Augmented Reality Enhanced: 3D Crowd Reconstruction from a Single Viewpoint

Xiaohan Sun
Trinity College Dublin
Dublin, Ireland
sunx4@tcd.ie

Xiaosong Yang
Bournemouth University
Bournemouth, UK
xyang@bournemouth.ac.uk

***Abstract*—Reconstructing human figures from a single viewpoint has long intrigued researchers, particularly for augmented reality (AR) applications. While significant progress has been made in single-human body reconstruction, densely populated scenes with substantial occlusions pose complex challenges. This paper introduces 3DCrowd+, an advanced two-stage methodology for 3D reconstruction of human meshes in crowded environments. Building on the 3DCrowdNet framework, our model refines HRNet 2D pose estimation and integrates Lite-HRNet with Shuffle Block and CoordAttention modules, achieving robust feature extraction and lightweight performance. 3DCrowd+ combines an attention mechanism with a model pruning algorithm, demonstrating high accuracy and efficiency on various datasets. This research bridges the gap between complex crowd scenes and detailed 3D reconstruction, offering a promising solution for precise crowd modeling in AR environments.**

***Index Terms*—3D Reconstruction, Augmented Reality (AR), Crowd Modeling, Pose Estimation, Computer Vision**

I. Introduction

Recent advancements in artificial intelligence-generated content (AIGC) have revolutionized digital imaging and reconstruction, particularly in reconstructing 3D character meshes from 2D photos, impacting industries like gaming and film. For example, Unreal Engine's MetaHuman Animator [4] allows users to create digital human depictions easily. This progress is vital for augmented reality (AR) and the metaverse, but applying it to crowd scenarios remains complex due to diverse poses, interactions, and frequent occlusions. The main challenges in 3D crowd estimation are maintaining spatial resolution and managing occlusions.

Preserving spatial resolution is essential as scene complexity increases, but models often struggle with high-resolution individual depiction. Managing occlusions in dense scenes, where individuals overlap or conceal each other, is also challenging. Existing models [31, 15] often have to choose between capturing a scene's expanse and preserving individual details, frequently missing the precise attributes and postures of obscured individuals.

This paper introduces the 3DCrowd+ human reconstruction network, inspired by existing models. It integrates HRNet [29] with the Shuffle Block module [17] and CoordAttention [9], maintaining high-resolution portrayals while processing image attributes of individuals in crowds. Using these 2D depictions, the model predicts individual depth, combining this information into a specialized 3D network. The result is a mesh capable of accurately handling occlusions within crowds. The following sections will explore this 3D reconstruction approach and its implications for AR applications.

II. Related Work

A. Multi-person Pose Estimation

Deep learning-based multi-person pose estimation operates under two paradigms: top-down and bottom-up.

1) Top-down Paradigm: The top-down strategy detects individuals first, then assesses each person's pose [30, 5, 29]. It is effective for single-person pose estimation but can be slow and struggles in crowded environments due to overlapping bounding boxes and occlusions.

2) Bottom-up Paradigm: The bottom-up approach [14, 2] identifies all body parts or keypoints first, then links them to individuals. OpenPose [20] exemplifies this method, which is more efficient as it bypasses initial individual segmentation. However, it can be challenging to correctly associate keypoints in dense crowds with significant occlusions. These methods often lack shape data, crucial for detailed applications.

B. Multi-person Pose and Shape Estimation

Pose estimation focuses on body joints, while shape estimation captures individual physique contours. Extracting a 3D body mesh from a single RGB image is challenging due to limited 3D information and various distortions like background, lighting, and clothing texture.

Hogg et al. [8] introduced the WALKER model, translating images into textual human attributes. Subsequent research [23, 7] used iterative optimization to refine 3D body models based on 2D annotations. Recent efforts combine parametric human models with deep learning [28], reducing data requirements by leveraging parametric blueprints. Lassner et al. [13] used convolutional neural networks to detect keypoints for 3D reconstruction but still required extensive annotated data. Researchers are now using 2D pose datasets [22, 6] to minimize keypoint detection needs.

1) Parametric Body Models: Key parametric models in 3D reconstruction include SCAPE [1] and SMPL [18], offering detailed triangular mesh representations. SMPL is more precise and compatible with rendering engines, capable of deriv-

ing 3D keypoints directly from its surface. It has significantly advanced 3D human shape reconstruction.

2) Multi-stage vs. Single-stage: Some 3D models use two-stage frameworks, reconstructing each detected person individually. Jiang et al. [12] developed CRMH, refining it with penetration and perceptual loss but struggled with accurate dimension reconstructions. Choi et al. [3] introduced 3DCrowdNet to handle occlusion using HRNet, effective in dense environments but computationally demanding. Single-stage solutions [34, 25, 26] offer efficient end-to-end mesh recovery but rely on low-resolution inputs, limiting high-resolution capabilities and yielding only relative depths.

III. Methodology

A. Model

This research introduces 3DCrowd+, a resilient two-stage method for 3D human mesh reconstruction in dense, occluded settings. 3DCrowd+ modifies the HRNet 2D pose estimation architecture from 3DCrowdNet by integrating Lite-HRNet for a lightweight structure. By combining Shuffle Block [19] and CoordAttention, we maintain feature extraction capabilities while achieving a lightweight framework, enabling precise 3D mesh estimation in crowded environments.

1) Baseline Networks:

- HRNet [2]: HRNet preserves high-resolution features using parallel pathways, avoiding spatial detail loss from downsampling and upsampling. It combines outputs from each stage for accurate 2D models that support detailed 3D meshes in 3DCrowd+.
- Lite-HRNet [33]: Lite-HRNet is a streamlined version of HRNet, optimized for human pose estimation. It reduces channel dimensions and redundant computations while maintaining spatial data integrity, merging multi-resolution representations efficiently to balance efficacy and efficiency.
- CoordAttention [9]: CoordAttention enhances feature extraction by integrating x and y coordinate data, improving positional precision and reducing confusion in overlapping figures without significant computational overhead.

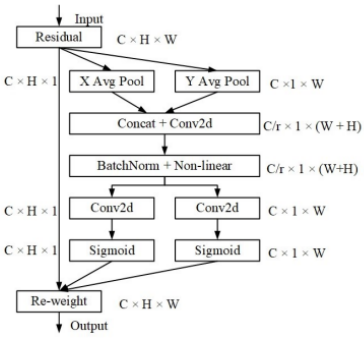

Fig. 1. CoordAttention Structure [9]

- Shuffle Block [35]: Shuffle Block, part of ShuffleNetV1, reduces computational complexity and parameters using Pointwise Group Convolutions and Channel Shuffle. Ma et al. [19] improved this with the Channel Split technique, creating a dual-branch structure. One branch undergoes 1x1 and 3x3 convolutions, while the other remains untouched. Outputs are fused and channels are rearranged using Channel Shuffle.

2) Structural Innovations: To balance feature extraction and efficiency, we introduce the Shuffle Attention Block by merging Shuffle Block with CoordAttention. Conditional Channel Weight (CCW) [36] is also incorporated to optimize time complexity. The structure is visualized in Figure 3.

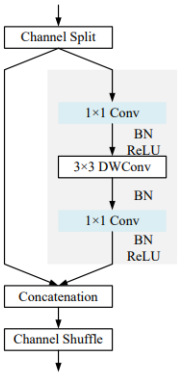

Fig. 2. Shuffle Block Structure [19]

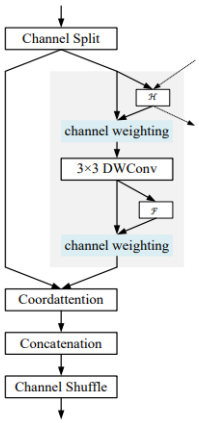

Fig. 3. The presented Shuffle Attention Block structure

The conditional channel weighting method has a complexity that is linear to the number of input channels. Time complexity is significantly lower compared to 1×1 point-by-point convolution. Furthermore, the method effectively replaces the role of 1x1 point-by-point convolution using weights H and F as a bridge of information exchange between channels from other branches and the resolution. In HRNet, the s stage has s parallel branches and the corresponding s weight maps W, W_2, ...,W_s. The weight maps of all channels at different branching rates are computed cross-channel by the lightweight module H (Cross-resolution Weight Computation), which can be represented as Figure 4. X denotes the different branch input feature maps, X_1 is the maximum branching rate feature map and X_s is the minimum resolution feature map.

$$(W_1, W_2, \ldots, W_s) = \mathcal{H}_s(X_1, X_2, \ldots, X_s)$$

Fig. 4. Cross-resolution weight formula

To produce a specific output size $X_{\prime}$, we begin by generating the input feature map using adaptive average pooling. This output size is then spliced to obtain the weighted feature map by following the steps described in Figure 5.

Therefore the use of the lightweighting module H (Spatial Weight Computation) to accept weight maps from different

$$(X_1', X_2', \ldots, X_s') \rightarrow Conv. \rightarrow RELU \rightarrow Conv. \rightarrow Sigmoid \rightarrow (W_1', W_2', \ldots, W_s')$$

Fig. 5. Weighted feature map formula

branches with different resolutions can serve to exchange information across channels and resolutions. After the output of the lightweighting module H has been convolved with a depth of 3×3 as an input into the module F. The module calculates the spatial weights for each of the different splitting rates by calculating the weight vectors w_s; which are the same for all positions, and the spatial weights depend on all the pixels of the input channel, which can be represented as Figure 6.

$$w_s = \mathcal{F}_s(X_s)$$

Fig. 6. Spatial weight computation formula

The implementation flow of module F can be represented as Figure 7.

$$X_s \rightarrow GAP \rightarrow FC \rightarrow RELU \rightarrow FC \rightarrow sigmoid \rightarrow w_s$$

Fig. 7. Global channel capture formula

It can be seen that the use of the on structure can exploit the correlation between the input feature map passes to capture the channel scale dependencies from the global. The computation introduced by the Shuffle Attention Block is only marginally higher than the base Shuffle Block. Its design not only allows adaptive weighting from multi-resolution inputs and network channels but also excels at harnessing cross-channel information both spatially and directionally. By integrating the Shuffle Attention Block into Stages 2-4, we've crafted the lightweight pose estimation network central to this study, with its architecture presented in Figure 8.

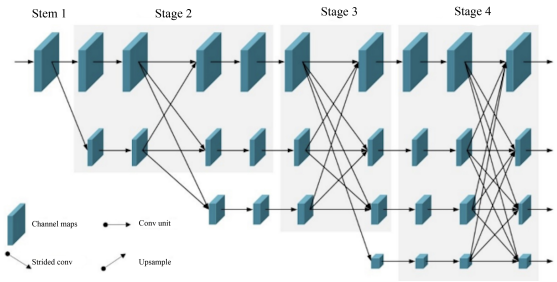

Fig. 8. Human posture estimation network structure

3) Stage Integration: 3DCrowd+ adopts a dual-phase approach. In the first phase, it leverages the lightweight human pose estimation network presented in this study for 2D human pose recognition. During the subsequent 3D human mesh reconstruction phase, 3DCrowd+ is built upon the foundational architecture of 3DCrowdnet. Its primary goal is to discern depth and shape disparities within the focal figure, relying on the image features extracted from the 2D phase. To sustain the spatial activation specific to the target figure, 3DCrowd+ incorporates the union-based regressor from 3DCrowdnet. This ensures the distinction of individual features amidst a crowd. The methodology zeroes in on isolating image features rooted in the joint positions of the focal person, enabling it to differentiate from non-target individuals. Consequently, in scenarios with dense crowds and frequent interpersonal occlusions, 3DCrowd+ can adeptly pinpoint and reconstruct the 3D human mesh.

B. Dataset

The MS COCO 2017 dataset [27] was used to train the human pose estimation network in 3DCrowd+. COCO provides annotations for 17 keypoints, including joints like the nose, eyes, shoulders, and ankles. It offers a wide range of human postures, activities, interactions, and occlusions, making it a robust training environment. Each keypoint annotation includes (x, y) coordinates and visibility indicators, which help model human postures even in crowded scenes. The diversity and challenges in COCO allow for a thorough evaluation of the model's resilience and precision. For testing, we used the 3D Poses in the Wild dataset [16], which offers precise 3D pose annotations in natural environments. It includes 60 sequences with over 51,000 frames featuring 7 actors in 18 different outfits. This dataset provides valuable diversity for testing and training. Unlike HumanEva [24] and H3.6M [11], which focus on controlled indoor settings, 3D Poses in the Wild offers accurate 3D pose data in complex outdoor scenarios.

IV. Experimental Results and Analysis

To validate the efficacy of the algorithm developed in this thesis, the necessary software platforms were installed based on the algorithm's requirements and prevailing conditions. An experimental framework was constructed within the testing environment, and pertinent datasets were chosen. These measures were taken to guarantee a stable and consistent experimental environment, thereby ensuring accurate validation of the algorithm.

A. Experimental Environment

All experiments in this paper are based on the Linux operating system, and the specific experimental environment configuration is shown in the following Figure 9.

Operating System	Windows10
Experimental Platforms	Desktop-DCKTIN9
Processor	Intel®CoreTMi7-8700 CPU@3.20GHz×12
Memory	32G
Graphics Card	Nvidia 3060
Video Memory	20G

Fig. 9. Experimental Environment

The experiments use PyTorch deep learning framework, and the specific steps for building the system environment are shown below.

- Installation of a Linux dual system on top of a Windows 10 system.

- Installation of Miniconda virtual environment and configuration of header files, library file settings.
- Install Pytorch 1.7.1, Python 3.7.3 and its dependencies.

B. Model Training

Figures 10 show the accuracy curve and model loss curve of the proposed model with 280 batches of training. From the figure, it can be seen that the model loss starts to converge at 170 batches of training, and the final model accuracy can reach about 82%.

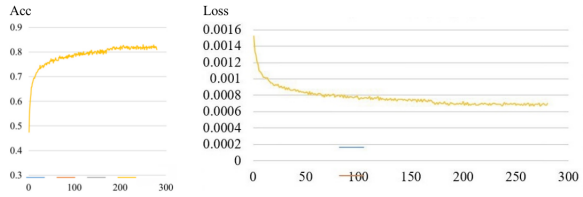

Fig. 10. Accuracy and Loss after trained 280 epoch

C. Evaluation Standards

We employ both quantitative and qualitative methods for evaluation. Quantitatively, we benchmark the refined 2D pose and 3D mesh estimation networks in 3DCrowd+ against existing high-performing models. Qualitatively, we compare 3D mesh outputs from 3DCrowd+ with the original 3DCrowdNet.

1) Quantitative Evaluation: We benchmarked our algorithms against state-of-the-art models using the COCO2017 dataset. Larger networks included Hourglass, CPN, SimpleBaseLine [32], and HRNet; smaller ones included MobileNetV2 [21], Shuffle NetV2, and Small HRnet. The results are shown in Figure 11.

With a 256×192 input image, our model has 1.76M parameters and 0.42GFLOPs, achieving an AP score of 67.3. For a 384x288 input, it maintains 1.76M parameters with 0.95GFLOPs, reaching an AP of 70.3, demonstrating high accuracy with a lightweight structure.

Model	Input size	#Params	GFLOPs	AP
Large networks				
Hourglass	256×192	25.1M	14.3	66.9
CPN	256×192	27M	6.2	68.6
SimpleBaseline	256×192	34M	8.9	70.4
HRNet	256×192	28.5M	7.1	74.4
Small networks				
MobileNetV2	384×288	9.6M	3.33	67.3
Shuffle NetV2	384×288	7.6M	2.87	63.6
Small HRNet	384×288	1.3M	1.21	56
Ours	**256×192**	**1.76M**	**0.42**	**67.3**
Ours	**384×288**	**1.76M**	**0.95**	**70.3**

Fig. 11. Posture estimation networks

To evaluate 3D mesh estimation, we use Mean Per Joint Position Error (MPJPE), Procrustes-Aligned MPJPE (PA MPJPE), and Mean Per Joint Position Error Posture (MPVPE). We compare our model with previous methods [10, 12, 26] and the baseline 3DCrowdNet using real-world 3D pose datasets. The new 3DCrowd+ model shows improvements of 3.25% in MPJPE, 5.92% in PA MPJPE, and 1.81% in MPVPE over 3DCrowdNet, demonstrating its efficacy in 3D pose estimation.

Model	MPJPE↓	PA MPJPE↓	MPVPE↓
SPIN	121.2	69.9	144.1
Pose2Mesh	124.8	79.8	149.5
ROMP	104.8	63.9	127.8
3DCrowdNet	89.3	59.1	110.7
Ours	**86.4**	**55.6**	**108.7**

Fig. 12. 3d mesh reconstruction networks [3]

2) Qualitative Evaluation:

- A side-by-side qualitative comparison between our 3DCrowd+ model and the BASELINE 3DCrowdNet is presented in Figure 13. When both models are tested using the specified "3D Poses in the Wild" dataset, it's evident that 3DCrowd+ offers a more resilient 3D mesh representation in complex, crowded scenarios. Furthermore, it demonstrates superior performance in capturing spatial relationships.

Fig. 13. Validation results of 3D Poses in the Wild dataset

- In Figure 14, this paper further evaluates the effectiveness of the 3DCrowd+ model in various application scenarios by selecting multi-person images of different sizes and in different scenarios.

V. Conclusion and Further Works

This thesis presented 3DCrowd+, an enhanced version of 3DCrowdNet, designed for 3D reconstruction in crowded scenes. By integrating an attention mechanism and model pruning algorithm, 3DCrowd+ performs robustly on field datasets with lightweight parameters.

However, 3DCrowd+ has limitations, such as difficulties in capturing subtle details like hand nuances and inconsistencies in character positioning on a uniform ground plane (Figure 14). Future work will address these issues by enhancing

Fig. 14. Verification results of real application scenarios

adaptive models for specific character detection and refining character bit value predictions. Integrating this method into AR devices will allow user perception experiments, aiding in optimization of user interactions with reconstructed scenes. Future research will focus on:

- Specialized Hand Detection: Developing a network dedicated to hand detection to improve accuracy.
- Uniform Ground Plane Constraints: Creating strategies to maintain consistent character positioning in larger scenes.
- Reducing Dependency on Pre-trained Models: Decreasing reliance on SMPL models for greater robustness.
- Enhancing Detection of Smaller Characters: Improving sensitivity to smaller characters for comprehensive scene representation.

In conclusion, 3DCrowd+ shows significant advancements in 3D reconstruction from crowded scenes, but further refinement is needed. By incorporating this method into AR devices and conducting user perception experiments, we aim to enhance user interaction and experience in augmented reality environments.

REFERENCES

[1] Dragomir Anguelov et al. "SCAPE: Shape Completion and Animation of People". In: *ACM Trans. Graph.* 24.3 (July 2005), pp. 408–416. ISSN: 0730-0301. DOI: 10.1145/1073204.1073207. URL: https://doi.org/10.1145/1073204.1073207.

[2] Bowen Cheng et al. "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5386–5395.

[3] Hongsuk Choi et al. "Learning to estimate robust 3d human mesh from in-the-wild crowded scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1475–1484.

[4] Epic. *Metahuman: Realistic person creator*. 2021. URL: https://www.unrealengine.com/en-US/metahuman.

[5] Hao-Shu Fang et al. "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[6] Zhijie Fang and Antonio M López. "Intention recognition of pedestrians and cyclists by 2d pose estimation". In: *IEEE Transactions on Intelligent Transportation Systems* 21.11 (2019), pp. 4773–4783.

[7] Peng Guan et al. "Estimating human shape and pose from a single image". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, pp. 1381–1388.

[8] David Hogg. "Model-based vision: a program to see a walking person". In: *Image and Vision computing* 1.1 (1983), pp. 5–20.

[9] Qibin Hou, Daquan Zhou, and Jiashi Feng. "Coordinate attention for efficient mobile network design". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 13713–13722.

[10] Buzhen Huang, Tianshu Zhang, and Yangang Wang. "Pose2UV: Single-Shot Multiperson Mesh Recovery With Deep UV Prior". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 4679–4692. DOI: 10.1109/TIP.2022.3187294.

[11] Catalin Ionescu et al. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (July 2014), pp. 1325–1339.

[12] Wen Jiang et al. "Coherent reconstruction of multiple humans from a single image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 5579–5588.

[13] Christoph Lassner et al. "Unite the people: Closing the loop between 3d and 2d human representations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6050–6059.

[14] Jia Li, Wen Su, and Zengfu Wang. "Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 11354–11361.

[15] Kun Li et al. “MILI: Multi-person inference from a low-resolution image”. In: *Fundamental Research* 3.3 (2023), pp. 434–441.

[16] Tsung-Yi Lin et al. “Your Title Here”. In: *Your Journal Here* Volume Here (Year Here), Pages Here.

[17] Sitong Liu et al. “Block shuffling learning for deep fake detection”. In: *arXiv preprint arXiv:2202.02819* (2022).

[18] Matthew Loper et al. “SMPL: A Skinned Multi-Person Linear Model”. In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16.

[19] Ningning Ma et al. *ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design*. 2020. arXiv: 1807.11164 [cs.CV].

[20] Daniil Osokin. “Real-time 2d multi-person pose estimation on cpu: Lightweight openpose”. In: *arXiv preprint arXiv:1811.12004* (2018).

[21] Mark Sandler et al. “Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation”. In: *CoRR* abs/1801.04381 (2018). arXiv: 1801.04381. URL: http://arxiv.org/abs/1801.04381.

[22] Wenkang Shan et al. “P-stmo: Pre-trained spatial-temporal many-to-one model for 3d human pose estimation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 461–478.

[23] Leonid Sigal, Alexandru Balan, and Michael Black. “Combined discriminative and generative articulated pose and non-rigid shape estimation”. In: *Advances in neural information processing systems* 20 (2007).

[24] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. “HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion.” In: *Int. J. Comput. Vis.* 87.1-2 (2010), pp. 4–27. URL: http://dblp.uni-trier.de/db/journals/ijcv/ijcv87.html#SigalBB10.

[25] Yu Sun et al. “Monocular, one-stage, regression of multiple 3d people”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11179–11188.

[26] Yu Sun et al. “Putting people in their place: Monocular regression of 3d people in depth”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13243–13252.

[27] SuperAnnotate. *Introduction to the coco dataset*. Aug. 2023. URL: https://opencv.org/blog/2021/10/12/introduction-to-the-coco-dataset/.

[28] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. “Indirect deep structured learning for 3d human body shape and pose prediction”. In: (2017).

[29] Jingdong Wang et al. *Deep High-Resolution Representation Learning for Visual Recognition*. 2020. arXiv: 1908.07919 [cs.CV].

[30] Shih-En Wei et al. “Convolutional pose machines”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 4724–4732.

[31] Hao Wen et al. “Crowd3D: Towards hundreds of people reconstruction from a single image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 8937–8946.

[32] Bin Xiao, Haiping Wu, and Yichen Wei. *Simple Baselines for Human Pose Estimation and Tracking*. 2018. arXiv: 1804.06208 [cs.CV].

[33] Changqian Yu et al. “Lite-hrnet: A lightweight high-resolution network”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 10440–10450.

[34] Jianfeng Zhang et al. “Body meshes as points”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 546–556.

[35] Xiangyu Zhang et al. *ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices*. 2018. arXiv: 1707.01083 [cs.CV].

[36] Quan Zhou et al. *Multi-modal medical image fusion based on densely-connected high-resolution CNN and hybrid transformer - neural computing and applications*. July 2022. URL: https://link.springer.com/article/10.1007/s00521-022-07635-1.