https://doi.org/10.2352/EI.2023.35.9.IPAS-297

AInBody: Are You in Shape?- An Integrated Deep Learning Model that Tracks Your Body Measurement

Nakyung Lee, Youngsun Cho, Minseong Son, Sungkeun Kwak, Jihwan Woo; CJ OliveNetworks AI Research; Seoul, Republic of Korea

Abstract

This paper presents AInBody, a novel deep learning-based body shape measurement solution. We have devised a user-centered design that automatically tracks the progress of the body by adequately integrating various methods, including instance segmentation, human parsing, and image matting. Our system guides a user's pose when taking photos by displaying the outline of the latest picture of the user, divides the human body into several parts, and compares before and after photos based on the body part-level segmentation results. The parsing performance has been improved through an ensemble approach and a denoising phase in our main module, called Advanced Human Parser. In evaluations, the proposed method was 0.1% to 4.8% better than the second best models in average precision in 3 out of 5 parts, and was 1.4% and 2.4% superior in mAP and mean IoU, respectively. Furthermore, the inference time of our framework takes approximately three seconds to process one HD image on an ordinary single GPU server, demonstrating that our structure can be applied to real-time applications.

Introduction

There are several ways to measure the human body. For example, scales can measure a person's weight, and BMI machines can measure their body mass index. In addition, people can save pictures of their bodies by taking photos to see their progress over time. When we compare a one-dimensional scalar's representation of the weight or BMI, it has less information present than a two-dimensional array's representation of a photo which is capable of providing more details(e.g., muscle shape, the visual balance of the full body) making it more suitable for an approximation of the human body. Thus, people, especially those who frequently exercise, track their progress by constantly taking pictures. However, some limitations make it challenging to take pictures while maintaining a consistent and repeatable pose, potentially leading to inconsistent comparisons.

We have subdivided these problems and attempted to find a solution that provides a more convenient and accurate way of tracking body shapes. As a result, we have identified three technical fields that are necessary to solve the problems *i.e.* Instance Segmentation, Human Parsing, and Image Matting. These methods have been incorporated into our system as key components.

Recently, as artificial intelligence-based research has been rapidly advancing, deep learning-based research in the field of computer vision has also been extensively developed, to the point where it could be considered saturated compared to when only conventional approaches were available. Therefore, we have naturally incorporated deep learning approaches into the design of our overall system architecture.

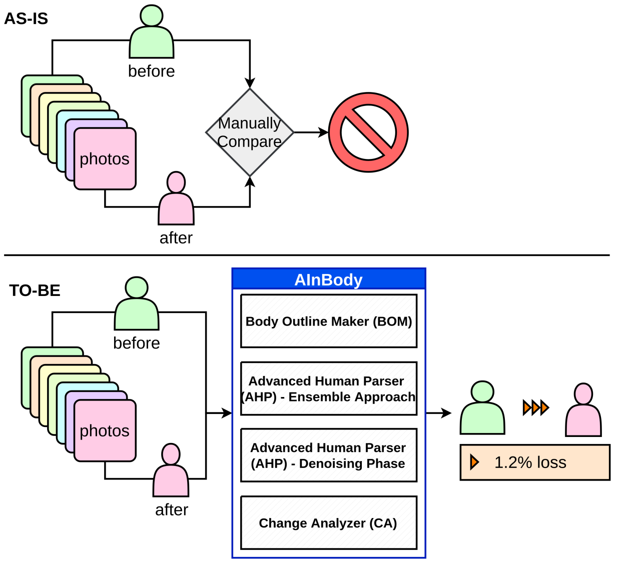

Figure 1. *The Comparison between the current method and the proposed approach (AInBody). AInBody automates the process of comparing body changes, which was previously done manually by humans.*

In short, in this paper, we suggest AInBody, a body shape tracking solution that guides the human pose into the pose of a previously taken picture, pixel-wisely segments the human body into several parts, and automatically compares the before and after photos. Fig 1 illustrates the advantages of our approach compared to the current method. Under our AInBody schema, people do not have to get into their nerves tracking photos of their body shapes.

First of all, we have developed a module for real-time pose guidance by using an instance segmentation model to identify and segment human figures in an image at the pixel level. This allows users to compare the pose being taken in real-time to previously stored poses. Moreover, to segment each body part, pixels, the smallest unit of an image, must be dealt with, so the two-dimensional human part segmentation method that has several publicly available datasets and is being actively studied has been chosen to be the basis. While three-dimensional human parsing may be more accurate, it is not practical for general users as it requires special cameras and specific environments to capture the necessary data. Using a network that receives a single two-dimensional image and spits out a three-dimensional body shape is also not ideal because the depth information should be predicted using 3D shape estimation methods like S. Saito *et al.* [1] which may result in aggravated error. Furthermore, to overcome the drawback of the segmentation model's inherent shortcomings that do not detect the contour precisely, we use an image matting method to accurately separate the background and foreground objects. The matting result is then merged into the parsing result to not only accurately detect the contour but also remove unnecessary noise or artifacts added from the human parsing model.

To sum up, our research has two main contributions. The first is the development of a real-time body progress tracking system that enables users to continuously measure their bodies. The second is the suggestion of an ensemble approach and a denoising phase to improve the performance of segmentation results, thereby contributing to the diversification of measuring tools.

Related Works

Instance Segmentation

Upon the arrival of deep learning era, research in the image segmentation field has been done rigorously. *Instance segmentation* is one of the most challenging tasks in segmentation since it predicts the class pixel by pixel in images and classifies each object individually. Compared to the quality of semantic segmentation models, obviously, the quality of instance segmentation marginally drops. Nevertheless, many instance segmentation models show reasonable results.

Mask-RCNN[2] model combines Faster R-CNN model and FCN to conduct instance segmentation. PANet[3] further developed Mask-RCNN model by suggesting new frameworks like bottom-up path augmentation. These frameworks enable the model to sustain low-level features in order to make the localization process more precise. These models are two-step models where it takes a great time in the feature localization process to create detailed masks.

Yolact[4] omits the localization process and creates prototype masks for the whole image. For every instance, this model predicts linear combination coefficients to make detailed segmentation of the human body. Unlike the two-step models mentioned above, it is more suitable for real-time inference because it performs major tasks in parallel. Consequently, Yolact was the best fit for our task because it is faster, thus, making real-time inference possible, which is indispensable for our task.

Human Parsing

Human parsing is a task that segments a human picture into various parts, including human body parts and clothes. Segmenting the human body accurately is a challenging and important task in this field. SCHP [5] proposed an iterative correction process to solve this challenge.

OSHP [6] pointed out that most human parsing models have difficulty discriminating the unseen classes and attempted to overcome this limitation using the 3-stage progressive model. Though the performance has improved, the confusion between unseen classes still exists as a shortage.

Since we are focusing on measuring human body parts and figuring out the change in each body part's shape, accurate segmentation of human body parts is the most crucial point. Even when models have the same structure but are trained on different datasets, the segmentation results can vary depending on the training data. This means that some models may be better at recognizing certain body parts than others. We take advantage of this characteristic by developing an ensemble model that combines multiple models and drops more accurate human part segmentation results.

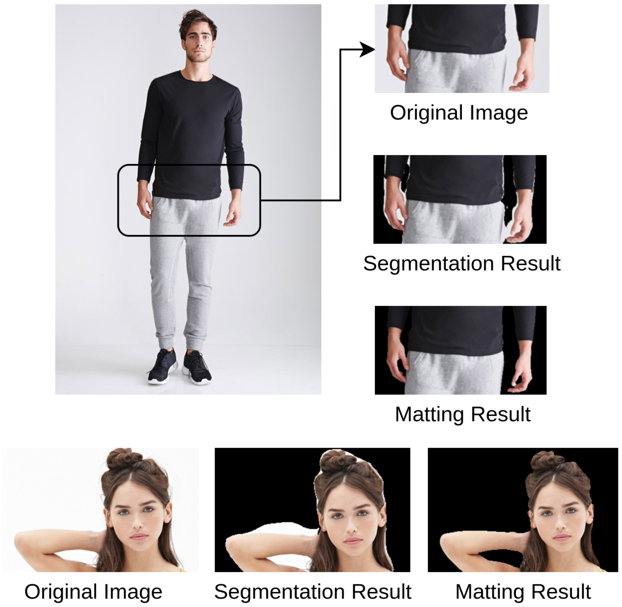

***Figure 2.** Comparing the segmentation and matting results. The matting method is better at accurately identifying the contour of the object compared to the segmentation method.*

Image Matting

Image matting is a task that separates the foreground and background by estimating the opacity of each pixel in the image. It is commonly used in pre/post-processing for photo and video editing. Compared to the segmentation approach that separates the binary mask (0, 1) without considering the opacity, image matting computes the level of opacity (between 0 and 1) for pixels around the edges of the foreground, resulting in more detailed and low-artifact images at the pixel level. Fig 2 clearly shows how the two are different.

Traditionally, the user have had to provide additional information, such as trimaps or hand-annotated hints, to support the estimation of opacity in image matting [7, 8]. However, there have been recent developments in image matting methods that do not require trimaps. One such method is PP-Matting[9], which is a trimap-free architecture that guarantees precise pixel-level classification and can reduce the artifact or noise at the boundaries of the image.

Proposed Method

The proposed method, AInBody, consists of three modules; *Body Outline Maker (BOM)*, *Advanced Human Parser (AHP)*, and *Change Analyzer (CA)*. In Figure 3, a flowchart is presented that details the process of using the service that includes our AInBody system, from start to finish. Figure 4 presents a high-level architecture of the proposed solution, showing the overall architecture and key components.

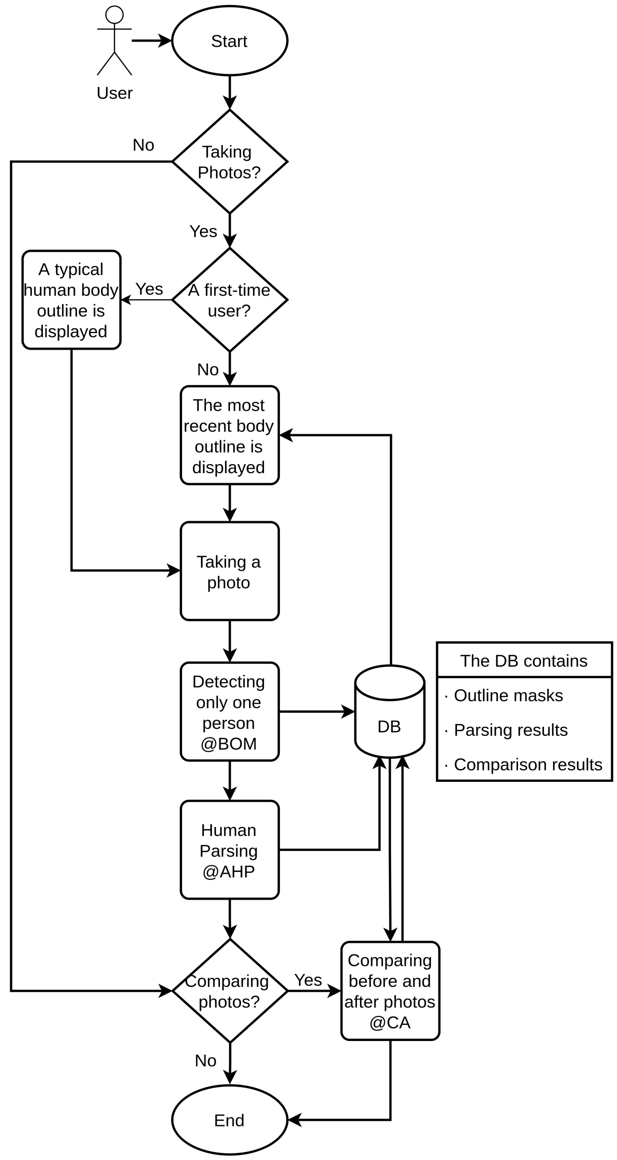

Figure 3. *A service flowchart of the proposed AInBody solution.*

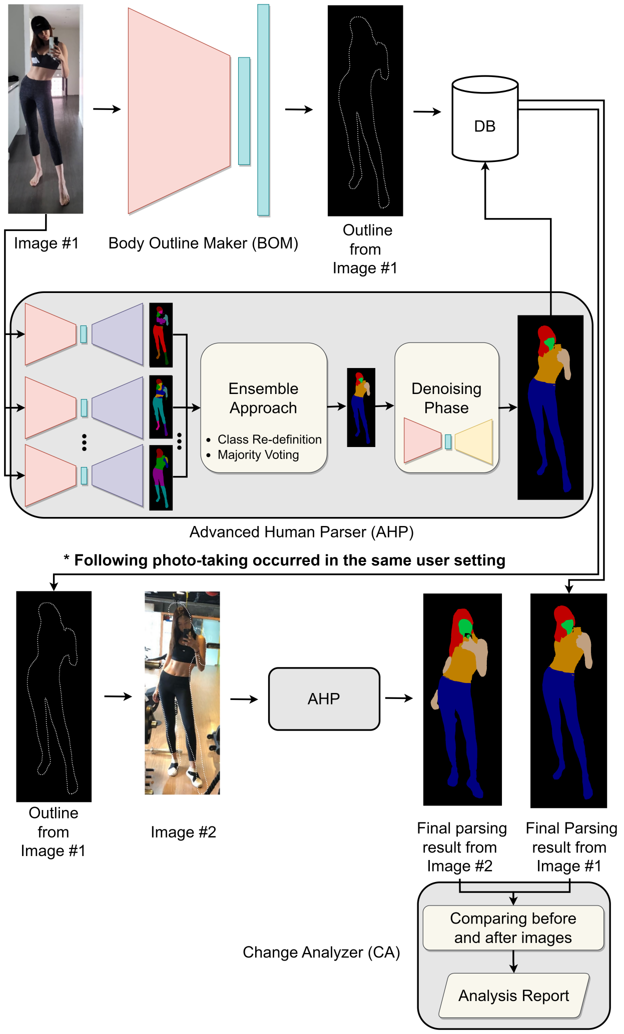

Figure 4. *A high-level architecture of AInBody illustrated with examples.*

Body Outline Maker (BOM)

Body Outline Maker (BOM) is a module that generates an outline of the human body based on a picture taken by the user. When a new picture is taken, the outline from the previous picture is displayed to the user as a pose guidance. This allows the user to see how their pose in the current picture compares to the pose in the previous picture. As described in Fig 4, the outline generated from the Image #1 is saved in a database. When the following picture is taken, the outline from the first picture is retrieved from the database and displayed as a reference for the user's pose.

We use the Yolact model with Resnet101-FPN as the backbone to generate the outline of the human body. When an input image is fed into this model, it classifies each pixel. Pixels that are identified as representing a 'Person' object are used to create polygons. The contours of all possible candidate regions within the image that could contain a person are then obtained. Since our system is designed to recognize a single user at a time, we calculate the area of each candidate region and select the silhouette with the largest area as the final outline. This allows us to focus on the single user and ignore any additional people present in the image.

BOM encourages users to continuously take pictures with consistent poses. This helps to reduce errors when comparing pictures. Moreover, it is able to provide real-time feedback to the user based on the overlap between the current pose and the previous pose. By calculating whether the overlap between the subject and the previous outline is equal to or greater than a certain threshold, the system can give instructions to the user in real-time, allowing the user to make adjustments to their pose as needed.

Advanced Human Parser (AHP)

Advanced Human Parser (AHP) segments a human body into several parts to obtain segmentation results. In *AHP*, SCHP[5] is applied as a basic model of human part segmentation, which has a typical encoder-decoder structure in CNN architecture and uses multiple pseudo-labeling to reduce the noise of ground

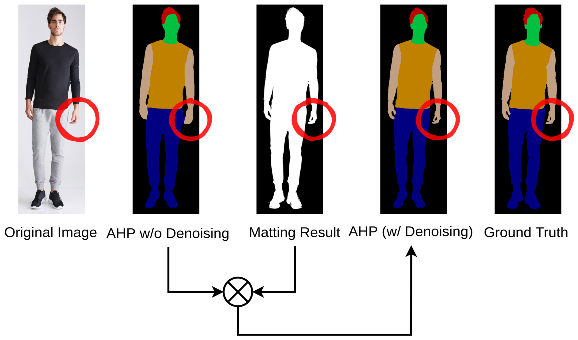

Figure 5. *The effectiveness of the denoising phase in AHP. After the parsing result and the matting result are merged, the merged result gets closer to the ground truth.*

truth for increasing prediction accuracy. We have observed, even when models have the same structure but are trained on different datasets, the segmentation results can differ based on the training data. This means that some models may be more proficient at identifying certain body parts than others. Therefore, we suggest an ensemble approach that combines multiple models. In practice, we trained three datasets (LIP[10], ATR[11], and Pascal-Person-Part[12]) with the same SCHP model.

The ensemble approach is attached after the inference processes of the multiple models. Our ensemble approach differs from traditional ensemble models in that we use a single model with different datasets. However, each of these datasets has its own distinct set of classes. In order to synchronize the datasets and make their classes identical, customized class re-definition is necessary for all of the datasets. In our implementation, we identified the essential body parts as Hair, Face, Arms, Torso, and Legs and therefore re-defined the classes to reflect these five parts. As an example, in order to re-define the classes of the ATR dataset (which has 18 classes) into five classes, pixels classified as pants would be mapped to the ”Legs” class, and pixels classified as shirts would be mapped to the ”Torso” class. To improve the accuracy of human parsing, the prediction results from multiple trained models are combined using a majority voting process. This process involves selecting the labels predicted by the majority of models for each pixel, and using those labels as the final prediction. If there is no majority, the prediction result of the Pascal model, which generally performes well in the reorganized class, is selected.

To further improve the accuracy of human parsing, we have added a denoising phase after the ensemble approach. Image matting is the main technique of the denoising phase. In practice, PP-Matting[9] is used. It takes an input image and seperates its background and object. The final output is generated by merging the matting masks and the parsing results. As shown in Figure 5, this denoising phase allows us to obtain more refined and accurate results than what was possible with the pre-trained human parsing network alone.

Change Analyzer (CA)

Change Analyzer (CA) compares two parsing results of each before and after photos. To compare the two results, the number of pixels is counted for each, and the percentage increase (%) for each body part is calculated to provide an analysis report of the change in body shape. The formulas for counting pixels and calculating the percentage increase (%) are represented in Formula 1 and 2.

Assume that, (x,y) is a two-dimensional image coordinate and if $f(x,y) = 1$, the function $pc(c)$ which counts the number of pixels for each body part c can be expressed as *i.e.*

$$pc(c) = \sum_{(x,y)\in c} f(x,y) = \sum_{(x,y)\in c} 1, \tag{1}$$

The percentage increase(%) for each body part c can be calculated as follows.

$$PercentageIncrease = \frac{pc_A(c) - pc_B(c)}{pc_B(c)} \times 100(\%), \tag{2}$$

where $pc_A(c)$ and $pc_B(c)$ are the number of pixels of body part c in the after image and the before image, respectively.

Experiments and Results

Datasets for training

The human parsing models in the ensemble approach of *AHP* were trained using three different datasets: LIP, ATR, and Pascal-Person-Part. These datasets were annotated seperately, so the types of body parts and the number of classes that can be trained differ. Table 1 provides more detailed information about the datasets.

Table 1. Three datasets used when training SCHP

Dataset Name	# Images (# Classes)	Classes
LIP[10]	50,462 (20)	Background, Hat, Hair, Glove, Sunglasses, Upper-clothes, Dress, Coat, Socks, Pants, Jumpsuits, Scarf, Skirt, Face, Left-arm, Right-arm, Left-leg, Right-leg, Left-shoe, Right-shoe
ATR[11]	17,706 (18)	Background, Hat, Hair, Sunglasses, Upper-clothes, Skirt, Pants, Dress, Belt, Left-shoe, Right-shoe, Face, Left-leg, Right-leg, Left-arm, Right-arm, Bag, Scarf
Pascal-Person-Part[12]	3,533 (7)	Background, Head, Torso, Upper Arms, Lower Arms, Upper Legs, Lower Legs

Datasets for Testing

To evaluate the proposed method, a new test set is needed. This is because the ensemble approach that combines human parsing models was trained on different datasets, and therefore the existing datasets cannot be used. To create this test set, we randomly sampled 45 images out of 44,096 images from DeepFashion-MultiModal[13], a recently published human dataset, and newly labeled them into five body parts: Hair, Face, Arms, Torso, and Legs. This test set is named ’DeepFashion-45-forAInBody’ and it consists of 25 female images and 20 male images in total.

Table 2. The result of comparing our solution to other methods

Method	mAP	Bkg (AP)	Face (AP)	Torso (AP)	Arms (AP)	Legs (AP)	Hair (AP)	mIoU
SCHP-LIP	92.99	98.50	92.22	86.37	89.48	97.51	93.84	78.59
SCHP-ATR	93.89	99.52	90.29	85.98	**94.38**	**98.53**	94.63	85.09
SCHP-Pascal	84.61	**99.72**	55.01	87.40	89.79	91.09	Nan	80.73
Ours (EA)	94.25	99.70	96.66	90.17	88.50	97.38	93.11	**87.47**
Ours (EA+DP)	**95.27**	99.23	**97.08**	**91.00**	91.33	98.25	**94.75**	86.54

Evalution Setup

Two experiments were conducted. The first experiment compares our method with the other models that were trained with a single dataset. The second is an ablation study that verifies the effectiveness of both the ensemble approach and the denoising phase.

Our architecture was implemented using Flask 1.1.2 with Python 3.8.5 and Pytorch 1.5.1, making it suitable for use in a server-client structure. The experiments were run on a server with an NVIDIA GeForce RTX 2070 Super with Max-Q Design.

Evaluation Metrics

To evaluate the performance of the segmentation result, we use both the IoU (Intersection over Union) and AP (Average Precision) as evaluation metrics, *i.e*.

$$IoU_c = \frac{Intersection_c}{Union_c} = \frac{TP_c}{(TP_c + FP_c + FN_c)}, \tag{3}$$

$$mIoU = \frac{1}{N}\sum_c IoU_c, \tag{4}$$

$$AP_c = \frac{TP_c}{(TP_c + FP_c)}, \tag{5}$$

$$mAP = \frac{1}{N}\sum_c AP_c, \tag{6}$$

where $c \in C$ denotes a body part class in a set of body part classes C, and N denotes the number of classes, with TP, FP, and FN indicating True Positive, False Positive, and False Negative, respectively.

Experiment #1 Testing Advanced Human Parser

To compare the performance of our AHP with other models, we conducted experiments using the the DeepFashion-45-forAInBody test set. These other models were trained with a single dataset. Table 2 shows our method outperformed the other models in terms of mean average precision (95.27%), part segmentation (face: 97.08%, torso: 91.00%, hair: 94.75), and average IoU (87.47%).

Table 3. The result of the ablation study

Method	mAP
Naive	93.89
w/ Ensemble Approach	94.25
w/ Ensemble Approach + Denoising Phase	**95.27**

Experiment #2 Ablation Study

In this experiment, we consider three methods: one where no method is applied (SCHP-ATR), one where the ensemble approach is used, and one where both the ensemble approach and the denoising phase are applied sequentially. Table 3 indicates that our proposed methods show higher accuracy in mAP compared to not using them. Specifically, using only the ensemble approach(94.25%) yields better results than not using it(93.89%), and using both the ensemble approach and the denoising phase together (95.27%) leads to the best performance.

Discussion

In our system, when the types of clothing worn are different, it can lead to increased errors in comparison. While limiting clothing choices could potentially be a solution, it is not a fundamental fix. Therefore, we aim to develop a method in the future that can reduce errors by predicting the body shape even when clothing is varied. Furthermore, we plan to continue researching ways to improve the performance of segmentation model itself and extend it to three-dimensional one.

Conclusion

In this research, we propose AInBody, a new system for tracking body shapes using deep learning-based instance segmentation, human parsing, and image matting methods. Our goal is to provide a more convenient and accurate way of tracking body progress. To achieve this, we have developed three modules: *Body Outline Maker*, which helps guide the pose of the user when the picture is taken, *Advanced Human Parser*, which provides more accurate parsing results, and *Change Analyzer*, which quantitatively compares before and after photos. We have demonstrated the feasibility of our framework by developing a well-structured architecture that meets our goals and can operate in real-time. Additionally, we have introduced a new method of measuring the human body by developing Advanced Human Parser, which performs better than existing approaches. Further work will focus on improving the segmentation model and providing a more informative visual representation for users. Lastly, since this study is about tracking the progress of the human body that changes over time, it can be used in health-care application services, and referenced in computer vision research on temporal image analysis.

References

[1] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization,” *CoRR*, vol. abs/1905.05172, 2019.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-

cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[3] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *CoRR*, vol. abs/1803.01534, 2018.

[4] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *ICCV*, 2019.

[5] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[6] H. He, J. Zhang, B. Thuraisingham, and D. Tao, "Progressive one-shot human parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1522–1530, 2021.

[7] G. Park, S. Son, J. Yoo, S. Kim, and N. Kwak, "Matteformer: Transformer-based image matting via prior-tokens," *arXiv preprint arXiv:2203.15662*, 2022.

[8] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," in *AAAI*, 2022.

[9] Y. Liu, L. Chu, G. Chen, Z. Wu, Z. Chen, B. Lai, and Y. Hao, "Paddleseg: A high-efficient development toolkit for image segmentation," 2021.

[10] "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 932–940, 2017.

[11] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, pp. 2402–2414, Dec 2015.

[12] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," 06 2014.

[13] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Author Biography

Nakyung Lee received her BS in computer science from Sookmyung Women's University (SWU, 2017) and her MS in computer science from Korea Advanced Institute of Science and Technology (KAIST, 2019). She is currently working at CJ OliveNetworks as an AI engineer, focusing on research in the areas of computer vision and image processing.

Youngsun Cho received his BS and MS in electronic engineering from Hanyang University. Since then he joined CJ OliveNetworks as an AI engineer and he works as a senior AI engineer now. His research area is computer vision.

Minseong Son received his BS in electrical and electronic engineering from Yonsei University (2018) and He is currently enrolled in the Master's Degree Program in Artificial Intelligence at the same university (2022). He is currently working at CJ OliveNetworks as an AI engineer and his research area is computer vision.

Sungkeun Kwak received his BS in Mathematics and Computer Science from Duke University (2018). He is currently working at CJ OliveNetworks as an AI engineer and his research area is computer vision.

Jihwan Woo is currently working at CJ OliveNetworks as a head of the AI Research. He is pursuing his research on providing AI services for art and media.