Supplementary Material for: Self-Supervised Multi-View Synchronization Learning for 3D Pose Estimation

Simon Jenni[0000−0002−9472−0425] and Paolo Favaro[0000−0003−3546−8247]

University of Bern, Switzerland
{simon.jenni,paolo.favaro}@inf.unibe.ch

1 Background Removal

Since background removal was shown to be important for shortcut prevention, we provide additional implementation details and analysis.

Implementation Details. We extract separate backgrounds for each subject and camera combination, resulting in a total of 7×4 background images (we observed small camera motion between different subjects). Background images are computed as the median image over the frame sequence corresponding to the action 'Walking-1'. To perform background removal, we compute the foreground mask by thresholding the absolute difference between the training example $x_\nu^{(i)}$ and the corresponding background image b_ν. In practice, the mask is given by pixels satisfying $sum(|x_\nu^{(i)} - b_\nu|, axis = -1) > 32$, where the sum is over the channel dimension. For background substitution we use this mask to combine the image $x_\nu^{(i)}$ with a randomly chosen background via alpha matting.
Qualitative Examples. We show a set of training images after background removal in Fig. 1. While we can observe some failure cases due to shadows or other lighting variations where backgrounds remain visible, the method works well enough to remove background cues while preserving the relevant foreground information.

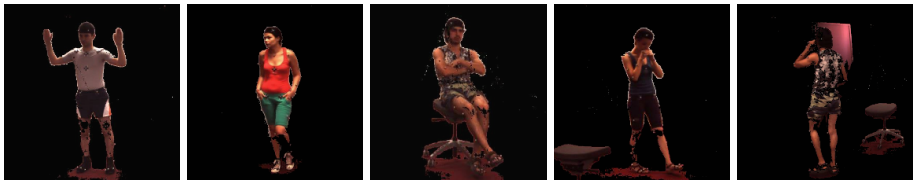

Fig. 1: **Examples of background removal.** We show some training images after applying our background removal. Background cues that could provide shortcuts to the self-supervised learning task are removed and image regions relevant to pose estimation are preserved.

Table 1: **Background removal sensitivity.** We compare the performance of a model trained with our background removal to a model trained using the ground-truth foreground masks provided with the dataset.

Method	MPJPE	NMPJPE	PMPJPE
Ground-truth mask	109.1	97.1	77.4
Ours	110.8	96.7	76.9

Table 2: **2D pose estimation experiments.** We compare our self-supervised pre-training to a randomly initialized baseline on 2D human pose estimation. Only annotations of training subject S1 are used.

Method	MPJPE
Baseline (random init.)	77.8
Ours (full fine-tuning)	53.6

Quantitative Evaluation. We evaluate the influence of the background removal quality on the performance of the model. We compare a model trained with our background removal to a model trained using the foreground masks provided with the Human3.6M data set. The comparison is shown in Table 1. We observe no significant difference in the models performance.

2 Evaluation of 2D Pose Estimation

To examine the influence of our pre-training strategy on other downstream tasks of interest we also transfer to 2D human pose estimation on Human3.6M. In this experiment we simplify the pre-processing considerably by only using random cropping (without centering on the subject) and without removing static backgrounds. The 2D pose annotations are provided in pixel coordinates of the unprocessed 1000×1000 resolution frames. The training is otherwise identical to our ablation experiment setup. We compare a model trained using our synchronization pretext task to a randomly initialized baseline in Table 2.