Unsupervised Cross-Modal Alignment for Multi-Person 3D Pose Estimation

Jogendra Nath Kundu*, Ambareesh Revanur*, Govind Vitthal Waghmare, Rahul Mysore Venkatesh, and R. Venkatesh Babu

Video Analytics Lab, Indian Institute of Science, Bangalore

Abstract. We present a deployment friendly, fast bottom-up framework for multi-person 3D human pose estimation. We adopt a novel neural representation of multi-person 3D pose which unifies the position of person instances with their corresponding 3D pose representation. This is realized by learning a generative pose embedding which not only ensures plausible 3D pose predictions, but also eliminates the usual keypoint grouping operation as employed in prior bottom-up approaches. Further, we propose a practical deployment paradigm where paired 2D or 3D pose annotations are unavailable. In the absence of any paired supervision, we leverage a frozen network, as a teacher model, which is trained on an auxiliary task of multi-person 2D pose estimation. We cast the learning as a cross-modal alignment problem and propose training objectives to realize a shared latent space between two diverse modalities. We aim to enhance the model's ability to perform beyond the limiting teacher network by enriching the latent-to-3D pose mapping using artificially synthesized multi-person 3D scene samples. Our approach not only generalizes to in-the-wild images, but also yields a superior trade-off between speed and performance, compared to prior top-down approaches. Our approach also yields state-of-the-art multi-person 3D pose estimation performance among the bottom-up approaches under consistent supervision levels.

1 Introduction

Multi-person 3D human pose estimation aims to simultaneously isolate individual persons and estimate the location of their semantic body joints in a 3D space. This challenging task can aid a wide range of applications related to human behavior understanding such as surveillance [56], group activity recognition [30], sports analytics [12], etc. Existing multi-person pose estimation approaches can be broadly classified into two categories namely, top-down and bottom-up. In top-down approaches [47,48,7,36], the first step is to detect persons using an off-the-shelf detector which is followed by predicting a 3D pose for each person using a single-person 3D pose estimator. Such approaches [47,48] are usually incapable of inferring absolute camera-centered distance of each human as they miss the global context. In contrast, the bottom-up approaches [34] first locate the body

*Equal contribution. | *Webpage*: https://sites.google.com/view/multiperson3D

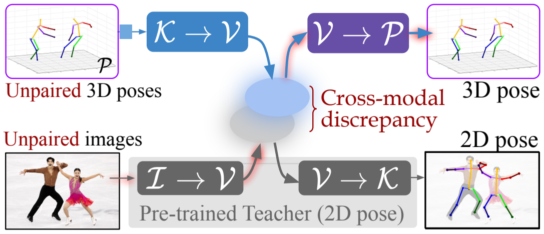

$\mathcal{K} \rightarrow \mathcal{V}$
$\mathcal{V} \rightarrow \mathcal{P}$
$\mathcal{P}$
Unpaired 3D poses
3D pose
Cross-modal discrepancy
Unpaired images
2D pose
$\mathcal{I} \rightarrow \mathcal{V}$
$\mathcal{V} \rightarrow \mathcal{K}$
Pre-trained Teacher (2D pose)

Fig. 1. We aim to realize a shared latent space $\mathcal{V}$ which embeds samples from varied input modalities *i.e.* the unpaired images and unpaired 3D poses. Auto-encoding pathway: $\mathcal{K} \rightarrow \mathcal{V} \rightarrow \mathcal{P}$. Distillation pathway: from $\mathcal{I} \rightarrow \mathcal{V} \rightarrow \mathcal{K}$ to camera projection of $\mathcal{I} \rightarrow \mathcal{V} \rightarrow \mathcal{P}$. Inference: $\mathcal{I} \rightarrow \mathcal{V} \rightarrow \mathcal{P}$ (red shadow).

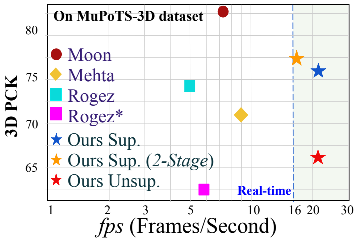


Fig. 2. We achieve a superior trade-off between speed and performance against the prior arts (Rogez[48], Rogez*[47], Mehta[34], Moon[36]). See Section 5

joints, and then assign them to each individual person via a keypoint grouping operation. The bottom-up approaches yield suboptimal results as compared to top-down approaches, but have a superior run-time advantage against top-down methods [18,46]. In this paper, we aim to leverage the computational advantage of bottom-up approaches while effectively eliminating the keypoint grouping operation via an efficient 3D pose representation. This results in a substantial gain in performance while maintaining an optimal computational overhead.

Almost all multi-person 3D pose estimation approaches access large-scale datasets with 3D pose annotations. However, owing to the difficulties involved in capturing 3D pose in wild outdoor environments, many of the 3D pose datasets are captured in indoor settings. This restricts diversity in the corresponding images (*i.e.* limited variations in background, attires and pose performed by actors) [14,15]. However, 2D keypoint annotations [19,20] are available even for in-the-wild multi-person outdoor images. Hence, several approaches aim to design 2D-to-3D pose lifters [4,32] by relying on an off-the-shelf, Image-to-2D pose estimator. Such approaches usually rely on geometric self-consistency of the projected 2D pose obtained from the lifter output, while imposing adversarial prior to assure plausible 3D pose predictions [4,16]. However, the generalizability of such approaches is limited owing to the dataset bias exhibited by the primary Image-to-2D pose estimator which is trained in a fully-supervised fashion.

Our problem setting. Consider a scenario where a pretrained Image-to-2D pose estimator is used for the goal task of 3D pose estimation. There are two challenges that must be tackled. First, a pretrained Image-to-2D estimator would exhibit a dataset bias towards the training data. Thus, the deployment of such a model in an unseen environment (*e.g.* dancers in unusual costumes) is not guaranteed to result in an optimal performance. This curtails the learning of the 2D-to-3D pose lifter, especially in the absence of paired images from the unseen environment. Second, along with the Image-to-2D model, one can not expect to be provided with its labeled training dataset owing to proprietary [37,29] or even memory [9,26] constraints. Considering these two challenges, the problem boils down to performing domain adaptation by leveraging the pretrained Image-to-2D network (*a.k.a* the teacher network) in an unsupervised fashion, *i.e. in the absence of any paired 2D or 3D pose annotations.* Further, acknowledging the limitations of existing 2D-to-3D pose lifters, we argue that the 3D pose lifter should access the latent convolutional features instead of the final 2D pose output; owing to its greater task transferability [28].

Though it is easy to obtain unpaired multi-person images, acquiring a dataset of unpaired multi-person 3D pose is inconvenient. To this end, we synthesize multi-person 3D scenes by randomly placing the single-person 3D skeletons in a 3D grid as shown in Fig. 3B. We also formalize a systematic way to synthesize single-person 3D pose by accessing plausible ranges of parent-relative joint angle limits provided by biomechanic experts. This eradicates our dependency even on an unpaired 3D skeleton dataset. Our idea of creating artificial samples stems from the concept of domain randomization [42,53] which is shown to be effective for generalizing deep models to unseen target environments. The core hypothesis is that the multi-person 3D pose distribution characterized by the artificially synthesized 3D pose scenes would subsume the unknown target distribution. Note that the proposed joint angle sampling would allow sampling of minimal implausible single-person poses as it does not adhere to the strong pose-conditioned joint angle priors formalized by Akhter *et al.* [2].

We posit the learning framework as a cross-modal alignment problem (see Fig. 1). To this end, we aim to realize a shared latent space $\mathcal{V}$, which embeds samples from varied input modalities [6], such as unpaired multi-person image $\mathcal{I}$ and unpaired multi-person 2D pose $\mathcal{K}$ (*i.e.* camera projection on multi-person 3D pose $\mathcal{P}$). Our training paradigm employs an auto-encoding loss on $\mathcal{P}$ (via $\mathcal{K} \to \mathcal{V} \to \mathcal{P}$ pathway), a distillation loss on $\mathcal{K}$ (via $\mathcal{I} \to \mathcal{V} \to \mathcal{P} \to \mathcal{K}$ pathway) and an additional adaptation loss (non-adversarial) to minimize the cross-modal discrepancy at the latent space $\mathcal{V}$. In further training iterations, we stop the limiting distillation loss and fine-tune the model on a self-supervised criteria based on the equivariance property [49] of spatial-transformations on the image and its corresponding 2D pose representation. Extensive experiments of our ablations and comparisons against prior arts establish the superiority of this approach. In summary, our contributions are as follows:

- We propose an efficient bottom-up architecture that yields fast and accurate single-shot multi-person 3D pose estimation performance with structurally infused articulation constraints to assure valid 3D pose output. In absence of paired supervision we cast the learning as a cross-modal alignment problem and propose training objectives to realize a shared latent space between two diverse data-flow pathways.
- We enhance the model's ability to perform even beyond the limiting teacher network as a result of the enriched latent-to-3D-pose mapping using artificially synthesized multi-person 3D scene samples.
- Our approach not only yields *state-of-the-art* multi-person 3D pose estimation performance among the prior bottom-up approaches but also demonstrates a superior trade-off between speed and performance.

2 Related Work

Multi-person 2D pose estimation works can be broadly classified into top-down and bottom-up methods. Top-down methods such as [5,39,11,54] first detect the persons in the image and then estimate their poses. On the other hand, bottom-up methods [38,44,13,3,40] predict the pose of all persons in a single-shot. Cao *et al.* [3] use a non-parametric representation Part Affinity Field (PAF) and Part Confidence Map (PCM) to learn association between 2D keypoints and persons in the image. Similarly, Kocabas *et al.* [18] proposed a bottom-up approach using pose residual network for estimating both keypoints and human detections simultaneously.

Many approaches have been proposed for solving the problem of single-person 3D human pose estimation [51,24,23,25,41,55]. Vnect [35] is the first realtime 3D human pose estimation work that infers the pose by parsing location-maps and joint-wise heatmaps. Martinez *et al.* [32] proposed an effective approach to directly lift the ground-truth 2D poses to 3D poses. Few methods have been proposed so far for Multi-person 3D pose estimation. In [47,48], Rogez *et al.* proposed a top-down approach based on localization, classification and regression of 3D joints. These modules are pipelined to predict the final pose of all persons in the image. Mehta *et al.* [34] proposed a single-shot approach to infer 3D poses of all people in the image using PAF-PCM representation. To handle occlusions, they introduced Occlusion Robust Pose Maps (ORPM) which allows full body pose inference under occlusions. Moon *et al.* [36] proposed the first top-down camera-centered 3D pose estimation. Their framework contains three modules: DetectNet localizes multiple persons in the image, RootNet estimates camera-centered depth of root joint and PoseNet estimates root relative 3D pose of the cropped person. In RootNet, they use pinhole camera projection model to estimate absolute camera-centered depth. Dabral *et al.* [7] proposed a 2D to 3D lifting based approach for camera-centric predictions. Rogez *et al.* [47,48] and Moon *et al.* [36] crop the detected person instances from the image and they do not leverage the global context information. All prior state-of-the-art works [47,48,34,7,36] require paired supervision. See Table 1 for a characteristic comparison against prior works.

Table 1. Characteristic comparison against prior works. **without paired supervision** implies the method does not need access to annotations.

Methods	Single shot	*w/o* paired supervision	Camera centric
Rogez [47]	✗	✗	✗
Mehta [34]	✓	✗	✗
Rogez [48]	✗	✗	✗
Dabral [7]	✗	✗	✓
Moon [36]	✗	✗	✓
Ours	✓	✓	✓

Cross-modal distillation. Gupta *et al.* [10] proposed a novel method for enabling cross-modal transfer of supervision for tasks such as depth estimation. They propose alignment of representations from a large labeled modality to a sparsely labeled modality. In [50], Spurr *et al.* demonstrated the effectiveness of cross-modal alignment of latent space for the task of hand pose estimation. In a related work [43], Pilzer *et al.* proposed an unsupervised distillation based depth estimation approach via refinement of cycle-inconsistency.

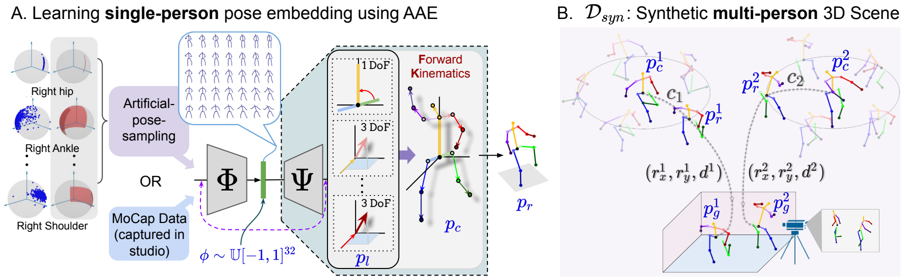
3 Approaches

Our prime objective is to realize a learning framework for the task of multi-person 3D pose estimation without accessing any paired data (*i.e.* images with the corresponding 2D or 3D pose annotations). To achieve this, we plan to distill the knowledge from a frozen teacher network which is trained for an auxiliary task of multi-person 2D landmark estimation. Furthermore, in contrast to the general top-down approaches in fully-supervised scenarios, we propose an effective single-shot, bottom-up approach for multi-person 3D pose estimation. Such an architecture not only helps us maintain an optimal computational overhead but also lays a suitable ground for cross-modal distillation.

3.1 Architecture

Aiming to design a single-shot end-to-end trainable architecture, we draw motivation from the real-time object detectors such as YOLO [46]. The output layer in YOLO divides the output spatial map into a regular grid of cells. The multi-dimensional vector at each grid location broadly represents two important attributes. Firstly, a confidence value indicating the existence of an object centroid in the corresponding input image patch upon registering the grid onto the spatial image plane. Secondly, a parameterization of the object properties, such as class probabilities and attributes related to the corresponding bounding box. In similar lines, for multi-person 3D pose estimation, each grid location of the output layer represents a heatmap indicating existence of a human pelvis location (or root) followed by a *parameterization* of the corresponding root-relative 3D pose. Here, the major challenge is how to parameterize root-relative human 3D pose in the efficient manner. We explicitly address it in the following subsection.

3.1.1 Parameterizing 3D pose via pose embedding. Root relative human 3D pose follows a complex structured articulation. Moreover, defining a parameterization procedure without accounting for the structural plausibility of the 3D pose would further add up to the inherent 2D to 3D ambiguity. Acknowledging this, we aim to devise a parameterization which selectively decodes anthropomorphically plausible human poses spanning a continuous latent manifold (see Fig. 3A). One of the effective ways to realize the above objective is to train a generative network which models the most fundamental form of human pose variations. Thus, we disentangle the root-relative pose p_r into its rigid and non-rigid factors. The non-rigid factor, also known as the canonical pose p_c is designed to be view-invariant. The rigid transformation is defined by the parameters c as required for the corresponding rotation matrix. In further granularity, according to the concept of forward kinematics [58], movement of each limb is constrained by the parent-relative joint-angle limits and the scale invariant fixed relative bone lengths. Thus, the unit vectors corresponding to each joint defined at their respective parent-relative local coordinate system [2] is regarded as the most fundamental form of 3D human pose which is denoted by p_l. Note that,


A. Learning **single-person** pose embedding using AAE

B. $\mathcal{D}_{syn}$: Synthetic **multi-person** 3D Scene

Fig. 3. A. Learning continuous pose-embedding on MoCap or Artificially sampled pose dataset. **B**. Creating $\mathcal{D}_{syn}$: Each canonical pose p_c is rigidly transformed through rotation and translation operation to form random 3D scenes.

the transformation $p_l \rightarrow p_c$ is a fully-differentiable series of forward kinematic operations. We train a generative network [21,22] following the learning procedure of adversarial auto-encoder $\{\Phi, \Psi\}$ (AAE [31]) on samples of p_l acquired from either a MoCap [1] dataset or via a proposed *Artificial-pose-sampling* procedure (see Fig. 3A). It is important to note that we consider a uniform prior distribution *i.e.* $\mathbb{U}[-1,1]^{32}$. This ensures that any random vector $\phi \in [-1,1]^{32}$ decodes (via Ψ) an anthropomorphically plausible human pose. (See Suppl)

In the proposed *Artificial-pose-sampling* procedure, we use a set of joint angle limits (4 angles *i.e.* the allowed range of polar and azimuthal angles in the parent relative local pose representation) provided by the biomechanic experts. The angle for each limb is independently sampled from a uniform distribution defined by the above range values (see the highlighted regions on the sphere for each body joint in Fig. 3A). Note that, the proposed joint angle sampling would allow sampling of minimal implausible single-person poses as it does not adhere to the pose-conditioned joint angle limits formalized by Akther *et al.* [2]. (See Suppl)

3.1.2 Neural representation of multi-person 3D pose. The last layer output of the single-shot latent to multi-person 3D pose mapper $\mathcal{H}$, denoted as s, is a 3D tensor of size $H \times W \times M$ (see block $\mathcal{M}$ Fig. 4B). The number of channels constitutes of 4 distinct components. The M dimensional vector for each grid location r^i constitutes of 4 distinct components viz, **a)** a scalar heatmap intensity indicating existence of a skeleton pelvis denoted as h^i, **b)** a 32 dimensional 3D pose embedding ϕ^i, **c)** 6 dimensional rigid transformation parameters c^i (sin and cos component of 3 rotation angles), and **d)** a scalar absolute depth d^i associated with the skeleton pelvis. Note that, the last 3 components are interpretable only in presence of a pelvis at the corresponding grid location as denoted by the first component. Here, ϕ^i is obtained through a *tanh* nonlinearity thus constraining it to decode (via frozen Ψ AAE from Section 3.1.1) only plausible 3D human poses.

The model accesses a set of 2D pelvis key-point locations belonging to each person in the corresponding input image, denoted as $\{r^i\}_{i=1}^{N}$. Here, N denotes the total number of persons. These spatial locations are obtained either as estimated

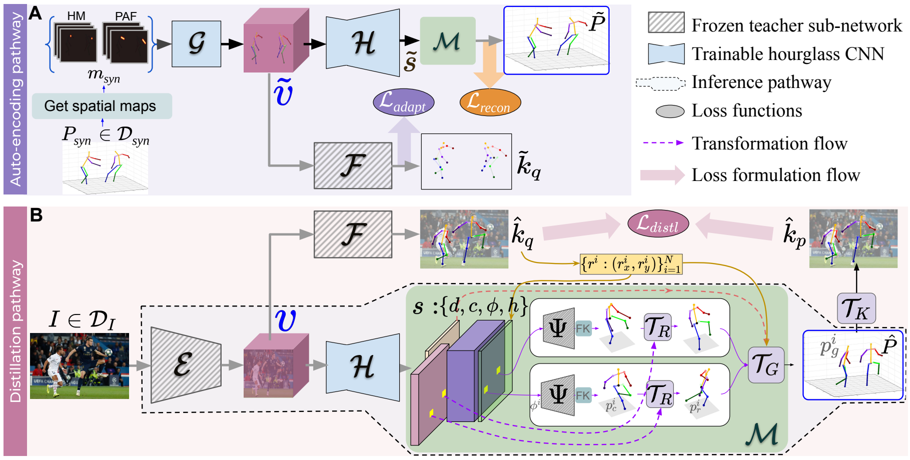


Fig. 4. Proposed data-flow pathways. Distillation is performed from the teacher, $\{\mathcal{E}, \mathcal{F}\}$ to the student $\{\mathcal{H}\}$. Weights of $\mathcal{H}$ and $\mathcal{F}$ are shared across both the pathways.

by the teacher network or from the ground-truth depending on its availability. For each selected location r^i, the corresponding ϕ^i and c^i are pooled from the relevant grid location to decode (via Ψ) the corresponding root-relative 3D pose, p^i_r. First, the canonical pose, p^i_c is obtained by applying forward kinematics (denoted as FK in Fig. 4B in module $\mathcal{M}$) on the decoded local vectors obtained from the pose embedding ϕ^i. Following this, p^i_r is obtained after performing rigid transformation using c^i, *i.e.* $\mathcal{T}_R$ in Fig. 4B. Finally, the global 3D pose scene, $\hat{P} = \{p^i_g\}^N_{i=1}$, is constructed by translating the root-relative 3D poses to their respective root locations in the camera centered global coordinate system, *i.e.* $\mathcal{T}_G$ in Fig. 4B. The 3D translation for each person i is obtained using (r^i_x, r^i_y, d^i), where r^i_x and r^i_y are the X and Y component obtained as a transformation of the spatial root location r^i. In Fig. 4B, the series of fixed (non-trainable) differentiable operations to obtain $\hat{P}$ from the CNN output s is denoted as $\mathcal{M}$. A weak perspective camera transformation, $\mathcal{T}_K$, of P provides us the corresponding multi-person 2D key-points denoted by $\hat{k}_p$.

Inference. During inference, (r^i_x, r^i_y) is obtained from the heatmap channel h predicted at the output of $\mathcal{F}$. We follow the non-maximum suppression algorithm inline with Cao *et al.* [3] to obtain a set of spatial root locations belonging to each person. Thus, the inference pathway during testing is as follows, $\hat{P} = \mathcal{M} \circ \mathcal{H} \circ \mathcal{E}(I)$.

3.2 Learning cross-modal latent space

We posit the learning framework as a cross-modal alignment problem. Moreover, we aim to realize a shared latent space, $\mathcal{V}$ which embed samples from varied modality spaces, such as multi-person image $\mathcal{I}$, multi-person 2D pose $\mathcal{K}$, and multi-person 3D pose $\mathcal{P}$. However, in absence of labeled samples (or paired samples) an intermediate representation of the frozen teacher network is treated as the shared latent embedding. Following this, separate mapping networks are trained to encode or decode the latent representation to various source modalities. Note that, the teacher network already includes the mapping of image to the latent space, $\mathcal{E}:\mathcal{I}\rightarrow\mathcal{V}$ and latent space to multi-person 2D pose, $\mathcal{F}:\mathcal{V}\rightarrow\mathcal{K}$. We train two additional mapping networks, viz. a) multi-person 2D pose to latent space, $\mathcal{G}:\mathcal{K}\rightarrow\mathcal{V}$ and b) latent space to multi-person 3D pose, $(\mathcal{M}\circ\mathcal{H}):\mathcal{V}\rightarrow\mathcal{P}$. Also note that, $(\mathcal{T}_K\circ\mathcal{M}\circ\mathcal{H}):\mathcal{V}\rightarrow\mathcal{K}$.

Available Datasets. We have access to two unpaired datasets viz. a) unpaired multi-person images $I\sim\mathcal{D}_I$ and b) unpaired multi-person 3D pose samples $P_{syn}\sim\mathcal{D}_{syn}$. Though it is easy to get hold of unpaired multi-person images, acquiring a dataset of unpaired multi-person 3D pose is inconvenient. Acknowledging this, we propose a systematic procedure to synthesize a large-scale multi-person 3D pose dataset from a set of plausible single-person 3D poses. A multi-person 3D pose sample constitute of a certain number of persons (samples of p_l^i) with random rigid transformations (c^i) placed at different locations ($i.e.$ r_x^i, r_y^i, d^i) in a 3D room. This is illustrated in Fig. 3B. Here, samples of p_l can be obtained either from a MoCap dataset or by following *Artificial-pose-sampling*.

Broadly, we use two different data-flow pathways as shown in Fig. 4. Here, we discuss how these pathways support an effective cross-modal alignment.

a) Cross-modal distillation pathway for $I\sim\mathcal{D}_I$. The objective of distillation pathway is to instill the knowledge of mapping an input RGB image to the corresponding multi-person 2D pose ($i.e.$ from the teacher network $\hat{k}_q=\mathcal{F}(v)$ where $v=\mathcal{E}(I)$) into the newly introduced 3D pose estimation pipeline. Here, $\hat{k}_q$ is obtained after performing bipartite matching inline with Cao *et al.* [3]. We update the parameters of $\mathcal{H}$ by imposing a distillation loss between $\hat{k}_q$ and the perceptively projected 2D pose $\hat{k}_p=\mathcal{T}_K\circ\mathcal{M}\circ\mathcal{H}(v)$, $i.e.$ $\boldsymbol{\mathcal{L}_{distl}=|\hat{k}_p-\hat{k}_q|}$.

b) Auto-encoding pathway for $P_{syn}\sim\mathcal{D}_{syn}$. In the auto-encoding pathway, the objective is to reconstruct back the synthesized samples of multi-person 3D poses via the shared latent space. Owing to the spatially structured latent representation, for each non-spatial P_{syn} we first generate the corresponding multi-person spatial heatmap (HM) and Part Affinity Map (PAF) inline with Cao *et al.* [3], denoted by m_{syn} in Fig. 4A. Note that m_{syn} represents the 2D keypoint locations of k_{syn} which is the obtained as the camera projection of the P_{syn}. Following this, we obtain $\tilde{P}=\mathcal{M}\circ\mathcal{H}(\tilde{v})$ where $\tilde{v}=\mathcal{G}(m_{syn})$. Parameters of both $\mathcal{G}$ and $\mathcal{H}$ are updated to minimize $\boldsymbol{\mathcal{L}_{recon}=|P_{syn}-\tilde{P}|}$.

c) Cross-modal adaptation. Notice that, $\mathcal{H}$ is the only common model updated in both pathways. Here, $\mathcal{L}_{distl}$ is computed against the noisy teacher prediction that too in the 2D pose space. In contrast, $\mathcal{L}_{recon}$ is computed against the true ground-truth 3D pose thus devoid of the inherent 2D to 3D ambiguity. As a result of this disparity, the model $\mathcal{H}$ differentiates between the corresponding input distributions, $i.e.$ between $\mathbb{P}(v)$ and $\mathbb{P}(\tilde{v})$, thereby learning separate strategies favouring the corresponding learning objectives. To minimize this discrepancy, we rely on the frozen teacher sub-network $\mathcal{F}$. We hypothesize that, the energy computed via $\mathcal{F}$, $i.e.$ $|\mathcal{F}(\tilde{v})-m_{syn}|$ would be low if the associated input distribution of $\mathcal{F}$, $i.e.$ $\mathbb{P}(v=\mathcal{E}(I))$ aligns with the output distribution of $\mathcal{G}$, $i.e.$ $\mathbb{P}(\tilde{v}=\mathcal{G}(m_{syn}))$.

Accordingly, we propose to minimize $\boldsymbol{\mathcal{L}_{adapt}} = |\boldsymbol{\mathcal{F}} \circ \boldsymbol{\mathcal{G}}(\boldsymbol{m}_{syn}) - \boldsymbol{m}_{syn}|$ to realize an effective cross-modal alignment.
Training phase-1 We update $\mathcal{G}$ and $\mathcal{H}$ to minimize all the three losses discussed above, *i.e.* $\mathcal{L}_{recon}$, $\mathcal{L}_{distl}$ and $\mathcal{L}_{adapt}$ each with different Adam [17] optimizers.

3.3 Learning beyond the teacher network

We see a clear limitation in the learning paradigm discussed above. The inference performance of the final model is limited by the dataset bias infused in the teacher network. We recognize $\mathcal{L}_{distl}$ as the prime culprit which limits the ability of $\mathcal{H}$ by not allowing it to surpass the teacher's performance. Though one can rely on $\mathcal{L}_{recon}$ to further improve $\mathcal{H}$, this would degrade performance in the inference pathway as a result of increase in discrepancy between v and $\tilde{v}$. Considering this, we propose to freeze $\mathcal{G}$ thereby freezing its output distribution $\mathbb{P}(\tilde{v} = \mathcal{G}(m_{syn}))$ in the second training phase.

Furthermore, in absence of the regularizing $\mathcal{L}_{distl}$ we use a self-supervised consistency loss to regularize $\mathcal{H}$ for the unpaired image samples. For each image I we form a pair (I, I') where $I' = T_s(I)$ is the spatially transformed version (*i.e.* image-flip, random-crop, or in-place rotation) of I. Here, T_s represents the differentiable spatial transformation. Next, we propose a consistency loss based on the equivariance property [49] of the corresponding multi-person 2D pose, *i.e.*

$$\mathcal{L}_{ss} = |T_s \circ \mathcal{T}_K \circ \mathcal{M} \circ \mathcal{H} \circ \mathcal{E}(I) - \mathcal{T}_K \circ \mathcal{M} \circ \mathcal{H} \circ \mathcal{E} \circ T_s(I)|$$

The above loss is computed at the root-locations extracted using the teacher network for the original image I. Whereas, for I' we use the spatial transformation T_s on the extracted root locations of the original image.
Training phase-2 We update the parameters of $\mathcal{H}$ ($\mathcal{G}$ is kept frozen from the previous training phase) to minimize two loss terms *i.e.* $\mathcal{L}_{recon}$ and $\mathcal{L}_{ss}$.

4 Experiments

In this section, we describe the experiments and results of the proposed approach on several benchmark datasets. Through quantitative and qualitative analysis, we demonstrate the practicality and performance of our method.

4.1 Implementation Details

First, we explain the implementation details of synthetic dataset creation. Next, we provide the training details for learning the neural representation.

3D skeleton dataset. *Artificial-pose-sampling* is performed by sampling uniformly from joint wise angle limits defined at local parent relative [2] spherical coordinate system (see Fig. 3A) *i.e.* $[\theta_1, \theta_2]$, and $[\gamma_1, \gamma_2]$. For example, right-hip joint $\theta_1 = \theta_2 = \pi$ (*i.e.* 1-DoF) and $\gamma_1 = \pi/3$, $\gamma_2 = 2\pi/3$ (See Suppl). Using these predefined limits, we construct a full 3D pose (via FK). A total of 1M poses are

Table 2. Quantitative analysis of different ablations of our approach on MuPoTS-3D. *Unpaired* means that there is no ground truth annotation available for an image. *Paired* means that there is a corresponding annotation available for an image. 3DPCK is Percentage of Correct 3D Keypoints predicted within 15cm. (higher 3DPCK is better). "sup." stands for supervision. MuCo-3DHP [34] is used in fifth column. Red color indicates that configuration is less preferable for low data regime. (*Best viewed in color*).

Methods	Artificial	MoCap	Paired **multi**	Composed **multi**	3DPCK
	Poses (Ψ_{arti})	Poses (Ψ_{mocap})	**person** 2D sup.	**person** 3D sup.	
Ours: Learning without any paired supervision. Using 2D predictions from teacher					
$\mathcal{L}_{distl}$ (no $\mathcal{D}_{syn}$)	✔	✗	✗	✗	53.3
$+\mathcal{L}_{recon}$	✔	✗	✗	✗	57.6
$+\mathcal{L}_{adapt}$	✔	✗	✗	✗	61.9
$+\mathcal{L}_{ss}$	✔	✗	✗	✗	64.2
Ours-Us	✗	✔	✗	✗	66.1
Ours: Weakly Supervised Learning Methods. Using paired 2D supervision only					
with Ψ_{arti}	✔	✗	✔	✗	66.4
Ours-Ws	✗	✔	✔	✗	67.9
Ours: Supervised Learning Methods. Using both paired 2D and 3D supervision					
No $\mathcal{D}_{syn}$	✗	✔	✔	✔	71.1
Ours-Fs	✗	✔	✔	✔	75.8

sampled for training Ψ_{arti}. Further, 100k synthetic multi-person pose scenes are created by sampling upto 4 single-person 3D poses per scene. Note that, the $\mathcal{D}_{syn}$ dataset can also utilize 3D poses from single-person 3D dataset such as Human 3.6M [14] and MPI-INF-3DHP [33], when accessible.

Training. First, we train a pose decoder (see Section 3.1.1) either on artificial pose dataset (Ψ_{arti}) or MoCap 3D dataset (Ψ_{mocap}). The AAE modules are trained using a batch size of 32, with a learning rate of 1e-4 using Adam optimizers till convergence (See Suppl). The decoder Ψ is frozen for rest of the training. For training the neural representation, we choose the pretrained network of Cao *et al.* [3] as the teacher network. We consider upto stage-1 "conv5-4-CPM" layer of [3] as $\mathcal{E}$. We concatenate the predictions of both heatmap and Part Affinity Field branches to obtain an embedding space of size 28×28×1024. We consider module $\mathcal{F}$ as from stage-1 "conv5-5-CPM" layer upto stage-2 "Mconv7-stage2" layer of [3]. Using this teacher model, we train the modules $\{\mathcal{H}, \mathcal{G}\}$ by minimizing the losses $\mathcal{L}_{distl}$, $\mathcal{L}_{recon}$, $\mathcal{L}_{adapt}$, $\mathcal{L}_{ss}$ using separate Adam optimizers for each of the losses. We use a learning rate of 1e-4 upto 100k iterations and 1e-5 for the following 500k iterations while using a fixed batch size of 8 throughout the training. Further, we use batches of images from $\mathcal{D}_{syn}$ and $\mathcal{D}_I$ in alternate iterations while training the network. The input image size for $\mathcal{D}_I$ is 224×224×3 and input PAF representation [3] is of shape $28 \times 28 \times 43$ for $\mathcal{D}_{syn}$. All transformations $\mathcal{T}_K$, $\mathcal{T}_R$, $\mathcal{T}_G$ have been implemented using TensorFlow and are designed to be completely differentiable end-to-end. We have trained the entire pipeline on a Tesla-V100 GPU card in Nvidia-DGX station (See Suppl).

Table 3. Comparison of 3DPCK$_{rel}$ on MuPoTS-3D sequences. Our methods are highlighted in gray background color. Underlined values indicate that our unpaired learning (*Ours-Us*) approach performs better on that sequence. *Ours-Fs* (fully-supervised) achieves state-of-the-art in bottom up methods. *Ours-Us* approach performs competitively even when compared with prior fully supervised approaches.

Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg
Accuracy for all groundtruths																					
Rogez[47]	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
Rogez[48]	**87.3**	61.9	67.9	74.6	**78.8**	48.9	58.3	59.7	**78.1**	**89.5**	69.2	73.8	**66.2**	56.0	**74.1**	82.1	78.1	72.6	73.1	61.0	70.6
Dabral[7]	85.1	67.9	**73.5**	**76.2**	74.9	52.5	65.7	63.6	56.3	77.8	76.4	70.1	65.3	51.7	69.5	**87.0**	82.1	**80.3**	**78.5**	**70.7**	71.3
Mehta[34]	81.0	60.9	64.4	63.0	69.1	30.3	65.0	59.6	64.1	83.9	68.0	68.6	62.3	59.2	70.1	80.0	79.6	67.3	66.6	67.2	66.0
Ours-Us	76.8	61.8	61.2	63.0	68.7	20.3	67.3	65.2	59.5	83.6	62.4	66.0	52.7	54.9	57.5	73.6	70.9	70.1	70.4	60.8	63.3
Ours-Ws	79.6	62.3	54.2	55.9	69.3	36.1	69.1	67.7	58.4	80.2	75.3	68.7	53.6	56.5	59.6	77.4	76.7	69.6	69.2	64.1	65.2
Ours-Fs	85.5	**84.1**	66.7	70.5	77.4	**68.6**	**74.8**	**77.9**	69.1	80.0	**78.4**	**75.4**	61.1	**60.9**	71.3	81.4	**85.1**	73.4	74.9	63.5	**74.0**
Accuracy only for matched groundtruths																					
Rogez[47]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Rogez[48]	**88.0**	73.3	**67.9**	**74.6**	**81.8**	50.1	60.6	60.8	**78.2**	**89.5**	70.8	74.4	**72.8**	64.5	**74.2**	84.9	85.2	78.4	75.8	74.4	74.0
Dabral[7]	85.8	73.6	61.1	55.7	77.9	53.3	**75.1**	65.5	54.2	81.3	**82.2**	71.0	70.1	67.7	69.9	**90.5**	**85.7**	**86.3**	**85.0**	**91.4**	74.2
Mehta[34]	81.0	65.3	64.6	63.9	75.0	30.3	65.1	61.1	64.1	83.9	72.4	69.9	71.0	72.9	71.3	83.6	79.6	73.5	78.9	90.9	70.8
Ours-Us	76.8	66.6	62.1	63.9	73.5	20.3	67.3	67.8	59.5	83.6	62.4	66.0	56.0	63.5	59.5	75.2	70.9	73.0	73.1	80.8	66.1
Ours-Ws	79.6	66.0	55.5	56.4	74.8	36.1	69.1	69.6	58.4	80.2	75.3	68.7	56.7	66.4	61.6	78.9	76.7	72.8	71.7	83.0	67.9
Ours-Fs	85.5	**86.5**	66.7	70.5	81.2	**68.6**	74.8	**79.5**	69.1	80.0	78.4	**75.4**	64.0	**68.6**	73.7	82.9	85.1	76.4	77.4	72.8	**75.8**

Table 4. Joint wise analysis of 3DPCK$_{rel}$ on MuPoTS-3D (higher is better). Underlined values indicate that our unpaired learning (*Ours-Us*) performs better on that joint

Methods	Hd.	Nck.	Sho.	Elb.	Wri.	Hip	Kn.	Ank.	Avg
Rogez[47]	49.4	67.4	57.1	51.4	41.3	84.6	56.3	36.3	53.8
Mehta[34]	62.1	81.2	77.9	57.7	47.2	**97.3**	66.3	47.6	66.0
Ours-Us	52.9	79.0	72.2	57.9	45.3	89.9	66.9	45.1	63.3
Ours-Ws	59.9	82.4	78.0	60.6	42.3	91.5	67.2	45.5	65.2
Ours-Fs	**63.4**	**85.5**	**84.2**	**70.4**	**56.8**	95.0	**78.2**	**59.0**	**74.0**

Table 5. We report Camera Centric absolute 3DPCK$_{abs}$ metric on MuPoTS-3D. B/U means Bottom-up. *fps* is runtime frames/second.

Methods	B/U	3DPCK$_{abs}$ (↑)	*fps* (↑)
Moon* [36]	✗	9.6	7.3
Moon [36]	✗	**31.5**	7.3
Ours-Us	✔	23.6	**21.2**
Ours-Ws	✔	24.3	**21.2**
Ours-Fs	✔	28.1	**21.2**

4.2 Ablation Studies

In order to study the effectiveness of our method, we perform extensive ablation study by varying levels of supervision, as shown in Table 2. For all the ablations, we have used MuCo-3DHP images [34] as $\mathcal{I}$. Depending on the supervision setting, we either access none (for unsup. setting), a small fraction (semi sup. setting) or a complete set (full sup. setting) of 3D annotations in MuCo-3DHP dataset.

Ours-Us (Using *Unpaired* images only): Our baseline model (see Table 2) trained without accessing any annotated labels gives an overall 3DPCK of 53.3. We observe that $\mathcal{L}_{adapt} + \mathcal{L}_{ss}$ gives a non-trivial boost of 4-6%. This demonstrates the importance of cross-modal alignment and self-supervised consistency.

Ours-Ws (*Weakly supervised*): When supervised weakly by 2D ground truth ($\mathcal{L}_{2D} = |k_p - \hat{k}_p|$), our approach obtains a 3DPCK of 67.9. Further, the performance of our approach that uses Ψ_{arti} is on par with our performance with Ψ_{mocap} indicating that ϕ_{arti} has rich representation space, equivalent to ϕ_{mocap}.

Ours-Fs (*Fully supervised*): When we access the full training dataset of MuCo-3DHP and impose a 3D reconstruction loss by using $\mathcal{L}_{3D} = |P - \hat{P}|$, we obtain a 3DPCK of 75.8, which is significantly better than the prior arts.

Table 6. Comparison of Absolute MPJPE (lower is better) on Human 3.6M evaluated on S9 and S11. The table is split into three parts: single-person 3D pose estimation approaches (No. 1 to 6), multi-person 3D pose estimation *top-down* approaches (No. 7 to 10), multi-person 3D pose estimation *bottom-up* approaches (No. 11 and 12). Our approach performs better than previous bottom-up multi-person pose estimation methods.

No.	Methods	Dir.	Dis.	Eat	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WaD.	WaP.	Avg
								Single-person approaches									
1.	Martinez [32]	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0	94.6	62.3	78.4	59.1	65.1	49.5	52.4	62.9
2.	Zhou [57]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2	111.6	64.1	65.5	66.0	51.4	63.2	55.3	64.9
3.	Sun [51]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
4.	Dabral [8]	44.8	50.4	44.7	49.0	52.9	43.5	45.5	63.1	87.3	51.7	61.4	48.5	37.6	52.2	41.9	52.1
5.	Hossain [45]	44.2	46.7	52.3	49.3	59.9	47.5	46.2	59.9	65.6	55.8	59.4	50.4	52.3	43.5	45.1	51.9
6.	Sun [52]	47.5	47.7	49.5	50.2	51.4	43.8	46.4	58.9	65.7	49.4	55.8	47.8	38.9	49.0	43.8	49.6
								Multi-person approaches									
7.	Rogez [47]	76.2	80.2	75.8	83.3	92.2	79.0	71.7	105.9	127.1	88.0	105.7	83.7	64.9	86.6	84.0	87.7
8.	Rogez [48]	55.9	60.0	64.5	56.3	67.4	71.8	55.1	55.3	84.8	90.7	67.9	57.5	47.8	63.3	54.6	63.5
9.	Dabral [7]	52.6	61.0	58.8	61.0	69.5	58.8	57.2	76.0	93.6	63.1	79.3	63.9	51.5	71.4	53.5	65.2
10.	Moon[36]	51.5	56.8	51.2	52.2	55.2	47.7	50.9	63.3	69.9	54.2	57.4	50.4	42.5	57.5	47.7	54.4
11.	Mehta[34]	58.2	67.3	61.2	65.7	75.8	62.2	64.6	82.0	93.0	68.8	84.5	65.1	57.6	72.0	63.6	69.9
12.	*Ours-Fs*	55.8	61.4	58.4	71.9	67.6	65.2	67.7	86.7	84.3	68.3	78.9	67.9	51.8	77.9	55.2	67.9

Table 7. 2D keypoint result comparison of our student model with teacher network on MuPoTS-3D. ↑ indicates that higher is better and ↓ indicates that lower is better.

Methods	IoU (↑)	2D-MPJPE (↓)	2D-PCK (↑)
Teacher (Cao [3])	60.1	38.0	66.6
$\mathcal{L}_{distl}$ (no $\mathcal{D}_{syn}$)	51.9	49.6	60.3
Ours-Fs	81.6	19.5	74.7

Table 8. Complexity analysis on MuPoTS-3D. B/U stands for bottom-up approach. ↑ indicates that higher is better and ↓ indicates that lower is better.

Methods	B/U	3DPCK (↑)	*fps* (↑)	Model size (↓)
Mehta [34]	✔	70.8	8.8	> 25.7M
Moon [36]	✗	82.5	7.3	34.3M
Ours-Fs	✔	75.8	21.2	17.1M

4.3 Datasets and Quantitative Evaluation

MuCo-3DHP Training Set and MuPoTS-3D Test Set. Mehta et.al [34] proposed creation of training dataset by compositing images from 3D single-person dataset MPI-INF-3DHP [33]. MPI-INF-3DHP is created by marker-less motion capture for 8 subjects using 14 cameras. MuPoTS-3D [34] is a multi-person 3D pose test dataset that contains 20 sequences capturing upto 3 persons per frame. Each of these sequences include challenging human poses and also capture real world interactions of persons. For evaluating multi-person 3D person pose, 3DPCK$_{rel}$ (Percentage of Correct Keypoints) is widely employed [47,34,36]. In the root-relative system, a joint keypoint prediction is considered as a correct prediction if the joint is present within the range of 15cm. For evaluating absolute location of human joints in camera coordinates, [36] proposed 3DPCK$_{abs}$ in which a prediction is considered correct when the joint is within the range of 25cm. In Table 3 we have compared the results of our method against the state-of-the-art methods. Our fully supervised approach yields state-of-the-art bottom-up performance (75.8 v/s Mehta [34] 70.8) while being faster than the top-down approaches. In Table 4 we present joint-wise 3DPCK on MuPoTS-3D dataset. We compare against [36] on 3DPCK$_{abs}$ metric in Table 5 as it is the only work that reported on 3DPCK$_{abs}$.

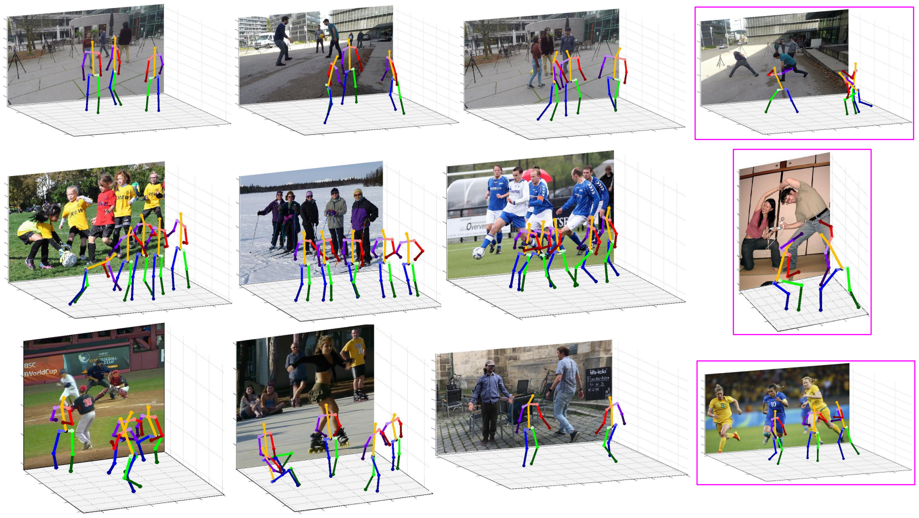


Fig. 5. Qualitative results on MuPoTS-3D (1st row), MS-COCO (2nd row), and "in-the-wild" images (3rd row) of our approach. Our approach is able to effectively handle inter-person occlusion and make reliable predictions for crowded images. Pink box highlights some failure cases. 1st row: presence of self-occlusion, 2nd row: rare multi-person interaction and 3rd row: joint location ambiguity.

Human 3.6M [14] This dataset consists of 3.6 million video frames of single person 3D poses that have been collected in laboratory setting. In Table 6, we show results on Protocol 2: MPJPE calculation on after alignment of root. As shown in Table 6, our approach outperforms bottom-up multi-person works (Mehta [34] 69.9 v/s Ours 67.9) and performs on par with top-down approaches (Rogez [48] 63.5 and Dabral [7] 65.2).

5 Discussion

Fast and accurate inference. In Table 8, we provide runtime complexity analysis of our model in comparison to prior works. All top-down approaches [36,47,48] depend on a person detector model. Hence these methods have low *fps* in comparison to bottom-up approaches (See Fig. 2). We outperform the previous bottom-up approach by a large margin in terms of 3DPCK, *fps* and model size. We achieve a superior real-time computation capability because our approach effectively eliminates the keypoint grouping operation usually performed in bottom-up approaches [3,34]. All *fps* numbers reported in Table 8 were obtained on a Nvidia RTX 2080 GPU. In Table 8, we also show the total number of parameters of the model used during inference time.

Is student network limited by teacher network? In Table 7 we report results of 2D pose estimation on both teacher model ($\hat{k}_q$) and student model ($\hat{k}_p$) by evaluating IoU, 2D-MPJPE and 2D-PCK on MuPoTS-3D dataset. We observe

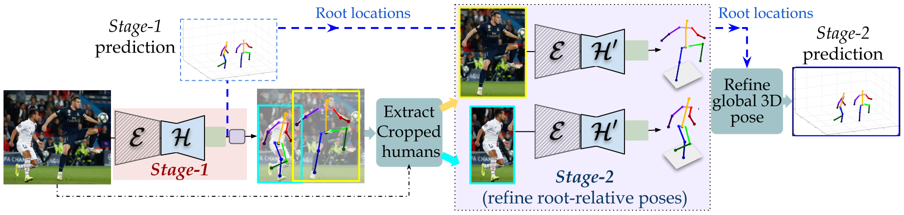


Fig. 6. A hybrid framework for two-stage refinement which treats *Stage-1* output as a person detector while *Stage-2* performs single-person 3D pose estimation.

that a student model trained by minimizing $\mathcal{L}_{distl}$ alone performs sub-optimally in comparison to the teacher. This result is not surprising as the student model is restricted by knowledge of the teacher model. However, in our complete loss formulation (*Ours-Fs*) our approach outperforms the teacher on the 2D task, validating the hypothesis that our approach can learn beyond the teacher network.

Qualitative results. We show qualitative results on the MS-COCO [27], MuPoTS-3D and frames taken from YouTube videos and other "in-the-wild" sources in Fig. 5. As seen in the Fig. 5, our model produces correct predictions on images with different camera viewpoints and on those images containing challenging elements such as inter-person occlusion. These qualitative results show that our model has generalized well on unseen images.

Two-stage refinement for performance-speed tradeoff. Top-down frameworks yield better performance as compared to the bottom-up approach while having substantial computational overhead [18]. To this end we realize a hybrid framework which would provide flexibility based on the requirement. For example, the current single-shot (or single-stage) operates in a substantial computational superiority. To further improve its performance, we propose an additional pass of each detected persons through the full pipeline (Fig. 6). Here, we train a separate $\mathcal{H}'$ for the single-person pose estimation task which is operated on the cropped image patches of single human instances obtained from the *Stage-1* predictions. By training the $\mathcal{H}'$ network we obtain a 3DPCK of 76.9 (v/s *Ours-Fs* 75.8) with a runtime *fps* of 16.6 (v/s *Ours-Fs* 21.2 *fps*). (See Table 3 and Fig. 2)

6 Conclusion

In this paper we have introduced an unsupervised approach for multi-person 3D pose estimation by infusing structural constraints of human pose. Our bottom-up approach has real-time computational benefits and can estimate the pose of persons in camera-centric coordinates. Our method can benefit from future improvements on 2D pose estimation works in a plug-and-play fashion. Extending such a framework for multi-person human mesh recovery and extraction of appearance related mesh texture remains to be explored in future.

Acknowledgement. This project is supported by a Indo-UK Joint Project (DST/INT/UK/P-179/2017), DST, Govt. of India and a WIRIN project.

References

1. CMU graphics lab motion capture database. available: http://mocap.cs.cmu.edu/ 6
2. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR (2015) 3, 5, 6, 9
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017) 4, 7, 8, 10, 12, 13
4. Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S., Rehg, J.M.: Unsupervised 3D pose estimation with geometric self-supervision. In: CVPR (2019) 2
5. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR (2018) 4
6. Chung, Y.A., Weng, W.H., Tong, S., Glass, J.: Unsupervised cross-modal alignment of speech and text embedding spaces. In: NeurIPS (2018) 3
7. Dabral, R., Gundavarapu, N.B., Mitra, R., Sharma, A., Ramakrishnan, G., Jain, A.: Multi-person 3D human pose estimation from monocular images. 3DV (2019) 1, 4, 11, 12, 13
8. Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., Jain, A.: Learning 3D human pose from structure and motion. In: ECCV (2018) 12
9. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: CVPR (2019) 2
10. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: CVPR (2016) 4
11. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. In: ICCV (2017) 4
12. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR (2016) 1
13. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV (2016) 4
14. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. TPAMI **36**(7), 1325–1339 (2013) 2, 10, 13
15. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015) 2
16. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018) 2
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014) 9
18. Kocabas, M., Karagoz, S., Akbas, E.: Multiposenet: Fast multi-person pose estimation using pose residual network. In: ECCV (2018) 2, 4, 14
19. Kundu, J.N., Ganeshan, A., MV, R., Babu, R.V.: Object pose estimation from monocular image using multi-view keypoint correspondence. In: ECCVW (2018) 2
20. Kundu, J.N., Ganeshan, A., MV, R., Prakash, A., Babu, R.V.: iSPA-Net: Iterative semantic pose alignment network. In: ACM Multimedia (2018) 2
21. Kundu, J.N., Gor, M., Babu, R.V.: BiHMP-GAN: Bidirectional 3D human motion prediction gan. In: AAAI (2019) 6
22. Kundu, J.N., Gor, M., Uppala, P.K., Babu, R.V.: Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In: WACV (2019) 6

23. Kundu, J.N., Patravali, J., Babu, R.V.: Unsupervised cross-dataset adaptation via probabilistic amodal 3D human pose completion. In: WACV (2020) 4
24. Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Babu, R.V., Chakraborty, A.: Self-supervised 3D human pose estimation via part guided novel image synthesis. In: CVPR (2020) 4
25. Kundu, J.N., Seth, S., Rahul, M., Rakesh, M., Babu, R.V., Chakraborty, A.: Kinematic-structure-preserved representation for unsupervised 3D human pose estimation. In: AAAI (2020) 4
26. Li, Z., Hoiem, D.: Learning without forgetting. TPAMI **40**(12), 2935–2947 (2017) 2
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 14
28. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML (2015) 3
29. Lopes, R.G., Fenu, S., Starner, T.: Data-free knowledge distillation for deep neural networks (2017) 2
30. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3D pose estimation and action recognition using multitask deep learning. In: CVPR (2018) 1
31. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015) 6
32. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV (2017) 2, 4, 12
33. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D human pose estimation in the wild using improved cnn supervision. In: 3DV (2017) 10, 12
34. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3D pose estimation from monocular rgb. In: 3DV (2018) 1, 2, 4, 10, 11, 12, 13
35. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3D human pose estimation with a single rgb camera. ACM TOG **36**(4), 1–14 (2017) 4
36. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In: ICCV (2019) 1, 2, 4, 11, 12, 13
37. Nayak, G.K., Mopuri, K.R., Shaj, V., Radhakrishnan, V.B., Chakraborty, A.: Zero-shot knowledge distillation in deep networks. In: ICML (2019) 2
38. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: NIPS (2017) 4
39. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016) 4
40. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: ICCV (2019) 4
41. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017) 4
42. Peng, X.B., Andrychowicz, M., Zaremba, W., Abbeel, P.: Sim-to-real transfer of robotic control with dynamics randomization. In: ICRA (2018) 3
43. Pilzer, A., Lathuiliere, S., Sebe, N., Ricci, E.: Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In: CVPR (2019) 4
44. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: CVPR (2016) 4

45. Rayat Imtiaz Hossain, M., Little, J.J.: Exploiting temporal information for 3D human pose estimation. In: ECCV (2018) 12
46. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016) 2, 5
47. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: CVPR (2017) 1, 2, 4, 11, 12, 13
48. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3D pose detection in natural images. TPAMI (2019) 1, 2, 4, 11, 12, 13
49. Schmidt, U., Roth, S.: Learning rotation-aware features: From invariant priors to equivariant descriptors. In: CVPR (2012) 3, 9
50. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: CVPR (2018) 4
51. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV (2017) 4, 12
52. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018) 12
53. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IROS (2017) 3
54. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV (2018) 4
55. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: A dual-source approach for 3D pose estimation from a single image. In: CVPR (2016) 4
56. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: CVPR (2017) 1
57. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017) 12
58. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: ECCVW (2016) 5