Multi-Person 3D Pose Estimation in Mobile Edge Computing Devices for Real-Time Applications

Md. Imtiaz Hossain, Sharmen Akhter, Md. Delowar Hossain, Choong Seon Hong and Eui-Nam Huh
Department of Computer Science and Engineering, Kyung Hee University
Yongin-si, South Korea
Email: {hossain.imtiaz, sharmen, delowar, cshong, johnhuh}@khu.ac.kr

2023 International Conference on Information Networking (ICOIN) | 978-1-6654-6268-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICOIN56518.2023.10049033

***Abstract*—In the last few years, real-time 3D pose estimation from RGB monocular images on Mobile Edge Computing devices has drawn immense attraction due to the ability to estimate, infer and transfer 3D motion and pose in VR, AR, gaming, animation and so on. However, estimating motion under occlusion is very challenging. Though a number of effective and efficient approaches have been proposed to deal with this issue, there is still a demand for robust occlusion-aware multi-person 3D pose and motion estimation under occlusion in real-world scenarios. In this paper, we propose a one-shot occlusion-aware real-time 3D pose estimation and inference approach called RRMP. Our proposed RRMP performs both 2D and 3D pose estimation and is composed of three sequential stages: 1) the residual to render a multi-level perspective for each individual people, 2) the initial stage, and 3) the refinement stages. As our goal is to estimate the pose in real-time for mobile edge computing devices, the RRMP is designed using Depthwise Separable Convolutions (DSCs) that perform with an average of 40 fps in real-time execution. Our extensive results and analysis depict that the proposed RRMP improves the performances of the existing state-of-the-art methods. Our RRMP technique can be deployed into any existing state-of-the-art works for further improving the robustness in terms of occlusion.**

***Index Terms*—RRMPs, Lightweight Architecture, Depthwise Separable Convolutions, Pose Inference, 3D Pose Estimations, Mobile Edge Computing, Residual Connection**

I. Introduction

For decades deep learning-based approaches have gained significant interest due to having remarkable performances on diverse tasks such as classification [16], recognition [9], segmentation [15] and so on, in both low and high-computational resource-based devices [11]. Real-time 3D multi-person pose estimation from a single monocular RGB image in mobile edge computing devices is a rigorous task. Though various methods have been proposed to provide satisfactory performances for this task, they struggle to estimate the pose and motion of occluded objects or humans while maintaining expected trade-offs between speed-vs-accuracy. In real-world multi-person scenarios, the human body is occluded by self or other humans which makes the existing works struggle in estimating pose with low latency.

For estimating multi-person 3D human pose in real-world scenarios most existing state-of-the-art works first decompose every individual human and process them individually as multiple single-person 3D pose estimation tasks [1]–[3]. During estimating the single-person pose estimation, the prediction for each person is combined at the post-processing stages. Dushyant et. al. [2] proposed a single-shot real-time 3D pose estimation technique that can successfully estimate 3D pose. Wang et. al. [5] and Chen et. al., [4] proposed a real-time pose estimation process in the 2D approach. [6] proposed a lightweight open-pose estimation technique for real-time applications in 2D space for CPU devices. Xnect [1] shows a top-down approach to successfully estimate human motion and transfer the motion to a 3D model.

In this work, we improve the outcome of 3D [2] and 2D [6] by introducing residual connection among multi-persons 2D spatial backbone features and multi-level perspective estimation (MPE) features to enhance the robustness in estimating the pose of the occluded humans. Similar to [2], firstly we extract the backbone features from RGB monocular images using the depthwise separable convolutions. The extracted 2D spatial features are then fed into the MPE block that is adopted from [2]. The combined features are then sent to the initial stage to get the keypoint heat maps and part affinity fields. The obtained keypoint heat maps and part affinity fields are then refined at the refinement stages. Finally, these refined features are employed to perform 3D pose inference and estimations. The 3D pose inference and estimation blocks estimate multi-person intense poses for occluded and unoccluded persons. The details of our proposed RRMP that stands for **R**esidual to **R**ender **M**ulti-Level **P**erspectives are described in section III. The whole network is trained on the MuCo-3DHP [2] dataset. The core contributions of this paper can be summarized as follows:

1) We combine the local and global multi-level perceptiveness to obtain multi-level features that increase the robustness in terms of occlusion.
2) Our proposed RRMP in between the backbone and MPE features enhances the quality of the features that enhance the performances in pose inference and estimation tasks.
3) We provide details experimental analysis about the effect of the proposed RRMP on real-time multi-person 3D human pose estimation tasks.

In section II we discuss the related works. The prospered method is discussed in section III. Results and discussions are described in section V. Finally, the conclusion is depicted in section VI.

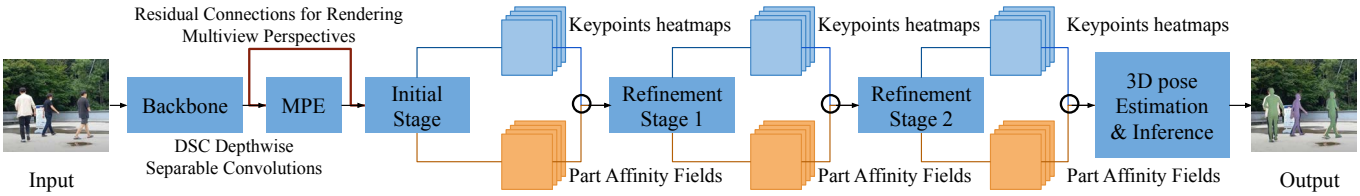


Fig. 1. Overview of our proposed technique RRPM

TABLE I
COMPARATIVE EVALUATION BETWEEN LCR-NET [22] AND OUR METHOD ON MUPOTS-3D DATASET.THE REPORTED RESULTS DENOTE OVERALL ACCURACY I.E., 3DPCK.

Model	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11
LCR-Net	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2
Ours	**80.6**	**58.1**	**63.7**	**60.3**	**67.7**	**27.8**	**47.6**	**53.2**	**37.6**	**81.9**	**55.3**
Model	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20		Total
LCR-Net	51.0	51.6	49.3	56.2	66.5	65.2	62.9	**66.1**	59.1		53.8
Ours	**53.3**	**54.6**	**53.8**	**61.3**	**69.1**	**67.6**	**63.4**	65.9	**61.5**		**59.2**

TABLE II
COMPARISON ON TEST SET OF MPI-INF-3DHP [2] DATASET. PCK AND AUC DENOTE THE PERCENTAGE OF CORRECT KEYPOINTS AND AREA UNDER THE CURVE FOR ALL ACTIVITIES, RESPECTIVELY.

Method	Sit	Crouch	Total	
	PCK	PCK	PCK	AUC
Mehta et al. [10]	74.8	73.7	75.7	39.3
LCR-Net [22]	58.5	69.4	59.7	27.6
VNect [3]	74.7	72.9	**76.6**	**40.4**
Ours	**75.3**	**73.8**	74.1	38.3

II. LITERATURE REVIEW

A. Single Person 3D Pose Estimation

Though prior single-person 3D pose estimation-based methods perform better on monocular RGB datasets, they struggle to generalize better due to training in the real-world scenarios for discriminative predictions [7], various poses, backgrounds, diverse appearances, and occlusions [2]. Existing single-person 3D pose estimation works can be divided into two steps: 1) estimating joints in 2D space, and 2) fitting them into 3D space [2], [4], [6], [7]. To utilize both the 2D and 3D annotations numerous works incorporated the SMPL body models with convolutional neural networks [8], [10], [12], [13]. However, instead of designing 2D and 3D pose individually, some works reasoned them jointly by using multi-stage belief maps [14]. The main challenge of these methods is estimating 2D joints via bounding boxes that fail to infer and estimate occluded objects perfectly. In our work, we focus on improving occlusion robustness. Dushyant et. al. [2] proposed an occlusion-robust method using the ORPMs block. We enhance that work by performing residual to render multi-level perspectives (RRMPs) that leverage the residual of the basic backbone features before feeding into the refinement stages.

B. Multi-Person 3D Pose Estimation

Simo-Serra, Edgar, et al. [17] and [2] provide good performing methods for multi-person 3D pose estimation from a single RGB image. Firstly, they perform person localization to obtain the bounding boxes for each individual by adopting the method in [18]. Then the bounding boxes are mapped to similar K-poses [19]. From K-poses the specified pose is identified by the score predicted by a classifier. A regression block then refines the pose. In principle, these methods localize the persons and perform pose estimation as well as inference for each person. The main challenge is that their methods fail to estimate and infer an occluded person's pose perfectly. Daniil Osokin et. al. [6] and [2] proposed a single-shot approach to localize 2D joints of multi-person and estimate 3D pose that facilitate pose estimation of occluded person. However, our work is different from these works in the perspective that, we perform residual to render multi-level perspective using RRMPs.

III. METHOD

The main goal of our work is to estimate and infer the multi-person 3D pose of the occluded people. We extend the performances of 2D pose estimation and single shot 3D pose inference via formulating location maps [20]. The location maps infer the 3D joints considering 2D pixel locations that facilitate robust 3D pose estimation. In this section we will discuss the 3D pose estimation of the occluded persons, **R**esidual for **R**endering **M**ulti-view **P**erspectives in backbone convolutional neural networks that use depthwise separable convolutions (DSCs) to provide a lightweight architecture, and the process of pose inference. The relations among the joints are considered as the directed graph (see Fig. 2).

A. Preliminaries

Let's consider $\mathcal{X}$ denotes a monocular RGB image of dimension $(W \times H \times 3)$ containing N number of persons. Firstly we estimate the pose $\mathcal{P} = \{\phi_i\}_{i=1}^{m}$ for each people where $\phi_i \in \mathrm{R}^{3\times n}$ denotes coordinates in 3-dimensional space for each body joints of ith person. We first localize and decompose the body into head, neck, pelvis, hand and so on. The edge and relation sets can be briefed as $\mathcal{L} =$

$\{\{wrist\_s, elbow\_s, shoulder\_s\}, \{ankle\_s, knee\_s, hip\_s\}\| \mathcal{S} \in \{left, right\}\}$. The joints locations are then encoded into latent space towards occlusion-robust pose maps indicated by $\mathcal{R} = \{\mathbf{R}_j\}_{j=1}^{n}$, here, $\mathbf{R}_j \in \mathrm{R}^{W \times H \times 3}$. The occlusion-robust maps store joints j as a set of 2D locations, such as:

$$l(j) = \{(u,v)_{neck}, (u,v)_{pelvis}\} \cup \{(u,v)_k\}_{k \in limb(j)}, \quad (1)$$

where,

$$limb(j) = \begin{cases} l, & \text{if } \exists l \in \mathrm{L} \text{ with j} \in l \\ head, & \text{if j = head} \\ \emptyset, \text{otherwise} & \end{cases} \quad (2)$$

here, in some cases, $l_{i1}(j) \cap l_{i2}(j) \neq \emptyset$.

Similar to the technique proposed in [2] we estimate part affinity fields (PAFs) $\mathcal{A} = \mathbf{A}_j \in \mathrm{R}^{W \times H \times 2^n}_{j=1}$ that indicates a 2D vector field [6], we utilize 2D joint heat-maps $\mathcal{A} = \{\mathbf{A}_j \in \mathrm{R}^{W \times H}\}_{j=1}^{n}$. For occlusion-aware Pose Estimation (OPE) and pose inference following the work in [2].

B. Residual for Rendering multi-level Perspectives (RRMPs)

We improve forward propagation and feature extraction to alleviate more local information through the networks. In pose estimation, the occlusion local mask plays important role in making accurate predictions [21]. As MPE in [2] includes multiple convolutional networks, it losses some fine-grained information. Being inspired by the ResNet [16], we perform residual connections that alleviate the information of the backbone to the initial stage. Our approach helps by propagating more information for the refinement task. For RRMPs, our approach learns identity function that improves generalization. Let's consider the output feature maps from the backbone network to be $\mathcal{F}_b$, RRMPs network as $\phi(\cdot)$ and output feature maps from RRMPs to be $\mathcal{R}$, as:

$$\mathcal{R} = \phi(\mathcal{F}_b) \quad (3)$$

Instead of feeding $\mathcal{R}$ as input to the initial stage, we deploy RRMPs for rendering multi-level perspectives. Our proposed approach contains both local and global features that actually denote multiple perspectives considering the level of features.

$$\mathcal{R}' = \mathcal{F}_b + \mathcal{R} \quad (4)$$

C. Pose Inference from RRPMs

Pose inference starts with detecting multiple people from the features of RRPMs by localizing 2D joint location $\mathcal{J}^{2D} = \{J_i^{2D}\}_{i=1}^{m}$ where, $J_i^{2D} = \{(x,y)_j^i\}_{j=1}^{n}$. The joint detection confidences $\mathcal{C}^2D = \{C_i^{2D} \in \mathrm{R}^n\}_{i=1}^{m}$ for each people i in the image. The part affinity fields and keypoint heat-maps associate explicit 2D joints to estimate the pose of the whole person similar to Cao et al. [23]. Finally, the 2D joints keypoints $\mathcal{J}^{2D}$ and the keypoints detection confidence $\mathcal{C}^2D$ along with RRMPs map $\mathcal{R}$ is exploited to infer and predict multi-person 3D pose in the image.

1) Multi-level Perspective: The read-out of 3D joint locations using occlusion robust pose maps that are obtained from the feature maps of ORPMs block, takes feature maps as input from the backbone that is a stack of sequential depthwise separable convolutions. ORPMs leverage comparatively global features to the refinement stages and finally 3D block predict the 3D pose estimation. In our work, we alleviate both local and global features via residual to render multi-level perspective blocks (RRMPs). As the feature maps of RRMPs contain both the local and global context compared to the feature maps of ORPMs, the refinement and 3D pose block infer and predict better the pose of the occluded persons. Similarly, the fine-grained and global feature helps keypoint heat maps and affinity part fields to perform better in finding local joints. However, our multi-level perspective block also helps to improve the regularization and gradient flow in backward propagation by improving the situation of vanishing gradient [16].

2) 2D Joint Validation: We estimate the joints location of 2D, $J^{2Dj}_i = (x,y)^i_j$ of person i as the joint validation in two considerations: (1) if the person is occluded then the confidence score is less than a threshold t_J, and (2) if the joint location point is not nearest to the read-out location, then the 2D joint location j^{2Di}_j can be validated as:

$$valid(J^{2Dj}_i) \Leftrightarrow C^{2Dj}_i > t_J \wedge ||s - J^{2Dj}_i||_2 \geqq t_D \quad \forall \hat{i} = [1:n], \hat{i} \neq i. \in \rho_{\hat{i}(j)} \quad (5)$$

The refinement stages and occlusion-aware pose inference technique with our RRMPs provide accurate poses even if the body parts are occluded.

IV. Experimental Setup

A. Training Details

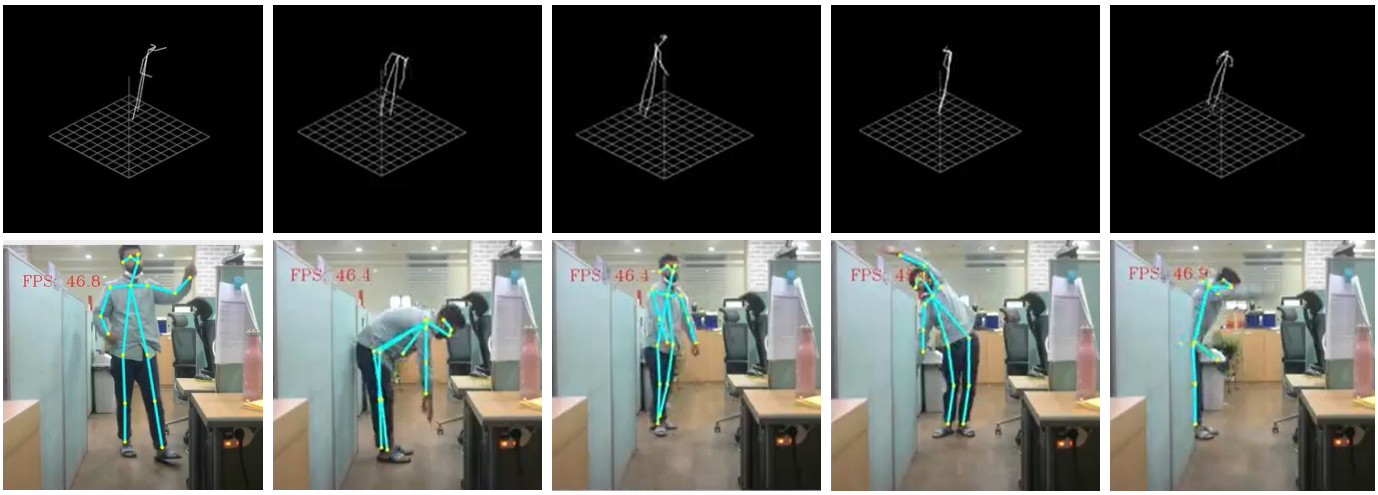
Following [2] we use ResNet56 as the backbone and employ two tails where one is for extracting the 2D features for computing affinity, addressed as 2D+affinity stream and another one for 3DPose stream. For extracting 2DPose and 3DPose we train our network using MS-COCO [24] and MPI-INF-3DHP [2] datasets, respectively. The branch that is trained on the MS-COCO dataset provides 2D heat-maps class $\mathcal{A}_{COCO}$ of the body joints in 2D space and part affinity fields $\mathcal{F}_{COCO}$. The second stream provides 3D RRMPs $\mathcal{R}_{COCO}$ and 2D heatmaps $\mathcal{A}_{MPI}$.

B. Loss Functions

$\mathcal{A}_{COCO}$, part affinity fields $\mathcal{F}_{COCO}$ and $\mathcal{A}_{MPI}$ are trained using L2 loss considering unit peak Gaussian's with the 2D joint regions, and using the framework offered by [23] as the ground truth, respectively. For training RRMPs, we adopt the ORPMs loss with the MuCo-3DHP dataset that is described in [2].

V. Results and Discussions

Our main goal is to improve the performance of the multi-person 3D pose estimation works by performing residual to render multi-level perspective (RRMPs). The main challenge



(a) Normal (b) Complex Pose (c) Self-Occluded (d) Partially-Occluded (e) Major-Occlusion

Fig. 2. Occlusion Robustness: Visualization of the qualitative results of our proposed method. The first row indicates the 3D space of the estimated poses. The second row shows the corresponding similarity with the real-time pose of humans. Each different column depicts the results in various situations. From this figure, we can see that the proposed method can estimate and infer the pose perfectly even if the person is self-occluded or occluded by another object. The output of the second rows shows the average framerates that are (40fps)

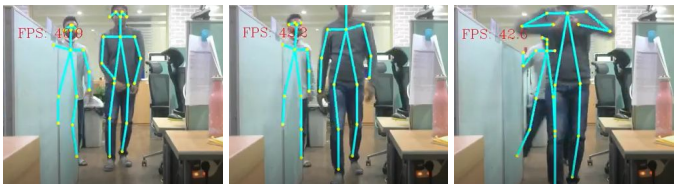


Fig. 3. Qualitative visualizations for multi-person pose estimation. The output of the second rows shows the average framerates that are (40fps)

in estimating pose is occlusion. Propagating both local and global perspectives towards refinement stages helps the regression block estimate and infer the occluded pose better. To validate our approach we perform extensive experiments on both single and multi-person pose estimation along with the robustness on occlusion.

A. Single Person Pose Estimation

Though our goal is to estimate the pose of multiple people we also evaluate the performance of our proposed RRMPs-based approach compared to the state-of-the-art methods on the MPI-INF-3DHP dataset for a single person. We also compare the results of three single-person pose estimations on MuCo-3DHP following [2]. From Table II, we notice that our method performs poorly compared to the VNect in 3D PCK and AUC, this is because of training on the hard dataset. However, our method outperforms LCRNet [22] in these scenarios and beats all the methods in sit (PCK) and Crouch (PCK).

B. Multi-Person Pose Estimation

In this section, we discuss the comparative results of our proposed method on the MuPoTS-3D [2] dataset. The competing results are shown in Table I. From the table, we can observe that for all the 20 sequence test sets which have available 3D annotated pose of MuPoTS-3D dataset, our proposed method beats LCR-Net [22] except at test set 19. In total, our method achieves the best results compared to the LCR-Net. The proposed method achieves 5.4% better 3DPCK on multi-person pose estimation than LCR-Net.

C. Occlusion Robustness and Real-time 3D Pose Estimation

We provide qualitative results of the occlusion robustness of our method in Figure 2 and 3. From the figure, we can see that our method performs better in inference and estimating the poses of the occlusion joints. To demonstrate the occlusion robustness we train our network on the MuToPS-3DHP dataset.

VI. Conclusion

In this work, we extend the work of 3D multi-person pose estimation performance by introducing residual to render the multi-level perspectives (RRPMs) to alleviate both local and global features to the refinement stages using depthwise separable convolutions (DSCs) based backbone networks. Our method is able to perform in real-time applications on mobile edge computing devices with high frame rates. As we alleviate both local and global features, the proposed technique performs better in single and multi-person pose estimation. Also, our method performs better than state-of-the-art methods under occluded joints. In future work, we extend our work in motion transfer from human to 3D model generation tasks for mobile edge computing devices in real-time applications.

Acknowledgment

This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT)

(No. 2202-0-00047, Development of Microservices Development/Operation Platform Technology that Supports Application Service Operation Intelligence) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A4A1018607).
** Corresponding author: Prof. Eui-Nam Huh.

REFERENCES

[1] Mehta, Dushyant, et al. "XNect: Real-time multi-person 3D motion capture with a single RGB camera." Acm Transactions On Graphics (TOG) 39.4 (2020): 82-1.
[2] Mehta, Dushyant, et al. "Single-shot multi-person 3d pose estimation from monocular rgb." 2018 International Conference on 3D Vision (3DV). IEEE, 2018.
[3] Mehta, Dushyant, et al. "Vnect: Real-time 3d human pose estimation with a single rgb camera." Acm transactions on graphics (tog) 36.4 (2017): 1-14.
[4] Chen, Ching-Hang, and Deva Ramanan. "3d human pose estimation= 2d pose estimation+ matching." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[5] Wang, Yangang, Baowen Zhang, and Cong Peng. "Srhandnet: Real-time 2d hand pose estimation with simultaneous region localization." IEEE transactions on image processing 29 (2019): 2977-2986.
[6] Osokin, Daniil. "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose." arXiv preprint arXiv:1811.12004 (2018).
[7] Bo, Liefeng, and Cristian Sminchisescu. "Twin gaussian processes for structured prediction." International Journal of Computer Vision 87.1 (2010): 28-52.
[8] Li, Sijin, Weichen Zhang, and Antoni B. Chan. "Maximum-margin structured learning with deep networks for 3d human pose estimation." Proceedings of the IEEE international conference on computer vision. 2015.
[9] Hossain, Md Imtiaz, et al. "Batch entropy supervised convolutional neural networks for feature extraction and harmonizing for action recognition." IEEE Access 8 (2020): 206427-206444.
[10] Mehta, Dushyant, et al. "Monocular 3d human pose estimation in the wild using improved cnn supervision." 2017 international conference on 3D vision (3DV). IEEE, 2017.
[11] Huh, Eui-Nam, and Md Imtiaz Hossain. "Brainware computing: Concepts, scopes and challenges." Applied Sciences 11.11 (2021): 5303.
[12] Mehta, Dushyant, et al. "Monocular 3d human pose estimation in the wild using improved cnn supervision." 2017 international conference on 3D vision (3DV). IEEE, 2017.
[13] Pons-Moll, Gerard, David J. Fleet, and Bodo Rosenhahn. "Posebits for monocular human pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
[14] Yang, Wei, et al. "3d human pose estimation in the wild by adversarial learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[15] Mo, Yujian, et al. "Review the state-of-the-art technologies of semantic segmentation based on deep learning." Neurocomputing 493 (2022): 626-646.
[16] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[17] Simo-Serra, Edgar, et al. "A joint model for 2d and 3d pose estimation from a single image." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
[18] Sarafianos, Nikolaos, et al. "3d human pose estimation: A review of the literature and analysis of covariates." Computer Vision and Image Understanding 152 (2016): 1-20.
[19] Rogez, Grégory, and Cordelia Schmid. "Mocap-guided data augmentation for 3d pose estimation in the wild." Advances in neural information processing systems 29 (2016).
[20] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European conference on computer vision. Springer, Cham, 2016.
[21] Heo, Miran, et al. "Integrating Pose and Mask Predictions for Multi-Person in Videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[22] Rogez, Gregory, Philippe Weinzaepfel, and Cordelia Schmid. "Lcr-net: Localization-classification-regression for human pose." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[23] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[24] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.