Ratio-and-Scale-Aware YOLO for Pedestrian Detection

Wei-Yen Hsu and Wen-Yen Lin

***Abstract*— Current deep learning methods seldom consider the effects of small pedestrian ratios and considerable differences in the aspect ratio of input images, which results in low pedestrian detection performance. This study proposes the ratio-and-scale-aware YOLO (RSA-YOLO) method to solve the aforementioned problems. The following procedure is adopted in this method. First, ratio-aware mechanisms are introduced to dynamically adjust the input layer length and width hyperparameters of YOLOv3, thereby solving the problem of considerable differences in the aspect ratio. Second, intelligent splits are used to automatically and appropriately divide the original images into two local images. Ratio-aware YOLO (RA-YOLO) is iteratively performed on the two local images. Because the original and local images produce low- and high-resolution pedestrian detection information after RA-YOLO, respectively, this study proposes new scale-aware mechanisms in which multiresolution fusion is used to solve the problem of misdetection of remarkably small pedestrians in images. The experimental results indicate that the proposed method produces favorable results for images with extremely small objects and those with considerable differences in the aspect ratio. Compared with the original YOLOs (i.e., YOLOv2 and YOLOv3) and several state-of-the-art approaches, the proposed method demonstrated a superior performance for the VOC 2012 comp4, INRIA, and ETH databases in terms of the average precision, intersection over union, and lowest log-average miss rate.**

***Index Terms*— Multiresolution fusion, pedestrian detection, ratio-aware, scale-aware.**

I. Introduction

PEDESTRIAN detection has been a long-standing problem in the field of computer vision. Therefore, many studies have attempted to solve this problem [1]. Many scholars have been working on this field in the past [56], [61], [62]. In recent years, due to developments in deep learning and convolutional neural network (CNN) imaging and audio processing, an increasing number of studies are focusing on pedestrian detection [2], [55], [56], [58], [60]. Pedestrian detection is a key objective in object detection. It aims to predict the bounding boxes of all pedestrians in images and has attracted widespread attention in the field of computer vision because of its increasing importance in applications such as self-driving cars, personnel re-identification, video surveillance, and robotics [3].

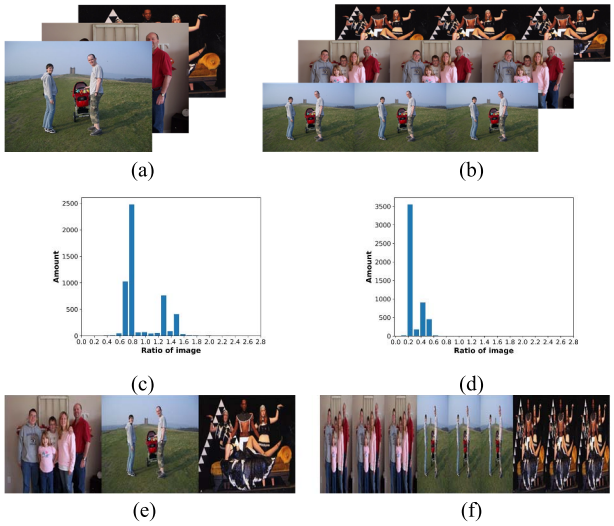

Fig. 1. (a) Sample images; (b) image aspect ratio distribution in the VOC2012 test sets; (c), (d) images resized by the detector prior to inputting the images; (e) images whose widths were three times their lengths (i.e., accurate panoramic images); and (f) image aspect ratio distribution in panoramic images, which differed considerably from that in images with an aspect ratio of 1.

However, many studies have applied deep learning methods to pedestrian, object detection, vehicle license plate location and person re-identification [4]–[10], [57], [59] without considering the effects of small pedestrian ratios and considerable differences in the aspect ratios of input images. Images with considerable differences in the aspect ratio are generally compressed to fixed lengths and widths to ensure consistent dimensions of the final output image. Although the state-of-the-art approaches generally work well in pedestrians with large size where they are near the camera, their performance becomes considerably worse when dealing with small-sized ones [54]. In addition, because aspect ratios are not considered when handling images, the detection results are often unfavorable, particularly for images with small pedestrian ratios or considerable differences in the aspect ratio (e.g., panoramic images), as displayed in Fig. 1. For images with considerable width, the use of pedestrian detection approaches causes the aspect ratios of pedestrians in resized images to become markedly distorted. To solve this problem, this study used YOLOv3 [9], which is an object detector with a fully

Manuscript received March 31, 2020; revised July 7, 2020 and October 13, 2020; accepted November 12, 2020. Date of publication November 26, 2020; date of current version December 8, 2020. This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST105-2410-H-194-059-MY3 and Grant MOST108-2410-H-194-088-MY3. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Charith Abhayaratne. *(Corresponding author: Wei-Yen Hsu.)*

Wei-Yen Hsu is with the Department of Information Management, National Chung Cheng University, Chiayi 62102, Taiwan, also with the Advanced Institute of Manufacturing With High-Tech Innovations, National Chung Cheng University, Chiayi 62102, Taiwan, and also with the Center for Innovative Research on Aging Society (CIRAS), National Chung Cheng University, Chiayi 62102, Taiwan (e-mail: shenswy@gmail.com; shenswy@mis.ccu.edu.tw).

Wen-Yen Lin is with the Department of Information Management, National Chung Cheng University, Chiayi 62102, Taiwan (e-mail: po1234263@gmail.com).

Digital Object Identifier 10.1109/TIP.2020.3039574

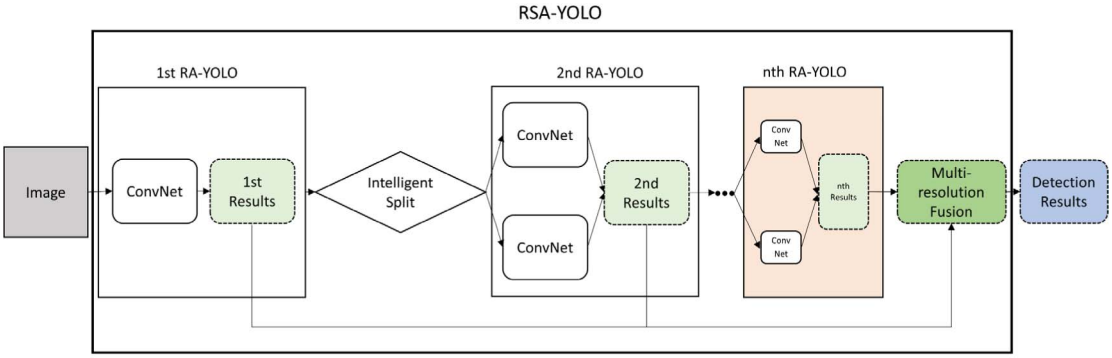

Fig. 2. Detection mechanism of RSA-YOLO. The first RA-YOLO block was used to obtain low-resolution PD info from the original images. Then, intelligent splits were used to cut local images. The next n – 1 RA-YOLO blocks were used to obtain high-resolution PD info from the local images. Finally, multiresolution fusion was used to integrate the PD info obtained from the images with various resolutions.

convolutional network (FCN) structure, to introduce image length and width information to networks. YOLOv3 was used because in the FCN structure, only the output dimensions of the last filters must be fixed. Thus, ratio-aware mechanisms can be incorporated initially to enable networks to use image length and width information more effectively, which allows original deep learning pedestrian detection approaches to demonstrate improved performance in all scenarios. Detecting small pedestrian ratios in images is difficult; therefore, most deep learning methods [6], [8]–[10] adopt a multiscaling method to improve detection results. However, using the multiscaling method when aspect ratios are imbalanced increases noise, thus creating subsequent problems. By effectively introducing ratio-aware mechanisms, original deep learning pedestrian detection approaches can demonstrate superior performance. Furthermore, multiresolution fusion can be used for scale-aware pedestrian detection without deforming or distorting pedestrian images.

To solve the abovementioned problems caused by small pedestrian ratios and considerable differences in the aspect ratio, this study developed the ratio-and-scale-aware YOLO (RSA-YOLO), which is based on YOLOv3 [9] but features several improvements. The RSA-YOLO method consists of two major components: ratio-aware YOLO (RA-YOLO) (that combines ratio-aware mechanisms and the original YOLOv3 [9]) and multiresolution fusion. RA-YOLO dynamically adjusts the hyperparameters of individual images to produce images with the most suitable hyperparameters, thereby solving the problem faced by other deep learning methods (i.e., fixed hyperparameters that cannot be adjusted freely). Multiresolution fusion effectively integrates pedestrian detection information (PD info) of various scales outputted by the RA-YOLO, thereby achieving a scale-aware effect. Furthermore, intelligent splits were applied to cut local images from original images and to help RA-YOLO capture PD info from images of various scales. Fig. 2 illustrates a simple diagram of the detection mechanism used in this study.

Overall, this study makes the following contributions. First, a novel RSA-YOLO model for pedestrian detection is proposed by incorporating a ratio-aware mechanism, intelligent splits, and a scale-aware (multiresolution fusion) mechanism into a unified architecture. Second, ratio-aware mechanisms are proposed to integrate information about the aspect ratio of images into the deep learning framework to enhance the pedestrian detection performance of images with considerable differences in the aspect ratio. Third, intelligent splits are proposed to suitably and iteratively cut images to help RA-YOLO obtain the PD info from images of various resolutions. Fourth, the original non-maximum suppression (NMS) [28] is to filter out the bounding boxes with high classification confidence from several candidate boxes and performed in single scale. Multiresolution fusion is proposed to obtain the multi-scale pedestrian information (i.e. multi-scale NMS) of an image through intelligent splits and thus achieve scale-aware effects that greatly improve the detection performance for all pedestrian ratios. Finally, RSA-YOLO is experimentally demonstrated to outperform the original YOLOs (i.e., YOLOv2 and YOLOv3) and several state-of-the-art approaches. RSA-YOLO exhibits optimal performance for the VOC 2012 comp4 [11], INRIA [12], and ETH [13] databases and for the VOC 2012 test sets (contain images whose width is three times the length; similar to panoramic images).

II. Related Work

A. Machine-Learning-Based Approaches

Early pedestrian detection approaches generally used handcrafted features together with classifier methods to detect pedestrians. For example, in 2005, Dalal proposed histogram of gradients + support vector machine (HOG + SVM)-based pedestrian detection algorithms at the International Conference on Computer Vision and Pattern Recognition [12]. The many pedestrian detection algorithms that have been proposed since 2005 are essentially extensions of [12], which makes Dalal's method a critical milestone in pedestrian detection. The HOG + AdaBoost-based pedestrian detection algorithm, which is a classic pedestrian detection algorithm, is also an extension of [12]. However, because the algorithms presented in [12] involve a substantial amount of calculation, using them in real-world applications is difficult. Thus, researchers have referred to Jones and Viola's classifier designs for human face detection [14] and have applied the AdaBoost classifier cascading method in pedestrian detection to dramatically increase the overall detection speed. Although HOG features have been applied with some success in several pedestrian detection algorithms, these features can only reflect pedestrians' shape and edge information. These features do not collect

information regarding the physical appearance of pedestrians, which makes the problem of pedestrian occlusion difficult to solve. To solve this problem, researchers have proposed the use of integral channel features (ICFs) [15]. ICFs [15] are used in combination with AdaBoost [16] to efficiently calculate and capture diverse image information, compute it in real time, and maintain high accuracy.

Some researchers have proposed using a deformable parts model [17] that divides pedestrians into various components (e.g., heads, shoulders, and torsos) for detection and later combines the detection results. This model is similar to the HOG and normally uses SVM and AdaBoost as classifiers. Although the aforementioned classic machine-learning-based approaches have demonstrated favorable performance, many problems remain unsolved. Dollar, Wojek, Schiele, and Perona (2012) evaluated several commonly used pedestrian detection algorithms and pointed out their advantages and disadvantages [27].

B. Deep-Learning-Based Approaches

Deep-learning-based approaches, particularly CNN object detection, have facilitated major breakthroughs in image detection. Since 2012, when Hinton's research team won the ImageNet image classification competition by using deep learning [18], many deep-learning-based approaches have been proposed for object detection [3]–[10], [19]–[23]. Classic deep learning object detection methods can be divided into two types according to their detection models: one-stage and two-stage detection models.

The two-stage model divides detection tasks into two stages. The first stage involves extracting several region proposals from images, and the second stage involves running classification networks on the region proposals to identify object categories in each region. Thus, the two-stage model is also called the "region-based method." The first deep learning method to adopt the two-stage model was the region-based CNN (R-CNN) [4]. The R-CNN has a major drawback. Its operations are time-consuming mainly because they are performed separately on each region proposal [4]. Furthermore, the fact that a large portion of the region proposals overlap makes the R-CNN inefficient. Therefore, fast R-CNN [5] was subsequently introduced. In this approach, base convolutional operations are shared, and all pictures are scanned on the basic network before being transferred to the R-CNN for calculation. This method considerably reduces the calculation time. Girshick (2015) noted that region proposals obtained from other methods are generated from conventional sliding windows and selective searches [24]. Because these algorithms for detection frame generation all require considerable calculation time, Ren *et al.* (2015) [6] used a region proposal network (RPN) to directly generate detection frames. This method considerably improved the generation speeds of detection frames and became the fundamental operation of the two-stage method. RPN networks replaced selective search algorithms to enable end-to-end completion of the detection task in neural networks. The inside–outside net framework (ION) for detecting objects in context with skip pooling and recurrent neural networks [25] was proposed based on [5] to introduce the concepts of outside net and inside net, where outside and inside refer to the exterior and interior of the region of interest (ROI), respectively. The outside net is mainly used to extract contextual information, and the inside net is used to achieve multiscaling fusion. Various scales were used for object detection. ROI pooling was performed on the shallow convolution layers of feature maps, which enabled the ION to demonstrate favorable detection ability for relatively small objects in images and to achieve optimal performance. Although faster R-CNN [6] exhibits a remarkable object detection performance, its detection accuracy is compromised by ROI-wise subnetworks after using the RPN. Dai *et al.* [31] identified this problem and attempted to find a suitable solution. They successfully overcame the contradictions between translation invariance in image classification and translation variance in object detection, enhanced the detection accuracy, and increased the detection speed by using position-sensitive score maps.

In the one-stage model, the region proposal extraction process is not used and prediction results are generated directly from images. The first deep learning method to adopt the one-stage model was the YOLO method [7], which transforms object detection into a unified end-to-end regression problem. The main advantage of one-stage detection methods such as YOLO [7] is their high speed. However, they demonstrate a low detection performance for small or overlapping objects. Thus, Redmon and Farhadi developed various improved methods [8], [9], one of which [9] was used in the present study. Other well-known one-stage algorithms include the single shot multibox detector (SSD) [10], which has an accuracy similar to that of two-stage algorithms but a substantially higher speed. Therefore, SSD [10] can be considered a groundbreaking one-stage method. Most subsequent one-stage algorithms are based on this method. The deconvolutional single shot detector (DSSD) [22] includes the advantages of [10] but replaces backbone networks and adds contextual information to strengthen the detection ability for small objects. It uses the deconvolution layer and skip connection to enable feature maps on shallow layers to offer superior representation.

However, previous methods usually resized images with various aspect ratios to a fixed aspect ratio. Thus, images with considerable differences in the aspect ratio did not generally demonstrate favorable detection results. Accordingly, this study introduced a simple and effective framework that used a ratio-aware mechanism to transfer length and width information to networks, which allowed the networks to make the most appropriate adjustments to each image. In addition, PD info of varying scales was combined to achieve the scale-aware effect and produce the final results. This method enabled networks to demonstrate a favorable detection performance for images with varying aspect ratios and pedestrians of varying scales.

III. Ratio-and-Scale-Aware YOLO (RSA-YOLO)

A. Overview of the Proposed Model

The proposed RSA-YOLO framework is a new framework based on YOLOv3 [9]. RSA-YOLO considers the image

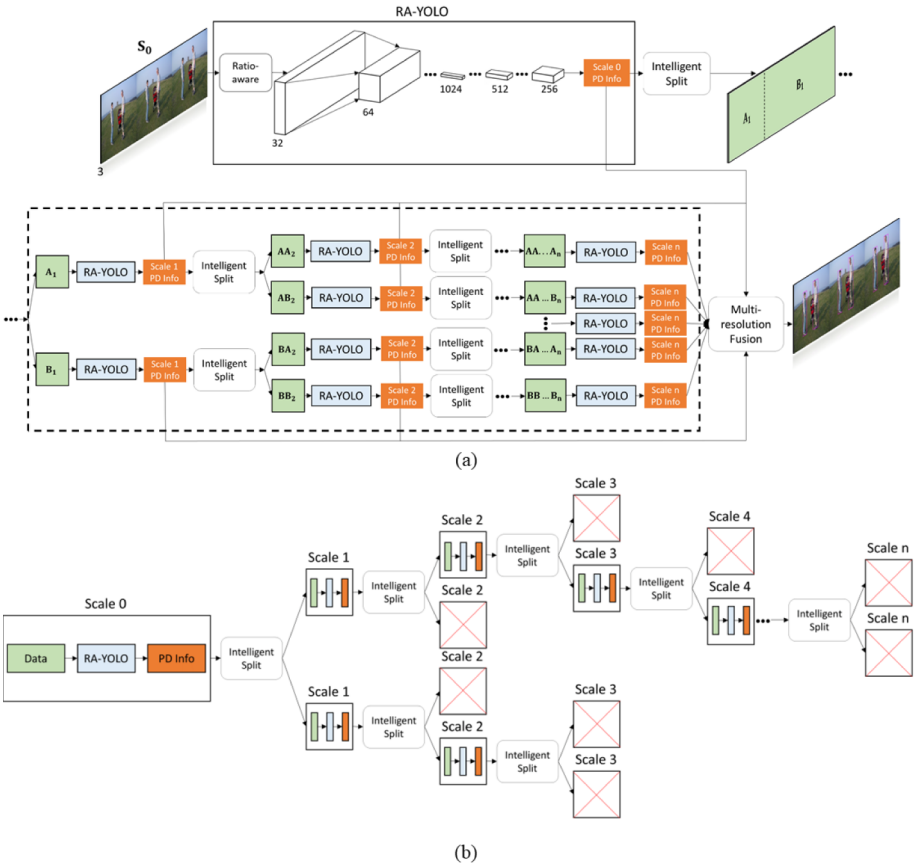

Fig. 3. (a) Complete RSA-YOLO framework. The whole image is first inputted to RA-YOLO to obtain the PD information of scale 0, and then fed into the Intelligent Split layer to divide into two sub-images (i.e. local images) with different sizes. The PD info of local images at different scales (scale 1, 2,…, and n) is then obtained iteratively. Finally, the PD info of various scales is integrated using multiresolution fusion obtain the final results that achieve the scale-aware effect; (b) Schematics of the Intelligent Splits. The data in scale 0 represents the original image, and the PD info of scale 0 (initial PD info) is then obtained through RA-YOLO. Subsequently, intelligent splits are performed to segment the images of previous stage into two local images/data of the next stage. The operations of intelligent splits are performed iteratively until the termination condition is met.

aspect ratio information and multiresolution PD info in images. Each network layer has its own PD info based on that of the upper layer, which results in high PD info resolution. In this study, PD info with varying resolutions was integrated to obtain the detection results.

B. Architecture of RSA-YOLO

Fig. 3(a) displays a detailed description of the network architecture of RSA-YOLO. RSA-YOLO uses ratio-aware mechanisms to automatically adjust the hyperparameters for optimal suitability in the images. In this study, the RA-YOLO model was used for this purpose. The original image $\mathbf{S_0}$ was processed using the first RA-YOLO, from which its PD info was obtained. Next, intelligent splits were used to divide this image into local images that could have different sizes. The PD info of the local images was then obtained similarly. The number of iterations of intelligent splits differed among images. Because the first RA-YOLO detected the entire image, image details were less recognizable and the obtained PD info had low resolution. The PD info with such resolution was referred to as Scale 0 PD info. Subsequent iterations produced increasingly small images with more detailed PD info. Fig. 3(a) indicates that the Scale PD info of subsequent layers had increased resolutions. Finally, the PD info of various scales was integrated using multiresolution fusion to achieve the scale-aware effect and the results were outputted. By integrating the PD info of multiple scales, the RSA-YOLO framework can accurately frame pedestrians of various scales in the given images.

C. Ratio-Aware Mechanism

Because images have different aspect ratios, forcing them to have the same aspect ratio deforms and distorts the pedestrians in them, resulting in subsequent detection errors or failure (Fig. 1). Not all images share the same original aspect ratio or have a ratio close to 1:1. Therefore, this study introduced ratio-aware mechanisms in the RSA-YOLO framework to enable the aspect ratio information of images to be incorporated into the neural network framework. Ratio-aware mechanisms enabled the dynamic adjustment of the aspect ratio parameters to match the parameters with those required by current images. When performing detection, networks require some of their hyperparameters (e.g., the hyperparameters of the input layer length and width) to be preset. In the following equations, h and w represent the hyperparameters of the length and width specified by the input layers, respectively, whereas H and W represent the length and width of the input images, respectively. After defining H and W of the images,

h and w can be obtained using Eqs. (1) and (2).The constant γ in YOLOv3 [9] can only take certain values because all convolution layers in [9] are multiples of 32. Therefore, γ is generally 32 or a multiple of 32. Moreover, L is the hyperparameter ceiling. In YOLOv3 [9], the standard L is 416, which indicates that input images are compressed into $L \times L$, and hyperparameters can be adjusted upward according to the requirements of the hardware device. In this study, L was an adjustable hyperparameter.

When $\mathrm{H} \leq \mathrm{W}$,

$$w = \begin{cases} \gamma, & if\ \mathrm{W} < \gamma \\ R\left(\frac{\mathrm{W}}{\gamma}\right) * \gamma, & if\ \gamma \leq \mathrm{W} < L \\ L, & if\ \mathrm{W} \geq L \end{cases}$$

$$h = \begin{cases} \gamma, & if\ \mathrm{W} < \gamma \\ \max\left(R\left(\frac{w * \frac{\mathrm{H}}{\mathrm{W}}}{\gamma}\right) * \gamma, \gamma\right), & if\ \gamma \leq \mathrm{W} < L \\ \max\left(R\left(\frac{w * \frac{\mathrm{H}}{\mathrm{W}}}{\gamma}\right) * \gamma, \gamma\right), & if\ \mathrm{W} \geq L \end{cases} \quad (1)$$

And when $\mathrm{H} > \mathrm{W}$

$$h = \begin{cases} \gamma, & if\ \mathrm{H} < \gamma \\ R\left(\frac{\mathrm{H}}{\gamma}\right) * \gamma, & if\ \gamma \leq \mathrm{H} < L \\ L, & if\ \mathrm{H} \geq L \end{cases}$$

$$w = \begin{cases} \gamma, & if\ \mathrm{H} < \gamma \\ \max\left(R\left(\frac{h * \frac{\mathrm{W}}{\mathrm{H}}}{\gamma}\right) * \gamma, \gamma\right), & if\ \gamma \leq \mathrm{H} < L \\ \max\left(R\left(\frac{h * \frac{\mathrm{W}}{\mathrm{H}}}{\gamma}\right) * \gamma, \gamma\right), & if\ \mathrm{H} \geq L \end{cases} \quad (2)$$

where

$$R(x) = \left\lfloor x + \frac{1}{2} \right\rfloor \quad (3)$$

The hyperparameters w and h calculated using Eqs. (1)-(3) are then immediately used to the input layers of the YOLO algorithm to obtain the pedestrian detection information of current scale. The entire process above is referred to as RA-YOLO in this study.

D. Intelligent Splits

In this study, RSA-YOLO integrated the PD info of images with various resolutions to produce the final results. Both low-resolution PD info (i.e., Scale 0 PD info) that contained the information of the entire images and high-resolution PD info (i.e., Scale 1–n PD info) that originated from intelligent splits were obtained. In this study, intelligent splits were performed using a two-step procedure. The first step involved filtering out outliers of the pedestrian bounding boxes, and the second step involved separating the input images into two subimages. Detection networks that rely only on the PD info of full images may potentially remove overlapping pedestrians during NMS operations, a problem addressed in [26]. Furthermore, pedestrian frames that are exceptionally large (i.e., exceeding certain ratios specified by the images) are likely to generate erroneous results. To solve this problem, this study used the outlier decision method to filter out the frames of relatively large pedestrian bounding boxes. The frames filtered out were wider or taller than those of most pedestrian bounding boxes. To prevent pedestrian bounding boxes from being filtered out, this study added the filtering condition of classification confidence to reduce the occurrence of misjudgment. In Eq. (4), this study used the Z-score method to determine whether the frames were outliers.

$$\mathrm{Outlier} = \begin{cases} \mathrm{TRUE}, & if\ (Z_h \geq \alpha\ or\ Z_w \geq \alpha)\ and\ C < \beta \\ \mathrm{FALSE}, & o.w. \end{cases} \quad (4)$$

$$\mathrm{Z-score} = \frac{x - \mu}{\sigma} \quad (5)$$

where Z_h is the height score (Z-score) of all the pedestrian bounding boxes, Z_w is the width score (Z-score) of all the pedestrian bounding boxes, C is the classification confidence of the pedestrian bounding boxes, and β is an adjustable parameter used as a reference value for decision-making. The Z_h and Z_w can be obtained using Eq. (5), where μ and σ represent the average and standard deviation, respectively. The samples of Z-scores usually range between 5 and 6 (or between -5 and -6 when converted). This study set $\mathrm{Z} \geq \alpha$ as the outlier condition. Moreover, α was an adjustable parameter, where a large α indicated a high outlier tolerance. $Z \leq \alpha$ may also be considered an outlier. The results of searching a public database revealed that the chance of having relatively small pedestrians in images was relatively high, and this finding is supported by [27]. Thus, this study retained pedestrian bounding boxes with $\mathrm{Z} \leq \alpha$. Frames identified as TRUE were filtered out during this process, and only the PD info of the remaining frames was used to segment subsequent images.

The input images were divided into two subimages. Before intelligent splits were performed, some original images exhibited shrinking due to RA-YOLO. Although these images still contained the PD info of the entire image, the decreased area data resulted in a small amount of image information, which lowered the resolution of the PD info. However, performing intelligent splits enabled RA-YOLO to be applied to the local images, thereby diminishing the decrease in area data. Although the PD info was obtained only from local images, the total amount of data of the entire region was relatively high, which resulted in the production of high-resolution PD info. After performing initial filtering in the previous stage, calculations were performed for the remaining pedestrian bounding boxes, and a method similar to that used to solve optimization problems was used to calculate the most suitable split locations for the images. Subsequently, the images were segmented to form local images, which were transferred to RA-YOLO to obtain PD info with relatively high resolution. Iterations were performed multiple times to produce PD info with high resolution, thereby enabling the capture of PD info with relatively small scales.

Algorithm 1 RSA-YOLO

Input: Scale 0 PD info *S0PI,* the upper-limit hyperparameter L, and the number of iterations Ωout
Output: the set of final detected bounding boxes D
Begin
$FSPI \leftarrow \emptyset$
$count \leftarrow 0$
$FSPI \leftarrow FSPI \cup \{ S0PI \}$
$SPI \leftarrow S0PI$
while $count < \Omega out$ **do**
  $SD \leftarrow IS (SPI)$
  if $max (width (SD), height (SD)) > L$ **then**
    $SPI \leftarrow RA\text{-}YOLO (SD)$
    $FSPI \leftarrow FSPI \cup \{ SPI \}$
    $count \leftarrow count + 1$
  end if
end while
$D \leftarrow MrFusion (FSPI)$
return D
End

E. Scale-Aware Effect With Multiresolution Fusion

The last stage of the RSA-YOLO operation was multiresolution fusion, which involved integrating all the Scale PD info obtained during the operation. When extracting the PD info of images, this study used an iterative mechanism to determine the optimal detection method for different images. Scale 0–n PD info obtained from the previous RA-YOLO was highly detailed. The specific operation methods are displayed in Fig. 3(b) and presented in Algorithm 1. After obtaining all the Scale PD info, this study employed a method similar to NMS [28] to merge the PD info of various scales, thereby achieving the scale-aware effect. Because the PD info of various resolutions is merged, this process is referred to as "multiresolution fusion." The fusion results were used as the final detection results and outputted by RSA-YOLO.

IV. Experiments

In addition to conducting evaluations using the VOC 2012 comp4 [11], INRIA [12], and ETH [13] databases of PASCAL VOC [29], this study stringed 5138 images with annotations in the PASCAL VOC 2012 test set into horizontal images with triple the original width. These images were used to display panoramic images labeled with the correct answers, and the test results of the panoramic images were used for evaluating the validity of the RSA-YOLO model. The goal was to prove the superior performance of RSA-YOLO for images with special aspect ratios.

A. Data Sets

1) *PASCAL VOC 2012:* PASCAL VOC 2012 included 16135 images, of which 5138 images were labeled with the correct answers. In images labeled with the correct answers, 7330 pedestrians were identified (an average of 1.43 pedestrians in each image). For images not labeled with the correct answers, the results were uploaded to the PASCAL VOC evaluation server comp4 to determine the average precision (AP). The dataset that had be uploaded to the PASCAL VOC evaluation server comp4 to obtain the test results was called VOC 2012 comp4, and the 5138 images with the correct answers were called VOC 2012.

TABLE I
Comparison of the AP in the Original YOLO Approach With That in Other Versions for VOC 2012 Comp4. Here, 07++12: 07 Trainval + 07 Test + 12 Trainval

Method	Data	Network	AP
YOLOv2-544[8]	07++12	DARKNET-19	81.3
YOLOv3-320	COCO TRAINVAL	DARKNET-53	85.1
YOLOv3-416	COCO TRAINVAL	DARKNET-53	87.4
YOLOv3-608	COCO TRAINVAL	DARKNET-53	87.3
YOLOv3-SPP	COCO TRAINVAL	DARKNET-53	82.7
RSA-YOLO	COCO TRAINVAL	DARKNET-53	**88.5**

2) *INRIA:* The INRIA Person dataset [12] was divided into the training and test sets. The training set comprised 614 positive samples and 1218 negative samples, and the test set included 288 images. This study only evaluated the 288 images in the test set and followed the standard evaluation metric. The log miss rate was averaged over the false positive per image in $[10^{-2}, 10^{0}]$ (denoted as MR).

3) *ETH:* The test sets of the ETH Pedestrian dataset [13] included 1804 images from three video clips. We followed the standard evaluation metric. The log MR was averaged over the false positive per image in $[10^{-2},10^{0}]$ (denoted as MR).

B. Implementation Details

This study applied ratio-aware mechanisms to implement the adaptive adjustment hyperparameters each time they entered the RA-YOLO model. The required parameter γ in the ratio-aware mechanism was set as 32, and L was set as 416. The upper-limit hyperparameters of L were set as 416 mainly because these hyperparameters are set as 416 in the standard configuration of YOLOv3 [9]. Thus, this study could be compared with the original YOLO approach in terms of performing calculations with similar magnitudes.

In the intelligent split method, outlier screening is performed in the first step for parameter calibration. The tolerance parameter for the α outlier was set at 4, and the β classification confidence threshold was set at 0.5. A neutral perspective was adopted to conduct the final screening for the pedestrian bounding boxes with $\alpha \geq 4$. Intelligent splits were performed iteratively to obtain the manual stop threshold Ω_{out} required for extracting high-resolution PD info. In this study, they are performed two iterations. Because the proposed algorithm was mainly based on methods similar to the NMS algorithm [28] for merging various types of resolutions to achieve the scale-aware effect, the NMS threshold parameters were adjusted in the final multiresolution fusion step. This parameter was set as 0.5 after calibration.

C. Comparison With State-of-the-Art Methods

1) VOC 2012 Comp4: After the results for the VOC 2012 comp4 [11] database were obtained, the original YOLO (YOLOv2 and YOLOv3) was compared with several state-of-the-art approaches. Then, the proposed method was compared with the original YOLO approach and several state-of-the-art approaches, as presented in Tables I and II. This

TABLE II

COMPARISON OF THE AP OF SEVERAL STATE-OF-THE-ART APPROACHES WITH THAT OF THE PROPOSED METHOD FOR VOC 2012 COMP4. HERE, 07+12: 07 TRAINVAL +12 TRAINVAL, 07+12+S: 07+12 PLUS THE SEGMENTATION LABELS, 07++12: 07 TRAINVAL +07 TEST+12 TRAINVAL, AND 07++12+COCO TRAINVAL35K: 07 TRAINVAL +07 TEST+12 TRAINVAL + COCO TRAINVAL 35K

Method	Data	Network	AP
Faster R-CNN[6]	12 TRAINVAL	VGG	67.0
Faster R-CNN[6]	07 TRAINVAL&TEST + 12TRAINVAL	VGG	70.4
Faster R-CNN[6]	07++12+COCO TRAINVAL	VGG	84.1
MLKP[30]	07++12	VGG	83.5
MLKP[30]	07++12	RESIDUAL-101	84.3
ION[25]	07+12+S	VGG	82.3
R-FCN[31]	07++12	RESIDUAL-101	84.4
SSD300[10]	07++12	VGG	82.6
SSD300[10]	07++12+COCO	VGG	85.6
SSD321[22]	TRAINVAL 07++12	RESIDUAL-101	81.5
DSSD321[22]	07++12	RESIDUAL-101	82.1
SSD512[10]	07++12	VGG	85.5
SSD512[10]	07++12+COCO	VGG	88.4
SSD513[22]	TRAINVAL 07++12	RESIDUAL-101	85.7
DSSD513[22]	07++12	RESIDUAL-101	86.4
RefineDet320[32]	07++12	VGG	78.1
RefineDet320[32]	07++12+COCO TRAINVAL	VGG	82.7
RefineDet512[32]	07++12	VGG	80.1
RefineDet512[32]	07++12+COCO TRAINVAL	VGG	85.0
RefineDet320+[32]	07++12	VGG	82.7
RefineDet320+[32]	07++12+COCO TRAINVAL	VGG	86.0
RefineDet512+[32]	07++12	VGG	83.5
RefineDet512+[32]	07++12+COCO TRAINVAL35K	VGG	86.8
RSA-YOLO	COCO TRAINVAL	DARKNET-53	**88.5**

study compared the four versions of the original YOLO approach, namely YOLOv3-320, -416, -608, and –SPP, and the previous version, namely YOLOv2-544. In Table I, RSA-YOLO demonstrated a superior test performance for the VOC 2012 comp4 database compared with the other versions of the original YOLO approach. After demonstrating the advantage of RSA-YOLO compared with the original YOLO approach, the results of RSA-YOLO were compared with those of several state-of-the-art approaches for VOC 2012 comp4, including faster R-CNN [6], SSD [10], DSSD [22], ION [25], MLKP [30], R-FCN [31], RefineDet [32]. The results revealed that RSA-YOLO achieved the same performance as SSD512 07++12+COCO trainval35k [10] and was superior to all state-of-the-art approaches.

2) VOC 2012: A total of 5138 images in VOC 2012 had the correct answers. To demonstrate the validity of RSA-YOLO for special aspect ratios, the VOC 2012 images were first

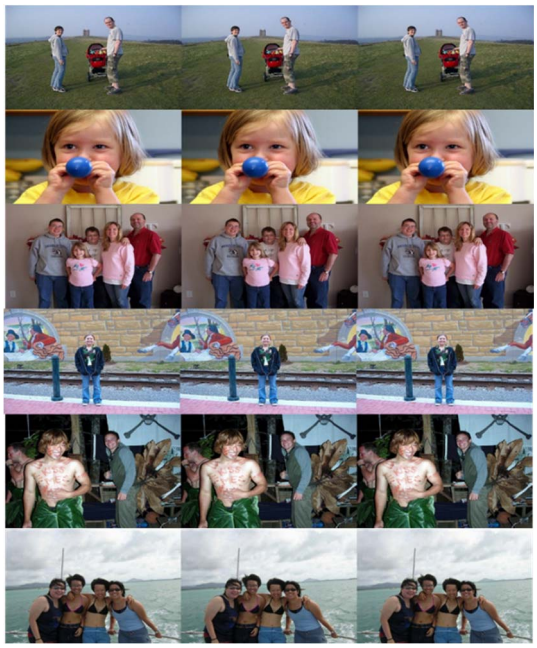

Fig. 4. Horizontally linked triple-width images in VOC 2012.

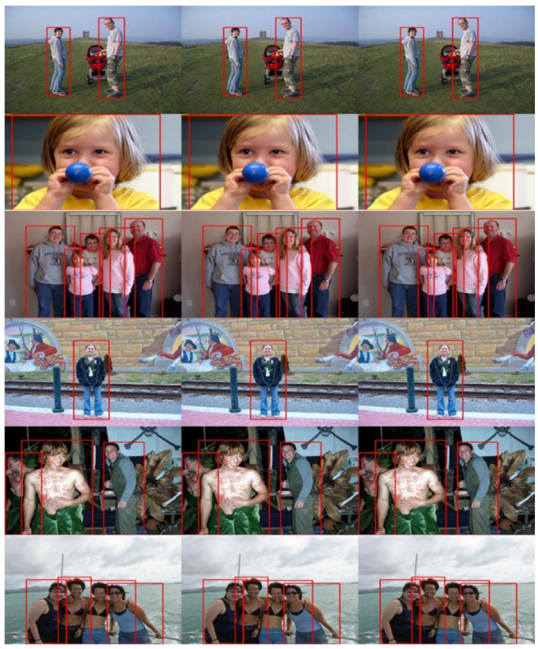

Fig. 5. Horizontally linked triple-width images with the correct answer in VOC 2012.

linked to form triple-width images in a horizontal direction and thus display panoramic images with the correct answers. The 7330 marked pedestrians in the original 5138 images therefore became $7330 \times 3 = 21990$ marked pedestrians after being linked into the triple-width images. Fig. 4 displays a few post-scaling images. In addition, to prove that the correct answers were not affected after scaling, this study presents examples of post- scaling images with the correct answers (Fig. 5). After examining numerous images, the proposed method was compared with the original YOLO approach at various versions (including 320, 416, 608, SPP), as presented

TABLE III

COMPARISON OF AP IN ORIGINAL YOLO APPROACH WITH THAT OF OTHER VERSIONS IN VOC 2012 (IMAGE WITH A WIDTH OF THREE TIMES IN THE HORIZONTAL DIRECTION)

Method	Person (AP)	PRAUC
YOLOv3-320	48.4	48.3
YOLOv3-416	54.0	54.0
YOLOv3-608	55.1	55.1
YOLOv3-SPP	52.5	52.5
RSA-YOLO	**74.2**	**74.1**

(a) IOU threshold = 0.6

Method	Person (AP)	PRAUC
YOLOv3-320	40.1	40.1
YOLOv3-416	45.8	45.8
YOLOv3-608	47.1	47.0
YOLOv3-SPP	47.2	47.2
RSA-YOLO	**70.7**	**70.7**

(b) IOU threshold = 0.7

Method	Person (AP)	PRAUC
YOLOv3-320	26.4	26.4
YOLOv3-416	31.8	31.8
YOLOv3-608	32.3	32.2
YOLOv3-SPP	36.4	36.4
RSA-YOLO	**62.3**	**62.3**

(c) IOU threshold = 0.8

Method	Person (AP)	PRAUC
YOLOv3-320	9.6	9.6
YOLOv3-416	12.9	12.8
YOLOv3-608	13.6	13.5
YOLOv3-SPP	19.5	19.5
RSA-YOLO	**40.5**	**40.4**

(d)

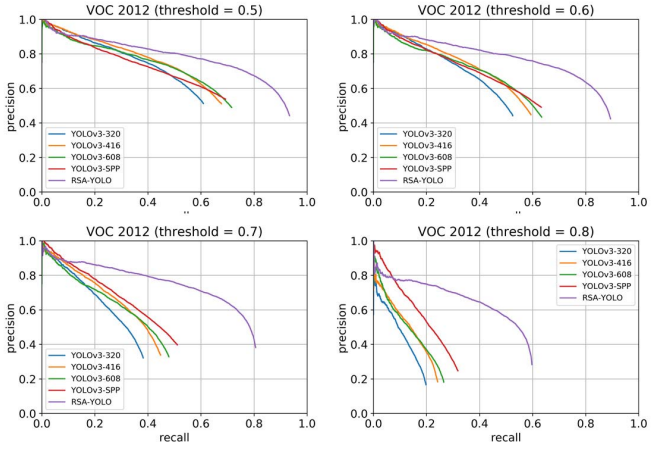

Fig. 6. Comparison of the RSA-YOLO and original YOLO approaches for various thresholds in VOC 2012 (images whose widths were three times the lengths). A cross-comparison with Table III reveals that RSA-YOLO demonstrated the most favorable performance for VOC 2012.

in Table III and Fig. 6. The experimental results indicate that the area under the precision-recall curve (PRAUC) and AP of the proposed method were considerably larger than those of the original YOLO approach (Fig. 6). Table III also reveals that RSA-YOLO provided the most favorable performance for various thresholds. In addition, to verify the performance of the proposed method, we compared it with several state-of-the-art approaches, including faster R-CNN [6], SSD [10], Mask R-CNN [19], RetinaNet [20], RefineDet [32], and RFB-Net [33], for the VOC 2012 images. The results revealed that the proposed RSA-YOLO method was superior to all

TABLE IV

COMPARISON OF THE AP OF SEVERAL STATE-OF-THE-ART APPROACHES WITH THAT OF THE PROPOSED METHOD FOR VOC 2012 (IMAGE WITH A WIDTH OF THREE TIMES IN THE HORIZONTAL DIRECTION)

Method	Data	Network	AP	PRAUC
Faster R-CNN[6]	07++12	VGG	62.9	62.9
Faster R-CNN[6]	07++12	ResNet-101	64.8	64.6
Mask R-CNN[19]	COCO TRAINVAL35K	ResNet-101-FPN	73.4	73.4
RetinaNet[20]	COCO TRAINVAL35K	ResNet-50	60.7	60.4
SSD300[10]	07++12	VGG	24.7	24.7
SSD300[10]	07++12+COCO	VGG	33.4	33.4
SSD512[10]	TRAINVAL35K	VGG	39.4	39.3
	07++12+COCO TRAINVAL35K			
RefineDet320[32]	07+12	VGG	52.4	52.1
RefineDet320[32]	07++12	VGG	54.8	54.8
RefineDet320[32]	07+12+COCO TRAINVAL35K	VGG	58.5	58.5
RefineDet320+[32]	07+12+COCO TRAINVAL35K	VGG	60.4	60.4
RefineDet320+[32]	07++12+COCO TRAINVAL35K	VGG	61.6	61.5
RefineDet512[32]	07+12	VGG	59.0	59.0
RefineDet512[32]	07++12	VGG	61.4	61.4
RefineDet512[32]	07+12+COCO TRAINVAL35K	VGG	63.3	63.3
RefineDet512+[32]	07+12+COCO TRAINVAL35K	VGG	65.4	65.3
RefineDet512+[32]	07++12+COCO TRAINVAL35K	VGG	66.0	65.9
RefineDet320[32]	COCO TRAINVAL35K	VGG	58.2	58.2
RefineDet512[32]	COCO TRAINVAL35K	VGG	63.0	63.0
RefineDet320[32]	COCO TRAINVAL35K	ResNet-101	58.8	58.7
RFBNet300[33]	07+12	VGG	26.9	26.9
RFBNet300[33]	COCO TRAINVAL35K	VGG	29.0	29.0
RFBNet512-E[33]	COCO TRAINVAL35K	VGG	32.6	32.6
RFBMobileNet300[33]	COCO TRAINVAL35K	MobileNet	23.8	23.8
RSA-YOLO	COCO TRAINVAL	DARKNET-53	**74.2**	**74.1**

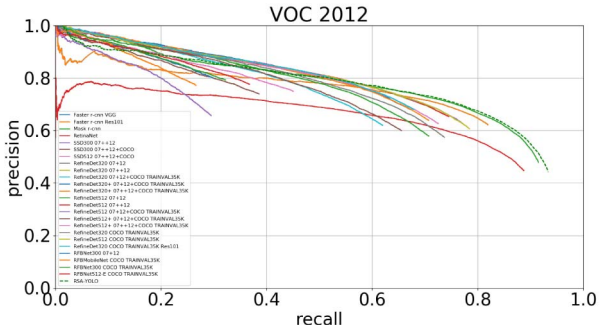

Fig. 7. Comparison of RSA-YOLO and several state-of-the-art approaches for VOC 2012 images whose widths were three times their heights. A cross-comparison with Table IV reveals that RSA-YOLO provided the most favorable performance for the triple-width VOC 2012 images.

state-of-the-art approaches, as listed in detail in Table IV. As displayed in Fig. 7, the proposed method provided superior results for VOC 2012 (images whose widths were thrice their lengths) compared with several state-of-the-art approaches. Moreover, to further validate that our method is robust to images with a considerably large aspect ratio, we tested our method for images whose widths were five and seven times their lengths and then compared it with several state-of-the-art

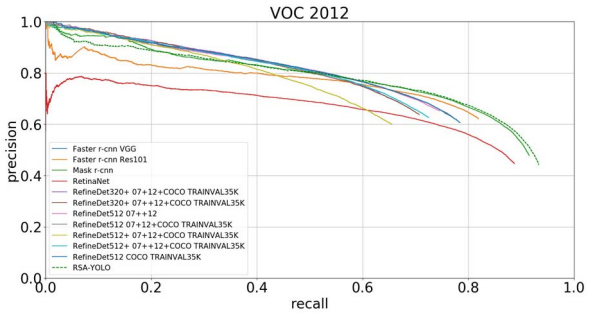

Fig. 8. Comparison of RSA-YOLO and several state-of-the-art approaches whose AP is equal to and above 60 for the VOC 2012 images whose widths were three times their lengths.

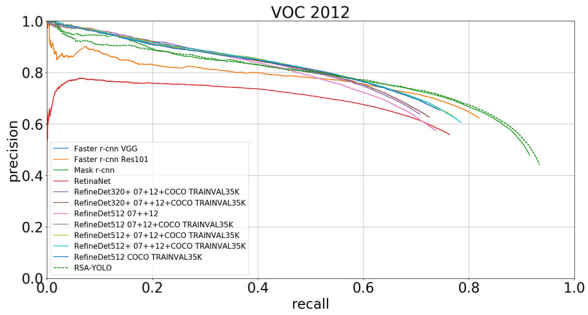

Fig. 9. Comparison of RSA-YOLO and several state-of-the-art approaches whose AP is equal to or above 60 for VOC 2012 images whose width is three times their length. A cross-comparison with Table V reveals that RSA-YOLO provided the most favorable performance for VOC 2012 images whose width was five times their lengths.

TABLE V

COMPARISON OF THE AP OF SEVERAL STATE-OF-THE-ART APPROACHES WITH THAT OF THE PROPOSED METHOD FOR VOC 2012 IMAGES WHOSE WIDTHS ARE FIVE TIMES THEIR LENGTHS. HERE, 07+12: 07 TRAINVAL+12 TRAINVAL, 07++12: 07 TRAINVAL+07 TEST+12 TRAINVAL, AND 07++12+COCO TRAINVAL35K: 07 TRAINVAL+07 TEST+12 TRAINVAL+COCO TRAINVAL35K

Method	Data	Network	AP	PRAUC
Faster R-CNN[6]	07++12	VGG	50.5	50.5
Faster R-CNN[6]	07++12	ResNet-101	55.5	55.4
Mask R-CNN[19]	COCO TRAINVAL35K	ResNet-101-FPN	67.7	67.7
RetinaNet[20]	COCO TRAINVAL35K	ResNet-50	54.8	54.5
RefineDet320+[32]	07+12+COCO TRAINVAL35K	VGG	31.8	31.8
RefineDet320+[32]	07++12+COCO TRAINVAL35K	VGG	33.6	33.5
RefineDet512[32]	07++12	VGG	34.5	34.5
RefineDet512[32]	07+12+COCO TRAINVAL35K	VGG	34.7	34.7
RefineDet512+[32]	07+12+COCO TRAINVAL35K	VGG	39.2	39.2
RefineDet512+[32]	07++12+COCO TRAINVAL35K	VGG	39.6	39.5
RefineDet512[32]	COCO TRAINVAL35K	VGG	33.9	33.9
RSA-YOLO	COCO TRAINVAL	DARKNET-53	**68.9**	**68.9**

approaches [6], [19], [20], [32] whose APs were close to that of our method for VOC 2012. The comparison results of the PRCURVE chart are illustrated in Fig. 8.

First, we performed pedestrian detection on the VOC 2012 images whose widths were five times their lengths. The results are presented in Table V and Fig. 9. Table V indicates that the AP and PRAUC of the proposed method for the VOC 2012 images whose width were five times their lengths were still larger than those of the state-of-the-art approaches. As presented in Table V and Fig. 9, several state-of-the-art approaches provided misdetected results for most pedestrians because they warped the input images to a fixed and incorrect ratio. Next, we performed pedestrian detection on images whose widths were seven times their lengths. The results are presented in Table VI and Fig. 10.

TABLE VI

COMPARISON OF THE AVERAGE PRECISION (AP) OF SEVERAL STATE-OF-THE-ART APPROACHES WITH THAT OF THE VOC 2012 (IMAGE WITH A WIDTH OF SEVEN TIMES IN THE HORIZONTAL DIRECTION). HERE, 07+12: 07 TRAINVAL+12 TRAINVAL, 07++12: 07 TRAINVAL+07 TEST+12 TRAINVAL, 07++12+COCO TRAINVAL35K: 07 TRAINVAL+07 TEST+12 TRAINVAL+COCO TRAINVAL35K

Method	Data	Network	AP	PRAUC
Faster R-CNN[6]	07++12	VGG	37.0	37.0
Faster R-CNN[6]	07++12	ResNet-101	43.1	43.1
Mask R-CNN[19]	COCO TRAINVAL35K	ResNet-101-FPN	44.1	44.1
RetinaNet[20]	COCO TRAINVAL35K	ResNet-50	46.8	46.4
RefineDet320+[32]	07+12+COCO TRAINVAL35K	VGG	15.7	15.6
RefineDet320+[32]	07++12+COCO TRAINVAL35K	VGG	16.7	16.7
RefineDet512[32]	07++12	VGG	17.8	17.8
RefineDet512[32]	07+12+COCO TRAINVAL35K	VGG	14.8	14.8
RefineDet512+[32]	07+12+COCO TRAINVAL35K	VGG	20.4	20.4
RefineDet512+[32]	07++12+COCO TRAINVAL35K	VGG	20.4	20.4
RefineDet512[32]	COCO TRAINVAL35K	VGG	14.2	14.2
RSA-YOLO	COCO TRAINVAL	DARKNET-53	**64.4**	**64.4**

Table VI indicates that the proposed method had the highest AP and PRAUC for the VOC 2012 images whose widths were seven times their lengths. The difference between the AP of the proposed and second-best approaches (RetinaNet [20]) was 17.6, whereas the difference between the PRAUC of the proposed and RetinaNet approaches was 18.0.

3) INRIA: This study assessed the proposed model using the INRIA Person dataset [12]. Fig. 11 displays the experiment results. The proposed method was compared with the original YOLO approach [9] and several state-of-the-art approaches, namely HOG [12], VJ [34], NAMC [35], SCCPriors [36], InformedHaar [37], LDCF [38], Franken [39], Roerei [40], SketchTokens [41], SpatialPooling [42], RPN + BF [43], PCN [44], and F-DNN [45]. For the INRIA Person dataset,

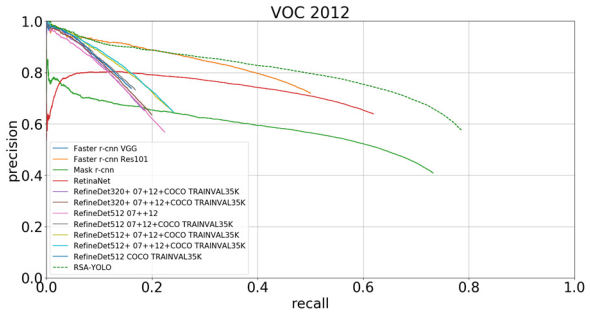

Fig. 10. Comparison of the RSA-YOLO and several state-of-the-art approaches in the VOC 2012 (image with a width of seven times in the horizontal direction). A cross-comparison with Table VI reveals that RSA-YOLO demonstrated the most favorable performance in the VOC 2012 (image with a width of seven times in the horizontal direction).

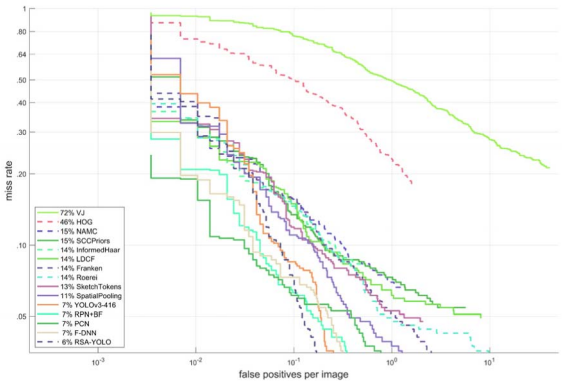

Fig. 11. RSA-YOLO had a lowest log-average miss rate of 6% in the INRIA Person data set, thus outperforming the original YOLO and other state-of-the-art approaches.

TABLE VII

COMPARISON OF THE MISS RATE OF SEVERAL STATE-OF-THE-ART RELEVANT APPROACHES WITH THAT OF THE INRIA

Method	Miss rate(%)
LDCF[38]	14
Roerei[40]	14
SpatialPooling[42]	11
LDCF+PBF[60]	13
MT-LDCF[61]	11
MCF[62]	9
NNNF[56]	10
RSA-YOLO	**6**

the proposed model demonstrated a 6% miss rate, which was superior to that of several state-of-the-art approaches [9], [43], [44], [45]. In addition to three relevant works (LDCF [38], Roerei [40], SpatialPooling [42]) from the journal, as shown in Fig. 11, four related papers (LDCF+PBF [60], MT-LDCF [61], MCF [62], and NNNF [56]) published in this journal are also a measure of comparison with our method, as listed in Table VII.

Taiana *et al.* (2013) [46] indicated that the original INRIA Person dataset provided relatively few recognizable pedestrian labels. Thus, they proposed a new annotation for the dataset in which all pedestrians with a height greater than 25 pixels were marked. This increased the number of annotated pedestrians

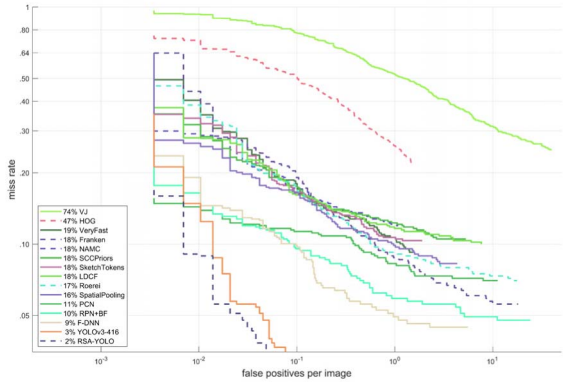

Fig. 12. RSA-YOLO had a lowest log-average miss rate of 2% in the INRIA Person data set after the use of annotated pedestrians, outperforming the original YOLO and other state-of-the-art approaches.

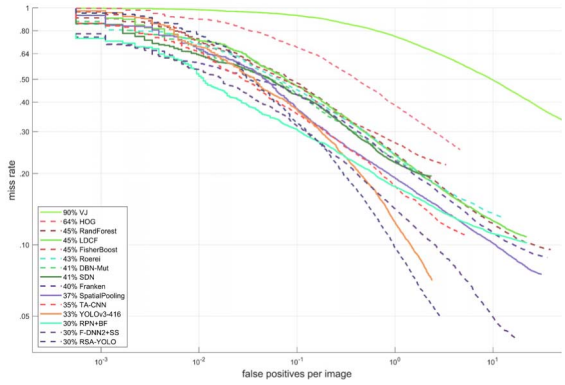

Fig. 13. RSA-YOLO had a lowest log-average miss rate of 30% in the ETH Pedestrian data set, rivaling those of several state-of-the-art approaches and outperforming that of the original YOLO approach.

from 589 to 878. These new annotated pedestrians were also used by the authors for subsequent studies. Accordingly, we performed model assessments using these annotated pedestrians. Fig. 12 illustrates the experimental results and compares the results obtained using the proposed model with those obtained using the original YOLO approach [9] and several state-of-the-art approaches (i.e., [12], [34]–[36], [38]–[45], and [47]). By using annotated pedestrians, the proposed model yielded a miss rate of 2%, which is superior to the miss rates of both the original state-of-the-art approach (9%) [43] and the original YOLO approach (3%) [9].

4) ETH: We performed assessments using the ETH Pedestrian dataset [13]. Fig. 13 displays the experiment results and compares the results obtained using the proposed model with those obtained using the original YOLO approach [9] and several state-of-the-art approaches (i.e., [12], [34], [38]–[40], [42], [43], and [48]–[52]). In the ETH Pedestrian dataset, the proposed model yielded a miss rate of 30%, which was similar to that of the original state-of-the-art approach [43], [53] and superior to that of the original YOLO approach (33%) [9].

D. Ablations Studies

This section investigates the different components of RSA-YOLO and the validity of its parameter configuration.

TABLE VIII

DETECTION RESULTS OBTAINED BY USING DIFFERENT L VALUES IN VOC2012 COMP4

Method	***L***	Person(AP)	Inference time(ms)
RSA-YOLO	320	87.4	32
RSA-YOLO	416	88.5	39
RSA-YOLO	608	88.6	64

TABLE IX

EFFECT OF USING OUTLIERS ON THE DETECTION RESULTS

Prediction from method	person (AP) Use outlier determination	
	Yes	No
RSA-YOLO	**88.5**	87.7

TABLE X

DETECTION RESULTS OBTAINED BY FUSING RSA-YOLO-SCALE 0 PD INFO, RSA-YOLO-SCALE 1–N PD INFO, OR RSA-YOLO IN VOC2012 COMP4

Prediction from method	Person (AP)
RSA-YOLO-Scale 0 PD Info	87.4
RSA-YOLO-Scale 1~n PD Info	87.6
RSA-YOLO	**88.5**

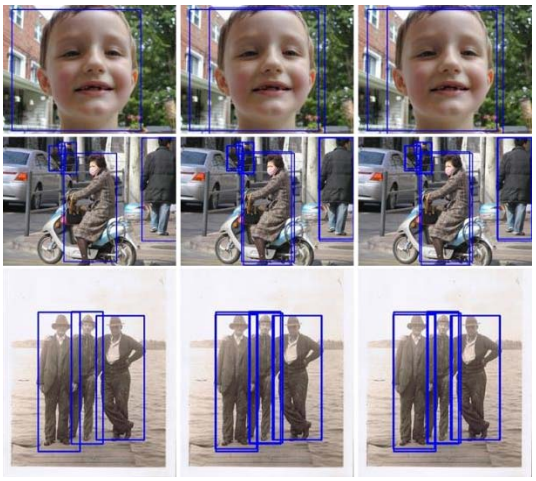

Fig. 14. Pedestrian detection results obtained using various NMS algorithms. The NMS algorithms used were (from left to right) RSA-YOLO + NMS, RSA-YOLO + soft-NMS G, and RSA-YOLO + soft-NMS L. The soft-NMS algorithm retained many unnecessary detection frames. The threshold values set for soft-NMS G and soft-NMS L were the standard values used in [26].

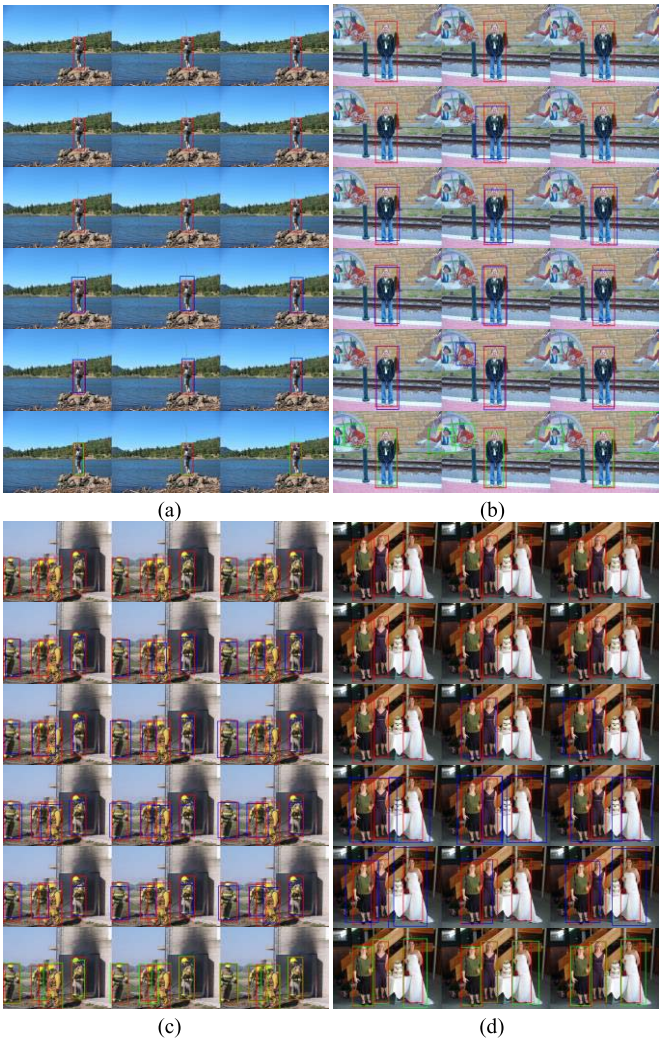

Fig. 15. Comparison of the pedestrian detection results obtained using RSA-YOLO with those obtained using the original YOLO approach. The uppermost input images present the correct answers (marked in red rectangles), whereas the remaining input images reveal the detection results obtained using YOLOv3-320, −416, −608, and -SPP (marked in blue rectangles) and RSA-YOLO (marked in green rectangles). RSA-YOLO successfully detected most pedestrians in panoramic images, whereas the original YOLO approach failed to detect the pedestrians.

Because RSA-YOLO achieved a similar accuracy for the detection performance in the VOC 2012 comp4 [11], INRIA [12], and ETH [13] datasets, the experiments described in the following subsections were performed in VOC 2012 comp4 and the differences in performance demonstrated by RSA-YOLO for various parameter settings and framework compositions were compared.

1) Sizes of L in Ratio-Aware Mechanisms: In RSA-YOLO, the upper limit of L had to be set initially. In most pedestrian detection approaches, a large L value corresponds to a high accuracy in framing pedestrians at various scales. Moreover, original images with large differences in the aspect ratio or those with a length or width less than L exhibit pedestrian deformation and distortion after image resizing, thus decreasing the detection accuracy. In this study, adjusting the size of L affected the hyperparameter ceiling after ratio-aware mechanisms were used. Nevertheless, ratio-aware mechanisms balanced themselves out between L and the image length and width. Thus, when L increased, the detection accuracy increased almost proportionally. To establish a balance between accuracy and speed, we set L as 416, which is similar to the standard of [9]. To verify the superiority of the proposed ratio-aware mechanisms, we used several L values, as listed in Table VIII (AP and inference time). Using an L value of 416 produced results similar to those in [9].

Fig. 16. Comparison of the pedestrian detection results obtained with the proposed RSA-YOLO method and those obtained with several state-of-the-art approaches whose AP was equal to or above 60 for VOC 2012 images whose widths were three times their lengths. The uppermost row displays the ground truth (marked in red rectangles), whereas the remaining rows illustrate the detection results obtained with faster R-CNN VGG [6], faster R-CNN Res101 [6], Mask R-CNN [19], RetinaNet [20], RefineDet320+07+12+COCO TRAINVAL35K [32], Refine-Det320+07++12+COCO TRAINVAL35K [32], RefineDet512 07++12 [32], RefineDet512 07+12+COCO TRAINVAL35K [32], RefineDet512+ 07+12+COCO TRAINVAL35K [32], RefineDet512+ 07++12+COCO TRAINVAL35K [32], RefineDet512 COCO TRAINVAL35K [32] (marked in blue rectangles), and RSA-YOLO (marked in green rectangles), respectively.

2) Outlier Identification: For RSA-YOLO, this study defined outliers prior to performing intelligent splits to reduce the occurrence of non-pedestrians being identified as pedestrians, which leads to unfavorable detection results. To demonstrate that defining outliers can improve detection results, this study assessed the detection results obtained when outliers were used and those obtained when outliers were not used (Table IX).

3) Intelligent Splits: This study predicted pedestrian bounding boxes using Scale 0 PD info after RA-YOLO. Moreover, PD info of local images cut using intelligent splits was obtained. Multiresolution fusion was later performed to merge the PD info (which had various resolutions) for producing the detection results. To demonstrate that intelligent splits can improve the detection results, we assessed the detection results obtained by fusing RSA-YOLO Scale 0 PD info, RSA-YOLO Scale 1–n PD info, and RSA-YOLO (Table X).

4) Multiresolution Fusion: In the final stage of RSA-YOLO, low- and high-resolution PD info was merged through multiresolution fusion to integrate the PD info of all resolutions. Thus, the scale-aware effect was achieved and the final detection results were produced. This study adopted an algorithm similar to the NMS algorithm [28] to combine images of various resolutions. Therefore, the NMS threshold had to be adjusted. This study adopted the soft-NMS [26] to integrate the final results. The use of soft-NMS L in VOC2012 comp4 produced the highest AP (88.7), and the use of soft-NMS G produced an AP similar to that of the original NMS method (88.5). However, the use of RSA-YOLO + NMS produced an AP almost identical to that produced by RSA-YOLO despite using fewer frames. Fig. 14 indicates that the soft-NMS algorithm retained overlapping pedestrian images. However, it diminished the suppression effect and required more time for calculation compared with the original NMS algorithm. Therefore, NMS [28] was selected as the

TABLE XI
EFFECT OF VARIOUS NMS ALGORITHMS ON THE DETECTION RESULTS GENERATED BY RSA-YOLO IN VOC2012 COMP4

Prediction from method	Person (AP)	# Boxes
RSA-YOLO + NMS	88.5	13206
RSA-YOLO + soft-NMS G	88.5	23339
RSA-YOLO + soft-NMS L	**88.7**	23339

algorithm for merging images of various resolutions, as indicated in Table X.

5) Schematics of the Detection Results: Fig. 15 illustrates the comparison of the detection results obtained using RSA-YOLO with those obtained using the original YOLO approach (i.e., YOLOv3). The results confirmed the superiority of RSA-YOLO in detecting pedestrians in panoramic images. In Figs. 16 (a)–(d), the uppermost input images provide the correct answers (marked in red rectangles), whereas the remaining input images present the detection results obtained using YOLOv3-320, -416, -608, and -SPP (marked in blue rectangles) and RSA-YOLO (marked in green rectangles).

The results indicated that RSA-YOLO successfully detected many pedestrians whom YOLOv3 failed to detect, particularly those with small scales or those that were overlapping. Fig. 15(a) indicates the importance of ratio-aware mechanisms. Because of the pedestrians in Fig. 15(a) are relatively slender, compressing them into squares using common resizing methods made them even more slender, which increased the difficulty of pedestrian detection. In Fig. 16, we present the pedestrian detection results of RSA-YOLO and several state-of-the-art approaches. The figure only displays the results of RSA-YOLO and approaches whose AP was equal to or above 60 for VOC 2012 images whose widths were three times their lengths. The results indicated that the proposed RSA-YOLO method successfully detected most pedestrians in images, whereas detection failed when using the state-of-the-art approaches.

V. CONCLUSION AND FUTURE WORK

This study proposed an RSA-YOLO model that incorporated aspect ratio information with deep learning and effectively integrated the PD info of various image resolutions to solve the problems associated with images that have various aspect ratios and small pedestrian ratios. The experimental results indicated that the RSA-YOLO generated superior detection results in common pedestrian datasets as well as for images with considerable differences in the aspect ratio. In future work, we will extend the proposed mechanisms to other different CNN structures, including one-stage and two-stage detectors, and the detection of various objects by readjusting the appropriate parameters or refining the corresponding mechanisms with less prior knowledge.

REFERENCES

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[2] H. Li, Z. Wu, and J. Zhang, "Pedestrian detection based on deep learning model," in *Proc. 9th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2016, pp. 796–800.

[3] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[4] 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.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[7] 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.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jul. 2017, pp. 7263–7271.

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

[10] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[11] M. Everingham and J. Winn, "The PASCAL visual object classes challenge 2012 (VOC2012) development kit," Pattern Anal., Stat. Model. Comput. Learn., Tech. Rep. 8, 2011. [Online]. Available: https://scholar.google.com/scholar?hl=zh-TW&as_sdt=0%2C5&q=The+PASCAL+visual+object+classes+challenge+2012+%28VOC2012%29+development+kit&btnG=host.robots.ox.ac.uk/pascal/VOC/voc2012/devkit_doc.pdf

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[13] W.-Y. Hsu, "Automatic pedestrian detection in partially occluded single image," *Integr. Comput.-Aided Eng.*, vol. 25, no. 4, pp. 369–379, Sep. 2018.

[14] M. Jones and P. Viola, "Fast multi-view face detection," Mitsubishi Electr. Res. Lab, Tech. Rep. TR-20003-96, 2003, vol. 3, no. 14, p. 1–10. [Online]. Available: https://scholar.google.com/scholar?hl=zh-TW&as_sdt=0%2C5&q=Fast+multi-view+face+detection&btnG=

[15] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. BMVC*, 2009, pp. 1–11.

[16] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Jpn. Soc. Artif. Intell.*, vol. 14, nos. 771–780, p. 1612, 1999.

[17] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

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

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

[21] T.-Y. Lin, P. Dollár, 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.

[22] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[24] 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.

[25] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[26] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5561–5569.

[27] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[28] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, 2006, pp. 850–855.
[29] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[30] H. Wang, Q. Wang, M. Gao, P. Li, and W. Zuo, "Multi-scale location-aware kernel representation for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1248–1257.
[31] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.
[32] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.
[33] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 385–400.
[34] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2002.
[35] C. Ţoca, M. Ciuc, and C. Pătraşcu, "Normalized autobinomial Markov channels for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 175.1–175.13.
[36] Y. Yang, Z. Wang, and F. Wu, "Exploring prior knowledge for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.
[37] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 947–954.
[38] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 424–432.
[39] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1505–1512.
[40] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3666–3673.
[41] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3158–3165.
[42] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 546–561.
[43] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–457.
[44] S. Wang, J. Cheng, H. Liu, and M. Tang, "PCN: Part and context information for pedestrian detection with CNNs," 2018, *arXiv:1804.04483*. [Online]. Available: http://arxiv.org/abs/1804.04483
[45] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.
[46] M. Taiana, J. C. Nascimento, and A. Bernardino, "An improved labelling for the INRIA person data set for pedestrian detection," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.* Berlin, Germany: Springer, 2013, pp. 286–295.
[47] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2903–2910.
[48] J. Marin, D. Vazquez, A. M. Lopez, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2592–2599.
[49] C. Shen, P. Wang, S. Paisitkriangkrai, and A. van den Hengel, "Training effective node classifiers for cascade classification," *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 326–347, Jul. 2013.
[50] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3222–3229.
[51] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 899–906.
[52] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5079–5087.
[53] X. Du, M. El-Khamy, V. I. Morariu, J. Lee, and L. Davis, "Fused deep neural networks for efficient pedestrian detection," 2018, *arXiv:1805.08688*. [Online]. Available: http://arxiv.org/abs/1805.08688
[54] X. Zhang, L. Cheng, B. Li, and H.-M. Hu, "Too far to see? Not really!—Pedestrian detection with scale-aware localization policy," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3703–3715, Aug. 2018.
[55] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. CVPR*, Jun. 2013, pp. 3626–3633.
[56] J. Cao, Y. Pang, and X. Li, "Pedestrian detection inspired by appearance constancy and shape symmetry," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5538–5551, Dec. 2016.
[57] W. Min, X. Li, Q. Wang, Q. Zeng, and Y. Liao, "New approach to vehicle license plate location based on new model YOLO-L and plate pre-identification," *IET Image Process.*, vol. 13, no. 7, pp. 1041–1049, May 2019.
[58] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.
[59] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.
[60] Y. Zhao, Z. Yuan, and B. Chen, "Accurate pedestrian detection by human pose regression," *IEEE Trans. Image Process.*, vol. 29, pp. 1591–1605, 2020.
[61] C. Zhu and Y. Peng, "A boosted multi-task model for pedestrian detection with occlusion handling," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5619–5629, Dec. 2015.
[62] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3210–3220, Jul. 2017.

Wei-Yen Hsu received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, in 2008. He is currently a Professor with the Department of Information Management, National Chung Cheng University, Chiayi, Taiwan. His research interests include image processing, pattern recognition, and machine learning. He has been a Founding Member of the Brain-Computer Interface (BCI) Society since 2015. He was a recipient of the Young Scholar Award of Taipei Medical University in 2011. He was also a recipient of the Young Scholar Award and the Outstanding Research Award of National Chung Cheng University in 2013 and 2019, respectively. He is an Academic Editor of *Medicine* journal and an Associate Editor of *BMC Medical Informatics and Decision Making* journal.

Wen-Yen Lin received the master's degree from the Department of Information Management, National Chung Cheng University, Chiayi, Taiwan, in 2020.