IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 21, 2021, accepted August 23, 2021, date of publication August 30, 2021, date of current version September 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108776

CNN-Based Mask-Pose Fusion for Detecting Specific Persons on Heterogeneous Embedded Systems

JEONGJUN LEE[1], (Graduate Student Member, IEEE), JIHOON JANG[1], (Graduate Student Member, IEEE), JINHONG LEE[1], DAYOUNG CHUN[2], (Graduate Student Member, IEEE), AND HYUN KIM[1], (Member, IEEE)

[1]Research Center for Electrical and Information Technology, Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea
[2]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Hyun Kim (hyunkim@seoultech.ac.kr)

This work was supported by the Research Program through Seoul National University of Science and Technology (SeoulTech).

ABSTRACT In recent times, numerous convolutional neural network (CNN) based detection models have been proposed and have shown excellent performance. However, because these models are generally developed to detect objects in class units (e.g., person, car), additional training processes with numerous datasets are required to find a specific object. This paper proposes a model that accurately detects specific persons by using top clothing color information without any additional training processes. The proposed method combines CNN-based instance segmentation and pose estimation, utilizing all the advantages of each technique. To avoid redundant computations, these two schemes are implemented as a filtering-based sequential operation structure. As a result, the proposed method has a 92.57% of accuracy in detecting a specific person with only a slight processing speed decrease. Furthermore, in this paper, the proposed model is efficiently ported on the heterogeneous embedded platform (i.e., NVIDIA Jetson AGX Xavier) with a parallel processing technique to maximize the hardware utilization.

INDEX TERMS AlphaPose, deep learning, embedded systems, instance segmentation, NVIDIA Jetson AGX Xavier, object detection, pose estimation, YOLACT.

I. INTRODUCTION

With the development of hardware accelerators like graphics processing units (GPUs), deep learning (DL) has become pwidely used in various computer vision (CV) tasks, such as image classification [1]–[3], object detection [4]–[7], segmentation [8]–[12], and pose estimation [13]–[17], and has shown remarkable performance. In particular, high-performance models that can accurately detect people in images have been actively proposed. However, practical applications require not only person class detection but also the identification of specific individuals such as missing persons or criminals [18], [19]. Notably, the number of reports of missing children is increasing every year [20], [21]. Nevertheless, the detection of a specific person in an image has not yet been fully automated, and manually checking images by mobilizing a large amount of manpower causes a considerable waste of labor.

To detect specific individuals, several prior studies have focused on DL-based facial recognition schemes [18], [22]–[24]; however, in practical environments, it may be necessary to identify specific persons in CCTV images. In such cases, there are significant limitations to performing accurate facial recognition, such as resolution and noise problems [25]. Moreover, these existing approaches require an additional training process using a large amount of specific dataset containing the characteristics of the person to be found, and consequently suffer from significantly low scalability and compatibility. In addition, although there is a considerable need for such specific person detection to be operated on embedded platforms in the future [26], existing studies have not considered operating on embedded applications.

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Shen.

To compensate for these limitations of the existing studies, this paper proposes a model that accurately detects specific persons in images based on the color of their upper-body garments. It should be noted that searching for a specific person using such color information is a popular approach. The contributions of this study can be summarized as follows: First, a mask-pose fusion model that combines instance segmentation and pose estimation is proposed for the accurate localization of the upper body. While it is possible to recognize the upper body of a standing person using segmentation schemes alone, this cannot be done for a person in a sitting or sleeping position. Therefore, in the proposed method, the upper body is detected by further applying pose estimation to the masking result obtained using segmentation. In particular, to eliminate unnecessary computations for pose estimation, we selectively apply the pose estimation by filtering the target candidates based on the segmentation results. Second, we propose a technique to maximize the utilization of hardware while porting the proposed mask-pose fusion model to heterogeneous embedded platforms. If the inference processes of both the instance segmentation and pose estimation networks are performed on a single GPU, increased latency is inevitably observed as compared to that of the inference process of a single network owing to the serial configuration of the two networks. To minimize this additional latency, both the GPU and deep learning accelerator (DLA) in the NVIDIA Jetson Xavier, a heterogeneous embedded platform, are utilized. In detail, instance segmentation and pose estimation are allocated to each device in parallel and are processed in a pipelined structure, thereby increasing the hardware utilization and avoiding the additional latency. Experimental results show that the proposed mask-pose fusion model (i.e., the first contribution) achieves an accuracy improvement of about 11.8% and 3.5% as compared to individual standalone segmentation and pose estimation models, respectively. Notably, in the detection of persons who are in sitting and sleeping positions, the combined model achieves an accuracy improvement of more than 42% compared to the standalone instance segmentation model. Moreover, the proposed parallel processing technique (i.e., the second contribution) achieves a speed improvement of 87.9% as compared to the serial processing of both networks.

The remainder of this paper is organized as follows: Section II presents the background of DL-based CV algorithms and embedded GPU platforms. Section III describes the structure of the proposed mask-pose fusion model in detail. Section IV provides evidence that the performance of the proposed model is superior to that of existing models. Finally, Section V concludes the paper.

II. BACKGROUND

A. DL-BASED CV ALGORITHMS FOR SPECIFIC PERSON DETECTION

In recent times, there has been a significant amount of research on various CV tasks, including object detection, semantic segmentation, instance segmentation, and pose estimation. All these approaches can be used for specific person detection. Their characteristics are as follows:

YOLOv3 [5], a representative model of object detection, predicts classes using binary cross-entropy loss, and creates anchor boxes through clustering to detect bounding boxes. However, it is not suitable for use as a color discrimination model for the upper body because it is not possible to determine the exact position of a person's upper body using the bounding box alone.

Semantic segmentation is a technique that creates a mask for each pixel by separating the object to be searched from the background. Deeplabv3+ [12] is a representative model of semantic segmentation, which configures a module into an encoder-decoder structure with intermediate connections which are similar to U-Net [27], and performs segmentation using atrous separable convolution. However, in semantic segmentation, because the masking results are output for each class rather than for each object, it is impossible to classify the two objects when the objects overlap. Therefore, it is not suitable for use as an upper-body garment color discrimination model.

Instance segmentation models simultaneously process both localization and segmentation tasks. Similar to semantic segmentation models, they can process images in pixel units. They can also solve the problem of overlapping objects, which is a problem in semantic segmentation, because a mask is created for each object rather than for each class. Mask-R-CNN [11] and YOLACT++ [10] are representative instance segmentation models. Mask-R-CNN [11] is a representative two-stage instance segmentation model which generates regions of interest (ROI) in the first stage and uses these ROIs in the second stage to classify the objects and perform the segmentation process. Although the two-stage instance segmentation method exhibits relatively high accuracy, real-time processing is difficult because the additional process of re-pooling the features for each ROI reduces the processing speed. On the other hand, YOLACT ++ [10] is a representative one-stage instance segmentation model which has an advantage that real-time operation is possible with acceptable accuracy degradation as compared to the two-stage method. This is achieved by implementing two subtasks, Protonet and prediction head, in parallel, without a localization step. Owing to these characteristics, one-stage methods are widely used in applications which are implemented on embedded platforms.

However, although these instance segmentation techniques can also accurately detect a person, it is difficult to predict the exact upper body position of the detected person. Using instance segmentation, it is possible to identify the upper body of a person in an upright position, however, it is impossible to do so for a person in a non-upright position. In other words, upper body prediction through instance segmentation has a problem in that the detection accuracy varies significantly according to the person's pose.

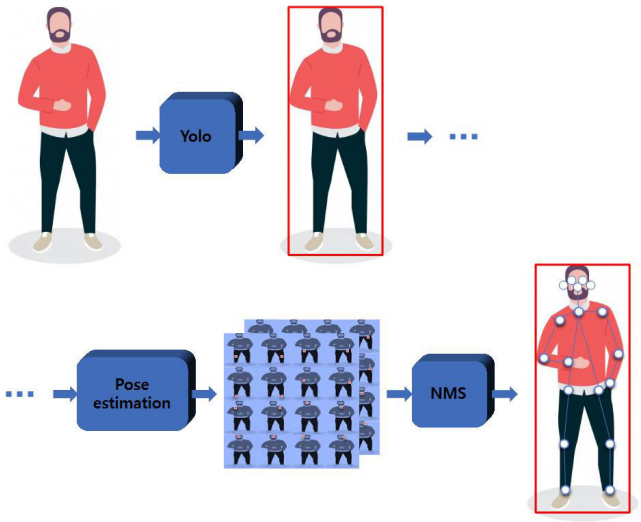


FIGURE 1. Pose estimation process of AlphaPose [17].

Pose estimation is a technique for estimating the position of a person's joints and feature points such as the head and shoulder. AlphaPose [17] is a representative pose estimation model which performs object detection on a person using YOLO [5], and then generates a heatmap using the pose estimation weight based on the detection result. Finally, as illustrated in Fig. 1, AlphaPose extracts 17 characteristic points from the body, such as the head, shoulders, and pelvis, using heatmap information. Because these pose estimation techniques search for joint coordinates, they have the advantage of being able to find the exact upper body position regardless of pose, using shoulder and pelvic coordinates. However, as the number of detected objects increases, the joint extraction process also increases linearly, leading to a decrease in the processing speed. This, in turn, makes real-time operation impossible in an embedded environment. In addition, pose estimation utilizes a box-based processing method rather than a pixel unit-based method, which prevents accurate detection of the upper body comparable to the instance segmentation model.

This paper proposes a mask-pose fusion model that combines the advantages of instance segmentation and pose estimation to achieve accurate upper body detection. A detailed description of the mask-pose fusion model is provided in the next section.

B. EMBEDDED PLATFORM: JETSON AGX XAVIER

The NVIDIA Jetson Xavier embedded platform is a heterogeneous system which contains a power-efficient dedicated DLA, CPU, and GPU. The DLA is a computational device with 2.5 tera floating point operations per second (TFLOPS) of computational performance in FP16 (16-bit floating point) format. Its performance is approximately 4.4 times lower than that of the Volta GPU in the Jetson Xavier, which has 11 TFLOPS of computational performance, but it is expected to be widely used for DL inference in the future owing to its high power efficiency. However, although the DLA with the limited whose computational performance can be utilized for DL inference of simple networks, it is difficult to support the real-time DL inference of practical networks (i.e., relatively deep networks) alone. Therefore, it is important to develop a method to efficiently utilize the DLA in combination with GPUs by taking advantage of the fact that the DLA can perform operations which are independent of a GPU. In this study, the mask-pose fusion model, which combines the advantages of instance segmentation and pose estimation, is accelerated by porting the model to the heterogeneous embedded board, NVIDIA Jetson Xavier, with maximized hardware utilization as discussed in the next section.

III. PROPOSED METHOD: CNN-BASED MASK-POSE FUSION

A. NETWORK STRUCTURE

This paper proposes a mask-pose fusion model that combines the representative instance segmentation model, YOLACT++ [10], and the representative pose estimation model, AlphaPose [17], to identify a specific person in real time using the precise position of the upper body. As mentioned in the previous section, YOLACT++ has a high processing speed and relatively high accuracy, however, its detection accuracy is heavily dependent on the person's pose. To solve this problem, the proposed network structure additionally applies AlphaPose, which can search for the exact upper body position regardless of pose, to the result of YOLACT++ masking. Accordingly, the masked result corresponding to the accurate upper-body position can be obtained based on the joint information. Notably, as both YOLACT++ and AlphaPose predict the masking and pose results, respectively, using box detection; it is possible to avoid redundant computations by sharing the box detection results from YOLACT++.

In addition, because most pose estimation methods, including AlphaPose, have the drawback of low algorithm execution speeds, the proposed structure adds a filtering process between YOLACT++ and AlphaPose. Essentially, in the proposed method, by filtering the candidates for pose estimation based on the masking result of YOLACT++, the pose estimation process is not performed on all the individuals in the image, which significantly mitigates the slowdown caused by pose estimation. Following the accurate and immediate detection of the upper body position, it is verified that the pixel values corresponding to the masked upper body results match the target color. For accurate color detection, HSV, which represents colors based on hue, saturation, and brightness, is used in order to express colors more clearly than the red, green, blue (RGB) color expression method. The HSV range of the six most popular upper-body garment colors (i.e., red, yellow, green, blue, black, and white) are constructed through repetitive manual processes. It should be noted that this process is very important for detecting a specific color, as the HSV range for each color has a significant impact on accuracy.

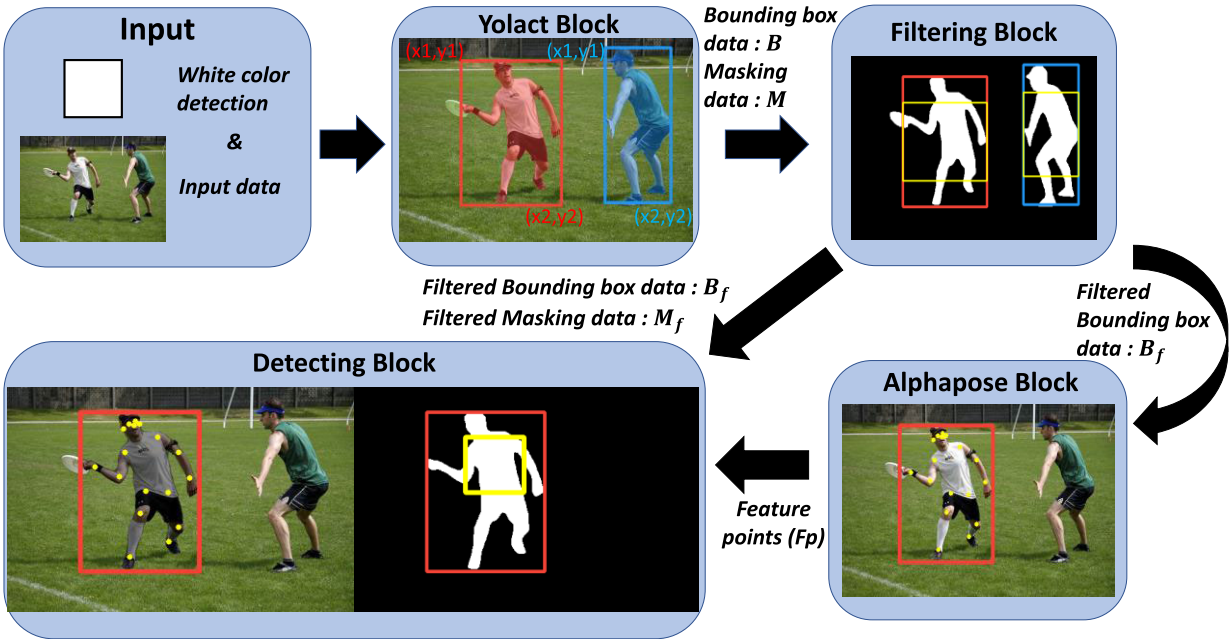


FIGURE 2. Overview of the mask-pose fusion model.

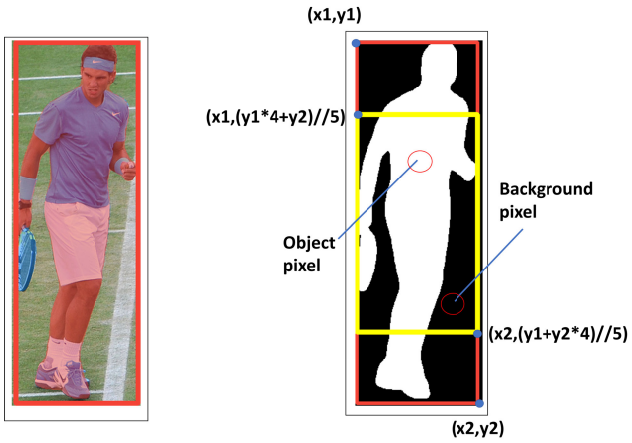


FIGURE 3. Example image of the proposed filtering method.

Fig. 2 presents the overall structure of the proposed model. It can be observed that when an image is given as an input to YOLACT++, the bounding box information, B, with N bounding box coordinates (i.e., (x1, y1) and (x2, y2)) and the masking information, M, with N pixel information (masking data) are generated. Subsequently, as presented in Fig. 3, a filtering process based on B and M is performed to eliminate redundant pose estimation computations. It should be noted that this filtering process can significantly alleviate the problem of AlphaPose, whose speed reduces linearly with an increase in the number of detected objects. In the filtering process, a filtering box is generated with the following coordinates for each bounding box of the detected object:

$$(x1, \frac{y1 * 4 + y2}{5}), \quad (x2, \frac{y1 + y2 * 4}{5}). \tag{1}$$

All pixels in the filtering box are divided into object and background pixels using the object's masking information, and filtering is performed using the HSV data of the object pixels, based on the following equation:

$$\frac{Dcp}{Op} < \alpha, \tag{2}$$

where Op is the number of pixels of the object in the box, and Dcp is the number of color pixels being searched for in the box. That is, objects whose proportion of the desired color pixels is lower than the predefined threshold, α, are filtered out. The performance and processing speed depend on the value of α, and in this study, the value of α is empirically set to 0.1. Following the filtering process, AlphaPose is performed only for the bounding box information of the filtered candidate group (B_f in Fig. 2).

As a result of pose estimation for the filtering blocks (i.e., result of AlphaPose blocks), 17 joint positions, called feature points (Fp), are obtained, including the head, shoulders, and hips of the detected person, along with filtered bounding box data, B_f, and filtered masking data, M_f. Based on this information, the upper-body position is detected, and the color of the upper-body garment is searched. Among these Fp, the information related to the upper body is the shoulders, pelvis, and elbows. Specifically, an upper body position box is created using a total of six coordinates consisting of two shoulder coordinates, two pelvic coordinates, and two elbow coordinates among 17 pieces of Fp information. Fig. 4 presents the results of detecting the upper body in various positions, including standing, sitting, and lying down positions. As illustrated in the figure, the shoulder and pelvic coordinates represent the upper and lower boundaries of the

FIGURE 4. Result of upper-body detection in various postures.

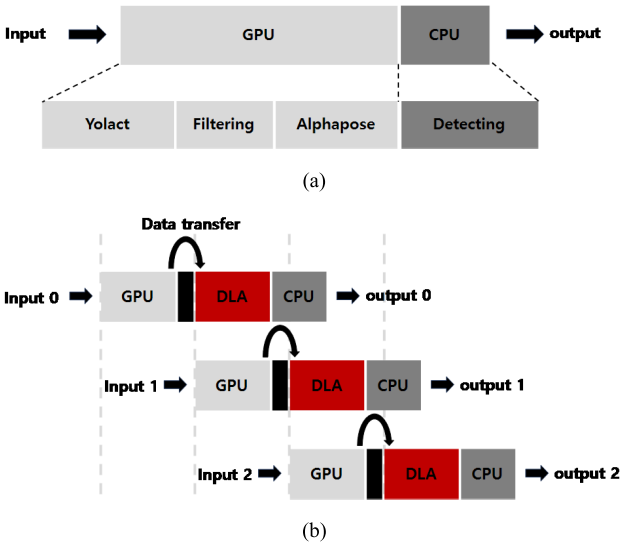


FIGURE 5. Time diagram of the proposed model in the heterogeneous platform. (a) Serial process using the GPU alone. (b) Pipelined structure using both the GPU and DLA in parallel. In both cases, the detecting task is post-processed on the CPU.

upper body, respectively. For the left and right boundaries, elbow or wrist coordinates can be used; however, when using wrist coordinates, the detection accuracy is lowered if the detected person is wearing a short-sleeved or sleeveless garment, and thus, the upper body position is estimated using the relatively stable elbow coordinates. Subsequently, as in the previously described filtering process, each upper body position box is classified into object pixels and background pixels based on M_f, and the ratio of the detecting color pixels (Dcp') to the object pixels (Op') is calculated with the HSV information. Finally, as presented in the following equation, the object is output as a detection target only when this ratio is greater than the predefined threshold, β, which is experimentally determined as 0.3.

$$\frac{Dcp'}{Op'} > \beta \tag{3}$$

B. NETWORK PORTING ON HETEROGENEOUS EMBEDDED PLATFORMS

To alleviate the increase in computation that occurs while merging two models (i.e., YOLACT++ and AlphaPose), this paper presents a technique for maximizing the hardware utilization of heterogeneous platforms that have been actively used in recent years. Specifically, the hardware structure is designed to parallelize the operation of each model in the GPU and DLA in the NVIDIA Jetson Xavier, and consequently minimize the processing speed degradation by hiding the latency caused by using both YOLACT++ and AlphaPose. The proposed mask-pose fusion model shares the bounding box output of YOLACT++ in both pose estimation and instance segmentation processes and includes the process of filtering pose estimation candidates based on the masking result of YOLACT++. Therefore, as illustrated in Fig. 5(a), two models must be processed serially (i.e., AlphaPose can only be executed after YOLACT++ has been executed). To improve this inefficient structure, the hardware structure is designed by additionally utilizing the DLA such that YOLACT++ is processed by the GPU and the AlphaPose is processed by the DLA, as presented in Fig. 5(b). It is difficult to improve the processing speed for a single frame in the proposed structure. However, when processing multiple frames, it is possible to process two devices in parallel with a pipelined structure to hide the time required for pose estimation in DLA, which maximizes the hardware utilization and improves the processing speed. Essentially, input_0 must be serially processed through the GPU and DLA, however, when the AlphaPose operation of input_0 is processed in the DLA, the GPU processes the YOLACT++ operation of input_1 in parallel, so that the GPU and DLA can be fully utilized.

To completely hide the AlphaPose operation with the proposed parallelization method, the inference time of the AlphaPose operation must be less than that of the YOLACT++ operation. However, YOLACT++ has a constant inference time owing to its frame unit processing structure, whereas AlphaPose has a variable inference time depending on the number of objects in a frame owing to its single-person pose estimation structure. Therefore, if the number of individuals detected in one frame exceeds a certain threshold, the computation time of the AlphaPose operation may be longer than that of the YOLACT++ operation, resulting in unnecessary idle time. Specifically, as presented in Fig. 5(b), if the DLA block of input_0 is longer than the GPU block of input_1, parallel processing of the DLA block of input_1 is impossible, which leads to a delay in the GPU block of input_2. To alleviate this problem, the method proposed herein uses

TensorRT [28] to reduce the computing time of AlphaPose and applies quantization from FP32 to FP16. As a result of this optimization, if B_f does not exceed eight people, the latency caused by the AlphaPose operation on the DLA can be completely hidden.

It should be noted that when parallelizing processing in a heterogeneous platform, data transfer latency occurs according to device switching. In other words, a device switching process is required for data transfer to process the AlphaPose operation in the DLA following the completion of the YOLACT++ operation in the GPU. Although GPU and DLA parallelism can be further subdivided, frequent device switching can create a bottleneck in the overall processing time. To prevent frequent device switching, as presented in Fig. 5(b), in the proposed structure, the filtering process following YOLACT++ is also processed by the GPU, and the data transfer time is minimized by allowing only one device switching process (i.e., the process of switching from the filtering operation to the AlphaPose operation). The detection task is processed by the CPU in consideration of the characteristics of its computation, and the CPU operation is also processed in parallel with the GPU and DLA to prevent an increase in latency.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL ENVIRONMENTS

To verify the performance of the proposed design, the accuracy and processing speed are evaluated on an RTX-2080 GPU with the COCO pre-trained weights of YOLACT++ and AlphaPose. To verify the accuracy of the upper-body garment color detection, a test set consisting of 313 images is created by extracting images containing people with various gestures from the COCO [29] and MPII [13] datasets, because colors are not labeled in these datasets. We directly labeled samples of six colors (i.e., white, red, blue, yellow, green, and black) in the test sets, and the accuracy is calculated according to the following formula, which is widely used in existing DL research:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}), \quad (4)$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.

B. PERFORMANCE EVALUATION

First, to demonstrate the need for a filtering process, the processing speed was compared in frames per second (FPS) based on the application of the filtering process, as presented in Table 1. Because the speed of the proposed model decreases linearly as the number of pose estimation processes increases, the reduction of redundant computations through the filtering process increases the processing speed from 10 FPS to 24 FPS, indicating that the filtering process is essential.

Tables 2 and 3 compare the detection accuracy and speed of the proposed method and existing standalone methods (i.e., YOLACT++ [10] and AlphaPose [17]), and the detection accuracy for people with irregular postures. The proposed model with the filtering technique demonstrates a detection accuracy of 92.57% and a processing speed of 24 FPS. Although YOLACT++ has the highest processing speed of 30 FPS, it shows a decrease in the detection accuracy by approximately 11.8% as compared to the proposed method, as presented in Table 2; in particular, the detection accuracy of people in non-standing positions is significantly low, as presented in Table 3. AlphaPose shows a decrease in the detection accuracy by approximately 3.5% compared to the proposed method, and has the slowest lowest processing speed of 13 FPS. In AlphaPose, the upper-body position can be accurately estimated regardless of the person's posture, however, because it processes images in box-based units, its performance is inferior to that of the proposed model, which processes images in units of pixels. In addition, because detection is performed using both shoulder and hip coordinates, it is difficult to detect individuals wearing outer garments such as jackets or coats.

TABLE 1. Speed (in FPS) according to the use of filtering process.

	Proposed model w/o filtering	Proposed model w/ filtering
FPS	10	24

TABLE 2. Detection accuracy and speed of the proposed method and existing standalone methods.

Model	Accuracy (%)	Speed (FPS)
Proposed	**92.57**	24
YOLACT++[10]	80.77	30
AlphaPose[17]	89.06	13

TABLE 3. Detection accuracy of specific persons with irregular postures.

	Proposed	YOLACT++ [10]	AlphaPose [17]
Accuracy	**87.48%**	45.20%	85.68%

Table 4 presents the detection accuracy for each upper-body garment color (i.e., white, red, blue, yellow, green, and black) using the three methods. It can be seen that the proposed model exhibits the highest detection accuracy regardless of the color. All three methods have low detection accuracy for black and white garments (colors that occupy a significant portion of the background), but the proposed method achieves relatively high accuracy in this case as well.

Table 5 presents the normalized processing speed results of performing specific person detection on the NVIDIA Jetson Xavier board. When YOLACT++ and AlphaPose are serially processed, the processing speed is approximately 60% lower than that of processing speed of YOLACT++ alone. However, applying the proposed parallelization and pipelining technique improves the processing speed by

FIGURE 6. Visual evaluation on the same image. (a) To detect red top, (b) To detect blue top, and (c) To detect white top.

FIGURE 7. Visual evaluation of various images from the COCO dataset. (a) Blue garment detection, (b) White garment detection, (c) Yellow garment detection, and (d) Red garment detection.

TABLE 4. Detection accuracy for each color in the proposed method and existing standalone methods.

Color	Proposed (%)	YOLACT++ [10] (%)	AlphaPose [17] (%)
White	92.74	81.59	87.59
Red	96.01	82.30	92.44
Blue	92.00	80.32	87.96
Yellow	94.60	82.94	92.20
Green	96.02	83.87	94.47
Black	84.05	73.65	79.75

TABLE 5. Normalized results of the processing speed on the NVIDIA Jetson Xavier board.

	YOLACT++	AlphaPose + YOLACT++ (Serial)	AlphaPose + YOLACT++ (Parallel)
Normalized Speed	1	0.396	0.745

approximately 87.9% over the serial structure. As a result, the proposed method can significantly improve the accuracy with only a slight speed degradation as compared to the processing speed of YOLACT++ alone.

C. VISUAL EVALUATION

Figs. 6 and 7 present the visual evaluation of the proposed model. Figs. 6(a), 6(b), and 6(c) present the results of detecting a person wearing a red, blue, and white upper-body garment, respectively, from a single image depicting five people in a seated position. Figs. 7(a), 7(b), 7(c), and 7(d) present the results of detecting a person wearing red, blue, white, and yellow upper-body garments, respectively, on various images from the COCO dataset. These results indicate that even if there are several people in the image, the proposed mask-pose fusion model can accurately detect a specific individual based on the color of their upper-body garment.

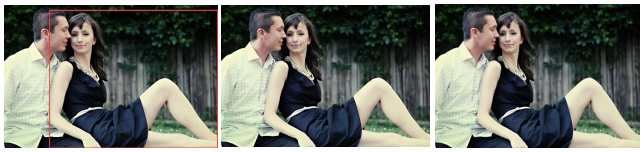
Fig. 8 presents the comparative result of visual evaluation for detecting a person with an irregular posture using three models: the proposed mask-pose fusion model, YOLACT++ [10], and AlphaPose [17]. It can be seen that YOLACT++ cannot detect an individual in a sleeping position, whereas the mask-pose fusion model and AlphaPose can detect such irregular postures. Fig. 9 presents the comparative results of the visual evaluation for detecting a person with irregular garment colors (i.e., a black jacket over white clothes). In this case, when white is set as the target color, all three models can successfully detect the person, however, when black is set as the target, AlphaPose fails to detect the person. This is because AlphaPose creates the

(a) (b) (c)

FIGURE 8. Comparison result of visual evaluation for detecting a person wearing a red top using three models. (a)Proposed. (b) YOLACT++. (c) AlphaPose.

(a) (b) (c)

FIGURE 9. Comparative result of visual evaluation for detecting a person wearing a black upper-body garment using three models. (a) Proposed model; (b) YOLACT++; (c) AlphaPose.



(a) (b) (c)

FIGURE 10. Comparative result of visual evaluation for detecting a person wearing a black upper body using three models. (a) Proposed model; (b) YOLACT++; (c) AlphaPose.

upper body position box using the shoulder and pelvis coordinates, resulting on the focus being on the individual's body. To alleviate this problem, the upper-body position box can be widened. However, in this case, there is a problem in that it is not possible to clearly distinguish between the object and the background. Fig. 10 presents the comparative result of the visual evaluation for detecting a person with irregular posture and color conditions. YOLACT++ fails to detect a person wearing a black upper-body garment, because it cannot accurately estimate the position of the upper body owing to the irregular posture. AlphaPose also fails to detect a person wearing a black top because it cannot accurately distinguish between the object and the background owing to the use of box-based units rather than pixels in the processing method. On the other hand, the proposed model, which has the advantages of both YOLACT++ and AlphaPose, succes sfully detects the target, as illustrated in the figure.

V. CONCLUSION

In previous studies, several high-performing person detection models have been proposed. However, these models have always required additional training processes using a specific dataset for each person who needs to be detected in a real-world situation. To address this problem, this study presents a mask-pose fusion model that can detect a specific person based on the color of their upper-body garments. By combining the instance segmentation and pose estimation models, the proposed model can detect a specific person with an accuracy of 92.57%, which is significantly higher than that of the existing standalone models. In addition, the proposed design minimizes redundant computations through the filtering process. It also maximizes hardware utilization by porting the proposed model on an embedded platform based on parallel processing, thereby mitigating the decrease in processing speed by 87.9%. There are several real-world scenarios in which it is necessary to find a specific person, such as when searching for a missing person. If the proposed model is implemented in practical situations, it is expected to significantly improve efficiency as it can accurately detect a specific person while eliminating unnecessary costs.

ACKNOWLEDGMENT

(Jeong Jun Lee and Ji Hoon Jang are co-first authors.)

REFERENCES

[1] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Med. Image Anal.*, vol. 54, pp. 10–19, May 2019.

[2] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, Swinoujście, Poland, May 2018, pp. 117–122.

[3] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621

[4] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015.

[5] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[6] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 502–511.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

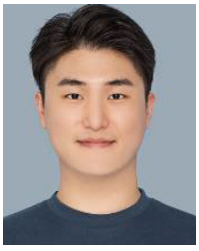
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 346–361.

[9] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9157–9166.

[10] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++: Better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 5, 2020, doi: 10.1109/TPAMI.2020.3014297.

[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[12] L. Cheih, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[13] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.
[14] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 468–475.
[15] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.
[16] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1347–1355.
[17] N. S. Abraham, R. A. Rajan, R. E. George, S. Gopinath, and V. Jeyakrishnan, "Finding missing child in shopping mall using deep learning," in *Advances in Smart System Technologies*, vol. 1163. Singapore: Springer, 2021, pp. 477–482.
[18] Y. Rao, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition and person re-identification," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 701–718, Jun. 2019.
[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[20] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.
[21] P. S. Chandran, N. B. Byju, R. U. Deepak, K. N. Nishakumari, P. Devanand, and P. M. Sasi, "Missing child identification system using deep learning and multiclass SVM," in *Proc. IEEE Recent Adv. Intell. Comput. Syst. (RAICS)*, Dec. 2018, pp. 113–116.
[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[23] S. Il Lee and H. Kim, "Instant and accurate instance segmentation equipped with path aggregation and attention gate," in *Proc. Int. SoC Design Conf. (ISOCC)*, Oct. 2020, pp. 320–321.
[24] A. Eden, C. M. Christoudias, and T. Darrell, "Finding lost children," in *Proc. IEEE Workshop Person-Oriented Vis.*, Jan. 2011, pp. 7–12.
[25] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8350–8360.
[26] M. G. West. (2018). *Pooling Resources to Fight Child Abuse and Abduction*. [Online]. Available: https://on.wsj.com/32R9Vgz
[27] NCIC. (2018). *Missing Person and Unidentified Person Statistics*. [Online]. Available: https://www.fbi.gov/file-repository/2018-ncic-missing-perso nandunidentified-person-statistics.pdf/view
[28] H. Kim, C. E. Rhee, and H. J. Lee, "A low-power video recording system with multiple operation modes for H.264 and light-weight compression," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 603–613, Apr. 2016.
[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
[30] J. Choi, D. Chun, H.-J. Lee, and H. Kim, "Uncertainty-based object detector for autonomous driving embedded platforms," in *Proc. 2nd IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Aug. 2020, pp. 16–20.



JEONGJUN LEE (Graduate Student Member, IEEE) received the B.S. degree in electrical and information engineering from Seoul National University of Science and Technology, Seoul, South Korea, in 2021, where he is currently pursuing the M.S. degree in electrical and information engineering. His research interests include algorithms and architectures of computer vision and deep learning.

JIHOON JANG (Graduate Student Member, IEEE) received the B.S. degree in electrical and information engineering from Seoul National University of Science and Technology, Seoul, South Korea, in 2021, where he is currently pursuing the M.S. degree in electrical and information engineering. His research interests include algorithms and architectures of deep learning and memory architecture.

JINHONG LEE received the B.S. degree in electrical and information engineering from Seoul National University of Science and Technology, Seoul, South Korea, in 2021. His research interests include image processing and computer vision.

DAYOUNG CHUN (Graduate Student Member, IEEE) received the B.S. degree in electronics engineering from Sogang University, Seoul, South Korea, in 2018. She is currently pursuing the integrated M.S. and Ph.D. degree in electrical and computer engineering with Seoul National University, Seoul. Her research interests include algorithms and architectures of deep learning and GPU architecture for computer vision.

HYUN KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 2009, 2011, and 2015, respectively. From 2015 to 2018, he was with the BK21 Creative Research Engineer Development for IT, Seoul National University, as a BK Assistant Professor. In 2018, he joined the Department of Electrical and Information Engineering, Seoul National University of Science and Technology, where he is currently working as an Assistant Professor. His research interests include algorithms, computer architecture, memory, and SoC design for low-complexity multimedia applications and deep neural networks.