3D Human Pose, Shape and Texture from Low-Resolution Images and Videos

Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, László A. Jeni, Fernando De la Torre

Abstract—3D human pose and shape estimation from monocular images has been an active research area in computer vision. Existing deep learning methods for this task rely on high-resolution input, which however, is not always available in many scenarios such as video surveillance and sports broadcasting. Two common approaches to deal with low-resolution images are applying super-resolution techniques to the input, which may result in unpleasant artifacts, or simply training one model for each resolution, which is impractical in many realistic applications.

To address the above issues, this paper proposes a novel algorithm called RSC-Net, which consists of a **R**esolution-aware network, a **S**elf-supervision loss, and a **C**ontrastive learning scheme. The proposed method is able to learn 3D body pose and shape across different resolutions with one single model. The self-supervision loss enforces scale-consistency of the output, and the contrastive learning scheme enforces scale-consistency of the deep features. We show that both these new losses provide robustness when learning in a weakly-supervised manner. Moreover, we extend the RSC-Net to handle low-resolution videos and apply it to reconstruct textured 3D pedestrians from low-resolution input. Extensive experiments demonstrate that the RSC-Net can achieve consistently better results than the state-of-the-art methods for challenging low-resolution images.

Index Terms—3D human pose and shape, low-resolution, neural network, self-supervised learning, contrastive learning.

---◆---

1 INTRODUCTION

3D human pose and shape estimation from 2D images is of great interest to the computer vision and graphics community. Whereas significant progress has been made in this field, it is often assumed that the input image is high-resolution and contains sufficient information for reconstructing the 3D human geometry in detail [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. However, this assumption does not always hold in practice, since people in applications of interest, such as surveillance cameras and sports videos, often have low resolutions [13], [14], [15], [16], [17], [18]. As a result, existing algorithms trained with high-resolution images are prone to fail when applied to low-resolution inputs as shown in Figure 1.

In this paper, we study the so-far unexplored problem of estimating 3D human pose and shape from low-resolution images where we find two major challenges. First, the resolutions of the input images in real scenarios vary in a wide range, and a network trained for one specific resolution does not always work well for another. One might consider overcoming this problem by simply training different models, one for each image resolution. However, this is impractical in terms of memory and training computation. Alternatively, one could super-resolve the images to a sufficiently large resolution, but the super-resolution step often brings in unpleasant artifacts, which can result in significant domain gap between the super-resolved images and real high-

resolution images, and thereby causes low reliability of 3D estimation. To address this issue, we propose a resolution-aware deep neural network for 3D human pose and shape estimation that is robust to different image resolutions. Our network builds upon two main components: a feature extractor shared across different resolutions and a set of resolution-dependent parameters to adaptively integrate the different-level features.

Another challenge we encounter is due to the fact that high-quality 3D annotations are hard to obtain, especially for in-the-wild data, and only a small portion of the training images have 3D ground truth labels [4], [7], which brings difficulties for the training process. Whereas most training images have 2D keypoint labels, they are usually not sufficient for training the network due to the inherent ambiguities in the 2D-to-3D mapping. This weak-supervision problem is further accentuated in our task, as the low-resolution 3D estimation is not well constrained and has a large solution space due to limited pixel observations. Therefore, directly training low-resolution models with weak supervision does not typically achieve good results. Inspired by the self-supervised learning [19], [20], we propose a directional self-supervision loss to remedy the above issue. Specifically, we enforce the consistency across the outputs of the same input image with different resolutions, such that the results of the higher-resolution images can act as guidance for lower-resolution input. We will show that this strategy significantly improves the 3D estimation results.

In addition to enforcing output consistency, we also devise an approach to enforce consistency of the feature representations across different resolutions. Nevertheless, we find that the commonly-used mean squared error is not effective in measuring discrepancies between high-dimensional fea-

- *X. Xu, L. Jeni, and F. De la Torre are with Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. F. De la Torre is also with Facebook Reality Lab (Oculus). E-mail: xuxiangyu2014@gmail.com, laszlojeni@cmu.edu, ftorre@cs.cmu.edu.*
- *H. Chen is with Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. E-mail: hchen@cmu.edu.*
- *F. Moreno-Noguer is with Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, 08028, Spain. E-mail: fmoreno@iri.upc.edu.*
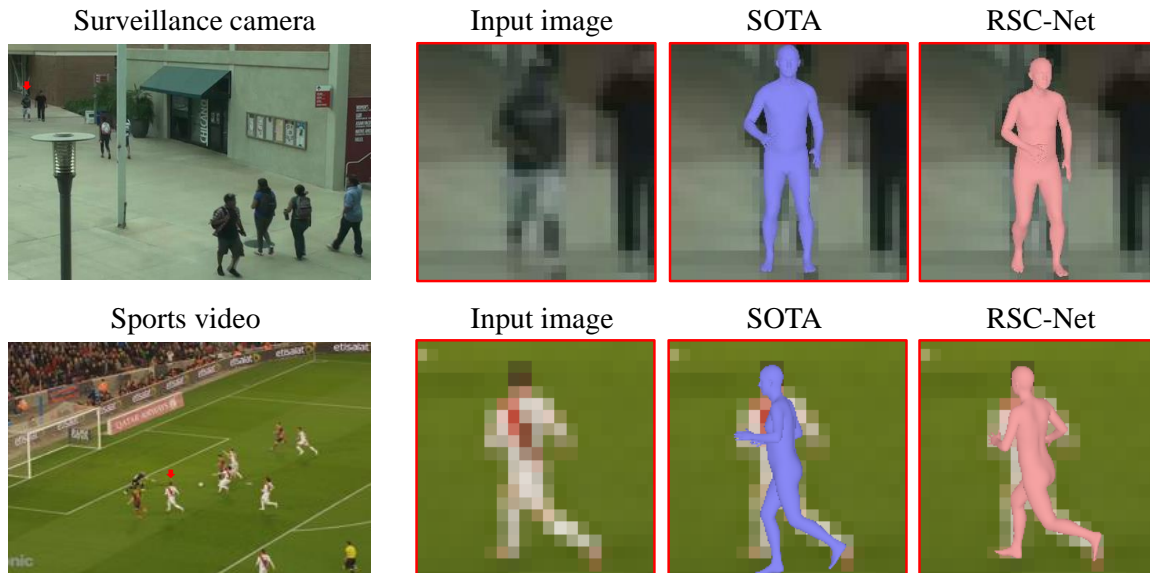
Fig. 1. 3D human pose and shape estimation from real low-resolution images captured from a surveillance camera (top) and a sports video (bottom). SOTA method [7] that works well for high-resolution images performs poorly at low-resolution ones (note the orientation of the estimated human face and the posture of the arms and legs).

ture vectors. Instead, we adapt the contrastive learning [21], [22], [23] which aims to maximize the mutual information across the feature representations at different resolutions, and encourages the network to produce better features for the low-resolution input.

To summarize, we make the following contributions in this work. First, we study the unexplored problem of 3D human pose and shape estimation from low-resolution images and present a simple yet effective solution called RSC-Net, which is based on a novel resolution-aware network that can handle arbitrary-resolution input with one single model. To the best of our knowledge, there is no prior work that solves this practically important problem. Second, we propose a self-supervision loss to address the issue of weak supervision. Third, we introduce contrastive learning for more effectively enforcing the consistency of the feature vectors across different resolutions. Extensive experiments demonstrate that the proposed method outperforms the state-of-the-art algorithms on challenging low-resolution inputs and achieves robust performance for high-quality 3D human pose and shape estimation.

A preliminary version of this work is published in [24]. Here we further extend the paper in the following aspects. First, we provide a more thorough analysis of the proposed RSC-Net with more qualitative examples and more quantitative evaluations on benchmark datasets. Second, while the main contribution in this work is a single-frame model for 3D human pose and shape estimation, we also extend the RSC-Net to handle low-resolution videos by adding a simple temporal post-processing step. In addition, we show an important application of the RSC-Net, *i.e.* reconstructing textured 3D human from a low-resolution pedestrian image, by introducing an effective texture estimation network.

2 RELATED WORK

We first review the state-of-the-art methods for 3D human pose and shape estimation and then discuss the low-resolution image recognition algorithms that are related to this work.

2.1 3D Human pose and shape Estimation

Recent years have witnessed significant progress in the field of 3D human pose and shape estimation from a single image [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [25], [26], [27], [28], [29], [30], [31], [32]. Existing methods for this task can be broadly categorized into two classes. The first kind of approaches generally splits the 3D human estimation process into two stages: first transforming the input image into new representations, such as human 2D keypoints [1], [2], [3], [8], [9], [26], human silhouettes [2], [8], [9], body part segmentations [1], [32], UV mappings [25], and optical flow [26], and then regressing the 3D human parameters [33] from the transformed outputs of the last stage either with iterative optimization [2], [3] or neural networks [1], [8], [9], [26]. For instance, Omran *et al.* propose a novel approach called Neural Body Fitting (NBF) which creatively integrates a statistical body model within a CNN, successfully leveraging both semantic body part segmentation and constraints of human bodies. As these methods map the original input images into simpler representation forms which are generally low-dimensional (*e.g.* sparse [34], [35]) and can be easily rendered, they can exploit a large amount of synthetic data (or data captured in constrained scenes) for training where there are sufficient high-quality 3D labels. However, these two-stage systems are error-prone, as the errors from early stage may be accumulated or even deteriorated [4]. In addition, the intermediate results may throw away valuable information in the image such as context. More importantly, the task of the first stage, *i.e.* to estimate the intermediate representations, is usually difficult

for low-resolution images, and thereby, the aforementioned two-stage models are not suitable to solve our problem of low-resolution 3D human pose and shape estimation.

Without relying on new representations, the second kind of approaches can directly regress the 3D parameters from the input image [4], [5], [6], [7], [10], [11], [27], where most methods are based on deep neural networks. While being concise and not requiring the estimation of intermediate results, these methods usually suffer from the problem of weak supervision due to a lack of high-quality 3D ground truth. Most existing works focus on this problem and have developed different techniques to solve it. As a typical example, Kanazawa *et al*. [4] include a generative adversarial network (GAN) [36] to constrain the solution space using the prior learned from 3D human data. However, we find the GAN-based algorithm less effective for low-resolution input images where substantially fewer pixels are available. Kolotouros [7] *et al*. integrate the optimization-based method [3] into the training process of the deep network to more effectively exploit the 2D keypoints. While achieving good improvements over [4] on high-resolution images, [7] cannot be easily applied to low-resolution input, as the low-resolution network cannot provide good initial results to start the optimization loop. In addition, it significantly increases the training time. Different from the above methods, we propose a low-resolution 3D human pose and shape estimation algorithm using a single image as input. We propose a self-supervision loss and a contrastive feature loss which effectively remedy the problem of insufficient 3D supervision.

Note that temporal information has also been exploited to enforce the temporal consistency of the 3D estimation results [5], [6], [27], which however requires high-resolution video input. In this work, we show that the proposed RSC-Net can be easily extended to exploit multi-frame information from low-resolution videos.

2.2 Low-Resolution Image Recognition

While there is no prior work for low-resolution 3D human pose and shape estimation, there are some related approaches to process low-resolution inputs for other image recognition tasks, such as 2D body pose estimation [16], face recognition [37], [38], image classification [14], image retrieval [39], [40], and object detection [41], [42]. Most of these methods address the low-resolution issue by enhancing the degraded input, in either the image space [14], [38], [41] or the feature space [37], [39], [40], [42]. One typical image-space method [41] applies a super-resolution network which is trained to improve both the image quality (*i.e.* per-pixel similarity such as PSNR) and the object detection performance. However, the loss functions for higher PSNR and better recognition performance do not always agree with each other, which may lead to inferior solutions. Moreover, the super-resolution model may bring unpleasant artifacts (*e.g.* over-smoothing and checkerboard patterns [13]), resulting in domain gap between the super-resolved and the real high-resolution images. Unlike the image enhancement based approaches, feature enhancement based methods [37], [39], [40], [42] are not distracted by the image quality loss and thus can better focus on improving the recognition

performance. As a representative example, Ge *et al*. [37] use mean squared error (MSE) to enforce the similarity between the features of low-resolution and high-resolution images, obtaining good results for face recognition. Different from the above approaches, Neumann *et al*. [16] propose a novel method for low-resolution 2D body pose estimation by predicting a probability map with Gaussian Mixture Model, which, however, cannot be easily extended to 3D human pose and shape estimation. In this work, we apply the feature enhancement strategy to low-resolution 3D human pose and shape estimation. Instead of using MSE for measuring feature similarity, we introduce the contrastive learning [21] which can more effectively maximize the mutual information across the features of different resolutions. In addition, we handle different-resolution inputs with a resolution-aware neural network.

3 ALGORITHM

We study the problem of 3D human pose and shape estimation from a single low-resolution image. We propose a simple yet effective solution (RSC-Net) for this problem, which consists of a resolution-aware network and new training losses. An overview of the RSC-Net is shown in Figure 2. In this section, we first introduce the 3D human representation model (Section 3.1) and the baseline network for 3D human estimation from a single 2D image (Section 3.2). We then describe the proposed resolution-aware model (Section 3.2) as well as the self-supervision loss (Section 3.3) and the contrastive learning strategy for training the network (Section 3.4). In addition, we introduce a temporal recurrent module to extend the proposed single-frame model to low-resolution videos (Section 3.5). We also show how the RSC-Net can be used to reconstruct textured human from a low-resolution pedestrian image (Section 3.6).

3.1 3D Human pose and shape Representation

We represent the 3D human body using the Skinned Multi-Person Linear (SMPL) model [33]. The SMPL is a parametric model which describes the body pose and shape with two sets of parameters β and θ, respectively. The body shape is represented by a basis in a low-dimensional shape space learned from a training set of 3D human scans, and the parameters $\beta \in \mathbb{R}^{10}$ are coefficients of the basis vectors. The body pose is defined by a skeleton rig with $K = 24$ joints including the body root, and the pose parameters $\theta \in \mathbb{R}^{3K}$ are the axis-angle representations of the relative rotation between different body parts as well as the global rotation of the body root. With β and θ, we can obtain the 3D body mesh:

$$M = f_{\text{SMPL}}(\beta, \theta), \qquad (1)$$

where $M \in \mathbb{R}^{N \times 3}$ is a triangulated mesh with $N = 6890$ vertices.

Similar to prior works [4], [7], we can predict the 3D locations of the body joints X with the body mesh using a pre-trained mapping matrix $W \in \mathbb{R}^{K \times N}$:

$$X \in \mathbb{R}^{K \times 3} = WM. \qquad (2)$$

With the 3D human joints, we can then use a perspective camera model to project the body joints from 3D to 2D.
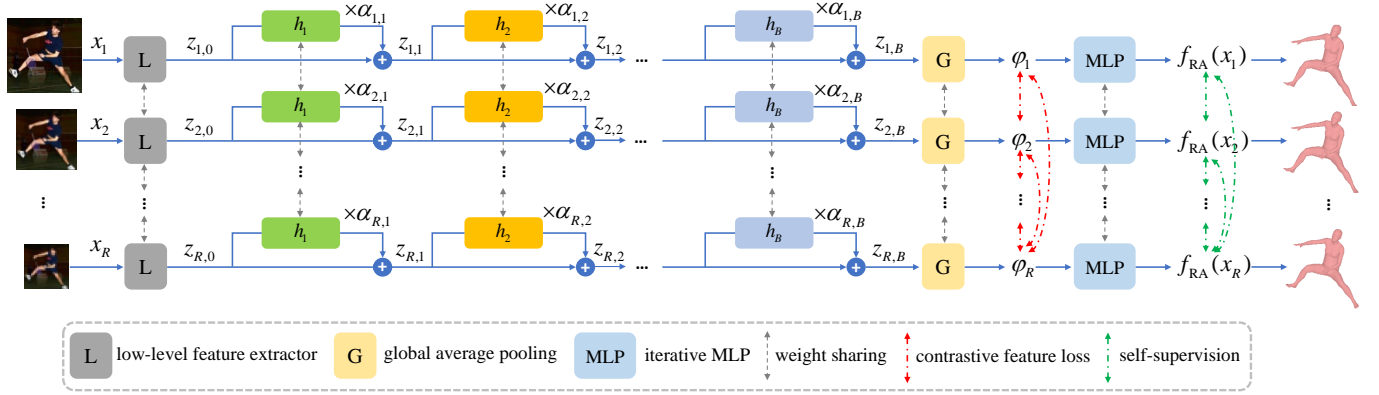
Fig. 2. Overview of the proposed RSC-Net. The resolution-aware network f_{RA} is trained with a combination of the basic loss (omitted in the figure for simplicity), self-supervision loss and contrastive feature loss. The modules with the same colors are shared across different resolutions, while the matrix α is resolution-dependent. $\{x_i\}_{i=1}^{R}$ represents the different resolution versions of the same image, where R is the number of resolutions considered in this work. We resize the different resolution inputs with bicubic interpolation before feeding them into the network. Note that in real implementation, we split the R resolutions into P ranges to ease the training burden (see the text in Section 3.2). During the training phase, we jointly train the RSC-Net with all different resolution images. During inference, we first decide the resolution range of the input image and then choose the suitable row of parameters in α for usage in the network.

Assuming the camera parameters are $\delta \in \mathbb{R}^3$ which define the 3D translation of the camera, the 2D keypoints can be written as:

$$J \in \mathbb{R}^{K \times 2} = f_{\text{project}}(X, \delta), \qquad (3)$$

where f_{project} is the perspective projection function [43].

3.2 Resolution-Aware 3D Human Estimation

Baseline network. Similar to the existing methods [4], [7], we use the deep convolutional neural network (CNN) for 3D human estimation, where the ResNet-50 [44] is employed to extract features from the input image. The building block of the ResNet (*i.e.* ResBlock [45]) can be formulated as:

$$z_k = z_{k-1} + h_k(z_{k-1}), \qquad (4)$$

where z_k represents the output feature maps of the k-th ResBlock, and h_k is the nonlinear function used to learn the feature residuals, which is modeled by several convolutional layers with ReLU activation [46]. The ResNet stacks B ResBlocks together, and recursively we will have:

$$
\begin{aligned}
z_B &= z_{B-1} + h_B(z_{B-1}) \\
&= z_{B-2} + h_{B-1}(z_{B-2}) + h_B(z_{B-1}) \\
&= z_{B-3} + h_{B-2}(z_{B-3}) + h_{B-1}(z_{B-2}) + h_B(z_{B-1}) \\
&\cdots \\
&= z_0 + \sum_{k=1}^{B} h_k(z_{k-1}), \qquad (5)
\end{aligned}
$$

where z_0 is the low-level feature maps extracted from the input image x with convolutional layers, and z_B is a combination of different level residual maps from all the ResBlocks. Note that we do not explicitly consider the downsampling ResBlocks in Eq. 5 for clarity. With the output features of the ResNet, we can use global average pooling to obtain a feature vector φ and employ an iterative MLP for regressing the 3D parameters β, θ, δ similar to [4], [7]. Specifically, the iterative MLP regresses the parameters in an iterative error feedback loop, where progressive changes are made to the

current estimate. Thus, the network can learn residuals to recurrently refine the initial output, leading to more accurate results. We refer to [4] for more details.

Resolution-aware network. The baseline network is originally designed for high-resolution images with input size of 224×224 pixels, whereas the image resolutions for human in real scenarios can be much lower and vary in a wide range. A straightforward way to deal with these low-resolution inputs is to train different networks for all possible resolutions and choose the suitable one for each test image. However, this is not practical for real applications given the high computational complexity and memory requirement.

To solve this problem, we propose a resolution-aware network, and the main idea is that the different-resolution images with the same contents are largely similar as shown in Figure 2, and can share most parts of the feature extractor. And only a small amount of parameters are needed to be resolution-dependent to account for the characteristics of different image resolutions. Towards this end, instead of directly combining the different level features as in Eq. 5, we learn a matrix α to adaptively fuse the residual maps from the ResBlocks for each resolution as shown in Figure 2. This allows different resolution inputs to have tailored feature extractors for better 3D estimation. Specifically, we formulate the output of the proposed resolution-aware network as:

$$z_{i,B} = z_{i,0} + \sum_{k=1}^{B} \alpha_{i,k} h_k(z_{i,k-1}), \quad i = 1, 2, \ldots, R, \qquad (6)$$

where i is the index for different image resolutions, and larger i indicates smaller image. $i = 1$ corresponds to the original high-resolution input. $\alpha \in \mathbb{R}^{R \times B}$, where R denotes the number of all the image resolutions considered in this work. $z_{i,k}$ and $\alpha_{i,k}$ respectively represent the output and the fusion weight of the k-th ResBlock for the i-th input resolution. According to Eq. 6, the original ResBlock in Eq. 4 is modified as:

$$z_{i,k} = z_{i,k-1} + \alpha_{i,k} h_k(z_{i,k-1}). \qquad (7)$$

Note that we use a slightly different notation here compared with Eq. 4 and Eq. 5 which do not have the index i for image resolution, as the baseline network is not resolution-aware and applies the same operations to different resolution inputs. This new formulation of the ResBlock in Eq. 7 is in form similar to the concurrent work [47], and the difference is that we use α for resolution-aware feature extractors while [47] uses it for fast convergence in training networks.

To train the above network, we need to conduct the downsampling operation to each high-resolution image in the training dataset for $R-1$ times, such that each row of parameters in α have their corresponding training data. Whereas the original training datasets [48], [49], [50], [51], [52] are already quite large for the diversity of the training images, it will be further augmented by $R-1$ times, which significantly increases the computational burden of the training process. To alleviate the training issues, we divide all the R resolutions into P ranges and only learn one set of parameters for each range. We design the first resolution range to only have the original high-resolution image, and for the other ranges, we randomly sample a resolution in each range during each training iteration. The training images with different resolutions can be denoted as $\{x_i\}_{i=1}^P$ where the smaller images $x_2, x_3, \ldots, x_P$ are synthesized from the same high-resolution image x_1 with bicubic interpolation. With this strategy, the training set can be much smaller without losing diversity, and we can have a lower-dimensional matrix $\alpha \in \mathbb{R}^{P \times B}$, where the number of parameters can be reduced from RB to PB. During inference, we first decide the resolution range of the input image and then choose the suitable row of parameters in α for usage in the network.

Progressive training. Directly using different resolution images for training all at once can lead to difficulties in optimizing the proposed model since the network needs to handle inputs with complex resolution properties simultaneously. Instead, we train the proposed network in a progressive manner, where the higher-resolution images are easier to handle and thus first processed in training, and more challenging ones with lower resolutions are subsequently added. In this way, we alleviate the difficulty of the training process and the proposed model can evolve progressively. Note that this strategy is related to the curriculum learning algorithm [53] in that difficult examples are gradually emphasized. However, [53] does not explicitly consider multi-resolution training.

Basic loss function. Similar to the previous works [4], [7], the basic loss of our network is a combination of 3D and 2D losses. Suppose the output of the proposed network for input image x_i is $[\hat{\beta}_i, \hat{\theta}_i, \hat{\delta}_i] = f_{\mathrm{RA}}(x_i)$ where i is the resolution index, and $X_g, J_g, \beta_g, \theta_g$ are the ground truth 3D and 2D keypoints and SMPL parameters. The basic loss function can be written as:

$$L_{\mathrm{b}} = \sum_i \|[\hat{\beta}_i, \hat{\theta}_i] - [\beta_{\mathrm{g}}, \theta_{\mathrm{g}}]\|_2^2$$
$$+ \lambda_1 \|\hat{X}_i - X_{\mathrm{g}}\|_2^2 + \lambda_2 \|\hat{J}_i - J_{\mathrm{g}}\|_2^2, \qquad (8)$$

where $\hat{X}_i$ and $\hat{J}_i$ are estimated with Eq. 2 and Eq. 3, respectively. λ_1 and λ_2 are hyper-parameters for balancing different terms. Note that while all the training images have

2D keypoint labels J_g in Eq. 8, only a limited portion of them have 3D ground truth X_g, β_g, θ_g. For the training images without 3D labels, we simply omit the first two terms in Eq. 8 similar to [4], [5], [7].

3.3 Self-Supervised Loss

The 3D human pose and shape estimation is usually posed as a weakly-supervised problem as only a small part of the training data has 3D labels, and this is especially the case for in-the-wild images where accurate 3D annotations cannot be easily captured. This issue gets even worse for the low-resolution images, as the 3D estimation is not well constrained by limited pixel observations and requires strong supervision signal during training to find a good solution.

To remedy this problem, we propose a self-supervision loss to assist the basic loss for training the resolution-aware network f_{RA}. This new loss term is inspired by the self-supervised learning algorithm [19] which improves the training by minimizing the MSE between the network predictions under different input augmentation conditions. For our problem, we naturally have the same input with different data augmentations, *i.e.* the different-resolution images synthesized from the same high-resolution image. Thus, the self-supervision loss can be formulated by enforcing the consistency across the outputs of different image resolutions:

$$\sum_{i,j} \|f_{\mathrm{RA}}(x_i) - f_{\mathrm{RA}}(x_j)\|_2^2. \qquad (9)$$

However, a major difference between our work and the original self-supervision method [19] is that we are generally more confident in the predictions of the higher-resolution images, while [19] treats the results under different input augmentations equally. To exploit this prior knowledge, we improve the loss in Eq. 9 and propose a directional self-supervision loss:

$$L_{\mathrm{s}} = \sum_{i,j} w_{i,j} \|\bar{f}_{\mathrm{RA}}(x_i) - f_{\mathrm{RA}}(x_j)\|_2^2,$$
$$w_{i,j} = \mathbb{1}(j - i > 0) \cdot (j - i), \qquad (10)$$

where $w_{i,j}$ is the loss weight for an image pair (x_i, x_j), and it is nonzero only when x_i has higher-resolution than x_j. $\bar{f}_{\mathrm{RA}}$ represents a fixed network, and the gradients are not back-propagated through it such that the lower-resolution image x_j is encouraged to have similar output to higher-resolution x_i but not vice versa. In addition, since higher-resolution results usually provide higher-quality guidance during training, we give a larger weight to larger resolution difference by the term $(j-i)$ in $w_{i,j}$. Note that we use all the resolutions that are higher than x_j as supervision in Eq. 10 instead of only using the highest resolution x_1, as the results of x_j and x_1 can differ from each other significantly for a large j, and the results of the resolutions between x_j and x_1 can act as soft targets during training. In [54], Hinton *et al.* show the effectiveness of the "dark knowledge" in soft targets, and similarly for low-resolution 3D human pose and shape estimation, we also find that it is important to provide the challenging input a hierarchical supervision signal such that the learning targets are not too difficult for the network to follow.

3.4 Contrastive Learning

While the self-supervision loss enforces the consistency of the network outputs across different image resolutions, we can further improve the model training by encouraging the consistency of the final feature representation φ encoded by the network, such that features of lower-resolution images are closer to those of higher-resolution images. Similar to Eq. 10, we have the feature consistency loss:

$$L_{\mathrm{f}} = \sum_{i,j} w_{i,j} g(\bar{\varphi}_i, \varphi_j), \tag{11}$$

where φ_i is the feature vector of the i-th resolution input image x_i, and $\bar{\varphi}$ denotes a fixed feature vector without gradient back-propagation. $w_{i,j}$ is identical to that in Eq. 10. The function g is used to measure the distance between two feature vectors, and a straightforward choice is the MSE as in Eq. 10. However, the extracted features φ usually have very high dimensions (*e.g.* 2048), and the MSE loss is not effective in modeling correlations of the complex structures in high-dimensional representations, due to the fact that it can be decomposed element-wisely, *i.e.* assuming independence between elements in the feature vectors [21], [55]. Moreover, unimodal losses such as MSE can be easily affected by noise or insignificant structures in the features, and it would be desirable to have a loss function exploiting more global structures [21].

Towards this end, we propose a contrastive feature loss similar to [21], [22], [23], [55] to maximize the mutual information across the feature representations of different resolutions. The main idea behind our contrastive loss is to encourage the feature representation to be close for the same sample with different resolutions and far for different image samples. Mathematically, the contrastive function can be written as:

$$g(\bar{\varphi}_i, \varphi_j) = -\log \frac{\exp(s(\bar{\varphi}_i, \varphi_j)/\tau)}{\exp(s(\bar{\varphi}_i, \varphi_j)/\tau) + \sum_{q \in \mathcal{Q}} \exp(s(q, \varphi_j)/\tau)}, \tag{12}$$

where s represents the cosine similarity function, and τ is a temperature hyper-parameter [23]. φ_i, φ_j are the features of the same input with different resolutions. $\mathcal{Q}$ is a queue of data samples, which is constructed and progressively updated during training, and $\varphi_i, \varphi_j \notin \mathcal{Q}$. We use a method similar to [22] to update the queue, *i.e.* after each iteration, the current mini-batch is enqueued, and the oldest mini-batch in the queue is removed. Supposing the size of the queue is $|\mathcal{Q}|$, the contrastive loss is essentially a $(|\mathcal{Q}| + 1)$-way softmax-based classifier which classifies different resolutions (φ_i, φ_j) as a positive pair while different contents (q, φ_j) as a negative pair. As the feature extractor of the higher resolution image does not have gradients in Eq. 12, the proposed loss function enforces the network to generate higher-quality features for the low-resolution input image.

Our final loss is a combination of the basic loss, self-supervision loss, and contrastive feature loss: $L_{\mathrm{b}} + \lambda_{\mathrm{s}} L_{\mathrm{s}} + \lambda_{\mathrm{f}} L_{\mathrm{f}}$, where λ_{s} and λ_{f} are hyper-parameters.

3.5 Extension for Handling Video Input

While this work mainly focuses on 3D human pose and shape estimation from a single low-resolution image, it
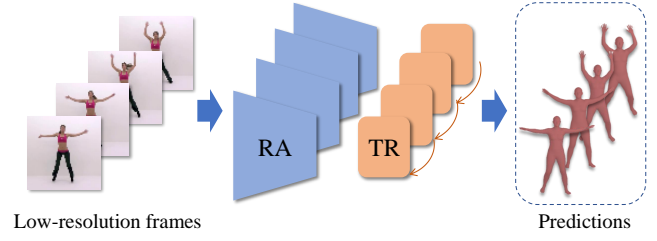

Fig. 3. Extension for handling video input. We first use the proposed RSC-Net to extract features for each frame and then apply a Temporal Recurrent module (TR) [6] to improve the per-frame results.

is a common practice to exploit the temporal information to further improve the predictions when videos are available [5], [6]. In this section, we show that we can easily extend the RSC-Net to handle low-resolution video input by incorporating a temporal post-processing step.

Specifically, we employ a strategy similar to the VIBE algorithm [6], which first performs per-frame estimation of the 3D human pose and shape with a pre-trained single-image model and then uses a temporal recurrent module composed of a Gated Recurrent Unit (GRU) [56] and an MLP to aggregate multi-frame information. We adapt this method to deal with low-resolution videos by adding the temporal recurrent module to the proposed resolution-aware network, as shown in Figure 3. We first use the pre-trained resolution-aware network to extract feature vectors φ from each frame of the input video, and then apply the recurrent module to the features to encode the temporal relations across frames, which facilitates generating more accurate and temporally-coherent SMPL outputs. Similar to [6], we also use a GAN loss to train the temporal recurrent module, where a motion discriminator takes a sequence of SMPL predictions as input and learns to discriminate between generated motion and ground-truth motion. Such an adversarial learning scheme [36] enforces more constraints on the ill-posed low-resolution estimation problem and enhances both the consistency and feasibility of the generated motion sequences.

3.6 Texture Reconstruction for Low-Resolution Human

To further demonstrate the potential of our approach, we propose a human texture estimation network which, together with the RSC-Net, can be used to reconstruct textured 3D human body from a single low-resolution pedestrian image (an overview is shown in Figure 4(a)).

Network structure. Similar to the state-of-the-art human texture estimation method [57] which is originally designed for high-resolution input, we use an encoder-decoder framework with concatenative skip-connections to predict the desired UV map from low-resolution images as shown in Figure 4(b), where we add an extra upsampling layer to the baseline model of [57] to encourage more details in the reconstructed UV map.

While the basic encoder-decoder model of [57] works reasonably well in practice, it has the potential problem that the fully convolutional structure only exploits the information from a local receptive field. Note that the contents of the input image of the texture network and the output UV

(a) Overview of the reconstruction process

(b) Architecture of the texture estimation network
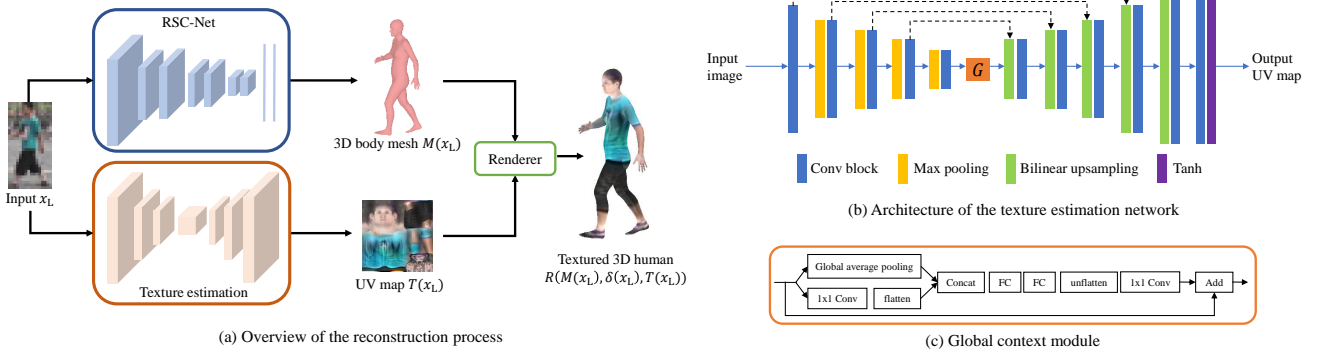
(c) Global context module

Fig. 4. Textured 3D human reconstruction from a single low-resolution image. We show an overview of the reconstruction process in (a). Note that the RSC-Net in (a) also outputs the camera parameter $\delta(x_\mathrm{L})$ which is omitted in the figure for clarity. The detailed architecture of the proposed texture estimation network is illustrated in (b). The main difference between the baseline texture estimation network [57] and our model is the global context module G detailed in (c). The "flatten" in (c) represents the feature flattening layer which reshapes the feature map into a one-dimensional vector. "unflatten" is the inverse-operation reshaping the global feature vector into feature maps. "FC" is the fully-connected layer.
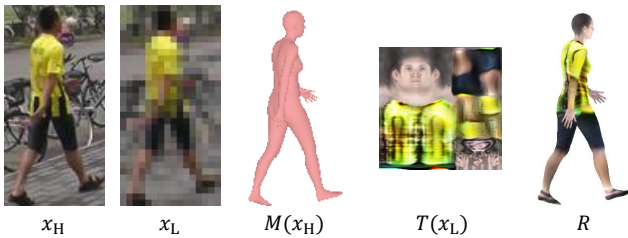

x_H x_L $M(x_\mathrm{H})$ $T(x_\mathrm{L})$ R

Fig. 5. Intermediate results for training the texture estimation network. x_H is a high-resolution image from the training dataset, and x_L is the corresponding low-resolution version synthesized with bicubic interpolation. M and T represent the predicted 3D body mesh and UV map. R is the textured output rendered with Pytorch3D [58].

map are significantly misaligned, *e.g.* the legs in the bottom of the input are located at the top of the UV map as shown in Figure 4(a). A successful solution to this problem should be able to grasp the global context information such that it can better spatially re-arrange the output pixels. To this end, we introduce a global context module in Figure 4(c), where we combine two types of mechanisms for global information aggregation, *i.e.* global averaging pooling [44], and feature flattening [59] which reshapes the input feature maps into a one-dimensional vector. The feature vectors from the pooling and flattening layers are first concatenated and then processed by two fully-connected layers such that the information from different spatial locations can be aggregated and better exploited for UV map prediction. As the input feature maps of the global context module have quite high dimensionality, we use a 1×1 convolution layer to reduce the feature maps before sending them to the flattening layer. The feature dimensionality is recovered with another 1×1 convolution layer after the unflattening layer in Figure 4(c) which reshapes the processed vectors back to feature maps. We finally fuse the unflattened global image features back to the input features with point-wise summation. Note that we also tried to fuse the features with a concatenation layer which, however, does not work well in our experiments.

Training process. With a high-resolution human image x_H (Figure 5) from the training dataset, we can synthesize a low-resolution image x_L with bicubic interpolation.

We first use the high-resolution image x_H as input of the proposed RSC-Net to obtain a high-quality 3D body mesh $M(x_\mathrm{H})$ as well as its corresponding camera parameters $\delta(x_\mathrm{H})$. Then we predict the UV map $T(x_\mathrm{L})$ for the low-resolution image with the texture estimation network, and the textured human image can be reconstructed by $R(M(x_\mathrm{H}), \delta(x_\mathrm{H}), T(x_\mathrm{L}))$ using a differentiable renderer R. Different from [57] where the OpenDR [60] is used as the renderer, we employ the more modern tool Pytorch3D [58] for differentiable rendering. Finally, the texture estimation network can be trained with the Re-identification (ReID) loss of [57]:

$$\|f_\mathrm{Re}(R(M(x_\mathrm{H}), \delta(x_\mathrm{H}), T(x_\mathrm{L}))) - f_\mathrm{Re}(x_\mathrm{H})\|_2, \quad (13)$$

where f_Re represents a pre-trained ReID network that is explicitly trained to minimize the distance between images of the same identity. As the human identity is mostly characterized by the body textures, minimizing the ReID loss can effectively encourage the rendered output to have similar appearance as the high-resolution image x_H, while alleviating the effect of the misalignment between the ground truth and output. During inference for a low-resolution image, the textured human can be reconstructed by $R(M(x_\mathrm{L}), \delta(x_\mathrm{L}), T(x_\mathrm{L}))$ as show in Figure 4(a).

4 EXPERIMENTS

We first describe the implementation details of the proposed algorithm. Then we compare the RSC-Net with the state-of-the-art 3D human pose and shape estimation approaches on different benchmark datasets. In addition, we also provide evaluations on low-resolution video input. We finally present quantitative and qualitative results to evaluate the proposed texture estimation network.

4.1 Implementation Details

Single-frame 3D human pose and shape. We train our model and the baselines using a combination of 2D and 3D datasets similar to previous works [4], [7]. For the 3D datasets, we use Human3.6M [49] and MPI-INF-3DHP [50] with ground truth of 3D keypoints, 2D keypoints, and SMPL
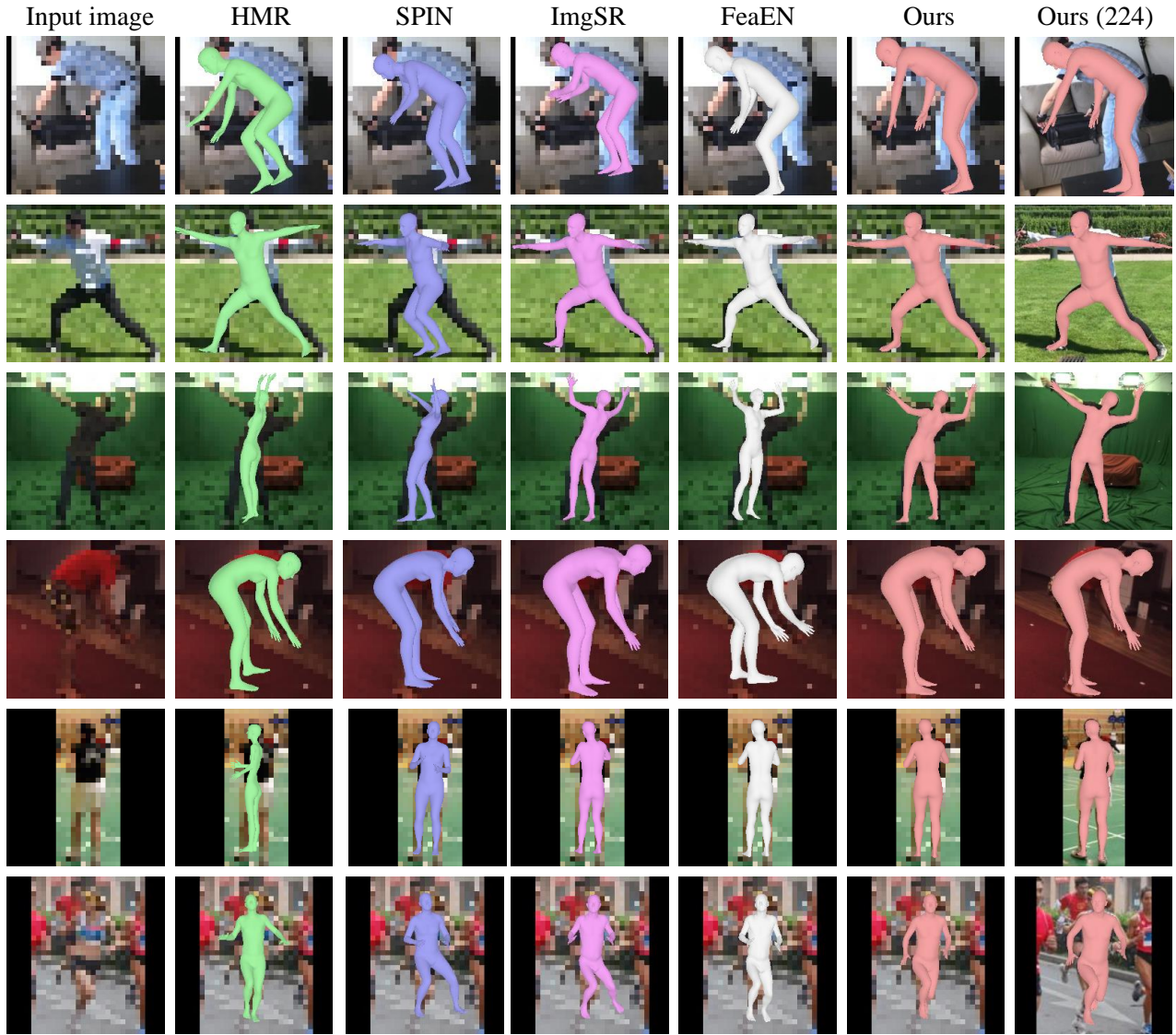
Input image	HMR	SPIN	ImgSR	FeaEN	Ours	Ours (224)

Fig. 6. Visual comparisons with the state-of-the-art methods on challenging low-resolution input. The input image has a resolution of 32×32. The results of high-resolution images are also included as references.

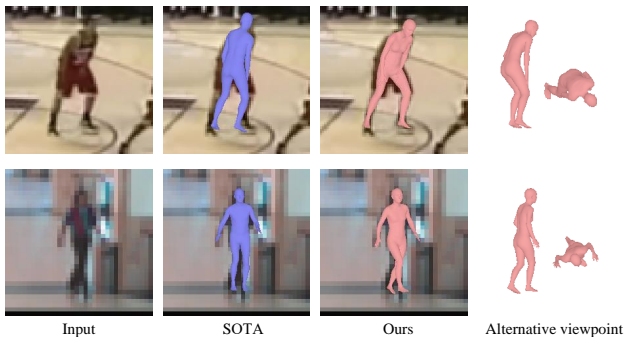| Input	SOTA	Ours	Alternative viewpoint

Fig. 7. Visual examples of real low-resolution images obtained from the Internet. SOTA method [7] that works well for high-resolution images performs poorly at low-resolution ones.

parameters. These datasets are mostly captured in constrained environments, and models trained on them do not generalize well to diverse images in real world. For better performance on in-the-wild data, we also use 2D datasets including LSP [61], LSP-Extended [62], MPII [51], and MS COCO [52], which only have 2D keypoint labels. Similar to prior works [4], [7], we assume the human bounding box is known, and the target human is located at the center of the bounding box. We crop the human regions from the images and resize them to 224×224. Images with significant occlusions or small human are discarded from the dataset. We consider human image resolutions ranging from 224 to 24. As introduced in Section 3.2, we split all the resolutions into $P = 5$ ranges: $\{224, (224, 128], (128, 64], (64, 40], (40, 24]\}$, where the first range corresponds to the original high-resolution image x_1. We obtain the lower-resolution images by downsampling the high-resolution images and resize them back to 224 with bicubic interpolation. During training, we apply data augmentations to the images including Gaussian noise, color jitters, rotation, and random flipping. For the loss functions, we set $\lambda_1 = 5$, $\lambda_2 = 5$, $\lambda_s = 0.1$, and $\lambda_f = 0.1$. For contrastive learning, we set the size of the queue as 8192 and $\tau = 0.1$ in Eq. 12. Following [23],

TABLE 1
Results on the 3DPW [48], MPI-INF-3DHP [50], and H36M [49] datasets. The proposed method compares favorably against the baseline approaches for challenging low-resolution (32×32) images.

Methods	3DPW		MPI-INF-3DHP		H36M	
	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA
HMR	142.29	79.73	133.25	85.30	93.08	63.96
SPIN	137.61	78.44	127.27	83.38	86.17	57.35
ImgSR	146.58	81.07	125.91	83.52	87.23	57.06
FeaEN	143.51	77.21	124.99	81.80	93.88	62.70
Ours	**117.12**	**67.59**	**115.80**	**78.68**	**79.59**	**53.56**

(a) VIBE

(b) Ours-single
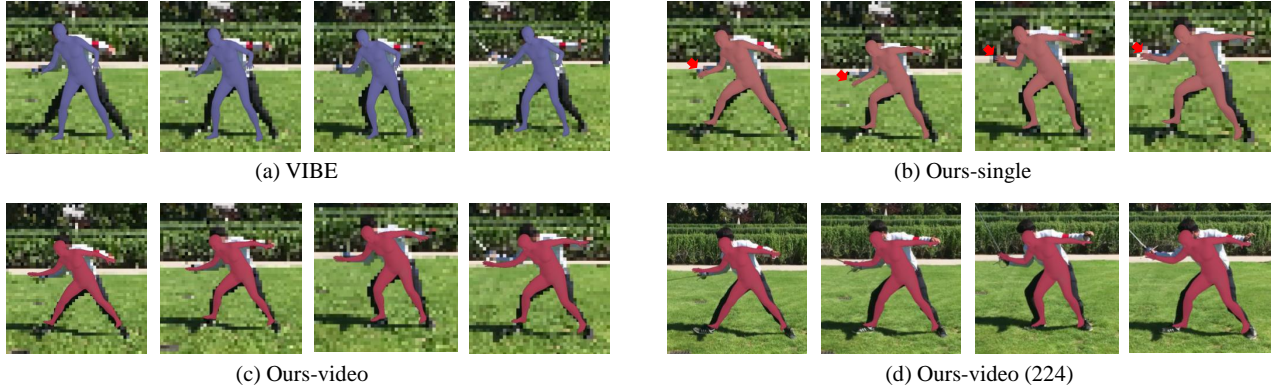
(c) Ours-video

(d) Ours-video (224)

Fig. 8. Qualitative evaluation of the proposed method on a low-resolution video sequence from 3DPW [48]. Our video model (c) generates more accurate and temporally smoother predictions. The result of the corresponding high-resolution video (d) is also included as a reference. Note that the results of the single image model (b) is not as temporally-coherent (*e.g.* the hand and body size).

the similarity function s is equipped with a 2-layer MLP to better transform and measure the features. As in [6], we initialize the baseline networks and our model with the parameters of [7]. We use the Adam algorithm [63] to optimize the network with a learning rate 5e-5.

Similar to [6], we conduct evaluations on a large in-the-wild dataset 3DPW [48] with 3D joint ground truth to demonstrate the strength of our model in an in-the-wild setting. We also provide results for constrained indoor images using the MPI-INF-3DHP [50] and H36M [49] datasets. Following [4], [6], [7], we compute the procrustes aligned mean per joint position error (MPJPE-PA) and mean per joint position error (MPJPE) for measuring the 3D keypoint accuracy. We mainly perform the evaluations on a challenging low resolution of 32×32. To analyze the robustness of our method on more diverse resolutions, we also compare the results of the resolutions 176, 96, 52, and 32 in Section 5.1, which are the middle points of the divided training resolution ranges.

Extension to low-resolution videos. We use a 2-layer GRU and a 3-layer MLP as the temporal recurrent module. We follow a similar paradigm as in VIBE [6] to extract features before training the temporal recurrent module. The offline feature extraction allows larger batch size and faster training. For the video datasets, we use the training set of InstaVariety [5], MPI-INF- 3DHP [50], 3DPW [48], and Human3.6M [49]. In addition, we use the large-scale 3D human motion dataset AMASS [64] as the ground truth motion for adversarial training. We refer to VIBE [6] for more implementation details.

Texture reconstruction. Each Conv block of our texture estimation network consists of two convolutional layers

followed by Batch Normalization [65] and ReLU activation [46]. For the ReID loss, we use the model in [66] as the pedestrian re-identification network for its simplicity and efficiency. Similar to [57], we use 1401 person identities of the Market-1501 dataset [67] for training, and the remaining 100 identities are used for testing. We also use the SURREAL dataset [68] to train the network such that the predicted texture can have a plausible-looking face [57]. To synthesize the low-resolution data, we downsample the images to a height of 32 pixels which are further resized to 128 before fed into the network. Due to the limitation of the SMPL model, the reconstructed human and the input image are not always well-aligned, and thus the pixel-wise evaluation metrics, such as PSNR and SSIM, cannot effectively measure the quality of the output image. To solve this problem, we use the rescaled ReID loss (Eq. 13) for evaluation which assesses the results from a semantic (identification) perspective and thus can better focus on the texture similarity of human bodies without affected by misalignment issues. In addition, we also use the LPIPS [69] to evaluate the perceptual quality of the reconstructed textures.

4.2 Evaluation on Low-Resolution Single Image

We compare against the state-of-the-art 3D human pose and shape estimation methods HMR [4] and SPIN [7], by fine-tuning them on different resolution images with the same training data and settings as our model. In addition, since no previous approach has focused on the problem of low-resolution 3D human pose and shape estimation, we adapt the low-resolution image recognition algorithms to our task as new baselines, including both image super-resolution based [41] and feature enhancement based [39]. For the image super-resolution based method (denoted as ImgSR),

(a) VIBE

(b) Ours-single
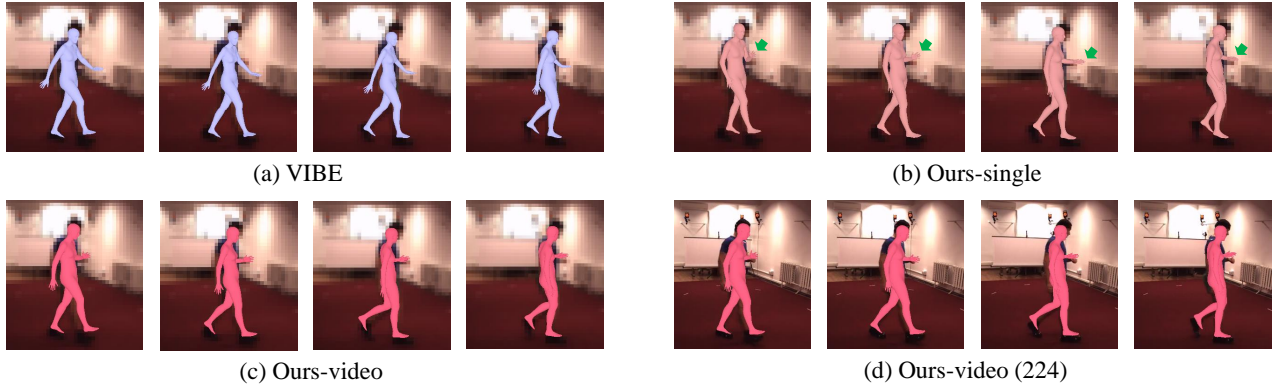
(c) Ours-video

(d) Ours-video (224)

Fig. 9. Qualitative evaluation of the proposed method on a low-resolution video sequence from H36M [49]. Our video model (c) generates more accurate and temporally smoother predictions. The result of the corresponding high-resolution video (d) is also included as a reference. Note that the results of the single image model (b) is not as temporally-coherent (*e.g.* the movement of the hand).

we first use a state-of-the-art network RDN [70] to super-resolve the low-resolution image, and the output is then fed into SPIN [7] for regressing the SMPL parameters. Similar to [41], the network is trained to improve both the perceptual image quality and the 3D human pose and shape estimation accuracy. For feature enhancement (denoted as FeaEN), we apply the strategy in [39] which uses a GAN loss to enhance the discriminative ability of the low-resolution features for better image retrieval performance. Nevertheless, we find the WGAN [71] used in the original work [39] does not work well in our experiments, and we instead use the LSGAN [72] combined with the basic loss (Eq. 8) to train a stronger baseline network.

As shown in Table 1, the proposed method generates consistently better results than the baseline approaches on the 3DPW, MPI-INF-3DHP, and H36M datasets for challenging low-resolution (32×32) images. Note that we achieve significant improvement over the baselines on the 3DPW dataset, which demonstrates the effectiveness of the proposed method on the challenging in-the-wild images. Similar to [4], [7], we also evaluate our approach on the auxiliary task of 2D human body segmentation on the test images of LSP. The proposed algorithm compares favorably against the baseline methods as shown in Table 2. Furthermore, we provide a qualitative comparison against the baseline models in Figure 6, where the proposed method generates higher-quality 3D human estimation results on the challenging low-resolution input. We also provide visual examples of real low-resolution images obtained from the Internet in Figure 1 and 7, which further shows that the proposed algorithm can well generalize to images in real world.

4.3 Evaluation on Low-Resolution Videos

In Section 3.5, we introduce a post-processing module to exploit the temporal information of videos. To evaluate the performance of the proposed multi-frame network on low-resolution videos, we use the sequences of the 3DPW [48] and the H36M [49] datasets that are downsampled to 32×32 pixels. We follow VIBE [6] to pre-process the test data, where the severely-occluded frames are also included to ensure the continuity of the input frames. For fair comparisons, we re-train the baseline network VIBE [6] on low-resolution videos with the same training data and training settings as our models. As shown in Table 3, the proposed video

TABLE 2
Results on the LSP dataset [61]. "H-acc" and "H-f1" denote the accuracy and f1 score of the human silhouettes, while "P-acc" and "P-f1" denote those of body part segmentation.

Methods	H-acc	H-f1	P-acc	P-f1
HMR	89.33	0.8273	84.73	0.5145
SPIN	90.00	0.8384	87.09	0.5998
ImgSR	89.75	0.8308	87.07	0.5955
FeaEN	89.57	0.8287	86.71	0.5843
Ours	**90.18**	**0.8412**	**87.43**	**0.6159**

TABLE 3
Quantitative evaluation on low-resolution videos. "ACC" and "ACC-ERR" represent the acceleration and the acceleration error, respectively.

	Methods	MPJPE	MPJPE-PA	ACC	ACC-ERR
3DPW	VIBE [6]	135.05	77.94	52.99	54.07
	Ours-single	141.87	75.04	90.61	91.12
	Ours-video	**123.14**	**72.64**	**48.63**	**49.58**
H36M	VIBE [6]	122.16	82.19	72.67	68.48
	Ours-single	135.09	75.72	170.35	162.33
	Ours-video	**84.10**	**60.04**	**63.23**	**60.80**

model attains more accurate predictions than VIBE, which demonstrates that the RSC-Net can be well extended to handle low-resolution video input.

To measure the temporal consistency of the results, we also report the acceleration which represents the second-order derivative of the 3D joint motion, and the acceleration error that is calculated by the difference between ground-truth and predicted 3D acceleration of every joint. As shown in Table 3, the proposed multi-frame model ("Ours-video") can achieve better temporal coherence than the single-frame network ("Ours-single"), which shows the effectiveness of the temporal recurrent module. In addition, we provide qualitative comparisons in Figure 8 and 9, where our video model generates more accurate and temporally smoother results.

4.4 Evaluation of the Texture Estimation Network

In Table 4, we compare our method with the baseline model [57] for low-resolution human texture reconstruction. Our texture estimation network with the global module is

TABLE 4
Quantitative evaluation of the proposed texture estimation network against the baseline model [57] on the test set of Market-1501 [67]. The original "Baseline" uses the HMR [4] for 3D body mesh prediction and the OpenDR [60] for differentiable rendering. We re-train the baseline with the 3D body mesh predicted by our RSC-Net, which is denoted as "Baseline†". We further replace the OpenDR [60] with the Pytorch3D [58], leading to an even stronger baseline, *i.e.* "Baseline*".

Metrics	Baseline	Baseline†	Baseline*	TexGlo
ReID loss	60.61	60.41	60.21	**59.89**
LPIPS	0.2170	0.2156	0.2120	**0.2094**

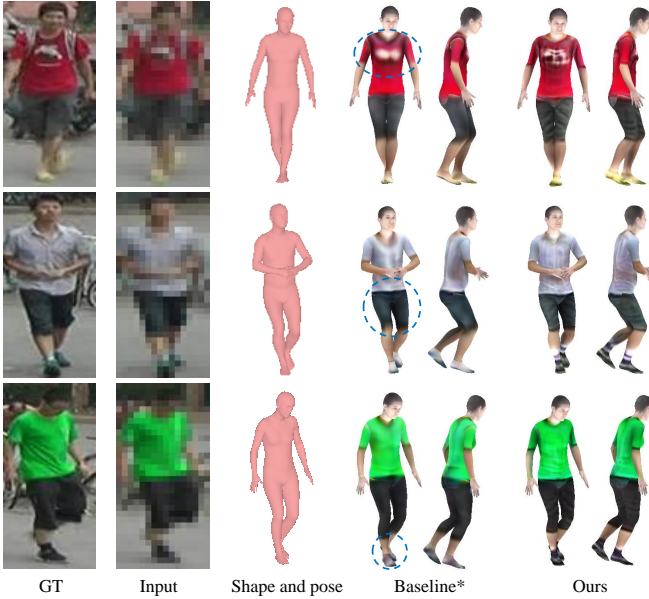GT   Input   Shape and pose   Baseline*   Ours

Fig. 10. Qualitative evaluation of the texture estimation network. HR represents the corresponding high-resolution image.

named as TexGlo. The original "Baseline" uses the HMR [4] for 3D body mesh prediction and the OpenDR [60] for differentiable rendering, whereas we use the proposed RSC-Net and the Pytorch3D renderer [58]. For better comparison, we re-train the baseline with the 3D body mesh predicted by our network ("Baseline†"). We further replace the OpenDR with the Pytorch3D, which leads to an even stronger baseline, *i.e.* "Baseline*". As shown in Table 4, our method can achieve better results (lower ReID errors and LPIPS) than all the baselines, which demonstrates the effectiveness of the proposed texture estimation network.

For more comprehensive analysis, we further conduct a user study to evaluate the texture results. This study uses 10 images randomly selected from the test set of Market-1501 [67], and 21 subjects are asked to compare the textures reconstructed by "Baseline*" and our algorithm. 92.4% of the answers prefer the textures generated by the proposed network, which shows the effectiveness of our method. In addition, we also present visual examples in Figure 10, where the proposed method recovers more vivid colors and finer details.

5 ANALYSIS AND DISCUSSION

In this section, we first analyze the robustness of the proposed RSC-Net for different resolution images. We then present a comprehensive ablation study to demonstrate

TABLE 5
Comparison of the MPJPE on high-resolution images. "SPIN-o" represents the original SPIN [7] model specifically trained for high-resolution images, whereas "SPIN" is trained on all the different resolution images. Numbers in bold indicate the best in each column, and underscored numbers indicate the second.

Methods	3DPW		MPI-INF-3DHP	
	224	32	224	32
SPIN-o	<u>97.18</u>	149.31	**103.39**	147.67
SPIN	113.24	<u>137.61</u>	108.80	<u>127.27</u>
Ours	**96.60**	**117.12**	<u>103.50</u>	**115.80**

the effectiveness of the resolution-aware network, the self-supervision loss, the contrastive learning loss. Finally, we provide more analysis of our texture estimation network.

5.1 Robustness to Different Resolutions

As introduced in Section 3.3 and 3.4, the RSC-Net is trained by enforcing the scale consistency of the predicted results and features. As a result, the proposed algorithm is able to achieve robust performance on different resolution images as shown in Figure 11, where our method achieves consistently better results than the state-of-the-art approaches.

In addition, we also conduct evaluations on high-resolution images to demonstrate the robustness of the proposed method. As shown in Table 5, our RSC-Net is able to compete with the state-of-the-art high-resolution model [7] ("SPIN-o" in Table 5) on high-resolution input. Further, the proposed method significantly outperforms "SPIN-o" on low-resolution images, as "SPIN-o" has only been trained for high-resolution images whereas our network is trained for all the different resolution images. On the other hand, while fine-tuning "SPIN-o" on more diverse resolutions (*i.e.* "SPIN" in Table 5) can achieve better results on low-resolution input, the performance of SPIN suffers from a significant drop on high-resolution images. In other words, SPIN needs to sacrifice its performance on high-resolution images to achieve good results on other resolutions, which further demonstrates the robustness of the proposed algorithm. In addition, we present visual examples in Figure 12 where our method generates high-quality 3D body pose and shape for different resolutions.

5.2 Ablation Study

In Table 6, we provide an ablation study using the 3DPW dataset to evaluate the proposed resolution-aware network, the self-supervision loss, and the contrastive feature loss. We first compare the proposed resolution-aware network with the baseline model ResNet-50 [4], [44]. As shown by "RA" (resolution-aware) and "Ba" (baseline) in Table 6, our network can obtain slightly better results than the baseline network with the basic loss (Eq. 8) as loss function. Further, we can achieve a more significant improvement over the baseline when adding the self-supervision loss (Eq. 10) and the contrastive feature loss (Eq. 11) for training, *i.e.* "RA+SS" *vs.* "Ba+SS" and "RA+SS+CL" *vs.* "Ba+SS+CL" in Table 6, which further demonstrates the effectiveness of the resolution-aware structure.

Second, we use the self-supervision loss in Eq. 10 to exploit the consistency of the outputs of the same input
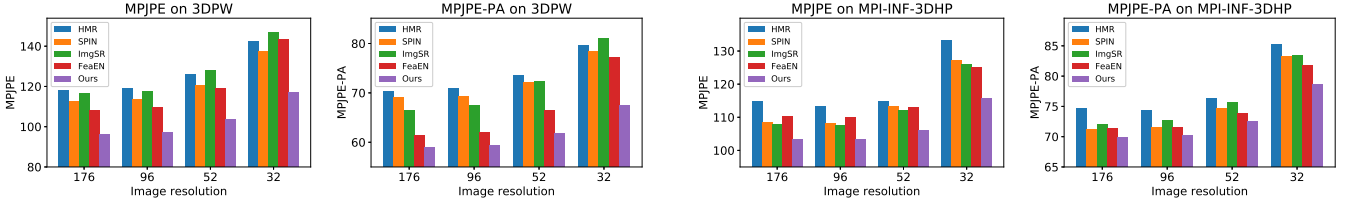
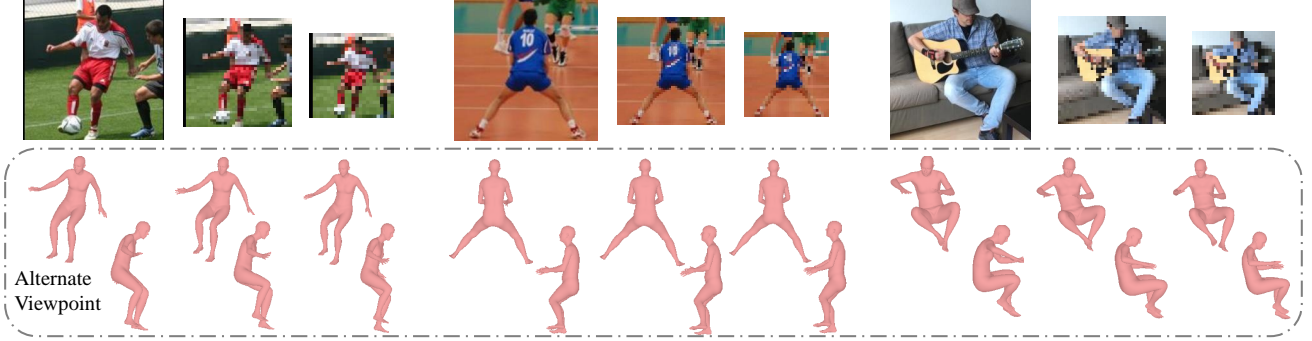Fig. 11. Robustness of the RSC-Net on different low-resolution images.

Fig. 12. The proposed method can achieve robust 3D human pose and shape estimation for different resolution inputs. The resolutions for the images are 224, 52, and 32.

TABLE 6
Ablation study of the proposed method. Ba: baseline network with basic loss function, RA: resolution-aware network with basic loss function, SS: self-supervision loss, MS: MSE feature loss, CD: cosine distance feature loss, CL: contrastive learning feature loss.

Methods	MPJPE				MPJPE-PA			
	176	96	52	32	176	96	52	32
Ba	112.26	115.18	124.88	143.63	65.04	66.41	71.12	79.43
Ba+SS	107.51	109.58	116.54	128.88	62.32	63.27	66.78	72.49
Ba+SS+CL	106.06	107.75	112.44	123.63	63.25	64.07	66.04	70.86
RA	111.55	112.18	118.70	135.29	64.53	68.88	68.01	75.49
RA+SS	102.56	104.18	110.17	124.23	60.17	60.84	63.71	69.87
RA+SS+MS	105.96	106.15	111.33	124.85	60.90	61.76	64.55	70.40
RA+SS+CD	104.95	105.96	111.41	125.08	61.29	61.91	64.30	70.17
RA+SS+CL	**96.36**	**97.36**	**103.49**	**117.12**	**58.98**	**59.34**	**61.81**	**67.59**

TABLE 7
Analysis of the alternative training strategies. PT: Progressive Training, SS-o: original self-supervision loss, SS-h: only using the highest-resolution for supervision.

Methods	MPJPE				MPJPE-PA			
	176	96	52	32	176	96	52	32
w/o PT	105.11	106.60	113.41	127.05	61.46	62.22	65.47	71.30
w/ SS-o	143.31	142.32	145.61	156.25	77.75	77.51	79.06	82.97
w/ SS-h	104.16	105.24	109.94	122.01	62.46	62.73	64.47	68.89
full model	**96.36**	**97.36**	**103.49**	**117.12**	**58.98**	**59.34**	**61.81**	**67.59**

image with different resolutions. By comparing "RA+SS" against "RA" in Table 6, we show that the self-supervision loss is important for addressing the weak supervision issue of 3D human pose and shape estimation and thus effectively improves the results. The comparison between "Ba+SS" and "Ba" also leads to similar conclusions.

In addition, we propose to enforce the consistency of the features across different image resolutions. However, a normally-used MSE loss does not work well as shown in "RA+SS+MS" of Table 6, which is mainly due to that the uni-modal losses are not effective in modeling the correlations between high-dimensional vectors and can be easily affected by noise and insignificant structures in the embedded fea-

tures [21]. In contrast, the proposed contrastive feature loss can more effectively improve the feature representations by maximizing the mutual information across the features of different resolutions, and achieve better results as in "RA+SS+CL" of Table 6. Note that we adopt the cosine similarity function in the contrastive feature loss (Eq. 12) similar to prior methods [21], [22], [55]. Alternatively, one may only use the cosine distance for measuring the distance of two features instead of using the whole contrastive loss (Eq. 12). Nevertheless, this strategy does not work well as shown by "RA+SS+CD" in Table 6, which demonstrates the effectiveness of the proposed contrastive learning scheme.

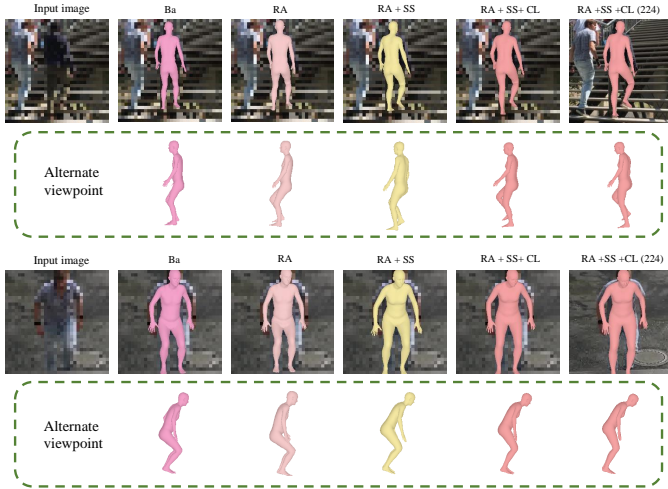Furthermore, we present several visual examples in Fig-

Fig. 13. Visual examples which show the effectiveness of the three components of the RSC-Net, *i.e.* the resolution-aware network, the self-supervision loss, and the contrastive learning feature loss.

TABLE 8
Effectiveness of the proposed texture estimation network. "Up" represents the extra upsampling layer. "Glo" is the global context module. "Glo-C" and "Glo-S" represent fusing the global features with concatenation and point-wise summation, respectively.

Methods	w/o Up & Glo	w/o Glo	w/ Glo-C	w/ Glo-S
ReID loss	60.21	60.14	60.25	**59.89**

ure 13, which intuitively illustrate that all the three components of the RSC-Net (RA, SS, and CL) essentially improve the estimated 3D human pose and shape.

Analysis of training strategies. We also provide a detailed analysis of the alternative training strategies of our model. First, as described in Section 3.2, we train our model as well as the baselines in a progressive manner to deal with the challenging multi-resolution input. As shown in the first row of Table 7 ("w/o PT"), directly training the model for all image resolutions without the progressive strategy leads to degraded results.

Second, the original self-supervision loss (Eq. 9) treats the images under different augmentations equally, while we are generally more confident in the high-resolution predictions. Therefore, we propose a directional self-supervision loss in Eq. 10 to exploit this prior knowledge. As shown in the second row of Table 7 ("w/ SS-o"), using the original self-supervision loss (Eq. 9) is not able to achieve high-quality results, as the network can minimize Eq. 9 by simply degrading the high-resolution predictions without improving the low-resolution ones (note the significant performance gap between "w/ SS-o" and "full model"). In addition, we provide hierarchical supervision for low-resolution images in Eq. 10 which can act as soft targets during training. As shown in Table 7, only using the highest-resolution predictions as guidance ("w/ SS-h") cannot produce as good results as the proposed approach ("full model").

5.3 More Analysis of the Texture Estimation Network

As introduced in Section 3.6, the main difference between the baseline [57] and the proposed texture estimation net-
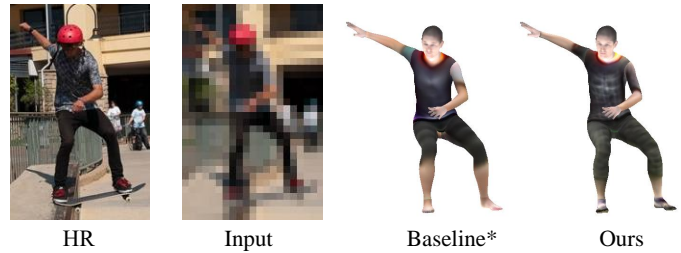

Fig. 14. Limitation of the texture estimation network. The proposed method is not able to recover the textures well as the image has a rare color around the human head.

work is the global context module which can capture the information of the human body globally and thereby generate higher-quality UV maps. As shown in Table 8, the network without the global context module ("w/o Glo") does not perform as well as our method ("w/ Glo-S"). In addition, we use an extra upsampling layer before the output layer of the encoder-decoder architecture (Figure 4(b)), which allows the network to generate more details in the predicted UV map. The effectiveness of this design has also been shown in Table 8 ("w/o Up & Glo"). Note that we use the point-wise summation to fuse the extracted global context information back to the input feature maps in Figure 4(c), as we empirically find that the normally-used concatenation layer ("w/ Glo-C") does lead to satisfying results.

Limitation. Although the proposed algorithm is able to predict high-quality texture maps for low-resolution humans, the training process of this model is based on the ReID loss in Eq. 13, which essentially relies on a pre-trained pedestrian re-identification network. Thus, it has the same limitations as the state-of-the-art high-resolution texture estimation method [57] and is likely to fail when the input is not a common pedestrian scene. As shown in Figure 14, the proposed method is not able to recover the textures well as the image has a rare color around the human head. Our future work will focus on developing a texture estimation network that can better generalize to more diverse images.
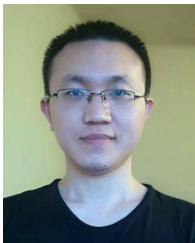
6 CONCLUSION

In this work, we have studied the challenging problem of predicting 3D human pose and shape from low-resolution images and presented an effective solution for it. We propose a resolution-aware neural network which can deal with different resolution images using a single model. For training the proposed network, we propose a directional self-supervision loss which can exploit the output consistency across different resolutions to remedy the issue of lacking accurate 3D labels. In addition, we introduce a contrastive feature loss which is more effective than MSE for measuring high-dimensional feature vectors and helps learn better feature representations. Our method performs favorably against the state-of-the-art methods on different resolution images and achieves high-quality results for low-resolution 3D human pose and shape estimation. Further, while we focus on learning 3D human pose and shape in this work, the proposed methodology is generic, and we expect its application in other low-resolution recognition problems.

REFERENCES

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single rgb camera," in *CVPR*, 2019. 1, 2

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *CVPR*, 2018. 1, 2

[3] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*, 2016. 1, 2, 3

[4] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018. 1, 2, 3, 4, 5, 7, 8, 9, 10, 11

[5] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *CVPR*, 2019. 1, 2, 3, 5, 6, 9

[6] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *CVPR*, 2020. 1, 2, 3, 6, 9, 10

[7] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *ICCV*, 2019. 1, 2, 3, 4, 5, 7, 8, 9, 10, 11

[8] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "Siclope: Silhouette-based clothed people," in *CVPR*, 2019. 1, 2

[9] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *CVPR*, 2018. 1, 2

[10] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, "3dpeople: Modeling the geometry of dressed humans," in *ICCV*, 2019. 1, 2, 3

[11] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *ICCV*, 2019. 1, 2, 3

[12] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3d human reconstruction from a single image," in *ICCV*, 2019. 1, 2

[13] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *ICCV*, 2017. 1, 3

[14] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *CVPR*, 2016. 1, 3

[15] K. Nishibori, T. Takahashi, D. Deguchi, I. Ide, and H. Murase, "Exemplar-based human body super-resolution for surveillance camera systems," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014. 1

[16] L. Neumann and A. Vedaldi, "Tiny people pose," in *ACCV*, 2018. 1, 3

[17] X. Xu, Y. Ma, and W. Sun, "Towards real scene super-resolution with raw images," in *CVPR*, 2019. 1

[18] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR*, 2011. 1

[19] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2017. 1, 5

[20] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017. 1

[21] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv:1807.03748*, 2018. 2, 3, 6, 12

[22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020. 2, 6, 12

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020. 2, 6, 8

[24] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. De la Torre, "3d human shape and pose from a single low-resolution image with self-supervised learning," in *ECCV*, 2020. 2

[25] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *ICCV*, 2019. 2

[26] C. Doersch and A. Zisserman, "Sim2real transfer learning for 3d human pose estimation: motion to the rescue," in *NeurIPS*, 2019. 2

[27] J. Y. Zhang, P. Felsen, A. Kanazawa, and J. Malik, "Predicting 3d human dynamics from video," in *ICCV*, 2019. 2, 3

[28] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints," in *CVPR*, 2018. 2

[29] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *CVPR*, 2017. 2

[30] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A joint model for 2d and 3d pose estimation from a single image," in *CVPR*, 2013. 2

[31] E. Simo-Serra, C. Torras, and F. Moreno-Noguer, "3d human pose tracking priors using geodesic mixture models," *IJCV*, vol. 122, no. 2, pp. 388–408, 2017. 2

[32] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *International Conference on 3D Vision (3DV)*, 2018. 2

[33] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, p. 248, 2015. 2, 3

[34] X. Xu, Y. Ma, and W. Sun, "Learning factorized weight matrix for joint filtering," in *ICML*, 2020. 2

[35] J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2021. 2

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014. 3, 6

[37] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *TIP*, vol. 28, no. 4, pp. 2051–2062, 2018. 3

[38] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *ACCV*, 2018. 3

[39] W. Tan, B. Yan, and B. Bare, "Feature super-resolution: Make machine see more clearly," in *CVPR*, 2018. 3, 9, 10

[40] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *ICCV*, 2019. 3

[41] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," *arXiv:1803.11316*, 2018. 3, 9, 10

[42] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *CVPR*, 2017. 3

[43] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 4, 7, 11

[45] ——, "Identity mappings in deep residual networks," in *ECCV*, 2016. 4

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010. 4, 9

[47] T. Bachlechner, B. P. Majumder, H. H. Mao, G. W. Cottrell, and J. McAuley, "Rezero is all you need: Fast convergence at large depth," *arXiv preprint arXiv:2003.04887*, 2020. 5

[48] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018. 5, 9, 10

[49] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *TPAMI*, vol. 36, no. 7, pp. 1325–1339, 2013. 5, 7, 9, 10

[50] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017. 5, 7, 9

[51] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014. 5, 8

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014. 5, 8

[53] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009. 5

[54] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015. 5

[55] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019. 6, 12

[56] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase rep-

resentations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014. 6

[57] J. Wang, Y. Zhong, Y. Li, C. Zhang, and Y. Wei, "Re-identification supervised texture generation," in *CVPR*, 2019. 6, 7, 9, 10, 11, 13

[58] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Pytorch3d," https://github.com/facebookresearch/pytorch3d, 2020. 7, 11

[59] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014. 7

[60] M. M. Loper and M. J. Black, "Opendr: An approximate differentiable renderer," in *ECCV*, 2014. 7, 11

[61] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation." in *BMVC*, 2010. 8, 10

[62] ——, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*, 2011. 8

[63] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2014. 9

[64] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *ICCV*, 2019. 9

[65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015. 9

[66] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018. 9

[67] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015. 9, 11

[68] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017. 9

[69] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. 9

[70] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *CVPR*, 2018. 10

[71] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017. 10

[72] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017. 10

Francesc Moreno-Noguer is a Scientific Researcher of the Spanish National Research Council at the Institut de Robotica i Informatica Industrial. His research interests are in Computer Vision and Machine Learning, with topics including human shape and motion estimation, 3D reconstruction of rigid and nonrigid objects and camera calibration. He received the Polytechnic University of Catalonia's Doctoral Dissertation Extraordinary Award, several best paper awards (e.g. ECCV 2018 Honorable mention, ICCV 2017 workshop in Fashion, Intl. Conf. on Machine Vision applications 2016), outstanding reviewer awards at ECCV 2012 and CVPR 2014, and Google and Amazon Faculty Research Awards in 2017 and 2019, respectively. He has (co)authored over 100 publications in refereed journals and conferences (including 9 IEEE Transactions on PAMI, 5 Intl. Journal of Computer Vision, 21 CVPR, 8 ECCV and 7 ICCV).

László A. Jeni is Systems Faculty in the Robotics Institute at Carnegie Mellon University, Pittsburgh, PA, USA. He specializes in computer vision and computational behavior science, specifically in areas of modeling, understanding, and synthesizing human behavior using diverse sensors. He received his MSc degree in Computer Science from Eötvös Loránd University, Hungary, and his Ph.D. in 2012 from the University of Tokyo, Japan. To his credit he has over 50 peer reviewed publications. His honors include Best Paper Awards at the IEEE Conference on Human System Interaction (HSI'2011) for work on validating observations of human activity and at the Conference on Automatic Face and Gesture Recognition (FG'2015) for work on dense 3D face alignment from 2D video. He has organized computational face challenges and workshops at ECCV, ICCV, and FG.

Fernando De la Torre received the MSc and PhD degrees in Electronic Engineering from the La Salle School of Engineering at Ramon Llull University, Barcelona, Spain, in 1996, and 2002. Since 2005, he is faculty in the Robotics Institute at Carnegie Mellon University, and currently is Associate Research Professor. He has published more than 200 peer-reviewed papers in top conferences and journals. His research interests include the fields of computer vision and machine learning. He is currently directing the Human Sensing Laboratory at Carnegie Mellon University. His work won the best student paper award in IEEE Computer Vision and Pattern Recognition 2012. In 2014 he founded FacioMetrics LLC, which was acquired by Facebook in 2016.

Xiangyu Xu is a postdoc fellow in the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He was a research scientist at SenseTime from 2018 to 2019. He got the Ph.D. degree in the Department of Electronic Engineering, Tsinghua University in 2018. He received the B.Eng degree in the Department of Electronic Engineering, Tsinghua University in 2013, and the B.Ec degree in the National School of Development, Peking University in 2015. He was a visiting Ph.D. student at University of California, Merced and Harvard University. His research interest includes image processing, computer vision, and machine learning.

Hao Chen is a master student in Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA, USA. He received the B.Eng degree of Electrical and Electronics at School of Engineering, University of Edinburgh and B.Eng degree of Automation at Department of Electrical Communication and Automation Department, Tianjin University. His research interest lies in computer vision and deep learning.