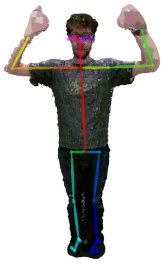
PanopTOP: a framework for generating viewpoint-invariant human pose estimation datasets

Nicola Garau[1], Giulia Martinelli[1], Piotr Bródka[1], Niccolò Bisagno[1], and Nicola Conci[1]

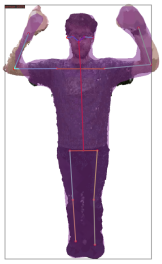
{*nicola.garau, giulia.martinelli-2*}@*unitn.it*, piotrbrodka95@gmail.com,
{niccolo.bisagno,nicola.conci}@unitn.it
[1]University of Trento, Via Sommarive, 9, 38123 Povo, Trento TN

Abstract

Human pose estimation (HPE) from RGB and depth images has recently experienced a push for viewpoint-invariant and scale-invariant pose retrieval methods. Current methods fail to generalize to unconventional viewpoints due to the lack of viewpoint-invariant data at training time. Existing datasets do not provide multiple-viewpoint observations and mostly focus on frontal views. In this work, we introduce PanopTOP, a fully automatic framework for the generation of semi-synthetic RGB and depth samples with 2D and 3D ground truth of pedestrian poses from multiple arbitrary viewpoints. Starting from the Panoptic Dataset [15], we use the PanopTOP framework to generate the PanopTOP31K dataset, consisting of 31K images from 23 different subjects recorded from diverse and challenging viewpoints, also including the top-view. Finally, we provide baseline results and cross-validation tests for our dataset, demonstrating how it is possible to generalize from the semi-synthetic to the real-world domain. The dataset and the code will be made publicly available upon acceptance.

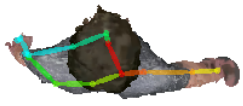


(a) OpenPose front

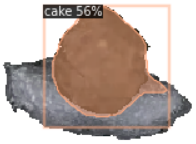


(b) MaskRCNN front

(c) HMR front



(d) OpenPose top



(e) MaskRCNN top

(f) HMR top

Figure 1: OpenPose, MaskRCNN and Human Mesh Recovery baselines (front, top views). All the methods perform very well on front and side views (Fig. 1a, 1b, 1c). However, when dealing with top-view images, current methods fail to correctly retrieve the human pose (Fig. 1d, 1f) or to even recognise the object as a human body (Fig. 1e).

Dataset	RGB	Depth	Top-view	Multi-View	2D Pose GT	3D Pose GT	Camera parameters
PanopTOP31K	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
ITOP	N	**Y**	**Y**	**Y**	N	**Y**	**Y**
EVAL	N	**Y**	N	N	N	**Y**	N
TVPR	**Y**	**Y**	**Y**	N	N	N	N
TVPR 2	**Y**	**Y**	**Y**	N	N	N	N
K2HPD	N	**Y**	N	N	N	**Y**	N
UBC3V	N	**Y**	N	**Y**	N	**Y**	**Y**
Human3.6M	**Y**	N	N	**Y**	N	**Y**	**Y**

Table 1: Reference datasets for multi-view and viewpoint-invariant networks training. Only few datasets propose true top-view ground truth data, and most of them mainly focus on depth images, discarding the RGB component.

1. Introduction

In the field of human pose estimation (HPE), depth and RGB sensors are commonly employed in a wide range of applications, from robotics to immersive entertainment and from surveillance to smart spaces. [10, 2, 9]. Such a diverse application range requires cameras to capture humans from a wide variety of different angles. Thus, HPE frameworks should be able to retrieve the body pose from multiple different viewpoints. Currently, existing human pose estimation methods [10, 35, 2, 16, 21, 20] achieve good performances from many different camera viewpoints, but the most challenging ones. As shown in Fig. 1, good performances are achieved when retrieving the human pose from front-view images, and poor results can be obtained when dealing, for example, with the top-view. In this paper, we introduce a complete pipeline, called PanopTOP, which is based on computer graphics and that allows generating new RGB, depth, and pose samples from arbitrary viewpoints from the raw point cloud data. Our work aims at solving the following issues:

- ground truth alignment: we provide pixel-perfect aligned RGB and depth images regardless of the viewpoints, as well as the 2D and 3D ground truth pose;
- we encourage the usage of true multi-view cameras, allowing to obtain ground truth data from virtually every viewpoint and specifically the top-view one;
- our method employs a full pinhole camera model, allowing us to customize every aspect of the camera parameters, including intrinsic and extrinsic parameters. The remaining data (RGB, depth images, and reprojected 2D joints ground truth) is automatically changed according to the changes in the camera parameters.

To prove the effectiveness of our PanopTOP framework, we introduce PanopTOP31K, a training dataset specifically built for viewpoint invariant human pose estimation from depth and RGB images, consisting of 31K images of 23 different subjects recorded from diverse and challenging viewpoints. By using our PanopTOP method, it is possible to configure virtual cameras while fixing the existing 3D ground truth. To the best of our knowledge, PanopTOP31K is the first dataset that provides both top-view RGB images, as well as the corresponding 3D and 2D pose ground truth. Annotated poses in top-view RGB datasets are not available because of the difficulty of annotation, mostly due to occlusions. In Table 1 we show how the PanopTOP31K dataset provides the most complete and diverse set of poses and ground truths when compared with similar datasets. We argue that the complete set of multi-view RGB and depth images along with 2D and 3D ground truth provide HPE researchers with the necessary data for the development of viewpoint-invariant frameworks.

We create the PanopTOP31K dataset starting from the six-degrees-of-freedom (6DoF) videos of the Panoptic dataset [15]. The dataset provides the pose ground truth for each video frame. Since 6DoF videos provide a 3D model of the scene, it is possible to generate a virtually infinite number of new 2D semi-synthetic RGB and depth images, in a bullet time fashion and from multiple viewpoints, simultaneously. In this way, we can create realistic videos on the fly, captured from different angles. To prove the suitability of the dataset for further developments, we show how baseline algorithms [25] trained on our PanopTOP31K can generalize on real data leading to improved results on multiple datasets.

Our contributions can be summarised through the following key steps:

i. We propose a method to generate new RGB and depth datasets with a virtually infinite number of semi-synthetic viewpoints, called PanopTOP.

ii. We propose a multi-view dataset, called PanopTOP31K, which consists of 31 thousand poses of 23 different subjects, rendered from the front, side, and top viewpoints in both RGB and depth domains.

iii. We provide baseline results for the novel Panop-

TOP31K on RGB images.

iv. We show the improvement of performances on multiple views given by our dataset cross-validating on different scenarios and viewpoints.

2. Related Work

In this section, we explore the state-of-the-art for human pose estimation in both RGB and depth domains, as well as the available datasets providing multiple viewpoints for each pose.

2.1. State-of-the-art methods

The most common form of HPE consists of solving the task of estimating 2D joints and their connection, starting from RGB images and videos. However, HPE can be carried out even in different domains, such as depth images [6, 24, 44], LiDARr [8] or even radio signals [47]. Additionally, human pose estimation has recently shifted towards 3D estimation, by lifting 2D human poses to 3D [46, 4, 29], as well as with end-to-end approaches [34, 31]. Among the most recent developments of HPE, we can find human mesh recovery, which deals with the problem of retrieving the human pose from images or videos in terms of a fully rigged 3D mesh [22, 19]. Granularity also plays an important role in HPE, with an increasing number of methods extending the pose to hand [12, 42], feet [43] or even face pose [14]. In recent years, research on human pose estimation has been focusing largely on single views, using either RGB [2, 10] or depth images [9, 25], as shown in Table 1.

HPE from depth images. Viewpoint-invariant HPE methods have been focusing exclusively on depth images [9, 25, 42] from top-view and side-view, and only a limited number of works address this problem. This might be due to the lack of datasets containing real or synthetic labeled depth data. Additionally, the majority of the depth-based datasets are small; this does not match the requirements of deep learning, which requires large amounts of data for proper training. In addition, they do not provide an accurate ground truth most of the time, rather automatic annotations, *i.e.* the position of the body joints is predicted using pose detectors such as [30]. The work by Shotton et al. [30] has been decisive in the human pose estimation from depth maps field, especially for its application in many successful commercial scenarios, such as the *Microsoft Kinect* and its SDKs. The authors propose a method for human pose estimation based on a *Random Forest* trained on a synthetic dataset (not publicly available), by classifying each pixel into body parts. The 3D position of joints is predicted from the labeled depth map with a local mode-finding approach based on Mean Shift. In very specific scenarios, such as strict front viewpoints, these methods obtain reasonable accuracy results and real-time performance, given the context of the application (gaming, interactive applications). In [11], Hernandez-Vela *et al.* propose an object segmentation framework using depth maps combining the use of Random Forest and Graph-cuts theory for the segmentation of human limbs in-depth maps. Firstly, Random Forest assigns a set of labeled probability for each depth sample belonging to a set of possible object labels. Then, with the use of Graph-cuts, the precedent procedure is optimized both locally, spatially, and temporally. Ye *et al.* in [45] extract a point cloud from a depth map, and after the point cloud has been cleaned, transformed in frame coordinates, the body pose is predicted.

HPE from RGB images. In literature we can find two classes of approaches that extract human pose from RGB images. The bottom-up methods [1, 3, 26] detect firstly the human parts and then locate them in each object, and top-down methods [48, 36, 39] locate the key points in the human body and then compose the single parts into a person. Tekin *et al.* in [33] recover the 3D pose of people from consecutive frames of a video. They use at the same time appearance and motion information and regress directly from short sequences of frames to 3D poses in the central one. However, this method is limited to image sequences. Most recent methods exploit 2D pose estimation using CNNs [39, 27, 41, 1, 32]. The well-known work done by Cao *et al.*[1] detects the 2D pose of multiple people in an image. This approach associates body parts with individuals through the use of a parametric representation called *Part Affinity Fields*. In the context of multi-person pose estimation, one of the most recent works is the one proposed by Duan et al. [5] that implements a solution named location-sensitive network (LSNet) that unifies three recognition tasks like object detection, instance segmentation, and human pose estimation. The authors also present a novel loss function called *cross-IOU loss* that calculates the cross-IOU of each anchor-landmark to approximate the global IOU between prediction and ground-truth.

2.2. Datasets

In the context of RGB images, there is a lack of datasets providing multiple viewpoints, and in particular the top-view viewpoint. For example, common large HPE datasets in literature such as Human3.6M [13] and the Panoptic Dataset [15] provide RGB images from multiple views, still lacking the top-view component. Other datasets, such as K2HPD Body Pose Dataset [38] and ITOP [9] only provide top-view and side-view depth images, lacking the matching RGB ground truth data. TVPR and TVPR2 datasets [23, 28] also provide a top-view point of the scene, but they do not provide information about 3D joints, making the dataset

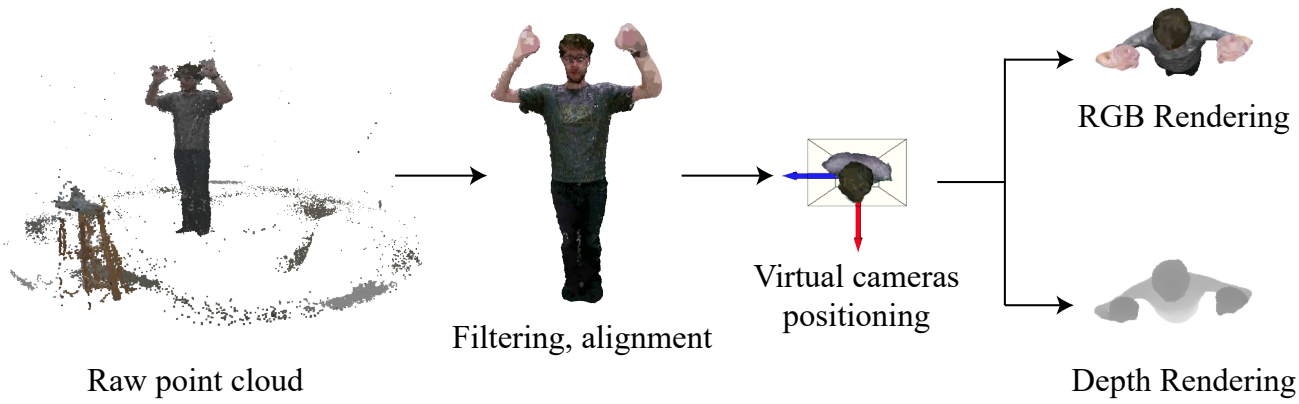


Figure 2: PanopTOP rendering process: the raw point cloud is filtered and aligned with the chosen coordinate system. Virtual cameras are positioned into the scene for simultaneous RGB + depth rendering.

not suitable to solve HPE-related tasks. The lack of multi-view datasets in the RGB domain leads to popular out-of-the-box networks, such as OpenPose [2] and MaskRCNN, [10] experiencing a big decrease in performance whenever the viewpoint is changed. A possible solution to the aforementioned annotation problem is to rely on fully synthetic datasets to generate data from custom viewpoints. However, many works [7, 37] show that the gap between visual realism in computer-generated and real-world images contributes to an even bigger gap in the models' prediction performance. Relying on photo-realistic rendering helps to mitigate this issue. However, many other aspects contribute to the overall perceived realism, especially when dealing with humans and human poses. Our perception of reality is influenced by unrealistic or inconsistent body proportions through time, fake lighting, physics, and slightly off skeletal movements. The same effect is shown when training HPE models on synthetic data and testing on real-world images, or vice versa. Other works [40] employ synthetic data to augment the already available datasets, but the same issues apply, even if to a lesser extent. In this work we adopt a hybrid solution, relying on real-world 3D human scans to generate new semi-synthetic data. We thus preserve the photo-realism of the rendered scene while maintaining all the advantages typical of fully synthetic approaches.

3. The PanopTOP framework

In this section we present PanopTOP, a fully automated pipeline for creating multi-view HPE datasets, starting from real-world 3D data and ground truth joints. Next, we employ the PanopTOP framework to create the PanopTOP31K dataset. The main advantages of our method are: (i) it automatically provides RGB and depth images, along with the 2D and 3D ground truth, requiring the user's input only when positioning the virtual cameras; (ii) it is highly customizable, meaning that it allows to precisely tune each camera, including their intrinsic and extrinsic parameters, as well as the desired RGB and depth output quality; (iii) it outputs data suitable for multiple tasks, such as 2D/3D human pose estimation, detection, segmentation, view synthesis and others in both RGB and depth domain.

The proposed framework generates RGB, depth, 2D and 3D joints ground truth data, starting from raw point clouds. We used the Panoptic dataset [15], although similar datasets can also be used. The complete pipeline is shown in Fig. 2 and includes the following steps:

1. point cloud retrieval and coordinate system setup;
2. skeleton-based point cloud filtering;
3. mesh reconstruction;
4. virtual cameras positioning;
5. hidden points removal;
6. rendering.

Each processing step is further detailed in the next subsections. The proposed method allows for a manifold of different configuration parameters, such as the number of virtual cameras to be used for the image generation and their global position, the density of point clouds and meshes, and the resolution of the output images. Furthermore, it also takes care of all the steps required to generate the dataset, from fetching the raw point clouds to saving the desired dataset to memory. Moreover, since our architecture leverages high-speed GPU operations, it could be used to automatically generate new batches on-the-fly, and directly use them as input to train a neural network, without saving them to memory in advance.

3.1. Point cloud retrieval and coordinate system setup

Starting from the Panoptic dataset [15], each point cloud is retrieved and then transformed to a center of origin, scale, and coordinate system of choice. For consistency, we adopt the same coordinate system and scale used in the ITOP dataset [9] by default. The center of origin is selected based on the ground truth pose data associated with the point cloud. At this stage, no additional filtering is employed, the output is a raw point cloud with many outliers and aligned to a chosen coordinate system.

3.2. Skeleton-based point cloud filtering

The ground truth 3D pose data is used to compute an axis-aligned 3D bounding box containing the subject. All the points that do not belong to the subject (Fig. 2) are outside of the bounding box, and are thus removed. An additional skeleton-based point cloud filtering based on the L2 distance d between the closest joint j in the 3D skeleton and each 3D point p_i is then applied:

$$d = \sqrt{\sum_{i=0}^{P} (p_i - j)^2} \tag{1}$$

Subsequently, statistical and sphere radius outliers are removed, only keeping the points that belong to the subject, thus greatly reducing noise.

3.3. Mesh reconstruction

Here we describe how 3D meshes can be reconstructed from the filtered point clouds if needed. They can be useful as ground truth for a more efficient 3D model analysis since the mesh provides more structured information and a smoother texture. The point cloud's vertex normals are estimated by looking for adjacent points and using covariance analysis to calculate their principal axis. Then the normalized point cloud is converted into a 3D mesh via Screened Poisson Surface Reconstruction [18]. Finally, surface subdivision and mesh smoothing are employed. These steps allow to obtain a smoother surface and thus a better rendering of both the RGB and depth outputs.

3.4. Virtual cameras positioning

A configuration file is designed to create virtual cameras with user-defined intrinsic and extrinsic parameters, for later rendering of both RGB images and depth maps. Users can also manually adjust the camera position in an interactive visualization window containing a preview of the rendering.

3.5. Hidden points removal

Optional hidden points removal is performed via Direct Visibility of Point Sets [17]. Since the input point cloud may be a combination of multiple viewpoints clouds, as in the Panoptic dataset, it may be necessary to remove occluded points to replicate the standard format of the majority of the other datasets. By default, we keep this option enabled to promote consistency with the ITOP dataset [9].

3.6. Rendering

RGB and depth images are finally rendered for each camera and for each point cloud sample in the dataset. The 2D and 3D ground truth of the scene is automatically generated starting from the 3D pose matrices and the extrinsic and intrinsic camera matrices. By default, we also convert the 19-joints skeleton to the 15-joints model to be consistent with the ITOP dataset [9].

4. The PanopTOP31K dataset

Our method shown in Sec. 3 allows for the generation of an arbitrary number of viewpoints for each point cloud. Starting from the Panoptic dataset, we apply our pipeline to generate a new dataset, called PanopTOP31K. The dataset contains approximately $30K$ RGB images, $30K$ depth maps, $10K$ filtered point clouds, and $10K$ 3D meshes from 23 different subjects recorded from the front, side, and top view ($\sim 10K$ RGB images for each viewpoint). Each RGB image and depth image have size 256×256 with depth 3. The provided pose ground truth employs the 15 joints skeleton model as in [9].

5. Experiments

In this section, we show the results obtained by some popular out-of-the-box human pose estimation networks, both for the front, side, and top view on our PanopTOP31K dataset on RGB images. Then, we validate our semi-synthetic dataset showing how it achieves good results when used for data augmentation and domain adaptation on depth images.

5.1. Baselines on RGB images

OpenPose [2, 1], MaskRCNN [10] and HMR [16, 22] are three popular methods for HPE in the RGB domain. We take the off-the-shelf pre-trained networks of all the baseline algorithms for testing on our new dataset, PanopTOP31K. As shown in Figs. 1a, 1b, 1c, we obtain good results on the side views. However, when dealing with top view images, all the methods fail in detecting the pose from more than 90% of the dataset images. For example, MaskRCNN misclassifies as 'cake' a subject as seen from the top view (Fig. 1e), while OpenPose fails to produce a coherent skeletal structure (Fig. 1d), despite being trained on the Panoptic dataset. A similar issue is encountered in HMR, which fails to correctly fit a mesh to the top-view image

	Experiment	Head	Neck	Shoulders	Elbows	Hands	Torso	Hips	Knees	Feet
(a)	**[I],[I],[I]**	99.50	99.60	99.05	**97.90**	**90.80**	***100.00***	98.55	95.20	87.15
(b)	**[I],[I],[P]**	96.60	97.90	93.80	76.10	63.60	97.80	89.90	84.60	46.50
(c)	**[I],[I+P],[P]**	97.20	98.10	95.45	77.15	59.10	98.00	90.25	70.20	35.80
(d)	**[I+P],[I+P],[P]**	**98.50**	**99.70**	**99.70**	***98.20***	***90.90***	**99.70**	**99.40**	95.80	***95.55***
(e)	**[P],[P],[P]**	**98.50**	**99.70**	**99.70**	97.80	90.85	99.60	99.35	***96.30***	95.45
(f)	**[P],[P],[I]**	99.50	99.50	98.10	93.90	61.45	99.30	94.85	75.45	26.80
(g)	**[P],[I+P],[I]**	99.60	99.80	97.95	94.00	66.60	99.50	94.45	83.55	59.20
(h)	**[I+P],[I+P],[I]**	***100.00***	***100.00***	***100.00***	97.80	90.35	***100.00***	***99.55***	***96.30***	**89.35**

Table 2: Percentages of correctly detected joints for the ITOP and PanopTOP31K datasets in our 8 conducted experiments. Each experiment is identified by a letter (**a-h**) and a data split **[train],[validation],[test]** (**P** = PanopTOP31K, **I** = ITOP). Each value represents the percentage of joints with L2 distance smaller than a threshold $T = 0.2m$ from the ground truth. The top scores for each joint regarding tests on the ITOP dataset are highlighted in **blue**, while the PanopTOP31K ones are highlighted in **green**. The top overall scores for each joint are in ***italic***.

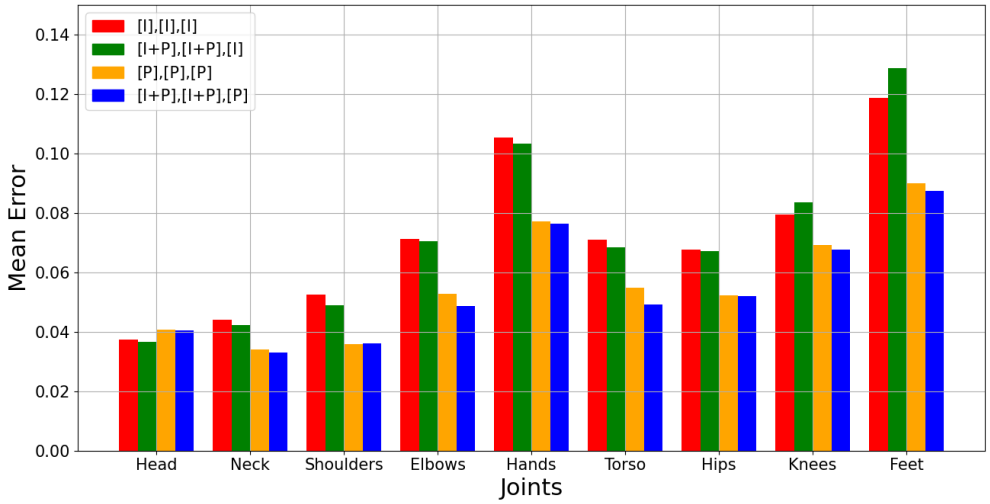


Figure 3: Mean per-joints errors in meters for ITOP and PanopTOP31K datasets, respectively, with (green, blue) and without (red, orange) training-wise augmentation. Red, green, yellow and blue bars correspond to experiments (a), (h), (e) and (d) respectively.

(Fig. 1f). This incorrect behavior explains that most human pose estimation networks are not trained to handle extreme viewpoints and thus they do not achieve viewpoint-invariance. Our method allows creating multi-view datasets for human pose estimation, that could be used to develop viewpoint invariant HPE networks in the RGB or depth domain.

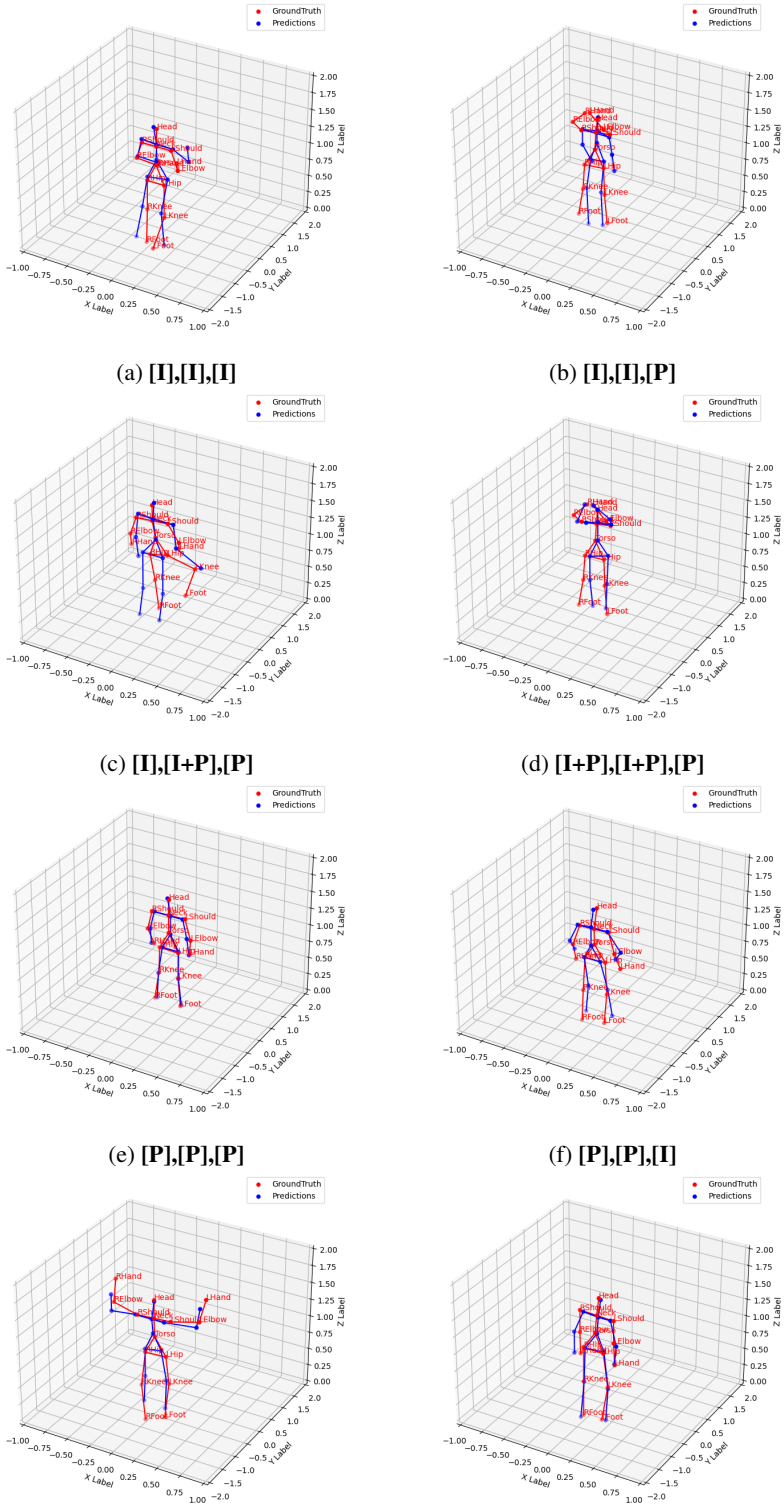
5.2. Dataset validation on depth images

We have shown how state-of-the-art methods for HPE on RGB images work on the PanopTOP31K dataset, failing in the case of top-view images. In this section, we focus on validating our dataset on depth images. Since human pose estimation already works well enough on front-view images, we focus our attention on a most difficult scenario, namely top-view. We use a vanilla version of the V2V network [25] on depth images for training and validating.

In the remainder of this section and in table 2 we use the following notation: **I** and **P** identify the ITOP and our PanopTOP31K dataset respectively. We use **[train],[validation],[test]** to indicate on which datasets the network has been trained, validated and tested.

We perform 8 different cross-validation experiments, from (a) to (h), as shown in Tab. 2. In Fig. 4 we show some qualitative results for the HPE task, while Tab. 2 and Fig. 3 report the percentage of correctly detected joints and the mean per-joint error respectively.

Experiment (a) shows how the network performs well when trained, validated, and tested on the ITOP dataset. At the same time, in (b), the same training and validation split shows a poor ability to generalize on a new dataset.



(a) **[I],[I],[I]** (b) **[I],[I],[P]**

(c) **[I],[I+P],[P]** (d) **[I+P],[I+P],[P]**

(e) **[P],[P],[P]** (f) **[P],[P],[I]**

(g) **[P],[I+P],[I]** (h) **[I+P],[I+P],[I]**

Figure 4:
Qualitative results on multiple **[train],[validation],[test]** data splits, corresponding to quantitative results in Table 2. As an example, the notation **[P],[I+P],[I]** means that the network has been trained on PanopTOP31K, validated on both ITOP and PanopTOP31K and tested on ITOP. Each experiment is identified by a letter (**a-h**).

In (c), we provide a diverse validation set, but we achieve a small gain and in some cases, we even worsen the performances on hands and feet estimation, as shown in Tab. 2. This happens because the ITOP dataset is not diverse enough and it tends to overfit the training data. Experiment (d) shows how adding our PanopTop31K split to the training set results in a substantial improvement in performances with respect to previous cases (a)(b)(c). The PanopTOP31K dataset is thus suitable for augmenting a real-world dataset and it leads to a performance improvement.

Experiment (e) proves that the network can correctly process our PanopTOP31K dataset with good results. In (f), we obtain much better results than (b), and in some cases, they are also comparable with (a). This proves how the PanopTOP31K dataset is more able to generalize to different data than the ITOP one, without overfitting. Adding the ITOP validation split as in (g) allows the network to improve its performances with respect to (f), thus proving the robustness to the overfitting of the network trained on our dataset. Finally, experiment (h) shows how the best performances on real data are obtained by augmenting a real-world dataset with our semi-synthetic one. Both (f) and (h) validate the ability of our semi-synthetic dataset to provide samples that are realistic enough for the network to generalize well on real-world depth images.

6. Conclusions

We presented PanopTOP, a new method for fully automatic multi-view datasets creation along with PanopTOP31K, the first multimodal RGB and depth dataset exhibiting challenging viewpoints for HPE. Our dataset allows for the training of viewpoint-invariant HPE networks from a manifold of data inputs (RGB images, depth images, point clouds, 3D meshes). Experiments on our semi-synthetic PanopTOP31K dataset show promising results on top-view HPE, obtaining comparable results with popular real-world datasets and improving network accuracy when used for data augmentation.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Bottom-up higher-resolution networks for multi-person pose estimation. *CoRR*, abs/1908.10357, 2019.

[4] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2262–2271, 2019.

[5] Kaiwen Duan, Lingxi Xie, Honggang Qi, Song Bai, Qingming Huang, and Qi Tian. Location-sensitive visual recognition with cross-iou loss. *CoRR*, abs/2104.04899, 2021.

[6] Andrea D'Eusanio, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Refinet: 3d human pose refinement with depth maps. 01 2021.

[7] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *European Conference on Computer Vision (ECCV)*, 2018.

[8] Michael Fürst, Shriya T. P. Gupta, René Schuster, Oliver Wasenmüller, and Didier Stricker. HPERL: 3d human pose estimation from RGB and lidar. *CoRR*, abs/2010.08221, 2020.

[9] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 160–177, Cham, 2016. Springer International Publishing.

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

[11] Antonio Hernandez-Vela, Nadezhda Zlateva, Alexander Marinov, Miguel Reyes, Petia Radeva, Dimo Dimov, and Sergio Escalera. Graph cuts optimization for multi-limb human segmentation in depth maps. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 726–732, 2012.

[12] Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. AWR: adaptive weighting regression for 3d hand pose estimation. *CoRR*, abs/2007.09590, 2020.

[13] 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, July 2014.

[14] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric CNN regression. *CoRR*, abs/1703.07834, 2017.

[15] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[16] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[17] Sagi Katz, Ayellet Tal, and Ronen Basri. Direct visibility of point sets. 26(3):24–es, July 2007.

[18] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3), July 2013.

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. *CoRR*, abs/1912.05656, 2019.

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[21] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *CoRR*, abs/1909.12828, 2019.

[23] Daniele Liciotti, Marina Paolanti, Emanuele Frontoni, Adriano Mancini, and Primo Zingaretti. *Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration*, pages 1–11. Springer International Publishing, Cham, 2017.

[24] Manuel J. Marín-Jiménez, Francisco J. Romero Ramírez, Rafael Muñoz-Salinas, and Rafael Medina Carnicer. 3d human pose estimation from depth maps using a deep combination of poses. *CoRR*, abs/1807.05389, 2018.

[25] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018.

[26] Alejandro Newell and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *CoRR*, abs/1611.05424, 2016.

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *CoRR*, abs/1603.06937, 2016.

[28] M. Paolanti, R. Pietrini, A. Mancini, E. Frontoni, and P. Zingaretti. Deep understanding of shopper behaviours and interactions using rgb-d vision. *Machine Vision and Applications*, 31(7-8), 2020.

[29] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. *CoRR*, abs/1904.01324, 2019.

[30] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304, 2011.

[31] Lucas Stoffl, Maxime Vidal, and Alexander Mathis. End-to-end trainable multi-instance pose estimation with transformers. *CoRR*, abs/2103.12115, 2021.

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *CoRR*, abs/1902.09212, 2019.

[33] Bugra Tekin, Xiaolu Sun, Xinchao Wang, Vincent Lepetit, and Pascal Fua. Predicting people's 3d poses from short sequences. 04 2015.

[34] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *CoRR*, abs/1911.07451, 2019.

[35] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[36] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013.

[37] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[38] Keze Wang, Shengfu Zhai, Hui Cheng, Xiaodan Liang, and Liang Lin. Human pose estimation from depth images via inference embedded multi-task learning. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 1227–1236, New York, NY, USA, 2016. Association for Computing Machinery.

[39] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *CoRR*, abs/1602.00134, 2016.

[40] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *CoRR*, abs/1804.06208, 2018.

[42] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 793–802, 2019.

[43] Tianxu Xu, Dong An, Yuetong Jia, and Yang Yue. A review: Point cloud-based 3d human joints estimation. *Sensors*, 21:1684, 03 2021.

[44] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys. Accurate 3d pose estimation from a single depth image. In *2011 International Conference on Computer Vision*, pages 731–738, 2011.

[45] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys. Accurate 3d pose estimation from a single depth image. In *2011 International Conference on Computer Vision*, pages 731–738, 2011.

[46] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. *CoRR*, abs/1904.03345, 2019.

[47] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018.

[48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019.