Available online at www.sciencedirect.com
ScienceDirect
Fundamental Research 00 (2023) 1–11
Procedia Computer Science

MILI: Multi-person Inference from a Low-resolution Image

Kun Li[a,*], Yunke Liu[a], Yu-Kun Lai[b], Jingyu Yang[a,*]

[a]*Tianjin University, Tianjin 300350, China*
[b]*Cardiff University, Cardiff CF24 4AG, United Kingdom*

Abstract

Existing multi-person reconstruction methods require the human bodies in the input image to occupy a considerable portion of the picture. However, low-resolution human objects are ubiquitous due to trade-off between the field of view and target distance given a limited camera resolution. In this paper, we propose an end-to-end multi-task framework for multi-person inference from a low-resolution image (MILI). To perceive more information from a low-resolution image, we use pair-wise images at high resolution and low resolution for training, and design a restoration network with a simple loss for better feature extraction from the low-resolution image. To address the occlusion problem in multi-person scenes, we propose an occlusion-aware mask prediction network to estimate the mask of each person during 3D mesh regression. Experimental results on both small-scale scenes and large-scale scenes demonstrate that our method outperforms the state-of-the-art methods both quantitatively and qualitatively. The code is available at `http://cic.tju.edu.cn/faculty/likun/projects/MILI`.

Keywords:
Multi-person reconstruction, Low-resolution human objects, End-to-end, Multi-task learning, Occlusion-aware prediction

1. Introduction

Multi-person reconstruction from a single image is of great importance in computer vision and computer graphics, which aims at estimating the 3D poses and shapes of all the people in an image. Existing methods [3, 4, 5, 6, 1] perform well in constrained experimental settings. However, these methods ignore some challenging situations, especially for low-resolution images, which are ubiquitous in the real world due to the limitations of cameras and transmission bandwidth. Existing methods tend to produce severely degraded results on low-resolution images, as shown in Fig. 1.

There are two challenges to achieve accurate and robust multi-person reconstruction from a low-resolution image: the first challenge is how to model the occlusions in multi-person scenes; the other challenge is how to deal with low-resolution images with limited information.

In this work, we propose MILI (Multi-person Inference from a Low-resolution Image), a two-stage framework for multi-person inference from a low-resolution image. To alleviate the occlusion problem in crowded scenes, we propose an occlusion-aware mask prediction network to estimate the mask of each person. With this multi-task learning setting, MILI can estimate more reasonable 3D meshes by leveraging the occlusion information guided by our mask prediction network. To exploit the limited information in the low-resolution image, we propose a restoration network by introducing the constraint of high-resolution images during training, which helps to extract richer information. With the restoration network, our model can learn how to

*Corresponding author

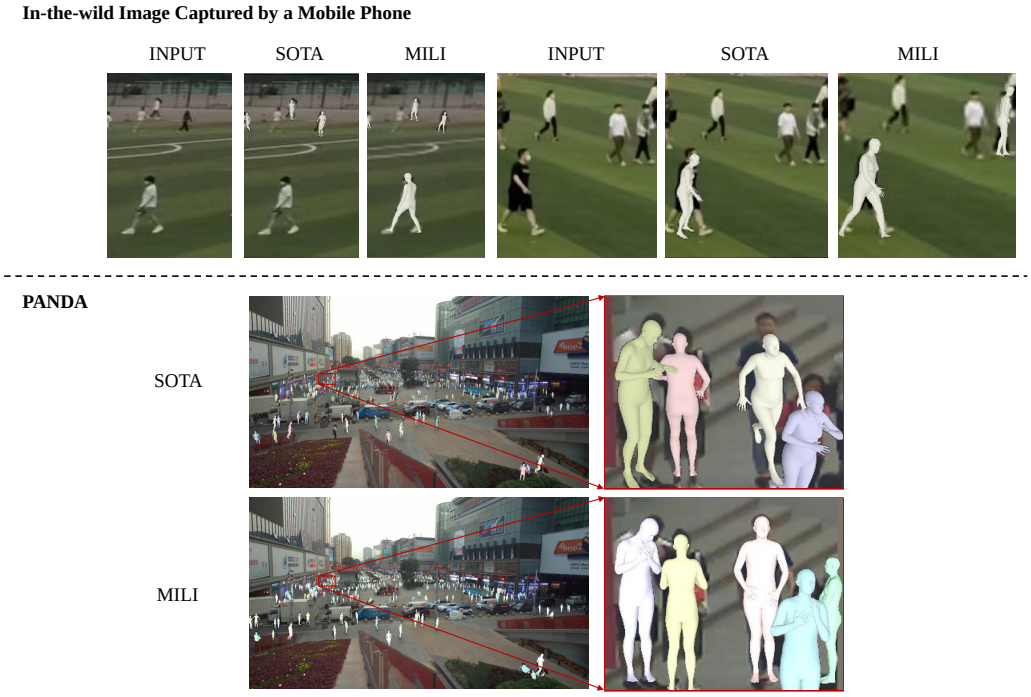

Figure 1. Given a low-resolution image, our method can achieve more accurate multi-person reconstruction compared with state-of-the-art method (SOTA) [1]. The inputs are captured by a mobile phone and downsampled from the PANDA dataset [2], respectively.

predict high-resolution information from low-resolution images, in favor of regressing more accurate 3D meshes. The code is available at `http://cic.tju.edu.cn/faculty/likun/projects/MILI`.

To summarize, our main contributions are as follows:

• We propose MILI, an end-to-end framework for multi-person reconstruction from a low-resolution image. To the best of our knowledge, MILI is the first framework that can regress accurate multi-person 3D meshes from a low-resolution image.

• We propose an occlusion-aware mask prediction network to estimate the mask of each person during 3D mesh regression, which helps alleviate the occlusion problem existing in crowded scenes. With this multi-task learning setting, our model can cope well with the occlusion problem.

• We design a restoration network for better training the low-resolution branch with the guidance of high-resolution images. We design a simple but effective loss to help perceive richer information from a low-resolution image and generate more accurate results.

2. Related Work

Multi-person 3D Pose Estimation. Existing work on 3D pose estimation can be categorized into two classes: top-down methods and bottom-up methods. Many approaches adopt the top-down framework due to the generality of Faster-RCNN [7], such as LCR Net [8], LCR Net++ [9] and 3DMPPE [10]. These methods directly regress the 3D pose from the feature of anchor-based proposals. To alleviate the ambiguity of directly estimating the 3D pose from a single image, Dabral *et al.* [11] decouple the reconstruction by regressing the 2D joints from the image and recovering 3D pose by 2D-to-3D lifting. Different from multi-stage methods, bottom-up methods are one-stage methods that estimate the poses of all persons in the image. Mehta *et al.* [12] propose occlusion-robust pose-maps to achieve better pose estimation results for the inputs with serious partial occlusions. To handle the ambiguity of absolute depth and scale in the scene, Zhen *et al.* [13] propose a depth-aware part association algorithm that regresses absolute 3D pose based on 2.5D representations. Considering many applications that require 3D pose estimation for a large number of people in the real world, Benzine *et al.* [14] present a novel method that handles different

scales of people in an image. However, none of the above methods can regress the shapes of persons, which are important for many downstream applications.

Multi-person 3D Pose and Shape Estimation. Multi-person reconstruction from a single image is challenging, and related work on this topic is rather limited. Existing methods adopt a parametric model SMPL [15], a low-dimensional vector, as the representation of the human mesh. Zanfir *et al.* [3] propose the first framework to estimate multiple persons by using 3D pose estimation as an intermediate result. Follow-up work [4] adds scene constraints to optimize the results. To ensure the depth consistency of all the people in the scene, CRMH [5] applies the interpenetration and depth ordering-aware loss to deal with the occlusion problem. Different from the top-down methods, BMP [6] proposes a one-stage approach to estimate more accurate depths by representing multiple persons as points in 3D space, which is suitable for dealing with occlusion situations. ROMP [1] adopts the body center heatmap and a mesh parameter map, and achieves state-of-the-art performance compared with previous work. However, all these methods use high-resolution images as inputs, and cannot adapt well to low-resolution inputs.

Low-Resolution Image Reconstruction. Since low-resolution images are ubiquitous in the real world, many researchers focus on different tasks using low-resolution images, *e.g.*, 2D pose estimation [16] and single-person reconstruction [17]. Neumann *et al.* [16] enhance 2D pose estimation from low-resolution images with probability maps of Gaussian models, which is hard to apply to 3D pose estimation. Xu *et al.* [17] propose the first method to regress a single-person mesh from a low-resolution image, which improves the accuracy of human reconstruction by applying self-supervision and contrastive learning in the feature domain. However, due to the occlusion in multi-person scenes, it is difficult to use the single-person reconstruction method to obtain reasonable multi-person reconstruction results, especially for crowded scenes.

In this paper, we propose an end-to-end multi-task framework, which regresses 3D poses and shapes of multiple persons from a real-world low-resolution image. Different from existing multi-person reconstruction methods [3, 4, 5, 6, 1], we design an occlusion-aware mask prediction network to alleviate the occlusion problem in multi-person scenes. With this multi-task setting, our model can cope well with the occlusion problem. Besides, to make up for lack of information, we propose a restoration network to improve the feature extraction by introducing the constraint of high-resolution images during training.

3. Method

We design MILI, an end-to-end multi-person inference framework from a low-resolution image, as shown in Fig. 2. MILI is trained with pair-wise low-resolution and high-resolution images, which detects all the persons with the proposed restoration network and regresses 3D poses and shapes of the detected persons under the multi-task setting of the proposed occlusion-aware mask prediction network (only for the low-resolution branch) and the SMPL estimation network [5]. Different from the existing work [5, 6, 1], MILI, as an end-to-end network, encourages the multi-person reconstruction from low-resolution images and significantly improves the robustness to occlusions with the occlusion-aware mask prediction network by refining the detection stage with segmentation. Meanwhile, to perceive more information from low-resolution images, the restoration network is proposed to guide the feature extraction of low-resolution images with the constraint of high-resolution images during training.

MILI consists of three aspects: a basic model (Sec. 3.1), an occlusion-aware mask prediction network (Sec. 3.2), and a restoration network (Sec. 3.3). To achieve a preliminary reconstruction from a low-resolution image, we use a two-branch architecture [17] to feed pair-wise images of high and low resolutions and then apply spatial attention, contrastive learning, and MSE (Mean Squared Error) loss to the network during training. To alleviate the occlusion problem, we design an occlusion-aware mask prediction network to improve the detection stage on the low-resolution image. The proposed restoration network generates effective features from low-resolution images under high-resolution guidance.

3.1. Basic Model

Resolution-aware Network. Existing multi-person reconstruction methods [5, 6, 1] usually fail to perform well on low-resolution images due to the limited information, as shown in Fig. 1. To perceive more information from low-resolution images, we design a two-branch network so that both low- and high-resolution images of the same scene are fed into the network during training, as shown in Fig. 2. To distinguish the features of images from different resolutions, we use an attention mechanism to extract the information from two branches respectively before the backbone, encouraging each branch to focus on the corresponding image features sufficiently. Since the main difference of two-resolution features is caused by spatial pixels, we em-

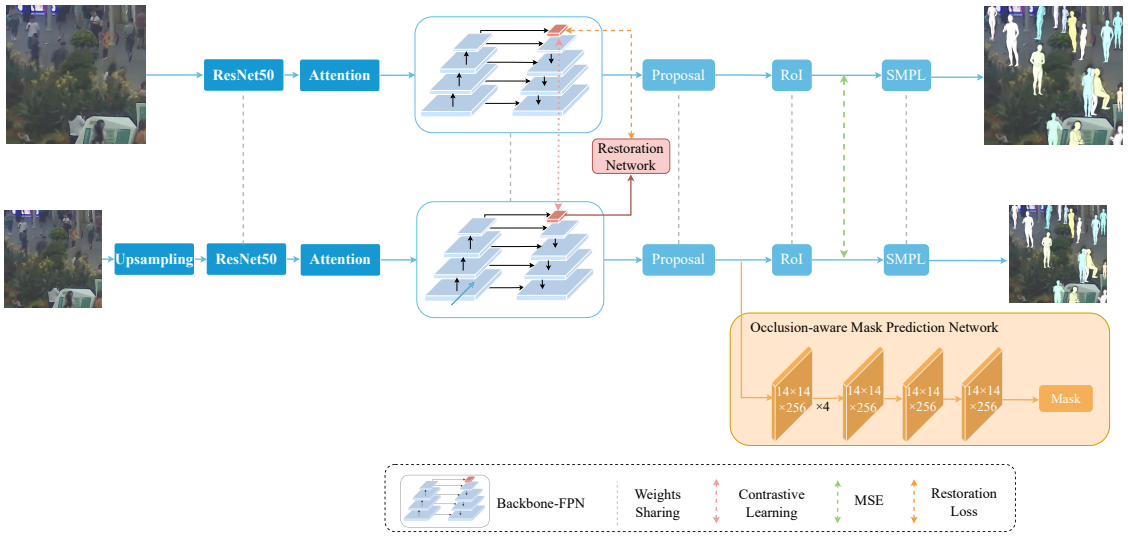

Figure 2. Overview of the proposed MILI, a two-stage multi-person reconstruction method. MILI is trained using pair-wise images at high resolution and low resolution from the same scene. Most of the two-branch network shares the parameters to better achieve the feature guidance from the high-resolution images, except for the occlusion-aware mask prediction network and the restoration network.

ploy the attention module on space. The model is formulated as follows:

$$y_i = R(\phi(x_i) \times x_i), \quad i \in \{0, 1\}, \tag{1}$$

where R represents the backbone of ResNet50 network [18], ϕ represents the spatial attention module [19], x_i represents the input, y_i represents the feature representation after backbone, and i is the branch index: "0" for the high-resolution branch and "1" for the low-resolution branch.

Contrastive Learning. Considering contrastive learning is widely used in image recognition [17, 20, 21, 22, 23] to encourage the consistency of features from the same scene, we use it to enforce the consistency of features encoded by the network across the different resolutions. As shown in Fig. 2, while there are many intermediate features, we adopt the top-level feature of the FPN (Feature Pyramid Network) [24] that consists of the most concentrated information with minimum computational complexity.

Constraint for Region of Interest (RoI). Through FPN [24], the samplings and bounding boxes are obtained, and thus each human mesh can be regressed via an SMPL estimation network. The most straightforward way is to restrain the features fed into the SMPL estimation network. However, the features consisting of multi-person samplings are hard to compare because of the complexity of person-person mapping between two resolutions. Observing that the bounding boxes directly affect the final parameter regression of human poses and shapes, we simply implement the MSE constraint on the bounding boxes that are in the form of RoI as a substitute, as shown in Fig. 2.

3.2. *Occlusion-aware Mask Prediction Network*

Although we can get a preliminary prediction of human meshes by the basic model, the occlusion problems are still severe in multi-person situations at low resolution. Compared with high-resolution images, low-resolution ones are a little blurry and lack high-frequency details, which is more difficult to estimate occluded persons. For example, the samplings of low-resolution images during the detection stage are inaccurate. To improve the accuracy of human detection, we design an occlusion-aware mask prediction network on the low-resolution branch. As shown in Fig. 2, the features obtained via the RoI module are fed into two networks: one is a 3D human reconstruction branch [5] to reconstruct the 3D model with multi-person losses, and the other is our proposed network to predict the image mask via human instance segmentation.

Different from [5], the features from the detection stage are better predicted with the human instance segmentation instead of the only 3D human reconstruction branch. Fig. 2 illustrates the network details. With the predicted bounding boxes from the RoI module, the occlusion-aware mask network finally obtains the masks of all the people.

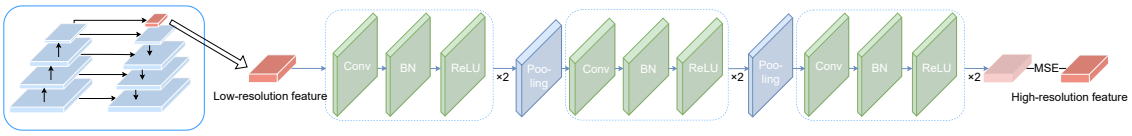

Figure 3. The details of the restoration network. The input is the feature from the top-level of FPN [24] and the output is compared with high resolution feature to calculate the loss.

3.3. Restoration Network

Although better features are obtained under the influence of high-resolution images in the basic network by the MSE constraint, further improvement can be achieved to get a more accurate human mesh by exploiting the high-resolution images during training. As the two branches of high resolution and low resolution share the parameters, and the high-resolution branch is fixed (*i.e.*, not back-propagated) during the training of the low-resolution branch, the network tends to focus on the low-resolution images gradually. Thus the prior features from the high-resolution branch are increasingly limited. The lack of feature information caused by the fewer pixels in low-resolution images finally leads to poor reconstruction. Inspired by the image restoration work [25] that introduces the guidance of high-resolution images with a skip connection network, we design a restoration network to improve feature extraction from low-resolution images. To reduce the computational complexity of the whole network, we restrain the top-level feature of FPN [24] compared to other levels. Fig. 3 illustrates the architecture of the restoration network, which is an encoder-decoder-like module. Through the convolution and pooling network, the low-resolution features are effectively guided by the high-resolution features, which greatly improves the performance and is demonstrated in the experiments (Sec. 4.4).

3.4. Loss Functions

In summary, the overall loss of the proposed network consists of three parts: L_{base}, L_{occ} and L_{res}, which are the losses of the basic model, the occlusion-aware mask prediction network and the restoration network, respectively.

Basic Loss. With the basic loss, the model achieves the preliminary reconstruction at low resolution. Given the general loss L_{mul} of multi-person reconstruction at high resolution [5], contrastive learning loss L_{cs}, and MSE constraint for RoI L_{RoI}, L_{base} is derived as:

$$L_{base} = L_{mul} + \lambda_{cs} \times L_{cs} + \lambda_{RoI} \times L_{RoI}, \tag{2}$$

where λ_{cs} and λ_{RoI} are the weights that balance the contributions of individual losses. Similar to [17], we define the contrastive learning loss as

$$L_{cs} = -\log \frac{\exp(\cos(\overline{x_h}, x_l)/\gamma)}{\exp(\cos(\overline{x_h}, x_l)/\gamma) + \sum_{q \in Q} \exp(\cos(q, x_l)/\gamma)}, \tag{3}$$

where x represents the feature of the top level, and the subscripts h and l represent the high-resolution and low-resolution branches, respectively. The distance between the high-resolution and low-resolution image features is measured by cosine of the angle between two vectors instead of MSE, which is more suitable for low-level features. Inspired by [17], $\overline{x_h}$ represents the fixed features of high resolution, and the gradients of the high-resolution branch are not back-propagated. Only low-resolution features x_l are encouraged to be more similar to high-resolution features x_h. γ is a temperature hyperparameter and Q is a queue of features of low-resolution images from other scenes, which are updated during training gradually. The main idea of this loss is to reduce the distance between the images from different scenes and make the low-resolution features closer to the high-resolution features. With the same features $\overline{x_h}$ and x_l, L_{RoI} is formulated as

$$L_{RoI} = \|\overline{x_h} - x_l\|_2^2 . \tag{4}$$

Loss of Occlusion-aware Mask Prediction Network. Occlusion-aware mask prediction network is designed to address the serious occlusion problems in low-resolution images, which is formulated as

$$L_{occ} = \sum_{p=1}^{P} \left\| \nu(x_p) - \tau_p \right\|_2^2 , \tag{5}$$

where P represents the number of detection boxes, $\nu(\cdot)$ represents the occlusion-aware mask prediction network, and x_p is the feature at low resolution. The instance segmentation of each person is predicted by $\nu(x)$, and we restrain the final mask with the ground truth via MSE. By restricting the segmentation of the recognized person, the accuracy of the detection stage can be improved to obtain a better human reconstruction result finally.

Loss of Restoration Network. To compensate for the limited information at low resolution, the guidance of high resolution is defined as

$$L_{res} = \|\overline{x_h} - \varphi(x_l)\|_2^2, \tag{6}$$

where $\phi(\cdot)$ represents the basic convolutional modules, and $\overline{x_h}$ represents the fixed feature of high-resolution branch that is not back-propagated, consistent with our basic model.

4. Experiments

In this section, we first introduce the datasets and implement details in Sec. 4.1 and Sec. 4.2, respectively, and then compare our method with state-of-the-art methods quantitatively and qualitatively in Sec. 4.3 and perform ablation studies to analyze the effects of different components of our approach in Sec. 4.4. Finally, we discuss the failure cases of our method in Sec. 4.5.

4.1. Datasets

***PANDA* [2]**: It is the only large-scale human-centric dataset which provides bounding boxes for the detection stage. Besides, we annotate 2D joints of human poses. We use *02 OCT Habour*, *05 Basketball Court* and *07 University Campus* as training set, and *10 Huaqiangbei* as test set. Because the original images have a gigapixel-level resolution, we crop the images into blocks with adaptively different sizes as input.

***Human3.6M* [26]**: It is an indoor dataset with a single person in each frame, which provides 3D pose annotations. Following *Protocol 1* of [27], ,we use *S1*, *S5*, *S6*, *S7* and *S8* for training.

***PII* [28]**: It is an in-the-wild dataset of multiple persons with 2D pose annotations. We use the training set for training.

***MPI-INF-3DHP* [29]**: It is a single person dataset with 3D pose annotations. We use *S1* to *S8* for training.

***COCO* [30]**: It is an in-the-wild dataset with 2D pose and instance segmentation annotations. We use the 2D poses for training. Meanwhile, the segmentation masks are adopted for the occlusion-aware prediction and SMPL estimation networks. We use the training set for training and the evaluation set for evaluation.

***MuPoTS-3D* [12]**: It is a multi-person dataset with 3D joint annotations for all the people in the scene. We use this dataset for evaluation.

***Panoptic* [31]**: It is a dataset with multiple people captured in a panoptic studio. We use this dataset for evaluation.

4.2. Implement Details

Since *PANDA* [2] is the only large-scale dataset that is closer to real-world outdoor scenes with a large number of people but without 3D joint annotations, we divide the datasets into two groups: the large-scale dataset (*PANDA* [2]) and the small-scale datasets (*Human3.6M* [26], *MPII* [28], *MPI-INF-3DHP* [29], *COCO* [30]). Our MILI is trained on two types of datasets, respectively. Because *PANDA* [2] has no segmentation labels, we add *COCO* [30] when training the SMPL estimation network. The high-resolution branch is adopted to guide the low-resolution branch during training, while the multi-person meshes are recovered from the low-resolution image only through the low-resolution branch during evaluation.

Setting Details. Since existing datasets are all high-resolution images, we downsample the inputs to low resolution, following [17]. Specifically, we set the low-resolution size at 286×176 and 208×128 for large-scale and small-scale datasets, respectively. Then, we uniformly resize the inputs to 832×512 by interpolation, keeping the same aspect ratio and padding with zero. For weight settings, γ is set to 0.01, and Q is set to of length 400. Due to the large RoI value, λ_{RoI} is set to $1e-8$ to balance the final result affected by the loss. Similar to [13], we first fine-tune the model at 832×512 as a result of the high-resolution branch, and it is not back-propagated when guiding the low-resolution branch during training. The final reconstruction is achieved by an SMPL estimation network [5]. The model is trained on a desktop with an NVIDIA RTX 3090 GPU with a batch size of 6 images.

Evaluation Metrics. To evaluate the reconstruction accuracy, we adopt mean per joint position error (MPJPE) and percentage of correct keypoints (PCK) on 3D poses.

4.3. Comparison

To demonstrate the effectiveness of the proposed model, we compare MILI with four state-of-the-art multi-person reconstruction methods: Zanfir *et al.* [3], CRMH [5], BMP [6] and ROMP [1]. For a fair comparison, we fine-tune both CRMH [5] and ROMP [1] on *PANDA* [2] and small-scale datasets, respectively. Note that the codes of Zanfir *et al.* [3] and BMP [6] are not publicly available, and all the results of them in the tables are from the original papers.

Results on Large-scale Dataset. Fig. 1 illustrates our reconstruction on an image from *10 Huaqiangbei* of *PANDA* [2], which demonstrates that our model can recognize the persons of different scales and recover reasonable human meshes. To illustrate more details of the

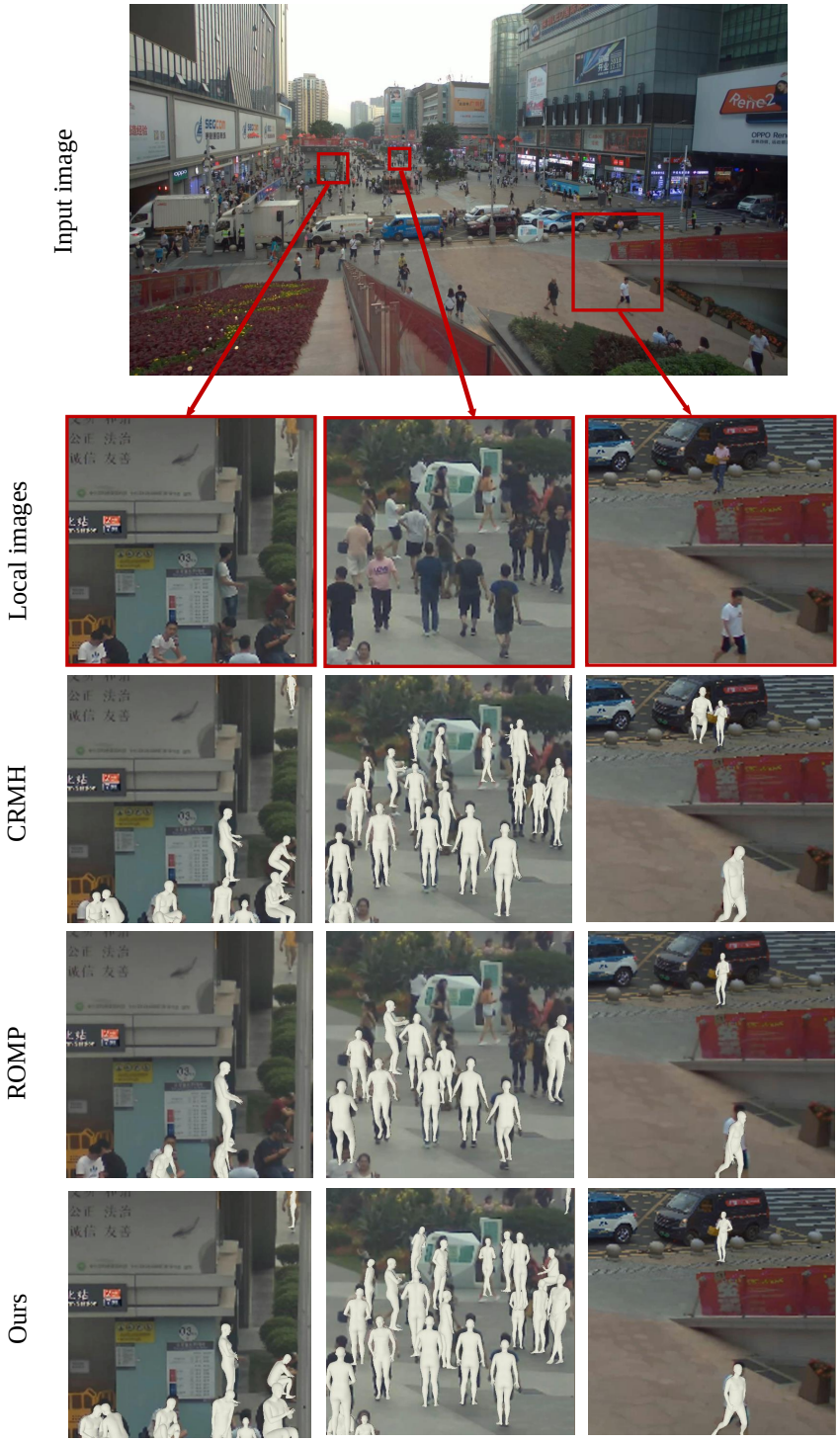

Figure 4. Qualitative results on low-resolution images of *PANDA* [2], compared with CRMH [5] and ROMP [1].

results, we visualize the meshes generated from different scales of blocks in Fig. 4. Compared with CRMH [5] and ROMP [1], our model achieves more accurate reconstructions from the low-resolution images with appropriate bounding boxes in different scales. Due to the lack of 3D annotations, no quantitative results are given on *PANDA* [2].

Results on Small-scale Datasets. Fig. 5 shows the qualitative results compared with the state-of-the-art methods on small-scale datasets. MILI performs better in complex in-the-wild scenes. Especially for very low-resolution images where people are very blurry, our model can detect the humans and predict the human poses and shapes that are more consistent with real situations. As shown in Tab. 1 and Tab. 2, our method achieves the state-of-the-art performance on *Panoptic* [31] and *MuPoTS-3D* [12]. For a fair comparison, we only compare with CRMH [5] and BMP [6] on *MuPoTS-3D* [12] since ROMP [1] uses this dataset as the training dataset. Tab. 2 illustrates that our method improves reconstruction accuracy by 14% and 2%, respectively.

We also visualize some meshes with bounding boxes estimated from low-resolution images of *COCO* [30] and *MuPoTS-3D* [12] in Fig. 6. The results suggest that our model can recover accurate human meshes even in blurry images with a large variation of person scales.

Method	*Haggling*	*Mafia*	*Ultim.*	*Pizza*	Mean
Zanfir *et al.* [3]	141.4	152.3	145.0	162.5	150.3
CRMH [5]	127.38	136.02	154.47	156.37	143.56
BMP [6]	120.4	132.7	**140.9**	147.5	135.4
ROMP [1]	134.47	161.51	157.67	164.05	154.43
Ours	**116.26**	**123.52**	141.24	**143.71**	**131.18**

Table 1. Results on *Panoptic* [31]. We use MPJPE (Mean Per Joint Position Error) as metric.

Method	CRMH [5]	BMP [6]	Ours
3DPCK	66.34	73.83	**75.42**

Table 2. Results on *MuPoTS-3D* [12].

4.4. Ablation Study

To verify the validity of the proposed model, we conduct ablation experiments on *COCO* [30] and *MuPoTS-3D* datasets [12] qualitatively, and on *Panoptic* [31] qualitatively and quantitatively.

Occlusion-aware Mask Prediction Network. As shown in Fig. 7, our full model can detect the humans of different scales. Without the occlusion-aware mask prediction network, the model tends to predict the bounding boxes with poor accuracy, resulting in wrong detection results. Quantitative results on *Panoptic* [31] are illustrated in Tab. 3. The full model achieves better performance, which demonstrates the effectiveness of our occlusion-aware mask prediction network.

Restoration Network. As shown in Fig. 8, the model without high-resolution feature guidance generates much less accurate poses and shapes, . Compared with that, the full model can regress the meshes with more appropriate scales by the restoration network with adequate access to feature information. As shown in Tab. 3, The full model obtains more accurate results, which demonstrates the effectiveness of our restoration network.

Method	w/o Occ.	w/o Res.	Full Model
MPJPE	133.13	142.01	**131.18**

Table 3. Ablation results on *Panoptic* [31]. Occ. is short for occlusion-aware mask prediction network, and Res. is short for restoration network, respectively.

4.5. Failure Cases

Although our model achieves promising reconstruction results, there are still two types of failure cases during evaluation: false detection and wrong unified pose for occlusion. As shown in Fig. 9 (left), MILI recognizes the human-like area as a human but there are actually no people. One possible reason is that the distribution of features without people is similar to that of areas containing people. As for occlusions, MILI tends to directly predict the reconstruction in a unified pose among all reconstructions, resulting a wrong sit-down pose, as shown in Fig. 9 (right). These problems will be further solved in future work.

5. Conclusion

In this paper, we design an end-to-end multi-person inference framework from a low-resolution image. Firstly, we propose a basic model to achieve a better reconstruction on low-resolution images. Then, we improve the model via an occlusion-aware mask prediction network and a restoration network. Our method achieves state-of-the-art performance on multiple benchmarks. Specifically, the bounding box results are significantly improved with the occlusion-aware mask prediction network. The low-resolution branch can get more effective features to reconstruct an accurate mesh by the restoration network, which encourages the features of low-resolution images to be generated under the effective guidance of a high-resolution branch during training. Comparison results on the large-scale

Figure 5. Qualitative results of CRMH [5], ROMP [1] and ours on *COCO* [30] and *MuPoTS-3D* [12].

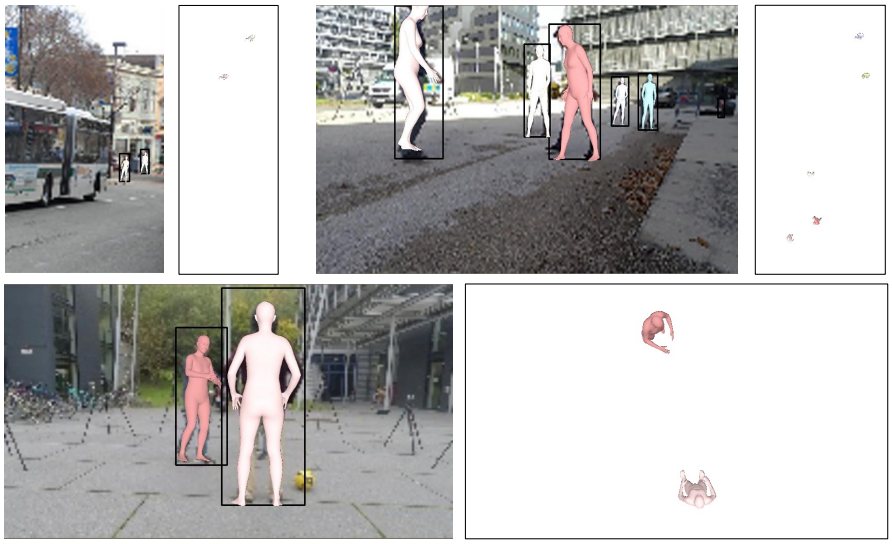

Figure 6. Qualitative results on *COCO* [30] and *MuPoTS-3D* [12]. We visualize the meshes with front and top viewpoints.

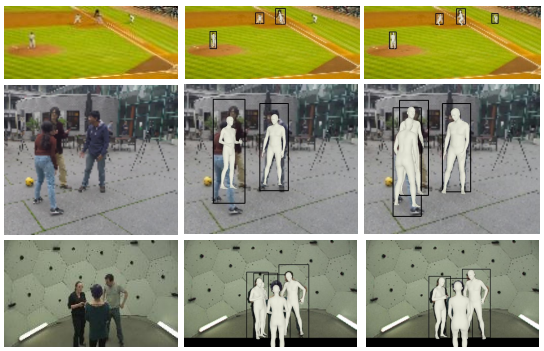

Figure 7. Qualitative results of the models without and with the occlusion-aware mask prediction network on *COCO* [30], *MuPoTS-3D* [12] and *Panoptic* [31] datasets (from top to bottom).

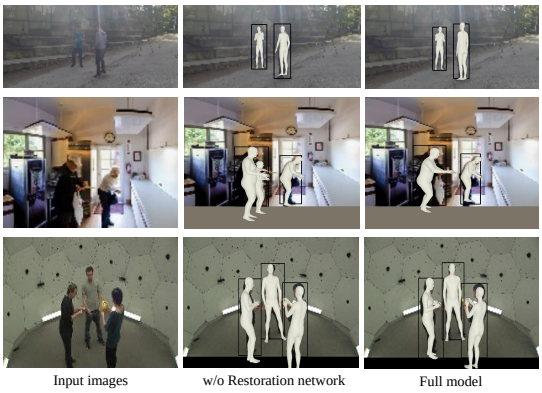

Figure 8. Qualitative results of the models without and with restoration network on *COCO* [30], *MuPoTS-3D* [12] and *Panoptic* [31] datasets (from top to bottom).

Figure 9. Failure cases on *PANDA* [2]. The left is an example with an extra person, and the right is a wrongly estimated pose caused by large occlusion.

and small-scale datasets demonstrate that our method is superior to the existing multi-person reconstruction methods in both detection and reconstruction.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (62122058, 62171317, and 62231018). We are grateful to Associate Editor and anonymous reviewers for their help in improving this paper.

References

[1] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, T. Mei, Monocular, one-stage, regression of multiple 3D people, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 11179–11188.

[2] X. Wang, X. Zhang, Y. Zhu, Y. Guo, X. Yuan, L. Xiang, Z. Wang, G. Ding, D. J. Brady, Q. Dai, PANDA: A gigapixel-level human-centric video dataset, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3268–3278.

[3] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, C. Sminchisescu, Deep network for the integrated 3D sensing of multiple people in natural images, in: Advances in Neural Information Processing Systems 31, 2018.

[4] A. Zanfir, E. Marinoiu, C. Sminchisescu, Monocular 3D pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2148–2157.

[5] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, K. Daniilidis, Coherent reconstruction of multiple humans from a single image, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5578–5587.

[6] J. Zhang, D. Yu, J. H. Liew, X. Nie, J. Feng, Body meshes as points, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 546–556.

[7] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Analysis and Machine Intelligence 39 (6) (2017) 1137–1149.

[8] G. Rogez, P. Weinzaepfel, C. Schmid, LCR-net: Localization-classification-regression for human pose, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3433–3441.

[9] G. Rogez, P. Weinzaepfel, C. Schmid, LCR-net++: Multi-person 2d and 3D pose detection in natural images, IEEE Trans. Pattern Analysis and Machine Intelligence 42 (5) (2019) 1146–1161.

[10] G. Moon, J. Y. Chang, K. M. Lee, Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 10133–10142.

[11] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, A. Jain, Multi-person 3D human pose estimation from monocular images, in: International Conference on 3D Vision, 2019, pp. 405–414.

[12] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, C. Theobalt, Single-shot multi-person 3D pose estimation from monocular

rgb, in: 2018 International Conference on 3D Vision, 2018, pp. 120–130.
[13] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, X. Zhou, SMAP: Single-shot multi-person absolute 3D pose estimation, in: European Conference on Computer Vision, 2020, pp. 550–566.
[14] A. Benzine, F. Chabot, B. Luvison, Q. C. Pham, C. Achrd, Pandanet : Anchor-based single-shot multi-person 3D pose estimation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6856–6865.
[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black, SMPL: a skinned multi-person linear model, ACM Trans. Graphics 34 (2015) 1–16.
[16] L. Neumann, A. Vedaldi, Tiny people pose, in: Asian Conference on Computer Vision, 2019, pp. 558–574.
[17] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, F. De la Torre, 3D human shape and pose from a single low-resolution image with self-supervised learning, in: European Conference on Computer Vision, 2020, pp. 284–300.
[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[19] H. Jie, S. Li, S. Gang, S. Albanie, Squeeze-and-excitation networks, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99).
[20] X. Wang, R. Zhang, C. Shen, T. Kong, L. Li, Dense contrastive learning for self-supervised visual pre-training, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3024–3033.
[21] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020, pp. 1597–1607.
[22] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[23] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, in: European Conference on Computer Vision, 2020, pp. 776–794.
[24] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[25] X. J. Mao, C. Shen, Y. B. Yang, Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections, Advances in Neural Information Processing Systems 29.
[26] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Analysis and Machine Intelligence 36 (7) (2013) 1325–1339.
[27] A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik, End-to-end recovery of human shape and pose, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7122–7131.
[28] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, Human pose estimation: New benchmark and state of the art analysis, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
[29] D. Mehta, H. Rhodin, C. Dan, P. Fua, C. Theobalt, Monocular 3D human pose estimation in the wild using improved cnn supervision, in: 2018 International Conference on 3D Vision, 2017, pp. 120–130.
[30] T. Y. Lin, M. Maire, S. Belongie, J. Hays, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[31] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, Y. Sheikh, Panoptic studio: A massively multiview system for social motion capture, in: IEEE/CVF International Conference on Computer Vision, 2015, pp. 3334–3342.