Face recognition for occluded face with mask region convolutional neural network and fully convolutional network: a literature review

Rahmat Budiarsa[1], Retantyo Wardoyo[2], Aina Musdholifah[2]
[1]Doctoral Program Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Yogyakarta, Indonesia

Article Info

Article history:

Received Sep 21, 2022
Revised Jan 27, 2023
Accepted Feb 3, 2023

Keywords:

Deep learning
Face recognition
Image recognition
Instance aware semantic segmentation
Mask region convolutional neural network

ABSTRACT

Face recognition technology has been used in many ways, such as in the authentication and identification process. The object raised is a piece of face image that does not have complete facial information (occluded face), it can be due to acquisition from a different point of view or shooting a face from a different angle. This object was raised because the object can affect the detection and identification performance of the face image as a whole. Deep leaning method can be used to solve face recognition problems. In previous research, more focused on face detection and recognition based on resolution, and detection of face. Mask region convolutional neural network (mask R-CNN) method still has deficiency in the segmentation section which results in a decrease in the accuracy of face identification with incomplete face information objects. The segmentation used in mask R-CNN is fully convolutional network (FCN). In this research, exploration and modification of many FCN parameters will be carried out using the CNN backbone pooling layer, and modification of mask R-CNN for face identification, besides that, modifications will be made to the bounding box regressor. it is expected that the modification results can provide the best recommendations based on accuracy.

Corresponding Author:

Retantyo Wardoyo
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science,
Universitas Gadjah Mada
Building C, 4th Floor, Sekip Utara, Bulaksumur, Senolowo, Sinduadi, Mlati District, Sleman Regency,
Yogyakarta Special Region 55281, Indonesia
Email: rw@ugm.ac.id

1. INTRODUCTION

Face recognition technology has been used in many ways, such as in the authentication and identification process. In the authentication process, facial recognition is used as a gateway for users to be able to access devices or systems, for example on laptops and smartphones. For authentication, it is more possible to use real time input using the camera. Face recognition is a combination of the face detection process and the face identification process. Face detection plays a role in the face localization process, namely the process to find the size and position of the face in the image [1].

Face recognition is one of the computer vision domains. Face recognition is never separated from detection [2]–[4] and identification of facial objects [5]–[7]. Most of the researchers conducted facial recognition research on a multi-object basis or in the input there were many faces, the obstacle in research usually lies in the detection which sometimes cannot detect many face objects in one frame [8]–[10]. Objects that cannot be detected are mostly because the information from the object cannot be perfectly received by the computer, or in other words the object is not all visible or only part of it [2], [11], [12].

The object that is lifted is a piece of face image that does not have complete facial information (occluded face) which only displays imperfect facial information, for example only half of the face is visible. A facepiece object (occlusion) that does not have all the facial image information can be due to acquisition from a different angle or shooting a face from a different angle. This object was appointed because it can affect the detection and identification performance of facial images as a whole [3], [4], [13]. In addition, research in the field of Face recognition can have a positive impact, such as in terms of security, assisting in searching and identifying one's data through image input. This research can also help many parties in facial recognition only the information is known from incomplete images, in other words the images are cut off.

The newest deep learning method for object segmentation is mask region convolutional neural network (mask R-CNN). Mask R-CNN can approach the human ability to perform segmentation classification of objects such as cars, animals, and humans. Mask R-CNN uses instance-aware semantic segmentation to label or classify objects, thereby distinguishing objects/instances within the same class. Mask R-CNN extends faster R-CNN by adding branches to predict object masks in parallel with existing branches for bounding box recognition [12].

This study will use the mask R-CNN method, this method was chosen because this method is proven to have good performance in object detection and recognition [4], [12], [14], in addition to therefore, this method can distinguish the same object in one image with instance-aware semantic segmentation or fully convolutional network (FCN). For example, in one image there are many objects in the form of cars, using this method we can identify an instance of each car object in the image at the pixel level. From its predecessor method, the mask R-CNN method is better [4].

The mask R-CNN method was used to prove whether this method was able to provide high accuracy in the classification of the input pieces of facial images. Modification of FCN is necessary because FCN is very necessary in the identification process, based on research [12] for images with many objects and a collection of different angles, there are still many that cannot be detected. To solve this problem, the research will conduct experiments on FCN modifications to perform segmentation with many parameters and many types of image sizes and quality to be inputted, so that later they can provide the best recommendations based on accuracy.

The use of facial data that does not have complete information or only in the form of a face image (occluded face) due to incomplete data acquisition or due to different facial angles. To solve this problem, one method that can be used is the mask R-CNN method with instance-aware semantic segmentation, but this method still has deficiencies in the segmentation or FCN section which results in a decrease in the accuracy of identifying facial objects with incomplete facial information (occluded face). Besides that, modifications will also be made to the bounding box regressor by removing the anchors that are outside the image.

2. RESEARCH METHOD

This literature study was based on several articles retrieved from Scopus. These articles will explain the research that has been done on face detection and face recognition. The articles collected them through the two stages explained in the following subsections.

2.1. Selection stage

Face recognition is one of the computer vision domains. Face recognition is never separated from detection [2]–[4] and identification of facial objects [5]–[7]. Most of the researchers conduct facial recognition research on a multi-object basis or in the input there are many faces, the obstacle in research usually lies in detection which sometimes cannot detect many facial objects in one frame [8]–[10]. Objects that cannot be detected are mostly because the information from the object cannot be perfectly received by the computer, or in other words the object is not all visible or only part of it [2], [11], [12]. There are several methods of detection and recognition of traditional image objects such as Viola-Jones [15], Hog+ support vector machine (SVM) [16], and deformable part model (DPM) [17].

There are many models and methods that can be used in facial recognition, one of which is deep learning. Convolutional neural networks [18] is one of the many methods available in deep learning that can be used for facial recognition [19]–[21] both in real-time [22] or not in real-time [23]. There is a study that builds a model based on normalized features extracted by deep CNN [24]. In addition, there is the convolutional channel feature (CCF) [25], which combines the advantages of the filtered channel feature and the CNN architecture, which has lower computational and storage costs than standard CNN methods.

The development of the first CNN method is was named region convolutional neural networks (R-CNN) [26], where R-CNN is done by finding regions or parts of images that can be objects, using the proposal region method, then each region will have a CNN for feature extraction. From this R-CNN research, other methods were obtained such as fast R-CNN [27], faster R-CNN [28], and the latest is mask R-CNN [12] where this method combines faster R-CNN by adding a new branch called mask to perform segmentation and called FCN [11], [29].

Several studies made modifications to the FCN section [30]–[32]. The multi scale FCN (MSFCN) [30] trains different face models for different face scales and these models share convolutional features, which can increase the computational efficiency of the network. utilizes generative adversarial networks (GAN) [33] to generate high-resolution small faces and MS-FCN composites to ensure the network uses adequate facial information to study scale-invariant facial features [34]. CV full convolutional neural network (CV-FCNN) for the classification of synthetic aperture radar (SAR) targets, which only contains convolutional layers in the hidden layer [31]. The purpose of replacing pooling and fully connected layers in a complex-valued convolutional neural network (CV-CNN) [35] with a convolution layer is to avoid complex pooling operations and prevent overfitting. Another method introduced mix-FCN to detect object locations and classify object types from images automatically [32], combined the spatial relation module and channel relation module for use in FCN for aerial images [36], improve the work of mask RCNN by improving the model's network structure by reducing the number of layers in the residual network, and adjusting the internal structure to strengthen the regularization of the model, increase the generalization ability, and then adjust the anchor box parameters and the loss of the loss function. to improve detection and segmentation accuracy [37].

In recent years, there have been significant advances in object detection using R-CNN-based methods. the faster R-CNN model using the WIDER dataset and verified performance on the FDDB and IJB-A datasets [38], improved the faster R-CNN framework through various architectural strategies, including multi-scale training, hard negative mining, and feature-blending on faster R-CNN [39], proposed the faster R-CNN method for face detection at different scales with low resolution data [2], proposed a cascaded backbone branch of the FCN (BB-FCN), and they used local facial landmarks to perform R-CNN-based face detection [40]. Face R-CNN [41], [42] adapts faster R-CNN for face detection and uses center loss [43] to pursue discriminatory features. Face R-FCN [44] based on R-FCN [45] eliminates the impact of non-uniform contributions from each part of the face by re-weighting embedding responses on score maps after the pooling region of interest layer (ROI). A study for multi-angle faces using R-FCN with multi-scale training [13]. FDNet [46] designed a deformable layer to make the studied features robust to scale variance.

Face recognition using deep convolutional neural network architecture is the method of choice for face recognition [8]–[10]. Deep CNN [18] maps facial images based on pixels into features that must have a small distance between classes and a large distance between classes. There are two main lines of research conducted to practice deep CNN facial recognition. Some researchers train multiple class classification divisions that can separate different identities within the training set, such as by using the softmax classification [8], [9], and others study embedding directly, such as triplet loss [10].

Other studies related to facial recognition [5]–[7] have been proposed to increase the discriminatory power of loss function that occurs in this softmax. Softmax has also been widely developed including by normalizing class-variant margin (CVM) [47], mis-classified vector [48], and sphere margins [49]. Pioneered center loss, the Euclidean distance between each feature vector and its center class, to obtain intra-class cohesiveness [43]. At the same time, the dispersion between classes is guaranteed by the combined softmax loss. However, updating training through an actual center was extremely difficult as the number of face classes available for training had recently increased dramatically.

Mask R-CNN [12] is a method that uses instance-aware semantic segmentation to label segmentation or classify objects, so that this method can distinguish objects/instances in the same class. Mask R-CNN are usually used for detection [14], [50] and object recognition [14], [51]. Mask R-CNN extends faster R-CNN by adding a new branch to predict the mask object in parallel with the existing branch for bounding box recognition. The segmentation on mask R-CNN [12] is FCN [11] which uses a convolutional neural network to convert image pixels into pixel categories. The FCN changes the height and width of the middle layer features by remapping to the input image size via the transformed convolution layer, so that the predictions have a one-to-one comparison correspondence with the input image in spatial dimensions (height and width). Because FCN uses a combination of convolutional layers, it is possible to do other combinations to get better segmentation than before.

In previous research, not many have given novelty to the segmentation section, even though segmentation is a very important part. Previous research, which focused more on loss function, and object and face detection, did not focus on face identification and recognition with imperfect data. So, in this study, it is proposed to develop the mask R-CNN method with instance-aware semantic segmentation and modify the FCN segmentation process in developing a face recognition model that uses face image information as a whole or only in pieces. In addition, this research will conduct segmentation experiments with many parameters and provide the best recommendations based on accuracy.

2.2. Analysis stage

Mask R-CNN [12] tries to expand the object detection capabilities using bounding boxes, with dense pixel-wise predictions, to provide a more complex understanding of an image. CNN is widely used for feature extraction from an image in the form of feature maps. Mask R-CNN uses these feature maps as input for the FCN, which generates a matrix. The resulting matrix is 1 for all pixel locations that are part of the object and 0 for all other locations. This matrix is known as the binary mask. With the binary mask obtained and the classification results and bounding boxes from faster R-CNN, Mask R-CNN can produce precise segmentation instances.

In this paper, we propose to use VGG16 as a CNN model. VGG16 is a convolutional neural network model proposed by Simonyan and Zisserman from the University of Oxford in the paper "very deep convolutional networks for large-scale image recognition" [52]. This model achieves a top-5 test accuracy of 92.7% on ImageNet, a dataset of more than 14 million images belonging to 1000 classes. It is one of the well-known models submitted to ILSVRC-2014. It made improvements over AlexNet by replacing large kernel sized filters (11 and 5 in the first and second convolutional layers, respectively) with multiple 3×3 kernel sized filters one after another. CNN VGG-16 itself consists of 16 CNN layers. The VGG-16 itself only uses 3×3 CONV stride 1 and 2×2 MAX POOL stride 2.

Region proposal network (RPN) is a FCN [11] which takes image input for any size and output in the form of a box from the object proposal that has an objectivity score [28]. RPN uses 9 types of the anchor with 3 ratios and 3 scales. The ratio is 1:1, 1:2, and 2:1. The scales are 128, 256, and 512. These scales and ratios are very important for overcoming the difference in ratios and scales. Since the 2,400 kernel window uses these 9 anchors, we have 21,600 anchors. Two methods are used to maximize the number of anchors, namely ignoring the cross-boundary anchors or anchors outside the image. The second method is to use non-max suppression (NMS). The way it works is on the positive intersecting objects to carry out many operations on the same object. From these anchors, the anchor with the maximum value is selected. That way, the number will become 2,000 anchor, which is more efficient. The formula for the RPN is as (1),

$$IoU=\frac{pixels(A \cap Gt)}{pixels(A \cup Gt)} \tag{1}$$

Explanation:
IoU : The ratio of the intersection of A with Gt to A union Gt,
A : Anchor or prediction box,
Gt : Ground truth boxes.

Intersection over union (IoU) stands for intersection over union where if IoU 0.7 then the image in the Anchor is considered an object, while if IoU < 0.7 then the image in the Anchor is considered not an object. RPN will use the 13th Conv layer to generate map features. Its size is 512, where 256 kernel windows are obtained from positive anchors and an equal number from negative anchors. 512 kernel windows are obtained from 40×60=2,400 kernel window/anchor locations, where later 2,400 anchors will justify the anchor location of the object and not. Object range values are 0 to 1, as well as non-objects. This value will be used for classification.

RPN loss function is a function to measure how big the error in the prediction is. The loss function is very useful for defining or giving positive labels to objects. In this case, 2 types of error measurement are carried out, namely the object/non-object classification and the loss function for box regression. The equation used in the RPN loss function is:

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \frac{1}{N_{reg}} \sum_i L_{reg}(t_i, t_i^*) \tag{2}$$

$$L_{cls}(p_i, p_i^*) = -\, p_i^* \log(p_i) - (1 - p_i^*) log(1 - p_i) \tag{3}$$

$$L_{reg}(t_i, t_i^*) = \begin{cases} 0.5(t_i - t_i^*)^2, & If\ |t_i - t_i^*| < 1 \\ |t_i - t_i^*| - 0.5, & otherwise \end{cases} \tag{4}$$

The formulation of classification loss function L_{cls} is similar to (2), except that the foreground and background classification loss $L_{cls}(p_i, p_i^*)$ changes to multi-class classification loss. The equation used in the classification of loss functions:

$$L_{cls}(p_i, p_i^*) = \sum_i^C p_i^* \log(p_i) \tag{5}$$

Total loss function:

$$L_{total}(p_i, p_i^*) = L_{reg}(t_i, t_i^*)+L_{cls}(p_i, p_i^*) \tag{6}$$

Explanation:
p_i : The predicted probability that the object contains,
p_i^* : Label with a value of 1 for a positive anchor and 0 for a negative anchor,
$\{p_i\}$: { p_1, p_2, ...} the set of predictive probabilities containing objects,
N_{cls} : Size in minibact (512),
N_{reg} : Anchor numbers or number of anchors in minibact (512),
t_i : The four coordinates of the bounding box,
t_i^* : The coordinates of the ground-truth window labeled positive,
$\{t_i\}$: $\{t_1$, t_2, ...} set of four bounding-box coordinates,
$L_{cls}(p_i, p_i^*)$: Loss function foreground and background classification of each anchor,
$L_{reg}(t_i, t_i^*)$: Loss function regression,
N_{cls}, λ : (constant value),
N_{reg} : Hyperparameters for adjusting the weight between two losses,
C : Number of all categories,
i : Index ancors in the mini-bath.

FCN uses a convolutional neural network to convert image pixels into pixel categories [27]. FCN changes the height and width of the middle layer features by mapping back to the input image size via the transformed convolutional layer. The predictions have a one-to-one ratio correspondence with the input image in spatial dimensions (height and width). Given the spatial dimension position, the output from the channel dimension will be a prediction of the pixel category according to the location.

FCN uses dense prediction, namely pixel-wise class labeling, to label image pixels in determining segmentation classes. All pixels are predicted one by one. Because it uses dense prediction, FCN has a prediction sensor that is the same size as the original image. To predict the size of the FCN, the closer to the output, the smaller the size, but the deeper the prediction is. After the prediction, the segmentation will be carried out.

Bilinear interpolation is performed using linear interpolation. first in one direction and again in the other. Although each step is linear in sample values and position, the overall interpolation is not linear but quadratic across the sample locations. Bilinear interpolation is one of the basic resampling techniques in deep learning, computer vision, and image processing. Bilinear interpolation is also called bilinear filtering or bilinear texture mapping. Because the prediction layer class is not the same size as the prediction sensor, we generate up-simple back 32 times by inserting 31 paddings, which is initialized using bilinear interpolation. After getting a sensor or the same size as the prediction, a convolution is carried out. The result is the sensor layer that will be used for prediction.

In mask R-CNN [12], we use FCN-8s. FCN-8s uses 3rd, 4th, and 5th pooling (conv7). Up-sample Conv-7 four times and up-sample again twice for pooling-4 so that it has the same size as pooling 3. After all three poolings have a size of 28×28 like pooling-3, join the three pooling using the 1×1 convolution. Which results in a deep depth of 21 with dimensions of 28×28. Up-sample back 8 times to get the insert padding, then convolution and up-simple back to get a more suitable segmentation to the image.

On mask R-CNN, you must train the proposal object, object detector, and so on to get n to n segmentation sizes. For the results, FCN and mask R-CNN have a difference. Mask R-CNN is better in segmentation because they change ROI Pool to ROI Align form with 10%-50% accuracy improvements and decouple masks and class predict.

The FCN [11], [12] loss function uses a formula or equation per pixel-softmax loss function. Matrix multiply+bias offset is a network with 3 classes. Input from the network is performed an exponential operation. Then normalization is performed. In the learning process, the process of minimizing the value of cross-entropy loss (softmax) is carried out with the formula:

$$H(p, q) = -\sum_x p(x) \log q(x) \tag{7}$$

Explanation:
$p(x)$: Ground-truth probability (0/1).
$\log q(x)$, L_i : Predictive probability.

Attempts are made to reduce the loss function by minimizing cross-entropy to get less value or loss function. This is a learning process in this method. The loss function is performed on all existing pixels.

Region of interest is a sample in a “data set” that is identified for a specific purpose [53]. The ROI concept is generally used in many application areas. For example, in medical imaging, a tumor's boundaries can be defined on an image or volume to measure its size. The endocardial border can be defined on the image, possibly during different cardiac cycle phases, for example, end-systole and end-diastole, to assess cardiac function. In “geographic information systems” (GIS), ROI can be taken literally as a selection of polygons from a 2D map. In “computer vision” and “optical character recognition”, ROI defines the object's boundaries under consideration. ROI is also often used in face detection or object detection.

The size of the input image after CNN VGG16 will have the initial size divided by 32 and the object image. In ROI Pool, the value of objects or images with size with the number of floats will be converted into an integer. Because the size of the object after the CNN VGG16 process and the size that softmax and regression boxes can accept is 7×7 on the fully connected layer, resizing must be done using max-pooling with the object size divided by 7 (for example, 20/7=2.86 divided by 2). Because there is quantization that occurs twice, information is lost from input to output. This is fine for classifications such as the fast and faster R-CNN.

There is a difference between ROI Pool and ROI Align. ROI Align does not round up for results from CNN VGG16, so they are still using floats. For example, to get an image at coordinates 2.97 is done by bilinear interpolation. We use Mask R-CNN Architecture in Figure 1.

In this study, an experiment will be conducted to perform segmentation FCN with many parameters and many types of image size and quality to be inputted, so that later it can provide the best recommendations based on accuracy. The calculation of the loss function is carried out on the RPN and FCN sections. The parts that were modified are as:

a) Box regression, in this section an overhaul will be carried out by removing anchors that are outside the size of the image and only using anchors that are in the image. The difference can be seen as in Figure 2.

The blue line in the Figure 2 is the ground truth box while the orange is the anchor. Figure 2(a) shows the original Mask R-CNN where anchors that are outside the image are also used. Figure 2(b) shows a modified Mask R-CNN that will be used by eliminating the use of anchors that are outside the image to save running time. This research model using Figure 2(b).

b) Mask Brach, in the mask section, FCN will be used where FCN will be modified by experimenting with using different pooling from the existing ones to get the possibility of increasing segmentation accuracy. FCN experiments can be seen in Figure 3.

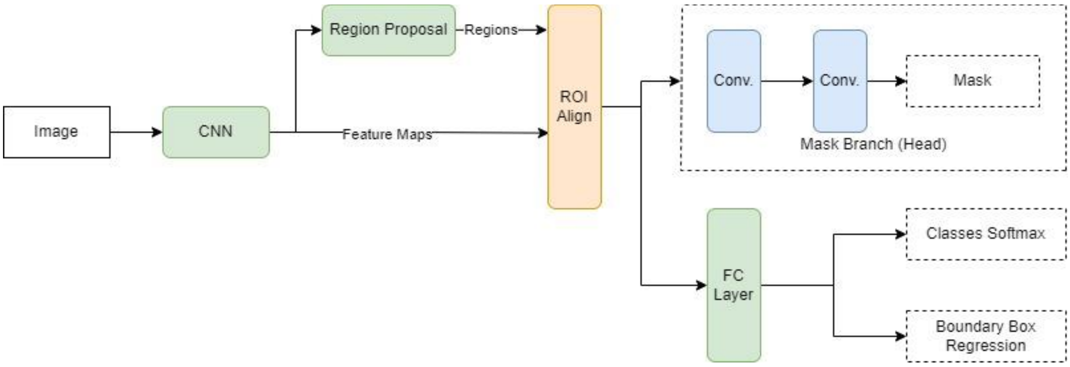


Figure 1. Mask R-CNN architecture

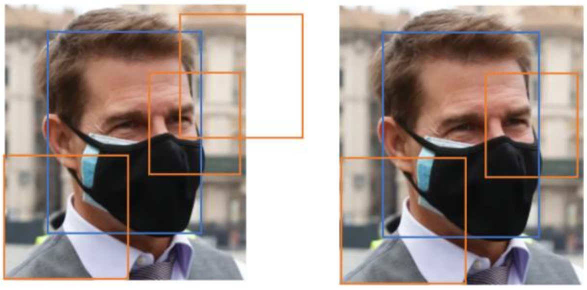

(a) (b)

Figure 2. Comparison of anchor usage (a) with anchor outside the image and (b) without anchor outside the image

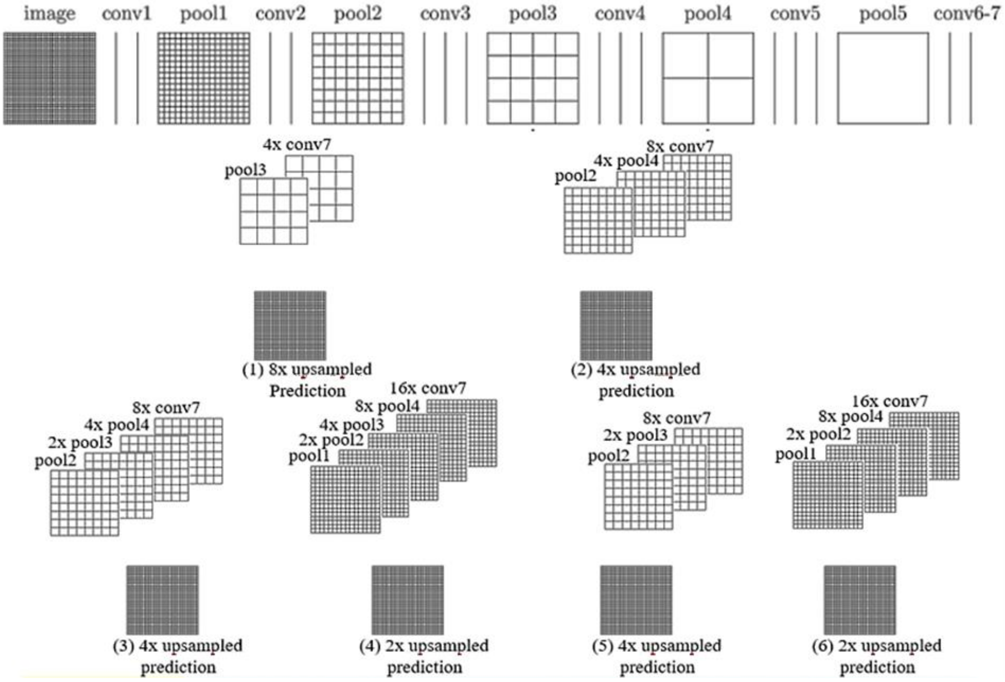


Figure 3. Modified FCN

The FCN section in Figure 3, various segmentation experiments or combinations of FCN with many parameters will be carried out and provide the best recommendations based on accuracy. Besides that, various image forms, both sizes and resolutions, will be used. Testing and training are carried out using various resolutions and image sizes with the aim of seeing the results of facial recognition.

3. RESULTS AND DISCUSSION

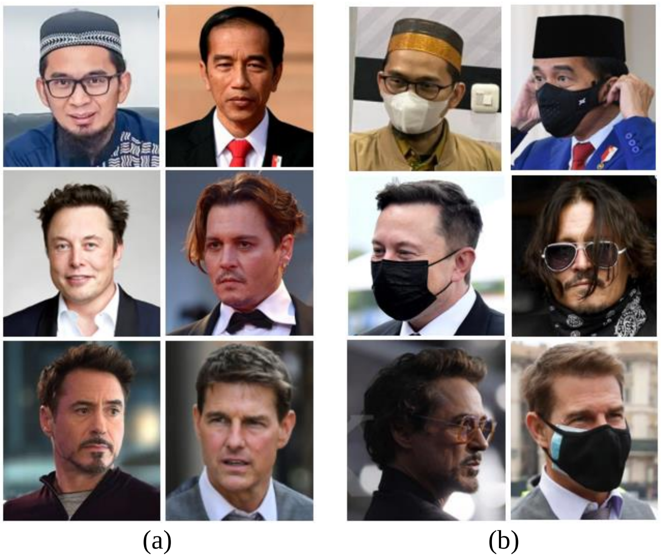
This paper focuses on the method ability to recognition and search for faces that have similarities to the input, where input can only be done with data on some faces. This dataset consists of various resolutions which will later be used to determine whether the segmentation method and parameters used affect the resolution or not. The amount of data used is expected to increase efficiency in the implementation of mask R-CNN. The data used does not have to be a complete face image (occluded face). The recommended data in this study is facial section data (occluded face) show in Figure 4. Figure 4(a) shows the dataset displaying all the facial information. in Figure 4(b) dataset that does not show some facial information such as the use of masks and sunglasses as well as shooting at an angle that only shows half the face (occluded face). Figure 4(a) is used for the model. whereas, Figure 4(b) is used for model and inputs.

Lin *et al.* [4], the good face detection results from the G-mask and mask R-CNN methods from comparison with other methods using the same, namely the FDDB, AFW, and WIDER FACE dataset. However, at the speed and efficiency of running time, the faster R-CNN is still the fastest. Running time of deference method in Table 1.

The method proposed from the research [4] in Table 1 compared with several major methods including multi-scale CNN (MSCNN) [54], contextual multi-scale region-based CNN (CMS-RCNN) [55], ScaleFace [56], multitask cascade CNN [57], and Faceness-WIDER [52]. The G-Mask method's precision-recall curve on the WIDER FACE benchmark of the G-mask method obtained 0.902 AP in the easy subset, 0.854 AP in the medium subset, and 0.662 AP in the hard subset. Compared to the advanced MSCNN method, the proposed method's AP value was only 0.014 lower in the easy subset and 0.049 lower in the moderate subset. There are some gaps between the G-mask method and MSCNN on the hard subset.

The reason may be that the MSCNN methods employ a series of strategies for small-scale face detection, and thus they can handle more challenging cases. Despite this, the G-mask method still delivers promising performance, which shows the G-Mask method's effectiveness. From the previous explanation, Mask R-CNN can be a good method in terms of face detection and running time, which is almost the same as faster R-CNN in detecting faces. In this paper, an architecture is designed to overhaul the FCN so that face detection can be even better because segmentation plays an important role in face detection.

The face recognition system design proposed in this paper is shown in Figure 1. From Figure 1, we can see how the data flow is processed, starting from the incoming image data, then the CNN value will be calculated, then using the proposed region network, a feature map will be obtained. Then apply the ROI pooling layer obtained from the ROI Align calculation on the bounding boxes to bring all the RPN candidates on the feature map to the same size.



(a) (b)

Figure 4. Dataset (a) show all face information and (b) not show all face information

Table 1. Running time of deference method

Method	Running time (s)		
	FDDB	AFW	ChokePoint
Faster R-CNN [4]	0.30	0.32	0.28
G-Mask [4]	0.35	0.42	0.33
Mask R-CNN	0.32	0.35	0.33

Meanwhile, segmentation is done using the FCN method, which combines several layers on the feature maps and CNN VGG-16 to perform segmentation. Proposals are forwarded to fully connected layers to classify and display bounding boxes for objects. The final step is to join forces between mask branches, box regression, and classification to get results in facial recognition.

The recommended CNN backbone is CNN VGG-16. RPN will use the 13th layer of CNN calculations to generate a feature map and use (1), where it is determined that if the input image contains a face object, it will be rated as 1. For images other than the face object, it will be considered the background, for this face image, which will be detected and predicted. There can be many face objects in one image, using RPN and instance aware semantic segmentation can label the same object in one image, for example, there are two faces, labeling face 1 and face 2 will be done so that the 2 objects can be distinguished.

ROI align must obtain an ROI pooling layer with a feature map size that can be processed by classification and box regression. ROI align calculates the face image's image/pixel by not rounding for the CNN VGG-16 results, so the numbers obtained are using floats. In other words, there is no quantization of the image coordinates.

FCN here functions in the form of segmentation by labeling images using dense prediction, namely pixel-wise class labeling the results of FCN in the form of branch masks. Because the prediction layer class is not the same size as the prediction sensor, a simple back-up is generated by padding the insert 31 (32-1=31) times, which is initialized using bilinear interpolation. After getting a sensor or the same size as the prediction, a convolution is carried out. The result is the sensor layer that will be used for prediction.

This paper provides experimental recommendations for segmentation (FCN) with many parameters and many types of image sizes and quality to be input so that later we can provide the best recommendations based on accuracy and running time. Loss function calculations are performed on the RPN and FCN sections. The experimental design in Figure 5 proposed and carried out in this paper uses the complete architectural design of mask R-CNN, instance aware semantic segmentation, and modified FCN. As in Figure 1, the image input will be processed with CNN VGG16, which uses 6 states where the first state has the same size as the input image. For each subsequent state, it will have half the size of the previous state. Next, we will determine the class prediction layer, using the 7th convolution layer, which has dimensions of 7×7. After getting the class prediction layer, it will be up simple back 32 times by inserting padding (32-1=31) 31 times to get a pixel-wise prediction size that matches the size of the input image. The experiment will be combination with various segmentations or FCN with many parameters and provide the best recommendations based on accuracy and running time, show in Figure 5.

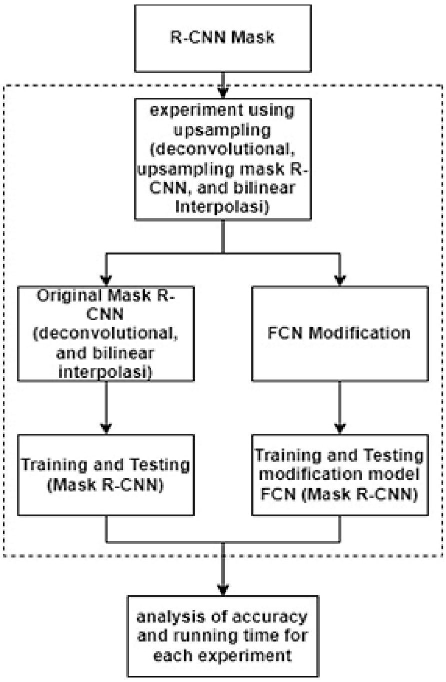

R-CNN Mask
experiment using upsampling (deconvolutional, upsampling mask R-CNN, and bilinear Interpolasi)
Original Mask R-CNN (deconvolutional, and bilinear interpolasi)
FCN Modification
Training and Testing (Mask R-CNN)
Training and Testing modification model FCN (Mask R-CNN)
analysis of accuracy and running time for each experiment

Figure 5. Experimental design

In this experimental, different image shapes will also be used, both in size and resolution. Testing and training are carried out using various solutions and image sizes to see facial recognition results. The output of this study is that in addition to performing facial recognition, it also obtains the results of a combination of FCN and image quality based on facial recognition accuracy and running time, which will later be useful as references for further research.

FCN modification for face recognition, as shown in Figure 3. FCN 4s is one of several modifications that will be made in our experiment. It is hoped that this modified FCN will have a good impact on facial recognition segmentation.

4. CONCLUSION

This paper discusses several efficient architectures in face or facial recognition based on face detection architectures, such as the main method of mask R-CNN and several overhaul methods, such as G-MASK, MSCNN, CMSRCNN, ScaleFace, multitask Cascade CNN, and Faceness-WIDER. This research will focus on research cases with occluded facial objects. The experiment using segmentation FCN with many parameters and many types of image size and quality to be inputted, so that later it can provide the best recommendations based on accuracy. The calculation of the loss function is carried out on the RPN and FCN sections. The parts that were modified is mask branch and box regression. For that reason, we made architecture and design to overhaul mask R-CNN and FCN by designing other pooling merging experiments that have never been used, such as FCN 8s (Merging FCN pooling 7, 6, 5, and 4), to make image segmentation better. It is hoped that this architecture will provide efficient results in face detection and time efficiency, both using high-quality image input and low-quality images. This architecture is also used for face recognition, while most research conducted with mask R-CNN is face detection.

ACKNOWLEDGEMENTS

The author would like to thank Gadjah Mada University and Pendidikan Magister menuju Doktor untuk Sarjana Unggulan (PMDSU) who have provided facilities and funded the research.

REFERENCES

[1] K. Oliver, *Witnessing: beyond recognition*. Minneapolis, USA: University of Minnesota Press, 2001.

[2] W. Wu, Y. Yin, X. Wang, and D. Xu, “Face detection with different scales based on faster R-CNN,” *IEEE Transactions on Cybernetics*, vol. 49, no. 11, pp. 4017–4028, Nov. 2019, doi: 10.1109/TCYB.2018.2859482.

[3] O. Cakiroglu, C. Ozer, and B. Gunsel, “Design of a deep face detector by mask R-CNN,” in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, Apr. 2019, pp. 1–4, doi: 10.1109/SIU.2019.8806447.

[4] K. Lin *et al.*, “Face detection and segmentation based on improved mask R-CNN,” *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1–11, May 2020, doi: 10.1155/2020/9242917.

[5] J. Deng, Y. Zhou, and S. Zafeiriou, “Marginal loss for deep face recognition,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 2006–2014, doi: 10.1109/CVPRW.2017.251.

[6] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6738–6746, doi: 10.1109/CVPR.2017.713.

[7] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, “Range loss for deep face recognition with long-tailed training data,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5419–5428, doi: 10.1109/ICCV.2017.578.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708, doi: 10.1109/CVPR.2014.220.

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Procedings of the British Machine Vision Conference 2015*, 2015, pp. 41.1-41.12, doi: 10.5244/C.29.41.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[11] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[13] Y. Du and Q. Wang, “Multi-angle face detection based on improved RFCN algorithm using multi-scale training,” in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, Apr. 2021, pp. 319–322, doi: 10.1109/ICSP51882.2021.9408676.

[14] J. Yu, M. Wu, C. Li, and S. Zhu, “A street view image privacy detection and protection method based on mask-RCNN,” in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Dec. 2020, pp. 2184–2188, doi: 10.1109/ITAIC49862.2020.9338847.

[15] Y.-Q. Wang, “An analysis of the viola-jones face detection algorithm,” *Image Processing On Line*, vol. 4, pp. 128–148, Jun. 2014, doi: 10.5201/ipol.2014.104.

[16] H. S. Dadi and G. K. M. Pillutla, “Improved face recognition rate using HOG features and SVM classifier,” *IOSR Journal of Electronics and Communication Engineering*, vol. 11, no. 04, pp. 34–44, Apr. 2016, doi: 10.9790/2834-1104013444.

[17] H.-G. Chung and E.-S. Kim, “Improved recognition of far objects by using DPM method in curving-effective integral imaging,” *The Journal of Korea Information and Communications Society*, vol. 37, no. 2A, pp. 128–134, Feb. 2012, doi: 10.7840/KICS.2012.37A.2.128.

[18] R. Vaillant, “Original approach for the localisation of objects in images,” *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 141, no. 4, pp. 26–30, 1994, doi: 10.1049/ip-vis:19941301.

[19] Y. Li, G. Wang, L. Lin, and H. Chang, “A deep joint learning approach for age invariant face verification,” in *CCF Chinese Conference on Computer Vision*, 2015, pp. 296–305.

[20] J. Liu, C. Fang, and C. Wu, “A fusion face recognition approach based on 7-Layer deep learning neural network,” *Journal of Electrical and Computer Engineering*, vol. 2016, pp. 1–7, 2016, doi: 10.1155/2016/8637260.

[21] H. El Khiyari and H. Wechsler, “Face recognition across time lapse using convolutional neural networks,” *Journal of Information Security*, vol. 07, no. 03, pp. 141–151, 2016, doi: 10.4236/jis.2016.73010.

[22] P. K B and M. J, “Design and evaluation of a real-time face recognition system using convolutional neural networks,” *Procedia Computer Science*, vol. 171, pp. 1651–1659, 2020, doi: 10.1016/j.procs.2020.04.177.

[23] M. Nimbarte and K. K. Bhoyar, “Biased face patching approach for age invariant face recognition using convolutional neural network,” *International Journal of Intelligent Systems Technologies and Applications*, vol. 19, no. 2, pp. 103–124, 2020, doi: 10.1504/IJISTA.2020.107216.

[24] R. Ranjan, V. M. Patel, and R. Chellappa, “HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, Jan. 2019, doi: 10.1109/TPAMI.2017.2781233.

[25] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Convolutional channel features,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 82–90, doi: 10.1109/ICCV.2015.18.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[27] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[28] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networkss,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[29] J. Ji, X. Lu, M. Luo, M. Yin, Q. Miao, and X. Liu, “Parallel fully convolutional network for semantic segmentation,” *IEEE Access*, vol. 9, pp. 673–682, 2021, doi: 10.1109/ACCESS.2020.3042254.

[30] Y. Bai, W. Ma, Y. Li, L. Cao, W. Guo, and L. Yang, “Multi-scale fully convolutional network for fast face detection,” in *Procedings of the British Machine Vision Conference 2016*, 2016, pp. 51.1-51.12, doi: 10.5244/C.30.51.

[31] L. Yu, Y. Hu, X. Xie, Y. Lin, and W. Hong, “Complex-valued full convolutional neural network for SAR target classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 10, pp. 1752–1756, Oct. 2020, doi: 10.1109/LGRS.2019.2953892.

[32] T. He, Y. Liu, C. Xu, X. Zhou, Z. Hu, and J. Fan, “A fully convolutional neural network for wood defect location and identification,” *IEEE Access*, vol. 7, pp. 123453–123462, 2019, doi: 10.1109/ACCESS.2019.2937461.

[33] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, “Finding tiny faces in the wild with generative adversarial network,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 21–30, doi: 10.1109/CVPR.2018.00010.

[34] I. Goodfellow *et al.*, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[35] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, “Complex-valued convolutional neural network and its application in polarimetric SAR image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017, doi: 10.1109/TGRS.2017.2743222.

[36] L. Mou, Y. Hua, and X. X. Zhu, “Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020, doi: 10.1109/TGRS.2020.2979552.

[37] Q. Zhang, X. Chang, and S. B. Bian, “Vehicle-damage-detection segmentation algorithm based on improved mask RCNN,” *IEEE Access*, vol. 8, pp. 6997–7004, 2020, doi: 10.1109/ACCESS.2020.2964055.

[38] H. Jiang and E. Learned-Miller, “Face detection with the faster R-CNN,” in *12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, May 2017, pp. 650–657, doi: 10.1109/FG.2017.82.

[39] X. Sun, P. Wu, and S. C. H. Hoi, “Face detection using deep learning: An improved faster RCNN approach,” *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018, doi: 10.1016/j.neucom.2018.03.030.

[40] L. Liu, G. Li, Y. Xie, Y. Yu, Q. Wang, and L. Lin, “Facial landmark machines: A backbone-branches architecture with progressive representation learning,” *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2248–2262, 2019, doi: 10.1109/TMM.2019.2902096.

[41] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, “Detecting faces using region-based fully convolutional networks,” *Prepr. arXiv.1709.05256*, Sep. 2017.

[42] Q. Chen, F. Shen, Y. Ding, P. Gong, Y. Tao, and J. Wang, “Face detection using R-FCN based deformable convolutional networks,” in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2018, pp. 4165–4170, doi: 10.1109/SMC.2018.00706.

[43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Computer Vision - ECCV*, 2016, pp. 499–515.

[44] H. Wang, Z. Li, X. Ji, and Y. Wang, “Face R-CNN,” *Prepr. arXiv.1706.01061*, Jun. 2017.

[45] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” *Prepr. arXiv.1605.06409*, May 2016.

[46] C. Zhang, X. Xu, and D. Tu, “Face detection using improved faster RCNN,” *Prepr. arXiv.1802.02142*, Feb. 2018.

[47] W. Zhang, Y. Chen, W. Yang, G. Wang, J.-H. Xue, and Q. Liao, “Class-variant margin normalized softmax loss for deep face recognition,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4742–4747, Oct. 2021, doi: 10.1109/TNNLS.2020.3017528.

[48] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, “Mis-classified vector guided softmax loss for face recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12241–12248, Apr. 2020, doi: 10.1609/aaai.v34i07.6906.

[49] G. Han, C. Chen, Z. Xu, and S. Zhou, “Sphere margins softmax for face recognition,” in *2020 39th Chinese Control Conference (CCC)*, Jul. 2020, pp. 7041–7046, doi: 10.23919/CCC50068.2020.9188526.

[50] Z. Zhou, M. Zhang, J. Chen, and X. Wu, “Detection and classification of multi-magnetic targets using mask-RCNN,” *IEEE Access*, vol. 8, pp. 187202–187207, 2020, doi: 10.1109/ACCESS.2020.3030676.

[51] Y.-P. Huang, T.-H. Wang, and H. Basanta, “Using fuzzy mask R-CNN model to automatically identify tomato ripeness,” *IEEE Access*, vol. 8, pp. 207672–207682, 2020, doi: 10.1109/ACCESS.2020.3038184.

[52] S. Yang, P. Luo, C. C. Loy, and X. Tang, “Faceness-net: Face detection through deep facial part responses,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1845–1859, Aug. 2018, doi: 10.1109/TPAMI.2017.2738644.

[53] R. Brinkmann, *The art and science of digital compositing (cdrom)*, 1th ed. Morgan Kaufmann-Academic Press, 1999.

[54] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in *Computer Vision – ECCV 2016*, 2016, pp. 354–370.

[55] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, “CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection,” in *Deep Learning for Biometrics*, 2017, pp. 57–79.

[56] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, “Face detection through scale-friendly deep convolutional networks,” *Prepr. arXiv.1706.02863*, Jun. 2017.

[57] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

BIOGRAPHIES OF AUTHORS

Rahmat Budiarsa is currently pursuing his doctoral program in the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Yogyakarta, Indonesia. He took bachelor's degree (S.Kom.) in informatic engineering program, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, in 2019 and Master (M.Cs.) in Computer Science and Electronics Department, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Yogyakarta, Indonesia in 2020 (PMDSU Master and Ph.D. Program). His research areas of interest include machine learning, artificial intelligence, and deep learning. He can be contacted at email: rahmat.budiarsa09@mail.ugm.ac.id.

Retantyo Wardoyo is a lecturer and a researcher at the Department of Computer Science, Universitas Gadjah Mada. He obtained his bachelor's degree from Mathematics in Universitas Gadjah Mada, Indonesia. He obtained his master's degree in Computer Science at the University of Manchester, UK, and His doctoral degree from Computation at the University of Manchester Institute of Sciences and Technology, UK. His research interests include intelligent systems, reasoning systems, expert systems, fuzzy systems, vision systems, group DSS and Clinical DSS, medical computing and computational intelligence. He can be contacted at email: rw@ugm.ac.id.

Aina Musdholifah is a lecturer and a researcher at the Department of Computer Science, Universitas Gadjah Mada. She obtained bachelor's degree from Computer Science, in Universitas Gadjah Mada, Indonesia. She obtained master's degree in Computer Science at the Universitas Gadjah Mada, Indonesia, and Her doctoral degree from Computer Science Universiti Teknologi Malaysia, Malaysia. Her research interests include genetics algorithm, fuzzy logic, bioinformatics, softcomputing, and machine learning. She can be contacted at email: aina_m@ugm.ac.id.