International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

2D Image Based Digital Anthropometry Using Deep Learning Approach

Ravindra B. Gadhiya[1] and Dr. Nilesh B. Kalani[2]

Submitted: 06/05/2024 **Revised**: 19/06/2024 **Accepted**: 26/06/2024

Abstract: Anthropometry is a tool which is widely used for human body parts measurement across diverse field of science. There are several conventional tools available for measurement like measure tape, clippers etc. These conventional anthropometric devices are quick being changed via way of means of modern AI based systems. Digital anthropometry (DA) is a relatively new technique for measuring the dimensions of human body parts. Estimating the pose of a human with the assist of a photograph or a video has these days acquired extensive interest from the medical community. An aim of the research work is to introduce Deep learning concept in digital anthropometry and to develop a novel 2D image based digital measurement system which is more efficient to deal with various limitations of existing techniques. Here for body parts measurement, advanced models of the segmentation and pose estimation is employed to get better results. Also, existing models for anthropometry is implemented. The analysis and comparison of the results with the other methods is presented for better understanding.

Keywords *- Digital Anthropometry, Semantic Segmentation, DeepLabV3+, Pose Estimation, BlazePose*

INTRODUCTION

Today, anthropometric measurement is used in a remarkably wide range of areas, e.g. the fitness industry, the fashion industry, the clothing industry, and ergonomic product design. In today's era, several digital anthropometry systems with various algorithms are available, with pros and cons. The huge demand for digital anthropometry originates from various fields. Also, the digital anthropometry system can compensate for cost, operating complexity, and required operating time with better accuracy and reliability. Ultimately, the measurement system should satisfy the needs of the operating person and the person who is being measured.

In the proposed work, we have studied various semantic segmentation models and selected DeepLabV3+ model for generating segmented images of people. Also, several algorithms of pose estimation have been examined, and based on their performance, we have selected the BlazePose model for identifying landmarks on the human body.

In the Yidong & Yigang (2019) method, the authors have used DeepLab model and the OpenPose model for generating the human body parts measurement. In this research, we have implemented the Yidong & Yigang (2019) method and generated the output. Also, we give a comparative analysis of the results from the proposed method and the Yidong & Yigang (2019) method.

SEMANTIC SEGMENTATION

2.1. Deeplabv3+

Different architectures are available to tackle the issue of segmentation. Semantic segmentation is a broad field that requires increased precision and accuracy. We have studied the various segmentation models shown in Table 1. Here, we have selected DeepLabV3+ for image segmentation of human image in our application based on the performance.

For the purpose of segmentation, a network has employed an encoder-decoder design or an atrous spatial pyramid pooling layer [1]. The encoder-decoder architecture is shown in fig. 1. Atrous Spatial Pyramid Pooling, or ASPP, is used to encode the multi-scale contextual data. The encoder-decoder retrieves both spatial and local data building design. The two approaches stated above are combined in the DeepLabv3+. With the use of ASP (Atrous Separable Convolution) and MAX (Modified Aligned Xception), a faster and more effective network has been developed for semantic segmentation. We may infer from the comparison above that DeepLabv3+ is a dependable technique for semantic segmentation. A thorough explanation of the DeepLabv3+ model is given here. The atrous separable convolution, which consists of both a point-wise and depth-wise convolution. The encoder-decoder architecture is displayed in fig. 1. The encoder in this model is the DeepLabv3 model.

[1]*Research Scholar, Electronics and Communication Department, School of Engineering, RK University, Rajkot, Gujarat, India*
[2]*Professor, School of Engineering, RK University, Rajkot, Gujarat, India*
[1]*ravindra.gadhiya@gmail.com*

Table 1. Comparison of various models for segmentation

Architecture	Performance (mIoU)
Deep Layer Cascade (LC) [2]	82.7
TuSimple [3]	83.1
Large Kernel Matters [4]	83.6
Multipath-RefineNet [5]	84.2
ResNet-38 MS COCO [6]	84.9
PSPNet [7]	85.4
IDW-CNN [8]	86.3
CASIA IVA SDN [9]	86.6
DIS [10]	86.8
DeepLabv3 [11]	85.7
DeepLabv3-JFT [11]	86.9
DeepLabv3+ (Xception) [1]	87.8
DeepLabv3+ (Xception-JFT) [1]	89.0

A version with the highest performance includes an altered Xception spine with several layers and impressive depth-wise separable convolutions. For the convolution challenge (1x1), the result from the atrous spatial pyramid pooling is provided. It has then been up-sampled using four factors. The outcome of the encoder spine convolution has also been mixed with the previous result and improved with each additional 1x1 convolution. The feature maps produce two convolution layers (3x3) at the next level, and then they are up sampled using four factors to provide an ultimate image with accurate segmentation.

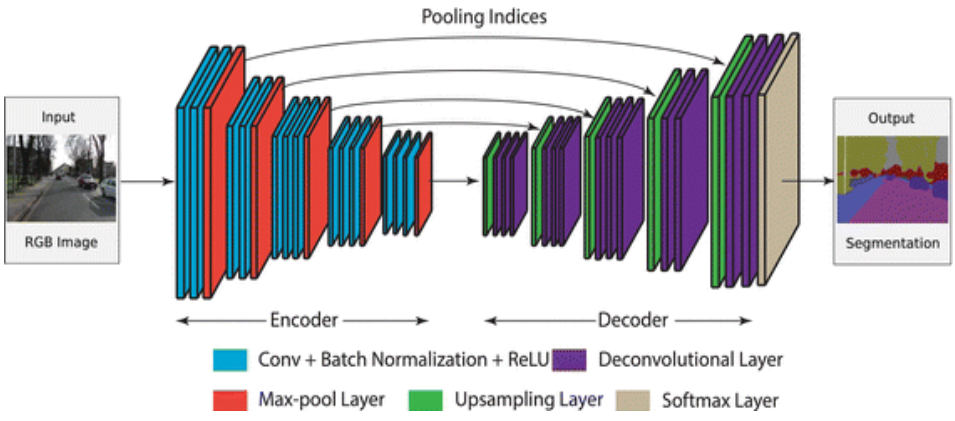

Fig. 1 Encoder decoder architecture

Pose Estimation

Pose estimation is a job that involves tracking, detecting, and associating points on various body parts using sensor data, 2D images, or videos. Human position identification from an image is a highly confined task because of the single-to-multiple mapping component that is inherent in this topic [12]. Additionally, CNN and RNN based Deep learning frameworks have been used to accomplish motion recognition utilizing RGB-D data [13]. The geometry and movement data that human posture estimation provides for the human body have been applied to a vast array of applications (e.g., augmented reality (AR), movement analysis, virtual reality (VR), healthcare, and human-machine interface). The high degree of accuracy as well as seamless performance are essential for any pose estimation system. Real-time applications now require improved posture estimation model performance due to technological developments. Nowadays, for image processing, the most

effective models are convolutional neural networks. Consequently, state-of-the-art methods frequently concentrate on tailoring CNN architecture for human pose inference.

In traditional object recognition, people are only seen as a box. Using position detection and tracking, computers can be trained to interpret human body language. On the other hand, conventional posture tracking systems are not feasible due to their slowness or inability to withstand occlusions. Real-time pose tracking and high-performance detection will power some of the biggest developments in computer vision. By tracking a person's position in real time, for instance, computers will be able to develop a more detailed and organic understanding of their behaviour. This will significantly affect several industries, including driverless vehicles. The majority of autonomous vehicle collisions now occur due to "robotic" driving, when an autonomous vehicle suddenly stops, causing a collision with a human driver. Real-time human posture detection and tracking enables computers to better comprehend and anticipate pedestrian behaviour, enabling more naturalistic driving.

Deep learning has been shown to outperform conventional computer vision algorithms in a variety of tasks such as image segmentation and object detection, thanks to the recent and rapid emergence of deep learning solutions

Consequently, posture estimation occupations have seen tremendous advancements and enhancements because of deep learning basis algorithms. All methods for posture estimation may be categorized using both top-down and bottom-up techniques. Bottom-up methods assess each body joint separately before combining them into a single position. Top-down methods estimate body joints inside the bounding boxes that have been detected by using a person detector as a starting point. Here, we use BlazePose architecture for finding landmarks on human body for anthropometric measurement creation.

BlazePose, a streamlined CNN framework, is optimized for instantaneous inference and tailored to assess an individual's body posture on mobile platforms. Utilizing the Pixel 2 phone's architecture, it achieves over 30 frames per second and delivers thirty-three (33) distinct landmarks or key points representing the body during inference. Illustrated in fig. 2, these 33 key points form the core output of BlazePose, rendering it exceptionally suitable for real-time tasks such as sign language recognition and fitness monitoring.

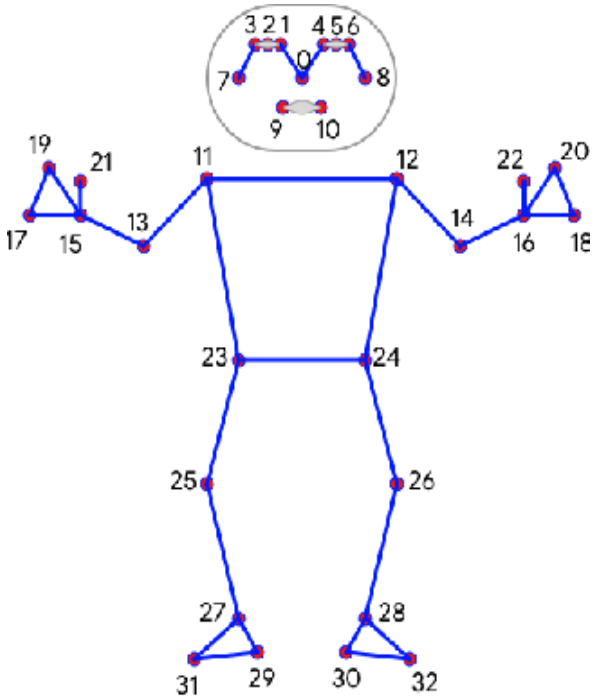

Fig. 2 BlazePose model – key points

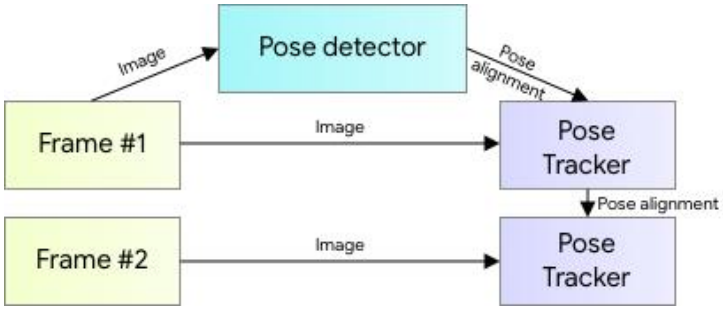

Fig. 3 BlazePose model

Newell accomplished this by linking multiple hourglass modules end-to-end within a single hourglass, enabling recurrent bottom-up and top-down inference across various scales [17]. In BlazePose, an encoder-decoder structure is employed to predict heatmaps for all joints, followed by a 2D encoder that condenses to the coordinates of all joints. Moreover, the heat-map branch is eliminated during inference in this version to reduce its weight for smartphone compatibility. For inference, the authors utilized a detector-tracker configuration, demonstrating effective real-time performance across various tasks such as hand landmark prediction and dense face landmark prediction. BlazePose architecture is illustrated in fig. 3. The pipeline involves initially employing a lightweight frame pose detector, followed by a pose tracker network.

The tracker forecasts the key-factor coordinates, determining a person's presence in the current frame, along with pinpointing the subtle area of interest within that frame. When the tracker detects no individuals present, it prompts the detector task to rerun on the subsequent frame. The BlazeFace model produces six facial key-factor coordinates, including those for eye centers, ear region, mouth center, and nose tip, facilitating estimation of face rotation (roll angle) and the prediction of axis-aligned face rectangles. This translation and rotation invariance enable the provision of a rotated face rectangle to the subsequent task-specific stages of the video processing pipeline, ensuring continuity in the subsequent processing steps [18].

Experimental Setup

For the measurement of a person's body parts, we have used a high-end system with an inbuilt camera. The anthropometry model requires input such as front and lateral images of people. Also, the height, weight, and gender of the person whose body parts are being measured. The Blazepoze model identifies the landmarks in the input images. The segmentation algorithm generates silhouette of the input images. The algorithm combines the segmented image with body part landmarks and applies mathematical formulae to generate human body part's measurement. The steps of digital anthropometry are displayed as a flowchart in fig. 4.

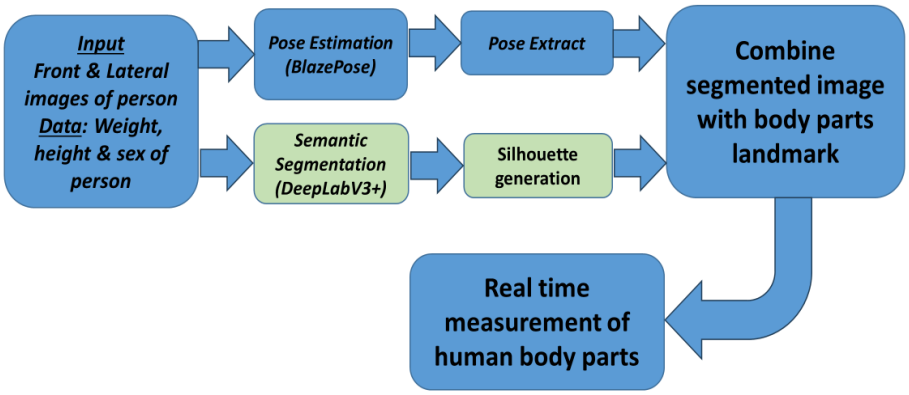

Fig. 4 Flow chart of the body parts measurement process

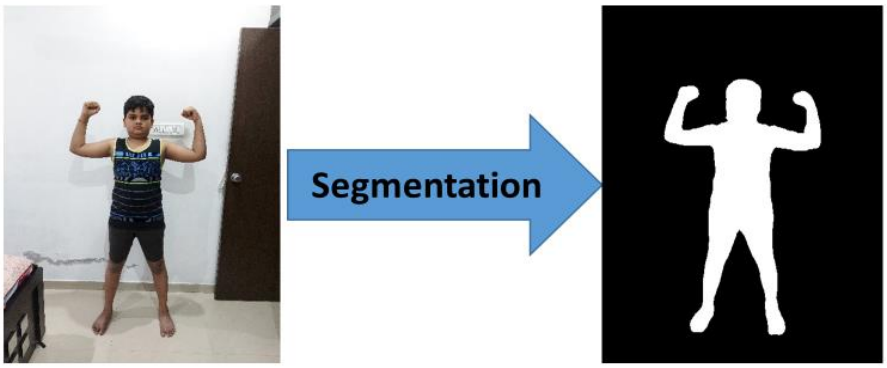

Fig. 5 Segmentation output (DeepLabV3+) of front image of person

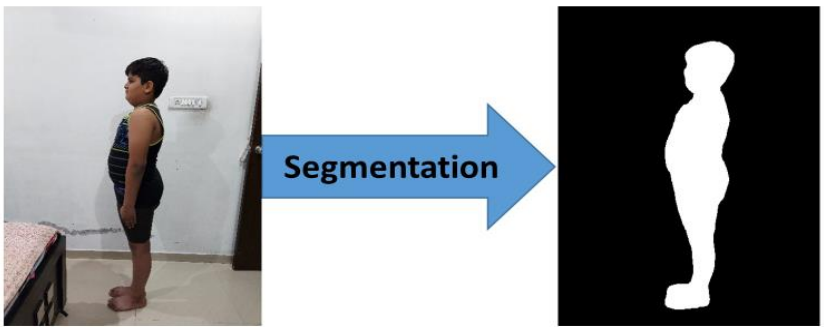

Fig. 6 Segmentation output (DeepLabV3+) of lateral image of person

Fig. 7 Pose Estimation output (BlazePose) of front image of person

Fig. 8 Pose Estimation output (BlazePose) of front image of person

The segmentation output of DeepLabV3+ model is shown in the fig. 5 and fig. 6 for front and lateral image of a person respectively. Fig. 7 and fig. 8 displays the output results of Blazepose algorithm for the same input images.

We performed the experiment on 32 males and 26 females with 140cm to 170cm height range. Also, we have tested the results with manual measurements to verify the effectiveness. We have also implemented Yidong & Yigang (2019) method to generate output. This method uses OpenpPose algorithm for identify the landmarks on body parts, and DeepLabV3 model for segmentation. The comparative analysis of proposed method with the Yidong & Yigang (2019) is presented in the Table. 2. Also

graphical representation is shown in the fig. 9. It shows that the proposed method can out performed over Yidong & Yigang method.

RESULTS AND DISCUSSION

By observing the comparison table, it is clear that the proposed method has performed efficiently in individual body part's measurement compared to Yidong & Yigang (2019) method. The overall performance of average parentage is 97.8301 by proposed method whereas 95.6137 by older method. The proposed method uses only one camera whereas Yidong & Yigang (2019) method used two cameras which has to be placed at 90 degrees to each other. Also, the background complexity is reduced by the proposed method compared to other method.

Table 2. Comparison of results

Body Parts	Proposed Method Accuracy (%)		Yidong and Yigang Method (2019) Accuracy (%)	
		Average		Average
Waistline	99.5949	97.8301	97.9497	95.6137
Leg Length (Belt to bottom leg)	99.1094		97.0717	
Arm Circumference (bicep)	95.4286		91.0273	
Neck Size	97.1875		96.4063	

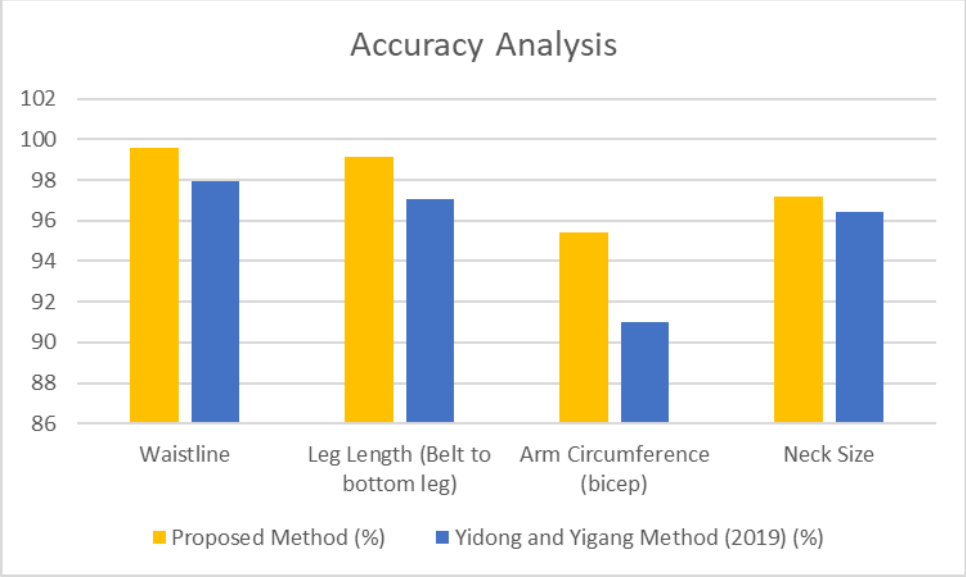

Fig. 9 Comparative analysis of accuracy

CONCLUSION

The experiments have been carried out with proposed method and Yidong & Yigang (2019) method. Also, verification of results has been done with manual measurement in this research work. It is clear that the proposed method more suitable for 2D image based digital anthropometry.

References

[1] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, pp. 833–851, 2018, doi: 10.1007/978-3-030-01234-2_49.

[2] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not All Pixels Are Equal: Difficulty-Aware Semantic Segmentation via Deep Layer Cascade," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3193-3202, 2017.

[3] . Wang et al., "Understanding Convolution for Semantic Segmentation," *Proceedings - 2018 IEEE Winter Conference on Applications of Computer*

Vision, WACV 2018, vol. 2018-Janua, pp. 1451–1460, 2018, doi: 10.1109/WACV.2018.00163.

[4] C. Peng, X. Zhang, G. Yu, G. Luo and J. Sun, "Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1743-1751, doi: 10.1109/CVPR.2017.189.

[5] G. Lin, A. Milan, C. Shen and I. Reid, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5168-5177, doi: 10.1109/CVPR.2017.549.

[6] Wu, Zifeng, C. Shen, and A. Van Den Hengel. "Wider or deeper: Revisiting the resnet model for visual recognition." *Pattern recognition 90 (2019)*: 119-133.FLEX Chip Signal Processor (MC68175/D), *Motorola*, vol. 15, no. 3, pp. 250-275, 1996.

[7] Zhao, Hengshuang, J. Shi, Xiaojuan Qi, X. Wang, and J. Jia. "Pyramid scene parsing network." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881-2890. 2017.

[8] Wang, Guangrun, P. Luo, L. Lin, and X. Wang. "Learning object interactions and descriptions for semantic image segmentation." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5859-5867. 2017.

[9] Fu, Jun, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu. "Stacked deconvolutional network for semantic segmentation." *IEEE Transactions on Image Processing*, pp. 01-12. 2019.

[10] G. Wang, P. Luo, L. Lin and X. Wang, "Learning Object Interactions and Descriptions for Semantic Image Segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5235-5243, doi: 10.1109/CVPR.2017.556.

[11] Chen, L. Chieh, G. Papandreou, F. Schroff, and H. Adam. "Rethinking atrous convolution for semantic image segmentation." *arXiv preprint arXiv*:1706.05587 (2017).

[12] Gong, Wenjuan, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E. Zahzah. "Human pose estimation from monocular images: A comprehensive survey." *Sensors 16, no. 12 (2016): 1966*.

[13] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding*, vol. 171, no. Wanqing Li, pp. 118–139, 2018, doi: 10.1016/j.cviu.2018.04.007.

[14] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Trans Pattern Anal Mach Intel*, vol. 43, no. 1, pp. 172–186, 2021, doi: 10.1109/TPAMI.2019.2929257.

[15] Bazarevsky, Valentin, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. "Blazepose: On-device real-time body pose tracking." *arXiv preprint arXiv:2006.10204 (2020)*.

[16] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014, doi: 10.1109/CVPR.2014.214.

[17] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912 LNCS, pp. 483–499, 2016, doi: 10.1007/978-3-319-46484-8_29.

[18] Bazarevsky, Valentin, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann. "Blazeface: Sub-millisecond neural face detection on mobile GPUs." *arXiv preprint arXiv:1907.05047 (2019)*.