} = \|h - \hat{h}\|_1$ is the loss between predicted (h) and ground truth ($\hat{h}$) jump height, and $\mathcal{L}_{jcls}$ is the cross-entropy loss for the jump class. For all experiments, we set $\omega_{2d} = 10$, $\omega_{3d} = 10$, $\omega_{bl} = 0.5$, $\omega_{jht} = 0.4$, and $\omega_{jcls} = 0.2$.

Global Position To estimate the global position of the player we need the camera parameters of the input image. Since NBA courts have known dimensions, we generate a synthetic 3D field and align it with the input frame. Similar to [53, 16], we use a two-step approach. First, we provide four manual correspondences between the input image and the 3D basketball court to initialize the camera parameters by solving PnP [40]. Then, we perform a line-based camera optimization similar to [53], where the projected lines from the synthetic 3D court should match the lines on the image. Given the camera parameters, we can estimate a player's global position on (or above) the 3D court by the lowest keypoint and the jump height. We cast a ray from the camera center through the image keypoint; the 3D location of that keypoint is where the ray-ground height is equal to the estimated jump height.

Fig. 4. Mesh generation contains two sub networks: IdentityNet and SkinningNet. IdentityNet deforms a rest pose template mesh (average rest pose over all players in the database), into a rest pose personalized mesh given the image. SkinningNet takes the rest pose personalized mesh and 3D pose as input and outputs the posed mesh. There is a separate SkinningNet per body part, here we illustrate the arms.

4.2   Mesh Generation

Reconstruction of a complete detailed 3D mesh (including deformation due to pose, cloth, fingers and face) from a single image is a key technical contribution of our method. To achieve this we introduce two sub-networks (Fig. 4): *IdentityNet* and *SkinningNet*. IdentityNet takes as input an image of a player whose rest mesh we wish to infer, and outputs the person's rest mesh by deforming a template mesh. The template mesh is the average of all training meshes and is the same starting point for any input. The main benefit of this network is that it allows us to estimate the body size and arm span of the player according to the input image. SkinningNet takes the rest pose personalized mesh and the 3D pose as input, and outputs the posed mesh. To reduce the learning complexity, we pre-segment the mesh into six parts: head, arms, shirt, pants, legs and shoes. We then train a SkinningNet on each part separately. Finally, we combine the six reconstructed parts into one, while removing interpenetration of garments with body parts. Details are described below.

IdentityNet. We propose a variant of 3D-CODED [22] to deform the template mesh. We first use ResNet [29] to extract features from input images. Then we concatenate template mesh vertices with image features and send them into an AtlasNet decoder [23] to predict per vertex offsets. Finally, we add this offset to the template mesh to get the predicted personalized mesh. We use the L1 loss between the prediction and ground truth to train IdentityNet.

SkinningNet. We propose a TL-embedding network [19] to learn an embedding space with generative capability. Specifically, the 3D keypoints $K_{pose} \in R^{35 \times 3}$ are processed by the pose encoder to produce a latent code $Z_{pose} \in R^{32}$. The rest pose personalized mesh vertices $V_{rest} \in R^{N \times 3}$ (where N is the number of vertices in a mesh part) are processed by the mesh encoder to produce a latent code $Z_{rest} \in R^{32}$. Then Z_{pose} and Z_{rest} are concatenated and fed into a fully

connected layer to get $Z_{pred} \in R^{32}$. Similarly, the ground truth posed mesh vertices $V_{posed} \in R^{N \times 3}$ are processed by another mesh encoder to produce a latent code $Z_{gt} \in R^{32}$. Z_{gt} is sent into the mesh decoder during training while Z_{pred} is sent into the mesh decoder during testing.

The Pose encoder is comprised of two linear residual blocks [43] followed by a fully connected layer. The mesh encoders and shared decoder are built with spiral convolutions [13]. See supplementary material for detailed network architecture. SkinningNet is trained with the following loss:

$$\mathcal{L}_{skin} = \omega_Z \mathcal{L}_Z + \omega_{mesh} \mathcal{L}_{mesh} \tag{2}$$

where $\mathcal{L}_Z = \|Z_{pred} - Z_{gt}\|_1$ forces the space of Z_{pred} and Z_{gt} to be similar, and $\mathcal{L}_{mesh} = \|V_{pred} - V_{posed}\|_1$ is the loss between decoded mesh vertices V_{pred} and ground truth vertices V_{posed}. The weights of different losses are set to $\omega_Z = 5$, $\omega_{mesh} = 50$. See supplementary for detailed training parameters.

Combining body part meshes. Direct concatenation of body parts results in interpenetration between the garment and the body. Thus, we first detect all body part vertices in collision with clothing as in [48], and then follow [60, 61] to deform the mesh by moving collision vertices inside the garment while preserving local rigidity of the mesh. This detection-deformation process is repeated until there is no collision or the number of iterations is above a threshold (10 in our experiments). See supplementary material for details of the optimization.

	HMR [34]	CMR [37]	SPIN [36]	Ours(Reg+BL)	Ours(Loc)	Ours(Loc+BL)
MPJPE	115.77	82.28	88.72	81.66	66.12	**51.67**
MPJPE-PA	78.17	61.22	59.85	63.70	52.73	**40.91**

Table 1. Quantitative comparison of 3D pose estimation to state of the art. The metric is mean per joint position error with (MPJPE-PA) and without (MPJPE) Procrustes alignment. Baseline methods are fine-tuned on our NBA2K dataset.

	HMR [34]	SPIN [36]	SMPLify-X [48]	PIFu [58]	Ours
CD	22.411	14.793	47.720	23.136	**4.934**
EMD	0.137	0.125	0.187	0.207	**0.087**

Table 2. Quantitative comparison of our mesh reconstruction to state of the art. We use Chamfer distance denoted by CD (scaled by 1000, lower is better), and Earth-mover distance denoted by EMD (lower is better) for comparison. Both distance metrics show that our method significantly outperforms state of the art for mesh estimation. All related works are retrained or fine-tuned on our data, see text.

5   Experiments

Dataset Preparation. We evaluate our method with respect to the state of the art on our NBA2K dataset. We collected 27,144 meshes spanning 27 subjects performing various basketball poses (about 1000 poses per player). PoseNet training requires generalization on real images. Thus, we augment the data to 265,765 training examples, 37,966 validation examples, and 66,442 testing examples. Augmentation is done by rendering and blending meshes into various random basketball courts. For IdentityNet and SkinningNet, we select 19,667 examples from 20 subjects as training data and test on 7,477 examples from 7 unseen players. To further evaluate generalization of our method, we also provide qualitative results on real images. Note that textures are extracted from the game and not estimated by our algorithm.
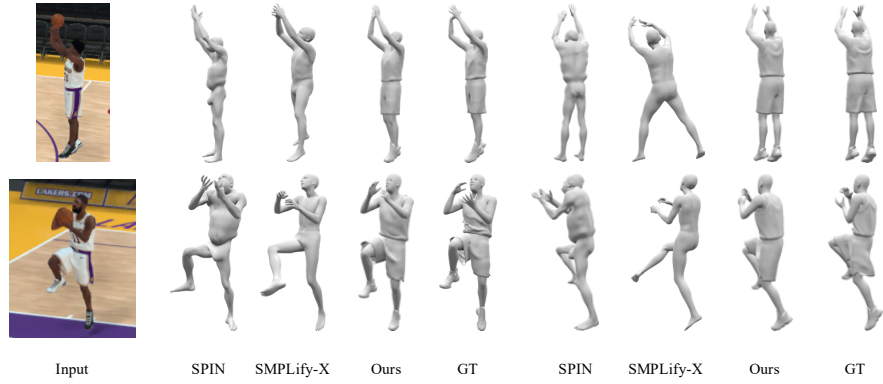
5.1   3D Pose, Jump, and Global Position Evaluation

We evaluate pose estimation by comparing to state of the art SMPL-based methods that released training code. Specifically we compare with HMR [34], CMR [37], and SPIN [36]. For fair comparison, we fine-tuned their models with 3D and 2D ground-truth NBA2K poses. Since NBA2K and SMPL meshes have different topology we do not use mesh vertices and SMPL parameters as part of the supervision. Table 1 shows comparison results for 3D pose. The metric is defined as mean per joint position error (MPJPE) with and without procrustes alignment. The error is computed on 14 joints as defined by the LSP dataset [32]. Our method outperforms all other methods even when they are fine-tuned on our NBA2K dataset (lower number is better).
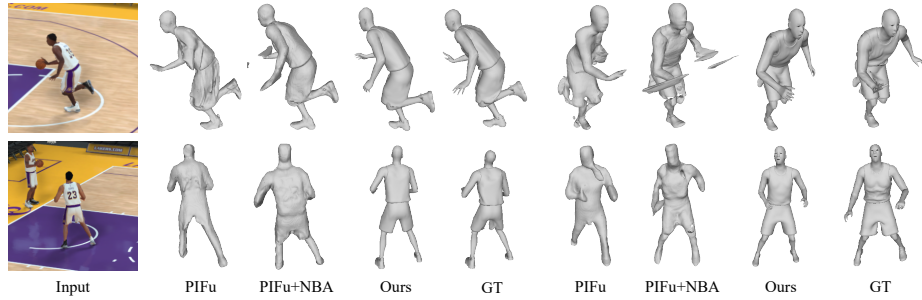
To further evaluate our design choices, we compare the location-map-based representation (used in our network) with direct regression of 3D joints, and also evaluate the effect of bone length (BL) loss on pose prediction. A direct regression baseline is created by replacing our deconvolution network with fully connected layers [43]. The effectiveness of BL loss is evaluated by running the network with and without it. As shown in Table 1, both location maps and BL loss can boost the performance. In supplementary material, we show our results on global position estimation. We can see that our method can accurately place players (both airborne and on ground) on the court due to accurate jump estimation.

5.2   3D Mesh Evaluation

Quantitative Results. Table 2 shows results of comparing our mesh reconstruction method to the state of the art on NBA2K data. We compare to both undressed (HMR [34], SMPLify-X [48], SPIN [36]) and clothed (PIFu [58]) human reconstruction methods. For fair comparison, we retrain PIFU on our NBA2K meshes. SPIN and HMR are based on the SMPL model where we do not have groundtruth meshes, so we fine-tuned with NBA2K 2D and 3D pose. SMPLify-X is an optimization method, so we directly apply it to our testing examples. The meshes generated by baseline methods and the NBA2K meshes do not have

| Input | SPIN | SMPLify-X | Ours | GT | SPIN | SMPLify-X | Ours | GT |

Fig. 5. Comparison with SMPL-based methods. Column 1 is input, columns 2-5 are reconstructions in the image view, columns 6-9 are visualizations from a novel view. Note the significant difference in body pose between ours and SMPL-based methods.

| Input | PIFu | PIFu+NBA | Ours | GT | PIFu | PIFu+NBA | Ours | GT |

Fig. 6. Comparison with PIFu[58]. Column 1 is input, columns 2-5 are reconstructions in the image viewpoint, columns 6-9 are visualizations from a novel view. PIFu significantly over-smooths shape details and produces lower quality reconstruction even when trained on our dataset (PIFu+NBA).

| Input | Prediction | Garment details | Input | Prediction | Garment details |

Fig. 7. Garment details at various poses. For each input image, we show the predicted shape, close-ups from two viewpoints.

one-to-one vertex correspondence, thus we use Chamfer (CD) and Earth-mover (EMD) as distance metrics. Prior to distance computations, all predictions are aligned to ground-truth using ICP. We can see that our method outperforms both undressed and clothed human reconstruction methods even when they are trained on our data.
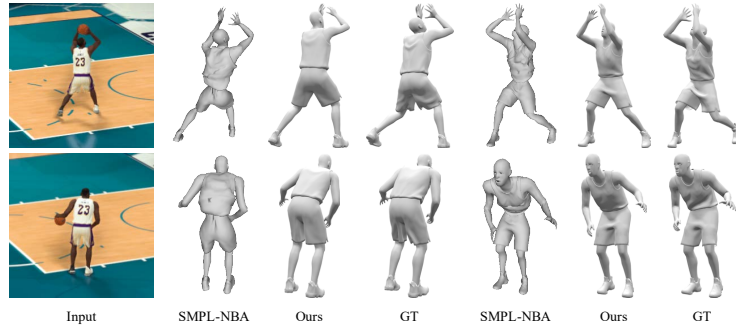

Fig. 8. Results on real images. For each example, column 1 is the input image, 2-3 are reconstructions rendered in different views. 4-5 are corresponding renderings using texture from the video game, just for visualization. Our technical method is focused only on shape recovery. *Photo Credit: [1]*

Qualitative Results. Fig. 5 qualitatively compares our results with the best performing SMPL-based methods SPIN [36] and SMPLify-X [48]. These two methods do not reconstruct clothes, so we focus on the pose accuracy of the body shape. Our method generates more accurate body shape for basketball poses, especially for hands and fingers. Fig. 6 qualitatively compares with PIFu [58], a state-of-the-art *clothed* human reconstruction method. Our method generates detailed geometry such as shirt wrinkles under different poses while PIFu tends to over-smooth faces, hands, and garments. Fig. 7 further visualizes garment details in our reconstructions. Fig. 8 shows results of our method on real images, demonstrating robust generalization. Please also refer to the supplementary pdf and video for high quality reconstruction of real NBA players.

5.3   Ablative Study

Comparison with SMPL-NBA. We follow the idea of SMPL [41] to train a skinning model from NBA2K registered mesh sequences. The trained body model is called SMPL-NBA. Since we don't have rest pose meshes for thousands of different subjects, we cannot learn a meaningful PCA shape basis as SMPL did. Thus, we focus on the pose dependent part and fit the SMPL-NBA model to 2000 meshes of a single player. We use the same skeleton rig as SMPL to drive the mesh. Since our mesh is comprised of multiple connected parts, we initialize the skinning weights using a voxel-based heat diffusion method [17]. The whole training process of SMPL-NBA is the same as the pose parameter training of

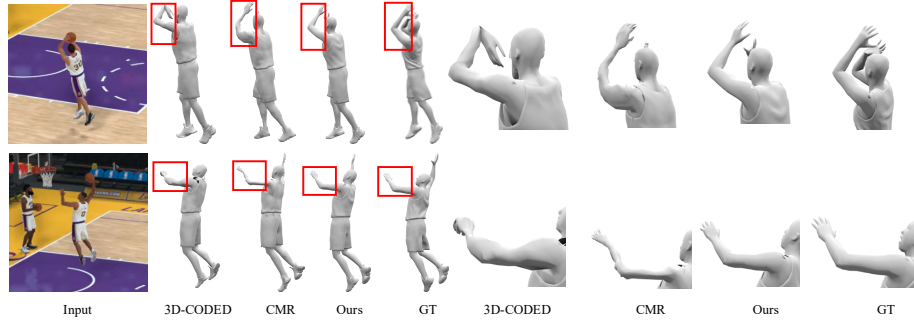| Input | SMPL-NBA | Ours | GT | SMPL-NBA | Ours | GT |

Fig. 9. Comparison with SMPL-NBA. Column 1 is input, columns 2-4 are reconstructions in the image view, columns 5-7 are visualizations from a novel viewpoint. SMPL-NBA fails to model clothing and the fitting process is unstable.

SMPL. We fit the learned model to predicted 2D keypoints and 3D keypoints from PoseNet following SMPLify [12]. Fig. 9 compares SkinningNet with SMPL-NBA, showing that SMPL-NBA has severe artifacts for garment deformation – an inherent difficulty for traditional skinning methods. It also suffers from twisted joints which is a common problem when fitting per bone transformation to 3D and 2D keypoints.

	CMR [37]	3D-CODED [22]	Ours
MPVPE	85.26	84.22	**76.41**
MPVPE-PA	64.32	63.13	**54.71**

Table 3. Quantitative comparison with 3D-CODED [22] and CMR [37]. The metric is mean per vertex position error in mm with (MPVPE-PA) and without (MPVPE) Procrustes alignment. All baseline methods are trained on the NBA2K data.

Comparison with Other Geometry Learning Methods. Fig. 10 compares SkinningNet with two state of the art mesh-based shape deformation networks: 3D-CODED [22] and CMR [37]. The baseline methods are retrained on the same data as SkinningNet for fair comparison. For 3D-CODED, we take 3D pose as input instead of a point cloud to deform the template mesh. For CMR, we only use their mesh regression network (no SMPL regression network) and replace images with 3D pose as input. Both methods use the same 3D pose encoder as SkinningNet. The input template mesh is set to the prediction of IdentityNet. Unlike baseline methods, SkinningNet does not suffer from substantial deformation errors when the target pose is far from the rest pose. Table 3 provides further quantitative results based on mean per vertex position error (MPVPE) with and without procrustes alignment.

Fig. 10. Comparison with 3D-CODED [22] and CMR [37]. Column 1 is input, columns 2-5 are reconstructions in the image view, columns 6-9 are zoomed-in version of the red boxes. The baseline methods exhibit poor deformations for large deviations from the rest pose.

6   Discussion

We have presented a novel system for state-of-the-art, detailed 3D reconstruction of complete basketball player models from single photos. Our method includes 3D pose estimation, jump estimation, an identity network to deform a template mesh to the person in the photo (to estimate rest pose shape), and finally a skinning network that retargets the shape from rest pose to the pose in the photo. We thoroughly evaluated our method compared to prior art; both quantitative and qualitative results demonstrate substantial improvements over the state-of-the-art in pose and shape reconstruction from single images. For fairness, we retrained competing methods to the extent possible on our new data. Our data, models, and code will be released to the research community.

Limitations and future work This paper focuses solely on high quality shape estimation of basketball players, and does not estimate texture – a topic for future work. Additionally IdentityNet can not model hair and facial identity due to lack of details in low resolution input images. Finally, the current system operates on single image input only; a future direction is to generalize to video with temporal dynamics.

References

1. Getty Images, https://www.gettyimages.com
2. Intel True View, www.intel.com/content/www/us/en/sports/technology/true-view.html

3. RenderDoc, https://renderdoc.org
4. RenderPeople, https://renderpeople.com
5. USA TODAY Network, https://www.commercialappeal.com
6. VISUAL CONCEPTS, https://vcentertainment.com
7. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
8. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
9. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: IEEE International Conference on Computer Vision (ICCV). IEEE (2019)
10. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM transactions on graphics (TOG). vol. 24, pp. 408–416. ACM (2005)
11. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
12. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: 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)
13. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
14. Calagari, K., Elgharib, M., Didyk, P., Kaspar, A., Matuisk, W., Hefeeda, M.: Gradient-based 2-d to 3-d conversion for soccer videos. In: ACM Multimedia. pp. 605–619 (2015)
15. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
16. Carr, P., Sheikh, Y., Matthews, I.: Pointless calibration: Camera parameters from gradient-based alignment to edge images. In: WACV (2012)
17. Dionne, O., de Lasa, M.: Geodesic voxel binding for production character meshes. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 173–180. ACM (2013)
18. Germann, M., Hornung, A., Keiser, R., Ziegler, R., Würmlin, S., Gross, M.: Articulated billboards for video-based rendering. In: Computer Graphics Forum. vol. 29, pp. 585–594. Wiley Online Library (2010)
19. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: European Conference on Computer Vision. pp. 484–499. Springer (2016)
20. Grau, O., Hilton, A., Kilner, J., Miller, G., Sargeant, T., Starck, J.: A free-viewpoint video system for visualization of sport scenes. SMPTE motion imaging journal **116**(5-6), 213–219 (2007)
21. Grau, O., Thomas, G.A., Hilton, A., Kilner, J., Starck, J.: A robust free-viewpoint video system for sport scenes. In: 2007 3DTV conference. pp. 1–4. IEEE (2007)

22. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: 3d-coded: 3d corre-spondences by deep deformation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 230–246 (2018)
23. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: A papier-mâché approach to learning 3d surface generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 216–224 (2018)
24. Guillemaut, J.Y., Hilton, A.: Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. IJCV (2011)
25. Guillemaut, J.Y., Kilner, J., Hilton, A.: Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes. In: ICCV (2009)
26. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
27. Habermann, M., Xu, W., , Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics, (Proc. SIGGRAPH) (2019)
28. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
30. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 International Conference on 3D Vision (3DV). pp. 421–430. IEEE (2017)
31. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (2013)
32. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference (2010), doi:10.5244/C.24.12
33. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8320–8329 (2018)
34. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018)
35. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Computer Vision and Pattern Regognition (CVPR) (2019)
36. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
37. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)
38. Krähenbühl, P.: Free supervision from video games. In: CVPR (2018)
39. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6050–6059 (2017)

40. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o (n) solution to the pnp problem. International journal of computer vision **81**(2), 155 (2009)
41. 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), 248 (2015)
42. 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: Proceedings of the European Conference on Computer Vision (ECCV). pp. 601–617 (2018)
43. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649 (2017)
44. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision (3DV). pp. 506–516. IEEE (2017)
45. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4), 44 (2017)
46. Moon, G., Chang, J., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: The IEEE Conference on International Conference on Computer Vision (ICCV) (2019)
47. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: Siclope: Silhouette-based clothed people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4480–4490 (2019)
48. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
49. Pavlakos, G., Kolotouros, N., Daniilidis, K.: Texturepose: Supervising human mesh estimation with texture consistency. In: ICCV (2019)
50. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 459–468 (2018)
51. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. ACM Transactions on Graphics, (Proc. SIGGRAPH) **34**(4), 120:1–120:14 (Aug 2015)
52. Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3DPeople: Modeling the Geometry of Dressed Humans. In: ICCV (2019)
53. Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., Seitz, S.: Soccer on your tabletop. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4738–4747 (2018)
54. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: ICCV (2017)
55. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016)
56. Robinette, K.M., Blackwell, S., Daanen, H., Boehmer, M., Fleming, S.: Civilian American and European Surface Anthropometry Resource (CAESAR), final report. volume 1. summary. Tech. rep., SYTRONICS INC DAYTON OH (2002)
57. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)

58. Saito, S., , Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. arXiv preprint arXiv:1905.05172 (2019)
59. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)
60. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. In: Symposium on Geometry processing. vol. 4, pp. 109–116 (2007)
61. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.P.: Laplacian surface editing. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing. pp. 175–184 (2004)
62. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 529–545 (2018)
63. 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)
64. Weng, C.Y., Curless, B., Kemelmacher-Shlizerman, I.: Photo wake-up: 3d character animation from a single photo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5908–5917 (2019)
65. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
66. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV) (2018)
67. Xu, F., Liu, Y., Stoll, C., Tompkin, J., Bharaj, G., Dai, Q., Seidel, H.P., Kautz, J., Theobalt, C.: Video-based characters: Creating new human performances from a multi-view video database. ACM Trans. Graph. **30**(4), 32:1–32:10 (Jul 2011). https://doi.org/10.1145/2010324.1964927, http://doi.acm.org/10.1145/2010324.1964927
68. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. ACM Trans. Graph. (2018)
69. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2148–2157 (2018)
70. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)