PoP-Net: Pose over Parts Network for Multi-Person 3D Pose Estimation from a Depth Image

Yuliang Guo* Zhong Li† Zekun Li‡ Xiangyu Du Shuxue Quan
Yi Xu
OPPO US Research Center, InnoPeak Technology, Inc.
{yuliang.guo, zhong.li, zekun.li, xiangyu.du, shuxue.quan, yi.xu}@oppo.com

Abstract

In this paper, a real-time method called PoP-Net is proposed to predict multi-person 3D poses from a depth image. PoP-Net learns to predict bottom-up part representations and top-down global poses in a single shot. Specifically, a new part-level representation, called Truncated Part Displacement Field (TPDF), is introduced which enables an explicit fusion process to unify the advantages of bottom-up part detection and global pose detection. Meanwhile, an effective mode selection scheme is introduced to automatically resolve the conflicting cases between global pose and part detections. Finally, due to the lack of high-quality depth datasets for developing multi-person 3D pose estimation, we introduce Multi-Person 3D Human Pose Dataset (MP-3DHP) as a new benchmark. MP-3DHP is designed to enable effective multi-person and background data augmentation in model training, and to evaluate 3D human pose estimators under uncontrolled multi-person scenarios. We show that PoP-Net achieves the state-of-the-art results both on MP-3DHP and on the widely used ITOP dataset, and has significant advantages in efficiency for multi-person processing. MP-3DHP Dataset and the evaluation code have been made available at: https://github.com/oppo-us-research/PoP-Net.

1. Introduction

Human pose estimation plays an important role in a wide variety of applications, and there is a rich pool of literature for human pose estimation methods. Categorizations of existing methods can be made from different dimensions. There are methods mostly relying on a single image to predict human poses [35, 23, 17] and others based on multiple cameras [27, 5]. Some methods are capable of predicting multiple poses [3, 16, 37] while others are focusing on single person [35, 19, 23]. Some methods estimate 3D poses [30, 13, 22, 36, 28] while others predict 2D poses [35, 19, 21]. Methods can also be classified by the input, while most methods use RGB images [3, 21, 17, 2], some use depth maps [14, 32, 36]. Specifically, this paper focuses on multi-person 3D pose estimation from a depth image.

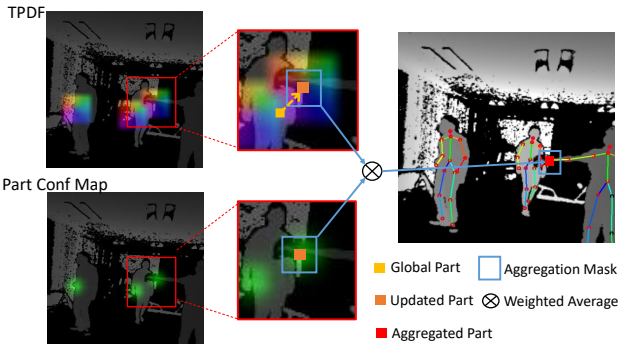

Figure 1. **Our paradigm.** Part representations and global poses predicted from PoP-Net are explicitly fused via utilizing Truncated-Part-Displacement-Field (TPDF). A part predicted from the global pose is dragged towards a more precise bottom-up part position following a displacement vector. More reliable part position is further estimated via a part-confidence-weighted average of TPDF within the aggregation mask.

In the era of deep learning, a large pool of Deep Neural Networks (DNN)-based methods have been developed for multi-person pose estimation. Ideas from existing literature can be generally categorized into three prototypical trends. The simplest idea is to directly extend a single-shot object detector [12, 24, 25] with additional pose attributes, so that the network can output human poses. Such single-shot regression can be very efficient, but has low part accuracy, as shown in Figure 2 (Yolo-Pose+), because a long-range inference for part locations is involved in a center-relative pose representation. The second type builds

*Contact Author. The work was done when Guo was with OPPO.
†Contact Author.
‡The work was done when Li was an intern with OPPO.

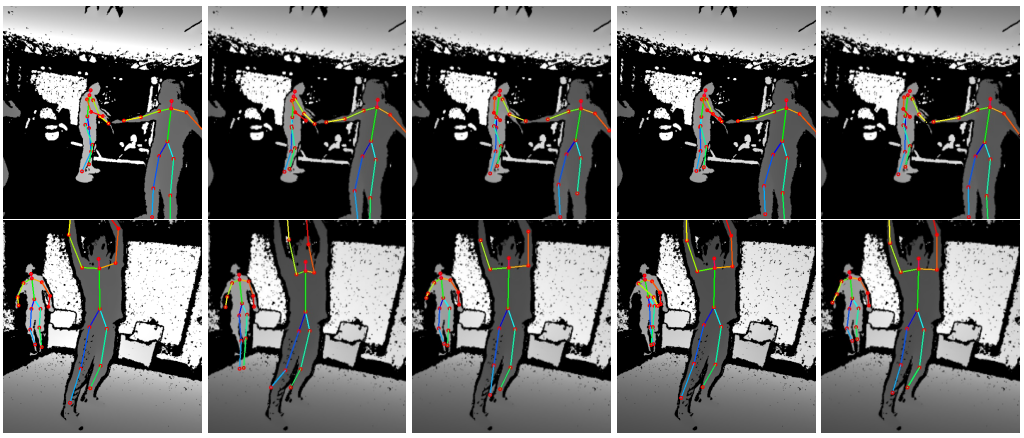

Figure 2. **Visual comparison of prototypical methods:** methods are compared on two examples from MP-3DHP testing set.

on a two-stage pipeline where the first stage detects object bounding boxes, and the second stage estimates the pose within each [21, 7, 36]. Two-stage methods can be very accurate, as shown in Figure 2 (Yolo-A2J), but not as efficient when more human beings appear in an image. In addition, more sophisticated design is required to solve the compatibility issue between pose estimation and bounding box detection [7, 24]. The third idea is to detect human poses from part[1] detection and association [10, 18, 3, 14, 16]. Although part detection can be rather efficient, solving the part association problem is usually time consuming. OpenPose [3] gained its popularity for introducing an efficient solution to solve the association, resulting in a network benefiting from both the single-shot pipeline with high efficiency and the part-based dense representation with high positional precision. However, a pure bottom-up method does not have a global sense, so that it is rather sensitive to occlusion, truncation, and ambiguities in symmetric limbs (Figure 2 Open-Pose+). Moreover, dependency on the bipartite matching in assembling parts presents the part predictions from differentiable integration with global reasoning, which may systematically block a solution from end-to-end training [16].

Developing a depth image-based multi-person 3D pose estimation from a well-established RGB image-based 2D method [21, 26, 3] is conceptually simple for two reasons: some schemes to handle multi-person detection can be shared; 3D information is partially available from the input [6, 36]. As a result, a depth image-based approach may not necessarily require as many sophisticated designs as methods aiming to estimate 3D poses from a single RGB image [17, 16]. However, RGB-based methods and depth-based methods usually focus on different challenges. While some RGB-based methods spend more effort in differentiating people cluttered in a large scene [3, 2], depth-based methods [14, 36] focus more on recovering highly accurate 3D poses of fewer people presented in a closer range. In practice, to develop a depth-based method that robustly predicts 3D poses in a multi-person scenario is very challenging due to noisy depth inputs and heavy occlusions.

In this paper, we present a method called Pose-over-Parts Network (**PoP-Net**) to estimate multiple 3D poses from a depth image. As illustrated in Figure 1, the main idea of PoP-Net is to explicitly fuse the predicted bottom-up parts and top-down global poses[2]. This fusion process is enabled by a new intermediate representation, called Truncated-Part-Displacement-Field (TPDF), which is a vector field that records the vector pointing to the closest part location at every 2D position. TPDF is utilized to guide a structural valid global pose towards more positionally precise part location so that the advantages of global pose and local part detection can be naturally unified.

At the same time, we release a comprehensive depth dataset, named Multi-Person 3D Human Pose Dataset (MP-3DHP), to facilitate the development of 3D pose estimation methods generalizable to novel background and unobserved multi-person configurations in real-world applications. Although there are a decent amount of RGB datasets [11, 9, 1, 15, 31, 34] in prior art, there are limited high-quality depth datasets. The released dataset is constructed to cover most of the essential aspects of visual variance related to 3D human pose estimation, and particularly to promote the development of multi-person methods.

The contribution of this paper is four fold. First, we introduce an efficient framework that predicts multiple 3D poses in a single shot. Second, we propose a new part-level representation called TPDF, which enables an explicit fusion of global poses and part-level representations. Third, we introduce a mode selection scheme that automatically resolves the conflict between local and global predictions.

[1]The definitions of 'part' and 'joint' are interchangeable in this paper.

[2]Poses predicted from a single-shot network where each pose contains a full set of body parts

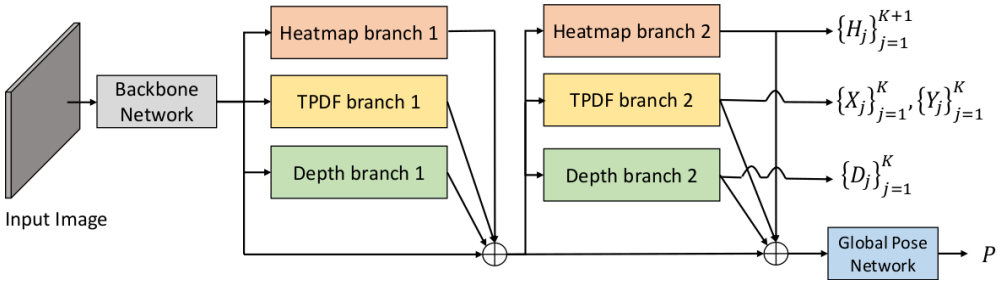

Figure 3. **PoP-Net** is composed of a backbone network, three functional branches, and a global pose network. The functional branches are organized in two stages with split and merge. PoP-Net outputs three part-level maps and a global pose map.

Finally, we release a comprehensive depth image-based dataset to facilitate the development of multi-person 3D pose estimation methods to tackle real-world challenges.

2. Pose-over-Parts Network

In this paper, we present a new method, called Pose-over-Parts Network (PoP-Net), for multi-person 3D pose estimation from a depth image. Our method first uses an efficient single-shot network to predict part-level representations and global poses, and then fuses the positionally precise part detection and structurally valid global poses in an explicit way.

The pipeline of PoP-Net is composed of a backbone network, a global pose network, and three functional branches: heatmap branch, depth branch, and TPDF branch, as illustrated in Figure 3. The two-stage split-and-merge design is inspired by OpenPose [3] and mostly follows the simplified version applied to depth input [14]. PoP-Net outputs three sets of part maps from the second stage of functional branches and an anchor-based global pose map from the global pose network.

Supposing a human body includes K body parts, the heatmap branch outputs a set of part confidence maps $\{H_j\}_{j=1}^{K+1}$, where each H_j from the first K maps indicates the confidence of a body part occurring at each discrete location, and the last indicates background confidence. The depth branch outputs a set of maps $\{D_j\}_{j=1}^{K}$, where each D_j encodes the depth values associated with part j.

The core of our method is a new part-level representation called Truncated Part Displacement Field (TPDF). For each part type j, TPDF records a displacement vector pointing to the *closest* part instance at every 2D position. The TPDF branch outputs TPDFs represented in a set of x-axis displacement maps $\{X_j\}_{j=1}^{K}$ and a set of y-axis displacement maps $\{Y_j\}_{j=1}^{K}$. The novelty of the proposed TPDF is two fold: (1) it encodes the displacement field involving multiple parts of the same type in a single map, and (2) a *truncated* effective range is adopted, which is critical for training CNN models that are able to handle multi-body scenarios. If truncated range is not applied to the part displacement field involving part instances from multiple bodies, the training of convolutional kernels will be confused by image patches similar in appearance but associated with highly different vectors, because a pair of displacement vectors whose origins are close to each other but pointing to different part instances may have large difference in X, Y values. The effectiveness of applying the truncated range is analyzed in detail in Section 4.3.

Compared with previous methods which predict person-wise part displacements [21, 36], TPDF operates at image level. In consequence, PoP-Net not only handles multiple bodies in one pass but also happens to be less sensitive to proposal error. While compared with the Part Affinity Field introduced in OpenPose [3], TPDF is free from the heavy bipartite matching process, and uses a simple fusion process to take advantage of both global poses and bottom-up part detections. As a result, PoP-Net shows increased robustness to handling truncation, occlusion and multi-person conflict compared with OpenPose.

Finally, a global pose map P is regressed from the global pose network. The global pose network is a direct extension from Yolo2 [25], where both bounding box attributes and additional 3D pose attributes are regressed with respect to the anchors associated with each grid. A set of predicted global poses are then extracted via conducting NMS on the global pose map P.

2.1. Training

PoP-Net is trained end-to-end via minimizing the total loss $\mathcal{L}$ which is the sum of heatmap loss $\mathcal{L}_h$, depth loss $\mathcal{L}_d$, TPDF loss $\mathcal{L}_t$, and global pose loss $\mathcal{L}_p$. As shown in Figure 3, losses corresponding to the functional branches are contributed from multiple stages of the network. Specifically, the loss function can be written as:

$$\mathcal{L} = \mathcal{L}_h + \mathcal{L}_d + \mathcal{L}_t + \mathcal{L}_p \tag{1}$$

$$\mathcal{L}_h = \sum_{s=1}^{S} \sum_{j=1}^{K+1} \left\| H_j^s - H_j^* \right\|_2^2 \tag{2}$$

$$\mathcal{L}_d = \sum_{s=1}^{S} \sum_{j=1}^{K} W_j^d \cdot \left\| D_j^s - D_j^* \right\|_2^2 \tag{3}$$

$$\mathcal{L}_t = \sum_{s=1}^{S}\sum_{j=1}^{K} W_j^t \cdot (\left\|X_j^s - X_j^*\right\|_2^2 + \left\|Y_j^s - Y_j^*\right\|_2^2) \quad (4)$$

$$\mathcal{L}_p = W^p \cdot \left\|P - P^*\right\|_2^2 \quad (5)$$

where s is the stage index, and S indicates the total number of stages. H_j^*, D_j^*, X_j^*, Y_j^*, and P^* indicate the ground-truth maps while W_j^d, W_j^t, and W^p indicate the point-wise weight maps in the same dimension as the corresponding ground-truth maps. Specifically, no weight maps are applied to heatmap loss so that the foreground and background samples are treated with equal importance.

The architecture of each network component and the preparation process of ground-truth maps and weight maps are illustrated in detail in Appendix A. In summary, the backbone and the heatmap branch follow the structure proposed in the simplified OpenPose for depth data [14]. The depth and TPDF branches adopt the same number of layers as the heatmap branch but with customized feature dimensions to balance the efficiency and robustness. The global pose network follows a similar architecture as the layers after the backbone in Yolo2 [25], and uses a similar anchor-based representation.

2.2. Fusion process

TPDF enables an explicit fusion of part representations and global poses. As illustrated in Figure 1, a 2D part predicted from a global pose located at (x_j, y_j) is updated to a new position $(\bar{x}_j, \bar{y}_j)$ following the displacement vector in the predicted TPDF of part j, such that $\bar{x}_j = x_j + X_j(x_j, y_j), \bar{y}_j = y_j + Y_j(x_j, y_j)$. To improve accuracy, weighted aggregation is later applied to estimate the final 2D position $\{(\hat{x}_j, \hat{y}_j)\}_{j=1}^K$ and depth $\{\hat{Z}_j\}_{j=1}^K$, as illustrated in Figure 1. Specifically, X_j, Y_j, and D_j within a mask M centered at the updated integer position $(\lfloor \bar{x}_j \rfloor, \lfloor \bar{y}_j \rfloor)$ is averaged by using H_j as aggregation weights, which leads to the following equations:

$$\hat{x}_j = \lfloor \bar{x}_j \rfloor + \frac{\sum_{(u,v)\in M} H_j(u,v) \cdot X_j(u,v)}{\sum_{(u,v)\in M} H_j(u,v)} \quad (6)$$

$$\hat{y}_j = \lfloor \bar{y}_j \rfloor + \frac{\sum_{(u,v)\in M} H_j(u,v) \cdot Y_j(u,v)}{\sum_{(u,v)\in M} H_j(u,v)} \quad (7)$$

$$\hat{Z}_j = \frac{\sum_{(u,v)\in M} H_j(u,v) \cdot D_j(u,v)}{\sum_{(u,v)\in M} H_j(u,v)}. \quad (8)$$

Predicted $\{(\hat{x}_j, \hat{y}_j)\}_{j=1}^K$ and $\{\hat{Z}_j\}_{j=1}^K$ are transformed to 3D positions given known camera intrinsic parameters. Compared with [20] which fuses multi-range offset maps in additional networks to recover instance segmentation, our solution relies on concise representations and an explicit fusion to recover 3D human poses.

2.3. Resolving conflicting cases

There are conflicting cases when multiple human bodies occlude each other or a global pose falls out of the effective range of a TPDF. To resolve them, a mode selection scheme is carefully designed. The scheme utilizes the part confidence maps $\{H_j\}_{j=1}^{K+1}$ from the heatmap branch and the part visibility attributes from the global pose network.

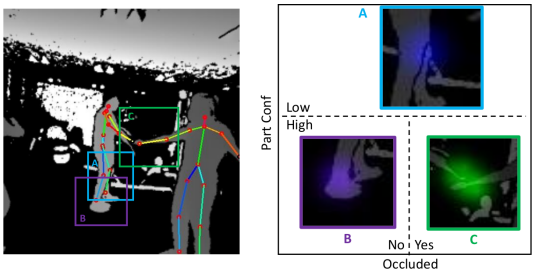

Figure 4. **Conflicting cases to resolve in fusion.** Part confidence maps for the marked regions are visualized to illustrate three conflicting cases to resolve. A: The confidence of left knee is low. B: The confidence of right foot is high without ambiguity. C: The confidence of occluded right hand is high but hallucinated by the same part from another person.

As illustrated in Figure 4, there are in total three cases to consider respectively: (A) when the part confidence H_j is low at a global part position, the global detection is used directly, which is usually observed when the position of a part is not accessible due to truncation or occlusion; (B) when the part confidence is high and no occlusion from another instance of the same part is involved, the presented fusion process is applied; and (C) a challenging case may occur when the part confidence is high but is impacted by occlusion from another instance of the same part type. Fortunately, since the part depth map is prepared following a z-buffer rule, a significant difference between the global part depth and the part depth map will be observed in this case. Therefore, part visibility attributes $\{v_j\}_{j=1}^K$ can be used as indicators of case (C), which can be integrated into the global pose representation. At last, the visual results of PoP-Net are shown in Figure 5 on a set of multi-person testing samples.

3. MP-3DHP: Multi-Person 3D Human Pose Dataset

Due to the lack of high-quality depth datasets for 3D pose estimation, we constructed Multi-Person 3D Human Pose Dataset (MP-3DHP) to facilitate the development of 3D pose estimation targeting real-world multi-person challenges. There are a few existing depth datasets for human pose estimation, but the data quality and diversity is rather limited. DIH [14] and K2HPD [32] include a decent amount of data but are limited to 2D poses. ITOP [6] is a widely

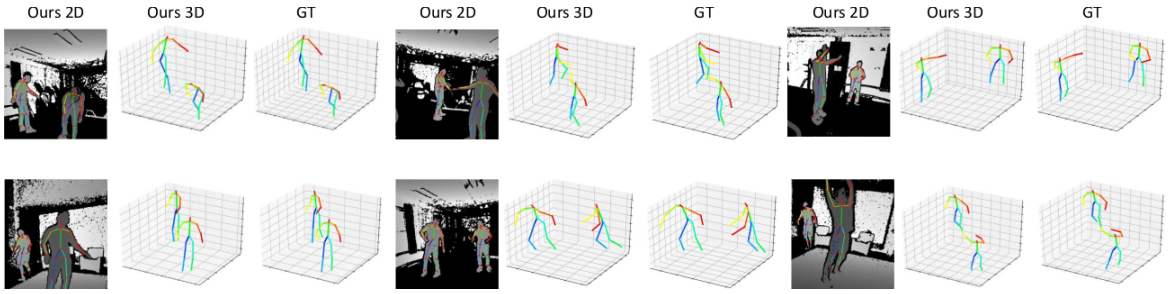

Figure 5. **PoP-Net visual results.** Predictions in 2D, 3D, and the ground-truth are visualized on six examples from MP-3DHP.

tested depth dataset for 3D pose estimation. However, data from ITOP is strictly limited to single person, clean background, and low diversity in object scales, camera angles, and pose types.

MP-3DHP is designed to effectively cover the essential variations in human poses, object scales, camera angles, truncation scenarios, background scenes, and dynamic occlusion. Because collecting sufficient data to fully represent multi-person configurations combined with different background scenes is intractable due to combinatorial explosion, real data is only collected to cover the variations not achievable from data composition or data augmentation. Specifically, our training data collection focuses on single-person data involving varying poses, different object scales, varying camera ray angles, and additional background-only data covering different types of scenes. The remaining types of data variation are covered via data composition and data augmentation utilizing collected human segments. The testing set focuses on multi-person data under uncontrolled real-world scenarios. Figure 6 shows examples from the training set, the background scenes and the multi-person testing set respectively.

Set	Img	Sbj	Loc	Ori	Act	Sn	L+
train	176828	13	4+	4	10+	1	seg
val	32719	2	4+	4	10+	1	seg
bg	8680	0	0	0	0	8	no
test	4484	5	0	0	free	4	mp

Table 1. **MP-3DHP summary.** The total number of images (Img), human subjects (Sbj), recording locations (Loc), self-orientations (Ori), action types (Act), scenes (Sn) are summarized. Additional label type (L+) indicates whether a set has segmentation (seg) or multi-person (mp) labels.

3.1. Construction procedure

We utilized Azure Kinect to record human depth videos and automatically extracted 3D human poses associated with each depth image. Overall, 20 human candidates were involved in the recording procedure; 15 of them were recorded individually to construct the training set, while the remaining five were recorded in multi-person sessions to produce the multi-person testing set. For the training set, each candidate was recorded with a clean background in four trials at four different locations within the camera frustum. In each trial, a candidate was asked to perform 10 predetermined actions while facing four different orientations spanning 360° and an additional short sequence of free-style movements towards the end. A classic graph-cut based method was applied to produce human segments for the training set. In addition, background images were recorded separately with moderate camera movements from eight different scenes. For the testing set, the remaining five people were recorded while performing random actions with different combinations in four different scenes. Table 1 shows the statistics of our MP-3DHP dataset.

To collect reliable 3D pose ground-truth, human annotations are integrated into our pseudo-automatic data collecting system to sift out those unqualified samples. Specifically, we used two calibrated and synchronized cameras mounted on a solid bar in data capture, one is from Azure Kinect and the other is from a commercial cellphone. A 3D pose output from Azure Kinect is only selected as ground-truth when its projections to both views are visually correct.

4. Experiments

We experimentally substantiate the superiority of PoP-Net in depth image-based multi-person 3D pose estimation on two datasets, under two evaluation metrics, compared to prior state-of-the-art methods.

4.1. Experimental setup

Datasets: A depth-based 3D pose estimation method is evaluated on two benchmarks: the MP-3DHP and ITOP [6] datasets. MP-3DHP includes highly diverse 3D human data and provides reliable human segments to enable background augmentation and multi-person augmentation. The evaluation on MP-3DHP aims to determine a method's capability in handling real-world challenges in multi-person 3D pose estimation. Meanwhile, ITOP is a widely tested dataset for single-person 3D pose under highly controlled environment. We report results on ITOP to compare with prior state-of-the-arts on a simplified task.

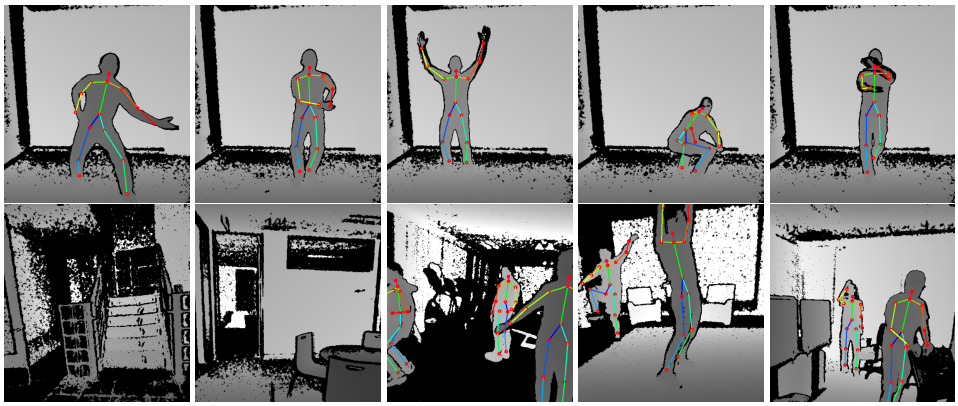

Figure 6. **MP-3DHP examples.** (Top) Five single-person examples recorded from different locations from the training set. (Bottom) Two examples of background scenes, and three examples from multi-person testing set.

Evaluation metric: A method is evaluated in PCK and mAP with different focuses. First, PCK is a measurement that focuses on pose estimation without considering redundant detections. In our experiments, PCK is calculated as the average percentage of accurate key points on the best-matched predictions to the ground-truth poses, where the best match is based on the IOU between 2D bounding boxes. Second, the mAP metric introduced in MPII [1] dataset is applied, which integrates both the object detection and pose estimation accuracy in an overall score. For both PCK and mAP, 0.5-head size rule is applied for 2D while 10-cm rule is applied for 3D.

Competing methods: A few prototypical methods have been compared with our method: (a) Yolo-Pose+ represents a typical top-down method, which is a pose estimation network extended from yolo-v2 [25] and implemented by us; (b) Open-Pose+ is a pure bottom-up method that is an extension from RPM [14], a simplified version of Open-Pose [3]) that we extended with a depth branch to predict 3D poses; (c) A2J [36] represents the state-of-the-art two-stage pose estimation from a depth image. To conduct a fair comparison, Yolo-Pose+, Open-Pose+, and PoP-Net are implemented using as many identical modules as possible. Specifically, Yolo-Pose+ is composed of the backbone network, the global pose network used for PoP-Net, and five additional intermediate 3×3 convolutional layers with $256d$ features in between. Open-Pose+ integrates the same backbone network, heatmap branch, depth branch as PoP-Net, and an additional Part Affinity Field (PAF) branch proposed in [3]. The post process of Open-Pose+ follows the original work to use bipartite matching upon PAF to assemble detected parts into human bodies, and in addition reads the depth branch output to produce 3D poses. For A2J [36], we use the identical network presented in the original paper to reproduce the results. Because A2J needs to work with given bounding-boxes for the multi-person case, we provide the predicted bounding boxes from Yolo-Pose+ to A2J so that the bounding-box quality is comparable to the other methods.

It is worth mentioning that Asure Kinect is not considered in comparison because it is impossible to retrain the algorithm for a fair comparison. It is also worth noting that recent methods integrated with the attention scheme [4, 29, 33] are not considered in comparison because they lack core designs tailored to the depth-image problem toward high 3D precision.

Implementation details: The input depth images are resized to 224×224 for Yolo-Pose+, Open-Pose+, and PoP-Net. The images are cropped and resized to 288×288 for A2J. Yolo-Pose+, Open-Pose+, PoP-Net are trained via standard SGD optimizer for 100 epochs, while A2J is trained via Adam optimizer following the original paper. Yolo-Pose+, Open-Pose+, and PoP-Net use two anchors with size 6×12 and 3×6, respectively. Following the study provided in [3], all the functional branches both in PoP-Net and Open-Pose+ use $S = 2$ stages for an optimal balance in efficiency and accuracy. PoP-Net uses the TPDF truncated range $r = 2$ in training, an optimal setting found in ablation study. At inference, the aggregation mask M used in fusion is practically defined as a 5×5 square. An identical data augmentation process is applied to the training of each method. The basic data augmentation process is applied to both MP-3DHP and ITOP, which includes random rotation, flipping, cropping, and a specific depth augmentation described in Appendix B. While the multi-person augmentation is only applied on MP-3DHP.

4.2. MP-3DHP dataset

On MP-3DHP dataset, each method is trained on the training set with multi-person augmentation on top of basic data augmentation. Given the provided human segments,

background augmentation can be applied by superimposing the human mask region from a training image onto a randomly selected background image. Meanwhile, multi-person augmentation can also been applied by superimposing multiple human segments onto a random background scene following z-buffer rule, which is described in more detail in Appendix C. In testing, a method is evaluated on four different datasets representing different levels of challenges: (1) the validation set (**Simple**), (2) the background augmented (**BG Aug**) set constructed from validation set , (3) the multi-person augmented (**MP Aug**) set constructed from validation set, and (4) the real multi-person (**MP Real**) testing set including challenging real-world recordings.

Evaluation results on MP-3DHP are shown in Table 2. As observed that all the methods' performance drop significantly from BG Aug to MP Aug, which indicates the multi-person occlusion is a major challenge. Meanwhile, PoP-Net achieves the state-of-the-art almost on every testing set, and significantly surpasses other methods under the most challenging metric: 3D mAP on the MP Real test. Open-Pose+ shows marginal advantages in 2D mAP in certain cases because it could benefit from higher recall by recovering isolated parts without seeing a whole body. However Open-Pose+ is more erroneous in depth prediction under occlusion due to its pure bottom-up pipeline, hence its 3D mAP drops significantly. A2J, on the other hand, shows marginal advantage in 3D PCK in certain tests, which can be interpreted as that the global weighted aggregation could leverage the full context within ROI to infer the depth of an occluded part. However, A2J appears to be rather sensitive to the quality of predicted ROIs; therefore its mAP performance is not optimal.

Visual results of PoP-Net are shown in Figure 5 for the real testing set, while additional qualitative comparisons on the most challenging cases are provided in Appendix G.

4.3. Ablation study

An ablation study has been conducted on MP-3DHP to analyze the effectiveness of PoP-Net components and the specific data augmentation methods. Because our motivation is to solve multi-person 3D pose estimation, we only conduct the ablation study on MP-3DHP, which to our knowledge is the only dataset providing multi-person 3D pose labels. In this section, we analyze the effectiveness of fusing pose and parts, the effectiveness of PoP-Net components. Additional analysis on the data augmentation can be found in Appendix D.

Effectiveness of fusing pose and parts: The effectiveness of fusing part representations and global poses has been studied. As shown in Table 3, evaluation has been done on four different testing sets from MP-3DHP, and separately on: (1) the 2D global poses predicted from the global pose network (**2D Glb**), (2) the final 2D poses after fusion (**2D Fuse**), (3) the 3D global poses predicted from the global pose network (**3D Glb**), and (4) the 3D poses computed from 2D fused poses and predicted depth of parts (**3D Fuse**). In addition, the upper bound of 3D poses computed from ground-truth 2D poses and predicated depth (**3D UB**) has been reported to illustrate the importance of depth prediction. As observed, 2D and 3D poses after fusion constantly improve upon the direct predictions from the global pose network, and the margin increases on more challenging sets. Meanwhile, the accuracy of 3D pose prediction drops more significantly compared with 2D pose prediction on more challenging sets, where the upper bound is far from ideal. Improving depth prediction under multi-person occlusion is expected a main task in the future.

Test	Method	2D PCK	3D PCK	2D mAP	3D mAP
Simple	Yolo-Pose+	0.957	0.910	0.926	0.847
	Open-Pose+	0.967	0.915	0.967	0.893
	Yolo-A2J	0.959	0.924	0.936	0.868
	PoP-Net	**0.978**	**0.947**	**0.974**	**0.926**
BG Aug	Yolo-Pose+	0.956	0.904	0.923	0.834
	Open-Pose+	0.911	0.916	0.969	0.885
	Yolo-A2J	0.964	0.927	0.941	0.871
	PoP-Net	**0.982**	**0.947**	**0.977**	**0.924**
MP Aug	Yolo-Pose+	0.872	0.777	0.799	0.651
	Open-Pose+	0.887	0.765	**0.870**	0.667
	Yolo-A2J	0.876	**0.819**	0.803	0.707
	PoP-Net	**0.906**	0.808	0.863	**0.708**
MP Real	Yolo-Pose+	0.734	0.607	0.616	0.449
	Open-Pose+	0.805	0.641	**0.802**	0.558
	Yolo-A2J	0.837	**0.724**	0.744	0.574
	PoP-Net	**0.839**	0.708	0.799	**0.606**

Table 2. **Evaluation on MP-3DHP.** Competing methods are evaluated on four testing sets. The best method is marked in bold black while the second best is marked in blue.

Metric	Test	2D Glb	2D Fuse	3D Glb	3D Fuse	3D UB
PCK	Simple	0.968	0.978	0.939	0.947	0.956
	BG Aug	0.970	0.982	0.938	0.947	0.953
	MP Aug	0.898	0.906	0.794	0.808	0.846
	MP Real	0.798	0.839	0.681	0.708	0.756
mAP	Simple	0.963	0.974	0.917	0.926	0.931
	BG Aug	0.965	0.977	0.915	0.924	0.927
	MP Aug	0.849	0.863	0.701	0.708	0.735
	MP Real	0.755	0.799	0.582	0.606	0.622

Table 3. **Ablation study on fusing pose and parts.**

Effectiveness of PoP-Net components: To find an optimal configuration of the pipeline, we analyzed the effects of the truncated range r on TPDF, the effectiveness of depth prediction (**D Pred**) and the conflict resolving scheme (**Sol C**),

Sol C	D Pred	TPDF	2D PCK	3D PCK	2D mAP	3D mAP
w	w\o	r = 2	0.839	0.418	0.799	0.293
w\o	w	r = 2	0.833	0.696	0.797	0.598
w	w	r = 1	0.827	0.678	0.782	0.560
w	**w**	**r = 2**	**0.839**	**0.708**	**0.799**	**0.606**
w	w	r = 3	0.825	0.681	0.777	0.567
w	w	r = 5	0.821	0.670	0.780	0.557
w	w	r = 10	0.797	0.654	0.755	0.543
w	w	r = inf	0.647	0.549	0.581	0.443

Table 4. **Ablation study on PoP-Net components.**

as reported in Table 4. A few important conclusions can be drawn from the results. First, depth prediction plays an important role in recovering reliable 3D poses, which leads to about 30% improvements in 3D metrics compared with using the raw depth directly. Second, the introduced conflict resolving scheme leads to about 1% improvement in 3D metrics. Third, applying an appropriate truncated range is critical in learning reliable models to predict part displacement vectors in multi-person scenarios. On the one hand, if the range is too limited, a global pose may not fall in the effective range of TPDF so that the positional accuracy can not be improved. On the other hand, a displacement field without truncation (r=inf) or with large truncated range would confuse the learning. As illustrated in Figure 7, near those sharp boundaries, similar image patches could be associated with drastically different flow vectors. This leads to a degenerate regression problem.

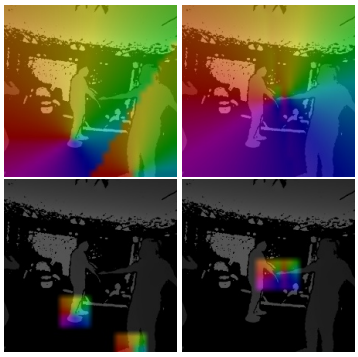

Figure 7. **PDF vs. TPDF.** Full-range PDF examples (top) are compared with the TPDF (bottom) with truncated range $r = 2$ at $\frac{h}{8} \times \frac{w}{8}$ feature resolution. Fields associated with the right ankle and right waist channels are visualized in columns.

4.4. ITOP dataset

PoP-Net is compared with competing methods on ITOP dataset. Because ITOP dataset is limited to single-person and clean background, PCK and mAP measurements are mostly identical. Therefore, we only report PCK metrics on ITOP dataset. To conduct a fair comparison, a method is trained and tested under two setups separately. One is based on provided ground-truth bounding boxes and the other directly uses the full image, as shown in Table 5. It can be observed that PoP-Net consistently outperforms Open-Pose+ and Yolo-Pose+ by a significant margin. Compared with A2J, PoP-Net is slightly worse in 3D and better in 2D, which is consistent with the PCK results reported on MP-3DHP.

Exp Setup	Method	2D PCK	3D PCK
GT Bbox	A2J	0.905	**0.891**
	PoP-Net	**0.914**	0.882
Full Image	Yolo-Pose+	0.833	0.787
	Open-Pose+	0.876	0.778
	Yolo-A2J	0.873	**0.854**
	PoP-Net	**0.890**	0.843

Table 5. **Evaluation on ITOP dataset (front-view).** Methods are evaluated on ITOP dataset with and without GT bounding boxes.

4.5. Running speed analysis

The efficiency of each method is measured in FPS on multi-person test (2-3 people). The calculation of the running speed considers necessary post-processing to achieve the final set of multiple 3D poses and the bounding box prediction time for a two-stage method. As shown in Table 6, the running speed of PoP-Net almost triples A2J and doubles Open-Pose+ on a single RTX 2080Ti GPU. The observation is as expected because OpenPose+ involves heavier post process and A2J's cost scales up with the number of humans. More detailed efficiency analysis is provided in Appendix E.

	Yolo-Pose+	Open-Pose+	A2J	PoP-Net
FPS	223	48	32	91

Table 6. **Running speed on multi-person data.**

5. Conclusion

In this paper, we introduce PoP-Net for multi-person 3D pose estimation from a depth image. PoP-Net predicts part maps and global poses in a single pass and explicitly fuses them via utilizing the proposed Truncated Part Displacement Field (TPDF). Conflicting cases are effortlessly resolved in a rule-based process given part visibility and confidence out of the network. Meanwhile, a comprehensive 3D human depth dataset called MP-3DHP is released to facilitate the development of methods for real-world multi-person challenges. In experiments, PoP-Net achieves state-of-the-art results on MP-3DHP and ITOP datasets with significant advantage in 3D mAP and running speed in processing multi-person data.

Acknowledgements
This work was supported by OPPO US Research Center.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

[2] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6855–6864. IEEE, 2020.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[4] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5669–5678, 2017.

[5] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Mykhaylo Andriluka, Christoph Bregler, Bernt Schiele, and Christian Theobalt. Marconi - convnet-based marker-less motion capture in outdoor and indoor scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(3):501–514, 2017.

[6] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Fei-Fei Li. Towards viewpoint invariant 3d human pose estimation. In *14th European Conference Computer Vision (ECCV)*, pages 160–177, 2016.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014.

[10] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *European Conference on Computer Vision Workshops*, pages 627–642, 2016.

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *13th European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *14th European Conference on Computer Vision (ECCV)*, pages 21–37, 2016.

[13] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017.

[14] Ángel Martínez-González, Michael Villamizar, Olivier Canévet, and Jean-Marc Odobez. Real-time convolutional networks for depth-based human pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 41–47, 2018.

[15] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516, 2017.

[16] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: real-time multi-person 3d motion capture with a single RGB camera. *ACM Trans. Graph.*, 39(4):82, 2020.

[17] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: real-time 3d human pose estimation with a single RGB camera. *ACM Trans. Graph.*, 36(4):44:1–44:14, 2017.

[18] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Annual Conference on Neural Information Processing Systems*, pages 2277–2287, 2017.

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *14th European Conference on Computer Vision (ECCV)*, pages 483–499, 2016.

[20] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *15th European Conference on Computer Vision (ECCV)*, volume 11218, pages 282–299, 2018.

[21] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7307–7316, 2018.

[23] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017.

[24] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[25] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[26] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[27] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *14th European Conference on Computer Vision (ECCV)*, pages 509–526, 2016.

[28] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(5):1146–1161, 2020.

[29] Nicholas Santavas, Ioannis Kansizoglou, Loukas Bampis, Evangelos G. Karakasis, and Antonios Gasteratos. Attention! A lightweight 2d hand pose estimation approach. *CoRR*, abs/2001.08047, 2020.

[30] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016.

[31] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *15th European Conference on Computer Vision (ECCV)*, volume 11214, pages 614–631, 2018.

[32] Keze Wang, Shengfu Zhai, Hui Cheng, Xiaodan Liang, and Liang Lin. Human pose estimation from depth images via inference embedded multi-task learning. In *ACM Conference on Multimedia Conference*, pages 1227–1236, 2016.

[33] Xiangyang Wang, Jiangwei Tong, and Rui Wang. Attention refined network for human pose estimation. *Neural Process. Lett.*, 53(4):2853–2872, 2021.

[34] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless C. Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *CoRR*, abs/1905.07718, 2019.

[35] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.

[36] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2J: anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 793–802, 2019.

[37] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: single-shot multi-person absolute 3d pose estimation. In *16th European Conference Computer Vision (ECCV)*, volume 12360, pages 550–566, 2020.