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).

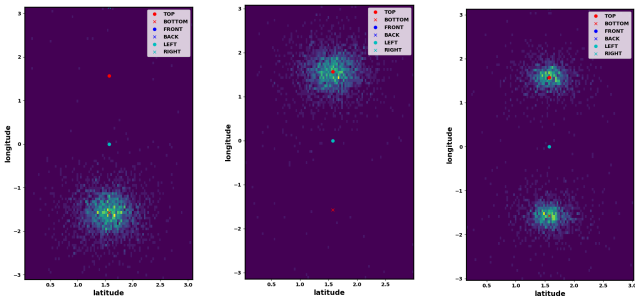
Examples of images from all sets of the dataset are in the Fig. 10, Fig. 11 and Fig. 12.

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 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.

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.



(a) (bottom) (b) (top) (c) (top+bottom)

Figure 7. Camera viewpoint distributions for various RePoGen datasets.

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.

Figure 8. Images from the RePoGen (bottom) dataset.

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-

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

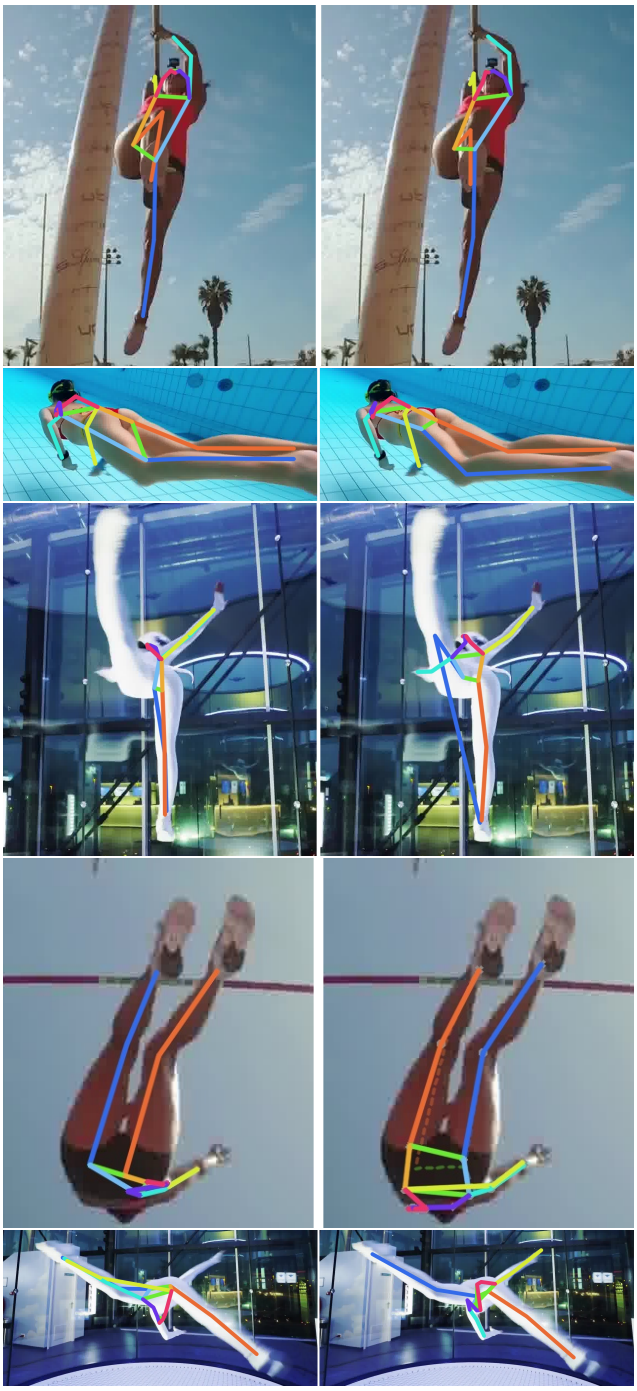


Figure 11. Examples from the RePo bottom val set. ViTPose-s estimates when trained on COCO (left) and on RePoGen data (right). Colors – right hand, right leg, left hand and left leg

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

References

[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 *European Conference on Computer Vision*, 2014. 1

[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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2