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

Vinkle Srivastav, Keqi Chen, Nicolas Padoy

▶ **To cite this version:**

Vinkle Srivastav, Keqi Chen, Nicolas Padoy. SelfPose3d: Self-Supervised Multi-Person Multi-View 3d Pose Estimation. IEEE/CVF 2024 Conference on Computer Vision and Pattern Recognition (CVPR 2024), Jun 2024, Seattle, WA, United States. hal-04626539

HAL Id: hal-04626539
https://hal.science/hal-04626539

Submitted on 26 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Vinkle Srivastav[1,2*] Keqi Chen[1*] Nicolas Padoy[1,2]

[1]University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France
[2]IHU Strasbourg, France

srivastav@unistra.fr keqi.chen@unistra.fr npadoy@unistra.fr

Abstract

We present a new self-supervised approach, SelfPose3d*, for estimating 3d poses of multiple persons from multiple camera views. Unlike current state-of-the-art fully-supervised methods, our approach does not require any 2d or 3d ground-truth poses and uses only the multi-view input images from a calibrated camera setup and 2d pseudo poses generated from an* off-the-shelf *2d human pose estimator. We propose two self-supervised learning objectives: self-supervised person localization in 3d space and self-supervised 3d pose estimation. We achieve self-supervised 3d person localization by training the model on synthetically generated 3d points, serving as 3d person root positions, and on the projected root-heatmaps in all the views. We then model the 3d poses of all the localized persons with a bottleneck representation, map them onto all views obtaining 2d joints, and render them using 2d Gaussian heatmaps in an end-to-end differentiable manner. Afterwards, we use the corresponding 2d joints and heatmaps from the pseudo 2d poses for learning. To alleviate the intrinsic inaccuracy of the pseudo labels, we propose an adaptive supervision attention mechanism to guide the self-supervision. Our experiments and analysis on three public benchmark datasets, including Panoptic, Shelf, and Campus, show the effectiveness of our approach, which is comparable to fully-supervised methods. Code is available at* https://github.com/CAMMA-public/SelfPose3D.

1. Introduction

The task of estimating 3d poses for multiple persons using a few calibrated cameras is a challenging computer vision problem [10, 18, 32, 52, 57]. A significant part of this challenge lies in identifying and matching the same person across different camera views. The solutions developed so far generally use one of the two paradigms: *learning-based methods* and *optimization-based methods*. The learning-based methods develop novel deep-learning models and use 3d ground-truth poses for both training the models and establishing person correspondences across different views [39, 52, 57, 59, 63]. The accurate 3d ground-truth poses are typically generated using a dense camera system [32]. In contrast, the optimization-based methods formulate the 3d pose reconstruction as a mathematical optimization task, primarily focusing on aligning and matching the 2d poses across different camera views to infer 3d poses using triangulation within the framework of multi-view geometry [10, 18, 31–33, 49]. The 2d poses are estimated using off-the-shelf 2d human pose detectors [5, 55]. These methods apply geometric and spatial constraints in the optimization loop to ensure the anatomical plausibility and consistency of the inferred 3d poses. Although these methods do not require 3d ground-truth poses, their effectiveness is somewhat limited compared to the fully-supervised learning-based methods, see Table 1.

In this paper, we explore the possibility of combining the strengths of both paradigms. Specifically, we investigate whether it's feasible to utilize a learning-based model for multi-view, multi-person 3D pose estimation and simultaneously eliminate its dependence on 3D ground-truth poses by incorporating geometric and appearance constraints, drawing inspiration from optimization-based methods.

We propose, *SelfPose3d*, a self-supervised learning-based approach to estimate the 3d poses of multiple persons from a few calibrated cameras without using any 2d or 3d ground-truth poses. Our approach requires only *2d pseudo poses* obtained using an off-the-shelf 2d pose detector [55]. Learning 3d poses without 3D ground-truth poses would require suitable supervisory signals to train a learning-based model. We follow the *learning-by-projection* paradigm, where the main idea is to learn the 3d output by comparing the projected bottleneck 3d output against the 2d input features [12].

We consider VoxelPose [57] as a learning-based method and use its output 3d poses as a bottleneck representation. To recover the accurate underlying 3d poses, we propose us-

*co-first authors with equal contributions.

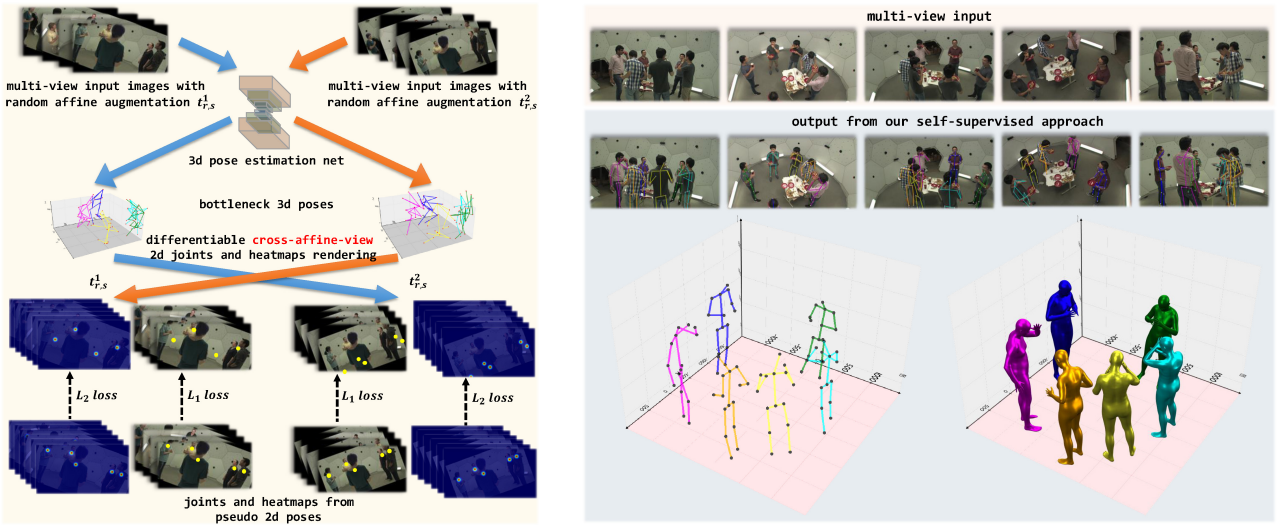

Figure 1. Our self-supervised approach, called SelfPose3d, estimates multi-person 3d poses from multi-view images and pseudo 2d poses generated using an *off-the-shelf* 2d human pose estimator. We propose a self-supervised learning objective that generates differentiable and geometrically constrained 2d joints and heatmaps across multiple views from bottleneck 3d poses. On the right, we show 3d pose outputs from our approach along with estimated body meshes (using SMPL body mesh fitting on 3d poses [4, 41]) and the projected 2d poses.

ing *differentiable multi-view 2d representations* and *cross-affine-view consistency*. In particular, given a multi-view input image, we apply two random affine augmentations and pass them to the VoxelPose. It generates the bottleneck 3d poses corresponding to each affine augmented multi-view image. To enforce the model to learn and reason in the spatial dimension, we project the bottleneck 3d poses onto each view, obtaining 2d joints, and rendering them into spatial 2d heatmap representations in an end-to-end differentiable way. We further put tight geometric constraints by cross-affine-view operation, *i.e*. the bottleneck 3d poses from the 1st affine augmented multi-view image is mapped and rendered in the 2nd affine augmented multi-view image space and vice versa. Finally, we use the affine transformed 2d joints and heatmaps from the *2d pseudo poses* to enable the geometrically constrained learning, with L_1 and L_2 losses respectively.

As the *2d pseudo poses* contain non-negligible noises (mostly due to occlusions, see Figure 3), we propose *adaptive supervision attention* to guide our model to focus on more reliable regions. We apply two strategies towards L_1 joint loss and L_2 heatmap loss; for L_1 joint loss supervision, we employ hard attention, where we ignore the one view with the largest absolute error for each multi-view image set; for L_2 heatmap loss supervision, we employ soft attention using a lighter backbone to process each view, obtaining same-size attention heatmaps. During L_2 loss computation, we compute the element-wise product of the attention heatmaps and the square error before averaging. To avoid obtaining zero attention, which the model tends to do, we add a regularization term, where we create tensors of all ones as the attention heatmap labels and use L_2 loss as the attention loss.

Finally, specific to our choice of learning-based method, *i.e*., VoxelPose, which uses a voxel-based 3d root localization model to localize the persons in space using ground-truth 3d root joints (mid-hip joint), we use a simple but effective strategy to localize persons in space. Specifically, we randomly place 3d points in 3d world-space and project them to each view using the given camera parameters, subsequently rendering the projected 2d points as heatmap representations. This generates a synthetic dataset containing 3d points (roots) and their corresponding rendered multi-view root-heatmaps. We then use this dataset to train a 3d root localization model that takes multi-view root-heatmaps as input and predicts the 3d roots as output. We further regularize the model by enforcing invariant constraints between pairs of affine augmented root-heatmaps coming from the real multi-view input.

Evaluation on three 3d pose benchmarks datasets, Panoptic [32], Shelf [1] and Campus [1], along with extensive ablation studies on the Panoptic [32] dataset, show the effectiveness of our approach. Our approach reaches a performance comparable to learning-based fully-supervised approaches and performs significantly better than optimization-based approaches. Moreover, SMPL body mesh fitting [4, 41] on our estimated 3d poses generates geometrically plausible body shapes (see Figure 1 and Figure 4).

We summarize our contributions as follows: 1) We address the challenging multi-person multi-view 3d person pose estimation problem using a self-supervised approach

without any 2d or 3d ground truth. 2) We propose *self-supervised 3d pose estimation* by using a new method to recover geometrically constrained 2d joints and heatmap representations from the bottleneck 3d poses. 3) We propose *adaptive supervision attention* to address the misinformation caused by the inaccurate pseudo labels. 4) We propose *self-supervised 3d root localization* to estimate the 3d root location utilizing synthetic 3d roots and the corresponding rendered multi-view root heatmaps.

2. Related work

In this section, we briefly review current works related to fully-supervised learning-based methods for 3d pose estimation, optimization-based methods for 3D pose estimation, and self-supervised learning.

Fully-supervised methods: Monocular 3d pose estimation [22, 43, 45, 47, 50, 56, 62, 65] is an ill-posed problem due to depth ambiguities as multiple 3d poses can produce same 2d pose projection. Having access to multi-view cameras can alleviate such depth ambiguities achieving the state-of-art results on benchmark datasets [23, 27, 28, 39, 51, 52, 57, 59, 63]. For single-person scenes, these approaches exploit multi-view geometry [24] to either fuse the visual features [27, 51], perform triangulation on heatmaps [28, 53], or use pictorial structural models for 3d reconstruction [48, 51]. The multi-person scene offers extra complexity due to the variability in the number of person in each view and the unknown cross-view correspondence. Existing multi-person multi-view approaches are based on volumetric paradigm [52, 57, 64], or direct regression [63] based on transformers [6, 58, 66]. Despite their good performance, these approaches rely on ground-truth 3d poses, which are generated using dense camera systems [32].

Optimization-based 3d pose estimation: For the multi-person and multi-view scenario, optimization-based approaches use an *off-the-shelf* person-id detector across all the views to solve the correspondence and triangulation problem and temporal refinement along with training a reinforcement learning agent to find the best camera locations for 3d pose reconstruction [49]. More recent approaches utilize multi-view 3d reconstruction in the optimization loop inferring 3D poses that are geometrically and spatially coherent [10, 18, 33].

Self-supervised learning: Self-supervised learning can be broadly classified into self-supervised *representation* learning and self-supervised *task* learning. Self-supervised *representation* learning aims to use large-scale unlabeled data to learn generic feature representations. The recent promising results from these approaches have started to surpass the fully-supervised baselines for various downstream tasks [7, 11, 26, 54]. Self-supervised task learning aims to learn a particular downstream task without using ground truth labels and has been applied to 2d pose estimation [29, 30], single-person 3d pose estimation [9, 19, 34, 36, 38], and surface correspondences estimation [3]. Self-supervised approaches for 3d pose estimation have primarily been developed for single-person scenarios. Given 2d poses, estimated by utilizing advances in the 2d pose estimation methods [5, 13, 14, 21, 35, 42, 44, 46, 55, 60], these approaches use the supervisory signals generated from multi-view geometry [34], video constraints [38], or adversarial learning [9, 19, 36].

Our work proposes a learning-based approach to model the 3d poses as bottleneck representations and recover geometrically constrained and spatially accurate 2d joints and heatmap representations in an end-to-end differentiable manner.

3. Methodology

3.1. Problem overview

Given a training dataset of multi-view images $\mathcal{D} = \{\mathbf{x}|\mathbf{y}^*\}$ where $\mathbf{x} \in \mathcal{R}^{C\times 3\times H\times W}$ is a multi-view image set from C cameras with height H and width W, and $\mathbf{y}^* \in \mathcal{R}^{C\times P\times J\times 2}$ represents the 2d pseudo poses for P persons with J joints, the goal is to learn a deep learning model that estimates the 3d poses $\mathcal{Y} \in \mathcal{R}^{P\times J\times 3}$ of all the P persons from the multi-view input image $\mathbf{x}$. It is to be noted that P can vary in each camera view due to occlusion and noisy pseudo 2d pose estimation. For simplicity in the notation, we keep the same variable P.

Fully-supervised approaches rely on 3d ground-truth poses, while we only have 2d pseudo poses $\mathbf{y}^*$. Therefore, after obtaining 3d poses $\mathcal{Y} \in \mathcal{R}^{P\times J\times 3}$ following traditional approach, we propose to project the poses to each view obtaining 2d poses $\mathbf{y} \in \mathcal{R}^{C\times P\times J\times 2}$, and train the model from 2d pseudo poses $\mathbf{y}^* \in \mathcal{R}^{C\times P\times J\times 2}$.

In the following, we present our self-supervised approach based on fully-supervised VoxelPose [57]. We first generate pseudo 2d poses, and then propose *self-supervised 3d root localization*, *self-supervised 3d pose estimation*, and an *adaptive supervision attention* to learn 3d poses in a self-supervised manner, without modifying the original VoxelPose structure.

3.2. Generating pseudo 2d poses

To circumvent the dependence on the ground-truth 2d poses, we generate the 2d pseudo poses on the training dataset using Mask R-CNN [25] to generate person bounding boxes followed by using HRNet [55] to generate 2d poses of each detected person bounding box. The two-stage approach is chosen based on its state-of-art performance on the COCO dataset [40]. We first pre-train the 2d CNN backbone $\mathrm{heatmap\_net}_{\mathrm{2d}}$ with pseudo 2d poses.

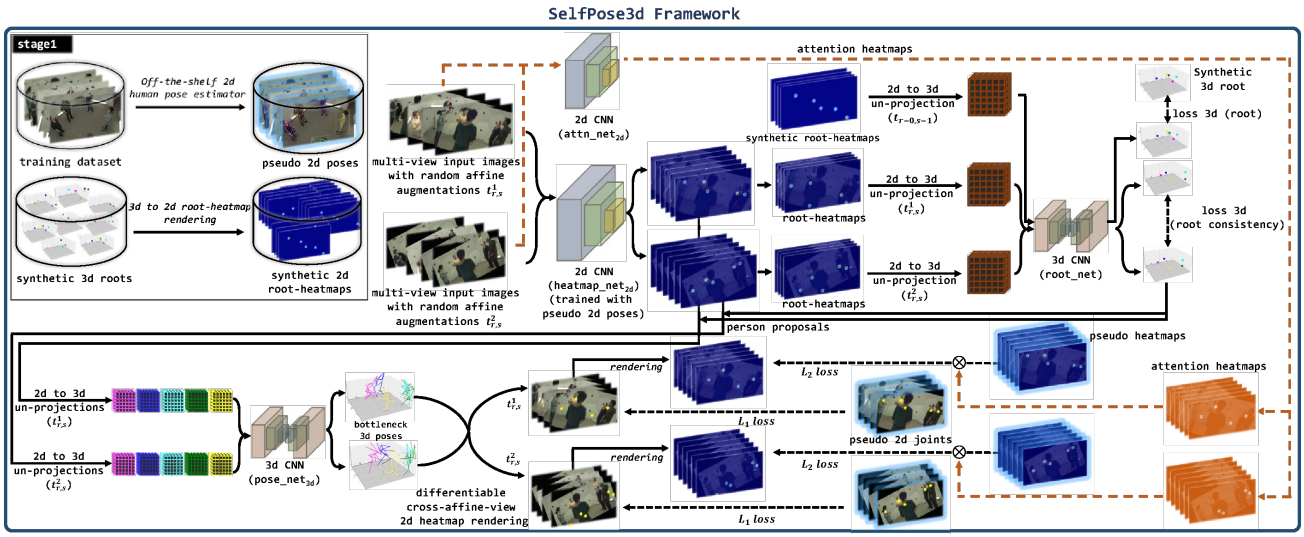

Figure 2. Illustrating our self-supervised SelfPose3d approaches for multi-view multi-person 3d pose estimation. Instead of using ground-truth 3d poses for learning, we propose self-supervised learning objectives to localize 3d roots (mid-hip location of the person) and estimate their 3d poses. We utilize a synthetic 3d roots dataset, two different affine transformations on the multi-view input images ($t^1_{r,s}, t^2_{r,s}$ parametrized by rotation r and scale s), a differentiable cross-affine-view 2d joints and heatmaps rendering from the bottleneck 3d poses, and an adaptive supervision attention mechanism to automatically learn the 3d poses in world-space.

3.3. Self-supervised 3d root localization

Given 2d multi-view heatmaps from all the views and all the joints $\mathcal{H} \in \mathcal{R}^{C\times J\times \frac{H}{4}\times \frac{W}{4}}$ estimated using a 2d backbone model $\text{heatmap\_net}_{\text{2d}}$, we use [28] to construct a discretized 3d feature volume $\mathcal{F} \in \mathcal{R}^{J\times X\times Y\times Z}$ for each joint by un-projecting the 2d multi-view heatmaps to 3d space:

$$\mathcal{P}_{\text{unproj}}(\text{cam}, \text{center}, t_{r,s}) \colon \mathcal{H} \longrightarrow \mathcal{F}, \tag{1}$$

To localize persons' root (mid-hip) joint in 3d space without using 3d ground truth, we hypothesize that 2d multi-view heatmaps of the root location $\mathcal{H}_{\text{root}} \in \mathcal{R}^{C\times \frac{H}{4}\times \frac{W}{4}}$ are sufficient for 3d root localization (see Sec. 7.6 for verification). Then we generate the 3d feature volume for the root location $\mathcal{F}_{\text{root}} \in \mathcal{R}^{X\times Y\times Z}$ using $\mathcal{H}_{\text{root}}$ using Eq. (1), which has the same dimensions as predicted root-volumes $\mathcal{G}$. This allows us to establish a one-to-one relationship between 3d root-volumes $\mathcal{G}$ and 2d multi-view root-heatmaps $\mathcal{H}_{\text{root}}$. We generate a synthetic root dataset $\mathcal{D}_{\text{root}} = \{\mathcal{G}^{\text{syn}*}_i | \mathcal{H}^{\text{syn}}_{\text{root\_}i}\}^N_{i=1}$ where $\mathcal{G}^{\text{syn}*}_i \in \mathcal{R}^{X\times Y\times Z}$ contains the root-volumes of randomly placed 3d points, and $\mathcal{H}^{\text{syn}}_{\text{root\_}i}$ is the corresponding 2d multi-view heatmaps generated by projecting the random 3d points to each view using camera parameters cam. After unprojecting $\mathcal{H}^{\text{syn}}_{\text{root}}$ to $\mathcal{F}^{\text{syn}}_{\text{root}}$, and passing it through $\text{root\_}net$ obtaining $\mathcal{G}^{\text{syn}}$, we compute the L_2 loss error as the synthetic root loss $l_{\text{root\_syn}}$:

$$l_{\text{root\_syn}} = \mathcal{L}_2(\mathcal{G}^{\text{syn}}, \mathcal{G}^{\text{syn}*}) \tag{2}$$

To further regularize root_net on the real-world 2d multi-view root-heatmaps, we propose the root consistency loss. Given a multi-view training image set x^0, we apply two affine transformations $(t^1_{r,s}, t^2_{r,s})$ with random rotation and scaling (r, s) to generate two affine transformed multi-view images (x^1, x^2). We pass x^0, x^1 and x^2 through $\text{heatmap\_net}_{\text{2d}}$, construct the root feature volumes using Eq. (1) with corresponding affine transformation parameters, and finally obtain $\mathcal{G}^0$, $\mathcal{G}^1$ and $\mathcal{G}^2$ through root_net. Since $\mathcal{G}^1$ and $\mathcal{G}^2$ are invariant to the applied affine transformations $t^1_{r,s}$ and $t^2_{r,s}$, we use $\mathcal{G}^0$ as the baseline to compute the L_2 loss error between $\mathcal{G}^0$, $\mathcal{G}^1$ and $\mathcal{G}^2$ as the root consistency loss $l_{\text{root\_}C}$:

$$l_{\text{root\_}C} = \mathcal{L}_2(\mathcal{G}^0, \mathcal{G}^1) + \mathcal{L}_2(\mathcal{G}^0, \mathcal{G}^2) \tag{3}$$

We train root_net by minimizing Eq. (2) and Eq. (3). We generate the person proposals $\{\text{root}_i\}^K_{i=1}$ by applying non-maximum suppression (NMS) and thresholding on $\mathcal{G}^2$ ($\mathcal{G}^1$ would also work).

3.4. Self-supervised 3d pose estimation

Given pseudo 2d poses $\mathbf{y}^*_{\text{2d}}$, person proposals $\{\text{root}_i\}^P_{i=1}$, and 2d multi-view heatmaps $\mathcal{H}^1, \mathcal{H}^2$ predicted using $\text{heatmap\_net}_{\text{2d}}$, we describe our self-supervised 3d pose estimation approach.

The person proposals $\{\text{root}_i\}^P_{i=1}$ are used to generate the 3d feature volumes *i.e* $\{\mathcal{F}^1_i\}^P_{i=1} = \{\mathcal{P}_{\text{unproj}}(\text{cam}, r_i, t^1_{r,s})\}^P_{i=1}$ and $\{\mathcal{F}^2_i\}^P_{i=1} =$

$\{\mathcal{P}_{\text{unproj}}(\text{cam}, r_i, t^2_{r,s})\}^P_{i=1}$ corresponding to the person feature volumes for each affine augmented multi-view input image x^1 and x^2, respectively. $\{\mathcal{F}^1_i\}^P_{i=1}$ and $\{\mathcal{F}^2_i\}^P_{i=1}$ are passed through $\text{pose\_net}_{\text{3d}}$ and soft-argmax [8] to estimate the 3d poses $\mathcal{Y}^1, \mathcal{Y}^2 \in \mathcal{R}^{P\times J\times 3}$. These 3d poses serve as a bottleneck representation.

Given the camera parameters cam and the affine transformation parameters $t^1_{r,s}, t^2_{r,s}$, we project the bottleneck 3d poses in cross-affine-view *i.e.* $\mathcal{Y}^1$ are projected to x^2 image space using $t^2_{r,s}$ to generate multi-view 2d poses $\mathbf{y}^2 \in \mathcal{R}^{C\times P\times J\times 2}$, and $\mathcal{Y}^2$ are projected to the x^1 image space using $t^1_{r,s}$ to generate multi-view 2d poses $\mathbf{y}^1 \in \mathcal{R}^{C\times P\times J\times 2}$.

We propose to render $\mathbf{y}^1$ and $\mathbf{y}^2$ 2d poses in the 2d heatmap representation. The heatmap representation encodes the per-pixel likelihood of a body joint and has been a vital component to enable the state-of-the-art 2d pose estimation approaches [55]. Generating heatmap representation has essentially been a pre-processing step where state-of-the-art approaches quantize the 2d joints before generating the heatmap [55]. However, this quantization step is non-differentiable and could cut the backward gradient flow. Zhang *et al*. [61] show that encoding floating point 2d joints into heatmap representation in their pre-processing step leads to improved performance. We propose to use the same differentiable approach in an online way to render the projected 2d joints into the heatmap representation.

We render $\mathbf{y}^1$ and $\mathbf{y}^2$ into heatmap representation to generate 2d multi-view heatmaps $\mathcal{H}^1$, and $\mathcal{H}^2$, respectively. We apply the affine transformations $t^1_{r,s}, t^2_{r,s}$ on the pseudo 2d poses $\mathbf{y}^*_{\text{2d}}$ to generate pseudo 2d multi-view joints $\mathbf{y}^{1*}_{\text{2d}}, \mathbf{y}^{2*}_{\text{2d}}$ and heatmaps $\mathcal{H}^{1*}, \mathcal{H}^{2*}$. Then, we compute the L_2 loss between heatmaps as the pose heatmap loss $l_{\text{pose\_H}}$:

$$l_{\text{pose\_H}} = \mathcal{L}_2(\mathcal{H}^1, \mathcal{H}^{1*}) + \mathcal{L}_2(\mathcal{H}^2, \mathcal{H}^{2*}) \tag{4}$$

After training with $l_{\text{pose\_H}}$ preliminarily, we propose to add the L_1 loss between 2d joints to further fine-tune the model. For each view, we employ the Hungarian algorithm [37] to obtain the optimal assignment between $\mathbf{y}$ and $\mathbf{y}^*_{\text{2d}}$, where the matching cost is the mean absolute error. Based on the optimal assignment, we obtain the L_1 loss as $l_{\text{pose\_J}}$. Then we use $l_{\text{pose\_H}}$ and $l_{\text{pose\_J}}$ together to train the whole network, where λ is a manually defined weight:

$$l_{\text{pose\_J}} = \mathcal{L}_1(\mathbf{y}^1, \mathbf{y}^{1*}_{\text{2d}}) + \mathcal{L}_1(\mathbf{y}^2, \mathbf{y}^{2*}_{\text{2d}}) \tag{5}$$

$$l_{\text{pose\_3d}} = l_{\text{pose\_H}} + \lambda l_{\text{pose\_J}} \tag{6}$$

As the network needs to reason about the 2d joint locations in spatial dimension, it implicitly solves the person correspondence problem. Training $\text{pose\_net}_{\text{3d}}$ with 3d pose loss $l_{\text{pose\_3d}}$ performs decently, but to achieve even better results, we introduce the adaptive supervision attention.

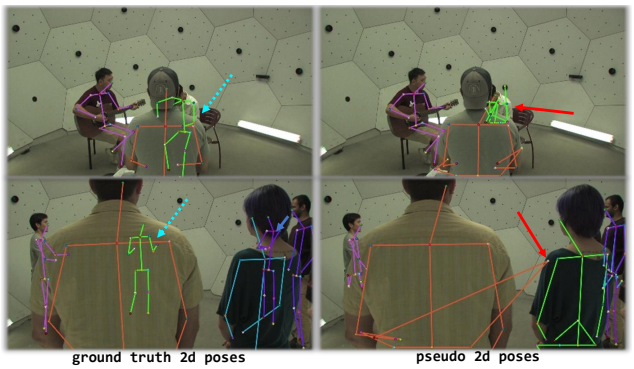

Figure 3. Comparing ground-truth 2d poses generated by projecting the ground-truth 3d poses to each multi-view image and our pseudo 2d poses generated by running HRNet human pose estimation model [55] on the training dataset. Pseudo 2d poses contain localization errors due to occlusion (see the red arrows), and ground-truth 2d poses exist for partially or even entirely occluded persons (see the blue dotted arrows).

3.5. Adaptive supervision attention

Traditional L_1 and L_2 losses treat each label equally, which is sub-optimal in two aspects: (1) the 2d human pose detector may generate inaccurate labels due to occlusions (see the red arrows in Figure 3); (2) the 3d-2d projection will output 2d joints in certain views even when the person is entirely occluded (see the blue dotted arrows in Figure 3). Therefore, we propose to employ attentions to adaptively guide the supervision process.

For L_2 loss supervision, we use the soft attention. Specifically, we use ResNet-18 to extract the visual features of the views, followed by deconvolutional layers to obtain the attention heatmaps $\mathcal{A}$ (see $\text{attn\_net}_{\text{2d}}$ in Figure 2). Then we compute the element-wise product of $\mathcal{A}$ and the square error before averaging, as the new loss $l^{\text{attn}}_{\text{pose\_H}}$:

$$l^{\text{attn}}_{\text{pose\_H}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{A}_i \otimes (\mathcal{H}_i - \mathcal{H}^*_i)^2 \tag{7}$$

To avoid that $\mathcal{A}$ becomes zero, we add a regularization term. We create tensors of all ones $\mathbb{1}$ as the attention heatmap labels, and compute L_2 error as the attention loss l_{attn}. If l_{attn} becomes zero, $l^{\text{attn}}_{\text{pose\_H}}$ degrades to non-attentive version.

$$l_{\text{attn}} = \mathcal{L}_2(\mathcal{A}, \mathbb{1}) \tag{8}$$

For L_1 loss supervision, we use the hard attention. For each input with K views, we compute the L_1 loss of each view, find the view with the largest loss, and ignore it when averaging the final loss $l^{\text{attn}}_{\text{pose\_J}}$:

$$j = \arg\max_i \mathcal{L}_1(\mathbf{y}_i, \mathbf{y}^*_{i,\text{2d}}) \quad (i = 1, 2, ..., K) \tag{9}$$

$$l^{\text{attn}}_{\text{pose\_J}} = \frac{1}{K-1} \sum_{i=1, i\neq j}^{K} \mathcal{L}_1(\mathbf{y}_i, \mathbf{y}^*_{i,\text{2d}}) \tag{10}$$

In general, the final 3d pose loss $l^{\text{attn}}_{\text{pose\_3d}}$ is as follows, where λ and σ are manually defined weights:

$$l^{\text{attn}}_{\text{pose\_3d}} = l^{\text{attn}}_{\text{pose\_H}} + \lambda l^{\text{attn}}_{\text{pose\_J}} + \sigma l_{\text{attn}} \tag{11}$$

We train $\text{pose\_net}_{\text{3d}}$ by minimizing $l^{\text{attn}}_{\text{pose\_3d}}$. Our self-supervised approach is visually described in Figure 2.

3.6. Implementation details

Training strategies For the Panoptic dataset, similar to VoxelPose [57], we first train $\text{heatmap\_net}_{\text{2d}}$ for 20 epochs with pseudo 2d poses. We use the Adam optimizer with an initial learning rate of 1e-4, which decreases to 1e-5 and 1e-6 at the 10th and 15th epochs, respectively. Then, we train the root_net for 1 epoch, followed by end-to-end joint training of the whole network for 5 epochs using only the L_2 loss, with a learning rate of 1e-4. Afterwards, we add L_1 loss to train the whole network for another 5 epochs with a learning rate of 5e-5. λ and σ in Eq. (11) are set to 0.01 and 0.1 respectively.

We use the random rotation between $-45°$ to $45°$ and random scale between -0.35 to 0.35. We also apply spatial augmentations using *rand-augment* [16] and *rand-cutout* [17] using python image library[1]. The rand-augment consist of "contrast-jittering", "auto-contrast", "equalize", "color-jittering", "sharpness-jittering", and "brightness-jittering", and the rand-cutout places random square boxes of sizes between 20 to 40 pixels at random locations in the image. We use the SMPL model and optimization-based body fitting approach[2][4, 41] to estimate body mesh parameters.

Inference pipeline During inference, we input the multi-view RGB images, and obtain the estimated 3d poses in an end-to-end pipeline. For each view, the backbone generates corresponding 2d heatmaps for cuboid construction. Then, given constructed cuboid of the whole space, the root_net predicts root joint locations of all persons. Finally, the pose_net outputs the regressed 3d locations of each joint for every cuboid proposal of the root joints.

4. Experiments

4.1. Datasets and evaluation metrics

We conduct experiments on three benchmark datasets: *Panoptic* [32], *Campus* [1], and *Shelf* [1].

[1] https://github.com/jizongFox/pytorch-randaugment

[2] https://github.com/JiangWenPL/multiperson/tree/master/misc/smplify-x

	Methods	AP_{25}	AP_{50}	AP_{100}	AP_{150}	$\text{Recall}_{@500}$	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

Table 1. Result on the Panoptic dataset (FS = fully-supervised, OB = optimization-based, SS = self-supervised).

	Methods	Shelf				Campus			
		Actor 1	Actor 2	Actor 3	Average	Actor 1	Actor 2	Actor 3	Average
FS	Ershadi *et al.* [20]	93.3	75.9	94.8	88.0	94.2	92.9	84.6	90.6
	Wu *et al.* [59]	99.3	**96.5**	97.3	**97.7**	-	-	-	-
	MvP [63]	99.3	95.1	97.8	97.4	**98.2**	**94.1**	97.4	96.6
	VoxelPose [57]	99.3	94.1	97.6	97.0	97.6	93.8	**98.8**	**96.7**
	VoxelPose* [57]	**99.5**	93.5	97.8	96.9	93.1	86.5	93.2	90.9
OB	3DPS [2]	75.3	69.7	87.6	77.5	93.5	75.7	84.4	84.5
	MvPose [18]	98.8	94.1	97.8	96.9	97.6	93.3	98.0	96.3
SS	SelfPose3d	97.2	90.3	**97.9**	95.1	92.5	82.2	89.2	87.9

Table 2. Results (in PCP) on Shelf and Campus datasets (FS = fully-supervised, OB = optimization-based, SS = self-supervised, * = reproduced results). SelfPose3d is trained from the pseudo 3d poses from the Panoptic training set.

The Panoptic dataset is a large-scale dataset captured inside a dome environment containing multiple persons performing daily social activities. We conduct extensive experiments on this dataset to evaluate and assess various components of our approach. We use the same data sequences for training and testing as VoxelPose [57] except that our training set doesn't include '160906_band3'. In other words, we are only using 9 multi-view videos for training (the '160906_band3' video is not available due to the broken images on the source website). We use the five HD camera images (3, 6, 12, 13, 23) to train and report the performance in our experiments. We use Average Precision (AP), Recall, and Mean Per Joint Position Error (MPJPE) in millimeters (mm) as evaluation metrics (higher AP and lower MPJPE are better) [57].

The Shelf and Campus are two multi-person datasets capturing activities in the indoor and outdoor environments, respectively [1]. We use the same training and test split as [1, 57]. As used in [1, 57], we use the Percentage of Correct Parts (PCP) as evaluation metrics for these two datasets.

Panoptic Table 1 shows the results on the challenging Panoptic dataset. All the fully-supervised approaches utilizing 2d and 3d ground-truth 3d poses reach nearly the same performance. SelfPose3d, without using any 3d or 2d ground-truth poses, achieves comparable results to fully-supervised approaches. Nevertheless, there still exists a gap compared to the fully-supervised VoxelPose model (96.4 *v.s* 98.3 AP_{50} and 24.5 *v.s* 17.7 MPJPE). However, unlike VoxelPose, which relies on heat maps from all the joints to es-

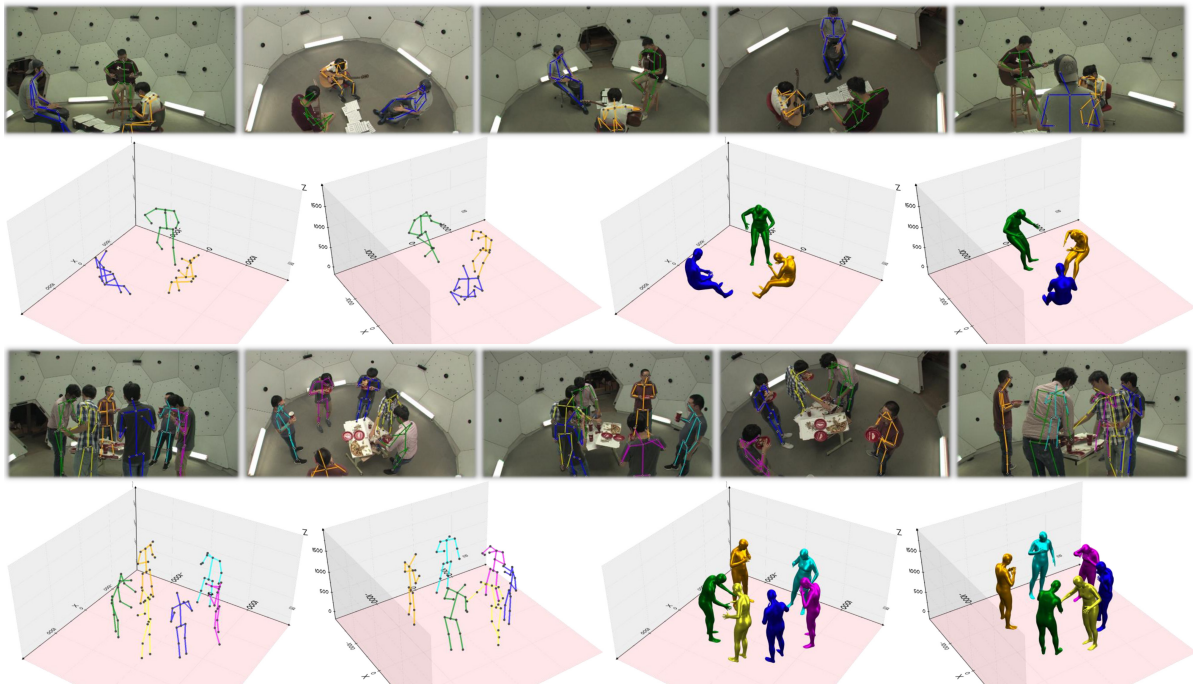

Figure 4. Qualitative results for the 3d pose estimations, 2d projections on the multi-view images, and estimated SMPL body shapes on some example images from the Panoptic dataset

.

timate 3d roots, we only use root-heatmaps to do the same. This results in the reduction of the input channel from 15 (number of keypoints) to 1 for the root_net, making our approach computationally faster.

We also compare our approach with optimization-based baselines from Pirinen *et al*. [49] and Dong *et al*. [18]. These non-learning-based approaches fail to capture the multi-person interaction in a complex scene from a few sparse multi-view cameras. Our learning-based self-supervised approach achieves much better performance. It is to be noted that Pirinen *et al*. evaluate their approach on two multi-person sequences, whereas we evaluate on four multi-person sequences.

Shelf and Campus We compare our approach with the state-of-the-art methods on the Shelf and Campus dataset. VoxelPose uses the 3d ground-truth from the Panoptic dataset to train their approach to these datasets due to noisy and incomplete 3d ground-truth poses. For a fair comparison with VoxelPose, we use the pseudo 3d poses (by running SelfPose3d on the Panoptic training set) and train on these two datasets in a fully supervised manner. As shown in Table 2, our approach using pseudo 3d poses from the Panoptic dataset also reaches the same performance as the fully-supervised approaches.

Qualitative visualizations Figure 4 shows 3d pose estimation results from our SelfPose3d approach on the challenging Panoptic dataset. Without using any 3d ground-truth, we can see that SelfPose3d is robust to occlusions and multiple persons while correctly identifying the person identities across all the views (see the corresponding 2d projections in Figure 4). We also show the qualitative results for the SMPL body mesh fitting [4, 41] on the estimated 3d poses. All these results demonstrate both the effectiveness and extendability of SelfPose3d. Please see the supplementary for more results.

4.2. Ablation studies

Ground-truth 2d poses *v.s* pseudo 2d poses As shown in Table 3, when we use the ground-truth 2d poses in our self-supervised framework, 3d reconstruction error significantly reduces. To inspect the better performance when using the ground-truth 2d poses, we qualitatively compare the ground-truth 2d poses with the pseudo 2d poses on some example training images. Pseudo 2d poses contain localization errors due to occlusion, whereas ground-truth 2d poses exist for partially or even entirely occluded persons as shown in Figure 3. As the ground-truth 2d poses are generated by projecting the ground-truth 3d poses to each multi-view image, they serve as a suitable proxy for the 3d poses, thereby reaching a performance close to the fully-supervised approaches. However, obtaining the ground 2d

2d poses	AP_{50}	AP_{100}	MPJPE
ground-truth	98.8	99.6	19.9
pseudo	96.4	98.5	24.5

Table 3. The ground-truth 2d poses in our self-supervised framework decrease the 3d reconstruction error and reach the performance close to the fully-supervised approaches.

cross-affine-view consistency	affine augs	AP_{50}	AP_{100}	MPJPE
		86.0	96.2	34.7
	✓	83.3	97.5	33.3
✓	✓	93.8	98.1	29.3

Table 4. Affine augmentations and *cross-affine-view* consistency significantly improves the 3d pose reconstruction accuracy. All three models are trained for two epochs with frozen backbone and frozen *root_net* and no attention.

L_1 loss	L_2 loss	AP_{25}	AP_{50}	AP_{100}	MPJPE
✓		-	-	-	-
	✓	43.8	95.8	98.2	25.7
✓	✓	**55.1**	**96.4**	**98.5**	**24.5**

Table 5. Training using L_1 and L_2 pose losses together achieves the best performance.

poses in this way would be as challenging as acquiring the ground-truth 3d poses.

Importance of cross-affine-view consistency and affine augmentations We also examine the effect of affine augmentations on the multi-view images and cross-affine-view consistency when generating differentiable 2d representations from the bottleneck 3d poses. As shown in Table 4, the affine augmentations and cross-affine-view consistency significantly improve the 3d pose reconstruction as they provide necessary geometric constraints during training.

Analysis of L_2 and L_1 pose losses We conduct experiments to analyze the use of L_1 and L_2 pose losses. As shown in Table 5, using L_2 and L_1 losses together can obtain better results than using L_2 loss solely. Also, using L_1 loss solely doesn't converge due to the label noises.

Importance of adaptive supervision attention We also examine the necessity of adaptive supervision attention. Table 6 shows that supervision attention for both L_1 and L_2 losses are necessary for training.

Influence of different 2d human pose estimation models Finally, we show how pseudo 2d poses generated from different 2d human pose estimation models affect the performance. As shown in Table 7, models that perform well on the COCO dataset [40] also generate better pseudo 2d poses for the Panoptic dataset, helping SelfPose3d to achieve better performance.

L_1 loss attention	L_2 loss attention	AP_{25}	AP_{50}	AP_{100}	MPJPE
		32.5	94.1	97.8	28.5
✓		37.9	95.8	98.0	26.3
	✓	47.4	**96.6**	98.2	25.0
✓	✓	**55.1**	96.4	**98.5**	**24.5**

Table 6. Training using L_1 and L_2 loss supervisions together achieves the best performance.

Method for 2d pseudo pose generation	AP_{50}	AP_{100}	MPJPE	Keypoint AP on COCO-val[40]
Keypoint R-CNN (R-101) [25]	89.2	97.6	31.9	66.1
HRNet (w48 384x288) [55]	93.8	98.1	29.3	76.3

Table 7. Comparing different models for generating pseudo 2d poses. Models that perform well on the COCO dataset [40] also generate better pseudo 2d poses for the Panoptic dataset, helping SelfPose3d to achieve better performance.

for the Panoptic dataset, helping SelfPose3d to achieve better performance.

5. Conclusion

We present a self-supervised approach, called *SelfPose3d*, to address the challenging problem of multi-view multi-person 3d human pose estimation. Unlike current state-of-the-art methods that use difficult-to-acquire 3d ground-truth poses to train the model, SelfPose3d requires only multi-view input images and an *off-the-shelf* 2d human pose detector. We propose a novel self-supervised learning objective that aims to recover 2d joints and heatmaps under different affine transformations from the bottleneck 3d poses. We further improve the performance of our approach by integrating adaptive supervision attention to address the misinformation caused by the inaccurate 2d pseudo labels from the *off-the-shelf* 2d human pose detector. We conduct extensive experiments on large-scale benchmark datasets, assess various components of our approach, and show that SelfPose3d reaches a performance on par with the well-established fully-supervised baselines. We visualize the 3d pose reconstruction in the complex multiple-person scenes and show that body shape meshes fitted on the estimated 3d poses look geometrically plausible under different viewpoints.

6. Acknowledgements

This work was partially supported by French state funds managed by the ANR under references ANR-20-CHIA-0029-01 (National AI Chair AI4ORSafety), ANR-10-IAHU-02 (IHU Strasbourg), ANR-18-CE45-0011-03 (OptimiX), and by BPI France (project 5G-OR). This work was also granted access to the servers/HPC resources managed by CAMMA, IHU Strasbourg, Unistra Mesocentre, and GENCI-IDRIS [Grant 2021-AD011011638R3].

References

[1] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, pages 1669–1676, 2014. 2, 6

[2] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *TPAMI*, 38(10):1929–1942, 2015. 6

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020. 3

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578. Springer, 2016. 2, 6, 7

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 1, 3

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3

[8] Olivier Chapelle and Mingrui Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010. 5

[9] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, M. V. Rohith, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. *CoRR*, abs/1904.04812, 2019. 3

[10] He Chen, Pengfei Guo, Pengfei Li, Gim Hee Lee, and Gregory Chirikjian. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. In *European Conference on Computer Vision*, pages 541–557. Springer, 2020. 1, 3

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[12] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in neural information processing systems*, 32, 2019. 1

[13] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 3

[14] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 3

[15] Rohan Choudhury, Kris M Kitani, and László A Jeni. Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14750–14760, 2023. 6

[16] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 6

[17] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[18] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *CVPR*, pages 7792–7801, 2019. 1, 3, 6, 7

[19] Dylan Drover, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 3

[20] Sara Ershadi-Nasab, Erfan Noury, Shohreh Kasaei, and Esmaeil Sanaei. Multiple human 3d pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12): 15573–15601, 2018. 6

[21] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3

[22] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8575–8584, 2021. 3

[23] Brian Gordon, Sigal Raab, Guy Azov, Raja Giryes, and Daniel Cohen-Or. Flex: Parameter-free multi-view 3d human motion reconstruction. *arXiv preprint arXiv:2105.01937*, 2021. 3

[24] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 8

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[27] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7779–7788, 2020. 3

[28] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. *arXiv preprint arXiv:1905.05754*, 2019. 3, 4

[29] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *Advances in neural information processing systems*, 31, 2018. 3

[30] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020. 3

[31] Hanbyul Joo, Hyun Soo Park, and Yaser Sheikh. Map visibility estimation for large-scale dynamic 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1122–1129, 2014. 1

[32] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 1, 2, 3, 6

[33] Abdolrahim Kadkhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3d human pose regression. *Machine Vision and Applications*, 32(1):1–14, 2021. 1, 3

[34] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019. 3

[35] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, pages 11977–11986, 2019. 3

[36] Yasunori Kudo, Keisuke Ogaki, Yusuke Matsui, and Yuri Odagiri. Unsupervised adversarial learning of 3d human pose from 2d joint locations. *arXiv preprint arXiv:1803.08244*, 2018. 3

[37] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[38] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6152–6162, 2020. 3

[39] Jiahao Lin and Gim Hee Lee. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11886–11895, 2021. 1, 3, 6

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 3, 8

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 6, 7

[42] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9034–9043, 2021. 3

[43] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 3

[44] William McNally, Kanav Vats, Alexander Wong, and John McPhee. Evopose2d: Pushing the boundaries of 2d human pose estimation using neuroevolution. *arXiv preprint arXiv:2011.08446*, 2020. 3

[45] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 3

[46] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, pages 2277–2287, 2017. 3

[47] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6951–6960, 2019. 3

[48] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, pages 1253–1262, 2017. 3

[49] Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 6, 7

[50] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6289–6298, 2017. 3

[51] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *ICCV*, pages 4342–4351, 2019. 3

[52] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15190–15200, 2021. 1, 3

[53] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *CVPR*, pages 6040–6049, 2020. 3

[54] Pierre H Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, et al.

Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020. 3

[55] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 1, 3, 5, 8

[56] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, pages 529–545, 2018. 3

[57] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, pages 197–212. Springer, 2020. 1, 3, 6, 2

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[59] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-based 3d multi-person pose estimation using multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11148–11157, 2021. 1, 3, 6

[60] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 3

[61] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020. 5

[62] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3d human pose estimation. *Advances in Neural Information Processing Systems*, 33:2408–2419, 2020. 3

[63] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3, 6

[64] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenyu Liu, and Wenjun Zeng. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *arXiv preprint arXiv:2108.02452*, 2021. 3

[65] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 3

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3