Simultaneous Person, Face, and Hand Detector Based on Improved YOLOv5

Duy-Linh Nguyen, Xuan-Thuy Vo, Adri Priadana, and Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering,
University of Ulsan,
Ulsan, Korea
ndlinh301@mail.ulsan.ac.kr, xthuy@islab.ulsan.ac.kr, priadana@mail.ulsan.ac.kr, and acejo@ulsan.ac.kr

2023 International Workshop on Intelligent Systems (IWIS) | 979-8-3503-0504-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/IWIS58789.2023.10284687

***Abstract*—Many person, face, and hand detectors have reached remarkable achievements in the last century. With the convolutional neural networks' emergence, applications based on these detectors were increasingly widely applied in practice. Following this trend, this paper develops a simultaneous person, face, and hand detector based on YOLOv5 network architecture. The research focuses on redesigning the backbone and neck with compact network architectures EfficientNet and CBAM attention technique. The proposed architecture is trained and evaluated on the fine-tuned Human-Parts dataset. As a result, this network reached 87.2% of mAP@0.5 and 58.6% of mAP@0.5:0.95. This experimet outperforms most of the original YOLOv5 network architectures and is comparable to large-scale YOLOv5 architectures and the latest YOLOv8 architecture.**

***Index Terms*—Attention module, Convolutional Neural Network (CNN), Hand detection, Human detection, Lightweight architecture, Face detection, YOLOv5.**

I. Introduction

The face and hands in the overall human body are very important organs. They perform most human actions to interact with other bodies and the external environment. Many applications have been developed and applied in the computer vision field such as Human Action [1], Human-Computer Interaction (HCI) [2], and Human Tracking [3]. In particular, the strong emergence of CNN with its outstanding advantages over the past decades has released a lot of robust detection algorithms in person detection, face detection, and hand detection. However, the methods only focus on detecting each object individually, ignoring their inherent relationship. This study considers human body, hand, and face detection as a multi-level object detection task [4] to explore more effective object detection in the overall context. It is also consistent with the ability to perform on multi-scale objects of the YOLOv5 network architecture. Therefore, this research proposes an improved YOLOv5 network for simultaneous person, face, and hand detection. This technique takes advantage of compact, efficient network architectures such as EfficientNet to optimize network parameters. Besides, the use of the Convolutional Block Attention Module (CBAM) mechanism to guide the model to learn salient informative features. This combination aims to enhance multi-scale object detection and deployment on real-time applications.

The novel contributions of this research are:

- Proposed a simultaneous person, face, and hand detector based on the YOLOv5 network with lightweight architectures and CBAM attention mechanism to optimize the network parameters and detection ability.
- Inspected and fine-tuned the Human-Parts dataset for the YOLO architecture family. This dataset was trained and evaluated with proposed networks, all variants of the YOLOv5, and several lightweight YOLOv8 networks to compare the performance.

The remaining of the paper is arranged as follows: Section II presents the methods relative to the person, face, and hand detection. Section III introduces the proposed method in detail. Section IV summarizes and analyzes the experiment. Finally, Section V concludes the issue and plans the direction of the future works.

II. Related Work

The related work section introduces the related methods applied in human, hand, and face detection. These techniques can be considered with single human part detection and simultaneous human part detection methods.

A. Person detection methods

The pedestrian detection in [5] combined low-level features like Haar, HOG, and CNN with high-level features. MultiContext Cascaded Convolutional Networks (MCCNN) [6] integrated multiple context regions at different scales to effectively capture multi-scale pedestrian features. Fast R-CNN with Context Aggregation Networks (FRCN-CAN) [7] incorporates context information by using context aggregation networks, which gather global scene context to improve pedestrian detection accuracy within the Fast R-CNN framework.

B. Face detection methods

Authors in [8] proposed a face detector based on one self-learning method and one bi-channel network for face detection challenges such as multi-view and occlusion. The study in [9] improved the InceptionV2 network to build a cascaded CNN for detecting small-sized target faces. For low computing devices, FaceBoxes [10] combined the Rapidly Digested Convolution Layers and Multiple Scale Convolution Layers in a powerful face detection suite. FlashNet [11] was

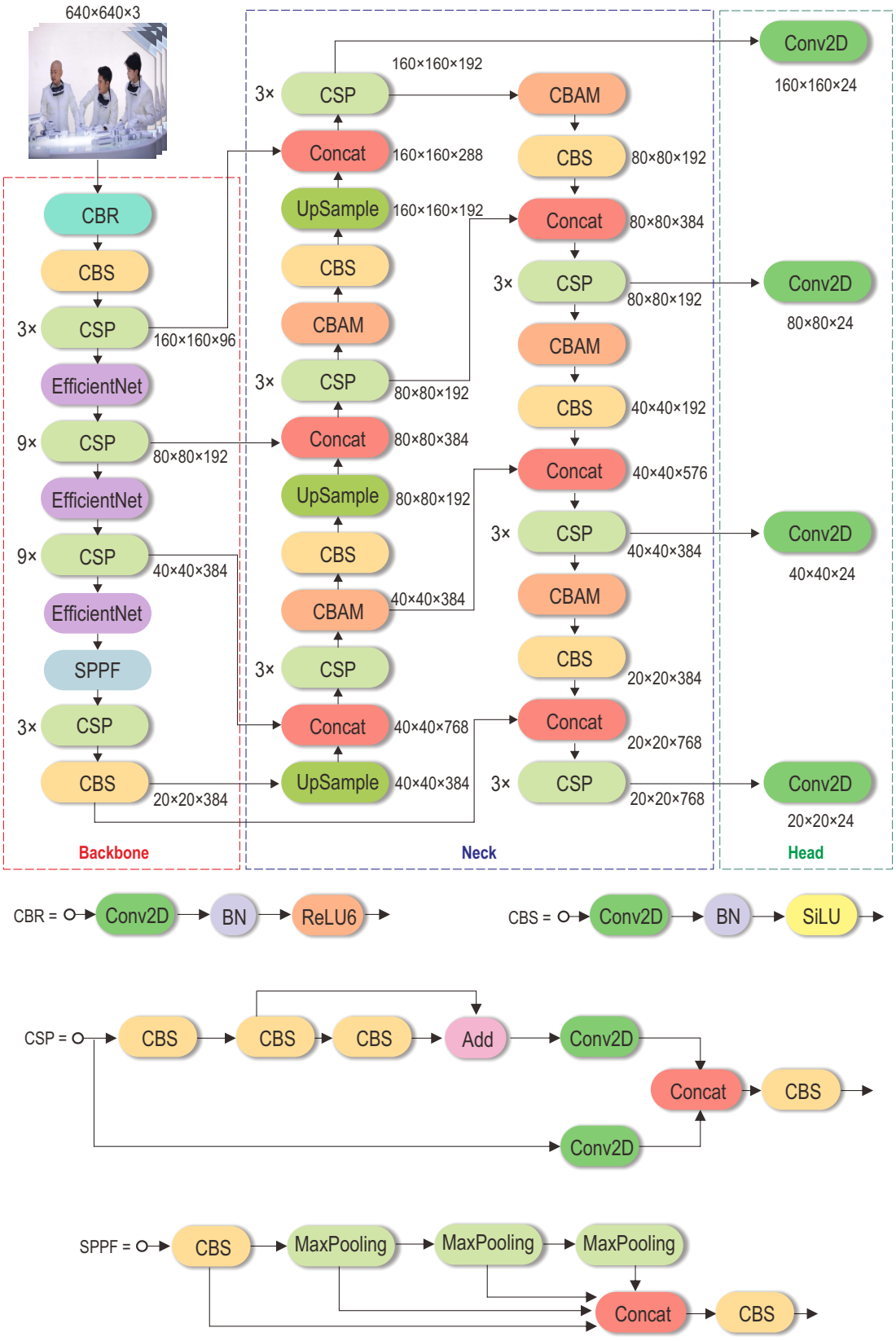

Fig. 1. The proposed network architecture overview and sub-modules.

developed from the MobileNetV2 network and free-anchor mechanism to optimize computation cost. This work [12] designed the lightweight CNN network architecture for an efficient face detector applied in surveillance systems.

C. Hand detection methods

The method in [13] introduced the one-stage CNN network for hand position and hand orientation detection. [14] applied the Faster-RCNN network to predict the position of the driver's hand on the steering wheel. A combination of the hand skin segmentation method and CNN in [15] to localize the hand position in different scenarios. Authors in [2] used the Mask-CNN network and new attention techniques for hand orientation and hand mask detection. [16] also proposed a hand detection reusing CNN and the attention module for robust hand localization ability.

D. Simultaneous human part detection methods

The authors in [17] presented a DID-Net (Detector-in-Detector network) for multi-level object detection, combining a lightweight parts detector after the Faster-RCNN network to deploy person, face, and hand detection. This work also collected a large dataset for this task, named the Human-Parts dataset.

III. Proposed Methodology

Fig. 1 shows the proposed network architecture overview. This network is comprised of three modules: backbone, neck, and detection head.

A. Backbone module

The Backbone module utilizes the architecture of the YOLOv5 backbone with some changes and replacements. Specifically, it retains the Cross Stage Partial (CSP), Spatial Pyramid Pooling Fast (SPPF) structures and incorporates lightweight blocks such as CBR (Convolution - BN - ReLU6), CBS (Convolution - BN - SiLU), and EfficientNet. In YOLOv5 architecture, the Focus block presents the advantages of feature extraction but it causes computational complexity and network parameters problems. From that observation, this work replaces the Focus block with another straightforward one, named the CBS block. This block is designed based on a convolution layer (1×1 Con2D) followed by a BN (Batch Normalization) and a ReLU6 activation function. Besides, the CONV block in the original YOLOv5 backbone is also replaced by the CBS block. The structure of CBS is the same as CBR but they are just different at activation function. The activation function in the CBS block is SiLU. The lightweight EfficientNet [18] block is designed quite simply with two convolution (1×1 Conv2D and 3×3 Conv2D) and one depthwise (3×3 DWConv2D) layers, interspersed with a BN and a ReLU6 activation function (for stride 2 case (Fig. 2 (b)). The same above design is for the stride 1 case (Fig. 2 (a)) but adds a skip connection to aggregate the input and output feature maps of the main branch using the addition operation. Its feature extraction process focuses on the channel dimension. The backbone module generates the multi-level informative feature map for the next module process. The structure of the backbone and sub-blocks are described in Fig. 1

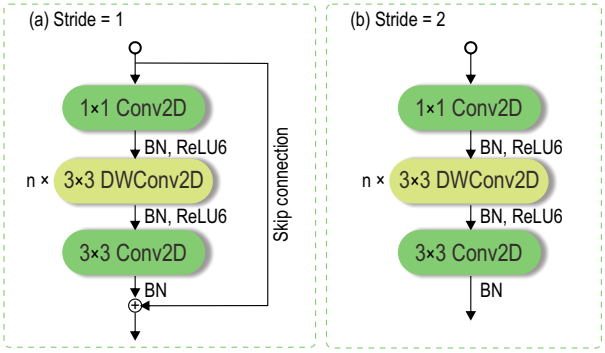

Fig. 2. The lightweight EfficientNet block.

B. Neck module

The Path Aggregation Network (PAN) [19] mechanism is still reused in the proposed neck module. This mechanism merges the current feature maps and previous upsampled feature maps using the concatenation operations. In addition, this work adds the CBAM [20] attention block after each CSP block to increase the focusing ability on salient features. The output of the neck module is four aggregated feature maps corresponding to the four scale levels of the object to be detected: large ($20 \times 20 \times 768$), medium ($40 \times 40 \times 384$), small ($80 \times 80 \times 192$), and very small ($160 \times 160 \times 192$). The CBAM structure is presented in Fig. 3

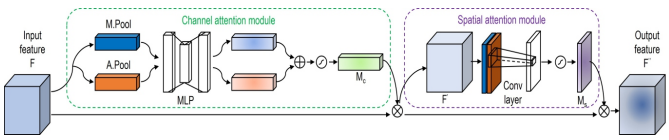

Fig. 3. The CBAM attention structure.

C. Detection head module

To build the detection head module, this study reutilizes three feature map levels like YOLOv5 for small, medium, and large object scales and adds one more level for a very small object scale. These four output feature maps go through four convolution operations to produce four detectors with dimensions $160 \times 160 \times 24$, $80 \times 80 \times 24$, $40 \times 40 \times 24$, and $20 \times 20 \times 24$ for very small, small, medium, and large object sizes, respectively. Each detector head applies three anchors of different sizes. Table I shows the details of each detector in the detection head module. The prediction coefficient of each detector is calculated as follows:

$$C = (5 + Class_n) \times Anchor = (5 + 3) \times 3 = 24, \quad (1)$$

where C denotes the detection coefficient, $Class_n$ denotes the number of classes, and $Anchor$ denotes the number of anchors.

TABLE I
THE DETAIL OF FOUR DETECTION HEADS.

Heads	Input	Anchor sizes	Ouput	Object
1	160 × 160 × 192	(5, 6), (8, 14), (15, 11)	160 × 160 × 24	Smallest
2	80 × 80 × 192	(10, 13), (16, 30), (33, 23)	80 × 80 × 24	Small
3	40 × 40 × 384	(30, 61), (62, 45), (59, 119)	40 × 40 × 24	Medium
4	20 × 20 × 768	(116, 90), (156, 198), (373, 326)	20 × 20 × 24	Large

TABLE II
THE PROPOSED NETWORK DETECTION RESULTS WITH EACH CLASS ON THE HUMAN-PARTS VALIDATION SET.

Class	Labels	P	R	mAP@0.5	mAP@0.5:0.95
all	14,354	87.9	83.7	87.2	58.6
person	4,796	91.4	78.6	87.7	63.2
face	3,728	89.3	90.8	90.7	63.5
hand	5,830	82.4	81.2	82.9	49.3

D. Loss function

This paper applies the loss function from YOLOv5 which is defined as follows:

$$Loss_{total} = \lambda_{box} Loss_{box} + \lambda_{obj} Loss_{obj} + \lambda_{cls} Loss_{cls}, \quad (2)$$

where $Loss_{box}$ is the bounding box regression loss is computed by using CIoU loss [21]. $Loss_{obj}$ and $Loss_{cls}$ are the object loss and the classification loss, respectively. They use the Binary Cross Entropy loss [22] to calculate. λ_{box}, λ_{obj}, and λ_{cls} are balancing parameters to control the overall loss.

IV. EXPERIMENTS

A. Dataset

The original Human-Parts dataset [17] includes 14,962 images with 106,879 annotations for three classes: person, face, and hand. The high-resolution images in this dataset were collected from IA-Challenger. The annotation format followed the PASCAL VOC dataset. During the experiment, this study recognizes that this dataset lost a lot of annotation when convert to YOLO format. Therefore, this work only uses 10,035 images including 8,038 images for the training phase and 1,997 images for the evaluation phase.

B. Experimental setup

The proposed network architecture is implemented based on YOLOv5 source code. This network was trained and evaluated on a Testla V100 GPU and tested on another GeForce GTX 1080Ti GPU. The input image size is 640 × 640 px. The learning rate is from 10^{-5} to 10^{-3}. The momentum is also assigned from 0.8 then gradually increases to 0.937. The Adam optimization is used for updating during training. The training phase uses 200 epochs and a batch size is 32. The balancing parameters are set as $\lambda_{box} = 0.05$, λ_{obj}=1, and $\lambda_{cls} = 0.5$. The training phase applies several data augmentation methods like translate, mosaic, flip, and scale. The inference time (ms) is implemented and reported with the same training input image size, and batch size while the IoU threshold and confidence threshold are set to 0.5.

C. Experimental results

The proposed network's performance is evaluated and compared with all versions of YOLOv5 (n, s, m, l, x) and two versions of YOLOv8 (s, n). These networks are trained from scratch on the Human-Parts dataset. Table III shows the comparison results. As a result, the proposed network reaches 87.2% of mAP@0.5 and 58.6% of mAP@0.5:0.95. With the mAP@0.5 metric, the proposed network's performance is slightly lower than that of YOLOv8s (0.5%↓) and YOLOv5l (0.1%↓) but it is superior to other network architectures. With the mAP@0.5:0.95 metric, the proposed network's performance is better than that of YOLOv8n (0.2%↑), YOLOv5n (4.9%↑), and YOLOv5s (1.3%↑) and slightly worse than the rest of the networks in YOLOv5 and YOLOv8 families. In terms of speed, the proposed network is faster than the large-size models (YOLOv5l (2.8 ms↑), YOLOv5x (12.9 ms↑)) but it is slower than small-size models (YOLOv5n (8.3 ms↓), YOLOv5s (6.4 ms↓), and YOLOv8n (4.7 ms↓)). Besides, its speed can compare with other model sizes like YOLOv5m and YOLOv8s in acceptable network parameters and computational complexity (GFLOPs). It has promising for application in real-time tracking and surveillance systems. Table II shows the detailed detection results for each class on the Human-Parts validation set. Fig. 4 presents several qualitative results on the Human-Parts dataset with different scenes, human poses, head poses, and hand poses. The comparison results between the proposed network and YOLOv5m in Fig. 5 prove that the proposed network is better than the YOLOv5m network architecture when detecting the hands and faces. For person detection ability, both are the same. From this comparison, it is obvious to see the balance in the person, face, and hand detection of the proposed network with acceptable computation complexity and network parameters. This allows the proposed model can be applied in real-time Human-Robot interaction or human tracking applications. However, this model is still affected by several factors that reduce the detection ability such as overlap, dense, and tiny objects.

D. Ablation studies

To evaluate the effectiveness of each proposed block in the overall network, this experiment conducts several ablation studies. Each block is replaced individually to build the new model and then trained and evaluated on the Human-Parts dataset. Table IV describes the obtained results. The comparison results demonstrate that the Focus block in YOLOv5 increases the network parameter but the mAP@0.5 and mAP0.5:0.95 reduce by 0.3% and 0.5%, respectively. Using EfficientNet and CBAM blocks had the same effect with mAP@0.5, however, mAP@0.5:0.95 large drops (18.7%↓) when the CBAM block is ignored. Adding a detector can increase the network parameters by two times but increase both mAP measurements. Finally, when comparing SPP and

TABLE III
THE COMPARISON RESULT BETWEEN THE PROPOSED NETWORK AND OTHER NETWORKS ON THE HUMAN-PARTS VALIDATION SET.

Models	Parameter	Weight (MB)	GFLOPs	mAP@0.5	mAP@0.5:0.95	Inf. time (ms)
YOLOv5x	86,186,872	173.1	204	87.2	60.9	22.4
YOLOv5l	46,119,048	92.8	107.8	87.3	60.3	12.3
YOLOv5m	20,861,016	42.2	48.0	86.9	59.3	6.7
YOLOv5s	7,018,216	14.4	15.8	86.7	57.3	3.1
YOLOv5n	1,767,976	3.8	4.2	85.4	53.7	1.2
YOLOv8n	3,006233	6.2	8.1	86.3	58.4	4.8
YOLOv8s	11,126,745	22.5	28.4	87.7	60.7	8.8
Our	**25,655,818**	**52.3**	**82.3**	**87.2**	**58.6**	**9.5**

Inf. time (ms): Inference time is evaluated on a GeForce GTX 1080Ti GPU.

Fig. 4. The qualitative result on Human-Parts validation set.

SPPF blocks, SPPF is better at mAP@0.5:0.95 (0.3%↑). From the above analysis, this research selects the last architecture to design the human, face, and hand detector. Therefore, this work selects the last architecture to train, evaluate, and report on vehicle detection capabilities.

V. CONCLUSION AND FUTURE WORK

This paper improves the YOLOv5 architecture network for simultaneous human, face, and hand detection. Research concerns redesigning backbone and neck modules using a combination of lightweight EfficientNet architecture and CBAM attention mechanism. On the other hand, this work refines

TABLE IV
ABLATION STUDIES WITH THE DIFFERENT MODULES IN THE PROPOSED NETWORK ON THE HUMAN-PARTS VALIDATION SET.

Modules	Networks					
CBR	Focus	✓	✓	✓	✓	✓
EfficientNet	✓		✓	✓	✓	✓
SPPF	✓	✓	✓	✓		✓
SPP					✓	
CBAM	✓	✓		✓	✓	✓
P2	✓	✓	✓		✓	✓
Parameter	25,659,658	22,117,210	22,113,370	10,904,160	25,655,818	25,655,818
Weight (MB)	87.5	45.1	45.1	22.3	52.3	52.3
GFLOPs	83.0	73.6	72.8	21.2	82.3	82.3
mAP@0.5	86.9	86.6	86.8	84.8	87.2	87.2
mAP@0.5:0.95	58.1	58.0	39.3	54.5	58.3	58.6

Fig. 5. The comparison result between proposed method and YOLOv5m on Human-Parts validation set.à

and distillates the images with the correct annotation for the YOLO architecture training task. With the network parameters optimization, computational complexity, inference speed, and small-size object detection ability, the proposed network has the potential to be deployed on low-computing devices. The detector will be developed with other novel attention mechanisms to boost small-size object detection in the future.

ACKNOWLEDGEMENT

This result was supported by the "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003).

REFERENCES

[1] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, L. Yuan, and Y.-G. Jiang, "Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning," 2023.

[2] J. Wang, R. Cheng, M. Liu, and P.-C. Liao, "Research trends of human–computer interaction studies in construction hazard recognition: A bibliometric review," *Sensors*, vol. 21, no. 18, 2021.

[3] L. C. Chang, S. Pare, M. S. Meena, D. Jain, D. L. Li, A. Saxena, M. Prasad, and C. T. Lin, "An intelligent automatic human detection and tracking system based on weighted resampling particle filtering," *Big Data and Cognitive Computing*, vol. 4, no. 4, 2020.

[4] K. Zhao, W. Zhang, and Y. Jiang, "Semantic interactions in multi-level objects segmentation," in *2010 International Conference on Computational and Information Sciences*, pp. 665–668, 2010.

[5] F. Takarli, A. Aghagolzadeh, and H. Seyedarabi, "Robust pedestrian detection using low level and high level features," in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, pp. 1–6, 2013.

[6] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *2013 IEEE International Conference on Computer Vision*, pp. 2056–2063, 2013.

[7] S. Zhang, R. Benenson, M. Omran, J. H. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?," *CoRR*, vol. abs/1602.01237, 2016.

[8] J. Li, L. Liu, J. Li, J. Feng, S. Yan, and T. Sim, "Toward a comprehensive face detector in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 104–114, 2019.

[9] X. Li, Z. Yang, and H. Wu, "Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks," *IEEE Access*, vol. 8, pp. 174922–174930, 2020.

[10] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Faceboxes: A CPU real-time face detector with high accuracy," *CoRR*, vol. abs/1708.05234, 2017.

[11] Y. Ge, Q. Wang, B. Sheng, and W. Yang, "Flashnet: A real-time anchor-free face detector," in *2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 441–446, 2020.

[12] M. D. Putro, D.-L. Nguyen, and K.-H. Jo, "An efficient face detector on a cpu using dual-camera sensors for intelligent surveillance systems," *IEEE Sensors Journal*, vol. 22, no. 1, pp. 565–574, 2022.

[13] X. Deng, Y. Zhang, S. Yang, P. Tan, L. Chang, Y. Yuan, and H. Wang, "Joint hand detection and rotation estimation using cnn," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1888–1900, 2018.

[14] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides, "Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 46–53, 2016.

[15] K. Roy, A. Mohanty, and R. R. Sahay, "Deep learning based hand detection in cluttered environment using skin segmentation," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 640–649, 2017.

[16] D.-L. Nguyen, M. D. Putro, X.-T. Vo, T.-D. Tran, and K.-H. Jo, "Robust hand detection based on convolutional neural network and attention

module," in *2022 International Workshop on Intelligent Systems (IWIS)*, pp. 1–6, 2022.
[17] X. Li, L. Yang, Q. Song, and F. Zhou, "Detector-in-detector: Multi-level analysis for human-parts," *CoRR*, vol. abs/1902.07017, 2019.
[18] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019.
[19] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *CoRR*, vol. abs/1803.01534, 2018.
[20] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *CoRR*, vol. abs/1807.06521, 2018.
[21] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *CoRR*, vol. abs/2005.03572, 2020.
[22] R. Rubinstein and D. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Information Science and Statistics, Springer New York, 2011.