Adaptive Fusion of Multi-Scale YOLO for Pedestrian Detection

WEI-YEN HSU,1,2,3 (Member, IEEE), AND WEN-YEN LIN1

1 Department of Information Management, National Chung Cheng University, Chiayi 62102, Taiwan
2 Advanced Institute of Manufacturing With High-Tech Innovations, National Chung Cheng University, Chiayi 62102, Taiwan
3 Center for Innovative Research on Aging Society (CI-RAS), National Chung Cheng University, Chiayi 62102, Taiwan

Corresponding author: Wei-Yen Hsu (shenswy@gmail.com)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST108-2410-H-194-088-MY3 and Grant MOST110-2221-E-194-027-MY3, and in part by the Center for Innovative Research on Aging Society (CIRAS), National Chung Cheng University, Taiwan, through the CYCH-CCU Joint Research Program under Grant 110-02076-01.

ABSTRACT
Although pedestrian detection technology is constantly improving, pedestrian detection remains challenging because of the uncertainty and diversity of pedestrians in different scales and of occluded pedestrian modes. This study followed the common framework of single-shot object detection and proposed a divide-and-rule method to solve the aforementioned problems. The proposed model introduced a segmentation function that can split pedestrians who do not overlap in one image into two subimages. By using a network architecture, multiresolution adaptive fusion was performed on the output of all images and subimages to generate the final detection result. This study conducted an extensive evaluation of several challenging pedestrian detection data sets and finally proved the effectiveness of the proposed model.

In particular, the proposed model achieved the most advanced performance on data sets from Visual Object Classes 2012 (VOC 2012), the French Institute for Research in Computer Science and Automation, and the Swiss Federal Institute of Technology in Zurich and obtained the most competitive results in a triple-width VOC 2012 experiment carefully designed by the present study.

INDEX TERMS
Pedestrian detection, multi-scale YOLO, adaptive fusion.

I. INTRODUCTION
Images of pedestrians at different scales, occlusions, and special aspect ratios often occur in practical applications, and this is a massive challenge for pedestrian detection. Most state-of-the-art pedestrian detection methods are dedicated to solving these problems [1], [2]. Although these methods detect pedestrians at different scales and with different occlusion problems excellently, they typically do not perform well when images have special aspect ratios.

Resizing the input image is the first step that all detection algorithms must perform during detection. However, for images with special aspect ratios, after being resized, most pedestrians in such images are greatly compressed or even disappear. The detector may be misled by the information in its detection window, which results in missing pedestrians that should be detected. As displayed in FIGURE 1, (a) is the sample image, (b) shows images with widths three times their original length, and (c–d) present images resized to a fixed size. The figure clearly indicates that pedestrians in the image may be excessively compressed or even disappear after the image is resized. Moreover, pedestrians with a similar color to the background are likely to be missed during detection. Therefore, this study explored methods of resolving such pedestrian detection problems.

Although deep learning has been widely used in pedestrian detection [3], [4], image recognition [5], place recognition [6], dimension reduction [7], object detection [8]–[10], [49]–[52], and pose recovery [11], [12], detection algorithms for special aspect ratio images remain lacking. More specifically, a multi-label learning approach [3] is proposed to jointly learn part detectors to capture partial occlusion patterns to handle the conditions of partial occlusions. Multiple built-in sub-networks [4] which detect pedestrians with scales from disjoint ranges are introduced. A HDWE model [5] is devised by integrating sparse constraints and an improved RELU operator to address click feature prediction from visual features. A SPE-VLAD layer and the WT-loss layer [6] are introduced by integrating with the VGG-16 network or ResNet-18 network to form a novel end-to-end deep neural
network that can be easily trained via the standard backpropagation method.

An unsupervised deep-learning framework LDFA [7] is presented for dimension reduction. An object detection framework based on ensemble deep learning [8] is proposed. CoCNN [9] is proposed to fuse color and disparity features from low to high layers in a unified deep model. Gated CNN [10] is introduced to integrate multiple convolutional layers for object detection. A novel pose recovery method [11] is presented using non-linear mapping with multi-layered deep neural network. A method is proposed to recover 3-D human poses from silhouettes [12].

To resolve the problems of uncertainty and diversity of pedestrians in special aspect ratios and of occluded pedestrian modes, this study used a single-shot object detector. You only look once (YOLO) v4 [13], as a base and initially imported the image length and width information into the network to alleviate the problem of image distortion after resizing. This study selected [13] as the base because in the network structure, only the output size of the last filter must be fixed. Therefore, the length and width information of the image can be easily imported and merged into the network for it to use information more effectively, thereby allowing the original deep learning pedestrian detection method to exhibit improved performance in all cases. In addition, because detecting small-scale pedestrians and occlusion problems in an image is difficult, most deep learning involves multiscale modeling to improve the detection results, including Faster R-CNN [14], Yolov3 [15], and SSD [16]. Although these algorithms achieve performance growth after multiscale modeling is added, this problem cannot be solved when encountering images with special aspect ratios, which cause subsequent detection problems. Therefore, the segmentation function and multiresolution adaptive fusion proposed by this study were used, and an image was detected through the divide-and-rule method, which can solve the problem of pedestrian detection for images with special aspect ratios and prevent the deformation and distortion of pedestrians in the image, which lead to poor detection performance.

The central idea of the segmentation function proposed in this study involves dividing the disjoint part of the pedestrian in the image into two subimages to separately extract features from the image. Subsequently, multiresolution adaptive fusion was used to complement and fuse the information given by the sample image and subimages to achieve the divide-and-rule pedestrian detection method proposed in this study.

The segmentation function and multiresolution adaptive fusion has the two following disadvantages: (1) The computational cost increases linearly with the number of subimages generated using the segmentation function; and (2) additional fusion time is required. However, controlling the overall time cost of segmentation to no more than two times the original method [13] can reduce difficulties encountered during pedestrian detection without excessively increasing time costs.

The contributions of this study are as follows. First, by integrating image length and width information into the network, the problem of image distortion after resizing was initially alleviated. Second, a segmentation function was applied in the neural network architecture to effectively segment the sample image to extract complementary information from multiple subimages. Third, the multiresolution adaptive fusion mechanism was incorporated into a unified architecture. This can integrate complementary information from multiple subimages to create a more satisfactory fusion method than by using any single image for pedestrian detection.

Finally, experiments revealed that the method proposed in this study outperforms the original YOLO and several state-of-the-art methods because it shows optimal performance for VOC 2012 comp4 [17], French Institute for Research in Computer Science and Automation (INRIA) [18], and Swiss Federal Institute of Technology in Zurich (ETH) [19] databases and the VOC 2012 test set (the triple-width image test set elaborately designed in this study).

The rest of the study is organized as follows. In Section 2, relevant work and literature on the latest methods are discussed. In Section 3, technical details of the model proposed are introduced. In Sections 4 and 5, numerous experiments we conducted are described, and the proposed method is applied to existing pedestrian data sets to prove its feasibility and performance. Finally, Section 6 concludes this study and provides a direction for future studies.

II. RELATED WORK

In the past few years, researchers from various countries have sought to improve pedestrian detection [20]. In recent years, people have become interested in neural networks because...
of the progress of neural network architecture and hardware equipment, and deep learning has also been popularized. In addition, the research team of Hinton et al. (2012) won the ImageNet competition by using deep learning [21], and this promoted the concept of deep learning and influenced many researchers to focus on this area to conduct many impressive studies, including Mask R-CNN [22], DSSD [23], Inside-Outside Net [24], convolutional learning [25], outline features neural network [26], R-CNN [27], Fast R-CNN [28], YOLOv1 [29], and RetinaNet [30].

A major drawback in [27] is that its detection operation is extremely time-consuming because classification for each regional proposal must be performed separately. In addition, because most region proposals overlap, the efficiency of [27] is even lower. Therefore, subsequent researchers introduced [14] and [28]. Faster regions with convolutional neural networks (R-CNN) [14] uses region proposal networks (RPNs) to generate a detection framework, which greatly improves the detection framework generation speed, and this has allowed it to become a cornerstone of R-CNN-type algorithms. Moreover, RPNs replaced most time-consuming and non-neural network selective searches in [27], [28], thus achieving end-to-end neural network detection.

Based on [28], some scholars have proposed a new structure through a different mindset: an inside–outside net framework [24]. The study proposed a neural network architecture that includes internal and external networks, which use skip connections and recurrent neural networks to detect objects in a given context in which the inside–outside net framework represents the internal and external region of interest. The outside net is mainly used to extract contextual information, and the inside net is used to achieve multiscale fusion. This operation enables the inside–outside net framework to demonstrate excellent detection capabilities for small objects in images.

The authors of YOLO [29] proposed a new idea. Unlike previous algorithms that use R-CNN [27], YOLO does not use region proposal extraction but directly generates prediction results from images, and this is known as the one-stage approach. The study [29] was the first to apply a one-stage model in deep learning, which converted object detection into a unified end-to-end regression problem. The difference between YOLO [29] and R-CNN [27] algorithms is typically a one-stage algorithm with higher detection speed. However, this typically results in low detection performance for small or overlapping objects. Therefore, the original authors of YOLO, Redmon and Farhadi, developed various improved methods for these problems [13], [15], which this study adapted [13] as a new method for improvement. Another one-stage algorithm is known as single-shot detection (SSD) [16], and most subsequent one-stage algorithms are based on this method and [29]. For instance, [23] is a variant of [16].

These methods typically adjust images directly with various aspect ratios to a fixed size before the next detection action can be performed. Therefore, images with different aspect ratios generally do not show satisfactory results. This study introduced a simple and effective framework that uses the transmission of length and width information to the network, thereby allowing the network to make the most appropriate adjustments to each image. In addition, the framework combines feature information provided by the sample image and segmented subimages to obtain the final result through multiresolution adaptive fusion. This method enables the network to have detection images of pedestrians with different aspect ratios and scales excellently.

III. PROPOSED NETWORK ARCHITECTURE

A. OVERVIEW OF THE PROPOSED MODEL

The network architecture proposed in this study is a new framework based on YOLOv4 [13]. It first considers the width and height information in the image, resizes the information accordingly, and sends it to the detection network. Subsequently, the subimages are obtained by segmenting the nonoverlapping area and sending it to the network for the feature information of the entire image and subimages to be acquired. Finally, multiresolution adaptive fusion was used to achieve the final fusion and output. In this study, feature information with different resolutions was integrated to obtain test results.

B. ARCHITECTURE OF THE PROPOSED MODEL

FIGURE 2 briefly describes the network architecture of the proposed model. This model first used the function resize with ratio to automatically adjust the network’s hyperparameters, which allowed the image to retain its original aspect ratio after resizing to adapt to images with various aspect ratios. Subsequently, the image was sent to the convolutional neural network (CNN) model to obtain a feature map of the entire image. The feature map was mapped back to the sample image to use this information to identify segmented points where the pedestrians do not intersect on the sample image. Moreover, the segmentation function was used to divide the sample image into subimages with different sizes. Similar to the steps used to obtain feature maps of an entire image, the subimages were sent to the CNN model to obtain the feature map of the subimages. Finally, multiresolution adaptive fusion was used to integrate the pedestrian detection information from each image and output the final result.

By integrating pedestrian detection information with multiple resolutions, the model proposed in this study can accurately frame pedestrians of various ratios in a given image.

C. RATIO RESIZING

Because each image has a different aspect ratio, to compress them to the same aspect ratio is by nature not an optimal method because this may cause the pedestrians in the image to be deformed or distorted, which results in subsequent detection errors or failures. As described in FIGURE 1, some subsequent errors may be caused by images with very different aspect ratios. Therefore, to alleviate problems from resizing, this study simply added aspect ratio information
In addition, to avoid filtering out frames that may contain pedestrians, this study simply added a screening condition for classification confidence to reduce the occurrence of false positives. This study determined the currently selected pedestrian frame is an outlier, outlier determination (OD), on the basis of the z-score,

\[
OD = \begin{cases} 
    \text{True}, & Z_h \geq a \text{ and } C \leq b \\
    \text{True}, & Z_w \geq a \text{ and } C \leq b \\
    \text{False}, & Z_h < a \text{ and } Z_w < a 
\end{cases}
\]

where \(Z_h\) and \(Z_w\) are the z-scores of the height and width of all frames in the input image, and \(C\) is the probability of whether the frame contains a pedestrian. In this equation, an adjustable hyperparameter of \(b\) was used as the probability determination value of whether the frame contains a pedestrian, and the value after converted the z-score of each pedestrian frame generally falls within \(-5\) to \(+5\). Hence, this study set the z-score benchmark value of \(Z \geq a\) as the outlier, where \(a\) is also an adjustable hyperparameter. A larger \(a\) indicates a higher tolerance of the algorithm to outliers. Although \(Z \leq a\) can be regarded as an outlier, studies have reasonably claimed that the probability of small-scale pedestrians appearing in images is greater. This fact has been described in [20], and thus this study retained the pedestrian frame of \(Z \leq a\). Pedestrian frames that were not filtered out were passed to the second stage for segmentation into subimages.

Segmentation can be simply seen as a process of segmenting the sample image. This study asserted that when the entire image shows poor detection performance, the divide-and-rule method can be applied, which involves detecting parts of a sample image and combining them. When the sample image before segmentation is put into the CNN model, the entire image shrinks more. Although a feature map of the entire image contains the complete pedestrian information, small regional data produce less obvious amounts of information for the entire image, which can only represent a low-resolution feature map. However, for segmented images in the CNN model, the shrinking of subimages occurs less. Although subimages only contain a small part of the entire image, the amount of information is richer than that of the entire image, thus it is a high-resolution feature map. This proposed model identifies the most suitable image segmentation point from the first-stage detection results. Ideally, the point in the center of the image not containing pedestrians was segmented into two images with the same length and width.

In this study, the method of identifying the optimal segmentation point is similar to the concept of support vector machine largest margins, which must abide by two principles: (1) being closest to the center and (2) being farthest from the left or right object frame or boundary. A specific description is presented in FIGURE 3. First, the algorithm identifies all points that excluded candidate frames (blue frame) on the horizontal axis of the input image according to the feature map and uses these points as the candidate segmentation points from the sample image during resizing to allow the network to adjust accordingly and retain the original ratio after resizing. Originally, [13] supported the flexibility needed to manually adjust the hyperparameters of the input layer, and this study transformed the manually adjusted hyperparameter process to produce a hyperparameter that could be adaptively adjusted through some simple calculations.

**D. SEGMENTATION FUNCTION**

The segmentation function proposed in this study mainly includes two steps. The first step filters the pedestrian frame in the feature map of the entire image, and the second step segments the sample image into subimages. The main consideration for adding the segmentation function to the model is that if the network only relies on the information given by the feature map of the entire image for detection, it may suppress some overlapping pedestrians due to nonmaximum suppression (NMS) [31] operations. This problem has been mentioned by other researchers [32]. Moreover, pedestrians may be excessively framed. For example, if the ratio exceeds a certain ratio of the image, the second step of this method may not identify the parts that do not intersect with the pedestrian, or the frame itself may produce the wrong result.

Therefore, this study first added some filtering conditions to filter out large pedestrian frames in the first step.
points. Subsequently, among these candidate segmentation points, only the point (black point) closest to the center of the image (red line) is retained, which corresponds to the first principle of this algorithm. Then, this black point is extended to the short side of the image until it reaches the nearest object candidate frame or image boundary as a candidate continuous point (purple point). Finally, the midpoint of these candidate continuous points (purple points) are connected in a vertical direction as the optimal segmentation line (green line) for this algorithm, which corresponds to the second principle of this algorithm. The purpose of selecting the midpoint with the largest margin as the segmentation point is to avoid a situation in which the object is segmented in the previous detection result. If no point in the entire image contains pedestrians, the segmentation stage is skipped, and the filtered pedestrian frame is directly output.

E. MULTI-RESOLUTION ADAPTIVE FUSION
The final stage in this study was multiresolution adaptive fusion. This fuses all feature maps encountered in the network, including the low-resolution feature map of the entire image and the high-resolution feature map of the subimages. This study adopted a method similar to that in [31] to fuse these feature maps from various images to achieve the final fusion step.

The threshold of NMS must be adjusted. In addition to adjusting the threshold, this study considered various NMS methods to fuse the proposed model. Apart from conventional NMS [31], this study used soft-NMS [32], and its detailed experimental content is described in the ablation analysis in Section 5. Finally, because all feature maps that appeared in the entire network process were fused, this model combined the contextual pedestrian detection information, which has a high pedestrian detection rate of various scales in the image.

IV. EXPERIMENTS
This chapter first selected three data sets commonly used to evaluate pedestrian detection performance in experiments: those from PASCAL VOC 2012 comp4 [17], INRIA [18], and ETH [19]. In addition, this study expanded 5138 images with ground truth in the PASCAL VOC 2012 test set horizontally to three times the length of images, making them images with special aspect ratios. These special images could be used to verify that the model proposed in this study can achieve excellent performance with various aspect ratio images.

A. EXPERIMENT PREPARATION
The experiments in this study were all based on the new network architecture developed by [13], as described in FIGURE 2, in which common objects in context (COCO) trainval [33] was used for training according to the standard configuration of [13]. We use stochastic gradient descent to train our model. The batch size of training is 16, accumulate is 4, momentum is 0.92, weight decay is 0.00045. The experiment is based on the 64-bit operating system Windows 10 Professional. The deep learning framework is Pytorch, and the GPU is a GeForce RTX 2080ti. The experiments use the official indicators on the ground, i.e., Average Precision (AP) and Log Average Miss Rate which are defined as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (2)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (3)
\]

where True Positives (TP) is the number of positive samples correctly identified as positive samples, False Positives (FP) is the number of negatives samples wrongly identified as positive samples, False Negatives (FN) is the number of positive samples wrongly identified as negatives samples.

The data set of pascal voc2012 [17] used in this experiment followed the AP calculation method described above, whereas the data sets of INRIA [18] and ETH [19] followed the standard evaluation metric. The log MR was averaged over the false positive per image in \([10^{-2}, 10^{3}]\) (denoted as MR).

B. IMPLEMENTATION DETAILS
Certain hyperparameters in the proposed model framework must be set manually, and these hyperparameters were created under careful experimentation to ensure that this method can obtain satisfactory results in actual environment.

The first stage of the segmentation function was outlier filtering. After experimental adjustment, this study set outlier tolerance parameter \(a\) to 4 and classification confidence threshold \(b\) to 0.5. A more neutral angle was used to filter the pedestrian frame with \(a \geq 4\). Finally, during multiresolution fusion, adaptive fusion, because the algorithm is mainly based on a variant of the NMS [31] algorithm, the NMS threshold must...
be adjusted to achieve the fusion effect, and this parameter was set to 0.5.

C. DATA SETS
This study briefly describes the open data set used in TABLE 1, and the detailed descriptions are as follows.

| Dataset name | Description |
|--------------|-------------|
| Pascal voc 2012 | [link](http://host.robots.ox.ac.uk/pascal/VOC/voc2012/) |
| voc 2012 comp4: 10992 images need to be uploaded to the evaluation server. |
| voc 2012: 7330 pedestrian annotation in 5138 |
| INRIA | [link](http://pascal.inrialpes.fr/data/human/) |
| Training data: 2416 pedestrian annotations in 614 and 1218 non-human images |
| Testing data: 1132 pedestrian annotations in 288 and 453 non-human images |
| ETH | [link](http://www.vision.ee.ethz.ch/~aess/dataset/) |
| Training data: 3780 pedestrian annotations in 853 images |
| Testing data: 14,167 pedestrian annotations in 1804 images |

1) PASCAL VOC 2012
Pascal voc 2012 [17] includes 16,135 images, of which 5138 images have ground truth. The data set is divided into 20 types, described as follows: Person, Bird, Cat, Cow, Dog, Horse, Sheep, Aeroplane, Bicycle, Boat, Bus, Car, Motorcycle, Train, Bottle, Chair, Dining table, Potted plant, Sofa, and TV/Monitor. This study concerns pedestrian detection, so the evaluation index of AP mentioned in the experiment only evaluated pedestrian accuracy. A total of 7330 pedestrians were included in the 5138 images with ground truth; thus, each image has an average of 1.43 pedestrians. For the remaining images that were not labeled with correct answers, these images were not validated by the algorithm, the results must be uploaded to the PASCAL VOC evaluation server to obtain AP. The partial data set that must be uploaded to the evaluation server to obtain the AP is known as voc 2012 comp4, and the 5138 images with ground truth are known as voc 2012.

2) INRIA
The INRIA Person dataset [18] 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.

3) ETH
The ETHEth Pedestrian dataset [19] was divided into the training and test sets. The training set comprised 853 images, and the testing set included 1804 images from three video clip which contains 14,167 pedestrians.

D. COMPARISON WITH STATE-OF-THE-ART

1) VOC 2012 COMP4
This study compared the proposed model with the original versions of YOLO (YOLOv3 [15] and YOLOv4 [13]). After VOC 2012 comp4 [17] obtained results superior to the original method, the model proposed in this study was compared to current state-of-the-art methods, including Faster R-CNN [14], SSD [16], deconvolutional SSD [23], ION [24], R-fcn [34], and RefineDet [35]. The detailed results are listed in TABLEs 2 and 3, and the model

| Method | Data | Network | AP |
|--------|------|---------|----|
| Faster R-CNN [14] | 12 TRAINVAL | VGG | 67.0 |
| Faster R-CNN [14] | 07 TRAINVAL & TEST + 12 TRAINVAL | VGG | 70.4 |
| Faster R-CNN [14] | 07 + 12 + COCO TRAINVAL | VGG | 84.1 |
| SSD300 [16] | 07 + 12 | VGG | 82.6 |
| SSD512 [16] | 07 + 12 | VGG | 85.5 |
| SSD321 [23] | 07 + 12 | RESIDUAL - 101 | 81.5 |
| DSSD321 [23] | 07 + 12 | RESIDUAL - 101 | 82.1 |
| SSD513 [23] | 07 + 12 | RESIDUAL - 101 | 85.7 |
| DSSD513 [23] | 07 + 12 | RESIDUAL - 101 | 86.4 |
| ION [24] | 07 + 12 + S | VGG | 82.3 |
| R-fcn [34] | 07 + 12 | RESIDUAL - 101 | 84.4 |
| RefineDet320 [35] | 07 + 12 + COCO TRAINVAL | VGG | 82.7 |
| RefineDet512 [35] | 07 + 12 + COCO TRAINVAL | VGG | 85.0 |
| RefineDet320+ [35] | 07 + 12 + COCO TRAINVAL | VGG | 86.0 |
| RefineDet512+ [35] | 07 + 12 + COCO TRAINVAL35K | VGG | 86.8 |
| Our Approach | COCO TRAINVAL | CSPDarknet-53 | 91.6 |
proposed by this study exhibited excellent performance in VOC 2012 comp4.

2) VOC 2012
This study isolated 5138 images with ground truth in Pascal voc 2012 [17] to VOC 2012. To prove that the model proposed in this study is effective in detecting pedestrians in images with special aspect ratios, this study first concatenated images in VOC 2012 horizontally into a triple-length image, and the ground truth images were expanded 3 times. Therefore, the original 7330 ground truth images became \(7330 \times 3 = 21,990\) pedestrians after being concatenated horizontally. FIGURE 4 displays the horizontally linked triple-length VOC2012 images. The pedestrians in the bottom of the images are small, and such images have high difficulty in pedestrian detection. Such images are difficult for some deep learning networks because they included small-scale pedestrians and large aspect ratios. Subsequently, this study compared the model proposed with the original YOLO methods. The detailed results are presented in TABLE 4 and FIGURE 5.

![FIGURE 4. Sample of horizontally linked triple-width images in VOC 2012.](image1)

According to the experimental results, the AP proposed by this study outperformed the original YOLO methods (FIGURE 5). In addition, TABLE 4 indicates that the proposed model provided optimal performance under various thresholds. To further prove the performance of this proposed model, this study compared it with several state-of-the-art methods, including Faster R-CNN [14], SSD [16], Mask R-CNN [22], RetinaNet [30], RefineDet [35], and RFBNet [36], by applying them to VOC 2012 images. The results revealed Our Approach was superior to all state-of-the-art methods, as indicated in TABLE 5.

3) INRIA
This study used the INRIA Person data set [18] to evaluate the proposed model by comparing it with YOLOv4 [13] and several state-of-the-art methods, namely Histogram of Oriented Gradients [18], VJ [37], InformedHaar [38], Locally Decorrelated Channel Features [39], Franken [40], Roerei [41], SketchTokens [42], SpatialPooling [43], RPN+ Boosted Forest [44], and Forest Deep Neural Networks [45]. As shown in FIGURE 6, the model proposed in this study missed only 6%
TABLE 5. Comparison of the AP of several state-of-the-art approaches with that of the proposed method for VOC 2012.

| Method                  | Data                  | Network k | AP   | PRA  |
|-------------------------|-----------------------|-----------|------|------|
| Faster R-CNN[14]        | 07++12                | VGG       | 62.9 | 62.9 |
| Faster R-CNN[14]        | 07++12                | ResNet    | 64.8 | 64.6 |
| SSD300[16]              | 07++12                | VGG       | 24.7 | 24.7 |
| SSD300+COCO TRAINVAL53K | 07++12+COCO TRAINVAL53K | VGG | 33.4 | 33.4 |
| SSD512[16]              | 07++12+COCO TRAINVAL53K | VGG | 39.4 | 39.3 |
| Mask RCNN[22]           | COCO             TRAINVAL53K | ResNet | -101 | -101 |
| RetinaNet[30]           | COCO             TRAINVAL53K | ResNet | -50  | -50  |
| RefineDet320[35]        | 07++12               | VGG       | 52.4 | 52.1 |
| RefineDet320[35]        | 07++12               | VGG       | 54.8 | 54.8 |
| RefineDet320[35]        | 07++12+COCO TRAINVAL53K | VGG | 58.5 | 58.5 |
| RefineDet320[35]        | 07++12+COCO TRAINVAL53K | VGG | 60.4 | 60.4 |
| RefineDet320[35]        | 07++12+COCO TRAINVAL53K | VGG | 61.6 | 61.5 |
| RefineDet512[35]        | 07++12+COCO TRAINVAL53K | VGG | 59.0 | 59.0 |
| RefineDet512[35]        | 07++12+COCO TRAINVAL53K | VGG | 61.4 | 61.4 |
| RefineDet512[35]        | 07++12+COCO TRAINVAL53K | VGG | 63.3 | 63.3 |
| RefineDet512[35]        | 07++12+COCO TRAINVAL53K | VGG | 65.4 | 65.3 |
| RefineDet512[35]        | 07++12+COCO TRAINVAL53K | VGG | 66.0 | 65.9 |
| RefineDet320[35]        | COCO             TRAINVAL53K | VGG | 58.2 | 58.2 |
| RefineDet320[35]        | COCO             TRAINVAL53K | VGG | 63.0 | 63.0 |
| RefineDet512[35]        | COCO             TRAINVAL53K | ResNet | 58.8 | 58.7 |
| RFBNet300[36]           | 07++12               | VGG       | 26.9 | 26.9 |
| RFBNet300[36]           | COCO             TRAINVAL53K | VGG | 29.0 | 29.0 |
| RFBNet152[36]           | COCO             TRAINVAL53K | VGG | 32.6 | 32.6 |
| RFBNetMobileNet[36]     | COCO             TRAINVAL53K | Mobile | 23.8 | 23.8 |
| Our Approach            | COCO             TRAINVAL53K | CSPDA | 77.3 | 77.1 |

FIGURE 6. Our approach had a lowest log-average miss rate of 6% in the INRIA Person data set, thus outperforming YOLOv4 and other state-of-the-art approaches.

FIGURE 7. Our approach had a lowest log-average miss rate of 2% in the INRIA Person data set after the use of annotated pedestrians, outperforming YOLOv4 and other state-of-the-art approaches.

of the INRIA Person data set [18] and outperformed all other methods.

Although this data set has a long history in the pedestrian detection field, some scholars have maintained that some pedestrian annotations provided by the original INRIA Person data set are missing. Therefore, [46] proposed a new annotation for the data set, which annotated all pedestrians with a height greater than 25 pixels and increased the total number of annotated pedestrians from 589 to 878. Therefore, this study used the new annotations to evaluate the proposed model. The experimental results on this new annotation in FIGURE 7 revealed that the method proposed in this study obtained optimal results and reduced the miss rate to 2%.

4) ETH

This study evaluated the proposed model by using the ETH Person data set [19], which was compared to YOLOv4 [13] and several state-of-the-art methods, namely [18], [37], [39]–[41], [43], [44], SDN [47], and task-assistant-CNN [48]. FIGURE 8 indicates that the model proposed in this study missed only 29% with the ETH Person data set [19], which outperformed all other methods.
FIGURE 8. Our approach had a lowest log-average miss rate of 29% in the ETH Pedestrian data set, rivaling those of several state-of-the-art approaches and outperforming YOLOv4.

E. LIMITATIONS
If feature maps that are candidate frames on the horizontal direction of input image are all overlapping, then the segment function will be not performed. The proposed method will work in a simpler manner, as shown in TABLE 7.

TABLE 7. Effect of using segmentation function on the detection results.

| Method                        | AP   |
|-------------------------------|------|
| Our Approach-original image feature | 90.5 |
| Our Approach-segmentation image feature | 90.7 |
| Our Approach                  | 91.6 |

V. ABLATION ANALYSIS
This section mainly studied the effectiveness and rationality of different components and parameter configurations of the model proposed in this study. All experiments in this chapter were conducted in VOC 2012 or VOC 2012 comp4, which manually concatenated the images into triple-length. As opposed to the previous section, this section compares the performance of the proposed model by various module components in detail.

A. SEGMENTATION FUNCTION
1) OUTLIER DETERMINATION
For the proposed model, this module was located in the first step of the segmentation function (filtering outliers). This step reduces situations in which nonpedestrians are identified as pedestrians. If this step is skipped, the test results are poor. To prove that this module can improve the detection results, this study lists the detection results obtained with and without filtered outliers in TABLE 6. The findings revealed that using filtered outlier modules can effectively improve detection accuracy.

2) SEGMENTATION
In this study, a module called segmentation function was used to divide the sample image into two to solve the problem of low pedestrian detection performance caused by small-scale pedestrians and special aspect ratios of the images through the divide-and-rule concept. Therefore, to prove that this module can improve the detection results, this study compared those that were obtained by using only features of the entire image, segmented image, and the fused features of the entire image and segmented image, as shown in TABLE 7.

B. MULTI-RESOLUTION ADAPTIVE FUSION
The final stage fused all feature maps in the entire network through multiresolution adaptive fusion to integrate the pedestrian detection information of all resolutions and achieve an improved method rather than simply using any single image for pedestrian detection. Therefore, to achieve the fusion effect, this study used feature maps similar to NMS [31] to fuse the network.

To select a suitable NMS algorithm, this study considered two commonly used NMS algorithms: NMS [31] and soft-NMS [32]. The results obtained from these two algorithms in VOC2012 are listed in TABLE 8. The AP generated by adding the proposed model and soft-NMS L in VOC2012 was the highest (91.8), and the AP generated using soft-NMS G produced similar results to the original NMS method (91.6). However, the output pedestrian frame illustrated that the proposed model with NMS obtained nearly the same AP as did the other two NMS algorithms when fewer output frames were used. Generally, although using soft-NMS L and soft-NMS G may improve the AP, they reduce the suppression effect and require more calculation time. Most importantly, this may result in numerous pedestrian frames output in the same location.

Therefore, after synthesizing visual data from TABLE 8, the NMS [31] was still selected as the algorithm for combining various resolution features.

C. INFERENCE TIME
Because the model proposed in this study is based on [13], the long-side maximum hyperparameters in [13] that can adjust the image input were retained. Nonspecifically indicated items in this study were similar to those in [13] in which the default long-side maximum value is 512, and this section lists the complete inference time of the methods proposed in TABLE 9.
VI. CONCLUSION

This study proposed a new pedestrian detection model, which follows the divide-and-rule concept. In addition to combining image ratio information into deep learning, it effectively integrated the pedestrian detection features of various resolutions in the network to solve pedestrian detection problems from the special aspect ratio and small-scale pedestrians. The experimental results indicated that the proposed model provides excellent detection results in popular pedestrian data sets and manually and horizontally expands triple-length images in this study.

In comparisons with the state-of-the-art approaches and YOLO, the proposed method achieves promising performance in average precision and log average miss rate for pedestrians with special aspect ratio and small scale, which remain challenging for the topic of pedestrian detection. Despite these advantages, our method still has some weaknesses. Some hyperparameters need to be selected manually, and the operation of segmentation function results in slightly increasing for inference time. In future work, these two problems will be considered by further presenting an approach that can automatically select the hyperparameters and a mechanism that is better than or even without segmentation function to speed up the operation.

Finally, the mechanism proposed in this study can be extended to apply to other CNN architectures or to detect various objects to further improve and enhance the performance of deep learning detection algorithms.

REFERENCES

[1] X. Wei, H. Zhang, S. Liu, and Y. Lu, “Pedestrian detection in underground mines via parallel feature transfer network,” Pattern Recognit., vol. 103, Jul. 2020, Art. no. 107155.
[2] C. Li, D. Song, R. Tong, and M. Tang, “Illumination-aware faster R-CNN for robust multispectral pedestrian detection,” Pattern Recognit., vol. 85, pp. 161–171, Jan. 2019.
[3] C. Zhou and J. Yuan, “Multi-label learning of part detectors for occluded pedestrian detection,” Pattern Recognit., vol. 86, pp. 99–111, Feb. 2019.
[4] W.-Y. Hsu, W.-Y. Lin, “Adaptive Fusion of Multi-Scale YOLO for Pedestrian Detection,” IEEE Trans. Image Process., vol. 30, pp. 1369–1381, 2021.
[5] J. Xu, W. Wang, H. Wang, and Y. Lu, “Hierarchical deep click feature prediction for fine-grained image recognition,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jul. 30, 2019, doi: 10.1109/TPAMI.2019.2932058.
[6] W.-Y. Hsu, “Automatic left ventricle recognition, segmentation and analysis,” in Proc. IEEE Int. Conf. Comput. Vis., vol. 27, no. 5, pp. 779–788, Jun. 2016.
[7] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DsDSSD: Deconvolutional single shot detector,” 2017, arXiv:1701.06659.
[8] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.
[9] L. Liu, D. Huang, “Receptive field block net for accurate and fast object detection,” in Proc. IEEE Int. Conf. Comput. Vis., vol. 189, no. 4, pp. 1097–1105, Dec. 2015.
[10] W.-Y. Hsu and W.-Y. Lin, “Ratio- and-scale-aware YOLO for pedestrian detection,” IEEE Trans. Image Process., vol. 30, pp. 934–947, 2021.
[37] P. Viola and M. Jones, “Robust real-time object detection,” Int. J. Comput. Vis., vol. 4, p. 4, Jul. 2001.

[38] 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.

[39] 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.

[40] M. Mathias, R. Benenson, T. Timofte, and L. V. Gool, “Handling occlusions with franken-classifiers,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 1505–1512.

[41] 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.

[42] 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.

[43] 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, Sep. 2014, pp. 546–561.

[44] 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, Oct. 2016, pp. 443–457.

[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., 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, Jun. 2013, pp. 286–295.

[47] 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.

[48] 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.

[49] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, “Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion,” IEEE Trans. Image Process., vol. 29, pp. 4296–4307, 2020.

[50] W.-Y. Hsu and C.-J. Chung, “A novel eye center localization method for multiview faces,” Pattern Recognit., vol. 119, Nov. 2021, Art. no. 108078.