IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 21, 2020, accepted January 3, 2021, date of publication January 13, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051424

Multiple-Clothing Detection and Fashion Landmark Estimation Using a Single-Stage Detector

HYO JIN KIM[1], DOO HEE LEE[1], ASIM NIAZ[2], CHAN YONG KIM[1], ASIF AZIZ MEMON[1], AND KWANG NAM CHOI[1]

[1]Department of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
[2]STARS Team, INRIA Sophia Antipolis, 06902 Sophia Antipolis, France

Corresponding author: Kwang Nam Choi (knchoi@cau.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2019R1F1A1062612, and in part by the HPC Support Project funded by the Ministry of Science and ICT and the National IT Industry Promotion Agency (NIPA) of Korea.

ABSTRACT Fashion image analysis has attracted significant research attention owing to the availability of large-scale fashion datasets with rich annotations. However, existing deep learning models for fashion datasets often have high computational requirements. In this study, we propose a new model suitable for low-power devices. The proposed network is a one-stage detector that rapidly detects multiple cloths and landmarks in fashion images. The network is designed as a modification of the EfficientDet originally proposed by Google Brain. The proposed network simultaneously trains the core input features with different resolutions and applies compound scaling to the backbone feature network. The bounding box/class/landmark prediction networks maintain the balance between the speed and accuracy. Moreover, a low number of parameters and low computational cost make it efficient. Without image preprocessing, we achieved 0.686 mean average precision (mAP) in the bounding box detection and 0.450 mAP in the landmark estimation on the DeepFashion2 validation dataset with an inference time of 42 ms. We obtained optimal results in extensive experiments with loss functions and optimizers. Furthermore, the proposed method has the advantage of operating in low-power devices.

INDEX TERMS Multiple-clothing detection, classification, object detection, landmark detection, single-stage detector.

I. INTRODUCTION

Fashion image analysis has attracted significant attention in the fields of image processing and computer vision. The fashion industry is one of the world's leading industries in terms of the job market and generates revenues with an annual growth of 8.4%. Revolution in computer vision is reshaping fashion trends in the society. However, the fashion image analysis is more challenging than the conventional image analysis due to significant degrees of variations in different styles, designs, and appearances. These variations often make detection and clothing retrieving tasks complicated and difficult. Classification, bounding box detection, and landmark estimation further add to the limitations of fashion image analysis. The accurate clothing detection is one of the performance measures for this analysis; however, technical challenges, such as the availability of large datasets and inference time, should be addressed.

To solve these problems, Huang *et al.* [1] and Kiapour *et al.* [2] looked for informative regions by detecting bounding boxes; Chen *et al.* [3] and Bossard *et al.* [4] detected human joints. Liu *et al.* [6] presented the fashion landmark concept, assuming that the clothing bounding boxes are fed as prior information for both training and testing sets. Fashion landmark detection is helpful to predict useful key points on the human body. The pipeline first detects the human body parts where the clothing regions are followed by synthesizing the clothing on that part [5]. This model could perform robust and discriminative representation for clothes. Liu *et al.* [7] further proposed a deep fashion alignment (DFA) framework

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

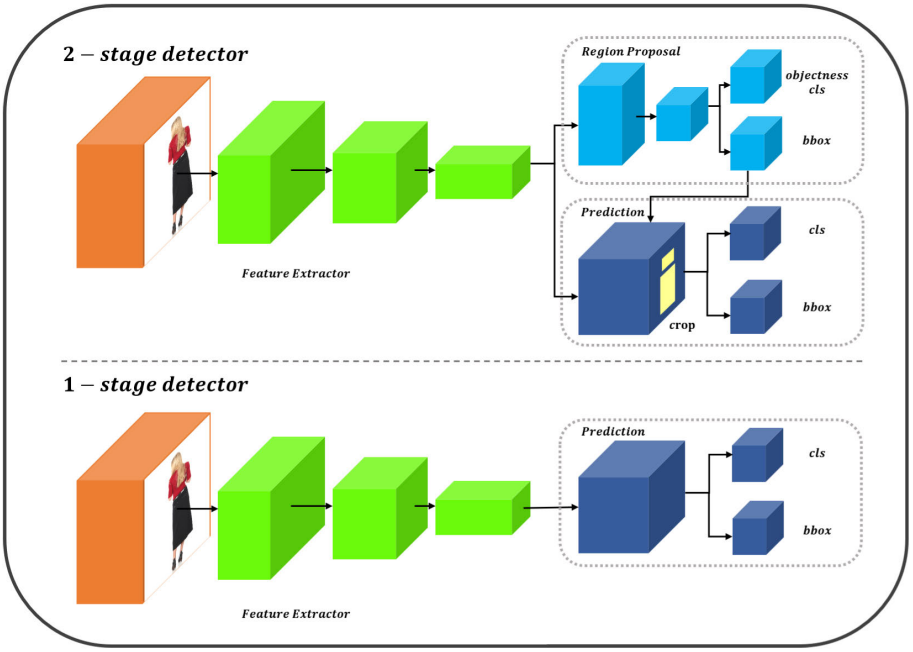


FIGURE 1. Examples of basic structures of a two-stage and single-stage detector. The two-stage detector has high accuracy but slow speed, as it contains the region proposal stage, whereas in the single-stage detector, the speed is high, but the accuracy is low due to imbalance of the foreground and background.

consisting of a three-stage convolutional neural network (CNN). The DFA successfully refined the predictions of each stage in the subsequent stages. However, both the approaches reported by Liu, *et al.* were computationally expensive, which made it limited for practice. Yan, *et al.*, [8] based on the Liu *et al.* DFA, proposed a deep landmark network, which could jointly estimate the landmarks and the bounding boxes, making it an applicable approach.

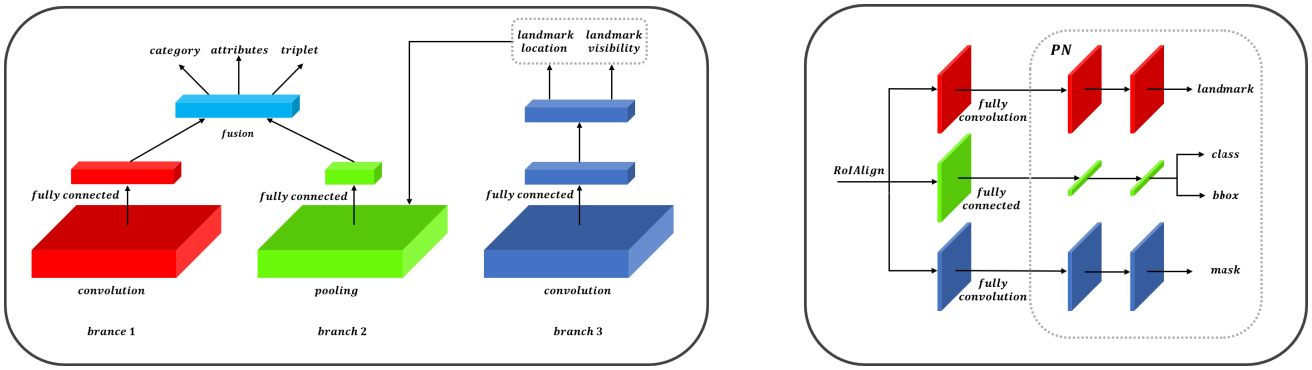
Object detection serves as the backbone of the fashion image analysis; it helps understand what the image contains and the location of this object. There are two different object detectors, i.e., a two-stage detector and single-stage detector [11]. To design and choose one of these detectors, certain factors must be considered carefully, such as localization, inference speed, and accuracy. Most object detection techniques are driven by region proposal methods [9] and region-based convolutional neural networks (R-CNNs) [10]. The two-stage detector identifies region proposals first, followed by classification. The single-stage detector [13], [23] executes both the region proposals and classification in parallel. For example, the two-stage detector, Faster R-CNN [12], has high accuracy and localization. In contrast, the single-stage detector, such as YOLO, shows high inference speed [13]. These factors make the two-stage detector more flexible and accurate than the single-stage detector, while the single-stage detectors are more efficient and have faster inference time than the former. Fig. 1 shows the basic structure of both the single-stage detector and the two-stage detector.

Most previous clothing detection and landmark detection studies have used the DeepFashion dataset, which had one item of clothing with 4 to 8 landmarks per image. Neural networks trained on such datasets cannot detect multiple clothes in images. Instead, we used the DeepFashion2 dataset, which includes more than 200,000 fashion images with annotations for 13 different classes. Each of its classes has specific keypoints with 294 unique landmarks.

Furthermore, we opted for a single-stage detector method as more feasible in real-world scenarios. The proposed network is an incremental improvement of EfficientDet by Google Brain [14]. Our network uses less graphical processing unit (GPU) resources, but has high inference speed and high detection accuracy.

This research focuses on detection and classification of multiple clothes. The main contributions of this study are as follows:

- We propose a single-stage detector that performs multiple clothing detection and landmark estimation in a fashion image with several complex feature attributes.
- This research was designed with a focus on solving the trade-off problem between accuracy and speed. The proposed method shows an accuracy of 0.686 mAP in bounding box detection, 0.450 mAP in landmark estimation task, and has a fast inference time of 42 ms in a single GPU. This shows the best balance compared to the existing methods.
- EfficientDet-based detectors are suitable for performing fashion image analysis in real-world applications because they consume less resources and are efficient with faster inference times.



(a) (b)

FIGURE 2. Illustration of exist model structure: a. FashionNet, b. Perception Network in Match R-CNN.

- This research can be a good example of training the network by focusing on a specific domain and applying computer vision to the real-world, and there is room for use in subsequent fashion image analysis research.

Experimental results and analysis confirm the efficiency of the proposed method.

The remainder of this paper is structured as follows. Section II covers the related background work. Section III describes the proposed method of detection and classification of multiple clothes using a single-stage detector. Section IV briefly explains the experimental setup. Section V describes the results and analysis of the experiment. Finally, Section VI presents the concluding remarks and suggestions for future work.

II. RELATED WORK

This section presents the background work related to the proposed method.

A. FashionNet

FashionNet [6] is a deep learning-based fashion image analyzer using a DeepFashion data set proposed by Liu in 2016. The network is designed based on VGG-16 and predicts landmark estimation and fashion attributes by modifying the last layer. Fig. 2 (a) shows the model structure of the last layer of FashionNet. The last convolution layer has three branches, the first one capturing the entire clothing item's global feature while the second capturing the local feature for the expected fashion landmark. The first and second branches' outputs are linked together to predict the clothing category attributes together and model the clothing pair. The last branch predicts the location and visibility of the landmarks. A key component of the FashionNet is the landmark pooling layer in the second branch that captures the landmark local feature. After apprehending the landmarks' visibility, perform max pooling inside the region to obtain a local feature map. This local feature map is stacked to form the final feature maps. This is similar to the RoI pooling layer introduced in [12], but it is meaningful to consider the interaction between fashion landmarks by connecting local feature without treating the pooled area independently.

FashionNet was meaningful because it used the first largescale fashion dataset called DeepFashion to perform a fashion image analysis and later contributed to the active research of fashion image analysis. However, the DeepFashion dataset has only the same number of landmarks for each clothing category, which is not enough to characterize many clothing categories. The Deepfashion dataset contains only one clothing per image, so the FashionNet cannot perform multi-clothes analysis.

B. MATCH R-CNN

Match R-CNN [18], proposed by Ge, is a fashion image analysis network using the DeepFashion2 dataset. Designed based on Mask R-CNN [39], it performs clothing detection, landmark estimation, and clothing segmentation. It consists of three networks called Feature Network(FN), Perception Network(PN), and Matching Network(MN). FN extracts the features and delivers the PN's output feature map through the RoIAlign proposed by [39]. Based on the feature maps received from the FN, the PN performs three predictions: landmark estimation, clothing detection, and segmentation. Fig. 2 (b) shows the structure of the PN. For bounding boxe and mask, it have the same structure as Mask R-CNN. Landmark estimation consists of a fully convolution network just like a clothing segmentation. MN is a matching network to learn the similarities of clothing as a network for clothing retrieval. Match R-CNN has contributed to further research as a guideline for performing fashion image analysis for multiple clothing. However, because it is designed based on the 2-stage object detector, Mask R-CNN, it is very inefficient in terms of inference time and resource consumption. This may limit the use of fashion image analysis in real-world applications.

C. DFA FRAMEWORK

Liu *et al.* [7] presented a DFA framework with a fashion landmark dataset of over 120,000 fashion images. Each image

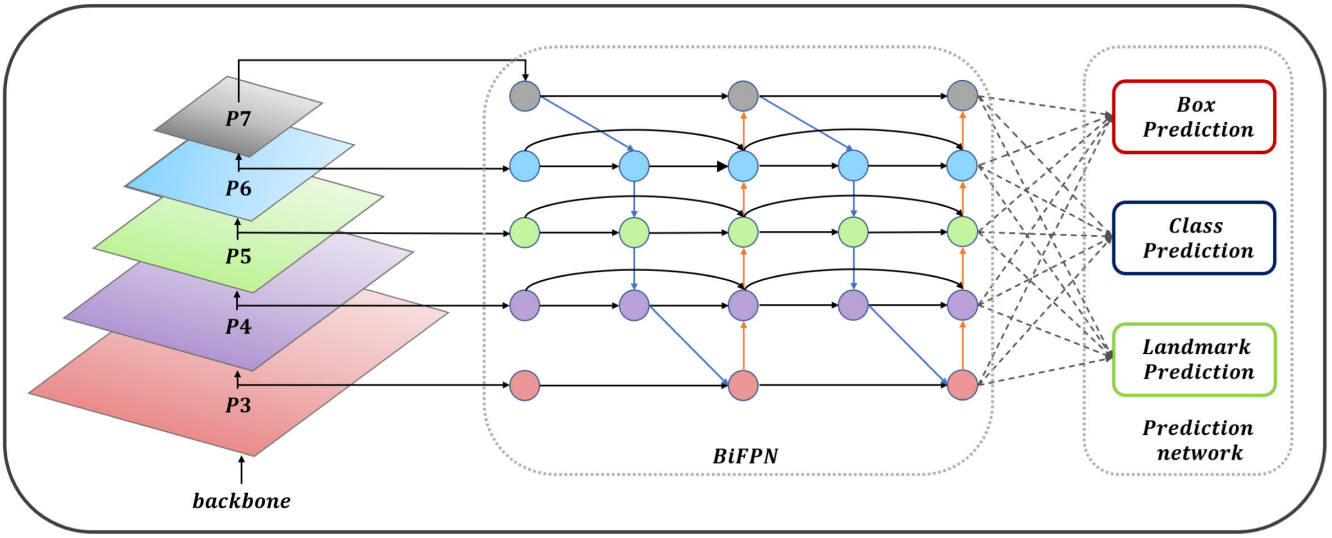


FIGURE 3. Overall network structure. It is based on EfficientDet-D1 and has 4 BiFPN layers because it has 1 coefficient. Feature maps of different sizes are all eventually entered into the prediction head to carry out the prediction.

had eight correctly labeled landmarks. Furthermore, images are separated into five subsets with respect to their ground-truth position and visibility. These subsets include normal, medium, and large poses; medium and large zoom-ins. Both the spatial and appearance domains of these subsets show vast variations (e.g., more than 30% images have large pose and zoom-in variations). To overcome these limitations, the DFA has three stages, wherein each stage refines the previous stage's predictions.

DFA deals with the full images and achieves better performance with less computational expense than other models such as DeepPose [21].

D. DeepMark

Sidnev *et al.* [24] presented a single-stage multiple-clothing detector based on CenterNet for clothing detection and landmark estimation. CenterNet [25] performs detection by estimating only one center point for each object. This assigns an anchor only with the position of the center point, not the box overlap, so only one anchor is used and the output has a high resolution. The advantage of using a single anchor prevents an imbalance between the positive and negative anchors to reduce the training time. The backbone used a stacked hourglass network and DLA-34 [40]. Hourglass network performs well in the keypoint estimation task, but the network is unsuitable for real-time inference because the network is deep. DLA-34 achieved a good balance between speed and accuracy.

However, DeepMark delivers directly from a backbone-through single feature map to the prediction head. It is efficient but does not take advantage of feature maps of various sizes and resolutions through the feature pyramid network (FPN). Using a structure called BiFPN, we assigned weight for each feature map with different resolution and noted the positive effects of this. The details can be found in Section III.

III. PROPOSED METHOD

Fashion image analysis should take advantage of the domain of fashion's features into account for scalability to the real world. Therefore, when designing a fashion image analysis network, we focused on the utilization of real applications. This should be an appropriate balance between accuracy and speed, and efficiency in terms of resource consumption. The EfficientDet proposed by Google Brain achieved state-of-the-art in the COCO dataset and is one of the best models to meet these conditions. The low number of parameters makes it efficient in terms of resource consumption, while at the same time being fast inference and high accuracy. Based on EfficientDet, we have modified the network globally to fit the fashion domain, added a prediction head, and designed a loss function for a new task called fashion landmark estimation.

This section details the overall network architecture, such as the proposed model's design direction, structure, and parameter setting. This section is divided into subsections describing the model, prediction head, loss function, and training.

A. MODEL

Speed and accuracy have a trade-off relationship, making it challenging to achieve high accuracy and functional efficiency. Therefore, the model must be carefully designed to maintain balances. EfficientDet [14] is a model that focuses on two tasks to maintain the optimal balance between speed and accuracy. We designed the network by modifying the EfficientDet, paying attention to this point. Fig. 3 shows our overall network structure.

We should first check if EfficientDet is suitable for specific work for fashion analysis only, i.e., fashion landmark estimation. Chen *et al.* [16] achieved excellent performance using a cascaded FPN to estimate key points from images of various sizes. The subnetwork, GlobalNet, composed of

FPN, solved the trade-off problem between high-resolution and low-resolution feature maps using element-wise sum after matching channels via 1×1 convolution operation. MultiPoseNet [26] also achieved good results by matching the number of channels after passing hierarchical CNN from the feature map over FPN. However, estimating fashion landmarks (294) with more keypoints than human key points (17) can vary significantly depending on the resolution of the input features. Therefore, there is a need to change the extent to which the input features contribute to the output features. Through the BiFPN structure using a cross-scale connection, we give additional weights to the input feature map to change the degree of contribution to the output feature map. This approach significantly contributes to performance by learning the importance of input features. The blue line in Fig. 3 is the weighted connection, and the orange line shows upsampling.

$$O = \Sigma_i \frac{w_i}{\epsilon + \Sigma_j w_j} \cdot I_i \tag{1}$$

This equation describes fast normalized fusion, which assigns weights. It is non-zero because it passes through the ReLU, and an epsilon of 0.0001 is added.

Backbone used EfficientNet [27] with pretrained weights. EfficientNet has an overwhelmingly smaller number of parameters than networks such as ResNet [28] and DenseNet [29], which are now widely used as feature extractors, but accuracy is also good. Hence, it is suitable for a method that focuses on the balance between accuracy and speed, and GPU resources can also be used efficiently. The stacked hourglass network [30] and HRNet [31], which are widely used in keypoints estimation, extract features with good accuracy by maintaining high resolution, but they were excluded from consideration because of insufficient speed.

B. PREDICTION HEAD

A prediction network has been designed to fit the fashion image domain. Fig. 4 shows the overall structure of the prediction network. It performs 3×3 convolution three times for each task by inputting a feature map Pi that has gone through BiFPN. Next, it adjusts the output channel by convolution with filter size 3, stride 1, and padding 1. Because it is a single-stage detector with no region proposal module, it uses nine anchors (of three sizes and three aspect ratios). The DeepFashion2 dataset has 13 clothing categories and 294 unique fashion landmarks. Therefore, the class prediction network's final output channel is ca($number\ of\ classes \times number\ of\ anchors$).

The boundary box prediction network predicts four coordinates (x_1, y_1, x_2, y_2) for regression, so it is $4a$. Landmark prediction networks require a large number of clothing landmarks to be estimated and should be carefully designed. First, the network estimates 294 landmarks and then fine-tunes the estimated landmark coordinates with an additional step of prediction, which is known as offset. The output channel for landmark estimation is $2ka$($number\ of\ landmarks \times number\ of\ anchors$), and the output channel for the feature map for offset$(\Delta x, \Delta y)$ estimation is $2ka$.

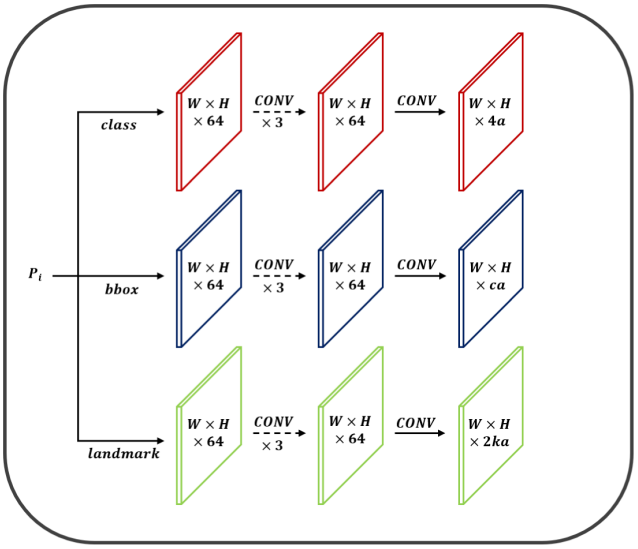


FIGURE 4. Structure of the prediction network. Each network performs classification, bounding box regression, and landmark estimation. We estimate 294 landmarks for each positive anchor and two additional offsets.

C. LOSS FUNCTION

A single-stage object detector has a higher speed but worse performance compared with the two-stage detector. The two-stage detector uses the region proposal module to separate the foreground and background to some extent. However, the single-stage detector uses a predefined anchor or grid, so it contains relatively large background areas. This creates a problem with class imbalance. To minimize the disadvantages of these single-stage detectors, we have employed the focal loss proposed by RetinaNet [32].

$$L_{cls} = -\alpha_t (1 - p_t)^{\gamma} log(p_t) \tag{2}$$

Focal loss minimizes loss renewal by giving less loss to the well-found case and maximizes loss renewal by giving a large loss for difficult cases. The alpha value was 0.25, and the gamma value was 2.

$$\begin{aligned} V &= \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \\ \alpha &= \begin{cases} 0, & if\ IoU < 0.5, \\ \dfrac{V}{(1 - IoU + V)}, & if\ IoU \geq 0.5. \end{cases} \\ L_{bbox} &= 1 - IoU + \frac{(px - gx)^2 + (py - gy)^2}{cw^2 + ch^2} + \alpha V \end{aligned} \tag{3}$$

$$L_{landmark} = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \tilde{y_i})^2}{N}}, \quad v > 0 \tag{4}$$

For bounding box regression, we use the complete IoU loss [33]. The C-IoU loss considers all overlapping areas, distances between center points, and aspect ratio. V is an equation in which aspect ratio was designed invariant

TABLE 1. Scaling configs for EfficientDet.

	Input size R_{input}	*Backbone Network*	*BiFPN* #*channels* W_{bifpn}	*BiFPN* #*layers* D_{bifpn}	*Box/class* #*layers* D_{class}
$D0\ (\phi=0)$	512	$B0$	64	3	3
$D1\ (\phi=1)$	640	$B1$	88	4	3
$D2\ (\phi=2)$	768	$B2$	112	5	3
$D3\ (\phi=3)$	896	$B3$	160	6	4
$D4\ (\phi=4)$	1024	$B4$	224	7	4
$D5\ (\phi=5)$	1280	$B5$	288	7	4
$D6\ (\phi=6)$	1280	$B6$	384	8	5
$D7\ (\phi=7)$	1536	$B6$	384	8	5
$D7$x	1536	$B7$	384	8	5

to regression scale, and alpha is the trade-off parameter, intended to maintain consistency of aspect ratio only in the bounding boxes that match well. We achieved better accuracy than L1 and L2 loss used in most object detectors. Details are given in Section IV. Fashion landmarks need to be predicted more carefully than a bounding box or human keypoints. L1 loss or smooth L1 loss is robust because of being less sensitive to outliers than L2, but low errors are almost ignored. Therefore, we designed the loss function in a sensitive but not non-intuitive because of the large difference. Rooted mean squared error is more intuitive than mean squared error (MSE) while paying close attention to the appearance of the outlier. v is the visibility of clothing landmarks. If visibility does not exist, it is not reflected in the loss. For landmark offset, the same equation as the landmark loss was applied. The total loss was optimized by minimizing the four-loss functions formulated as:

$$L_{tot} = L_{cls} + L_{bbox} + \lambda_{size} L_{landmark} + \lambda_{off} L_{off} \tag{5}$$

$$\lambda_{size} = 0.1 \ \text{ and } \ \lambda_{off} = 1.$$

D. TRAINING

EfficientDet is configured from D0 to D7 according to the coefficient so that lowering the value lowers the computational cost and decreases accuracy. In contrast, higher GPU resource consumption implies lower speed but higher accuracy. We have trained using two models, D0 and D1, because we aim to achieve near-real-time inference speed. Table 1 shows a detailed configuration of D0 and D1 models.

The input resolutions are 512×512, 640×640, and the backbone network uses EfficientNet-B0, B1. The batch size is 16, and the data augmentation only used flip. With the learning rate of $1e-3$ and the AdamP optimizer, we obtained better results than using the common Adam. The training was conducted using Nvidia Tesla V100 32GB GPU 3-way system.

IV. EXPERIMENTS

This section presents the techniques used in the experiment and their results. All experiments were performed on the publicly available DeepFashion2 dataset, which contains 191,961 images in the training set and 32,153 images in the validation set.

TABLE 2. Model performance based on EfficientDet-D0 and EfficientDet-D1.

	EfficientDet-D0	EfficientDet-D1
mAP_{box}	0.612	0.671
$mAP_{box}^{iou=0.50}$	0.734	0.792
$mAP_{box}^{iou=0.75}$	0.691	0.754
mAP_{pt}	0.392 0.363	0.470 0.441
$mAP_{pt}^{oks=0.50}$	0.679 0.655	0.736 0.722
$mAP_{pt}^{oks=0.75}$	0.490 0.410	0.550 0.462
$Inference\ time,\ ms$	37	42

A. BACKBONE

Fashion landmark estimation is complex and difficult because of the large spatial variation depending on the pose, occlusions, and style. In a single clothing landmark estimation, Lee *et al.* [15] tried to mitigate this using the global-local embedding module using a Gaussian map. In multi-human keypoints estimation, which is a similar task, Chen *et al.* [16] proposed to localize easy keypoints using the FPNs called the GlobalNet and then localize difficult keypoints through the RefineNet. Although there was a problem of low resolution when using FPN, it was solved element-wise using 1×1 *conv* after upsampling.

As a result of the previous two cases, we evaluated the feature extractor on the estimation of landmarks and keypoints. EfficientDet has models of different sizes, from D0 (faster and lighter) to D7 (larger and heavier). Among them, we compared D0 and D1, which are low-scale models for real-time inference. D1 has one additional level of BiFPN compared with D0; it has a larger input resolution, and the backbone uses EfficientNet-B1. Table 2 shows the results.

B. IoU LOSS

Most neural network-based object detectors typically perform direct bounding box regression of the center point coordinate, height, and width (x_{ct}, y_{ct}, w, h) of the bounding box using a loss function, such as MSE or L1/smooth L1.

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, & if\ |x| < 1 \\ |x| - 0.5, & otherwise, \end{cases} \tag{6}$$

For the anchor-based object detector, we estimate its offset ($x_{top\_left}, y_{top\_left}, x_{bottom\_right}, y_{bottom\_right}$). However, direct estimation of the coordinate at each point of the bounding box treats these points independently but does not consider the integrity of the bounding box itself. Therefore, several studies have focused on these problems. Recently, a new approach with intersection-over-union (IoU) loss was proposed. IoU is calculated by considering the area ranges of the predicted bounding box and the ground-truth bounding box. The IoU loss calculates the area of the predicted bounding box and the ground-truth bounding box to track the 4 vertices coordinates.

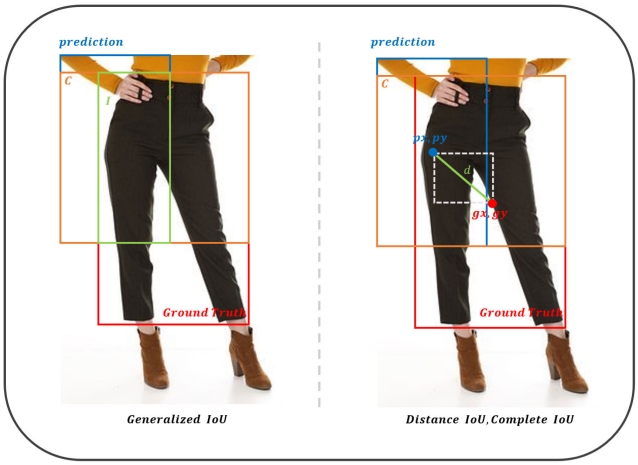


FIGURE 5. Example of the generalized IoU loss and distance/complete IoU method. The generalized IoU considers only the intersection and union between the ground-truth bounding box and the predicted bounding box. The distance IoU and the complete IoU consider the distance between the center coordinates. The complete IoU also considers the aspect ratio.

The IoU has a scale-invariance character. Therefore, it can solve the problem of increasing loss depending on a scale, such as the L1 and L2 loss function, which are the previous methods.

Generalized intersection-over-union (G-IoU) [34] loss considers the smallest enclosing convex object area. The method is to find a box of overlapping areas of the predicted bounding box and the ground-truth bounding box. This new box was proposed as a denominator to replace the denominator used in the previous IoU loss. The formula is as follows:

$$IoU = \frac{|I|}{|U|},$$
$$Generalized\ IoU = IoU - \frac{|C \backslash (P \cup G)|}{|C|}, \tag{7}$$

where I is the intersection between the predicted bounding box and the ground-truth bounding box, and U is the union. The distance intersection-over-union (D-IoU) loss [35] further considers the distance between the center point of the predicted bounding box and the center point of the ground-truth bounding box. The formula is as follows:

$$Distance\ IoU = 1 - IoU + \frac{(px - gx)^2 + (py - gy)^2}{cw^2 + ch^2}. \tag{8}$$

The complete intersection-over-union (C-IoU) loss [33] considers the overlapped area, the distance between center points, and the aspect ratio simultaneously.

We have applied smooth L1 and the above three methods to the proposed network. Table 3 shows the results. C-IoU outperformed other methods in bounding box regression.

C. COORDINATES CONVOLUTION LAYER

Various studies have been conducted to extract better image features. For example, the CNN performance was improved by adding several feature channels before the convolution step.

TABLE 3. Performance results according to the loss function.

	$smooth_{L1}$	*Generalized* IoU	*Distance* IoU	*Complete* IoU
mAP_{box}	0.671	0.672	0.675	0.676
$mAP_{box}^{iou=0.50}$	0.792	0.792	0.795	0.795
$mAP_{box}^{iou=0.75}$	0.754	0.754	0.756	0.759
mAP_{pt}	0.470 0.441	0.471 0.444	0.472 0.445	0.475 0.446
$mAP_{pt}^{oks=0.50}$	0.736 0.722	0.737 0.726	0.737 0.726	0.739 0.727
$mAP_{pt}^{oks=0.75}$	0.550 0.462	0.551 0.465	0.550 0.467	0.552 0.468

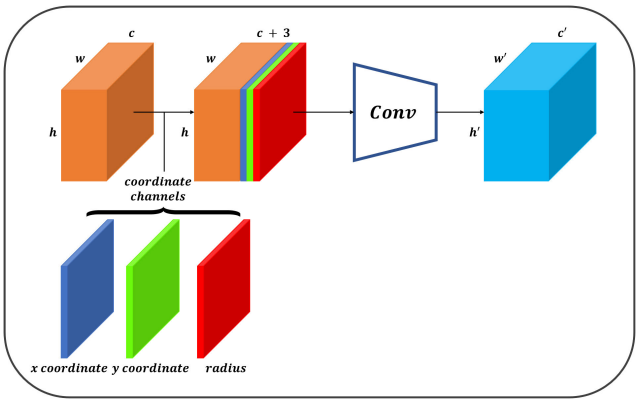


FIGURE 6. Coordinates convolution layer. The method is to add x, y coordinate and radius channels to the image and reflect them in the feature map.

Liu *et al.* [36] achieved good performance using a simple method. They experimented with basic data on coordinate transformation problem but could not obtain good results with general CNN. The workaround is the coordinate convolution, which adds coordinate information to the input image and adds it to the input channel. Fig. 6 shows this process. This method can achieve good performance by adding a few parameters and translation equivariant characteristics without impairing the speed. Hence, this approach can improve performance while minimizing the negative effect on the performance of existing models.

Coordinates are important features in fashion landmark estimation and bounding box regression. We added the radius channel and x, y coordinates before the image was input in the backbone network. Using this simple technique, we improved performance.

D. OPTIMIZER

Most object detection methods based on neural networks use the Adam optimizer. Adam's step size is not affected by gradient rescaling, and step size is bound even if the gradient grows, so any objective function allows for a stable descent for optimization. In addition, the step size can be adapted by referring to the previous gradient size. However, Adam is less normalized in L2 regularization because of its dependence on learning rate and weight decay. To solve

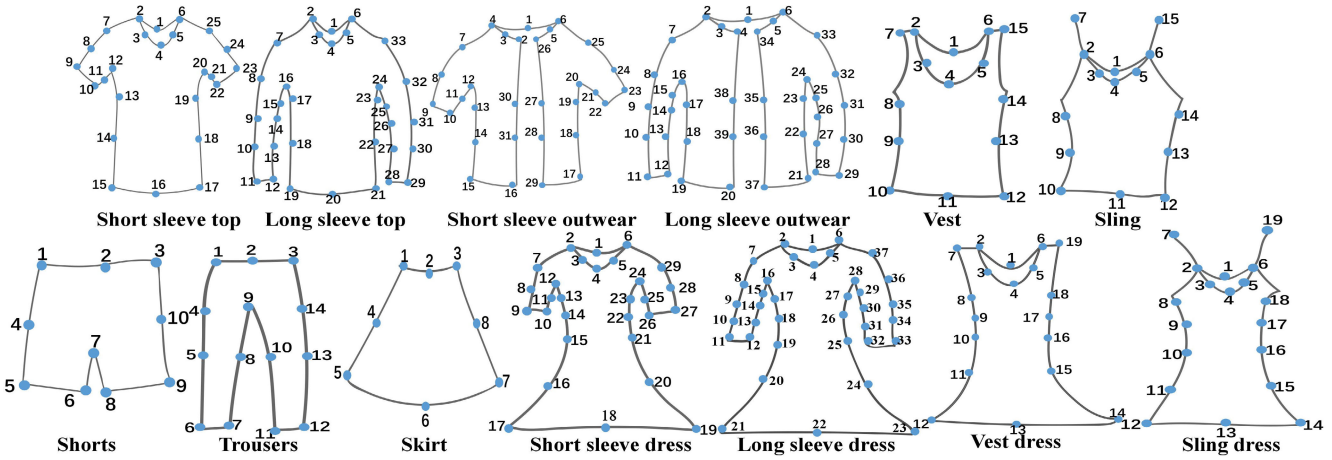


FIGURE 7. Definition of the DeepFashion2 dataset. It has 13 classes and different landmarks for each class. The number of landmarks in all classes is 294. The dataset contains the class name, bounding box coordinates, and landmark coordinates.

TABLE 4. Results of applying the coordinates convolution (CoordConv) layer to the complete IoU model.

	Normal	Coord Conv solution
mAP_{box}	0.676	0.686
$mAP_{box}^{iou=0.50}$	0.795	0.797
$mAP_{box}^{iou=0.75}$	0.759	0.765
mAP_{pt}	0.475 0.446	0.484 0.450
$mAP_{pt}^{oks=0.50}$	0.739 0.727	0.741 0.730
$mAP_{pt}^{oks=0.75}$	0.552 0.468	0.554 0.470
$Inference\ time,\ ms$	42	42

this problem, AdamW proposed by Loshchilov *et al.* [37] reflected the weight decay in the weight update equation so that the learning rate and weight decay were independent. In addition, they further proposed the AdamWR optimizer by adding a technique called a warm restart, which increases the learning rate in the middle of training and provides a chance to escape from the steep local minimum by creating a large weight update.

Moreover, Adam had scale-invariance in the normalization, wherein the weight of deep neural network was met with momentum, and the weight norm increased rapidly, causing the cumulative weight norm to slow the convergence of the descent. Recognizing that the cause of these problems is in normalized momentum-based-gradient descent, AdamP proposed by Heo *et al.* [38] eliminated the cumulative weight norm using the projection.

We conducted the experiment by applying Adam, AdamWR, and AdamP to our network. We obtained the best results using AdamP.

V. RESULTS

To verify the performance of the proposed method, another similar method the performance comparison with DeepMark and Match R-CNN was carried out. The mean average precision (mAP) was measured and compared for landmark estimation and bounding box regression. The inference time was also compared. In addition, this section shows examples of our test result images.

A. DeepFashion2 DATASET

In this research, DeepFashion2 dataset was used to train the proposed model. The DeepFashion2 dataset was compiled by Liu *et al.* in 2019 [18] with multipurpose benchmarks. These multipurpose benchmarks are clothing detection, landmark estimation, segmentation, and retrieval. DeepFashion2 fills the DeepFashion dataset gap as it contains multiple clothing items per image with rich annotations. The dataset contains more than 200,000 fashion images with 191,961 training samples and 32,153 testing samples.

B. COMPARISON WITH PREVIOUS METHODS

Our proposed method achieved the best results using the GPU system with three Nvidia Tesla V100 32 GB graphics card after training the model for 80epochs on the DeepFashion2 Challenge training set (191,961 images) with a batch of 48 images. We used EfficientDet MS COCO pretrained model for object detection as the checkpoint, and we have performed experiments with the EfficientNet backbone network.

Table 5 shows the results of the mAP comparison in the bounding box regression and landmark estimation between the DeepMark, Match R-CNN, and the proposed method. Match R-CNN in the DeepFashion2 paper shows good landmark estimation accuracy as a two-stage detector. However, Mask R-CNN, the base network of Match R-CNN, shows 44.4M parameters and 116 ms of the inference time. For Match R-CNN, the inference time would have increased further because of the addition of a landmark estimation network. In comparison, the number of parameters in the network proposed in this paper is only 7.3M, and the inference

TABLE 5. Results of performance comparison with the existing methods.

	Our Method	DeepMark				DeepFashion2 Match R-CNN
		DLA-34		Hourglass		
		Single	Fusion	Single	Fusion	
mAP_{box}	0.686	0.667	0.686	0.710	0.723	0.667
$mAP_{box}^{iou=0.50}$	0.797	0.788	0.802	0.821	0.827	0.814
$mAP_{box}^{iou=0.75}$	0.765	0.750	0.767	0.786	0.795	0.773
mAP_{pt}	0.484 0.450	0.469 0.432	0.513 0.448	0.582 0.512	0.601 0.532	0.641 0.563
$mAP_{pt}^{oks=0.50}$	0.741 0.730	0.730 0.716	0.746 0.730	0.784 0.774	0.793 0.784	0.820 0.805
$mAP_{pt}^{oks=0.75}$	0.554 0.470	0.547 0.455	0.563 0.473	0.660 0.571	0.683 0.599	0.728 0.641
$Inference\ time,\ ms$	42	35	216	76	315	-

FIGURE 8. Result of the proposed model. Demonstrations were conducted on images of different sizes and categories.

time is 42ms, which is much more efficient. The DeepMark-DLA34 backbone can be a good choice to solve the trade-off problem between accuracy and speed, but our proposed model is better in terms of accuracy. It has a lower accuracy than the result of the DeepMark-Hourglass backbone, but it performs much better in terms of inference time. Our method can be a best choice from the perspective of the balance of accuracy and speed without any image preprocessing. Moreover, the proposed method is suitable for low-power devices and real-time fashion image analysis.

C. EXAMPLE OF THE PROPOSED METHOD

This subsection describes the test result images using our trained model. Tests were conducted using Nvidia Tesla V100 32 GB single GPU and tested for random categories, including the images with single or multiple clothes. Fig. 7 shows the test results of the proposed network. Good results were achieved not only in single-clothing images but also in multiple-clothing images. However, we have confirmed that the results of the landmark estimation task are not good if the occlusion of clothing in the image is severe or too small.

Because there are few cases of fashion image analysis studies for multiple-clothing images, our result images can be a good reference in subsequent research.

VI. CONCLUSION

Herein, we proposed an approach as an adaptation of EfficientDet for the multiple-clothing detection and fashion landmark estimation. Focusing on the balance between inference time and accuracy, we have proposed a method that analyzes fashion images almost in real time on low-power devices. In addition, a new bounding box regression loss function, a different optimizer, and a strategy called coordinate convolution have been applied to the vanilla solution to increase the accuracy, which may be considered in other fashion image analysis models. The proposed method obtained an accuracy of 0.686 mAP_{box} and 0.450 mAP_{pt} with an inference time of 42 ms on the DeepFashion2 validation set for the two tasks, i.e., clothing detection and fashion landmark estimation. The proposed method is fast and accurate without image preprocessing. In a single stage, it performs efficient category classification, bounding box regression, and landmark estimation in parallel. Our proposed approach can contribute to future research on fashion images.

REFERENCES

[1] J. Huang, R. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1062–1070.

[2] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3343–3351.

[3] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany: Springer, 2012, pp. 609–623.

[4] L. Bossard, "Apparel classification with style," in *Proc. Asian Conf. Comput. Vis.*, Berlin, Germany: Springer, 2012, pp. 321–335.

[5] W.-H. Cheng, S. Song, C.-Y. Chen, S. Chusnul Hidayati, and J. Liu, "Fashion meets computer vision: A survey," 2020, *arXiv:2003.13988*. [Online]. Available: http://arxiv.org/abs/2003.13988

[6] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[7] Z. Liu, "Fashion landmark detection in the wild," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 229–245.

[8] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Unconstrained fashion landmark detection via hierarchical recurrent transformer networks," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 172–180.

[9] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[11] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[12] S. Ren, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[14] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[15] S. Lee, S. Oh, C. Jung, and C. Kim, "A global-local embedding module for fashion landmark detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3153–3156.

[16] 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.

[17] P. Li, Y. Li, X. Jiang, and X. Zhen, "Two-stream multi-task network for fashion recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3038–3042.

[18] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5337–5345.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[20] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[21] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[22] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

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

[24] A. Sidnev, A. Trushkov, M. Kazakov, I. Korolev, and V. Sorokin, "DeepMark: One-shot clothing detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3201–3204.

[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[26] M. Kocabas, S. Karagoz, and E. Akbas, "Multiposenet: Fast multi-person pose estimation using pose residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 417–433.

[27] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*. [Online]. Available: http://arxiv.org/abs/1905.11946

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] G. Huang, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[30] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 483–499.

[31] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 1, 2020, doi: 10.1109/TPAMI.2020.2983686.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[33] 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," 2020, *arXiv:2005.03572*. [Online]. Available: http://arxiv.org/abs/2005.03572

[34] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[35] Z. Zheng, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI*, 2020, pp. 12993–13000.

[36] R. Liu, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9605–9616.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*. [Online]. Available: http://arxiv.org/abs/1711.05101

[38] B. Heo, S. Chun, S. J. Oh, D. Han, S. Yun, G. Kim, Y. Uh, and J.-W. Ha, "AdamP: Slowing down the slowdown for momentum optimizers on scale-invariant weights," 2020, *arXiv:2006.08217*. [Online]. Available: http://arxiv.org/abs/2006.08217

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

[40] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

HYO JIN KIM received the B.S. degree in computer engineering from Daejin University, South Korea, in 2019. He is currently pursuing the M.S. degree with the Department of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea, as a Graduate Research Student.

Since 2019, he has been working as a Research Assistant with the Visual Image Media Lab, Chung-Ang University, under the supervision of Prof. Dr. Choi. His current research interests include object detection, semantic segmentation, landmark estimation, and fashion image analysis.

DOO HEE LEE received the B.S. degree in electrical engineering from Seokyung University, South Korea, in 2019. He is currently pursuing the M.S. degree with the Department of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea, as a Graduate Research Student.

Since 2019, he has been working as a Research Assistant with the Visual Image Media Lab, Chung-Ang University, under the supervision of Prof. Dr. Choi. His current research interests include object detection, semantic segmentation, and model optimization.

ASIM NIAZ received the B.S. degree in electrical and computer engineering from the COMSATS Institute of Information Technology, Pakistan, in 2016, and the M.S. degree from the Department of Computer Science and Engineering, Chung-Ang University, South Korea.

He is currently a Researcher with the STARS Team, INRIA Sophia Antipolis, France. His research interests include action recognition, video understanding, medical image analysis, and image segmentation.

CHAN YONG KIM received the B.S. degree in computer engineering from Daejin University, South Korea, in 2019. He is currently pursuing the M.S. degree with the Department of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea, as a Graduate Research Student.

Since 2019, he has been working as a Research Assistant with the Visual Image Media Lab, Chung-Ang University, under the supervision of Prof. Dr. Choi. His current research interests include 3D-object detection, object tracking, and autonomous driving.

ASIF AZIZ MEMON received the B.E. and M.E. degrees from the Mehran University of Engineering and Technology (UET), Jamshoro, Pakistan, in 2010 and 2015, respectively. He is currently pursuing the Ph.D. degree in application software from Chung-Ang University, Seoul, South Korea.

Since 2018, he has been working as a Research Assistant with the Visual Image Media Lab, Chung-Ang University, under the supervision of Prof. Dr. Choi. His research interests include image segmentation, image recognition, and medical imaging.

KWANG NAM CHOI received the B.S. and M.S. degrees from the Department of Computer Science, Chung-Ang University, Seoul, South Korea, in 1988 and 1990, respectively, and the Ph.D. degree in computer science from the University of York, U.K., in 2002.

He is currently a Professor with the School of Computer Science and Engineering, Chung-Ang University. His current research interests include motion tracking, object categorization, and 3D image recognition.