3D Human Pose & Mesh Reconstruction from 2D Image

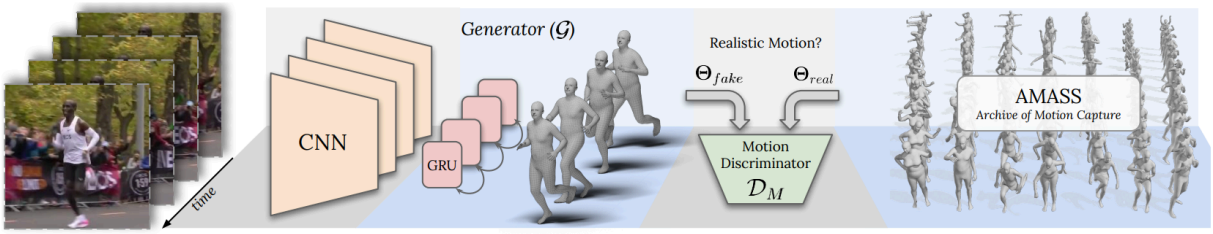

Image 1: VIBE Architecture. Kocabas et al., 2020

Abstract

A growing area of relevance in computer vision is the subject of 3D human pose and mesh reconstruction from 2D photos. This method aims to precisely derive a 3D human pose and mesh from a 2D RGB image. The subject of 3D human posture and mesh reconstruction is expanding quickly and has a wide range of possible uses. It may be applied to motion capture, augmented reality, and virtual reality, for instance. Given the inherent ambiguity and complexity required in interpreting 2D data for 3D reconstruction, this job offers a variety of difficulties, particularly in uncontrolled situations with varying lighting, viewpoints, and occlusions. This paper examines the most recent approaches, concentrating on deep learning-based ways to resolve this complex problem. Additionally, this paper discusses the shortcomings and suggests methods to enhance cutting-edge techniques for increased reconstruction accuracy.

1. Introduction

A fascinating topic of study in the science of computer vision is the reconstruction of 3D human pose and mesh from 2D images. This technique, which attempts to produce a 3D depiction of the human pose and mesh from 2D data, has uses in surveillance, healthcare, entertainment, virtual and augmented reality, and more. As per Martinez et al., 2017, reconstructing a 3D human position and model from 2D photos is still a difficult process, despite the field's tremendous progress. The inherent ambiguity of converting 2D data into a 3D context and the variety of real-world situations, which include various lighting conditions, views, and possible occlusions, are the main causes of this challenge. As per Pavlakos et al., 2018, CNNs and GANs are often used because of their capacity to learn complicated patterns and provide results of excellent quality. These models might be challenging to obtain because they frequently rely on huge amounts of annotated training data. In order to increase the precision of 3D mesh reconstruction, further study is being done in this area.

2. Related Work

2.1. 3DCrowdNet: Choi et al., 2022 created 3DCrowdNet, a deep learning system for predicting 3D human models from crowded real-world scenarios. In two methods, 3DCrowdNet overcomes the issues of 3D human mesh estimation from crowded settings. To begin, it employs a 2D posture estimator to extract picture information from a target subject. This reduces the impact of occlusion and other crowd-related difficulties. Second, it estimates the 3D human mesh using a joint-based regressor. By collecting characteristics from the target's joint positions, this regressor retains the spatial activation of a target. As per the paper by Choi et al., 2022, 3DCrowdNet was tested on the 3DPW dataset, which comprises crowded situations in the wild. On this dataset, it achieved state-of-the-art results.

method	MPJPE↓	PA-MPJPE↓	MPVPE↓
HMR [21]	130	76.7	-
GraphCMR [24]	-	70.2	-
SPIN [23]	96.9	59.2	116.4
I2L-MeshNet [35]	93.2	57.7	110.1
Pose2Mesh [7]	89.5	56.3	105.3
Song *et al.* [47]	-	55.9	-
Fang *et al.* [8]	85.1	54.8	-
TUCH [38]	84.9	55.5	-
ROMP [49]	91.3	54.9	108.3
3DCrowdNet (Ours)	**81.7**	**51.5**	**98.3**

Image 2: Figures showing the results of the 3DCrowdNet model. Choi et al., 2022

Choi et al., 2022's 3DCrowdNet bridges the domain gap by explicitly instructing a deep CNN to extract a crowded scene-robust feature using a commercially available 2D pose estimator. Then, it uses a joint-based regressor to maintain the target's spatial activation while regressing SMPL parameters to separate the target from other people. The SMPL layer receives the parameters and outputs a 3D mesh. To keep things simple, we just display picture feature sampling on two joints. Numbers in network layers represent the dimensions of the output channel. A stride size is indicated by the value in the maximum pooling layer. The channel dimension of the graph convolutional blocks is specified per joint.

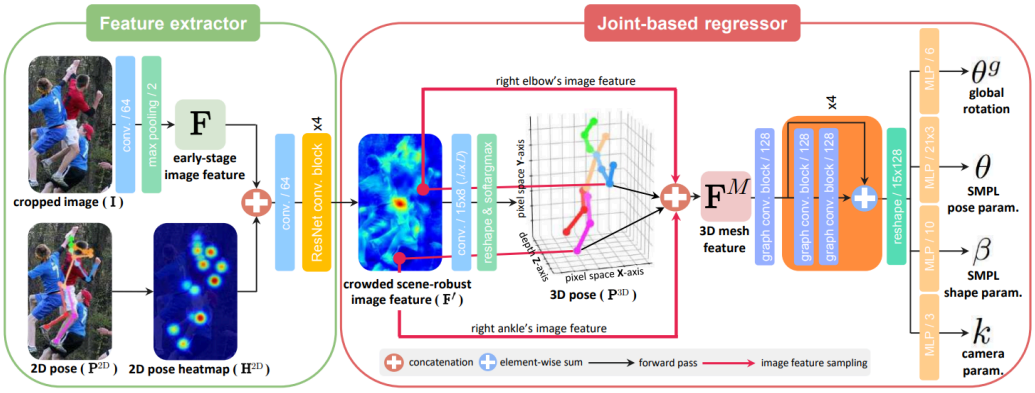

Image 3: 3DCrowdNet's Architecture by Choi et al., 2022

2.2. Pose2Mesh: Choi et al., 2021, introduced Pose2Mesh as a deep-learning model for 3D human posture and mesh reconstruction. According to the paper, the Pose2Mesh model generates a 3D pose and mesh of the human body from a 2D human posture, which may be either ground truth or estimated via a 2D pose estimator. Its graph convolutional network (GCN) design, which effectively captures the relationships between body joints for improved 3D pose prediction, is what makes this model distinctive. As per Choi et al., 2021, the Pose2Mesh approach also suggests a two-stage pipeline. The initial step of the procedure is estimating a 3D posture from the 2D input using a new graph convolutional auto-encoder. Using the projected 3D position and an SMPL model, the model predicts a full 3D human mesh in the second step. Pose2Mesh achieved state-of-the-art results on 3DPW and Human3.6M datasets.

method	Human3.6M		3DPW		
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPVPE
SMPLify [6]	-	82.3	-	-	-
Lassner et al. [34]	-	93.9	-	-	-
HMR [27]	88.0	56.8	-	81.3	-
NBF [47]	-	59.9	-	-	-
Pavlakos et al. [52]	-	75.9	-	-	-
Kanazawa et al. [28]	-	56.9	-	72.6	-
GraphCMR [32]	-	50.1	-	70.2	-
Arnab et al. [4]	77.8	54.3	-	72.2	-
SPIN [31]	-	**41.1**	-	59.2	116.4
Pose2Mesh (Ours)	**64.9**	46.3	**88.9**	58.3	106.3
Pose2Mesh (Ours)*	-	-	89.5	**56.3**	**105.3**

Image 4: Figures showing the results of Pose2Mesh on 3DPW and Human3.6M datasets. Choi et al., 2021

Choi et al., 2021 created Pose2Mesh in a cascaded architecture that includes PoseNet and MeshNet. PoseNet transforms a 2D human stance into a 3D human

pose. MeshNet estimates the 3D human mesh in a coarse-to-fine way using both 2D and 3D human postures. The mesh features are first processed at a coarse resolution and subsequently upsampled to a fine resolution during forward propagation.

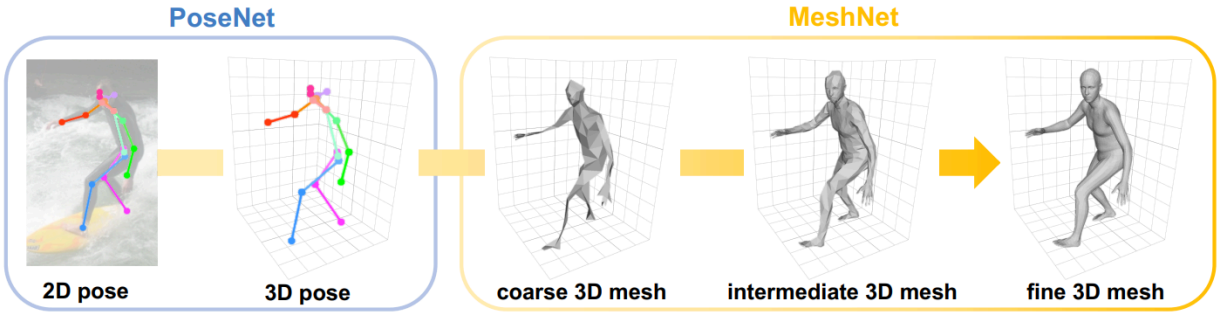

Image 5: Pose2Mesh Pipeline. Choi et al., 2021

2.3. I2L-MeshNet: Moon et al., 2020 introduced I2L-MeshNet, an image-to-lixel prediction network for accurate 3D human pose and mesh prediction from a single RGB image. I2L-MeshNet predicts the per-lixel probability using 1D heatmaps for each mesh vertex coordinate rather than explicitly regressing the parameters. This methodology offers various advantages over earlier methods. First, it retains the spatial connection between pixels in the input image, which is critical for accurate 3D pose and mesh estimation. Second, it models the prediction's uncertainty, which can make training more difficult but additionally result in more accurate findings. I2L-MeshNet performed decently well and achieved state-of-the-art scores.

methods	PA MPVPE	PA MPJPE	F@5 mm	F@15 mm	GT scale
Hasson et al. [10]	13.2	-	0.436	0.908	✓
Boukhayma et al. [5]	13.0	-	0.435	0.898	✓
FreiHAND [8]	10.7	-	0.529	0.935	✓
I2L-MeshNet (Ours)	**7.6**	**7.4**	**0.681**	**0.973**	✗

Image 6: Figures showing the performance of I2L-MeshNet. Moon et al., 2020

As per the paper by Moon et al., 2020, PoseNet and MeshNet make up I2L-MeshNet. From the given image, PoseNet generates three lixel-based 1D heatmaps of all human joints. PoseNet uses ResNet to extract picture features from the input image. The spatial scale is then increased by 8 times thanks to three upsampling units. Each upsampling module comprises a deconvolutional layer, a 2D batch normalization layer, and a ReLU function. The upsampled features are utilized to generate 1D human posture heatmaps based on lixels. The MeshNet and PoseNet have similar network architectures. MeshNet uses a pre-computed image feature from PoseNet and a 3D Gaussian heatmap instead of the input picture. A

fully-connected layer, a 1D batch normalization layer, and a ReLU function that transforms heatmaps to continuous joint coordinates comprise the building block. The final stage is to add a camera and mesh creation.

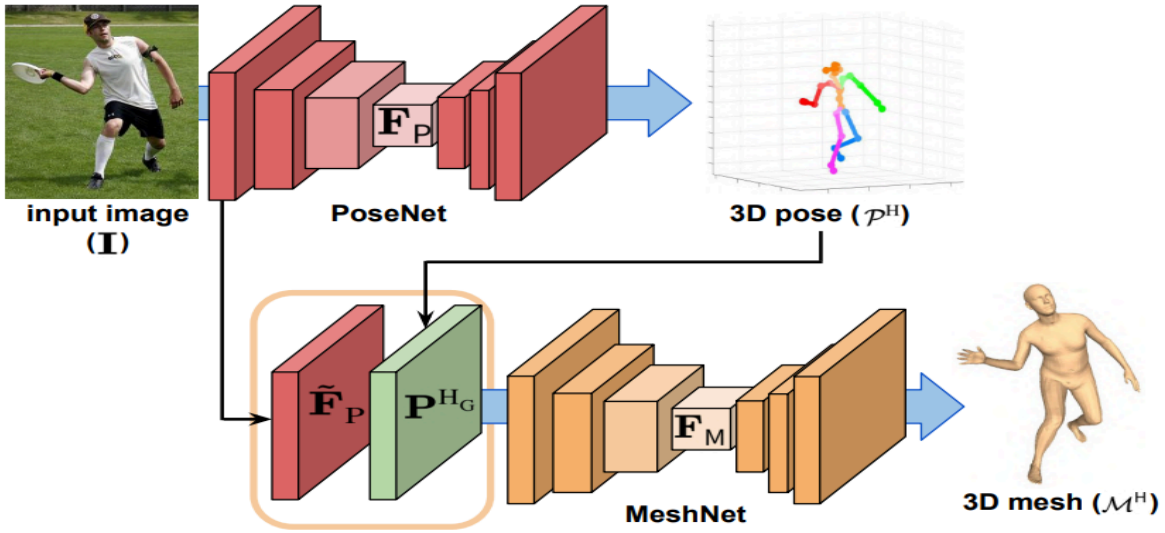

Image 7: I2L-MeshNet architecture. Moon et al., 2020

3. Key Problems

3.1. Ambiguity in 3D Pose from a Single 2D Image: As per Pavlakos et al., 2018, The depth information from the original 3D scene is lost in a 2D picture, resulting in substantial uncertainty in the different 3D positions that can yield the same 2D projection. Because of this inherent ambiguity, inferring the right 3D position from a single 2D photograph is difficult. Due to perspective uncertainty, an arm lifted forward may be perceived as raised sideways.

3.2. Occlusions: As per Marcard et al., 2019, Parts of the body are frequently obscured by other parts or objects in real-world circumstances. This makes determining the proper body stance and forms even more difficult. The main issue with occlusions is the loss of visual information. When a bodily part is obscured, its shape and location cannot be seen immediately in the picture. This can lead to uncertainty in pose assessment, making it more difficult to precisely estimate the 3D position of the obscured portion. Marcard et al., 2019 also noticed that occlusions frequently make detecting bodily components more difficult. This is particularly troublesome for systems that identify 2D body components before attempting to estimate the 3D position. An obscured part may not be identified at all or with poor confidence, resulting in mistakes in the following 3D posture estimate.

3.3. Variability in Human Shape and Clothing: As per Pavlakos et al., 2018, Human bodies come in a wide range of forms, sizes, and clothing styles. Clothing, in

particular, may significantly affect the look of body parts in a 2D picture, adding another degree of complication to producing 3D human posture and mesh. As per Bogo et al., 2016, because of the wide variation in body form, a pose estimation algorithm must be resistant to these variances and capable of reliably predicting postures for every human body shape. As per Pavlakos et al., 2018, clothing adds an additional degree of complication to the pose estimation problem. Different styles of clothes may substantially affect the look of body components in a 2D picture. Loose or baggy clothes might obscure the genuine contour of the body, making estimating the underlying position more difficult.

4. Chosen Topic

VIBE: Video Inference for Human Body Pose and Shape Estimation:

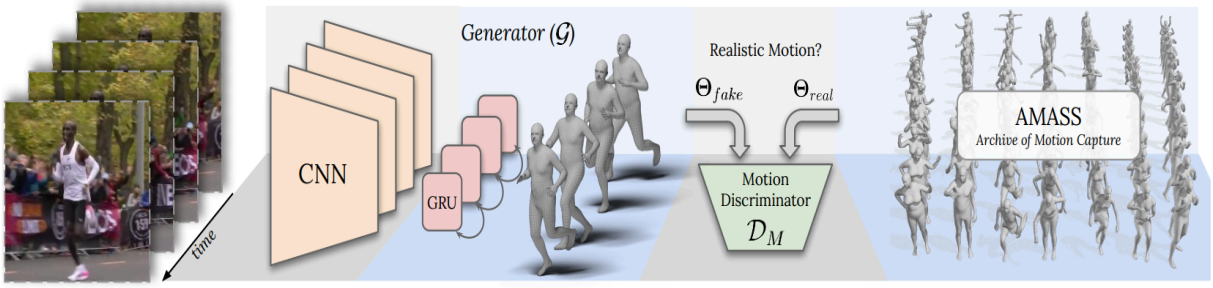

Image 8: VIBE Architecture. Kocabas et al., 2020

VIBE is a cutting-edge algorithm designed by Kocabas et al., 2020 to solve the challenge of estimating 3D human position and form from a single picture or video sequence. VIBE works by combining various deep learning approaches and models, each of which contributes to distinct elements of the problem at hand. The SMPL model (Bogo et al., 2016) lies at the core of VIBE. It is a learned statistical model that depicts a wide range of human body pose and mesh in a compact, low-dimensional, and continuous form. It generates a 3D mesh model of the human body using pose and shape characteristics, which VIBE utilizes to infer the 3D stance and form from a 2D picture.

As per the paper by Kocabas et al., 2020, VIBE takes a 2D picture or video sequence as input and detects the 2D human stance in the image using an off-the-shelf 2D pose estimator, such as SPIN, which is another crucial component in its design. The SPIN-detected 2D joints are then utilized to condition a temporal model, which is a recurrent module meant to capture temporal relationships across video frames. The output of the temporal model is input into a regressor, which calculates the SMPL parameters for each frame. This stage also includes an adversarial prior, which acts as a regularizer, encouraging the model to predict SMPL parameters that are likely to fall inside the SMPL parameter distribution learned from the training data. For

each frame, the anticipated SMPL parameters are utilized to create a 3D mesh representation of the body.

5. Suggested Improvements

5.1. Object Detection: As per Kocabas et al., 2020, detecting humans in input frames is frequently the first stage in VIBE's pipeline. This is commonly done with a commercially available 2D person detection model, which detects and localizes individuals in each frame by generating bounding boxes around each identified individual. This stage prepares the data for the pose estimation algorithm that follows. We may increase the precision, speed, and efficiency of object detection by incorporating **YOLOv5** into VIBE, hence improving the overall efficiency of the VIBE model.

As per the paper by Bochkovskie et al., 2020, **YOLOv5** employs a single-shot detection technique, allowing for real-time object recognition on both photos and videos. It achieves great accuracy by combining a deep convolutional neural network with anchor-based bounding box predictions. This design enables **YOLOv5** to recognize objects with varying sizes and aspect ratios successfully. Integrating **YOLOv5** with VIBE can give powerful object detection capabilities, allowing the model to recognize and track humans more effectively. **YOLOv5** outperformed YOLOv4 and other competing models with a mean average accuracy **(mAP) of 50.2%** on the **COCO dataset** in real time.

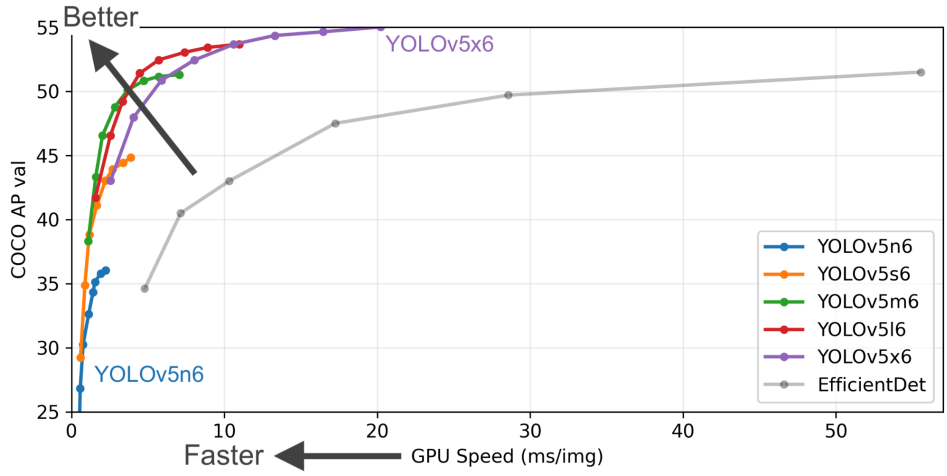

Image 9: Performance of YOLOv5 on COCO dataset. Ultralytics 2020

5.2. Pose Estimation: Pose estimation follows object detection. VIBE's pose estimation is not real-time, and therefore requires high-end technology to process, making the entire model computationally costly. **MoveNet** is a machine learning model that excels in estimating poses in real-time. It's fast, precise, and can estimate postures from a wide range of picture inputs, making it suited for a wide variety of applications. As per the research by Votel et al., 2021, **MoveNet** is intended

to improve accuracy as well as runtime performance. Because of this balance, it can conduct real-time posture estimation on devices with limited processing resources. The incorporation of **MoveNet** into the VIBE framework might greatly simplify the pose estimation procedure. MoveNet's quick 2D posture estimation may be used as an input to the VIBE model, decreasing the amount of expensive computation required for the initial pose estimation in VIBE. This might result in real-time 3D posture estimate with high accuracy. **MoveNet** fared extremely well with a mean average accuracy **(mAP) of 95.4%** on the **COCO dataset** in real-time.

Gender	% dataset	Keypoint mAP (Lightning)	Keypoint mAP (Thunder)
Male	46.0	90.2	93.7
Female	54.0	87.8	92.3

Age	% dataset	Keypoint mAP (Lightning)	Keypoint mAP (Thunder)
Young	87.6	89.1	93.3
Middle-age	10.5	89.3	91.5
Old	1.9	85.7	90.0

Skin Tone	% dataset	Keypoint mAP (Lightning)	Keypoint mAP (Thunder)
Darker	15.4	89.1	93.1
Medium	2.5	92.2	93.3
Lighter	82.1	92.9	95.4

Image 10: Performance of MoveNet on COCO dataset. Votel et al., 2021

6. Conclusion

To summarise, the reconstruction of 3D human position and mesh from 2D photos remains a difficult yet promising attempt in the area of computer vision. The inherent issues stem mostly from the ambiguity of translating 2D data to 3D, as well as the varied real-world settings such as variable lighting conditions, viewpoints, and probable occlusions. Despite these challenges, algorithms like CNNs and GANs have shown to be quite effective in learning complicated patterns and producing high-quality outputs. However, these models sometimes require significant labeled training data, which might be difficult to get. As a result, ongoing research and development are critical to overcoming these obstacles and improving the accuracy of 3D mesh reconstruction from 2D photos.

REFERENCES

- Julieta Martinez, 2017. A simple yet effective baseline for 3d human pose estimation [online]. *In:* Rayat Hossain, Javier Romero, James J. Little, eds., *Computer Vision and Pattern Recognition,* 2017. Available from: https://arxiv.org/abs/1705.03098 [Accessed 11 May 2023].
- Georgios Pavlakos, 2018. Learning to Estimate 3D Human Pose and Shape from a Single Color Image [online]. *In:* Luyang Zhu, Xiaowei Zhou, Kostas Daniilidis, eds., *Computer Vision and Pattern Recognition,* 2018. Available from: https://arxiv.org/abs/1805.04092 [Accessed 11 May 2023].
- Hongsuk Choi, 2022. Learning to Estimate Robust 3D Human Mesh from In-the-Wild Crowded Scenes [online]. *In:* Gyeongsik Moon, JoonKyu Park, Kyoung Mu Lee, eds., *Computer Vision and Pattern Recognition,* 2022. Available from: https://arxiv.org/pdf/2104.07300.pdf [Accessed 11 May 2023].
- Hongsuk Choi, 2021. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose [online]. *In:* Gyeongsik Moon, Kyoung Mu Lee, eds., *Computer Vision and Pattern Recognition,* 2021. Available from: https://arxiv.org/pdf/2008.09047.pdf [Accessed 12 May 2023].
- Gyeongsik Moon, 2020. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image [online]. *In:* Kyoung Mu Lee, eds., *Computer Vision and Pattern Recognition,* 2020. Available from: https://arxiv.org/pdf/2008.03713.pdf [Accessed 12 May 2023].
- Timo von Marcard, 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera [online]. *In:* Roberto Henschel, Michael J. Black, Bodo Rosenhahn, Gerard Pons-Moll, eds., *European Conference of Computer Vision*, 2018. Available from: https://virtualhumans.mpi-inf.mpg.de/papers/vonmarcardECCV18/vonmarcardECCV18.pdf [Accessed 12 May 2023].
- Federica Bogo, 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image [online]. *In:* Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, Michael J. Black, eds., *Computer Vision and Pattern Recognition,* 2016. Available from: https://arxiv.org/pdf/1607.08128.pdf [Accessed 13 May 2023].
- Muhammed Kocabas, 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation [online]. *In:* Nikos Athanasiou, Michael J. Black, eds., *Computer Vision and Pattern Recognition,* 2020. Available from: https://arxiv.org/pdf/1912.05656.pdf [Accessed 16 May 2023].
- Alexey Bochkovskiy, 2020. YOLOv5: Optimal Speed and Accuracy of Object Detection [online]. *In:* Chien-Yao Wang, Hong-Yuan Mark Liao, eds., *Computer Vision and Pattern Recognition,* 2020. Available from: https://arxiv.org/pdf/2004.10934.pdf [Accessed 16 May 2023].

- Ultralytics, 2020. YOLOV5-www.github.com-yolov5.png [photograph]. Available from: https://github.com/ultralytics/yolov5 [Accessed 17 May 2023].
- Ronny Votel, 2021. Next-Generation Pose Detection with MoveNet and TensorFlow.js [online]. *In:* Na Li, eds., *Tensorflow Blog*, 2021. Available from: https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html [Accessed 17 May 2023].