Multiview-Consistent Semi-Supervised Learning for 3D Human Pose Estimation (Supplementary)

1. Selection of Negatives

In this section, we elaborate on the process of creating a batch of *anchors* and *positves* such that the in-batch hard-mined *negatives* are not too similar to their corresponding *anchors*/*positives* as that can make the learning unstable. To this end, we group consecutive frames (already down-sampled temporally by factor of 5) into chunks of 4. At the time of creating a batch, we ensure *anchors*/*positives* are selected from different random chunks. We also randomly select 1 out of the 4 elements within a chunk. On the rare occasions where poses from different chunks are too similar, the parameter β (mentioned in Sec. 3.1 in the main paper) enables our framework to avoid choosing them as hard *negatives*.

2. Ablation on *Margin* Values

Our metric learning framework requires two tunable parameters, the margin m and minimum embedding distance threshold for a hard mined *negative*, β. In Tab. 1, we report our canonical pose estimation for different values m and β. The combination of $m = 0.6$ and $\beta = 0.3$ provides the best performance in the test set. Choosing higher margins leads to instability in $\mathcal{L}_{cnstr}$ as *negatives* has to be separated by larger distances while positives/*negatives* from large viewpoint variations need to remain close.

Hyper-Parameters	N-MPJPE	MPJPE
$m = 0.4,\ \beta = 0.2$	120.9	126.0
$m = 0.6,\ \beta = 0.3$	**111.9**	**121.0**
$m = 0.8,\ \beta = 0.4$	118.1	128.5
$m = 1.0,\ \beta = 0.5$	124.3	133.8

Table 1: Performance of our model with different margin values. We observe best performance for $m = 0.6$ and $\beta = 0.3$.

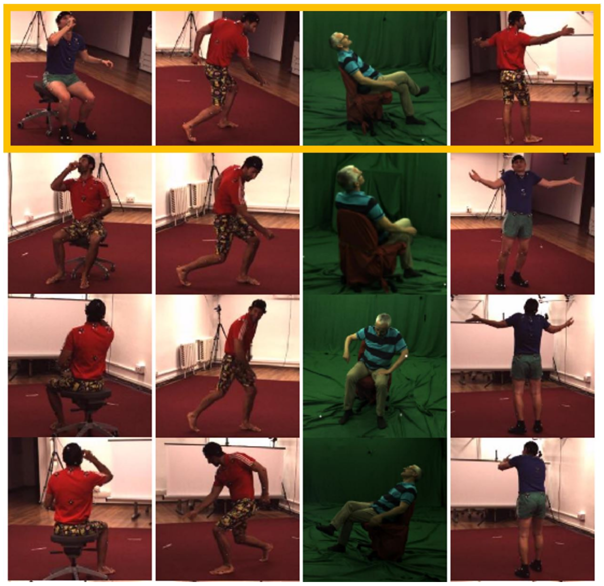


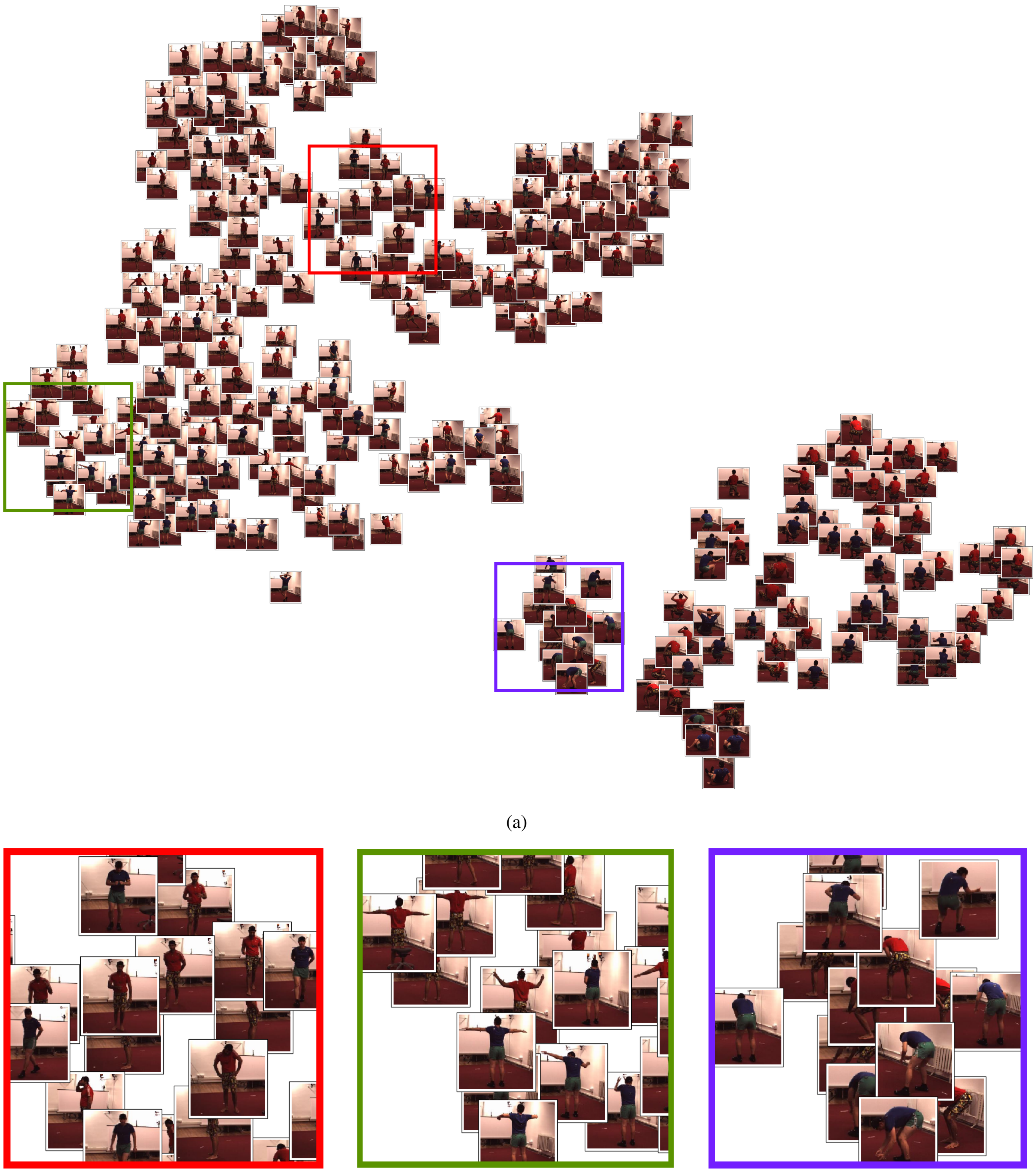
Figure 1: Qualitative image retrievals on Human 3.6M (S9, S11) and MPI-INF-3DHP (S7, S8) test sets. The first row represents query image and the rows below are the top 3 closest images in embedding space. For the left-most and right-most columns, the retrieval database is composed of images from different subject and viewpoint from that of query's. For the middle two columns, retrieval database is composed of images of same subject but different viewpoint from that of the query's. Note how the retrieved poses are very similar to query poses.

3. Qualitative Results For Pose Retrieval

In Fig. 1, we show qualitative image retrieval results based on embedding distance. We can clearly see that the closest images from other subjects and other viewpoints to the query image in embedding space share similar poses.

4. Visualisation of Embedding Space

In this section, we demonstrate the pose based clustering property of our embedding space by proving a 2D visualization of the same. In Fig. 2, we use the popular T-SNE [1]



(a)

(b)

Figure 2: (a) T-SNE plot of the our embedding space. The images are clustered according to the 3D human poses contain. (b) Zoomed view of three colored windows from Fig (a). Note: minor inconsistencies in the 2D visualisation is due to mapping the embedding from 128 dimension vector lying on the surface of a unit hyper sphere to 2D space.

dimension reduction method to map the embeddings to 2D. One cam observe images with similar poses in the world coordinate system are clustered together.

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. In *JMLR*, 2008. 1