SelfPose3d: Self-Supervised Multi-Person Multi-View 3d Pose Estimation

Supplementary Material

Video name	Number of persons	SelfPose3d				VoxelPose			
		AP_{25}	AP_{50}	AP_{100}	MPJPE	AP_{25}	AP_{50}	AP_{100}	MPJPE
160906_ian5	2	54.1	86.5	94.1	25.9	65.7	85.8	94.4	24.0
160422_haggling1	3	56.0	95.3	98.0	**23.8**	86.2	98.0	99.5	17.2
160906_band4	3	**58.6**	**98.9**	99.0	24.7	**98.1**	**99.6**	99.8	**15.4**
160906_pizza1	6	48.6	97.7	**99.7**	24.7	71.3	98.5	**99.9**	20.7
All videos	2-6	55.1	96.4	98.5	24.5	81.8	98.0	99.4	18.3

Table 8. Video-level test results on the Panoptic dataset.

7. Additional Experiments

7.1. Effect of number of persons

To evaluate the effect of different numbers of persons, we present the video-level results of SelfPose3d and VoxelPose on the Panoptic test set, as each video contains a different number of persons. As shown in Table 8, the variance of VoxelPose's performance is larger, and there is no strong correlation between the number of persons and the models' performance. We also observe that the occlusion is still the key factor because the video "160906_ian5" is of a kid playing with a woman, but he is heavily occluded due to his height, resulting in lower performance.

7.2. Cross-scene generalization

To test the cross-scene generalization ability of SelfPose3d, we compare it with fully-supervised VoxelPose [57] and MvP [63] in two directions.

From Panoptic to Campus/Shelf. In this part, SelfPose3d and VoxelPose are trained on the Panoptic dataset with 5 views, and then tested on the Campus and Shelf dataset without fine-tuning. For MvP, we use the provided best models. As shown in Table 9, SelfPose3d performs better than VoxelPose and MvP, showing better cross-scene generalization from a large dataset to a smaller dataset. The significant gap on the Campus dataset also shows that SelfPose3d is more robust to the number of camera views.

From Campus/Shelf to Panoptic. For SelfPose3d, we show the self-supervised learning result on the Panoptic dataset as it requires no 3D ground-truth labels. For VoxelPose and MvP, since they cannot be trained with Campus and Shelf datasets because of smaller dataset size and the noisy 3D ground-truth labels, we follow the original papers' training strategy, i.e., for VoxelPose, training using the synthetic Campus/Shelf dataset by randomly placing 3d poses of the Panoptic dataset in the Campus/Shelf 3D space; for MvP, using the provided MvP model, first trained on the Panoptic dataset and then fine-tuned on Shelf dataset. We test the above VoxelPose and MvP model, trained on the Campus/Shelf datasets, back on the Panoptic test set. As shown in Table 10, VoxelPose and MvP fail to detect any 3d pose, although they have used 3D ground-truth labels from the Panoptic dataset in the first place. In other words, they are severely overfitted on the camera poses of the training set. The experiment shows the ability of SelfPose3d to address large-scale unseen datasets.

Methods	Shelf (5 camera views)				Campus (3 camera views)			
	Actor 1	Actor 2	Actor 3	Average	Actor 1	Actor 2	Actor 3	Average
VoxelPose	99.5	93.5	97.8	96.9	93.1	86.5	93.2	90.9
VoxelPose*	94.6	91.4	97.5	94.5	0.0	0.3	0.0	0.1
MvP	99.3	95.1	97.8	97.4	98.2	94.1	97.4	96.6
MvP*	3.51	4.32	15.9	7.91	0.41	0.05	0.43	0.30
SelfPose3d	93.7	94.3	97.7	95.2	78.2	8.0	40.9	42.3

Table 9. Results (in PCP) on Shelf and Campus test set without fine-tuning. (1) SelfPose3d is trained on the Panoptic dataset without using GT labels. (2) VoxelPose and MvP are with fine-tuning, and VoxelPose* and MvP* are without fine-tuning.

Methods	AP_{25}	AP_{50}	AP_{100}	MPJPE
VoxelPose (Campus)	0.0	0.0	0.0	inf
VoxelPose (Shelf)	0.0	0.0	0.0	350.6
MvP (Shelf)	0.0	0.0	0.0	395.3
SelfPose3d	55.1	96.4	98.5	24.5

Table 10. Results on the Panoptic test set. (1) VoxelPose is trained on synthetic Campus/Shelf dataset. (2) MvP is firstly trained on the Panoptic dataset and then fine-tuned on Shelf dataset. (3) SelfPose3d is trained on the Panoptic dataset in a self-supervised way.

λ	AP_{25}	AP_{50}	AP_{100}	MPJPE
0.001	28.4	93.5	97.4	28.7
0.01	**33.6**	**95.1**	**97.7**	**27.7**
0.1	21.8	79.2	92.4	34.0
1.0	2.71	44.0	85.3	48.3

Table 11. Ablation study on λ in Eq. (11), where we train each model for 5 epochs without adding L_1 loss attention.

7.3. Ablation study on adding L_1 joint loss

As mentioned in Sec. 4.2, it is more likely to diverge when training the model using L_1 joint loss solely. However, based on the visualization of the output 3d poses in the training process (see Figure 5), we find that L_1 loss can help the model generate a human-shape pose much faster than L_2 loss in the early training stage. It is reasonable because L_1 loss provides a direct supervision on joint coordinates while L_2 loss doesn't. Thus we assume that L_1 loss is helpful for more precise prediction, and conduct an ablation study on merging it with L_2 loss in Table 11. Based on the results, we set λ in Eq. (11) to 0.01.

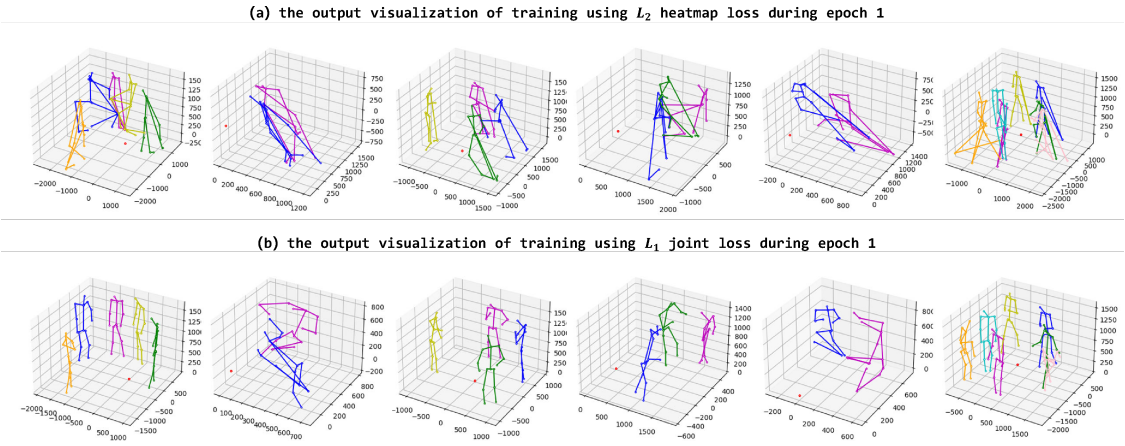

Figure 5. Comparing the visualization of the output 3d poses during epoch 1, using L_2 heatmap loss and L_1 joint loss respectively.

σ	AP_{25}	AP_{50}	AP_{100}	MPJPE
0.01	-	-	-	-
0.1	**36.6**	**95.1**	**97.9**	**26.6**
1.0	32.5	94.3	97.7	27.6

Table 12. Ablation study on σ in Eq. (11), where we train each model for 5 epochs using L_2 loss solely with ResNet-18 based $attn\_net_{2d}$.

Backbone	AP_{25}	AP_{50}	AP_{100}	MPJPE
ResNet-18	36.6	95.1	**97.9**	**26.6**
ResNet-34	**37.2**	**95.2**	97.7	26.9
ResNet-50	26.9	91.6	97.4	29.9
ResNet-50*	24.5	91.9	97.4	30.2

Table 13. Ablation study on the backbone network of $attn\_net_{2d}$, where we train each model for 5 epochs using L_2 loss solely with σ=0.1. * means shared backbone with $heatmap\_net_{2d}$.

7.4. Ablation study on L_2 loss attention

There are two aspects affecting the supervision attention for L_2 loss: the weight σ of l_{attn} in Eq. (11) and the backbone. We first use ResNet-18 as the backbone, and conduct experiments about σ in Table 12. When we set σ to 0.01, the model doesn't converge because the output of $\text{attn\_net}_{\text{2d}}$ is almost zero. Therefore, we set σ in Eq. (11) to 0.1.

Afterwards, we try to deepen the architecture of $\text{attn\_net}_{\text{2d}}$ backbone, and examine whether $\text{attn\_net}_{\text{2d}}$ and $\text{heatmap\_net}_{\text{2d}}$ can share weights. Table 13 shows that ResNet-18 is sufficient, and sharing weights degrades the performance.

7.5. Robustness of SelfPose3d

In order to test the robustness of our methods, we train SelfPose3d using fewer camera views of the Panoptic dataset. As shown in Table 14, the performance of SelfPose3d steadily reduces when we decrease the number of camera views to 3.

Methods	Views	AP_{25}	AP_{50}	AP_{100}	MPJPE
VoxelPose [57]	5	83.6	98.3	99.8	17.7
VoxelPose [57]	3	58.9	93.9	98.4	24.3
SelfPose3d (ours)	5	55.1	96.4	98.5	24.5
SelfPose3d (ours)	4	31.1	89.6	96.7	30.2
SelfPose3d (ours)	3	10.4	66.1	90.4	43.5

Table 14. Results on the Panoptic dataset with different number of camera views.

Method	*root_net* input	AP_{50}^{root}	AP_{100}^{root}	$MPJPE^{root}$
VoxelPose	all heatmaps	41.0	99.0	49.3
VoxelPose	root-heatmaps	34.0	99.0	50.0
SelfPose3d	root-heatmaps	35.2	92.3	54.9

Table 15. The rationale for using root-heatmaps as input to the *root_net* for 3d roots localization. Training VoxelPose model with only root-heatmaps obtains nearly the same performance. SelfPose3d trained using synthetic root-heatmaps with root consistency loss also reaches comparable performance. Here AP_{50}^{root}, AP_{100}^{root}, and $MPJPE^{root}$ are calculated only for the root joint.

7.6. Root localization with only root-heatmaps

We use the similar architecture compared to VoxelPose for our SelfPose3d approach. The only architectural change in the SelfPose3d w.r.t VoxelPose is using only the root-heatmaps as input to the root_net for root localization. This architectural change has enabled us to learn the root_net parameters from synthetic 3d roots. Table 15 shows the results for root localization using only the root-heatmaps *v.s* all the heatmaps for VoxelPose and SelfPose3d. We observed a minor decrease in the performance for both the approaches, confirming our hypothesis that using only 2d root-heatmaps is sufficient for 3d root localization.

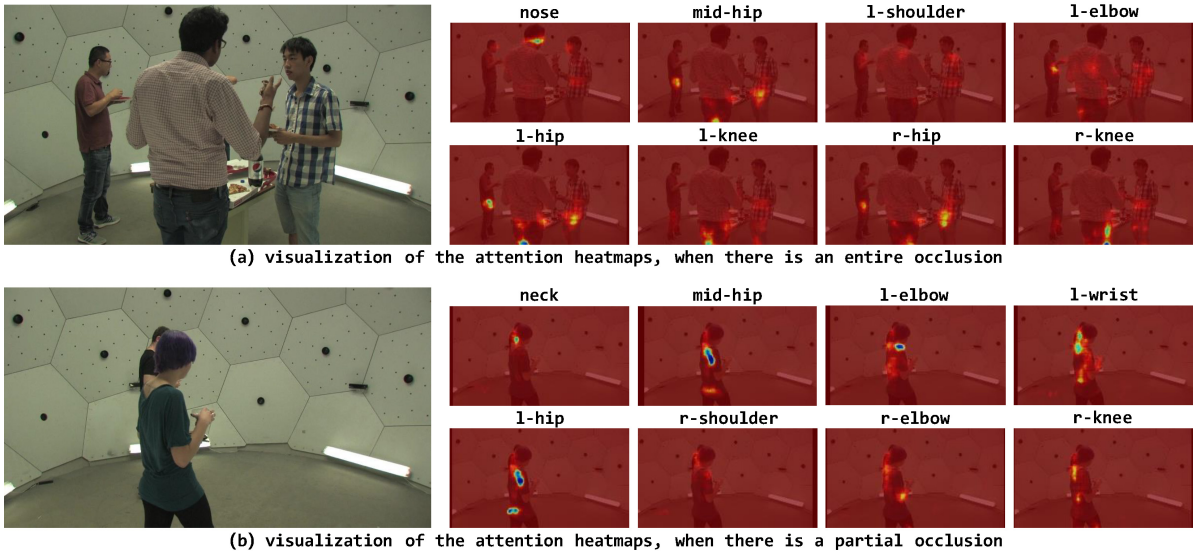

Figure 6. Visualization of the attention heatmaps. (a) The man in front of the suited man is entirely occluded, and we barely see the attention heatmaps focus on him. (b) The man is partially occluded, as we can see his head, shoulder and arm. The attention heatmaps are trying to infer the occluded part (*e.g.* mid-hip).

7.7. Attention heatmap visualization

To have a clearer view of the role that the attn_net_{2d} plays in SelfPose3d, we visualize the attention heatmaps of certain views in Figure 6. When there's an entire occlusion, attn_net_{2d} tends to ignore the occluded person. When there's a partial occlusion, attn_net_{2d} tends to infer the occluded part. The visualization explains the better performance when adding adaptive supervision attention.

7.8. Confidence threshold for pseudo labels

To investigate whether we need to filter out the pseudo labels with low confidence scores, we generate two sets of labels: the ones with no confidence threshold are called the soft labels, and the ones with a 0.7 confidence threshold on the joints are called the hard labels. We train our model with each label set under the same experiment setting, and the results are shown in Table 16. Our main takeaways are: (1) the model trained with hard labels performs slightly better at the end (especially on the AP_{25} index); (2) however, the model is more likely to collapse when we train it with hard labels. Therefore, we propose to train the model with soft labels at the beginning, and then fine-tune it with hard labels in the last 2 epochs. Table 16 shows that the proposed strategy can obtain the best result, with a stable training process.

7.9. Failure cases

Figure 7 shows some failure cases from our approach compared to the fully-supervised VoxelPose. Top row of Figure 7 shows two 3d poses for a single person. Pseudo 2d poses used in our approach contains the poses of the people outside the dome, whereas the ground truth 2d and 3d poses are curated to remove the persons outside the dome. Therefore, our approach tries to infer 3d poses for the persons outside the dome (see bottom row of Figure 7).

Method	Pseudo label category	AP_{25}	AP_{50}	AP_{100}	MPJPE
SelfPose3d	soft	51.6	**96.7**	**98.6**	24.8
	hard	54.2	96.4	**98.6**	24.6
	soft & hard	**55.1**	96.4	98.5	**24.5**

Table 16. Comparing the models trained with (1) soft pseudo labels solely, (2) hard pseudo labels solely, and (3) two sets of labels, respectively. For the soft & hard training, we only use the hard labels in the last 2 epochs.

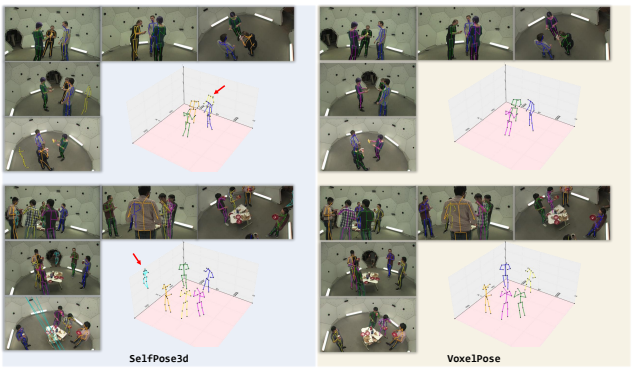

Figure 7. Failure cases from our approach compared to fully-supervised VoxelPose. The top row shows the two 3d poses for a single person, and the bottom row shows the 3d pose for a person outside the dome. Best viewed in color.