






PIRO: Permutation-Invariant Relational Network for Multi-person 3D Pose Estimation

Nicolas Ugrinovic¹^a, Adria Ruiz¹^b, Antonio Agudo¹^c, Alberto Sanfeliu¹^d and Francesc Moreno-Noguer¹^e

¹*Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain*
{nugrinovic, aagudo, asanfeliu, fmoreno}@iri.upc.edu

Keywords: Human pose estimation, 3D, single-view

Abstract: Recovering multi-person 3D poses from a single RGB image is an ill-conditioned problem due to the inherent 2D-3D depth ambiguity, inter-person occlusions, and body truncation. To tackle these issues, recent works have shown promising results by simultaneously reasoning for different individuals. However, in most cases this is done by only considering pairwise inter-person interactions or between pairs of body parts, thus hindering a holistic scene representation able to capture long-range interactions. Some approaches that jointly process all people in the scene require defining one of the individuals as a reference and a pre-defined person ordering or limiting the number of individuals thus being sensitive to these choice. In this paper, we overcome both these limitations, and we propose an approach for multi-person 3D pose estimation that captures long-range interactions independently of the input order. We build a residual-like permutation-invariant network that successfully refines potentially corrupted initial 3D poses estimated by off-the-shelf detectors. The residual function is learned via a *Set Attention* (Lee et al., 2019) mechanism. Despite of our model being relatively straightforward, a thorough evaluation demonstrates that our approach is able to boost the performance of the initially estimated 3D poses by large margins, achieving state-of-the-art results on two standardized benchmarks.


1 Introduction


Estimating 3D human pose from RGB images is a long-standing problem in computer vision, with broad applications in, e.g., action recognition, AR/VR, and human-robot interaction. With the important advancements of single person 3D pose from monocular images (Li and Chan, 2014; Mehta et al., 2017b; Zhou et al., 2017; Pavlakos et al., 2018; Sun et al., 2018; Martinez et al., 2017) as inspiration, the community has focused on extending this success to the multi-person setting. This setting introduces additional challenges to the single person setup due to inter-person occlusions, truncation, and depth/size ambiguity when estimating the root depth of each individual. In order to tackle these problems, sequential (Zanfir et al., 2018a; Cheng et al., 2021) and multi-camera systems (Dong et al., 2019; Lin and Lee, 2021; Tu


et al., 2020; Wu et al., 2021) have been exploited. In contrast, in this paper we aim to address the most constricting version of the problem: multi-people 3D pose estimation from one single view.


On the last couple of years, important advancements have been made in this area (Rogez et al., 2017; Rogez et al., 2019; Moon et al., 2019; Zhen et al., 2020; Wang et al., 2020; Mehta et al., 2020; Jiang et al., 2020; Sun et al., 2021). However, while these approaches have shown impressive results, the problem of multi-person 3D pose estimation remains quite challenging and not yet fully solved. So far, two main different paradigms dominate the existing works: top-down and bottom-up approaches. These approaches have different benefits and strengths, and often they exhibit a trade-off between root-relative pose precision and scale/depth accuracy, among others. None of these approaches fully exploit spatial relationships among individuals which are key to improve reconstructions, as we show in this work.


While it is true that bottom-up approaches are designed to capture and use interactions in a given image to produce 3D pose estimations, these interactions do

^a <https://orcid.org/0000-0002-1823-3780>

^b <https://orcid.org/0000-0001-7210-1378>

^c <https://orcid.org/0000-0001-6845-4998>

^d <https://orcid.org/0000-0003-3868-9678>

^e <https://orcid.org/0000-0002-8640-684X>

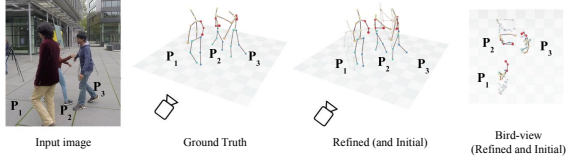


Figure 1: Given a set of potentially noisy input 3D poses, we leverage on the Set Transformer architecture (Lee et al., 2019) to compute a holistic encoding of all poses. This encoding which can take an arbitrarily large number of poses in any order, helps to predict a residual for each pose and refine the initial estimates. The approach is robust to large errors on the initial poses. Note how our refinement corrects the scale and translation of person P_2 . Our model also shows improvements in the root-relative pose (see main text).

not correspond to whole poses or whole individuals. In contrast, they refer to interactions between depths or positions between specific joints. A interesting examples of this approach are (Mehta et al., 2020; Zhen et al., 2020). (Mehta et al., 2020) captures interactions between groups of joints by enforcing a 3D intermediate representation that follows the kinematic chain, limiting these interactions to local regions of the body, while (Zhen et al., 2020) only captures relative-depth relations between joints from image cues. Furthermore, both of these approaches present refinement steps to complete missing occluded joints, refine the position of joints or get an estimate of the absolute depth per individual. They do this, however, in a per-individual basis without taking into account all the persons at the same time. On the other hand, top-down approaches do not take into account interactions at all.

A reduced number of works have proposed strategies to better exploit multi-person relations. For instance, (Wang et al., 2020) and (Jiang et al., 2020) try to remedy top-down approaches lack of global context by using ordinal depth losses between individuals. However, they use only ordinal information and reason about multiple people in a local neighborhood in a strictly pairwise manner. There are other works that acknowledge this limitation and similar to us they exploit the spatial relation among all individuals in the image. However, their applicability is limited by their design choices. For example, (Guo et al., 2021) is sensitive to permutations in the input order as it uses RNNs as their main architecture. (Cha et al., 2022) limit their model to a maximum of 3 persons, and (Fieraru et al., 2021) rely on direct supervision for modeling close-contact human interactions which comes from time-consuming and hard-to-obtain annotations.

In order to overcome the limitations of previous approaches, we propose a novel scheme to model people interactions in a holistic and permutation-invariant

fashion. We get inspiration from (Guo et al., 2021; Fieraru et al., 2021; Wang et al., 2020) that show that it is possible to exploit spatial information between individuals. We are equally inspired by from (Cha et al., 2022; Zhen et al., 2020; Mehta et al., 2020) that show that capturing interactions is possible and that refinement steps in a pose estimation pipeline are important to overcome challenges such as occlusion, truncation, and improved depth prediction. Furthermore, we build on top of a principled approach (Lee et al., 2019) and take advantage of two key characteristics of transformer attention models: their impressive capability of capturing relations between its inputs and their permutation-invariant property (set attention). This allows us to simultaneously process the poses of all individual and exploit contextual information in the form of intra and inter-person relationships. Figure 1 shows how our model refines the initial pose estimations and yields more correct ones, even under occlusions. By capturing the interactions between input poses, our model can improve the pose, translation, and scale of the people in the scene.

We thus pose our model as a refinement network capable of exploiting the spatial relations among all individuals. Our key insight is that people that share a common activity often have similar or correlated poses. We model an individual’s pose as an entity consisting of joints and consider the whole pose of a person to exploit global information and inter-person relations to improve initial 3D pose estimations. The approach is relatively straightforward, yet achieves results that significantly outperform the latest state of the art. Interestingly, our model runs efficiently and works with both top-down and bottom-up approaches, thus can be potentially used as a post-processing module on top of any pose estimation method with negligible computation overhead.

We perform extensive experiments on on MuPoTS-3D (Mehta et al., 2018), Panoptic (Joo et al., 2015) and show our model’s capabilities to perform under specific scenarios on the NBA2K (Zhu et al., 2020) dataset. We also carefully ablate the model’s capabilities of capturing interactions and present an extensive analysis of these interactions. In summary, our key contributions are the following: (1) We introduce a novel approach to capture the relationship among the 3D poses of multiple people. (2) Our model does not depend on the input order (i.e., permutation-invariant) and can handle an arbitrarily large number of people. (3) We present extensive experiments and analysis that validate our approach. (4) The proposed module is computationally efficient and could potentially be used along with any 3D pose detector.

2 Related work

Multi-person 3D Pose Estimation. We focus on the use of monocular static images. In this area, there exist two types of approaches. First, top-down (Rogez et al., 2019; Moon et al., 2019; Lin and Lee, 2020; Dabral et al., 2019) lead to more accurate root-relative pose results but are more sensitive to inter-person occlusions and truncation as they discard contextual information and focus individually on each person. Second, bottom-up (Mehta et al., 2017b; Mehta et al., 2018; Mehta et al., 2017a; Mehta et al., 2020; Zhen et al., 2020; Zhang et al., 2022; Qiu et al., 2022; Liu et al., 2022) which are more robust to occlusions and truncation as they use global reasoning over image information. However, they suffer scale variations and pose accuracy is compromised. Recently, a new trend to integrate both approaches has emerged (Wang et al., 2020; Khirodkar et al., 2022; Wang et al., 2022a; Jin et al., 2022).

Human-Human Interaction and Context. The idea of using human-human interaction information to improve 3D human pose estimation was first proposed by (Andriluka and Sigal, 2012). However, it was not until recently that the community shifted its attention to exploiting this information. (Jiang et al., 2020) and (Wang et al., 2020) exploit depth-order relationships between people. Though it has been shown that depth ordering losses help improve 3D human pose estimation (Pavlakos et al., 2018), they disregard the magnitude. As a consequence, this type of supervision is mostly coarse. (Cheng et al., 2021) propose a pose discriminator to capture two-person interactions. However, aside from using video as input, the discriminator captures interactions of only two people at a time. Our model captures the interaction information of all people in the scene at the same time. Closest to our approach, (Guo et al., 2021) and (Fieraru et al., 2021) propose interacting networks to exploit human-to-human interaction information.

Permutation invariant models We capture the interaction of people in a permutation-invariant manner, thus the order we input each person to our model does not matter. Therefore, we treat all the poses as elements of a set. (Santoro et al., 2017) proposes a relational network that allows capturing pairwise interactions of elements in a given set. However, we want to model higher-order interactions as the pose of a person may affect (or be affected by) not one but multiple other persons directly or indirectly. (Ma et al., 2018) use a Transformer (Vaswani et al., 2017) to model high-order interactions between objects. However,

they use mean-pooling to obtain aggregated features where interaction information may be lost. More suited to our problem, we choose (Lee et al., 2019) as our base architecture.

3 Method

The key idea of our approach is to implicitly capture the interaction information between all human body poses in a scene and use it to refine an initial set of 3D pose estimations. We represent this information with an *interaction embedding*. People in a scene are usually involved in a specific interaction and this constrains the range of possible poses. Thus, we argue that learning pose correlations can improve initial noisy estimations. Inspired by the effectiveness of Transformers (Vaswani et al., 2017) to capture correlations, we use the Set Transformer (Lee et al., 2019), a model unaffected by the order of its inputs. In short, we obtain an information-rich and permutation-invariant interaction embedding that captures human-to-human interactions and use it to refine each person’s pose.

3.1 Relational Network for Multi-person 3D Pose

Let $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ be an input RGB image with N people interacting in the same scene and $p^{1:N}$ to be the set of 3D joints corresponding to each person where $p^n \in \mathbb{R}^{J \times 3}$ with J number of estimated joints and n a number in the range $\{1, \dots, N\}$. These joints are obtained from an initial estimation using an off-the-shelf 3D pose estimator, such as (Moon et al., 2019). All N poses are assumed to be represented in absolute camera coordinates.

Given the previous definitions, we aim to improve each initial pose estimation p^n taking into account the pose of all the people present in the scene. These initial estimations could be inaccurate as they are sensitive to inter-person occlusions, self-occlusions, scale/depth ambiguity, and truncation. Although the latter does not originate from a lack of interaction information, we shall see that our model also deals with these cases. This is because, aside from modeling interaction information, our model also captures the error distribution of the initial estimation method.

The mapping between these initial estimations and the refined poses takes the form $q^{1:N} = \Phi(p^{1:N})$, where $q^{1:N}$ refers to the set of poses $\{q^1, \dots, q^n\}$ refined by exploiting the interactions, and $q^n \in \mathbb{R}^{J \times 3}$. The function Φ materializes as a neural-network capable of extracting interaction information between

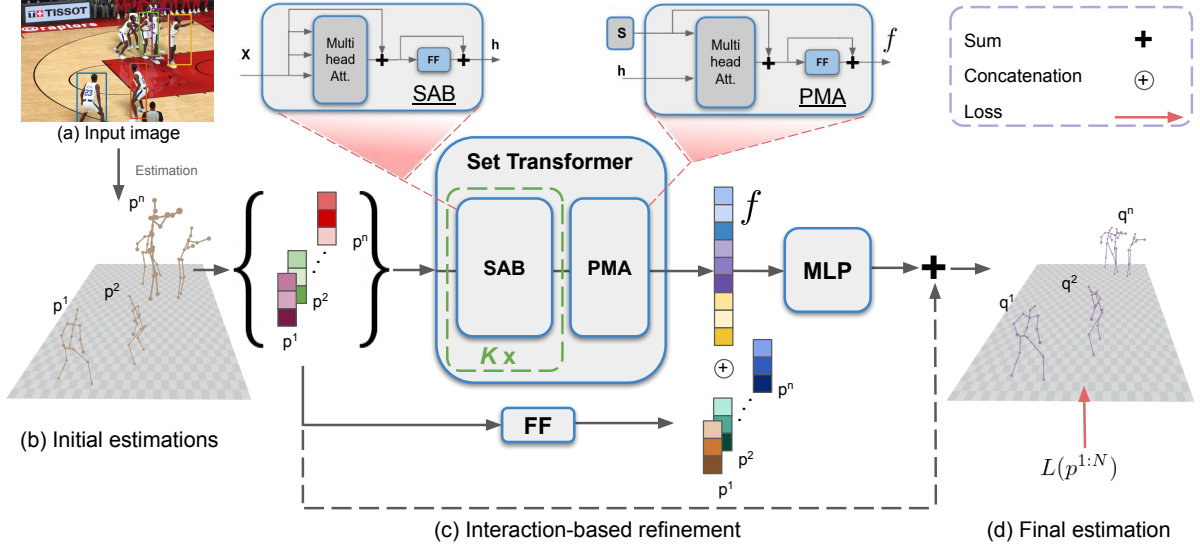


Figure 2: **Overview of our approach.** Given an input image (a), we first estimate the 3D keypoints as the initialization (b). Then, we input these initial estimations in the form of a set (hence the keys) to our interaction-based permutation-invariant model. We obtain the *interaction-based embedding* (f) and concatenate it with another embedding for each person. This *person* embedding is calculated as a projection from the input space to the same dimension as f via a feed-forward layer, denoted FF. We use an MLP network to get the corrections of the initial estimations (b) and compute the final estimations (d) by adding these corrections to the initial poses. We show the poses of the people with bounding boxes in the image just for clarity. Our model inputs all the poses in the scene.

the input poses $p^{1:N}$ and refining them.

For this purpose, we split the problem into two parts: First, we obtain an embedding (*interaction embedding*) able to capture the interactions of the scene; second, we use this to refine the initial poses. The *interaction embedding* f is a d -dimensional vector where d is a hyperparameter of our model. The embedding f , obtained via Set Transformer blocks, aims to globally capture interactions between people from the initial estimations $p^{1:N}$. Specifically,

$$f = \mathcal{G}(p^{1:N}), \quad (1)$$

where \mathcal{G} is a neural network composed of the Set Transformer elements described in the next section. Ideally, we are looking for a function \mathcal{G} capable of capturing the embedding regardless the input order and the number of people in the scene. If we explicitly take into account the *interaction embedding*, we can express the relation between the initial and refined estimations as:

$$q^{1:N} = \Phi(f, p^{1:N}), \quad (2)$$

where Φ is our full model described in more detail in Sec. 3.3 and depicted in Fig. 2.

3.2 Computing interaction embeddings

As stated before, our *interaction embedding* f should comply with two key requirements to model the interaction information: (1) being independent of the order of the input person’s body joints and (2) being able to process input scenes containing any number of people. Requirement (1) comes from the fact that we do not want the input order of our model to affect the pose refinement. Only the information regarding interaction between people’s body poses should affect the refinement. Both requirements are not easily satisfied by classical feed-forward neural networks, and recursive neural networks (RNNs) are sensitive to the input order (Vinyals et al., 2015). Thus, to get the desired interaction embedding, we base our model in an attention-based permutation-invariant neural network. For this purpose, we take components from the Set Transformer (Lee et al., 2019). Choosing an attention-based architecture for our model and working with sets as inputs allows us to: compute a variable number of input poses, disregard the input order, and naturally encode interaction between these elements. In this manner, we are able to attend to any person’s initial pose and obtain a rich permutation-invariant embedding that captures the interactions in the scene. We then use this feature to guide the pose

refinement process.

In our context, we treat the initial pose estimations $p^{1:N}$ as a set ((b) in Fig. 2). Thus, our model attends to each person’s joints and first generates an embedding for each person using a Set Attention Block (SAB). Later, these individual embeddings are aggregated in a learned fashion using a Pooling by Multi-head Attention (PMA) operation, providing us the interaction embedding f . For a formal definition of the SAB and PMA modules, we refer the reader to (Lee et al., 2019).

The SAB module used here emerges as an adaptation of the encoder block of the Transformer (Vaswani et al., 2017). To build the SAB, dropout and the positional encoding are discarded. This module uses self-attention to concurrently encode the input set. This allows to capture pairwise and higher-order relationships among instances during the encoding process. The output of the SAB contains information about pairwise interactions among the elements of the input set X and can be stacked K times by more of the same modules to capture higher than pairwise interactions. In our context, X is composed by the set of initial pose estimations $p^{1:N}$, thus, $X = \{p^1, \dots, p^n\}$.

To obtain a permutation-invariant feature, we use the Pooling by Multihead Attention (PMA) operation which aggregates the features obtained by the SAB. This constitutes a key step to make f permutation-invariant. The PMA operation aggregates the features by applying multi-head attention on a learnable set of k seed vectors $S \in \mathbb{R}^{k \times d}$. In our case, $k = 1$, as we only have one embedding to represent the whole scene. At the output of the PMA block, we find the desired *interaction embedding*. Under the definitions provided before, we define our embedding in the following manner:

$$f = \text{PMA}_1(\text{SAB}(X)). \quad (3)$$

3.3 Pose refinement via interaction information

Once we have the feature f , we use them to refine the initial 3D human pose estimations. For this, we employ an MLP that takes as input f concatenated with a projection of each person’s joints into a d -dimensional vector. This projection is done via a feed-forward layer, denoted FF. Finally, the MLP outputs a vector (δ) containing all the correction values needed to improve each of the initial joints locations in 3D space. We define correction vector as:

$$\delta^{1:N} = \Psi(f, p^{1:N}), \quad (4)$$

where our network Ψ is based on the SAB/PMA modules and the MLP in charge of decoding the interac-

tion embedding. The set-processing modules (SAB and PMA) generate the interaction embedding, as considered in Eq. ((1)). Adding this correction vector to our initial estimations, we can now compute the refined joints by the following relation:

$$q^{1:N} = p^{1:N} + \delta^{1:N}. \quad (5)$$

To guide the learning process, we optimize the whole network parameters by minimizing an L_2 loss over the final refined 3D joints and the ground truth joints:

$$L(p^{1:N}) = \frac{1}{N} \sum_{i=1}^N \|q^i - p_{GT}^i\|^2, \quad (6)$$

where p_{GT} denotes the 3D human pose ground truth. For more details regarding our final architecture, please refer to the supplementary material and the code of this paper.

4 Experiments

In this section, we present implementation details, the evaluation of our approach in comparison with other relevant SOTA, and an ablation study focusing on various types of interactions. Finally, we present an analysis of the refinement process and the computational complexity of our model. Our model has the advantage of being highly computationally efficient, lightweight and fast to train. We experiment on three datasets: MuPoTS-3D (Mehta et al., 2018), Panoptic (Joo et al., 2015) and NBA2K (Zhu et al., 2020). Additionally, we include qualitative results on COCO (Lin et al., 2014). We also use standard metrics for evaluation such as **MPJPE** (Ionescu et al., 2014) –which measures the accuracy of the 3D root-relative pose– and **3DPCK** (Mehta et al., 2016) with a threshold of 15cm, as it is standard in the literature (Guo et al., 2021; Zhen et al., 2020; Lin and Lee, 2020). Complementary to 3DPCK (from now on PCK), we use **AUC** (area under the curve) as a more complete metric. Additionally, **PCK_{abs}** is used to evaluate absolute camera-centered 3D human poses.

4.1 Implementation details

We optimize our model parameters using ADAM (Kingma and Ba, 2015) with a learning rate of 0.0001 in a single GTX 1080 Ti. To estimate the initial 3D human poses we use (Moon et al., 2019). Regarding training data, although generally used for this task, we discard the use of MuCo-3DHP. Given its synthetic nature, it does not contain real interactions between people. Instead,

Metric	All people			Matched people			Datasets training
	$PCK_{rel} \uparrow$	$PCK_{abs} \uparrow$	$AUC_{rel} \uparrow$	$PCK_{rel} \uparrow$	$PCK_{abs} \uparrow$	$AUC_{rel} \uparrow$	
SMAP (Zhen et al., 2020)*	73.5	35.4	-	80.5	38.7	42.7	MuCo-3DHP + COCO
RootNet (Moon et al., 2019)*	81.2	31.4	39.5	82.5	32.0	40.2	MuCo-3DHP + COCO
PI-Net (Guo et al., 2021)(w/ RootNet (Moon et al., 2019))*	82.5	-	-	83.9	-	-	MuCo-3DHP
Ours (w/ SMAP (Zhen et al., 2020))	79.3	40.8	-	86.1	44.2	44.3	MuPoTS Cross Validation
PI-Net (Guo et al., 2021)(w/ RootNet (Moon et al., 2019)) [†]	82.8	-	43.9	84.3	-	44.7	MuPoTS Cross Validation
Ours (w/ RootNet (Moon et al., 2019))	85.8	44.1	46.1	87.3	45.0	46.9	MuPoTS Cross Validation

Table 1: Quantitative comparison on the **MuPoTS-3D** (Mehta et al., 2018) dataset. Both PI-Net and our method use (Moon et al., 2019) for initialization. Additionally, we show our model’s performance when initialized with (Zhen et al., 2020). *Results shown for these methods are merely referential as they are not re-trained with the same data as ours. [†] Fine-tuned.

we use **MuPoTS-3D** which does contain real interactions. For fair comparison, we resource to k-fold cross-validation dividing the dataset into 10 folds, which is an accepted practice in the machine learning literature given a limited dataset. For evaluating on the **Panoptic** studio (Joo et al., 2015) dataset, we follow the evaluation protocol presented in (Zanfir et al., 2018a; Zanfir et al., 2018b). Finally, even though the **NBA2K dataset** is synthetic, it captures plausible interaction between players as opposed to MuCo-3DHP. For more details please refer to the supplementary material.

4.2 Comparison with state-of-the-art methods

We present a direct comparison with the method closest to ours, PI-Net (Guo et al., 2021), and show as reference other SOTA methods that also deal with multi-person 3D pose estimation (Wang et al., 2020; Lin and Lee, 2020; Zhen et al., 2020; Zanfir et al., 2018b). The quantitative results for **MuPoTS-3D** dataset are reported in Table 1. Here, we show results with both the initialization methods RootNet (Moon et al., 2019) and SAMP (Zhen et al., 2020). RootNet is also used by PI-Net (Guo et al., 2021) as initialization. We present, two rows referencing PI-Net (Guo et al., 2021). The first one, shows the results when training the model with MuCo-3DHP (Mehta et al., 2018) dataset, as reported in their work. For a fair comparison, we fine-tune the model with the MuPoTS-3D (Mehta et al., 2018) dataset and perform the same cross validation. We do not present other SOTA methods in this table because they were not trained with the same data which would lead to an unfair comparison. However, we present a direct comparison with them in table 2. As it can be seen, our model shows a 3.0% improvement over PI-Net when estimating the root-relative pose, 2.2% for AUC for all people and 2.1% (matched people). Also, worthy of notice, we remarkably outperform RootNet (Moon et al., 2019) in all metrics. Furthermore, we also show improvements

when using SMAP as initialization. While this version of our model improves notably w.r.t. SMAP, we note that the best results come from using RootNet for initialization. We argue that this is due that top-down approaches benefit more from additional global inter-person context as they do not originally exploit this information. In contrast, bottom-up approaches benefit in a less degree as they already incorporate different forms of contextual information.

The results for the **CMU Panoptic** dataset are shown in Table 2. We evaluate our method under MPJPE after root alignment following previous works (Zanfir et al., 2018a; Zanfir et al., 2018b). The dataset presents a challenging scenario as the majority of images contain several people at a time in a closed environment, severely affected by occlusion and truncation. Our method successfully reduces the interference of occlusions and truncation and improves by a large amount the initial estimations (fine-tuned RootNet (Moon et al., 2019)). To show how our model is able to deal with truncation, results of the metric calculated over all joints in the dataset are presented in Table 2, including those that are out of the image and are not visible. Most of these non-visible joint constitute cases of either truncation or occlusion. Our method improves over 30 mm. in average over the initial method when initialized with RootNet. Note that we also have a significant improvement (14.1 mm.) over both RootNet and 8.6mm over SMAP as the initial method if we account only for visible joints which is the standard practice. With regards to the SOTA in this dataset (HMOR (Wang et al., 2020)), we outperform the method by an overall of 5.3 mm. and 4mm. over DAS (Wang et al., 2022b) when initializing with RootNet and have a consistent improvement over all the actions. When we initialize with SMAP, we have an overall improvement 1mm. over HMOR and 0.6mm. over DAS. For qualitative comparisons on the Panoptic (Joo et al., 2015) dataset and using SMAP as initialization, please refer to Fig. 5 and the supplementary material.

For evaluating on the **NBA2K dataset**, we use the MPJPE without and with Procrustes Alignment

Method	Haggling	Mafia	Ultim.	Pizza	Mean ↓
RootNet (w/ all joints)	83.3	107.9	106.0	118.4	103.9
Ours (w/ all joints)	59.4	68.9	67.2	86.4	70.5
Zanfir <i>et al.</i> (Zanfir <i>et al.</i> , 2018b)*	72.4	78.8	66.8	94.3	78.1
RootNet (Moon <i>et al.</i> , 2019)	52.1	65.3	58.0	80.4	63.9
SMAP (Zhen <i>et al.</i> , 2020)	63.1	60.3	56.6	67.1	61.8
HMOR (Wang <i>et al.</i> , 2020)*	50.9	50.5	50.7	68.2	55.1
Liu <i>et al.</i> (Liu <i>et al.</i> , 2022)	55.2	55.0	50.4	61.4	55.0
Jin <i>et al.</i> (Jin <i>et al.</i> , 2022)	63.7	58.5	52.3	69.1	60.9
DAS (Wang <i>et al.</i> , 2022a)	53.3	51.2	49.1	61.5	53.8
Pi-Net (w/ RootNet (Moon <i>et al.</i> , 2019))	51.3	66.3	56.2	76.1	62.5
Ours (w/ SMAP (Zhen <i>et al.</i> , 2020))	49.3	53.7	48.7	61.2	53.2
Ours (w/ RootNet (Moon <i>et al.</i>, 2019))	42.0	50.3	47.3	59.4	49.8

Table 2: Evaluation on the **Panoptic** (Joo *et al.*, 2015) dataset. RootNet (Moon *et al.*, 2019) model was fine-tuned with CMU Panoptic data to provide a better initialization. The reported metric is **MPJPE** relative to the root joint and results are reported in mm. *The average of (Zanfir *et al.*, 2018b) and (Wang *et al.*, 2020) are recalculated following the standard practice in (Zanfir *et al.*, 2018a) and (Zhen *et al.*, 2020) (i.e. average over activities) for a more direct comparison.

(MPJPE-PA) as shown in Table 3. See how both methods that use interaction information from the scene (PI-Net and ours) are able to improve the results over the initial estimations (RootNet (Moon *et al.*, 2019)).

4.3 Interaction vs. no interaction

Having shown the effectiveness of our method at refining 3D poses, we continue with a careful analysis of our interaction component. Table 4 shows how the level at which we enforce the interaction to be learned affects the performance in comparison to the initial estimations. We define three different levels of interaction: (1) no interaction, (2) scene interaction, and (3) people interaction. The latter corresponds to our final method. For all the cases, we use the same architecture. However, we change the interaction levels by changing what we input to our method. To eliminate learning interactions (*no interaction*), we input each person’s pose individually as a unique and different set and not together. In this manner, it is impossible for the model to build an interaction-based embedding. At most, the model remains restricted to capture self-joint interactions. To enforce learning what we refer to as *scene interaction*, we make each joint in the scene a set by itself. Having each joint as set element instead of the whole person’s pose, we enforce a representation that can learn interactions between joints but without knowing which joint corresponds to which person. Thus, losing the sense of person as an “entity”. The results from Table 4 show that our method based on people’s interaction clearly outperforms other degrees of interaction consistently over both datasets. We report the results on the MuPoTS-3D and Panoptic datasets and use the MPJPE metric

only for simplicity.

Figure 3 gives us insight on what happens with the refinement on each level of interaction. Here, we present two images containing three persons each: high interaction between individuals (top image), low interaction (bottom image). Additionally, for each image we show three matrices, each regarding an interaction type presented above. Each matrix depicts the effect of the pose refinement when perturbing one joint over all of the joints of every person in the scene. This also includes the person whose joint has been perturbed. Each column represents the joint being perturbed and each row represents the affected joints. The perturbation applied to the joints is a displacement in the positive directions of x, y and z in the 3D space by 10 cm. The magnitude in each element in the matrix represents the maximum absolute value of the change in any of the 3D space coordinate direction in meters. With this setup, we can study how one person’s joint affects other person’s joints as well as its own. Here, we notice some key observations. (1) for the cases in which we enforce to learn interactions (people and scene), the effect of one person’s joint over other person’s joints (effect of interaction) is higher for the image on the top (high interaction) than for the one in the bottom (lower interaction). (2) in the case of people interaction, it can be clearly seen that one person’s joint affects in greater degree the pose of its own body in contrast to other people’s body. This is expected, as the model here has the notion of which joints correspond to which person. This does not happen in the scene interaction case. Also, (3) we can confirm that in the case of no-interaction, one person’s pose has no effect over others. The reader is referred to the supplemental material for additional examples.

4.4 Effects of the refinement over initial estimations

We show the effect of our model in refining the initial estimations. Our method can improve both absolute and root-relative poses while more effectively dealing with inter-person occlusions and truncations. This is achieved because our *interaction embedding* enables the model to reason directly in the 3D space, whereas other methods can only reason from 2D image cues. Furthermore, our loss encourages the model to learn a refinement for both the absolute and the root-relative poses. See Fig. 4. The three middle columns depict the initial estimation, the refined poses and the ground truth from a slightly rotated camera view along with a bird-view, respectively. From these views, we can appreciate the interaction between person 1 (P_1) and

Method	MPJPE [mm]					MPJPE-PA [mm]				
	Cory	Glen	Oscar	Tomas	Mean ↓	Cory	Glen	Oscar	Tomas	Mean ↓
RootNet (Moon et al., 2019)	154.3	167.7	159.3	136.2	154.1	115.8	137.5	122.4	103.9	119.7
Pi-Net [†] (Guo et al., 2021) (w/ RootNet (Moon et al., 2019))	136.6	155.3	140.2	119.8	137.8	109.7	129.2	111.5	96.2	111.5
Ours (w/ RootNet (Moon et al., 2019))	130.0	142.0	134.7	121.7	131.9	99.6	111.5	104.4	95.8	102.7

Table 3: Evaluation on the **NBA2K** (Zhu et al., 2020) dataset. We use the **MPJPE** metric. [†] This method has been fine-tuned with the same dataset and uses the same initial method (RootNet (Moon et al., 2019)) for fair comparison.

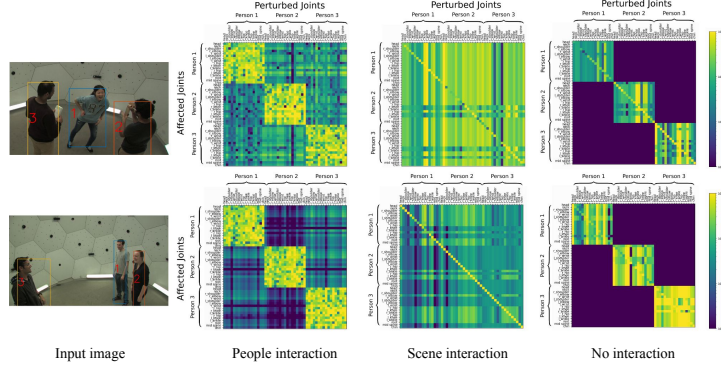


Figure 3: **Interaction analysis.** We show the effect of one joint over all other joints in the scene. Joints are grouped by person. We present 17 joints for each person. Each person’s number in the matrices corresponds to the number shown in the bounding box in the images. The magnitude of each matrix element represents the maximum displacement in 3D space measured in meters of the joints in each row caused by the corresponding joint in each column.

Dataset	Metric	MPJPE [mm]	MPJPE-PA [mm]
MuPoTS-3D	Initial (RootNet)	134.9	93.3
	Baseline – no interaction	136.4	95.1
	Baseline – scene interaction	134.5	96.1
	Ours – people interaction	104.8	79.7
CMU Panoptic	Initial (RootNet)	63.9	54.6
	Baseline – no interaction	57.0	47.4
	Baseline – scene interaction	58.2	45.1
	Ours – people interaction	49.8	40.5

Table 4: Importance of different levels of **interaction** in our model

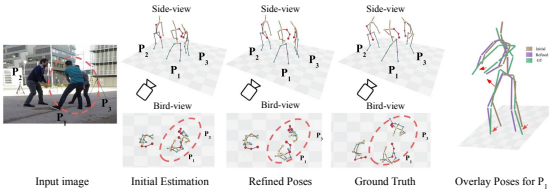


Figure 4: **Effects of the pose refinement.** From left to right: Input image, initial 3D pose estimations, refined poses, ground truth, and detail of the update on person P_1 ’s joints. For each estimation we include a bird-view so that absolute translation is better appreciated. The last column shows the root-relative pose improvement.

person 3 (P_3). The initial estimation does not take into account this interaction and, therefore, makes the mistake of overlapping the two bodies. Our model yields to more realistic estimations by exploiting these interactions. The rightmost picture shows the initial, refined and ground truth root-relative poses for P_1 . See how our estimations correct the initial joint positions taking them closer to the ground truth (highlighted in

red arrows). For example, the hands and ankle joints are closer to the ground truth than the initial estimations. The same happens with the hip joints. Additionally, Fig. 1 shows the input image, the ground truth and the refined poses overlapped with the initial estimations (in transparency) both in a tilted camera view and a bird-view. Here, we can also appreciate an interaction between the people that are about to hug (P_2 and P_3). Better seen in the bird-view, our refinement locates both persons in a more coherent way, whereas, the initial estimation places them further apart.

4.5 Qualitative Results

Qualitative results on the MuPoTS-3D, Panoptic, NBA2K and COCO datasets are shown in Fig. 5. Row 1 first column shows a case where people are closely interacting with each other (holding hands). Our model corrects the persons poses so their hands are closer together. Row 1 second column shows how our method corrects cases of severe truncation. Row 2 shows results on the NBA2K (Zhu et al., 2020) dataset and the last row shows as well results on images-in-the-wild from COCO (Lin et al., 2014) dataset. NBA2K (Zhu et al., 2020) presents several interactions in each scene. We show how our method improves over the SOTA (Moon et al., 2019), especially in cases where people need to be grouped closely to-

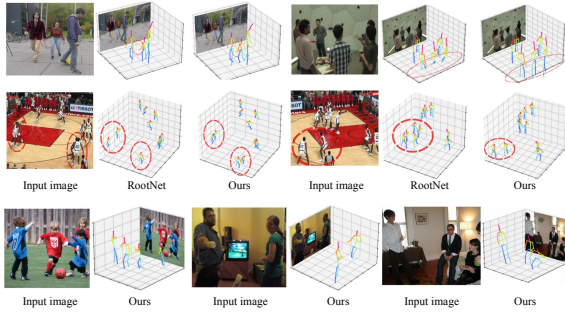


Figure 5: **Qualitative results.** We show results on the MuPoTS-3D (Mehta et al., 2018), Panoptic (Joo et al., 2015), NBA2K (Zhu et al., 2020) and COCO (Lin et al., 2014) datasets. Here we show cases of close interactions (first two rows), severe truncation (first row, second column), and our model on images in-the-wild (last row).

gether. Our method captures the interactions between the players and can determine which players should be grouped together. In each case, with dotted red circles, we show either a correctly located group of players (refined poses) or incorrectly placed players (initial poses).

4.6 Limitations

While we show consistent improvements over different initialization methods, our model is limited to these initializations. For example, our model is not able to recover the pose of an individual if this was not previously detected by the initial method as we based the refinements in spatial information and not image cues. The same happens when the initial estimates are severely corrupted. Handling severe occlusions is a very challenging task that can be attacked by using additional information such as temporal sequences, however, this requires a different approach and design and constitutes future work.

5 Conclusions

In this paper, we have proposed a novel algorithm to tackle the problem of multi-person 3D pose estimation from one single image. Building on the Set Transformer paradigm, we have introduced a holistic encoding of the entire scene, given an initial set of potentially noisy input 3D body poses. This encoding captures multi-person relationships, does not depend on the input order, and can represent an arbitrarily large number of inputs. We use it to refine the initial poses in a residual manner. A thorough evaluation shows that our approach provides state-of-the-art results on several benchmarks. Additionally, the pro-

posed module is computationally efficient and can be used as a post-processing step for any 3D pose detector in multi-person scenes to improve its accuracy and make it more robust to truncation and occlusions.

Acknowledgments

This work is supported by the project MoHuCo PID2020-120049RB-I00 funded by MCIN/AEI/10.13039/501100011033.

REFERENCES

- Andriluka, M. and Sigal, L. (2012). Human context: Modeling human-human interactions for monocular 3d pose estimation. In *International Conference on Articulated Motion and Deformable Objects*, pages 260–272. Springer.
- Cha, J., Saqlain, M., Kim, G., Shin, M., and Baek, S. (2022). Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 660–677. Springer.
- Cheng, Y., Wang, B., Yang, B., and Tan, R. T. (2021). Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7649–7659.
- Dabral, R., Gundavarapu, N. B., Mitra, R., Sharma, A., Ramakrishnan, G., and Jain, A. (2019). Multi-person 3d human pose estimation from monocular images. In *2019 International Conference on 3D Vision (3DV)*, pages 405–414.
- Dong, J., Jiang, W., Huang, Q., Bao, H., and Zhou, X. (2019). Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7792–7801.
- Fieraru, M., Zanfir, M., Szente, T., Bazavan, E., Olaru, V., and Sminchisescu, C. (2021). Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. In *Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021)*.
- Guo, W., Corona, E., Moreno-Noguer, F., and Alamedd-Pineda, X. (2021). Pi-net: Pose interacting network for multi-person monocular 3d pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2796–2806.
- Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large Scale Datasets and Predictive Methods for 3d Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

- Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., and Daniilidis, K. (2020). Coherent reconstruction of multiple humans from a single image. In *CVPR*.
- Jin, L., Xu, C., Wang, X., Xiao, Y., Guo, Y., Nie, X., and Zhao, J. (2022). Single-stage is enough: Multi-person absolute 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13086–13095.
- Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In *ICCV*.
- Khrodar, R., Tripathi, S., and Kitani, K. (2022). Occluded human mesh recovery. *arXiv preprint arXiv:2203.13349*.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. (2019). Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR.
- Li, S. and Chan, A. B. (2014). 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer.
- Lin, J. and Lee, G. H. (2020). Hdnet: Human depth estimation for multi-person camera-space localization. In *ECCV 2020*, page 633–648.
- Lin, J. and Lee, G. H. (2021). 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.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Liu, Q., Zhang, Y., Bai, S., and Yuille, A. (2022). Explicit occlusion reasoning for multi-person 3d human pose estimation. In *European Conference on Computer Vision*. Springer.
- Ma, C.-Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., and Graf, H. P. (2018). Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800.
- Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2016). Monocular 3d Human Pose Estimation In The Wild Using Improved CNN Supervision. *arXiv:1611.09813 [cs]*.
- Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2017a). Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE.
- Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.-P., Rhodin, H., Pons-Moll, G., and Theobalt, C. (2020). Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Trans. Graph.*
- Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. (2018). Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE.
- Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., and Theobalt, C. (2017b). Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14.
- Moon, G., Chang, J., and Lee, K. M. (2019). Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *The IEEE Conference on International Conference on Computer Vision (ICCV)*.
- Pavlakos, G., Zhu, L., Zhou, X., and Daniilidis, K. (2018). Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Qiu, Z., Yang, Q., Wang, J., and Fu, D. (2022). Dynamic graph reasoning for multi-person 3d pose estimation. *arXiv preprint arXiv:2207.11341*.
- Rogez, G., Weinzaepfel, P., and Schmid, C. (2017). Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441.
- Rogez, G., Weinzaepfel, P., and Schmid, C. (2019). Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*.
- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.
- Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y. (2018). Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Sun, Y., Bao, Q., Liu, W., Fu, Y., Michael J., B., and Mei, T. (2021). Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Tu, H., Wang, C., and Zeng, W. (2020). Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, pages 197–212. Springer.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

- Vinyals, O., Bengio, S., and Kudlur, M. (2015). Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.
- Wang, C., Li, J., Liu, W., Qian, C., and Lu, C. (2020). Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 242–259, Cham. Springer International Publishing.
- Wang, Z., Nie, X., Qu, X., Chen, Y., and Liu, S. (2022a). Distribution-aware single-stage models for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13096–13105.
- Wang, Z., Nie, X., Qu, X., Chen, Y., and Liu, S. (2022b). Distribution-aware single-stage models for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13096–13105.
- Wu, S., Jin, S., Liu, W., Bai, L., Qian, C., Liu, D., and Ouyang, W. (2021). 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.
- Zanfir, A., Marinoiu, E., and Sminchisescu, C. (2018a). Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *CVPR*.
- Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.-I., and Sminchisescu, C. (2018b). Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems 31*, pages 8410–8419. Curran Associates, Inc.
- Zhang, J., Wang, J., Shi, Y., Gao, F., Xu, L., and Yu, J. (2022). Mutual adaptive reasoning for monocular 3d multi-person pose estimation. *arXiv preprint arXiv:2207.07900*.
- Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., and Zhou, X. (2020). Smap: Single-shot multi-person absolute 3d pose estimation. In *Computer Vision – ECCV 2020*. Springer International Publishing.
- Zhou, X., Huang, Q., Sun, X., Xue, X., and Wei, Y. (2017). Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *The IEEE International Conference on Computer Vision (ICCV)*.
- Zhu, L., Rematas, K., Curless, B., Seitz, S., and Kemelmacher-Shlizerman, I. (2020). Reconstructing nba players. In *Proceedings of the European Conference on Computer Vision (ECCV)*.