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

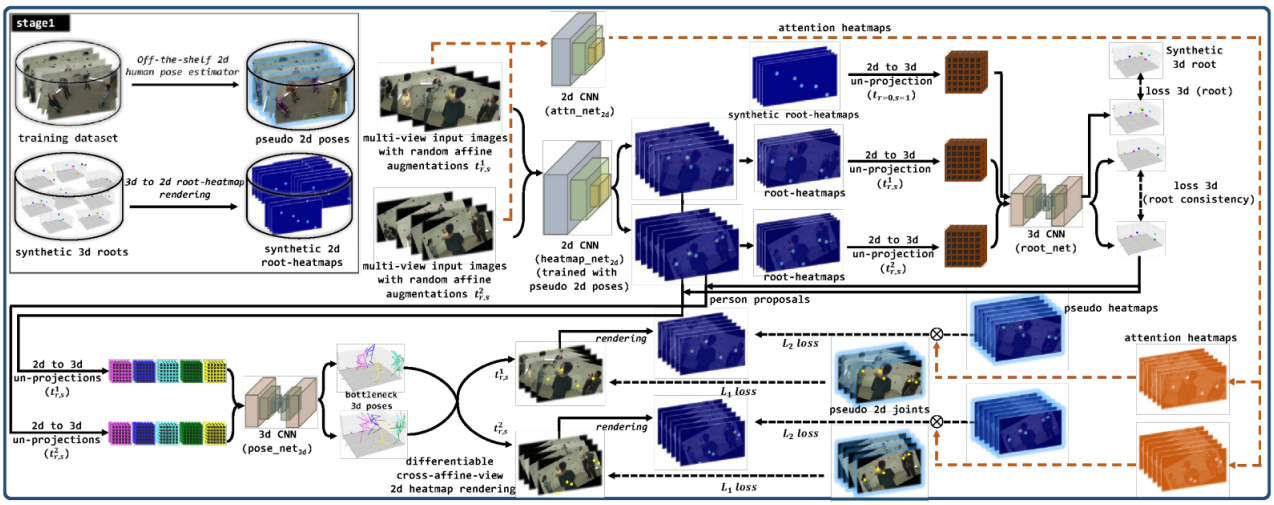
Vinkle Srivastav, Keqi Chen, Nicolas Padoy

Accepted for CVPR 2024 (Paper) (GitHub)

This paper introduces a new self-supervised approach for estimating 3d poses of multiple people from multiple camera views. This method does not require 2d or 3d ground-truth poses for training and only uses generated 2d poses from a calibrated multi-view camera setup. The model first localizes all of the humans in the 3d space and their 3d root joint, a mid-hip point. Then, the 3d pose is generated for each 3d root joint.

Motivation

Current state-of-the-art methods are either learning-based, which require 2d or 3d ground-truth poses from a dense camera system, or optimization-based, which don't rely on ground-truth but underperform in comparison. This method attempts to combine the strengths of both by training on 2D pose views from multiple cameras, which are easier to obtain than 3d ground-truth poses.



Method

1. **3D Root Localization**

   A 2d root joint heatmap is generated from each view using a 2D CNN. Then, the 2D heatmaps are triangulated to build a 3d feature volume for the person proposals using a 3D CNN.

2. **3D Pose Estimation**

   The 3d poses are predicted by passing the person proposals and 2d root heatmaps through a 3D CNN. During training, geometric constraints are enforced by projecting the 3d poses back into 2d space across affine-transformaed views.

3. **Adaptive Supervision Attention**

   Attention is used to guide the training in occluded scenarios. For example, the 2d pose estimation may generate innacurate labels when occluded and the 3d-to-2d projection will output 2d joints, even when the person is occluded from that view. Their solution was soft attention for heatmap supervision and hard attention for joint loss.

	Methods	AP_{25}	AP_{50}	AP_{100}	AP_{150}	$Recall_{@500}$	MPJPE[mm]
FS	VoxelPose [57]	83.6	98.3	**99.8**	**99.9**	98.8	17.7
	Lin *et al.* [39]	92.1	99.0	**99.8**	99.8	-	16.8
	MvP [63]	**92.3**	96.6	97.5	97.7	98.2	15.8
	Wu *et al.* [59]	-	-	-	-	98.7	15.8
	TEMPO [15]	89.0	**99.1**	**99.8**	**99.9**	-	**14.7**
OB	ACTOR [49]	-	-	-	-	-	168.4
	MvPose [18]	0.0	2.97	59.93	81.53	98.23	84.2
SS	SelfPose3d (ours)	55.1	96.4	98.5	99.0	**99.6**	24.5

Limitations

1. Occasionally hallucinates extra humans even in an isolated environment, as shown in the supplementary material compared to the fully-supervised VoxelPose
2. The performance of the off-the-shelf 2d pose estimator constrains the model's accuracy, as the model builds the resulting 3d pose from the initial multi-view 2d pose
3. May not handle heavily occluded environments where a person may not be visible to most of the cameras (trained on datasets that include multiple people and varying views but not occlusion-heavy environments)