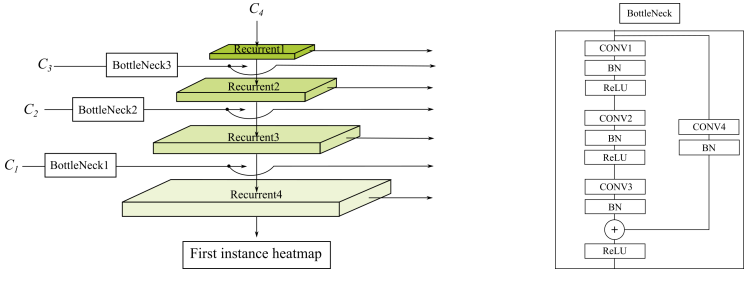
Supplementary Material

Serim Ryou
sryou@caltech.edu
Pietro Perona
perona@caltech.edu

Computational Vision Lab
California Institute of Technology
Pasadena, CA, USA

1 Person Decoder Architecture

Here we explain the detailed architecture of the person decoder.



(a) Person decoder (b) BottleNeck

Figure 1: **Person decoder.** We visualized the first time step of person decoder (a). To explain the details, we also showed the components of the bottleneck, which connects the part localizer to the person decoder.

Name	Input	Type	Spec	Output Size
BottleNeck1	C_1			64×64×128
.conv1		conv	1×1×256×64	
.conv2		conv	3×3×64×64	
.conv3		conv	1×1×64×128	
.conv4		conv	1×1×256×128	
BottleNeck2	C_2			32×32×256
.conv1		conv	1×1×512×128	
.conv2		conv	3×3×128×128	
.conv3		conv	1×1×128×256	
.conv4		conv	1×1×512×256	
BottleNeck3	C_3			16×16×512
.conv1		conv	1×1×1024×256	
.conv2		conv	3×3×256×256	
.conv3		conv	1×1×256×512	
.conv4		conv	1×1×1024×512	

Table 1: **BottleNeck specification**

© 2018. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Name	Input	Type	Spec	Output Size
Recurrent1	C_4	-	-	8×8×512
.convLSTM1		convLSTM	3×3×2048×512	
.convLSTM2		convLSTM	3×3×512×512	
Upsampling1	Recurrent1	upsampling	-	16×16×512
Concat1	Upsampling1, BottleNeck3	concat	-	16×16×1024
Recurrent2	Concat1	-	-	16×16×256
.convLSTM1		convLSTM	3×3×1024×256	
.convLSTM2		convLSTM	3×3×256×256	
Upsampling2	Recurrent2	upsampling	-	32×32×256
Concat2	Upsampling2, BottleNeck2	concat	-	32×32×512
Recurrent3	Concat2	-	-	32×32×128
.convLSTM1		convLSTM	3×3×512×128	
.convLSTM2		convLSTM	3×3×128×128	
Upsampling3	Recurrent3	upsampling	-	64×64×128
Concat3	Upsampling3, BottleNeck1	concat	-	64×64×256
Recurrent4	Concat3	-	-	64×64×64
.convLSTM1		convLSTM	3×3×256×64	
.convLSTM2		convLSTM	3×3×64×64	
Upsampling4	Recurrent3	upsampling	-	128×128×64
Output	Upsampling4	-	-	128×128×K
.conv1		conv	3×3×64×64	
.ReLU		ReLU	-	
.conv2		conv	3×3×64×K	
.ReLU		ReLU	-	
.conv3		conv	1×1×K×K	
.sigmoid		sigmoid	-	

Table 2: **Person decoder specification**

2 Qualitative Results

Here we show the qualitative results compared to the results of Mask-RCNN, and also provide some failure cases. As in Figure 2, Mask-RCNN frequently merges two instance into a single one when people are highly overlapped. Our method is robust at predicting distinct instance skeletons in crowd scenes.

Furthermore, we have collected a set of failure cases. Typical failure scenarios were 1) *Images with more than 5 people,* 2) *Assigning missing keypoints to a wrong person,* 3) *Cluttered scenes.* In order to show how the recurrent architecture influences the performance of the images with many people (more than 5), we show the step-by-step inference result that our network produces in Figure 3. Also, we visualize other failure cases in Figure 4.

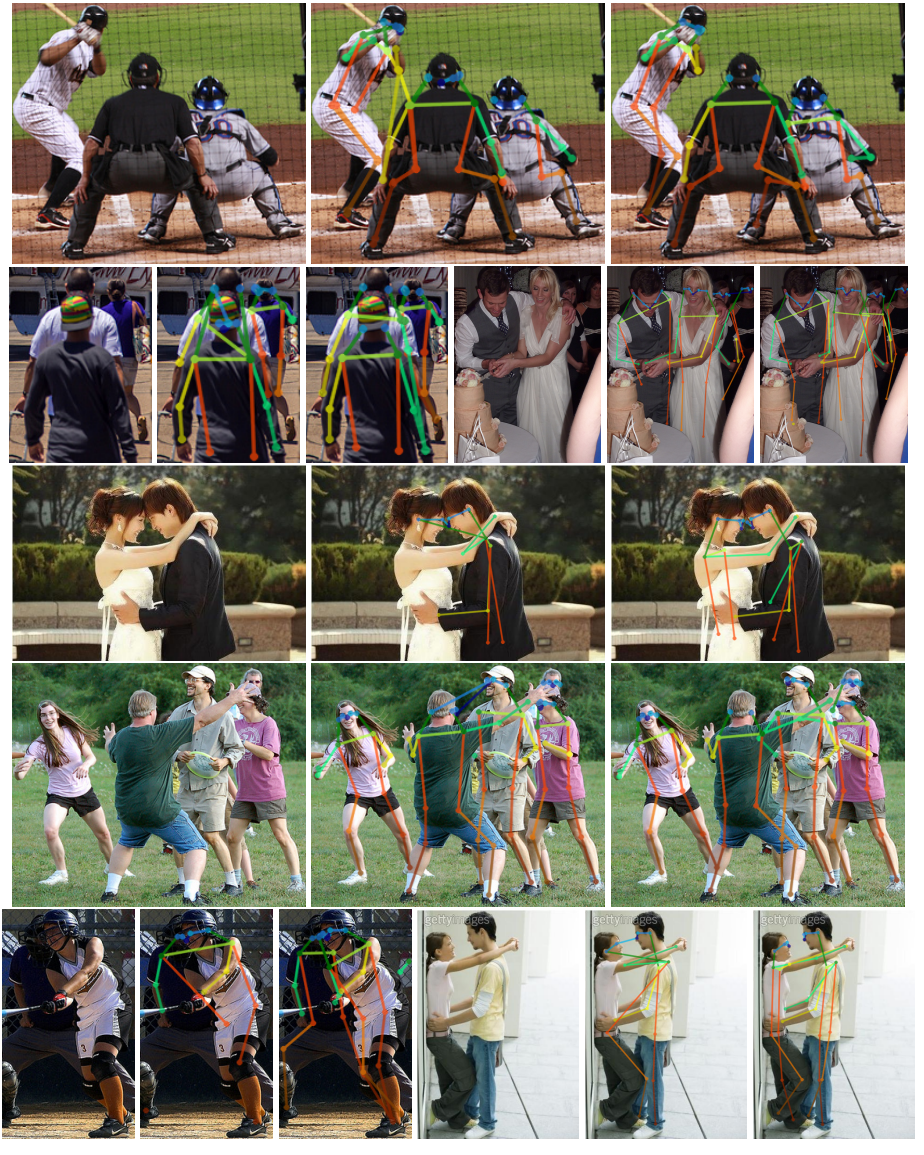


Figure 2: **Qualitative results compared to Mask-RCNN.** The images are ordered by input, result of Mask-RCNN, and ours, for each set of three images. Our method faithfully produces distinct keypoint skeletons in crowd scenes.

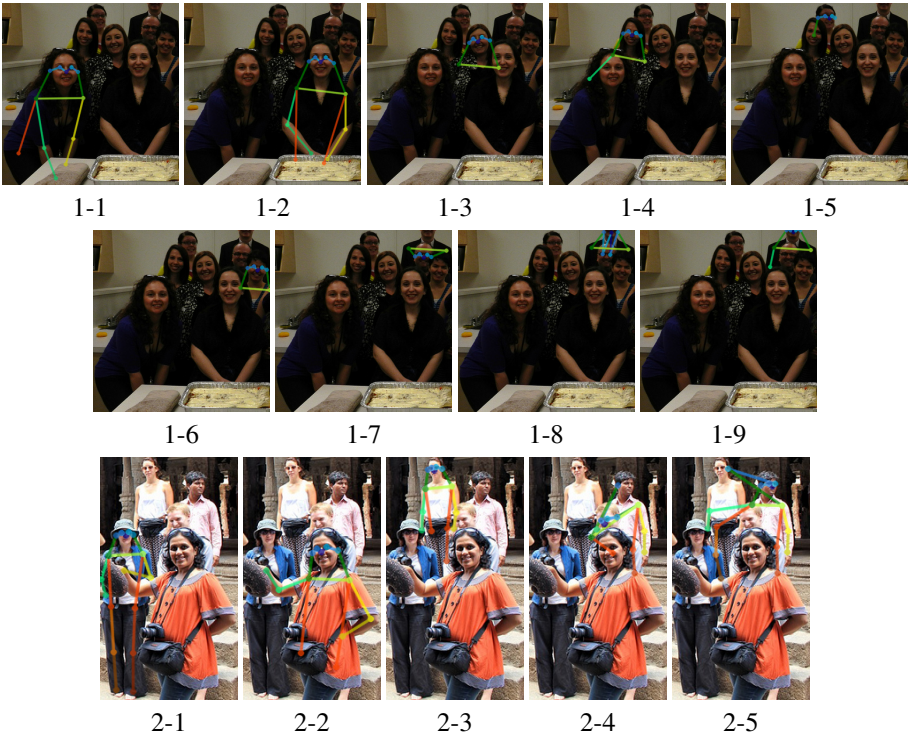
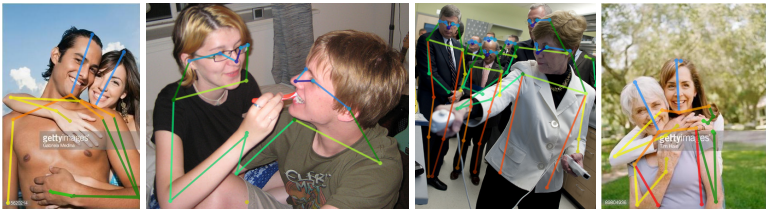


Figure 3: **Failure cases on images with more than 5 people.** Due to the recurrent architecture of our method, the network fails to predict the following sequences after it makes a mistake. 1-7, 1-8, 2-4, and 2-5 are the failure examples.



(a) Assigning missing keypoints to a wrong person

(b) Cluttered scenes

Figure 4: **Failure cases.** When the keypoint is occluded or missing, the network predicts part locations of other person (a). In cluttered scenes, it fails to assign correct joint identity to each individual or produces wrong detections.