3D Human Shape and Pose from a Single Low-Resolution Image with Self-Supervised Learning

Xiangyu Xu[1(✉)], Hao Chen[2], Francesc Moreno-Noguer[3], László A. Jeni[1], and Fernando De la Torre[1,4]

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
[2] Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA
[3] Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
[4] Facebook Reality Labs (Oculus), Pittsburgh, USA

Abstract. 3D human shape and pose estimation from monocular images has been an active area of research in computer vision, having a substantial impact on the development of new applications, from activity recognition to creating virtual avatars. Existing deep learning methods for 3D human shape and pose estimation rely on relatively high-resolution input images; however, high-resolution visual content is not always available in several practical scenarios such as video surveillance and sports broadcasting. Low-resolution images in real scenarios can vary in a wide range of sizes, and a model trained in one resolution does not typically degrade gracefully across resolutions. Two common approaches to solve the problem of low-resolution input are applying super-resolution techniques to the input images which may result in visual 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 Resolution-aware network, a Self-supervision loss, and a Contrastive learning scheme. The proposed network is able to learn the 3D body shape and pose across different resolutions with a single model. The self-supervision loss encourages scale-consistency of the output, and the contrastive learning scheme enforces scale-consistency of the deep features. We show that both these new training losses provide robustness when learning 3D shape and pose in a weakly-supervised manner. Extensive experiments demonstrate that the RSC-Net can achieve consistently better results than the state-of-the-art methods for challenging low-resolution images.

Keywords: 3d human shape and pose · Low-resolution · Neural network · Self-supervised learning · Contrastive learning.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58545-7_17) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12354, pp. 284–300, 2020.
https://doi.org/10.1007/978-3-030-58545-7_17

1 Introduction

3D human shape and pose 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 [1,2,6,21,22,24,25,34,40–42,52]. However, this assumption does not always hold in practice, since lots of images in real scenes have low resolutions, such as surveillance cameras and sports videos [35,36,38,46–48]. As a result, existing algorithms designed for high-resolution images are prone to fail when applied to low-resolution inputs as shown in Fig. 1. In this paper, we study the relatively unexplored problem of estimating 3D human shape and pose from low-resolution images.

There are two major challenges of this low-resolution 3D estimation problem. 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 results in visual artifacts, which leads to poor 3D estimation. To address this issue, we propose a resolution-aware deep neural network for 3D human shape and pose 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.

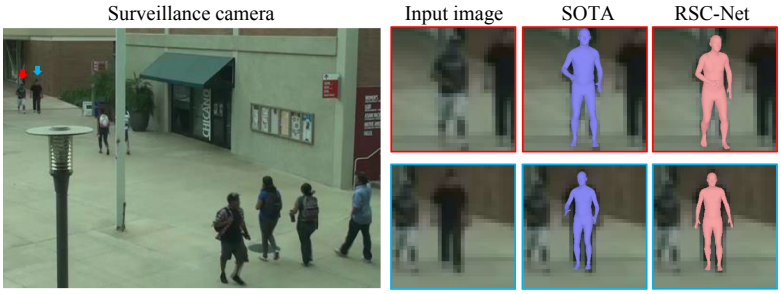

Fig. 1. 3D human shape and pose estimation from a low-resolution image captured from a real surveillance video. SOTA method [25] that works well for high-resolution images performs poorly at low-resolution ones.

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 [21,25], which complicates the training process. Whereas most training images have 2D keypoint

labels, they are usually not sufficient for predicting the 3D outputs due to the inherent ambiguities in the 2D-to-3D mapping. This 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 incomplete information typically does not achieve good results. Inspired by the self-supervised learning [26,44], 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. 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 feature vectors. Instead, we adapt the contrastive learning [7,14,39] 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 relatively unexplored problem of 3D human shape and pose estimation from low-resolution images and present a simple yet effective solution for it, called RSC-Net, which is based on a novel resolution-aware network that can handle arbitrary-resolution input with one single model. Second, we propose a self-supervision loss to address the issue of weak supervision. Furthermore, we introduce contrastive learning which effectively enforces the feature consistency 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 shape and pose estimation.

2 Related Work

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

3D human shape and pose estimation. Recent years have witnessed significant progress in the field of 3D human shape and pose estimation from a single image [1–3,6,9,21,22,24,25,34,40–42,49,50,52]. 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 [1,2,6,9,34,40], human silhouettes [2,34,40], body part segmentations [1], UV mappings [3], and optical flow [9], and then regressing the 3D human parameters [29] from the transformed outputs of the last stage either with iterative optimization [2,6] or neural networks [1,9,34,40]. As these methods map

the original input images into simpler representation forms which are generally sparse and can be easily rendered, they can exploit a large amount of synthetic data 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 [21]. 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 shape and pose estimation.

Without relying on new representations, the second kind of approaches can directly regress the 3D parameters from the input image [21,22,24,25,41,42,50], 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.* [21] include a generative adversarial network (GAN) [11] 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 [25] *et al.* integrate the optimization-based method [6] into the training process of the deep network to more effectively exploit the 2D keypoints. While achieving good improvements over [21] on high-resolution images, [25] 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. On the other hand, temporal information has also been exploited to enforce temporal consistency of the 3D estimation results, which however requires high-resolution video input [22,24,50]. Different from the above methods, we propose a 3D human shape and pose estimation algorithm using a single low-resolution image as input. We propose self-supervision loss and contrastive feature loss which effectively remedy the problem of insufficient 3D supervision.

Low-resolution image recognition. While there is no prior work for low-resolution 3D human shape and pose estimation, there are some related approaches to process low-resolution inputs for other image recognition tasks, such as 2D body pose estimation [35], face recognition [8,10,48], image classification [46], image retrieval [37,43], and object detection [12,27]. Most of these methods address the low-resolution issue by enhancing the degraded input, in either the image space [8,12,46] or the feature space [10,27,37,43]. One typical image-space method [12] 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, resulting in domain gap between the super-resolved and

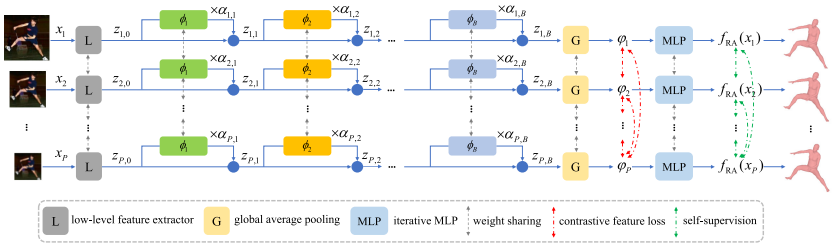

Fig. 2. Overview of the proposed algorithm. The resolution-aware network f_{RA} is trained with a combination of the basic loss (omitted in the figure for simplicity), self-supervision loss and contrastive feature loss. The modules with the same colors are shared across different resolutions, while the matrix α is resolution-dependent. Note that we resize the different resolution inputs $\{x_i\}$ to 224×224 with bicubic interpolation before feeding them into the network.

real high-resolution images. Unlike the image enhancement based approaches, the feature enhancement based methods [10,27,37,43] 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*. [10] use mean squared error (MSE) to enforce the similarity between the features of low-resolution and high-resolution images, which achieves good results for face recognition. Different from the above approaches, Neumann *et al*. [35] 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 shape and pose estimation. In this work, we apply the feature enhancement strategy to low-resolution 3D human shape and pose estimation. Instead of using MSE for measuring feature similarity, we introduce the contrastive learning [39] which can more effectively maximize the mutual information across the features of different resolutions. In addition, we handle different-resolution input with a resolution-aware neural network.

3 Algorithm

We study the problem of 3D human shape and pose estimation for a low-resolution image x. Instead of training different networks for each specific resolution, we propose a resolution-aware neural network f_{RA} which can handle the complex inputs with different resolutions. We first introduce the 3D human representation model and the baseline network for 3D human estimation with a single 2D image. Then we describe the proposed resolution-aware model as well as the self-supervision loss and the contrastive learning strategy for training the network. An overview of our method is shown in Fig. 2.

3.1 3D Human Representation

We represent the 3D human body using the Skinned Multi-Person Linear (SMPL) model [29]. The SMPL is a parametric model which describes the body shape and pose 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: $M = f_{\text{SMPL}}(\beta, \theta)$, where $M \in \mathbb{R}^{N \times 3}$ is a triangulated surface with $N = 6890$ vertices.

Similar to the prior works [21,25], we can predict the 3D locations of the body joints X with the body mesh using a pretrained mapping matrix $W \in \mathbb{R}^{K \times N}$:

$$X \in \mathbb{R}^{K \times 3} = WM. \tag{1}$$

With the 3D human joints, we use a perspective camera model to project the body joints from 3D to 2D. Assuming the camera parameters are $\delta \in \mathbb{R}^3$ which define the 3D translation of the camera, the 2D keypoints can be formulated as:

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

where f_{project} is the perspective projection function [13].

3.2 Resolution-Aware 3D Human Estimation

Baseline network. Similar to the existing methods [21,25], we use the deep convolutional neural network (CNN) for 3D human estimation, where the ResNet-50 [15] is employed to extract features from the input image. The building block of the ResNet (*i.e.*, ResBlock [16]) can be formulated as:

$$z_k = z_{k-1} + \phi_k(z_{k-1}), \tag{3}$$

where z_k is the output features of the k-th ResBlock, and ϕ_k represents the nonlinear function used to learn the feature residuals, which is modeled by several convolutional layers with ReLU activation [33]. The ResNet stacks B ResBlocks together, and the final output can be written as:

$$z_B = z_0 + \sum_{k=1}^{B} \phi_k(z_{k-1}), \tag{4}$$

where z_0 is the low-level features 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 (4) 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 [21,25].

Resolution-aware network. The baseline network is originally designed for high-resolution images with input size 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 impractical for real applications.

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 Fig. 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 (4), we learn a matrix α to adaptively fuse the residual maps from the ResBlocks for each input resolution as shown in Fig. 2, such that different resolutions can have suitable features for 3D estimation. Specifically, we formulate the output of the proposed resolution-aware network as:

$$z_{i,B} = z_{i,0} + \sum_{k=1}^{B} \alpha_{i,k} \phi_k(z_{i,k-1}), \quad i = 1, 2, \ldots, R, \tag{5}$$

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 (5), the original ResBlock in (3) is modified as: $z_{i,k} = z_{i,k-1} + \alpha_{i,k}\phi_k(z_{i,k-1})$. Note that we use a slightly different notation here compared with (3) and (4) 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.

Note that for training the above network, each high-resolution image in the training dataset needs to conduct the downsampling operation for $R-1$ times, such that each row of parameters in α have their corresponding training data. Whereas the original training datasets [4,18,28,31,32] 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 remedy the training issues and reduce the parameters in α, 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 $\{x_i, i = 1, 2, \ldots, P\}$ 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.

Basic loss function. Similar to the previous algorithms [21,25], 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 $X_g, J_g, \beta_{\mathrm{g}}, \theta_{\mathrm{g}}$ are the ground truth 3D and 2D keypoints and SMPL parameters. The basic loss function can be written as:

$$L_{\mathrm{b}} = \sum_i \|[\hat{\beta}_i, \hat{\theta}_i] - [\beta_{\mathrm{g}}, \theta_{\mathrm{g}}]\|_2^2 + \lambda_1 \|\hat{X}_i - X_{\mathrm{g}}\|_2^2 + \lambda_2 \|\hat{J}_i - J_{\mathrm{g}}\|_2^2, \tag{6}$$

where $\hat{X}_i$ and $\hat{J}_i$ are estimated with (1) and (2), 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 (6), 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 (6) similar to [21,22,25].

3.3 Self-Supervision

The 3D human shape and pose estimation is a weakly-supervised problem as only a small part of the training data has 3D labels, and it 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, which 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 [26] 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:

$$\sum_{i,j} \|f_{\mathrm{RA}}(x_i) - f_{\mathrm{RA}}(x_j)\|_2^2. \tag{7}$$

However, a major difference between our work and the original self-supervision method [26] is that we are generally more confident in the predictions of the higher-resolution images while [26] treats the results under different input augmentations equally. To exploit this prior knowledge, we improve the loss in (7) and propose a directional self-supervision loss:

$$\begin{aligned} L_{\mathrm{s}} &= \sum_{i,j} w_{i,j} \|\bar{f}_{\mathrm{RA}}(x_i) - f_{\mathrm{RA}}(x_j)\|_2^2, \\ w_{i,j} &= \mathbb{1}(j - i > 0) \cdot (j - i), \end{aligned} \tag{8}$$

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 (8) 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 [17], Hinton *et al.* show the effectiveness of the "dark knowledge" in soft targets, and similarly for low-resolution 3D human shape and pose 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 ones. Similar to (8), we have the feature consistency loss:

$$L_{\mathrm{f}} = \sum_{i,j} w_{i,j} g(\bar{\varphi}_i, \varphi_j), \tag{9}$$

where φ_i is the feature vector of the i-th resolution input image x_i, and $\bar{\varphi}$ denotes a fixed feature extractor without gradient back-propagation. $w_{i,j}$ is identical to that in (8). The function g is used to measure the distance between two feature vectors, and a straightforward choice is the MSE as in (8). However, the extracted features φ usually have very high dimensions, 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 [39,45]. Moreover, the unimodal losses such as MSE can be easily affected by the noise or insignificant

structures in the features, while a better loss function should exploit more global structures [39].

Towards this end, we propose a contrastive feature loss similar to [7,14,39,45] 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 image with different resolutions but far for different images. Mathematically, the contrastive function can be written as:

$$g(\bar{\varphi}_i, \varphi_j) = -\log \frac{\exp(s(\bar{\varphi}_i, \varphi_j)/\tau)}{\exp(s(\bar{\varphi}_i, \varphi_j)/\tau) + \sum_{q \in \mathcal{Q}} \exp(s(q, \varphi_j)/\tau)}, \tag{10}$$

where s represents the cosine similarity function, and τ is a temperature hyper-parameter. 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 [14] 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 (10), 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.

4 Experiments

We first describe the implementation details of the proposed RSC-Net. Then we compare our results with the state-of-the-art 3D human estimation approaches for different image resolutions. We also perform a comprehensive ablation study to demonstrate the effect of our contributions.

4.1 Implementation Details

We train our model and the baselines using a combination of 2D and 3D datasets similar to previous works [21,25]. For the 3D datasets, we use Human3.6M [18] and MPI-INF-3DHP [32] with ground truth of 3D keypoints, 2D keypoints, and SMPL 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 the 2D datasets including LSP [19], LSP-Extended [20], MPII [4], and MS COCO [28], which only have 2D keypoint labels. 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 Sect. 3.2, we split all the resolutions into $P = 5$

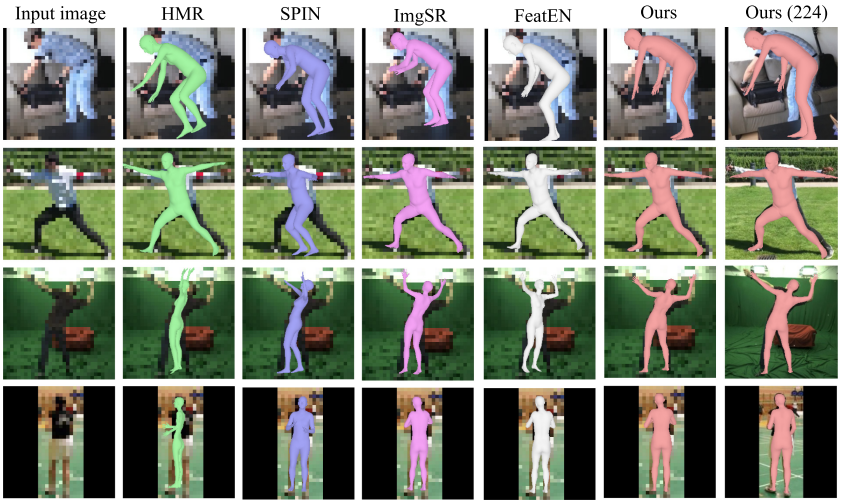

Fig. 3. Visual comparisons with the state-of-the-art methods on challenging low-resolution input. The input image has a resolution of 32×32. The results of high-resolution images are also included as a reference. All the baselines are trained with the same training data as our method.

Table 1. Quantitative evaluations against the state-of-the-arts on 3DPW [31].

Methods	MPJPE				MPJPE-PA			
	176	96	52	32	176	96	52	32
HMR	117.86	118.91	125.95	142.29	70.28	70.89	73.64	79.73
SPIN	112.72	113.60	120.71	137.61	69.20	69.40	72.21	78.44
ImgSR	116.47	117.74	127.78	146.58	66.62	67.48	72.34	81.07
FeaEN	107.97	109.42	119.08	143.51	61.37	62.13	66.62	77.21
Ours	**96.36**	**97.36**	**103.49**	**117.12**	**58.98**	**59.34**	**61.81**	**67.59**

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 (10) similar to [7]. As in [24], we initialize the baseline networks and our model with the parameters of [25]. We use the Adam algorithm [23] to optimize the network with a learning rate 5e-5. Similar to [24], we conduct evaluations on a large in-the-wild dataset 3DPW [31] 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 dataset [32]. Following [21,24,25], 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. To evaluate the performance of different image resolutions, we report results for the middle point of each resolution range, *i.e.*, 176, 96, 52, and 32.

4.2 Comparison to State-of-the-Art Methods

We compare against the state-of-the-art 3D human shape and pose estimation methods HMR [21] and SPIN [25] by fine-tuning them on different resolution images with the same training settings as our model. Since no previous approach has focused on the problem of low-resolution 3D human shape and pose estimation, we adapt the low-resolution image recognition algorithms to our task as new baselines, including both image super-resolution based [12] and feature enhancement based [43]. For the image super-resolution based method (denoted as ImgSR), we first use a state-of-the-art network RDN [51] to super-resolve the low-resolution image, and the output is then fed into SPIN [25] for regressing the SMPL parameters. Similar to [12], the network is trained to improve both the perceptual image quality and the 3D human shape and pose estimation accuracy. For feature enhancement (denoted as FeaEN), we apply the strategy in [43] 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 [5] used in the original work [43] does not work well in our experiments, and we instead use the LSGAN [30] combined with the basic loss (6) to train a stronger baseline network.

Table 2. Quantitative evaluations against the state-of-the-arts on MPI-INF-3DHP [32].

Methods	MPJPE				MPJPE-PA			
	176	96	52	32	176	96	52	32
HMR	114.89	113.27	114.82	133.25	74.77	74.45	76.35	85.30
SPIN	108.46	108.25	113.36	127.27	71.19	71.53	74.76	83.38
ImgSR	107.98	107.56	112.14	125.91	72.13	72.76	75.64	83.52
FeaEN	110.40	109.91	113.09	124.99	71.49	71.52	73.92	81.80
Ours	**103.36**	**103.39**	**106.04**	**115.80**	**70.01**	**70.27**	**72.56**	**78.68**

As shown in Table 1 and 2, the proposed method compares favorably against the baseline approaches on both 3DPW and MPI-INF-3DHP datasets for all the image resolutions. Note that we achieve significant improvement over the baselines on the 3DPW dataset as shown in Table 1, which demonstrates the effectiveness of the proposed method on the challenging in-the-wild images. We also provide a qualitative comparison against the baseline models in Fig. 3, where

the proposed method generates higher-quality 3D human estimation results on the challenging low-resolution input.

4.3 Ablation Study

We provide an ablation study using the 3DPW dataset in Fig. 4 and Table 3 to evaluate the proposed resolution-aware network, self-supervision loss, and contrastive feature loss. We first compare the proposed resolution-aware network with the baseline model ResNet50 [15,21]. As shown by "RA" and "Ba" in Table 3, our network can obtain slightly better results than the baseline network with the basic loss (6) as loss function. Further, we can achieve a more significant improvement over the baseline when adding the self-supervision loss (8) for training, *i.e.*, "RA+SS" *vs*. "Ba+SS", which further demonstrates the effectiveness of the resolution-aware structure.

Table 3. 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
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**

Second, we use the self-supervision loss in (8) to exploit the consistency of the outputs of the same input image with different resolutions. By comparing "RA+SS" against "RA" in Table 3, 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 show in "RA+SS+MS" of Table 3, which is mainly due to that the unimodal 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 features [39]. 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 3. Note that we adopt the cosine similarity in the contrastive feature loss (10) similar to prior methods [14,39,45]. Alternatively, one may only use the cosine distance function for measuring the distance of two features instead of using the whole contrastive loss (10). Nevertheless, this strategy does not work well as shown by "RA+SS+CD" in Table 3, which demonstrates the effectiveness of the proposed algorithm.

Analysis of training strategies. We also provide a detailed analysis of the alternative training strategies of our model. First, as described in Sect. 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 4 (*i.e.*, "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 (7) 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 (8) to exploit this prior knowledge. As shown in the second row of Table 4 (*i.e.*, "w/ SS-o"), using the original self-supervision loss (7) is not able to achieve high-quality results, as the network can minimize (7) by simply degrading the high-resolution predictions without improving the results of low resolution. In addition, we provide hierarchical supervision for low-resolution images in (8) which can act as soft targets during training. As shown in Table 4, only using the highest-resolution predictions as guidance (*i.e.*, "w/ SS-h") cannot produce as good results as the proposed approach (*i.e.*, "full model").

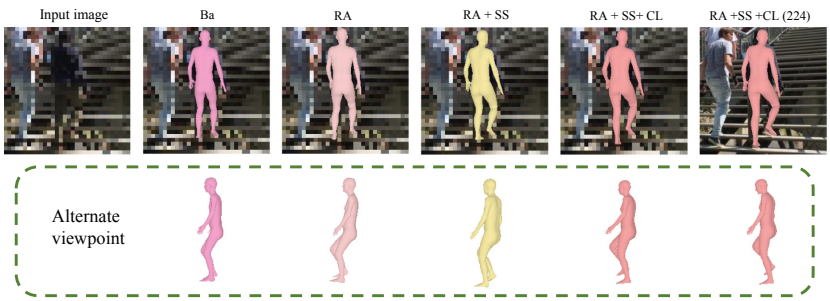

Fig. 4. Visual example which shows the effectiveness of the resolution-aware network, the self-supervision loss, and the contrastive learning feature loss.

Table 4. 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**

5 Conclusion

In this work, we study the challenging problem of low-resolution 3D human shape and pose estimation and present an effective solution, the RSC-Net. We propose a resolution-aware neural network which can deal with different resolution images with a single model. For training the network, we propose a directional self-supervision loss which can exploit the output consistency across different resolutions to remedy the issue of lacking high-quality 3D labels. In addition, we introduce a contrastive feature loss which is more effective than MSE for measuring high-dimensional 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 shape and pose estimation.

References

1. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: CVPR (2019)
2. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: CVPR (2018)
3. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: ICCV (2019)
4. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
5. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML (2017)
6. 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: ECCV (2016)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Cheng, Z., Zhu, X., Gong, S.: Low-resolution face recognition. In: ACCV (2018)
9. Doersch, C., Zisserman, A.: Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In: NeurIPS (2019)
10. Ge, S., Zhao, S., Li, C., Li, J.: Low-resolution face recognition in the wild via selective knowledge distillation. TIP **28**(4), 2051–2062 (2018)

11. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)
12. Haris, M., Shakhnarovich, G., Ukita, N.: Task-driven super resolution: Object detection in low-resolution images. arXiv:1803.11316 (2018)
13. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press (2003)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)
18. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI **36**(7), 1325–1339 (2013)
19. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC (2010)
20. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR (2011)
21. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
22. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019)
23. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014)
24. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR (2020)
25. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
26. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
27. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: CVPR (2017)
28. Lin, T.Y., et al.: Microsoft coco: Common objects in context. In: ECCV (2014)
29. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248 (2015)
30. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV (2017)
31. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
32. Mehta, D., et al.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV (2017)
33. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
34. Natsume, R., et al.: Siclope: Silhouette-based clothed people. In: CVPR (2019)
35. Neumann, L., Vedaldi, A.: Tiny people pose. In: ACCV (2018)
36. Nishibori, K., Takahashi, T., Deguchi, D., Ide, I., Murase, H.: Exemplar-based human body super-resolution for surveillance camera systems. In: International Conference on Computer Vision Theory and Applications (VISAPP) (2014)

37. Noh, J., Bae, W., Lee, W., Seo, J., Kim, G.: Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In: ICCV (2019)
38. Oh, S., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR (2011)
39. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv:1807.03748 (2018)
40. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: CVPR (2018)
41. Pumarola, A., Sanchez-Riera, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3dpeople: Modeling the geometry of dressed humans. In: ICCV (2019)
42. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV (2019)
43. Tan, W., Yan, B., Bare, B.: Feature super-resolution: Make machine see more clearly. In: CVPR (2018)
44. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NIPS (2017)
45. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
46. Wang, Z., Chang, S., Yang, Y., Liu, D., Huang, T.S.: Studying very low resolution recognition using deep networks. In: CVPR (2016)
47. Xu, X., Ma, Y., Sun, W.: Towards real scene super-resolution with raw images. In: CVPR (2019)
48. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: ICCV (2017)
49. 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: CVPR (2018)
50. Zhang, J.Y., Felsen, P., Kanazawa, A., Malik, J.: Predicting 3d human dynamics from video. In: ICCV (2019)
51. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)
52. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: ICCV (2019)