2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP) | 978-1-6654-7822-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IVMSP54334.2022.9816196

Monocular Weakly-Supervised Camera-Relative 3D Human Pose Estimation

Anestis Christidis, Christos Papaioannidis and Ioannis Pitas
Department of Informatics, Aristotle University of Thessaloniki, Greece
Email: canestis@csd.auth.gr, cpapaionn@csd.auth.gr, pitas@csd.auth.gr

***Abstract*—This paper presents a 3D human pose estimation framework based on Deep Neural Networks (DNNs). It builds upon existing weakly-supervised methods that predict 2D-3D correspondences and improves them by introducing a geometrical-alignment pre-processing step and a 3D skeleton-refinement post-processing step. The geometrical-alignment pre-processing step is applied on the ground-truth 3D human poses and transforms them, in order to enable the utilized 2D-to-3D skeleton mapping DNN to be efficiently trained in a weakly-supervised manner. The 3D skeleton-refinement post-processing step acts on the DNN outputs and enables the proposed 3D human pose estimation framework to predict the camera-relative 3D human poses. Experiments on the widely used public showed that the proposed framework managed to predict camera-relative 3D human poses with increased accuracy.**

I. Introduction

3D Human Pose Estimation (HPE) is a very important Computer Vision topic, especially for human-robot interaction, as it allows a robot to infer the exact location of all human limps relative to its position, which enables the successful interaction between them. Recent advances on autonomous systems, e.g., autonomous Unmanned Aerial Vehicles (UAVs), also require accurate 3D human pose estimation from monocular input, which consists in estimating the 3D coordinates of a predefined set of human body joints relative to the camera coordinate system, using a single RGB image as input. For example, in a real-world scenario where a human worker cooperates with a camera-equipped UAV, [1], [2], [3], the autonomous UAV is required to know the exact position of the human head, body and hands in the 3D world, in order to deliver a tool to the worker's hands while simultaneously keeping a safe distance from the worker's body and head.

Recent deep learning based 3D human pose estimation methods [4], [5], [6] utilized the power of Convolutional Neural Networks (CNNs) to directly predict 3D human poses from RGB images. However, their success usually depends on the availability of large-scale annotated datasets and often fail to successfully generalize to unseen test images, where the human is depicted under different background conditions (e.g., indoor/outdoor) or different camera viewpoints.

In order to overcome the first limitation, many approaches [7], [8], [9] utilized 2D skeletons to predict 3D human poses. That is, instead of directly predicting the 3D location of the human body joints from the RGB image, a two-step approach was adopted, where 2D skeletons are extracted from the input image using pretrained 2D body joints estimation methods [10] in the first step and in the second step, a Deep Neural Network (DNN) is trained to "lift" the extracted 2D skeletons to the corresponding camera-relative 3D skeletons. While this two-step approach managed to increase 3D human pose estimation accuracy, training the "lifting" DNN with 2D to 3D correspondences often leads to overfitting, since there is a limited availability of data annotated with their camera-relative 3D body joint locations.

In this direction, [11] trained a 2D-to-3D skeleton mapping DNN in a weakly supervised manner using Generative Adversarial Networks (GANs) [12], that was able to predict accurate 3D skeletons even from unseen camera viewpoints and motions.

While the weakly supervised training framework of [11] manages to overcome the overfitting problem, it requires all camera-relative 3D skeletons to be rotated, scaled and transformed in order to be aligned with a template, root-joint-relative 3D skeleton and thus, be suitable for the GAN training. As a result, the 2D-3D skeleton mapping DNN that is trained under the weakly supervised framework of [11] is only able to predicted such "aligned" 3D skeletons, rendering it incapable to estimate the actual camera-relative 3D body joints locations. In this direction, the proposed 3D human pose estimation framework incorporates a geometrical-alignment pre-processing step and a 3D skeleton-refinement post-processing step, which allow the trained DNN to accurately predict the camera-relative 3D body joints coordinates.

In short, the contributions of this paper are:

- a weakly-supervised framework to predict camera-relative 3D human poses from single RGB images,
- a geometrical-alignment pre-processing step that allows more efficient training of the 2D-3D skeleton mapping DNN,
- and a 3D skeleton-refinement post-processing step that enables estimating the camera-relative 3D human poses from the DNN outputs.

II. 2D skeleton-based 3D human pose estimation

Many recent 3D human pose estimation methods focus on directly estimating the 3D body joints coordinates from 2D skeletons in an end-to-end manner using DNNs [7], [13]. However, their impressive accuracy probably stems from memorization of the training dataset, which is typically very similar to the test set (e.g., regarding camera distance and viewpoints). To combat this, recent approaches [8], [9], [14] proposed online data augmentation methods to produce new 2D skeleton data during training and thus, strengthen the trained 2D-to-3D skeleton mapping DNN. For example, the augmentation

method of [8] is based on a genetic model that applies mutations to specific joints of the input 2D skeletons during training. Similarly, [9] utilized GAN based configuration with a Generator and a Discriminator network, where the first one generates 2D skeletons to augment the training set while the latter investigates the validity of the generated 2D skeletons. In order to generate realistic 2D skeletons, the Generator modifies three different aspects of input 2D skeletons: a) bone angle, which changes the angles between the bones themselves (e.g. upper with lower arm), b) bone length, which changes the size of the skeleton in a way that does not affect its validity or symmetry, and c) its global location by applying rigid transformations (rotation and translation).

In a different approach, the overfitting problem can be tackled by training the 2D-to-3D skeleton mapping DNN in a weakly-supervised manner, where using the ground-truth 3D human pose labels for directly training the 2D-to-3D skeleton mapping DNN is avoided. For example, a 3D body joints estimator from 2D skeleton data was trained in [11] with a weakly supervised adversarial learning approach, where the Discriminator network of a GAN was used to learn a distribution of 3D skeletons. Instead of forcing the 2D-to-3D skeleton mapping DNN to predict a specific 3D skeleton for each training input data point, it is tasked to map a distribution of 2D skeletons to a distribution of 3D skeletons, which are valid according to the Discriminator network. Besides the 2D-3D skeleton mapping DNN and the Discriminator, a third neural network is also used to infer the parameters of a weak perspective camera that are used to reproject the estimated 3D skeletons back to 2D, in order to obtain matching 2D and 3D skeletons. In a simlar manner, [15] also utilized a 2D-to-3D skeleton mapping DNN and GANs to generate multiple 3D human pose candidates that correspond to an input 2D skeleton. Then, the best candidate is selected in a post-processing step and is returned as the final estimation.

While these weakly-supervised 3D human pose estimation approaches managed to increase the 3D human pose estimation accuracy on unseen test data, they only predict plausible 3D human poses (3D human poses with accurate bone lengths and angles) that correspond to the input 2D skeletons. That is, they are not able to predict the real, camera-relative 3D human poses. In contrast, the proposed 3D human pose estimation framework is specifically designed to predict not only plausible 3D human poses but also the real camera-relative 3D human poses.

III. Proposed 3D human pose estimation framework

In this work, a weakly-supervised 3D human pose estimation framework is proposed, which is able to estimate camera-relative 3D skeletons from single RGB images. In the first stage, it utilizes a 2D human pose estimation CNN to extract 2D skeletons from the input images. The second, 2D-to-3D human pose lifting stage consists of the introduced geometrical-alignment pre-processing step, which is used only during the training phase to pre-process the ground-truth 3D human pose data that correspond to the extracted 2D skeletons, the 2D-to-3D skeleton mapping DNN that is tasked to predict a plausible 3D human pose for each extracted 2D skeleton and a 3D skeleton-refinement post-processing step, which utilizes the 3D human pose obtained by the DNN to compute the final camera-relative 3D skeleton.

Let $\mathbf{S}_{3D} \in \mathbb{R}^{3\times K}$ and $\mathbf{S}_{2D} \in \mathbb{R}^{2\times K}$ denote a 3D human pose and a 2D human pose, respectively, where K is the number of the selected body joints that comprise the human skeleton. Also, suppose that a 2D-to-3D human pose dataset can be constructed, where each dataset sample $\mathbf{s}_i = \{\mathbf{S}_{3D_i}, \mathbf{S}_{2D_i}, \mathbf{R}_i, \mathbf{t}_i\}$ consists of its annotated camera-relative 3D skeleton $\mathbf{S}_{3D_i}$, the corresponding 2D skeleton $\mathbf{S}_{2D_i}$ in pixel coordinates and the rotation matrix $\mathbf{R}_i \in \mathbb{R}^{3\times 3}$ and translation vector $\mathbf{t}_i \in \mathbb{R}^3$ that map $\mathbf{S}_{3D_i}$ to $\mathbf{S}_{3D_{w_i}}$ using:

$$\mathbf{S}_{3D_{w_i}} = \mathbf{R}_i^T(\mathbf{S}_{3D_i} - \mathbf{t}_i\mathbf{1}_K^T), \tag{1}$$

where $\mathbf{S}_{3D_{w_i}} \in \mathbb{R}^{3\times K}$ is the 3D human pose in a human body-centered coordinates system. Note that the 2D skeletons $\mathbf{S}_{2D_i}$ can be either provided from the dataset or be calculated offline using the employed 2D human pose estimation CNN. The proposed method utilizes the 2D-to-3D skeleton mapping DNN introduced in [11], which is trained in a weakly-supervised manner to predict $\mathbf{S}_{3D_{w_i}}$ using $\mathbf{S}_{2D_i}$ as input.

The geometrical-alignment pre-processing step is used only during the training phase of the proposed framework and first defines a template 3D skeleton $\mathbf{S}_{3D_t}$ by selecting a random camera-relative 3D skeleton $\mathbf{S}_{3D_j}$ from the dataset and centering it by using $\mathbf{R}_j$ and $\mathbf{t}_j$:

$$\mathbf{S}_{3D_t} = \mathbf{R}_j^T(\mathbf{S}_{3D_j} - \mathbf{t}_j\mathbf{1}_K^T). \tag{2}$$

Then, assuming that the camera intrinsic parameters $\mathbf{K}$ are known, the 3D skeleton of each training sample $\mathbf{s}_i$ is aligned to the template 3D skeleton with the following two steps: a) a Perspective-n-Point (PnP) algorithm is applied between $\mathbf{S}_{3D_t}$ and $\mathbf{S}_{2D_i}$ to obtain $\mathbf{R}'_i$ and $\mathbf{t}'_i$:

$$[\mathbf{R}'_i, \mathbf{t}'_i] = PnP(\mathbf{S}_{3D_t}, \mathbf{S}_{2D_i}, \mathbf{K}), \tag{3}$$

and b) $\mathbf{R}'_i$ and $\mathbf{t}'_i$ are used to align $\mathbf{S}_{3D_i}$ to $\mathbf{S}_{3D_t}$:

$$\mathbf{S}'_{3D_i} = \mathbf{R}_i'^T(\mathbf{S}_{3D_i} - \mathbf{t}'_i\mathbf{1}_K^T). \tag{4}$$

Note that only specific body joints (hips, shoulders, spine, neck) are used in the PnP algorithm. Algorithm 1 shows the geometrical-alignment pre-processing procedure. The aligned 3D skeletons $\mathbf{S}'_{3D_i}$ are the ones used to train the 2D-3D skeleton mapping DNN in the GAN-assisted weakly-supervised framework adopted from [11]. The key feature of the geometrical-alignment pre-processing step is that it applies rigid transformations that are fully reversible, allowing the calculation of the camera-relative 3D body joints coordinates from the "aligned" network outputs in a post-processing step.

The 3D skeleton-refinement post-processing step receives the aligned 3D skeleton $\hat{\mathbf{S}}'_{3D_i}$ that is predicted by the 2D-3D skeleton mapping DNN for input $\mathbf{S}_{2D_i}$ and calculates the final camera-relative 3D skeleton $\hat{\mathbf{S}}_{3D_i}$. This is accomplished by utilizing $\hat{\mathbf{S}}'_{3D_i}$ and $\mathbf{S}_{2D_i}$ in a PnP algorithm to obtain $\hat{\mathbf{R}}_i$ and $\hat{\mathbf{t}}_i$:

$$[\hat{\mathbf{R}}_i, \hat{\mathbf{t}}_i] = PnP(\hat{\mathbf{S}}'_{3D_i}, \mathbf{S}_{2D_i}, \mathbf{K}), \tag{5}$$

and using them to calculate the camera-relative 3D body joints coordinates according to:

$$\hat{\mathbf{S}}_{3D_i} = \hat{\mathbf{R}}_i\hat{\mathbf{S}}'_{3D_i} + \hat{\mathbf{t}}_i\mathbf{1}_K^T. \tag{6}$$

Algorithm 1 Geometrical-alignment pre-processing step

procedure GEOMETRICAL-ALIGNMENT($\mathbf{s}_i, \mathbf{s}_j$)
  $\mathbf{S}_{2D_i} \leftarrow get\_2d\_pose(\mathbf{s}_i)$
  $\mathbf{S}_{3D_i} \leftarrow get\_3d\_pose(\mathbf{s}_i)$
  $\mathbf{S}_{3D_j} \leftarrow get\_3d\_pose(\mathbf{s}_j)$
  $\mathbf{R}_j, \mathbf{t}_j \leftarrow get\_extrinsics(\mathbf{s}_j)$
  $\mathbf{K}_j \leftarrow get\_intrinsics(\mathbf{s}_j)$
  $\mathbf{S}_{3D_t} \leftarrow \mathbf{R}_j^T(\mathbf{S}_{3D_j} - \mathbf{t}_j\mathbf{1}_K^T)$
  $\mathbf{P}_{\mathbf{S}_{3D_t}} \leftarrow \mathbf{S}_{3D_t}[selected\_joints]$
  $\mathbf{P}_{\mathbf{S}_{2D_i}} \leftarrow \mathbf{S}_{2D_i}[selected\_joints]$
  $\mathbf{R}'_i, \mathbf{t}'_i \leftarrow PnP(\mathbf{P}_{\mathbf{S}_{3D_t}}, \mathbf{P}_{\mathbf{S}_{2D_i}}, \mathbf{K})$
  $\mathbf{S}'_{3D_i} \leftarrow \mathbf{R}'^T_i(\mathbf{S}_{3D_i} - \mathbf{t}'_i\mathbf{1}_K^T)$
  return $\mathbf{S}'_{3D_i}$
end procedure

Algorithm 2 3D skeleton-refinement post-processing step

procedure 3D SKELETON-REFINEMENT($\hat{\mathbf{S}}'_{3D_i}, \mathbf{S}_{2D_i}$)
  $\mathbf{K}_i \leftarrow get\_intrinsics(\hat{\mathbf{S}}'_{3D_i})$
  $\mathbf{P}_{\hat{\mathbf{S}}'_{3D_i}} \leftarrow \hat{\mathbf{S}}'_{3D_i}[selected\_joints]$
  $\mathbf{P}_{\mathbf{S}_{2D_i}} \leftarrow \mathbf{S}_{2D_i}[selected\_joints]$
  $\hat{\mathbf{R}}_i, \hat{\mathbf{t}}_i \leftarrow PnP(\mathbf{P}_{\hat{\mathbf{S}}'_{3D_i}}, \mathbf{P}_{\mathbf{S}_{2D_i}}, \mathbf{K}_i)$
  $\hat{\mathbf{S}}_{3D_i} \leftarrow \hat{\mathbf{R}}_i\hat{\mathbf{S}}'_{3D_i} + \hat{\mathbf{t}}_i\mathbf{1}_K^T.$
  return $\hat{\mathbf{S}}_{3D_i}$
end procedure

The 3D skeleton-refinement post-processing procedure is also described in Algorithm 2.

IV. Experimental Evaluation

In the proposed framework, the weakly supervised 3D human pose estimation method of [11] was adopted and incorporated into a single pipeline along with the geometrical-alignment pre-processing step and the 3D skeleton-refinement post-processing step. In addition, the 2D human pose estimation method of [16] was utilized to extract 2D skeletons from the input RGB images. The 2D-to-3D skeleton mapping DNN was trained for 20 epochs using the Adam [17] optimizer with initial learning rate of 0.0001 and batch size 160. Finally, the PnP algorithm proposed in [18] was used in all cases. All experiments were conducted using an *Ubuntu* machine, equipped with a *NVidia GTX 1080 Ti* graphics card.

The proposed method was trained and evaluated using the Human3.6M [19] dataset, where subjects are captured to perform everyday actions in an indoor environment. The typical train/test split was followed, where S1, S5, S6, S7, and S8 are used for training and subjects S9, S11 for testing. The camera intrinsic parameters $\mathbf{K}$ for all subjects are also provided with the dataset.

The evaluation metric used is the typically used Mean Per Joint Position Error (MPJPE) [11], which is the mean value of the euclidean distance in mm between the predicted and actual 3D poses. Following previous work [11], the MPJPE error is calculated for two evaluation protocols. Protocol-I calculates the MPJPE metric between the outputs of the 2D-to-3D mapping DNN and the ground-truth 3D human poses, while Protocol-II, first applies a rigid transformation on the network outputs to align them with the ground-truth 3D human poses using the Procrustes alignment method and calculates the MPJPE metric between the aligned predicted and the ground-truth 3D human poses.

TABLE I. Experimental results (MPJPE in mm) of all variations of the baseline RepNet and the proposed framework models using the ground-truth 2D skeletons provided with Human3.6M [19] both for human body-relative and camera-relative 3D human pose estimation and under both protocols.

Trained on GT 2D data	Human body-relative		Camera-relative	
	P-I	P-II	P-I	P-II
RepNet-GT [11]	211.35	40.40	172.58	40.40
RepNet-Det [11]	223.97	48.66	206.12	48.66
ours-GT	206.80	**40.04**	**166.40**	**40.04**
ours-Det	227.38	53.78	400.05	53.78
ours-Full	**202.13**	42.02	215.43	42.02

TABLE II. Experimental results (MPJPE in mm) of all variations of the baseline RepNet and the proposed framework models using 2D skeletons extracted by [16] both for human body-relative and camera-relative 3D human pose estimation and under both protocols.

Trained on 2D data [16]	Human body-relative		Camera-relative	
	P-I	P-II	P-I	P-II
RepNet-GT [11]	228.91	65.20	292.44	65.20
RepNet-Det [11]	230.17	54.43	**275.90**	54.43
ours-GT	**227.60**	66.76	318.15	66.76
ours-Det	234.32	59.11	336.30	59.11
ours-Full	229.19	**53.32**	319.86	**53.32**

The MPJPE metric for Protocol-I and Protocol-II is calculated both for the human body-centered coordinates system 3D human poses $\hat{\mathbf{S}}'_{3D_i}$ obtained directly by the 2D-to-3D mapping DNN and the camera-relative 3D poses $\hat{\mathbf{S}}_{3D_i}$ obtained by the 3D skeleton-refinement post-processing step to evaluate the proposed method in both cases. Moreover, three different sources for the 2D skeletons are examined to evaluate the proposed method resistance on noisy inputs, the ground-truth 2D skeletons provided by the Human3.6M dataset, the 2D skeletons provided by [20] and the 2D skeletons extracted offline using [16].

Different variations of the baseline RepNet [11] and the proposed 3D human pose estimation framework were evaluated. RepNet-GT denotes the baseline method trained on ground-truth data provided by Human3.6M dataset, while RepNet-Det denotes the baseline model trained on the 2D detections obtained by [16]. In a similar manner, ours-GT and ours-Det denote the proposed framework trained on the ground-truth and the extracted 2D skeletons, respectively. Finally, ours-Full denotes the proposed method trained using both ground-truth and extracted 2D skeletons as input. Note that in all cases, the 3D skeletons provided by the Human3.6M are used for GAN training, as described in [11] and the camera-relative 3D poses are obtained using the proposed 3D skeleton-refinement post-processing step.

Table I shows the comparison results when the ground-truth 2D skeletons are used as input to both the baseline and

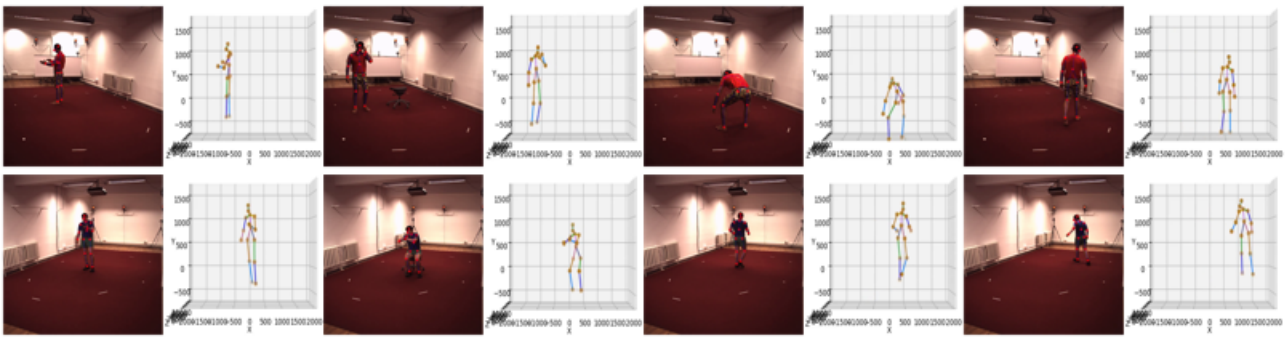

Fig. 1. Qualitative evaluation of the proposed 3D human pose estimation framework using random images from the Human3.6M [19] test set. First and second rows show the extracted 2D skeletons obtained by [16], as well as the corresponding camera-relative 3D human poses estimated from the proposed framework for test subjects S9 and S11, respectively.

TABLE III. EXPERIMENTAL RESULTS (MPJPE IN MM) OF ALL VARIATIONS OF THE BASELINE REPNET AND THE PROPOSED FRAMEWORK MODELS USING 2D SKELETONS EXTRACTED BY [20] BOTH FOR HUMAN BODY-RELATIVE AND CAMERA-RELATIVE 3D HUMAN POSE ESTIMATION AND UNDER BOTH PROTOCOLS.

Trained on	Human body-relative		Camera-relative	
2D data [20]	P-I	P-II	P-I	P-II
RepNet-GT [11]	260.04	118.41	1002.07	118.41
RepNet-Det [11]	238.74	85.82	**528.85**	85.82
ours-GT	255.63	119.86	1072.86	119.86
ours-Det	**233.66**	**83.54**	604.97	**83.54**
ours-Full	245.13	109.24	927.12	109.24

the proposed framework. "P-I" and "P-II" are used to denote that the MPJPE metric was calculated under Protocol-I and Protocol-II, respectively. The results show that the proposed method (ours-GT) outperforms the baseline RepNet-GT in both protocols and both for human body-relative and camera-relative 3D human pose estimation.

The comparison results between the baseline and the proposed framework when the 2D skeletons extracted by [16] are used as input can be seen in Table II. While the proposed method is able to outperform the baseline method when the evaluation Protocol-II is used, its performance is slightly decreased for Protocol-I. Nevertheless, the proposed 3D skeleton-refinement post-processing step allows both the baseline and the proposed method to accurately predict camera-relative 3D human poses.

For completeness, all variants of the proposed method and the baseline RepNet were evaluated using the 2D skeletons provided by [20] as input. This is to show that the proposed 3D human pose estimation framework can be used to predict camera-relative 3D human poses even from unseen 2D skeletons. The results reported in Table III show that the proposed framework managed to predict accurate 3D human poses, outperforming the RepNet baseline in most cases.

Finally, apart from the quantitative evaluation presented in Tables I-III, a qualitative evaluation of the proposed method was also conducted using random images from the Human3.6M test subjects S9 and S11. Test images along with the 2D skeletons extracted by [16], as well as the corresponding camera-relative 3D human poses obtained by the proposed framework are depicted in Fig. 1. It can be seen that in all cases the proposed framework is able to estimate accurate camera-relative 3D human poses, despite the fact that in some cases the extracted 2D skeletons contain noisy body joint detections (mostly due to body joint occlusions). This is achieved through the utilized GAN-assisted weakly-supervised training setting and the indirect supervision provided by the Discriminator network, which is trained efficiently using the aligned 3D skeletons $\mathbf{S}'_{3D_i}$ calculated by the proposed geometrical-alignment pre-processing step.

V. CONCLUSIONS

This paper presented a 3D human pose estimation framework that is able to predict camera-relative 3D human poses from 2D skeleton data. It incorporates a geometrical-alignment pre-processing step to prepare the training data, which are subsequently used to train a 2D-to-3D skeleton mapping DNN that is able to predict plausible 3D human poses that correspond to the input 2D skeletons. Finally, a 3D skeleton-refinement post-processing step is applied on the outputs of the 2D-to-3D skeleton mapping DNN to predict the final, camera-relative 3D human poses. Quantitative and qualitative evaluation showed that the proposed framework managed to predict accurate camera-relative 3D human poses, while it outperformed the baseline RepNet in terms of 3D human pose estimation accuracy. Finally, the proposed framework was able to predict accurate camera-relative 3D human poses even when noisy 2D skeletons were used as input.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871479 (AERIAL-CORE). This publication reflects the authors views only. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] F. Patrona, I. Mademlis, and I. Pitas, "An overview of hand gesture languages for autonomous UAV handling," in *Proceedings of the Aerial Robotic Systems Physically Interacting with the Environment (AIRPHARO)*, 2021.

[2] C. Papaioannidis, D. Makrygiannis, I. Mademlis, and I. Pitas, "Learning fast and robust gesture recognition," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2021.

[3] D. Makrygiannis, C. Papaioannidis, I. Mademlis, and I. Pitas, "Optimal video handling in on-line hand gesture recognition using deep neural networks," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021.

[4] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng, "Marker-less 3D human motion capture with monocular image sequence and height-maps," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[5] C. Luo, X. Chu, and A. Yuille, "Orinet: A fully convolutional network for 3D human pose estimation," *arXiv preprint arXiv:1811.04989*, 2018.

[6] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, 2017.

[7] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, "Cascaded deep monocular 3D human pose estimation with evolutionary training data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] K. Gong, J. Zhang, and J. Feng, "Poseaug: A differentiable pose augmentation framework for 3D human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] B. Wandt and B. Rosenhahn, "Repnet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[13] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3D human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[14] Z. Yuan and S. Du, "Jointpose: Jointly optimizing evolutionary data augmentation and prediction neural network for 3D human pose estimation," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 2021.

[15] L. Chen and H. L. Gim, "Weakly supervised generative network for multiple 3D human pose hypotheses," 2020.

[16] C. Papaioannidis, I. Mademlis, and I. Pitas, "Fast single-person 2D human pose estimation using multi-task Convolutional Neural Networks," *submitted*, 2022.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the PnP problem," *International Journal of Computer Vision*, vol. 81, no. 2, p. 155, 2009.

[19] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[20] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.