Supplementary Material for TRB: A Novel Triplet Representation for Understanding 2D Human Body

Anonymous ICCV submission

Paper ID 3786

1. TRB annotating

Recall that in the definition of TRB, in each keypoint triplet, there exists one skeleton keypoint and two corresponding contour keypoints. The proposed contour keypoints are annotated on three pose estimation datasets: MPI[1], LSP[4] and COCO[6]. In this part, annotation performance analysis is conducted using 5000 redundantly annotated images in COCO-trb. Besides that, more annotated examples on three datasets are provided.

1.1. Annotation performance Analysis

We introduce contour points as a new type of 2D human keypoints, they are defined on each side of the corresponding skeleton keypoint, locate on human boundary. Does this definition carries clear enough semantic meaning? To validate this, 5000 images were redundantly annotated by two groups of annotators. During annotation, skeleton is not provided as additional guidance. Following [10], to measure the clarity of contour keypoint i, we use standard deviation σ_i with respect to object scale s as the metric, which can be written as:

$$\sigma_i^2 = E[d_i^2/area] \tag{1}$$

In which d_i denotes the euclidean distance of the same contour keypoint annotated by different annotators, and $area$ is the size of ground-truth human bounding box, which represents the object scale. The average results on 5000 images was used as an approximate of the expectation. Please refer to Table 1 and Figure 1.

Comparing to the same metric for skeleton keypoints provided in [10]. Some characteristic of contour keypoints can be concluded: 1. The definition of contour keypoints is as clear as skeleton keypoints for labeling. 2. Medial contour points are less ambiguous than lateral contour points during labeling, thanks to the strong visual evidence on human medial boundary.

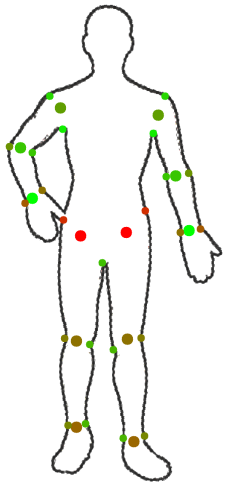

Figure 1. σ_i **of different keypoints.** Red points denote larger sigma and green points denote smaller sigma.

Table 1. σ_i of different keypoints

	Sho.	Elb.	Wri.	Hip	Knee	Ank.
Skeleton	0.079	0.072	0.062	0.107	0.087	0.089
Medial contour	0.065	0.072	0.087	0.073	0.075	0.075
Lateral contour	0.069	0.081	0.091	0.099	0.085	0.083

1.2. Samples of annotated images

TRB annotations on three datasets are visualized in Figure 2. The MPII dataset covers 410 different activities in daily life, which includes around 25K images containing over 40K people. LSP and extended LSPET contain 12K images. MPII has a relevant lower error tolerance than others for its evaluation metric. LSP contains more complex and rare poses with relevant low resolution. Most images in LSP and MPII contains whole body or whole upper body of the target person. In COCO, there exists pose instances in which only one small part is visible, like one leg or one arm. COCO is a more in-the-wild dataset and is considered as one of the most challenging 2D human pose benchmark.

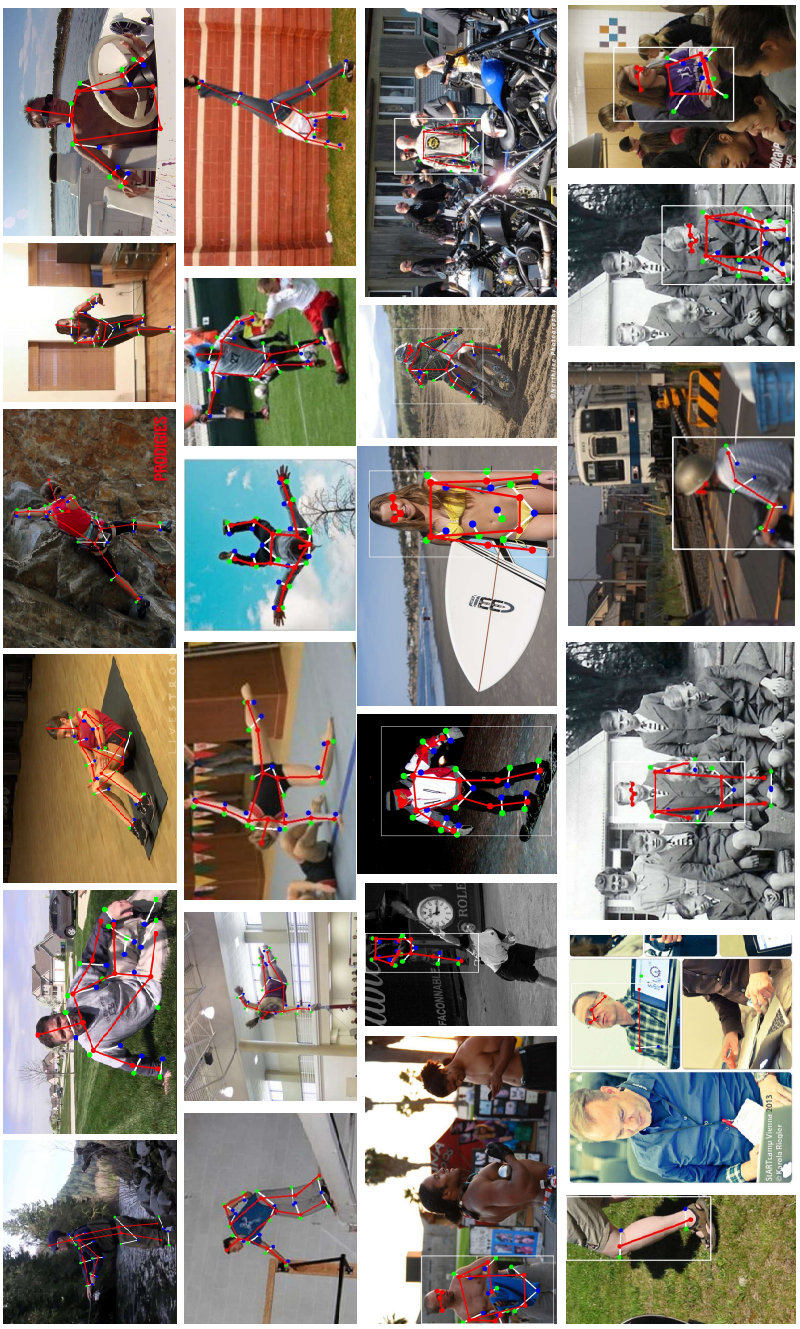

Figure 2. **Labeled examples from three datasets.** The first row displays images in MPII-trb dataset. The second row displays images in LSP-trb dataset. Last two rows display images in COCO-trb dataset, in which target person is denoted using white bounding boxes. Following images in the main paper, skeleton keypoints, medial contour keypoints, lateral contour points are red, blue and green respectively. (Best viewed in 4x)

2. Message Passing in TRB-Net

2.1. Formulation

Let $\mathbf{I}$ be an image, $\mathbf{s} = \{s_1, s_2, \cdots, s_N\}$ be the locations of N skeleton key points, $\mathbf{c} = \{c_1, c_2, \cdots, c_M\}$ be the locations of M contour key points. We could model the TRB estimation problem as an inferring process to conditional probability $p(\mathbf{s}, \mathbf{c}|\mathbf{I}, \Theta)$ parameterized by Θ, and relies on a Gibbs distribution:

$$\begin{aligned} p(\mathbf{s}, \mathbf{c}|\mathbf{I}, \Theta) &= \frac{e^{-E_n(\mathbf{s},\mathbf{c},\mathbf{I},\Theta)}}{S} \\ &= \frac{e^{-E_n(\mathbf{s},\mathbf{c},\mathbf{I},\Theta)}}{\sum_{\mathbf{s}\in S,\mathbf{c}\in C} e^{-E_n(\mathbf{s},\mathbf{c},\mathbf{I},\Theta)}} \end{aligned} \tag{2}$$

where $E_n(\mathbf{s}, \mathbf{c}, \mathbf{I}, \Theta)$ denotes the energy function. According to the previous work [2], the process could be implemented with neural networks.[1] According to the main paper, the probability can be re-formulated as:

$$\begin{aligned} p(\mathbf{s}, \mathbf{c}|\mathbf{I}, \Theta) &= \sum_{h_s}\sum_{h_c} p(\mathbf{s}, \mathbf{c}, h_s, h_c|\mathbf{I}, \Theta) \\ &= \frac{e^{-E_n(\mathbf{s},\mathbf{c},h_s,h_c,\mathbf{I},\Theta)}}{\sum e^{-E_n(\mathbf{s},\mathbf{c},h_s,h_c,\mathbf{I},\Theta)}} \end{aligned} \tag{3}$$

where

$$\begin{aligned} E_n(\mathbf{s}, \mathbf{c}, h_s, h_c, \mathbf{I}, \Theta) &= \sum_{(i,j)\in\epsilon_s} \delta_s[\psi_s(s_i, s_j)] \\ &+ \sum_{(p,q)\in\epsilon_c} \delta_c[\psi_c(c_p, c_q)] + \sum_{(i,p)\in\epsilon_{sc}} \delta_{sc}[\psi_{sc}(s_i, c_p)] \\ &+ \sum_{(i,k)\in\epsilon_{sh^s}} \psi_{sh^s}(s_i, h_k^s) + \sum_{(p,k)\in\epsilon_{ch^c}} \psi_{ch^c}(c_p, h_k^c) \\ &+ \sum_k \Phi(h_k^s, h_k^c) + \sum_k \gamma(h_k, I) \end{aligned}$$

In practice, we implement the terms in above Equation with the modules we proposed under the framework of CNNs. Specifically, $\sum_{(i,j)\in\epsilon_s} \delta_s[\psi_s(s_i, s_j)]$, $\sum_{(p,q)\in\epsilon_c} \delta_c[\psi_c(c_p, c_q)]$, $\sum_{(i,p)\in\epsilon_{sc}} \delta_{sc}[\psi_{sc}(s_i, c_p)]$ are corresponding to the *Triplet Constraints* we defined, the three terms respectively represents three types of important pairs of landmarks: skeleton-skeleton pair, contour-contour pair and skeleton-contour pairs. We implement these three terms using pairwise mapping. $\sum_{(i,k)\in\epsilon_{sh^s}} \psi_{sh^s}(s_i, h_k^s)$, $\sum_{(p,k)\in\epsilon_{ch^c}} \psi_{ch^c}(c_p, h_k^c)$ are two terms implemented by the multi-task network. $\Phi(h_k^s, h_k^c)$ is implemented by X-structure and direction convolution.

To present the difference of our message passing process with other models directly, we visualize the message propagation of different models in Figure 3 and message passing instance of our model in Figure 4.

[1] Please refer to the reference [2] for more theoretical derivation.

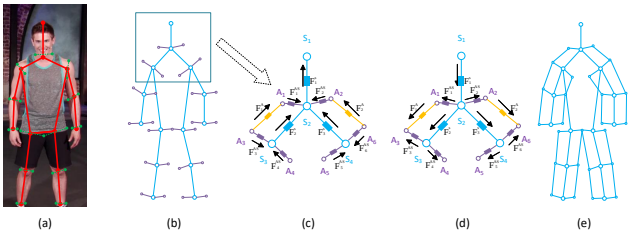

Figure 4. **Message Passing.** (a) is a person with annotated landmarks. (b) is the tree structured model of human with additional contour landmarks. (c,d) shows message passing on part of the graph in (b) with different directions. (e) demonstrates all possible message passing routes among all body landmarks.

2.2. Details of Warping in Pairwise Mapping

Here, we detail the warping operation mentioned in Sec.4 of the main paper. Using W to denote the estimated warping, H to denote the heatmap before warping, H' to denote the heatmap after warping. The forwarding can be formulated as

$$H'_{<i,j>} = H_{<i,j>+W_{<i,j>}} \tag{4}$$

$W_{<i,j>}$ denotes the corresponding value of warping to pixel $< i, j >$ on heatmap after warping. Assume its value is $< \Delta i, \Delta j >$, so that we have:

$$H'_{<i,j>} = H_{<i+\Delta i, j+\Delta j>} \tag{5}$$

However, $< i + \Delta i, j + \Delta j >$ may not be a coordinate located on the integer grid points, which brings difficulty to backward gradients. To issue the problem, we use linear interpolation to get an approximate value on this point. For convenience, we divide $< i + \Delta i, j + \Delta j >$ into integer part $< x, y >$ and decimal part $< \dot{x}, \dot{y} >$:

$$< i + \Delta i, j + \Delta j > = < x, y > + < \dot{x}, \dot{y} > \tag{6}$$

Hence we have:

$$\begin{aligned} H'_{<i,j>} =& (1-\dot{x})(1-\dot{y})H_{<\lfloor x\rfloor,\lfloor y\rfloor>} \\ &+ \dot{x}(1-\dot{y})H_{<\lfloor x\rfloor+1,\lfloor y\rfloor>} \\ &+ (1-\dot{x})\dot{y}H_{<\lfloor x\rfloor,\lfloor y\rfloor+1>} \\ &+ \dot{x}\dot{y}H_{<\lfloor x\rfloor+1,\lfloor y\rfloor+1>}. \end{aligned} \tag{7}$$

During backward pass, the gradient of one pixel on the warped heatmap will flow to both original heatmaps and the estimated warping. Examples are demonstrated below($W_{<i,j>}(0)$ and $W_{<i,j>}(1)$ denotes respectively x-component and y-component of estimated warping):

$$\begin{aligned} \frac{\partial H'_{<i,j>}}{\partial H_{<\lfloor x\rfloor,\lfloor y\rfloor>}} =& (1-\dot{x})(1-\dot{y}) \\ \frac{\partial H'_{<i,j>}}{\partial W_{<i,j>}(0)} =& (1-\dot{y})(H_{<\lfloor x\rfloor+1,\lfloor y\rfloor>} - H_{<\lfloor x\rfloor,\lfloor y\rfloor>}) \\ &+ \dot{y}(H_{<\lfloor x\rfloor+1,\lfloor y\rfloor+1>} - H_{<\lfloor x\rfloor,\lfloor y\rfloor+1>}) \end{aligned} \tag{8}$$

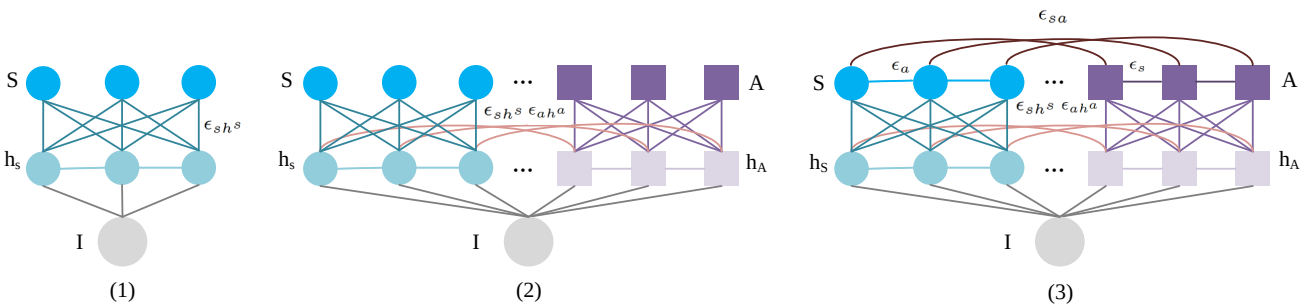

Figure 3. **Simplified message propagation models.** (1) only includes outputs and hidden states for skeleton keypoints. (2) includes outputs and hidden states for both skeleton keypoints and contour keypoints. (3) further adds *Triplet Constraints* beyond (2).

3. TRB Estimation

3.1. Training Details

In our experiments on LSP and MPII, SGD is used as our optimizing algorithm. We set momentum to be 0.9 and weight decay to be 1e-4. During training, the batch size is 32, while the whole training includes 250000 iterations. We use 8e-5 as initial learning rate, and decrease it by 2/3 at 160000 iteration and 200000 iteration. In our experiments on COCO, we use RMSprop as optimizing algorithm, with initial learning rate 5e-4. Our batch size is 96, the whole training includes 200000 iterations. Learning rate is decreased by 90% at 120000 iteration and 160000 iteration.

3.2. Additional Results for TRB Estimation

Table 2. Additional results on MPII_trb.

	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Ske.	Con.	Mean
Hourglass[9]	96.3	94.5	87.9	82.3	87.1	82.3	79.5	87.7	85.0	86.0
Simple Baseline[11] Res-50	96.3	94.9	87.5	82.2	87.1	83.9	80.3	88.0	84.8	85.9
Simple Baseline[11] Res-152	96.5	95.3	88.6	82.8	88.3	85.0	81.5	88.8	86.0	87.0
HRNet-W32 [5]	96.0	94.8	90.1	85.6	85.6	82.6	80.6	88.5	85.5	86.6
HRNet-W48 [5]	96.7	95.3	90.3	86.3	89.3	83.7	81.6	89.6	87.2	88.1
Cascaded AIOI [7]	96.7	95.0	88.4	82.9	87.7	83.5	80.0	88.3	85.5	86.5
TRB-Net	97.4	95.4	89.5	85.1	89.2	85.9	81.8	89.6	86.5	87.6

Deep high-resolution network[5], which is the latest state-of-the-art 2D skeleton keypoints estimation method was included into our comparison. Based on imagenet pretraining and high resolution feature learning, HRNet achieves good results on TRB estimation, which surpasses baseline used in this paper[7] a lot. In future, we will combine the proposed knowledge transfer scheme with up to date feature learning approach, to further improve current state-of-the-art of TRB estimation.

3.3. Skeleton Results on COCO test-dev

We further test the performance of the 2-stack hourglass based TRB-Net on skeleton estimation on COCO test-dev. The results are reported in Fig.5. The knowledge transferring scheme in TRB-Net boost skeleton performance by 2.4 AP with half training data and by 1.9 AP with full training data. Note that additional dataset and multi-scale testing were not used to produce this result, the best result of hourglass based model under the same conditions on COCO leader-board was 69.8 AP(reported by Raven-DL, with a 4-stack hourglass network). Under our finely tuned baseline, knowledge transfer modules in TRB-Net led to large improvement.

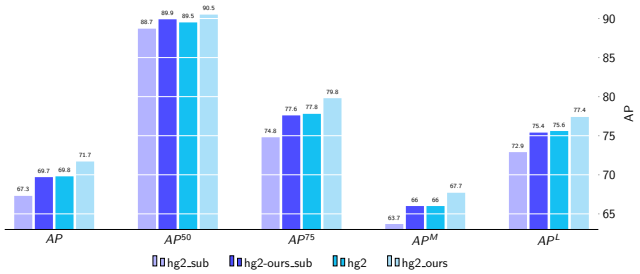

Figure 5. **Results on COCO test-dev.** 'sub' denotes using only half of the data for training. The results are obtained with single-scale testing and flipping.

3.4. Qualitative Results

We display some qualitative results on MPII-trb validation set in Figure 6. Based on three knowledge transfer module we proposed in TRB-Net, the semantic information of skeleton and visual evidence of contour are combined efficiently to benefit both tasks.

4. Human shape guided image generation

For experiment on DeepFashion[8], a 2-stack hourglass based TRB-Net trained on COCO was used to give out TRB prediction for all images. Then, following the same experiment setting in [3], we train a variational u-net for human shape guided image generation. Comparing to the original work, contour points in TRB are used as guidance additionally. We generate several demo videos for our proposed application — human shape editing. Frames of some videos are displayed in Figure. 7.

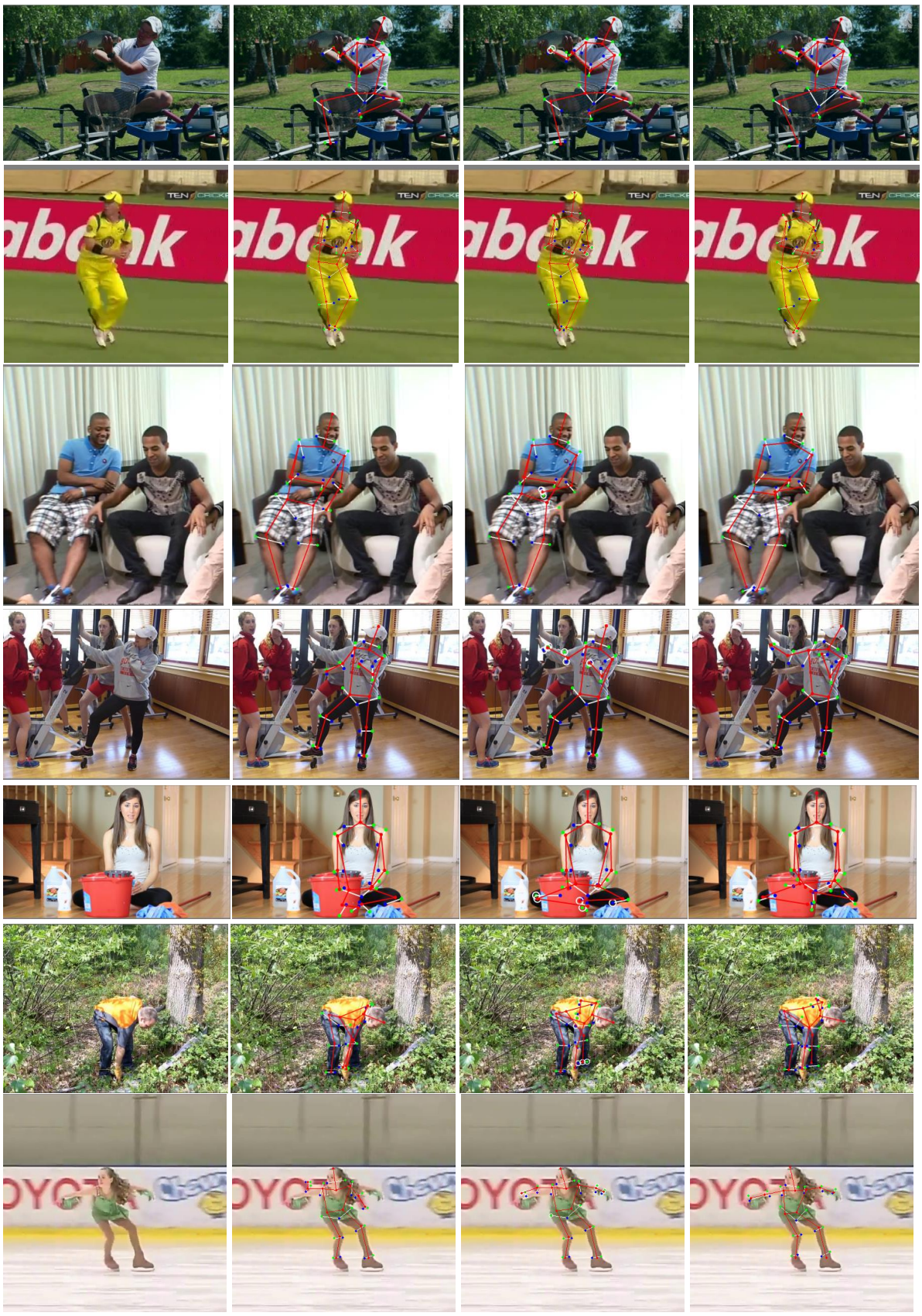

Figure 6. **Qualitative Results on MPII-trb validation.** The four columns denote original images, baseline prediction, TRB-Net prediction and ground-truth respectively. Wrong predictions corrected by TRB-Net are highlighted using white circles.(Best viewed in 4x)

Figure 7. **Frames in shape editing demo videos** Three different shape editing are performed to generate our demo videos. In the first column, upper leg contour points are edited to generate stronger legs. In the second column, upper body contour point are edited to generate stronger arms. In the third column, upper body contour point are edited to generate plump torsos.(Best viewed in 4x)

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1

[2] X. Chu, W. Ouyang, X. Wang, et al. Crf-cnn: Modeling structured information in human pose estimation. In *Advances in Neural Information Processing Systems*, pages 316–324, 2016. 3

[3] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 4

[4] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. 2010. 1

[5] D. L. Ke Sun, Bin Xiao and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*,

2019. 4

[6] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[7] W. Liu, J. Chen, C. Li, C. Qian, X. Chu, and X. Hu. A cascaded inception of inception network with attention modulated feature fusion for human pose estimation. In *AAAI*, 2018. 4

[8] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[9] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 4

[10] M. R. Ronchi and P. Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[11] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018. 4