Multi-Scale Structure-Aware Network for Human Pose Estimation

Lipeng Ke[1,2]; Ming-Ching Chang[1]; Honggang Qi[2]; Siwei Lyu[1]

[1]University at Albany, State University of New York, USA;
[2]University of Chinese Academy of Sciences, Beijing, China;

1. OVERVIEW

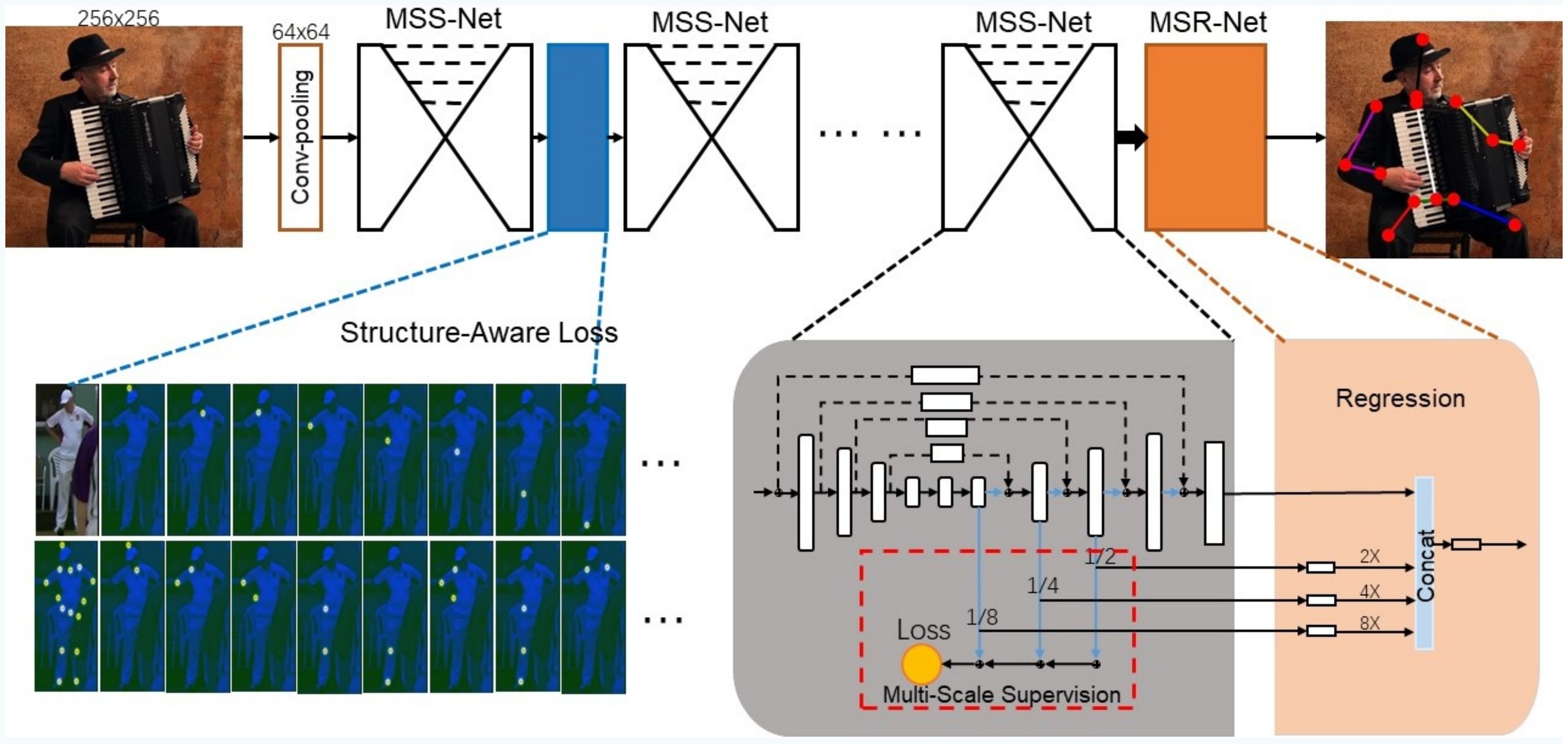

Fig. 1 Pipeline of Multi-Scale Structure-Aware Network for Human Pose Estimation

2. MOTIVATION

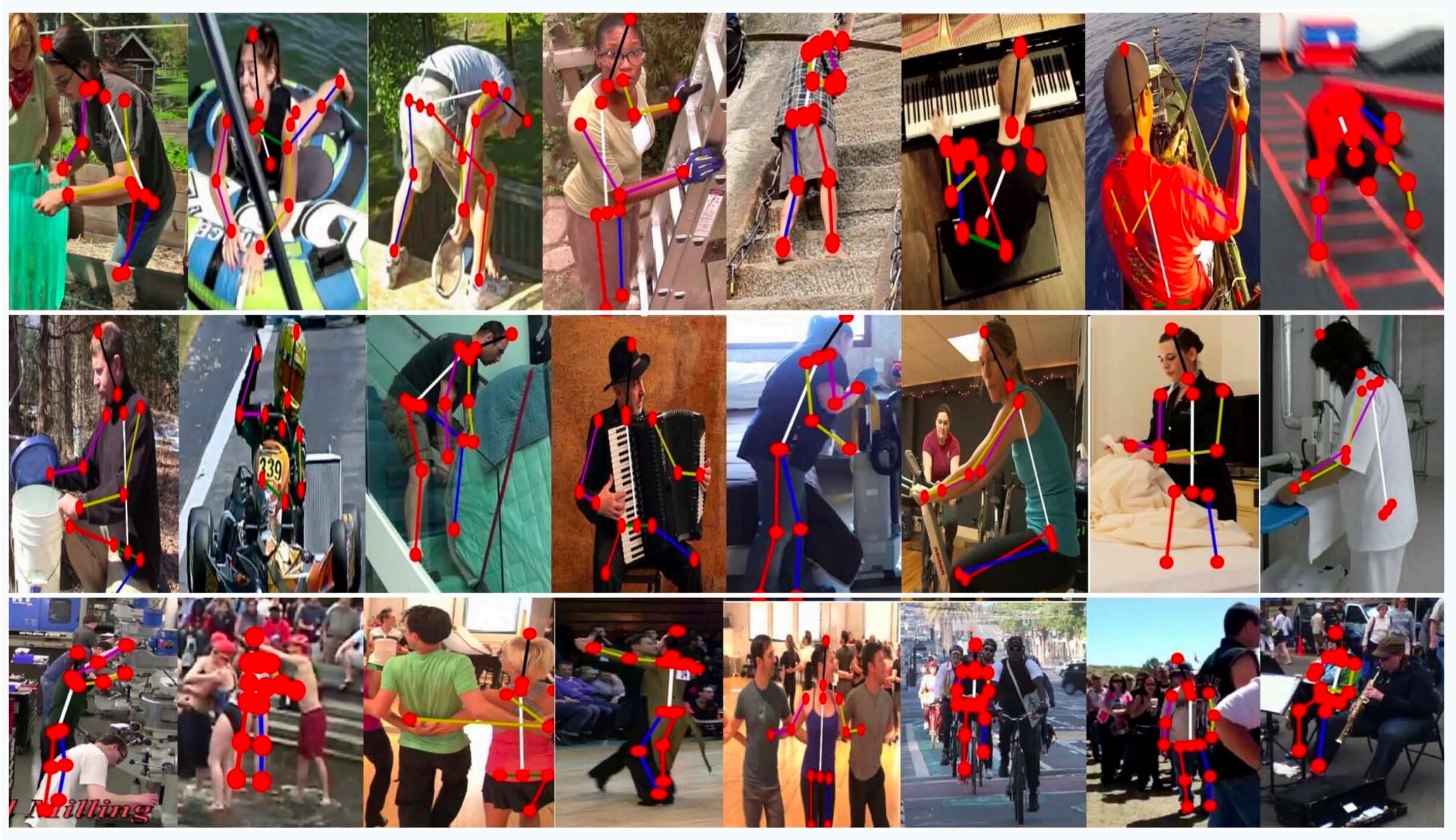

Fig. 2 State-of-the-art pose estimation networks face difficulties in diverse activities and complex scenes, which can be organized into three challenges: (top row) large scale varieties of body keypoints in the scenes, (middle row) occluded body parts or keypoints, (bottom row) ambiguities in matching multiple adjacent keypoints in crowded scenes.

Although great progress has been made, state-of-the-art DNN-based pose estimation methods still suffer from several problems such as, scale instability and insufficient structural priors (challenging cases are shown in Fig. 2)

In this paper, we propose a holistic framework to effectively address the drawbacks in the existing state-of-art hourglass networks. Our method is based on two neural networks: the multi-scale supervision network (MSS-net) and the multi-scale regression network (MSR-net).

3. CONTRIBUTION

The main contributions of this paper can be summarized as follows:

- We introduce the multi-scale supervision network (MSS-net) together with the multi-scale regression network (MSR-net) to combine the rich multi-scale features.
- Both the MSS-net and MSR-net are designed using a structure-aware loss to explicitly learn the human skeletal structures from multi-scale features.
- We propose a keypoint masking training scheme that can fine-tune our network pipeline by generating effective training samples for keypoint occlusions and cluttered scenes.

4. METHOD

4.1 Multi-Scale Supervision

We perform multiple layer-wise supervision at each of the deconv layers of the MSS-net, where each layer corresponds to a certain scale.

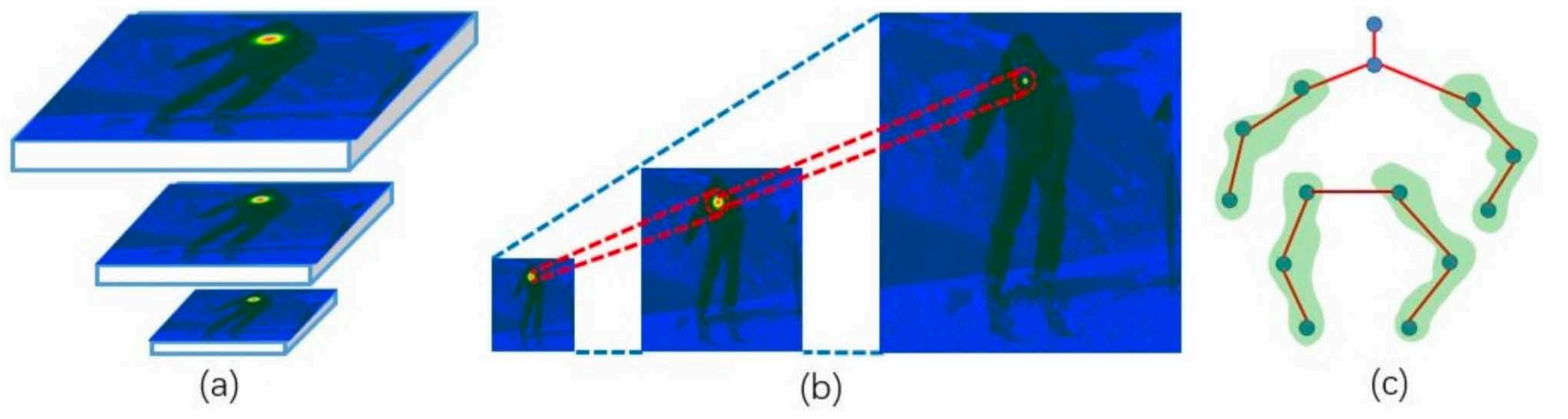

Fig. 3. In the MSS-net, the refinement of keypoint localization in up-sampling resolution works in analogy to the 'attention' mechanism used in the con- ventional resolution pyramid search. (a) shows the multi-scale heatmaps of the keypoint of the thorax. (b) shows the refinement of the keypoint heatmaps during the deconv up- sampling, where the location of the thorax is refined with increased accuracy. (c) shows our human skeletal graph with the visualization of keypoint connectivity links.

4.2 Multi-Scale Regression

The MSR-net takes the multi-scale heatmaps as input, and match them to the ground-truth keypoints at respective scales to effectively combine heatmaps across all scales to refine the estimated poses.

The efficacy of the multi-scale, high-order keypoint regression performed in the MSR-net is showed in Fig. 4.

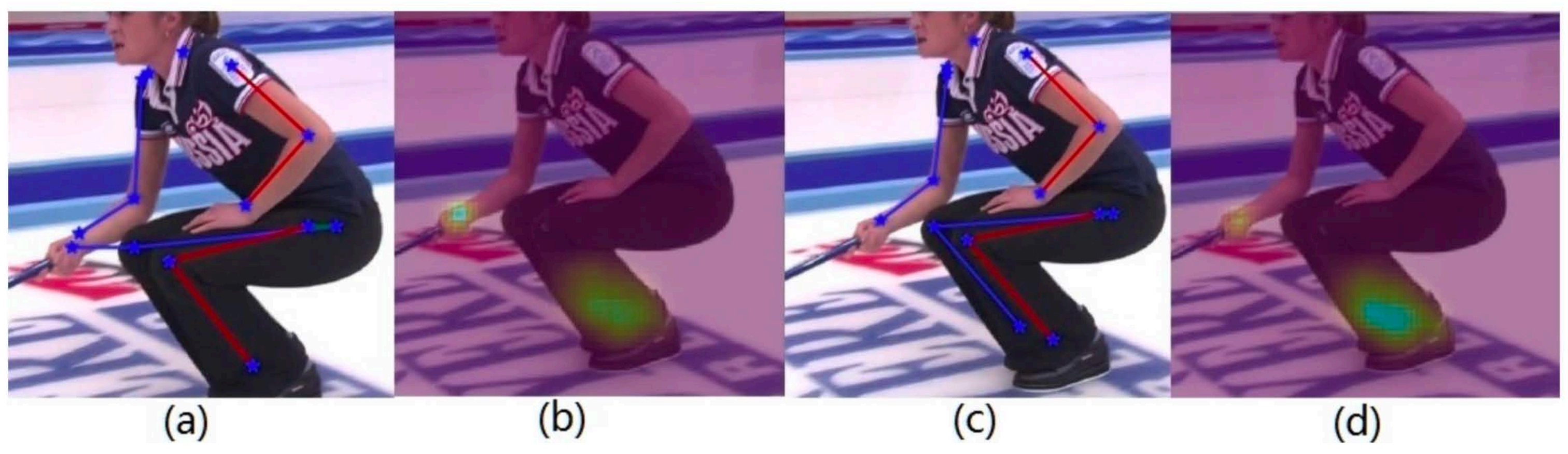

Fig . 4 Muti-scale keypoint regression to disambiguate multiple peaks in the keypoint heatmaps. (a-b) shows an example of (a) keypoint prediction and (b) heatmap from the MSS-net hourglass stacks, which will be fed into the MSR-net for regression. (c-d) shows (c) the output keypoint locations and (d) heatmap after regression. Observe that the heatmap peaks in (d) are more focused compared to (b).

4.3 Structure-Aware Loss

We design a structure-aware loss function following a graph to model the human skeletal structure showed in Fig. 3 (c), where single key-point is in blue dot, pair-wise key-point is connected by red line and triplet key-point is grouped by green map. Pair-wise and triplet key-point present the high order relationship in human skeleton. The loss at the i-th scale is formally defined as:

$$L^i_{SA} = \frac{1}{N}\sum_{n=1}^{N}||P^i_n - G^i_n||_2 + \alpha\sum_{i=1}^{N}||P^i_{S_n} - G^i_{S_n}||_2.$$

Each node $S_n \in S$ represent a body keypoint of the human skeleton and its connected keypoints, $n \in \{1, ..., N\}$. The first term of the equation represents individual keypoint matching loss. The second term represents the structural matching loss, where P_{S_n} and G_{S_n} are the combination of the heatmaps from individual keypoint n and its neighbors in graph S. Hyperparameter α is a weighing parameter balancing the two terms.

4.4 Keypoint Masking

We develop a novel keypoint masking data augmentation scheme to increase training data to fine-tune our networks.

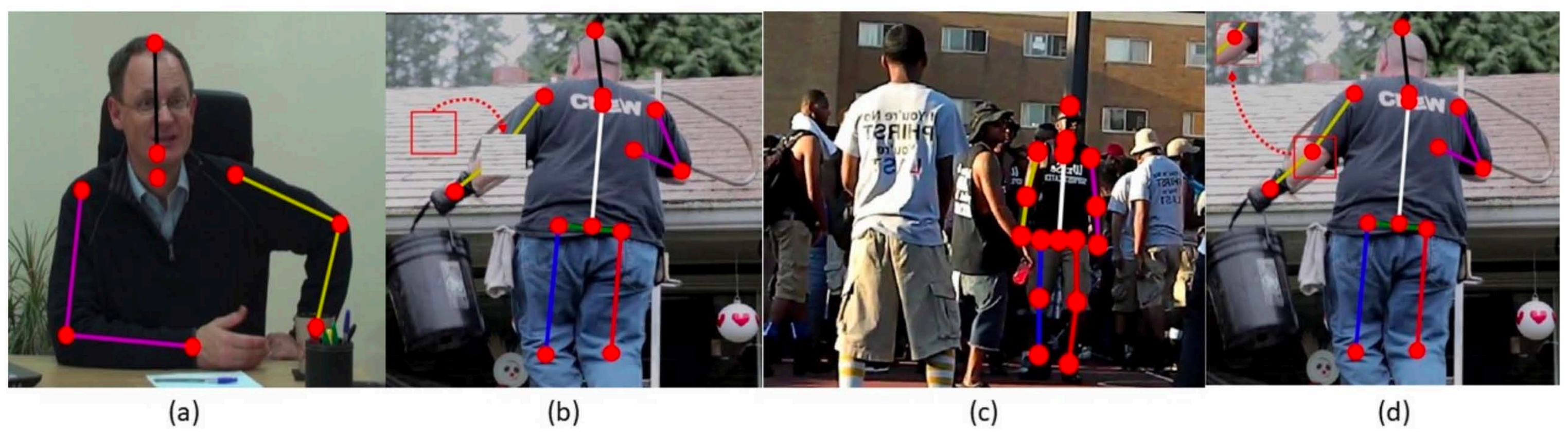

Fig. 5. Keypoint masking to simulate the hard training samples. (a) is a common case in human pose estimation, the keypoint (left-wrist) is occluded by an object, but it can be estimated from the limbs. (c) is another difficult case, where the nearby persons' keypoint can be mismatched to the target person. Thus there are two kind of keypoint masking, (b) is the background keypoint masking which crop a background patch and paste on a keypoint to simulate the keypoint invisible, (d) is the keypoint duplicate masking which crop a keypoint patch an

5. EXPERIMENT

Table 1. *Results on the FLIC dataset (PCK = 0.2)*

	Elbow	Wrist
Tompson *et al.* CVPR'15 [9]	93.1	92.4
Wei *et al.* CVPR'16 [11]	97.8	95.0
Newell*et al.* ECCV'16 [12]	99.0	97.0
Our model	**99.2**	**97.3**

Table 2. *Evaluation results on the MPII pose dataset (PCK^h = 0.5). Results were retrieved on 03/15/2018.*

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Our method	**98.5**	**96.8**	**92.7**	88.4	90.6	89.3	**86.3**	**92.1**	63.8
Chen *et al.* ICCV'17 [23]	98.1	96.5	92.5	**88.5**	90.2	**89.6**	86.0	91.9	61.6
Chou *et al.* arXiv'17 [24]	98.2	**96.8**	92.2	88.0	**91.3**	89.1	84.9	91.8	63.9
Chu CVPR'17 [13]	**98.5**	96.3	91.9	88.1	90.6	88.0	85.0	91.5	63.8
Luvizon *et al.* arXiv'17 [25]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2	63.9
Ning *et al.* TMM'17 [26]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2	63.6
Newell ECCV'16 [12]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9	62.9
Bulat ECCV'16 [21]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7	59.6
Wei CVPR'16 [11]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5	61.4
Insafutdinov ECCV'16 [27]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5	60.8
Belagiannis FG'17 [28]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1	58.8

6. DEMO

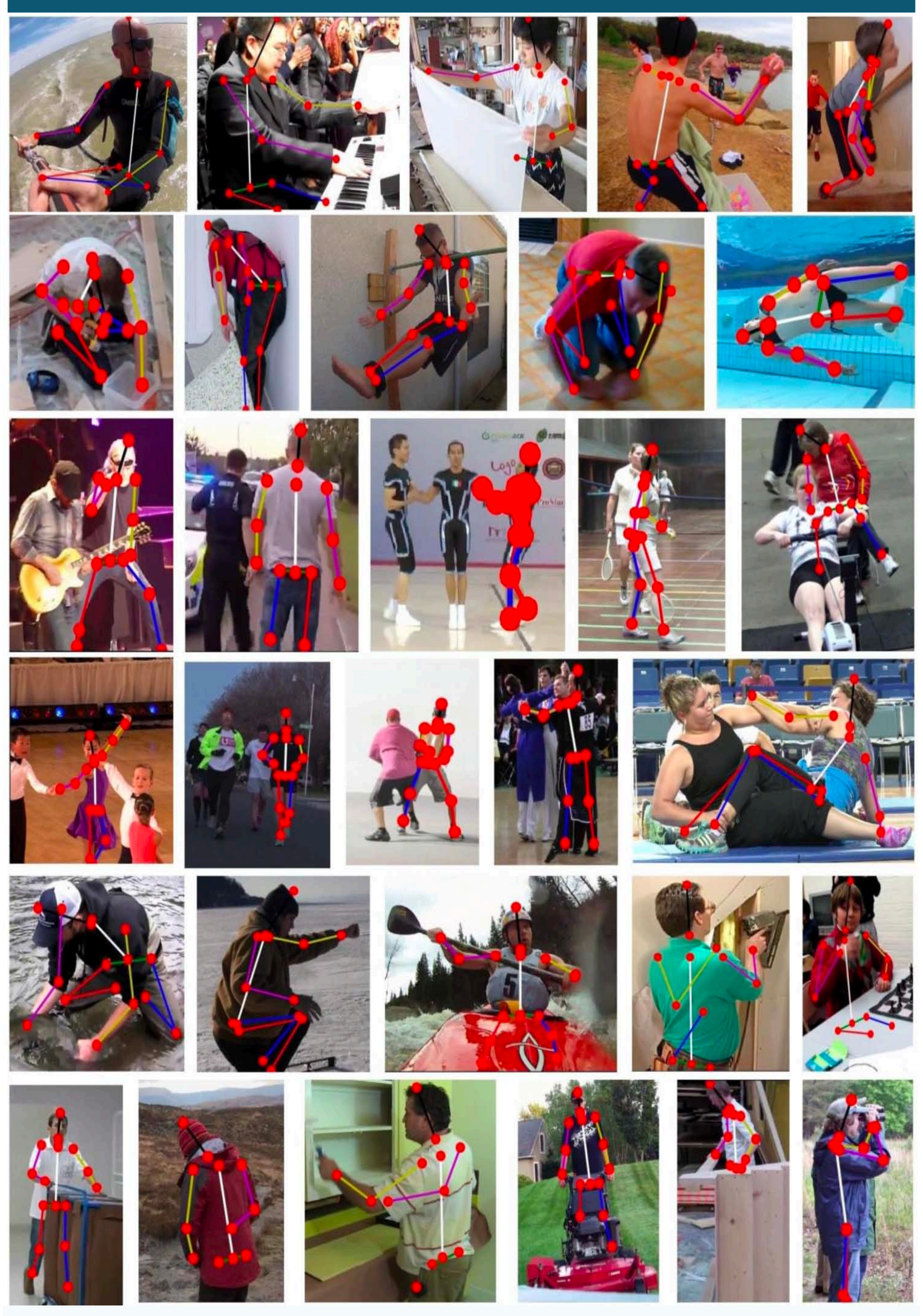

Fig. 6: Example of pose estimation results on the MPII dataset using our method. (row 1) Examples with significant scale variations for keypoints. (row 2,3) Examples with multiple persons. (row 4,5) Examples with severe keypoint occlusions.