VirtualPose: Learning Generalizable 3D Human Pose Models from Virtual Data

Jiajun Su[1,2], Chunyu Wang[5] *, Xiaoxuan Ma[2,3], Wenjun Zeng[6], and Yizhou Wang[2,3,4]

[1] Center for Data Science, Peking University
[2] Center on Frontiers of Computing Studies, Peking University
[3] Dept. of Computer Science, Peking University
[4] Inst. for Artificial Intelligence, Peking University
[5] Microsoft Research Asia
[6] EIT Institute for Advanced Study
{sujiajun, maxiaoxuan, yizhou.wang}@pku.edu.cn, chnuwa@microsoft.com,
zengw2011@hotmail.com

Abstract. While monocular 3D pose estimation seems to have achieved very accurate results on the public datasets, their generalization ability is largely overlooked. In this work, we perform a systematic evaluation of the existing methods and find that they get notably larger errors when tested on different cameras, human poses and appearance. To address the problem, we introduce *VirtualPose*, a two-stage learning framework to exploit the hidden "free lunch" specific to this task, *i.e.* generating infinite number of poses and cameras for training models at no cost. To that end, the first stage transforms images to abstract geometry representations (AGR), and then the second maps them to 3D poses. It addresses the generalization issue from two aspects: (1) the first stage can be trained on diverse 2D datasets to reduce the risk of over-fitting to limited appearance; (2) the second stage can be trained on diverse AGR synthesized from a large number of virtual cameras and poses. It outperforms the SOTA methods without using any paired images and 3D poses from the benchmarks, which paves the way for practical applications. Code is available at https://github.com/wkom/VirtualPose.

Keywords: Absolute 3D Human Pose Estimation

1 Introduction

Monocular 3D pose estimation has attracted much attention since it can benefit many applications. Most works [4, 5, 17, 19, 24, 25, 32, 38, 39] focus on a simpler sub-task of *relative* 3D pose estimation where only relative joint positions are estimated. Absolute 3D pose estimation needs to estimate the depth of a person's *root* joint in the camera coordinate system. This is more challenging because it is ill-posed and multiple entangled latent factors jointly determine the depth. As

---

* Corresponding author.

shown in Fig. 1, the relevant factors include at least the height of the person in neutral standing pose, its relative posture, its projection size in 2D, camera focal length, and camera view point. Some factors such as focal length may be assumed known in some cases but most others need to be implicitly estimated from images along with depth.

Many works [3, 15, 23, 30, 37, 42, 43, 47] propose to brute-forcely learn the mapping from images to depth. Although they have got good results on the public datasets, they have poor generalization ability. We are aware that other tasks also face the issue but the situation is very different for pose estimation. First, the pose datasets [10, 12, 20, 22, 31] have extremely limited variations in terms of cameras, human poses and appearance. Second, many data augmentation techniques are not applicable, *e.g.*, we can not change the view point neither the human poses in an image. So, addressing the generalization issue is non-trivial compared to other tasks. This was not identified as a serious issue previously because the current training and testing data [10, 12] are similar. In fact, even for the cross-dataset experiment (train on MuCo-3DHP [22] and test on MuPoTS-3D [22]), the camera views are also similar.

In this paper, we address the challenges by introducing an intermediate representation, termed as Abstract Geometry Representation (AGR), into the 3D pose estimation network. It is a bundle of multiple geometry representations that satisfy three requirements: (1) they are helpful for recovering absolute 3D poses, (2) can be synthesized from 3D poses, and (3) can be robustly estimated from images even when the model is trained on wild images rather than mocap datasets. As shown in Fig. 2, AGR splits a 3D pose estimator into two successive modules. The first module maps a raw image to AGR which, in current implementation, consists of human detection and joint localization results. Since it only handles 2D tasks which are barely affected by 3D projection geometry, we can train the model on the diverse 2D datasets such as COCO [16], which covers a large number of camera views, human poses and backgrounds, and apply extensive data augmentation. As a result, the module is robust to different factors and achieves desirable results on wild data.

The second module learns to regress 3D pose from AGR. Different from the previous works, we propose a novel training strategy which synthesizes a large number of paired <AGR, Pose> data from diverse camera views, poses and person positions to learn a generalizable model. It is worth noting that AGR suffers less from domain gaps than raw images. In some cases when we want to deploy a model for a fixed environment, *e.g.* installing a camera at home for elderly care, we can even generate virtual training data specifically for the environment, which as shown in our experiment, gets more accurate results. Combining the two modules, we get an accurate yet generalizable 3D pose estimator. It not only outperforms the state-of-the-art methods on the benchmark datasets but also achieves good results on our own videos collected in retail stores with cluttered background and severe occlusion.

We implement the above idea following the architecture of VoxelPose [33], as shown in Fig. 2. Given an image as input, we first estimate bounding boxes and
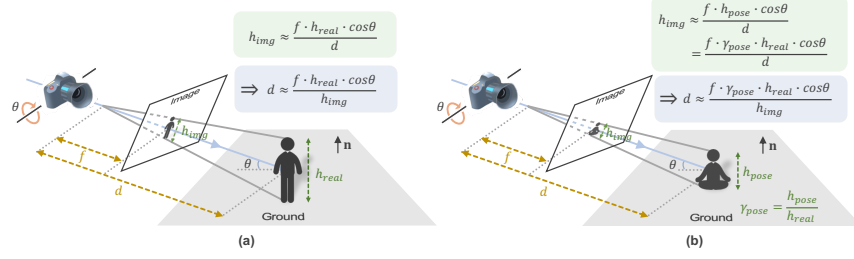
Fig. 1: Projection geometry of a pinhole camera model. (a) shows the case where a person is in a standing pose. (b) shows the case with a different pose.
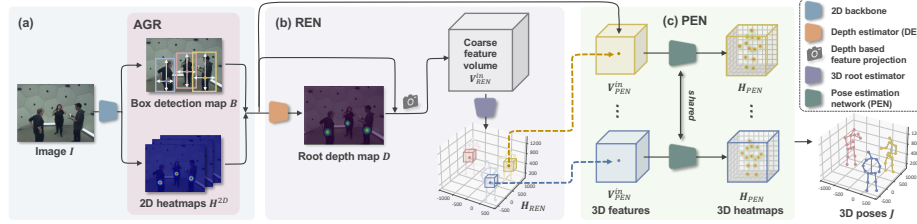

Fig. 2: Our 3D human pose estimation pipeline. (a) Given an image as input, it first estimates the AGR (*i.e.* bounding boxes and heatmaps) of all people. (b) Then, in Root Estimation Network (REN), it estimates the depth map of root joints from the AGR. The AGR and the depth map are integrated via depth based feature projection to estimate 3D root positions. (c) Finally, in Pose Estimation Network (PEN) for each root joint, we construct a feature volume around it to estimate a fine-grained 3D pose.

2D pose heatmaps as AGR using a CNN network. Then they are integrated into the 3D space to estimate the 3D positions of all persons in the image. Finally, for each person, we construct a feature volume around its position to estimate a fine-grained 3D pose.

In summary, our contributions are three-fold. First, we present the first systematic study on models' generalization ability which is largely overlooked in the previous works. We argue that it is important to purposely prevent models from over-fitting when designing and evaluating new models. Second, we present one possible way for learning generalizable models purely from virtual 3D poses and cameras which is made possible by AGR. Note that the decoupling strategy has been used in previous works [2,14,18,26,33,40] for different motivations. Our contribution lies in leveraging it in our training strategy for absolute 3D pose estimation with concrete designs. Finally, the method outperforms the existing methods without using paired images and 3D poses for training, which notably improves its applicability in practice.

2   Related Work

Datasets and Generalization Collecting motion capture datasets is cumbersome because it either uses multiview systems [10, 12, 20, 22, 31] or wearable IMUs [18, 45]. As a result, they usually have limited camera views, poses and actors. Some works propose to generate more data by changing the image backgrounds [22, 23] of the existing datasets, or using game engines [1, 46, 50]. However, the domain gap between synthetic and real images poses challenges for neural networks. We find that the models trained on these datasets lack generalization capability. Some works [36, 41, 48] address the generalization problem in relative pose estimation but they cannot be extended to the absolute task. There are some works in other areas that also decouple a network into different stages [18, 33, 40]. While we also follow this general idea, we have specific designs to apply it to the 3D absolute pose estimation task.

Minimizing Projection Error The methods of this class [6, 7, 28, 29] first estimate relative 3D human poses with sizes from an image in an end-to-end manner. Then based on the estimated 2D projection size and the focal length, they can coarsely compute the depth of each pose by minimizing an error between the 2D pose and the projection of the estimated 3D pose. The results of these methods heavily rely on the accuracy of relative 3D pose estimation. Some later works [36, 48] show that they have poor generalization results because they need to be trained on the (small) motion capture datasets [10, 20].

Ground Plane Geometry Some works such as [21] propose to analytically compute the real 3D size and depth from ground plane geometry. They assume that the camera is calibrated, people are standing on the ground and their feet are visible in the image. Then the absolute human height and depth can be calculated by projection geometry. However, these assumptions are not always true in practice which limits their applicability.

End-to-end Learning Some works [15, 23, 37, 47] directly estimate depth from images. We first explain why this is possible. Consider a simple case where a person with height h_{real} stands on the ground plane and a camera with focal length f is placed with pitch angle θ, as shown in Fig. 1(a). Then the height of the person in the image h_{img} is approximately $h_{img} \approx \frac{f \cdot h_{real} \cdot cos\theta}{d}$. In general, people may be in various poses, *e.g.* sitting on the ground as shown in Fig. 1(b). We define the height of the person perpendicular to the ground as h_{pose}, and a pose-dependent correction factor as $\gamma_{pose} = \frac{h_{pose}}{h_{real}}$, then we have $h_{img} \approx \frac{f \cdot \gamma_{pose} \cdot h_{real} \cdot cos\theta}{d}$, from which the depth d of the person can be computed.

These methods usually assume f is known and estimate normalized depth $d^{norm} = \frac{d}{f}$. It is challenging to estimate h_{real} from a single image. But since there are only few actors in the benchmark datasets, they are implicitly learned by the network. The rest factors of γ_{pose}, θ and h_{img} can be estimated from images when the training data is large and diverse. However, since most benchmark datasets are small, *e.g.* having limited camera poses, human poses and backgrounds,

the depth estimation models may easily over-fit to these datasets but can not generalize well on unseen images.

3   Generalization Study

In this section, we systematically evaluate the robustness of the existing methods to the variations of the key factors. In Section 3.1, we first introduce three representative baselines for estimating depth which are most frequently used by the existing 3D pose estimators [8, 15, 23, 34, 37, 47, 49]. Then we evaluate their robustness to the three factors including camera pose, image background and human pose.

3.1   Baselines and Datasets

Top-down Box Size based method (TBS) Given an image as input, it first localizes all people by bounding boxes and then estimates depth for each people. They assume that the 3D sizes of all people are constant and known, and geometrically compute the depth based on the 2D size of the bounding box. However, as discussed in Fig. 1(b), the 2D size is also dependent on the human pose. So they learn a correction factor from images. The factor is used to refine the estimated depth. Multiple methods [8, 23] follow this strategy. We choose RootNet [23] in our experiments.

Top-down Image Feature based method (TIF) It first detects all people in the input image. But instead of using the size of the detected bounding box to analytically compute depth, they use the estimated 2D heatmaps as attention masks to pool features from the image and directly predict the human depth. We choose HDNet [15] in our experiments.

Bottom-up Depth Regression based method (BDR) This method directly estimates a depth map from each image using a deep network. The depth of each joint can then be obtained from the depth map directly. Many works [34, 37, 47] follow this pipeline. We present a simple 3D pose estimator which has two branches for estimating 2D pose heatmaps and depth maps, respectively. The 3D pose can then be analytically computed from them.

Implementation Details In order to mitigate the influence of detection results, we use the GT human bounding box in TBS and TIF methods. The backbone in the BDR method is ResNet-50 [9]. In order to mitigate the influence of 2D pose estimation, we use the GT 2D root position to obtain root depth from the depth maps in BDR method. We conduct experiments on the CMU Panoptic [12], MuCo-3DHP and MuPoTS-3D [22] datasets.

CMU Panoptic [12] is a large-scale multi-person pose dataset captured in a studio with multiple cameras. We select data of four activities (Haggling, Mafia, Ultimatum, Pizza) as our training and test set. The training and testing splits are different in different ablative experiments and will be made clear as needed.
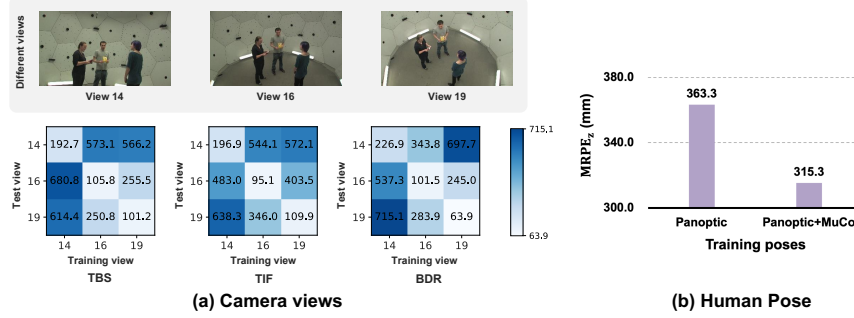
Fig. 3: (a) Cross-view results (MRPE$_z$) of the three models on the CMU Panoptic dataset. The x- and y-axes represent the camera index in training and test, respectively. (b) Depth estimation results (MRPE$_z$) when the training poses are from the CMU Panoptic dataset, or from the mixture of "Panoptic+MuCo" datasets. The MuPoTS-3D dataset is the test dataset.

MuCo-3DHP [22] is created by compositing randomly sampled 3D poses from single-person 3D human pose dataset MPI-INF-3DHP [20] to synthesize multi-person scenes. The images are synthesized by resizing and compositing the corresponding segmented images of each person according to their 3D positions. We use the data provided by [23] which contains images with green screen background and images with augmented background using COCO [16] dataset.

MuPoTS-3D [22] is a multi-person test set comprising 20 general real world scenes with GT 3D pose obtained with a multi-view marker-less motion capture system. It contains 5 indoor and 15 outdoor settings, varying human poses and camera poses, which is a commonly used test set to validate the models' generalization ability.

Table 1: Depth estimation results (MRPE$_z$) when we train the models on images with green-screen background (green) or with mixture of green-screen and augmented background (augmented), respectively.

Method	Training data	Test data	
		MuCo (green)	MuCo (augmented)
TBS	green	76.0	281.6
	augmented	73.3	82.5
TIF	green	57.3	168.8
	augmented	58.7	65.5
BDR	green	126.3	267.8
	augmented	141.7	148.9

Metrics We focus on evaluating the quality of the estimated depth of root joint in this section. We use the metric of $\mathrm{MRPE_z}$ (in mm) proposed in [23], which is the mean of the root position errors of all people in the z direction (*i.e.* the depth direction in the camera coordinate system).

3.2 Experimental Results

Camera Views We conduct the cross-view experiment on the CMU Panoptic [12] dataset using cameras of 14, 16 and 19 as shown in Fig. 3(a). We train the three baseline models on one of the cameras and test them on the rest. The results are shown in Fig. 3(a). All methods get several times larger depth estimation errors when they are trained and tested on different cameras. The results are reasonable because (1) camera angle affects image appearance and image features; (2) camera angle also affects the 2D size of the people in the image. The results suggest that the models learned on few camera views are barely generalizable. It is helpful to train the models on as many views as possible.

Image Background For each method, we train two models on the MuCo-3DHP [22] dataset, with the first on images with green screen background and the second with both green screen and augmented background. We select 70% of the dataset for training, and the rest for testing. Tab. 1 shows the results. When tested on the green screen subset of MuCo-3DHP, BDR (green) achieves a smaller error than BDR (augmented). It means that BDR probably learns some scene priors to reduce the ambiguity in estimation but it may also risk over-fitting. For example, BDR (green) gets a significantly larger error on the augmented subset of MuCo-3DHP. The results not only validate the limitations of those methods on small training datasets, but also suggest that we actually need a better evaluation protocol, *e.g.* training and testing on datasets with large differences, to convincingly evaluate the future works.

Human Pose We train two models either using poses only from the CMU Panoptic dataset, or from the combined CMU Panoptic and MuCo-3DHP datasets. We test the models on the MuPoTS-3D dataset. To keep other factors such as human appearance the same, we modify BDR to take pose heatmaps as input and output the root depth map $\boldsymbol{D}$ (this is actually part of our method as will be described). The results are shown in Fig. 3(b). The model trained on CMU Panoptic poses achieves a larger error meaning that the model trained on limited poses may have poor generalization results. While the conclusion is not surprising, it actually points out an overlooked problem that relying on limited mocap datasets to train 3D pose estimators will probably fail in real-world applications.

4 VirtualPose

In this section, we present our *VirtualPose*. To address the challenges, we introduce an intermediate representation, termed as Abstract Geometry Representation

(AGR, Section 4.1), into the 3D pose estimation network so that we can simultaneously leverage the diverse 2D datasets and large amounts of synthetic data. As shown in Fig. 2, AGR splits a 3D pose estimator into two successive modules. The first module maps a raw image to AGR which consists of human detection and joint localization results. Then the second module maps AGR to the corresponding 3D pose which is trained on synthesized <AGR, Pose> data. In particular, we present Root Estimation Network (REN, Section 4.2) to estimate 3D positions of the persons. Then for each person, Pose Estimation Network (PEN, Section 4.3) is proposed to estimate the 3D pose.

4.1   Abstract Geometry Representation

AGR is a general concept representing a bundle of geometry representations that satisfy three requirements: (1) they are helpful for recovering absolute 3D poses, and (2) can be robustly estimated from images even when the model is only trained on the wild 2D datasets rather than the 3D motion capture datasets, and (3) can be synthesized or rendered from 3D poses. In current implementation, we only use 2D pose heatmaps and human detection bounding boxes in AGR for simplicity. But more cues such as ordinal depth, occlusion relationship or segmentation mask can also be leveraged.

We adopt a simple architecture for estimating AGR from images following CenterNet [49]. As shown in Fig. 2(a), given an image $\boldsymbol{I}$ as input, the 2D backbone outputs the corresponding 2D pose heatmaps $\boldsymbol{H}^{2D} \in [0, 1]^{W \times H \times N}$ and box detection map $\boldsymbol{B} \in \mathcal{R}^{W \times H \times 4}$ where the four channels encode the distances from the human root joint to the four edges of the human bounding box, respectively. Here, N indicates the number of human joints, and W, H are the width and height of the output maps. The 2D human pose estimation and detection network is trained by:

$$\mathcal{L}_{2D} = \mathcal{L}_{heat}^{2D} + \lambda_{bbox}\mathcal{L}_{bbox} \tag{1}$$

$$\mathcal{L}_{heat}^{2D} = ||\boldsymbol{H}^{2D} - \widetilde{\boldsymbol{H}}^{2D}||_2^2 \tag{2}$$

$$\mathcal{L}_{bbox} = \sum_{p \in \mathcal{P}} ||\boldsymbol{B}_p - \widetilde{\boldsymbol{B}}_p||_1, \tag{3}$$

where $\widetilde{\boldsymbol{H}}^{2D}$ is the GT 2D pose heatmaps and λ_{bbox} is a hyper-parameter. The box supervision is only enforced at GT root joint locations $\mathcal{P}$ where $\boldsymbol{B}_p$ is the estimated box embedding for the p_{th} person and $\widetilde{\boldsymbol{B}}_p \in \mathcal{R}^4$ is the GT box embedding.

4.2   Root Estimation Network

We first estimate a coarse depth map for the root joint as shown in Fig. 2, which will be used to reduce the ambiguity when projecting 2D heatmaps to the 3D

space. To that end, the previously estimated 2D pose heatmaps $\boldsymbol{H}^{2D}$ are fed to the 2D CNN-based Depth Estimator (DE) to estimate a root depth map $\boldsymbol{D}$. It is trained by minimizing:

$$\mathcal{L}_{depth} = \sum_{p \in \mathcal{P}} |\boldsymbol{D}_p - \widetilde{\boldsymbol{D}}_p|, \tag{4}$$

where $\boldsymbol{D}_p$ and $\widetilde{\boldsymbol{D}}_p$ are the estimated and GT depth values of the root joint for the p_{th} person.

Then we estimate the 3D positions of the root joints by constructing a 3D representation. We discretize the 3D space in which people can freely move into $X \times Y \times Z$ voxels $\{\boldsymbol{G}_{x,y,z}\}$, and construct a 3D feature volume $\boldsymbol{V}_{REN}^{in} \in [0,1]^{X \times Y \times Z \times N}$ by inversely projecting 2D pose heatmaps of N body joints to the 3D space using camera parameters. Different from the previous works [33,44], for a heatmap vector at a pixel location within a bounding box, we only project it to a few voxels whose depths are similar to the estimated depth $\boldsymbol{D}_p$ of the root joint in the bounding box. For example, if a voxel $\boldsymbol{G}_{x,y,z}$ whose projected 2D pixel location (u,v) is within the p_{th} human bounding box, then the feature of $\boldsymbol{G}_{x,y,z}$ can be calculated as

$$\boldsymbol{V}_{REN}^{in}(x,y,z) = \boldsymbol{H}^{2D}(u,v) \exp(-\frac{(z - \boldsymbol{D}_p)^2}{2\sigma^2}), \tag{5}$$

where σ is empirically set to be 200. This is particularly important for monocular 3D pose estimation which helps reduce the ambiguity. The feature volume $\boldsymbol{V}_{REN}^{in}$ coarsely encodes the likelihood of each person's position in each voxel. We feed $\boldsymbol{V}_{REN}^{in}$ to a 3D network to estimate the corresponding 3D heatmaps $\boldsymbol{H}_{REN} \in [0,1]^{X \times Y \times Z}$ indicating the confidence of each voxel containing a root joint as shown in Fig. 2(b). We select the voxels with large confidence values in the estimated $\boldsymbol{H}_{REN}$ with Non-Maximum Suppression (NMS). We train the 3D network by minimizing:

$$\mathcal{L}_{REN} = ||\boldsymbol{H}_{REN} - \widetilde{\boldsymbol{H}}_{REN}||_2^2, \tag{6}$$

where $\widetilde{\boldsymbol{H}}_{REN} \in [0,1]^{X \times Y \times Z}$ is the GT 3D heatmap for root joint. We set X, Y and Z to be 80, 80 and 24, respectively.

4.3   Pose Estimation Network

After estimating the root joint position for each person, we will estimate a complete 3D pose for it. As shown in Fig. 2(c), we first construct a finer-grained feature volume centered at the estimated root joint position by inversely projecting the estimated 2D pose heatmaps $\boldsymbol{H}^{2D}$. The spatial size of the feature volume $\boldsymbol{V}_{PEN}^{in} \in [0,1]^{X' \times Y' \times Z' \times N}$ is set to be $2m \times 2m \times 2m$ which is sufficiently large to cover people in arbitrary poses. We set $X' = Y' = Z' = 64$, so approximately each voxel has a size of 30mm.

We feed $\boldsymbol{V}_{PEN}^{in}$ to a 3D Pose Estimation Network (PEN) to estimate 3D heatmaps $\boldsymbol{H}_{PEN} \in [0,1]^{X' \times Y' \times Z' \times N}$ for N joints, including the root joint. For

each joint k, the 3D location $\boldsymbol{J}_k$ can be obtained using the integration trick proposed in [32] to reduce quantization error:

$$\boldsymbol{J}_k = \sum_{x=1}^{X'} \sum_{y=1}^{Y'} \sum_{z=1}^{Z'} (x, y, z) \cdot \boldsymbol{H}_{PEN,k}(x, y, z). \tag{7}$$

We train PEN using the L_1 loss as in [33]:

$$\mathcal{L}_{PEN} = \frac{1}{N} \sum_{k=1}^{N} ||\boldsymbol{J}_k - \widetilde{\boldsymbol{J}}_k||_1. \tag{8}$$

PEN has the same network structure as the 3D network in REN, and the weights are shared for different people. We find that by estimating the absolute 3D locations of all joints, it further improves the absolute position of the root joint compared to the REN output.

5   Experiments

5.1   Implementation Details

General Details The 2D backbone for estimating boxes and 2D pose heatmaps is ResNet-152 [9]. The backbone for estimating depth map from heatmaps is ResNet-18 [9]. The input image is resized to 512×960 and the size of the resulting heatmap is 128×240. We adopt Adam [13] as optimizer, the learning rate is 1×10^{-4} and the batch size is 24. The 2D backbone network is trained for 40 epochs. The depth estimator, REN and PEN are trained end-to-end with synthetic heatmaps and their GT 3D poses for 20 epochs.

Training Data For the experiments on CMU Panoptic [12], the backbone is trained on the combined COCO [16] and CMU Panoptic datasets. When synthesizing AGR training data, the absolute 3D poses and camera views are the same as the original training data in Panoptic dataset. For the MuPoTS-3D [22] experiment, the backbone is trained on the combined COCO and MuCo-3DHP [22] datasets. When synthesizing AGR training data, the relative 3D poses and camera views are from MuCo-3DHP, and we randomly place the poses in 3D space to enhance the diversity of the training data.

5.2   Comparison to the State-of-the-arts

CMU Panoptic [12] Following [42], we use the video sequences of camera 16 and 30 as training and test data which consist of four activities, *i.e.* Haggling, Mafia, Ultimatum, Pizza. We show the results in Table 2. We achieve better results than previous methods except for HMOR [37] which uses the GT depth when estimating the absolute 3D pose. Besides, our method achieves the best performance in the Pizza sequence which is not included in the training data. This partially validates the generalization ability of our method.

Table 2: Comparison to the state-of-the-art methods on the CMU Panoptic dataset. The metric is MPJPE (mm). [37]† uses the GT depth when estimating absolute 3D poses so it is not fairly comparable to other methods. Our method does not use paired images and poses data for training but it still achieves smaller errors than other methods.

Method	Haggling	Mafia	Ultimatum	Pizza	Mean ↓
Li *et al.* (**HMOR**) [37]†	50.9	50.5	50.7	68.2	55.1
PoPa *et al.* [27]	217.9	187.3	193.6	221.3	203.4
Zanfir *et al.* [42]	140.0	165.9	150.7	156.0	153.4
Moon *et al.* (**RootNet**) [23]	89.6	91.3	79.6	90.1	87.6
Zanfir *et al.* [43]	72.4	78.8	66.8	94.3	78.1
Zhen *et al.* (**SMAP**) [47]	<u>63.1</u>	**60.3**	<u>56.6</u>	<u>67.1</u>	<u>61.8</u>
Ours	**54.1**	<u>61.6</u>	**54.6**	**65.4**	**58.9**

MuPoTS-3D [22] Our method is trained on the synthetic data where the 3D poses are from the MuCo-3DHP dataset. The rest methods are trained end-to-end on paired images and 3D poses from the MuCo-3DHP dataset. Following the standard practice on this dataset, the metric of percentage of correct keypoints (3DPCK) is used to measure the estimation results. The results are shown in Table 3. PCK_{abs} measures the accuracy of absolute pose and PCK_{root} measures the accuracy of the root joint. We can see that our method achieves significantly better depth and pose estimation results than the state-of-the-arts. It validates that our absolute 3D depth estimation method has strong generalization capability. We hope to emphasize the importance of the results because it means the method has the potential to be applied in the wild environments.

5.3   Ablation Study

In this section, we will evaluate the impact of our proposed modules and the training strategies to the estimation results of the root joint.

Root Estimation Network We first introduce two baselines. Baseline (a) analytically computes the 3D root position based on the 2D position and the estimated depth map. The baseline (b) uses the 3D root estimator to estimate the 3D root position on top of (a). The results on the CMU Panoptic and MuPoTS-3D datasets are shown in Table 4. By comparing the results of (a) and (b), we can see that the root depth error has a significant reduction which validates the effectiveness of the 3D root estimator.

Pose Estimation Network By comparing the results of (b) and (c) in Table 4, we can see that depth estimation can be notably improved by PEN, which leverages the rest of the body joints to refine the root joint. Another reason

Table 3: Comparison to the state-of-the-art methods on MuPoTS-3D. *Matched people* only computes accuracy for GT poses which are matched to predictions and *All people* computes accuracy for all GT poses in the dataset. The methods are not strictly comparable because they may have different backbones.

Method	Matched people		All people
	PCK_{abs} ↑	PCK_{root} ↑	PCK_{abs} ↑
Moon *et al.* (**RootNet**) [23]	31.8	31.0	31.5
Lin *et al.* (**HDNet**) [15]	35.2	39.4	-
Zhen *et al.* (**SMAP**) [47]	38.7	45.5	35.4
Veges *et al.* [35]	39.6	-	37.3
Sarandi *et al.* [30]	40.5	-	38.4
Li *et al.* (**HMOR**) [37]	-	-	43.8
Guo *et al.* [8]	39.6	-	39.2
Ours	**47.0**	**53.5**	**44.0**

for the improvement is that the quantization error is reduced by computing continuous root locations via the integral trick.

Depth Based Feature Projection As stated in Section 4.2, we construct the 3D feature volume by projecting 2D pose heatmaps based on the estimated depths and bounding boxes. We compare it to a baseline which naively projects the heatmaps to all voxels as in [11, 33] and find that MRPE and $MRPE_z$ of our method are significantly better than the baseline (204.0mm *vs.* 253.2mm, and 159.8mm *vs.* 210.5mm) on the MuPoTS-3D dataset.

Table 4: Ablation study on Root Estimation Network (REN) and Pose Estimation Network (PEN) in our method. We report the MRPE and $MRPE_z$ (mm) on the test set of CMU Panoptic and MuPoTS-3D dataset.

Method	REN	PEN	CMU Panoptic		MuPoTS-3D	
			MRPE ↓	$MRPE_z$ ↓	MRPE ↓	$MRPE_z$ ↓
(a) DE	2D	✗	113.9	104.1	282.0	245.4
(b) REN	2D+3D	✗	115.7	93.6	217.5	165.0
(c) Ours	2D+3D	✓	**97.0**	**86.0**	**204.0**	**159.8**

Training Data Generation Strategies We compare several training data generation strategies for different scenarios. First, when we have little knowledge about the testing camera view point, we can only uniformly sample camera views for the virtual camera to synthesize training data. The results are shown in Table

5. We can see that by training a universal model, our method achieves smaller depth estimation errors than the model trained on a single camera.

Second, if we know the camera view point in the testing environment, we can generate training data specifically for it. If we train a model for camera 19, then the testing error on camera 19 will be decreased significantly to 62.3mm. This is helpful when the camera is fixed, *e.g.* deployed at home for child or elderly care. In this case, we can get very accurate estimation results. We have similar observations for the pose factor. In particular, if we have the pose information for the people who are going to appear in the camera, then we can generate AGR using those poses. In that case, the testing accuracy will also be significantly improved. For example, if we train the model using the poses from the MuPoTS-3D dataset, then the PCK_{abs} will be improved from 44.0% to 50.4%.

Table 5: Impact of training data generation strategies. The metric is $MRPE_z$ (mm). We can see that by training a universal model, our method achieves smaller depth estimation errors than the model trained on a single camera.

Training Camera View	Test Camera view		
	14	16	19
14	-	170.3	294.6
16	454.6	-	198.6
19	621.3	385.1	-
Random	**273.8**	**134.0**	**155.2**

5.4   Qualitative Results

Fig. 4(a) shows some estimation results for images from the COCO dataset and MuPoTS-3D dataset. Since the camera parameters are not provided in the COCO dataset, we choose a general focal length (*i.e.* 1400) and assume the pitch angle of the camera is zero. We can see that our approach not only obtains accurate 3D pose for each person in the image but also estimates their absolute depth correctly in the 3D space. In particular, the model is robust to pose and background variations. For instance, in the baseball example, we get correct depth estimate for the person in sitting posture. Fig. 4(b) and (c) show some typical failure cases. In Fig. 4(b), the man in the red circle is occluded and truncated, while in Fig. 4(c), the little girl is much shorter than the people in the training dataset. In the future, to address the second issue, we will study the possibility of adding another parallel branch in the 2D network to estimate a correction factor to refine the estimated depth of each person.
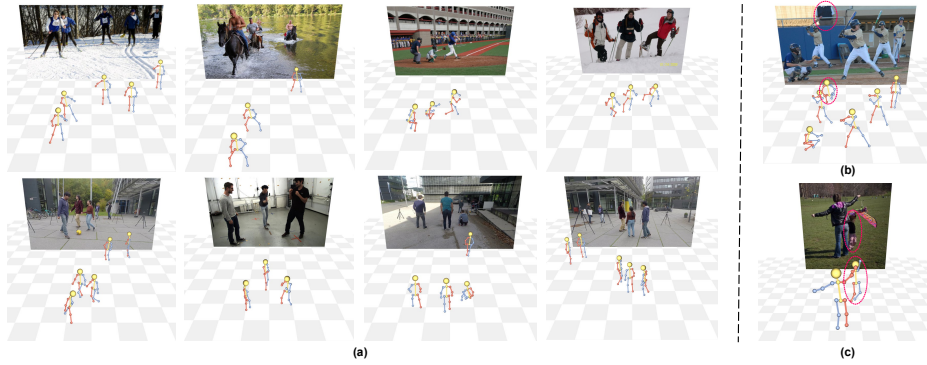
Fig. 4: (a) Some pose estimation results on the COCO (top) and MuPoTS-3D (bottom) datasets. (b) Typical failure cases due to occlusion and truncation. (c) Typical failure cases due to the person is shorter than those in the training dataset. As a result, the estimated depth is larger than GT.

6   Conclusion

In this work, we present a systematic study of the generalization problem of 3D pose estimation. We hope the study can inspire future works to consider the generalization aspect when designing and evaluating new models. Then we present *VirtualPose*, an approach for absolute 3D human pose estimation, which does not require paired images and 3D poses for training. In particular, part of the network is trained on abundant 2D datasets such as COCO and the rest are trained on synthetic datasets. The decoupling strategy helps the approach avoid over-fitting to small training data. As a result, it can be flexibly adapted to a new environment with minimal human effort. Our method achieves significantly better results than the existing methods on the benchmark datasets especially when the training and test datasets have different distributions.

6.1   Future Work

First, in the current implementation, AGR only consists of 2D joint heatmaps and person-level bounding boxes which are powerful to recover the absolute positions of the root joints. But we can actually explore finer-grained AGRs such as segmentation maps and occlusion relationship maps which can benefit more to relative pose estimation. Second, we only used the 3D poses from the Panoptic and MuCo-3DHP dataset for generating AGR training data. But we can actually generate more poses by manipulating joint angles which may further improve the results.

References

1. Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV. pp. 387–404. Springer (2020) 4
2. Chang, J.Y., Moon, G., Lee, K.M.: Absposelifter: Absolute 3d human pose lifting network from a single noisy 2d human pose. CoRR (2019) 3
3. Cheng, Y., Wang, B., Tan, R.: Dual networks based 3d multi-person pose estimation from monocular video. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 2
4. Ci, H., Ma, X., Wang, C., Wang, Y.: Locally connected network for monocular 3d human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1
5. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: ICCV. pp. 2262–2271 (2019) 1
6. Dabral, R., Gundavarapu, N.B., Mitra, R., Sharma, A., Ramakrishnan, G., Jain, A.: Multi-person 3d human pose estimation from monocular images. In: 3dv. pp. 405–414. IEEE (2019) 4
7. Fabbri, M., Lanzi, F., Calderara, S., Alletto, S., Cucchiara, R.: Compressed volumetric heatmaps for multi-person 3d pose estimation. In: CVPR. pp. 7204–7213 (2020) 4
8. Guo, Y., Ma, L., Li, Z., Wang, X., Wang, F.: Monocular 3d multi-person pose estimation via predicting factorised correction factors. Computer Vision and Image Understanding (CVIU) p. 103278 (2021) 5, 12
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 5, 10
10. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. PAMI **36**(7), 1325–1339 (2013) 2, 4
11. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: ICCV (2019) 12
12. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV. pp. 3334–3342 (2015) 2, 4, 5, 7, 10
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 10
14. Li, S., Ke, L., Pratama, K., Tai, Y.W., Tang, C.K., Cheng, K.T.: Cascaded deep monocular 3d human pose estimation with evolutionary training data. In: CVPR. pp. 6173–6183 (2020) 3
15. Lin, J., Lee, G.H.: Hdnet: Human depth estimation for multi-person camera-space localization. In: ECCV. pp. 633–648. Springer (2020) 2, 4, 5, 12
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 2, 6, 10
17. Ma, X., Su, J., Wang, C., Ci, H., Wang, Y.: Context modeling in 3d human pose estimation: A unified perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6238–6247 (2021) 1
18. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV. pp. 601–617 (2018) 3, 4

19. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV. pp. 2640–2649 (2017) 1

20. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV. pp. 506–516. IEEE (2017) 2, 4, 6

21. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. TOG **39**(4) (2020) 4

22. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 3DV. pp. 120–130. IEEE (2018) 2, 4, 5, 6, 7, 10, 11

23. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: ICCV. pp. 10133–10142 (2019) 2, 4, 5, 6, 7, 11, 12

24. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: CVPR. pp. 2823–2832 (2017) 1

25. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR. pp. 7025–7034 (2017) 1

26. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR. pp. 7753–7762 (2019) 3

27. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2d and 3d human sensing. In: CVPR. pp. 6289–6298 (2017) 11

28. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: CVPR. pp. 3433–3441 (2017) 4

29. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. PAMI **42**(5), 1146–1161 (2019) 4

30. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. IEEE Transactions on Biometrics, Behavior, and Identity Science **3**(1), 16–30 (2020) 2, 12

31. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV **87**(1-2), 4 (2010) 2, 4

32. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV. pp. 529–545 (2018) 1, 10

33. Tu, H., Wang, C., Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: ECCV. pp. 197–212. Springer (2020) 2, 3, 4, 9, 10, 12

34. Véges, M., Lőrincz, A.: Absolute human pose estimation with depth prediction network. In: IJCNN. pp. 1–7. IEEE (2019) 5

35. Véges, M., Lőrincz, A.: Multi-person absolute 3d human pose estimation with weak depth supervision. In: International Conference on Artificial Neural Networks. pp. 258–270. Springer (2020) 12

36. Wandt, B., Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: CVPR. pp. 7782–7791 (2019) 4

37. Wang, C., Li, J., Liu, W., Qian, C., Lu, C.: Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In: ECCV. pp. 242–259. Springer (2020) 2, 4, 5, 10, 11, 12

38. Wang, C., Wang, Y., Lin, Z., Yuille, A.L.: Robust 3d human pose estimation from single images or video sequences. IEEE transactions on pattern analysis and machine intelligence **41**(5), 1227–1241 (2018) 1
39. Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W.: Robust estimation of 3d human poses from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2361–2368 (2014) 1
40. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: 3d interpreter networks for viewer-centered wireframe modeling. International Journal of Computer Vision **126**(9), 1009–1026 (2018) 3, 4
41. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: CVPR. pp. 5255–5264 (2018) 4
42. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: CVPR. pp. 2148–2157 (2018) 2, 10, 11
43. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3d sensing of multiple people in natural images. NIPS **31**, 8410–8419 (2018) 2, 11
44. Zhang, Y., Wang, C., Wang, X., Liu, W., Zeng, W.: Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. T-PAMI (2022) 9
45. Zhang, Z., Wang, C., Qin, W., Zeng, W.: Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In: CVPR. pp. 2200–2209 (2020) 4
46. Zhang, Z., Wang, C., Qiu, W., Qin, W., Zeng, W.: Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. IJCV **129**(3), 703–718 (2021) 4
47. Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., Zhou, X.: Smap: Single-shot multi-person absolute 3d pose estimation. In: ECCV. pp. 550–566. Springer (2020) 2, 4, 5, 11, 12
48. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: ICCV. pp. 398–407 (2017) 4
49. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019) 5, 8
50. Zhu, L., Rematas, K., Curless, B., Seitz, S.M., Kemelmacher-Shlizerman, I.: Reconstructing nba players. In: ECCV. pp. 177–194. Springer (2020) 4