

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

Improving 2D Human Pose Estimation across Unseen Camera Views with Synthetic Data - Supplementary material

Anonymous WACV Algorithms Track submission

Paper ID 442

1. RePo Dataset

Here we describe a new RePo dataset of manually annotated real images. The dataset focuses on extreme poses in top and bottom views typically encountered in sports. Images come primarily from public sports videos on YouTube. The dataset is split into two parts - one focusing on the bottom view with 187 images, the other focusing on the top view with 91 images. Each part is divided into sets described in Tab. 1 in the paper.

One professional annotator annotated the whole dataset. We created a custom annotation environment allowing for an easier understanding of the scene necessary for annotating extreme views. Since the visibility of the joints is not defined in detail in the COCO dataset [1], we defined it as follows:

Visibility 0. The keypoint is not visible, and we cannot reliably tell its precise location. The words *reliably* and *precise* are crucial in situations where a keypoint is not visible. We can guess its location from the context but cannot be sure the guess was correct.

Visibility 1. The keypoint is not visible, but we can reliably tell its precise location from the context and other keypoints.

Visibility 2. The keypoint is visible in the image.

Further, as the extreme views pose additional challenges, we stick to the annotation of joint projection to the image plane if it is unambiguous. A typical example would be the ankle which is rarely visible from the bottom view (we see the heel instead). Without this relaxation of definition, almost no keypoints would be annotated as visible (visibility 2).

Examples of images from all sets of the dataset are in the Fig. 10, Fig. 11 and Fig. 12.

2. RePoGen Dataset

The RePoGen dataset is created with the proposed RePoGen method and was used to train the best-performing model. There are 3 variants of the RePoGen dataset, all

parameter	value
number of poses	1 500
number of views per pose	2
pose simplicity	uniform between (1.0, 1.5)
with texture	✓
with background	✓
default SMP pose	✗
uniform distribution	✗

Table 6. Parameters used for RePoGen dataset generation.

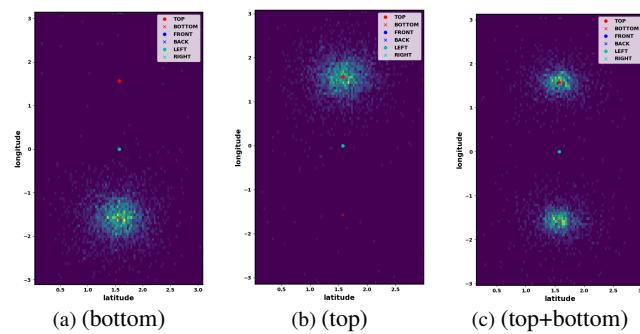


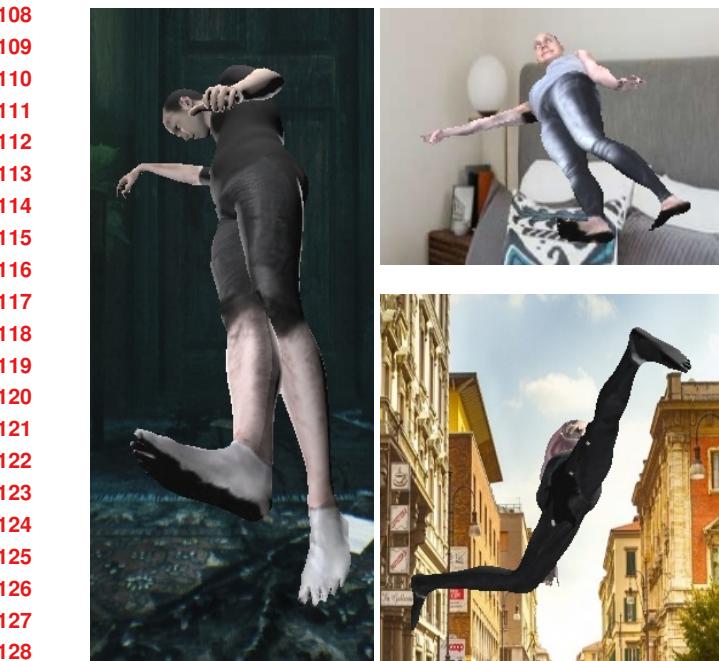
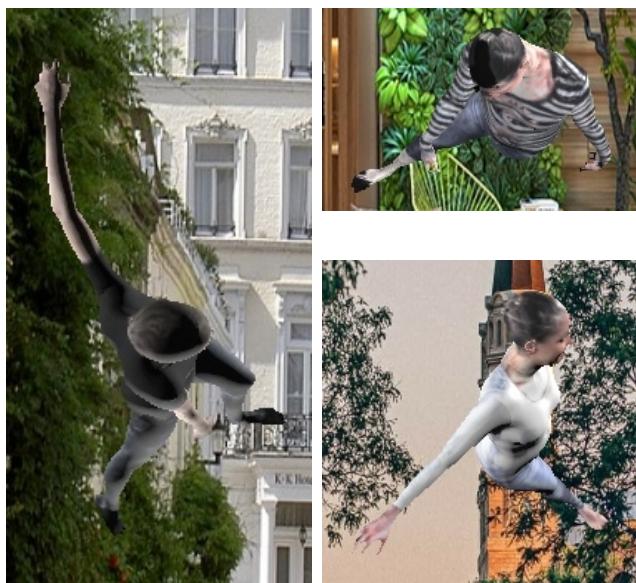
Figure 7. Camera viewpoint distributions for various RePoGen datasets.

meeting the description in Tab. 6. The parameter distinguishing the 3 variants is the camera viewpoints distribution. The RePoGen (bottom) and RePoGen (top) are sampled with a normal distribution centered around the bottom view and top view respectively. The RePoGen (top+bottom) is sampled from a combination of these two distributions. Visualization is in the Fig. 7.

The Fig. 8 and Fig. 9 contain images from the RePoGen dataset.

3. Additional results

We also offer additional results on the proposed RePo dataset. The RePo Bottom Test is in the Fig. 10, and Val set in Fig. 11.

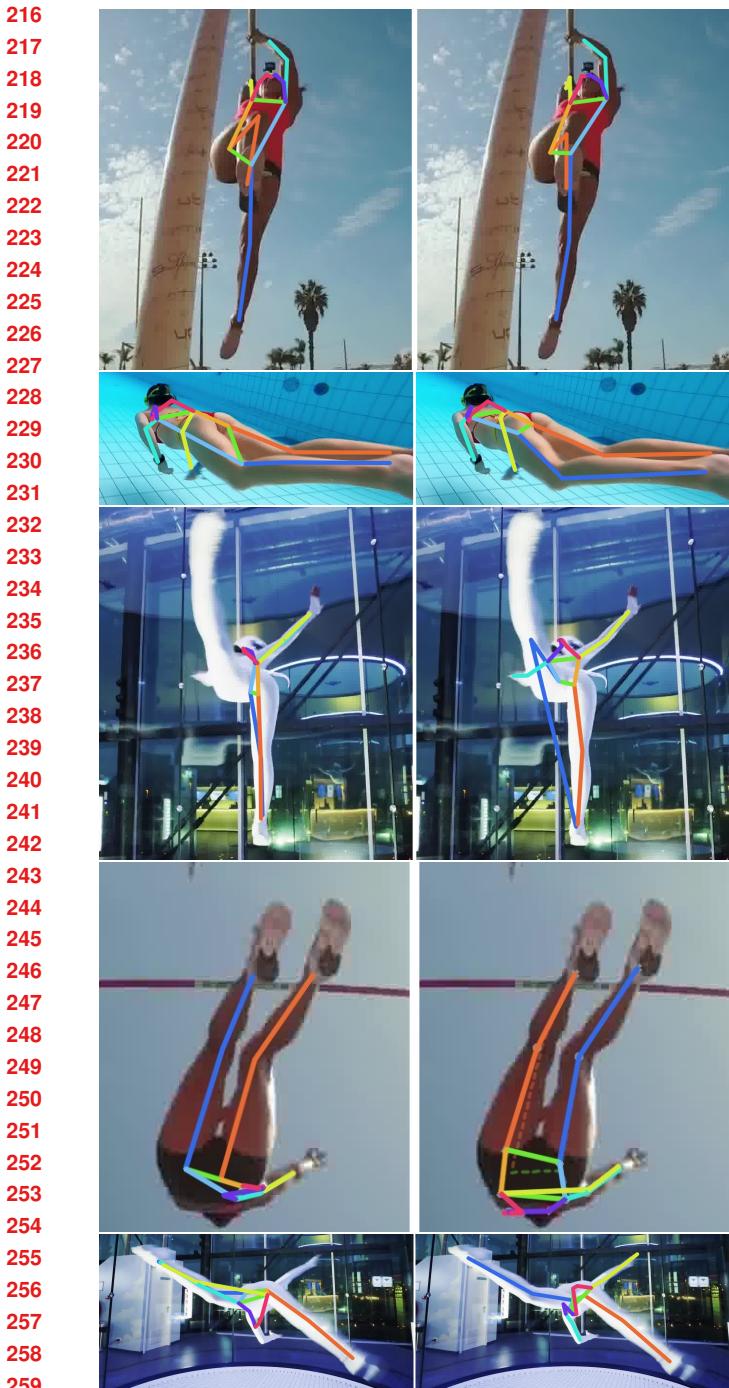
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
Figure 8. Images from the RePoGen (bottom) dataset.153
154
155
156
157
158
159
160
161
Figure 9. Images from the RePoGen (top) dataset.

The RePoGen struggles the most with head keypoints and strong motion blur. The Fig. 12 shows results on the Bottom Seq set where we show performance on a video. We show every third frame from a video.

4. Code

The code is available in the supplementary material. The repository builds on the SMPL-X project [2] and uses the same depen-

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
Figure 10. Examples from the RePo bottom test set. ViTPose-S estimates when trained on COCO (left) and on RePoGen data (right). Colors – right hand, right leg, left hand and left leg



260
261
262
263
264
265
266
267
268
269

Figure 11. Examples from the RePo bottom val set. ViTPose-s estimates when trained on COCO (left) and on RePoGen data (right). Colors – right hand , right leg , left hand and left leg

dencies. For more details, see the enclosed README.md file and the code itself.



Figure 12. Examples from the RePo bottom seq set. ViTPose-s estimates when trained on COCO (left) and on RePoGen data (right). Images from a consecutive sequence, taking every third frame. Colors – right hand , right leg , left hand and left leg

324	References	378
325		379
326	[1] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In <i>Eu- ropean Conference on Computer Vision</i> , 2014. 1	380
327		381
328		382
329		383
330	[2] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In <i>Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)</i> , 2019. 2	384
331		385
332		386
333		387
334		388
335		389
336		390
337		391
338		392
339		393
340		394
341		395
342		396
343		397
344		398
345		399
346		400
347		401
348		402
349		403
350		404
351		405
352		406
353		407
354		408
355		409
356		410
357		411
358		412
359		413
360		414
361		415
362		416
363		417
364		418
365		419
366		420
367		421
368		422
369		423
370		424
371		425
372		426
373		427
374		428
375		429
376		430
377		431