IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 20, 2019, accepted April 4, 2019, date of publication April 11, 2019, date of current version April 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910201

PedJointNet: Joint Head-Shoulder and Full Body Deep Network for Pedestrian Detection

CHIH-YANG LIN[1], (Member, IEEE), HONG-XIA XIE[2], AND HUA ZHENG[2]
[1]Department of Electrical Engineering, Yuan Ze University, Taoyuan 32003, Taiwan
[2]College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350007, China

Corresponding authors: Chih-Yang Lin (andrewlin@saturn.yzu.edu.tw) and Hua Zheng (hzheng@fjnu.edu.cn)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 106-2218-E-468-001, Grant MOST 107-2221-E-155-048-MY3, and Grant MOST 108-2634-F-008-001, and in part by the Special Funds of the Central Government Guiding Local Science and Technology Development under Grant 2017L3009.

ABSTRACT Pedestrian detection when occlusions exist represents a great challenge in real-world applications, including urban autonomous driving and surveillance systems. However, the head–shoulder feature of pedestrians, which is more stable and less likely to be occluded than other areas of the body, can be used as a complement to full body prediction to boost pedestrian detection accuracy. In this paper, we investigate the unique features of the head–shoulder and full body features belonging to pedestrians. Then, instead of using a popular general object detection framework like R-CNN series, SSD, or YOLO, we propose a novel pedestrian detection network, called PedJointNet, that simultaneously regresses two bounding boxes to localize the head–shoulder and full body regions based on a feasible object detection backbone. Moreover, unlike the traditional strategy of keeping the weights fixed for each attribute, we design an inbuilt mechanism to dynamically and adaptively adjust the relationships of the head–shoulder and full body predictions for more accurate pedestrian localization. We validate the effectiveness of the proposed method using the CUHK-SYSU, TownCentre, and CityPersons datasets. Overall, our two-pronged prediction approach achieves excellent performance in detecting both non-occluded and occluded pedestrians, especially under circumstances involving occlusion, as compared to other state-of-the-art methods.

INDEX TERMS Pedestrian detection, head-shoulder detection, adaptively adjusted weights.

I. INTRODUCTION

Pedestrian detection for automated vision systems has attracted intense interest from many areas, including robotics, surveillance, entertainment and care for the elderly or disabled. While research for detecting pedestrians has always been a hot spot within the computer vision field, variations in the appearance of pedestrians (e.g., different clothes, sizes, aspect ratios, and dynamic shapes) and varying environments, make occluded pedestrians a difficult situation to solve for.

In recent years, pedestrian detection methods [1]–[4] have been largely inspired by object detection algorithms based on deep learning. The two main approaches that have been proposed are two-stage and one-stage (end-to-end neural network) approaches. The popular two-stage R-CNN series includes R-CNN [5], Fast R-CNN [6] and Faster R-CNN [7]. As discussed in Zhang *et al.*'s study [3], region proposal network (''RPN'' for short in this paper) can serve as a stand-alone pedestrian detector, while the downstream classifier in Faster R-CNN may weaken the performance of pedestrian detection. That is why R-CNN-based pedestrian detectors usually retain RPN to generate region proposals [4], [8], [9]. One-stage style detectors, SSD [10] and YOLO [11] have also been applied to pedestrian detection for their high efficiency [12]. In general, one-stage neural networks have fast performance; however, two-stage neural networks can more easily achieve more robust performance.

Although the pedestrian detection methods mentioned above have demonstrated significant improvements compared to traditional methods using manual features [13], [14], occlusion remains an unsolved issue. According to the statistics in Caltech [15], at least 70% of pedestrians per video frame are occluded. Current algorithms for pedestrian detection usually focus on full body detection. When occlusions are present, such algorithms perform worse than part detectors [8], [16], [17] do.

Part detectors in pedestrian detection partition the full body into regions, and can help handle occlusions. A significant

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

disadvantage of these methods is that parts' weights are manually assigned rather than automatically learned, and easily produce suboptimal results. In DeepParts [16], several single part detectors are designed to learn separately for pedestrian detection. This idea ignores the correlation between different parts of the pedestrian. In bi-box regression method [8], the visible and full body regions are chosen as complementary selections. However, visible part annotations are rare in many pedestrian datasets.

The backbone networks in pedestrian detection networks [1], [8], [12], [9] typically derive from VGG16 [18] or ResNet50 [19], which are designed for image classification. Little research discussed specifically designing backbone networks for object detection with various scales until DetNet [20] was proposed. DetNet adds extra steps based on traditional image classification backbone network, while maintaining high spatial resolution of feature maps in deeper layers.

Inspired by the observations above, we put forward a novel method for pedestrian detection called PedJointNet, which regresses the full body and the head-shoulder parts using two bounding boxes per pedestrian. Instead of adopting the general object detection frameworks that are based on an image classification backbone, i.e. Faster R-CNN, SSD, or YOLO, we design a two-branch pedestrian detection architecture with unique backbone network. We investigate the inner characteristics of a pair of complementary features (full body and head-shoulder), and the different two-branch sub-networks are intended to produce correlative outputs for better detection results.

In addition, we explore how to design an efficient framework for detecting partially occluded pedestrians. Rather than adapt the backbone from popular general object detection architectures with VGG16 or ResNet50 backbones, we choose to modify the DetNet backbone for object detection. Then, we use the two-branch logic to simultaneously regress the head-shoulder and the full body predictions of pedestrians. The two attributes behave as complementary features to avoid missing pedestrian instances. In consideration of the various scales of pedestrians in real-world applications, we integrate pyramid module and atrous convolution to generate proposal boxes. With regard to the fusion of the two parts, we design an adaptive weighted loss layer to adjust their weights in order to lower the miss rate as much as possible.

In summary, our contributions lie in: (1) A novel two-part prediction model for pedestrian detection that incorporates the head-shoulder region and the full body region into a unified deep CNN architecture; (2) A novel backbone that is truly designed for pedestrian detection instead of adapting a common R-CNN backbone for image classification; (3) A two-part fusion scheme that is adaptively adjusted, and which adjusts the weights automatically during training to ensure the highest accuracy.

We validate the effectiveness of our approach on the CUHK-SYSU [21], TownCentre [34] and CityPersons [4] datasets. Our approach has comparable performance to the state-of-the-art method for non-occluded pedestrian detection, and it achieves the highest accuracy rate in detecting occluded pedestrians, especially under heavily occluded conditions. Figure 1 shows some experimental results of our proposed PedJointNet.

FIGURE 1. Detection examples of the proposed PedJointNet using CUHK-SYSU [22], TownCentre and CityPersons [4] datasets, where the blue and green boxes represent the predicted full body and head-shoulder parts respectively.

The paper is organized as below: Section II presents some work related to pedestrian detection. Section III explains the detail of the proposed algorithm and framework, Section IV demonstrates our experimental results and Section V summarizes our work.

II. RELATED WORK

A. COMMON DEEP R-CNN PEDESTRIAN DETECTION MODELS

Deep ConvNets have been widely modified for pedestrian detection and achieved promising results. Motivated by the success of the two-stage R-CNN series architecture, SAF-RCNN [9] developed a Scale-Aware Fast R-CNN framework for pedestrian detection. Zhou and Yuan [8] designed a bi-box regression to generate region proposals that may contain pedestrians and adopted the Fast R-CNN architecture for full body and visible part estimation. The CNN-based network is trained for pedestrian detection, and auxiliary semantic information is incorporated to enhance performance [23]. Brazil *et al.* [1] used semantic segmentation infusion layer to boost the pedestrian detection accuracy of Faster R-CNN. Wang *et al.* [2] designed the repulsion loss specifically for crowd scenes with a significant improvement in occlusion cases.

Unfortunately, the speed of two-stage frameworks is constrained by repeated feature extraction and evaluation in CNN. A potential solution is an one-stage detector, which simplifies the pipeline and directly uses the backbone for object prediction. As is typical of one-stage detectors, YOLO extracts features through the DarkNet backbone [11] and simplifies detection as a regression pipeline. Although the detection speed of YOLO is quite competitive, it is not

ideal for small target objects. Therefore, several upgraded versions, YOLOv2 [24] and YOLOv3 [25] have since been proposed. YOLOv3 adds a top-down multi-layer prediction to the network, and the softmax loss is replaced by logistic loss for better detection performance. Focal Loss proposed in RetinaNet [26] aims to solve the class imbalance led by the large foreground-background ratio. SSD adopts reduced VGGNet for feature extraction in a multi-layer way, aiming to handle variant object instance scales. While recent one-stage detection networks have not achieved competitive accuracy to two-stage detectors on popular pedestrian detection benchmarks, DSSD [27] introduces deconvolution (''transposed convolution'' in some literature) to enhance the correlation among top-down layers in the original SSD. ALFNet [12] has proposed Asymptotic Localization Fitting, which stacks a series of predictors to adapt SSD's default anchor boxes to improve pedestrian detection results.

B. PART DETECTORS TO HANDLE OCCLUSIONS IN PEDESTRIAN DETECTION

Recently, part detectors have become more prevalent in approaches to handle occlusion scenarios in pedestrian detection [16], [22], [28]. A common practice is to choose the desired parts suitable for various occlusion patterns, which will then be learned and assembled for pedestrian detection. The disadvantage of full body detection under instances of occlusion can be avoided to some extent this way. The Franken-classifiers in [29] can bias feature selection during the training of 16 classifiers for frequently occurring occlusions. Ouyang and Wang [30] constructed 20 component detector scores and incorporated the deformation layer into CNN. Various deformation operations handling occlusions can be easily inserted to the model with the deformation layer. However, part annotations are not available in this approach to learn the detectors, so the performance is limited. Part detectors learned in [28] share boosting decision trees to tie correlations among parts, seeking to reduce the computational cost. However, it does not perform as well as other top performing part detectors in either slightly occluded or non-occluded pedestrian detection cases.

One drawback of these approaches is that the weights of each part are manually designed and therefore may not be optimal. The correlation between these parts is not well explored in existing pedestrian detection literature, but their relationship is critical because these parts can serve as complementary attributes in the face of different occlusions and variations.

C. TWO-BRANCH ARCHITECTURE FOR PEDESTRIAN DETECTION

Some previous work closely relates to ours [1], [8], [9]. In [9], the proposed model introduces two built-in subnetworks that detect pedestrians with scales from disjoint ranges. It employs the gate function to handle various pedestrian scales. But the object proposals are generated using ACF detector, which is less efficient and accurate for pedestrian detection. The two branches in [8], i.e., full body and visible part, are used to produce complementary outputs. However, the visible parts are difficult to label in most applicants, and the softmax fusion operation has more limited performance than our automatically fusion methods do.

Brazil *et al.* [1] add a segmentation branch as a strong cue to boost pedestrian detection. RPN [7] was applied to generate pedestrian proposals in [1]. However, the fusion methods used in these approaches lack the ability to automatically adjust the weights between two branches, which may limit the performance. Furthermore, the sub-network architecture of both branches are exactly the same in [1], [8], [9], and there is no specific design for each attribute in the network architecture. We have explored the performance of head-shoulder and full body predictions in different sub-network architectures and designed targeted sub-networks for each. In addition, the adaptive weighted fusion loss scheme in our model can make the two attributes truly complementary, which is better than applying hand-crafted weights.

III. PROPOSED APPROACH

Our proposed framework produces two bounding boxes for each pedestrian, specifying the full body and head-shoulder part respectively. We modify a feasible architecture of DetNet [20] for object detection as our backbone network, specifically for pedestrian detection. For each image, region proposals potentially containing pedestrians are listed via an anchor box generation approach that is similar to [10]. In the next step, we integrate two kinds of feature pyramid modules to build two branches specifically for head-shoulder and full body prediction. Then, we build a novel sharing mechanism which can update the weight of each branch dynamically and adaptively through the corresponding generalization ability. Figure 2 depicts the overall architecture of our approach.

A. MODIFIED DETNET WITH STRONG PYRAMID FEATURE MODULES

Recent pedestrian detectors tend to rely on a backbone network, such as VGG16, ResNet, which are usually pre-trained for image classification task. Li *et al.* [20] indicated that the backbone network designed for image classification are sub-optimal for the localization objective, especially pedestrian detection. Although feature maps down-sampled by large strides, such as 32, in a typical image classification backbone have strong semantic information and a large receptive field, the boundary will be too blurry to obtain an accurate regression. Another drawback of a large stride in pedestrian detection network is that the features and contextual cues of small pedestrians are easily to be overlooked and large context information is integrated. This can explain why generic object detection models cannot do well in pedestrian detection. Thus, we need to figure out how to avoid down-sampling operations while keeping the larger receptive field in pedestrian detection.

To address these problems, the number of stages in our backbone architecture is directly designed for pedestrian

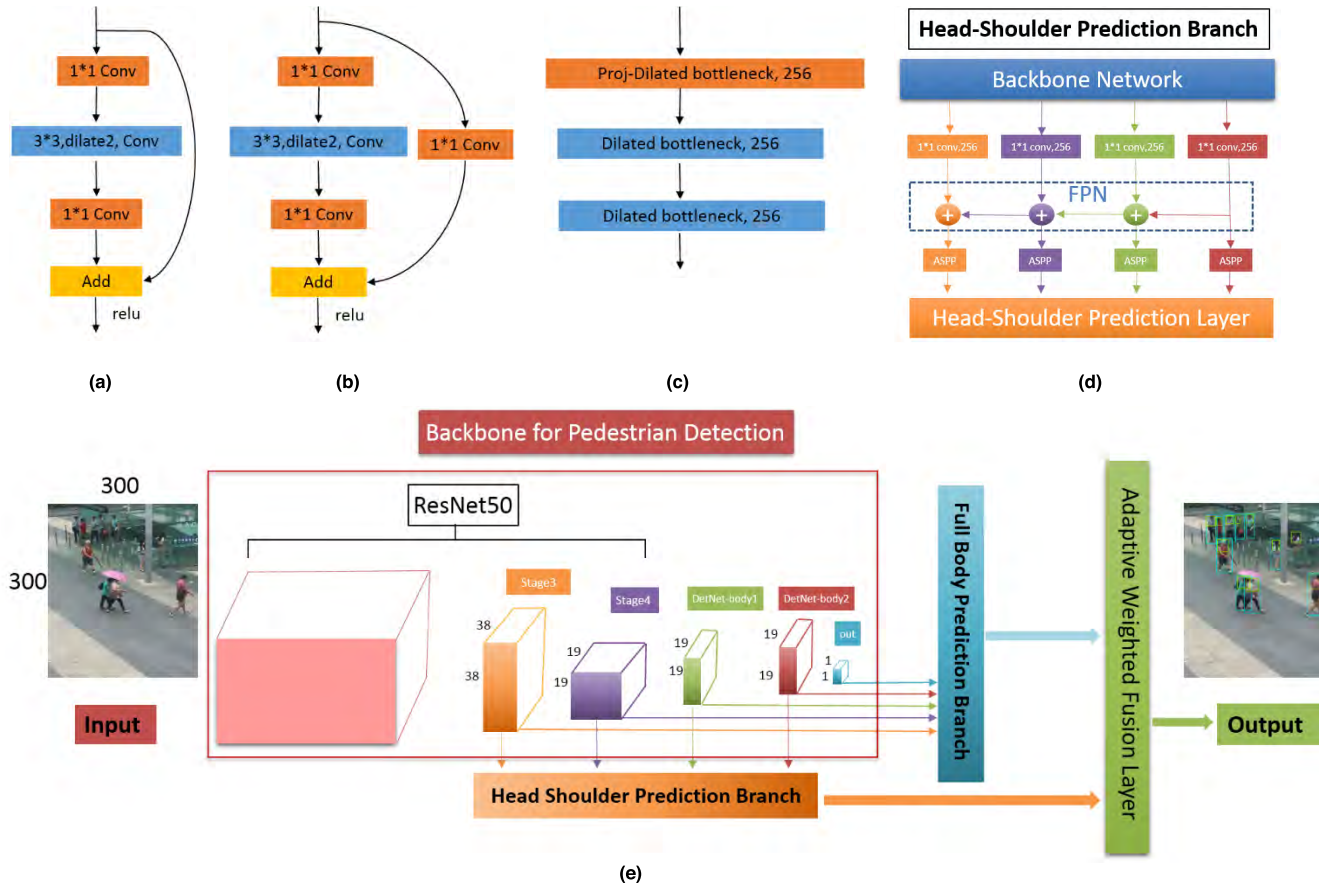


FIGURE 2. PedJointNet overall architecture. (a) Dilated bottleneck. (b) Proj-Dilated bottleneck (Dilated bottleneck with 1*1 conv projection). (c) DetNet-body architecture. (d) Details of head-shoulder prediction branch. (e) Overall architecture.

detection. Our backbone maintains a higher spatial resolution (i.e., 19 × 19) in the feature maps while preserving the large receptive field. Inspired by the architecture of DetNet [20], we extend its architecture for our backbone network. The details of our pedestrian detection backbone are illustrated below.

The first through fourth stages in our backbone are the same as in the original ResNet, and the extra *DetNet-body1*, *DetNet-body2*, and *out* are introduced in the backbone. Meanwhile, the spatial resolution is maintained just as in *stage 4*, *DetNet-body1, DetNet-body2*, as shown in Figure 2(e).

For *DetNet-body1*and *DetNet-body2*, we employ a dilated [28]–[30] bottleneck with the 1 × 1 convolution projection (as shown in Figure 2(b)) at the beginning of each stage. The experiment result shows the dilated bottleneck is important for pedestrian detection, especially for small head-shoulder detection.

Since the head-shoulder and full body scales vary between pedestrians, the sub-network needs more contextual cues from multi-level feature maps, and multi-scale information must be correctly encoded. The ability to control the spatial resolution of feature responses offers a useful guide for our pedestrian detection.

To address this issue, Feature Pyramid Network (FPN) [31] and Atrous Spatial Pyramid Pooling (ASPP) [32] are two ideas that are often used to enhance the network across feature maps learning. ASPP was proposed to concatenate multiple atrous-convolved features into a final feature representation by using 4 different dilation rates as in Figure 3(a). It allows us to enlarge the receptive field, (referred to as the "field of view" in some literature) and incorporate a larger context while maintaining the resolution of feature maps. FPN builds a U-shape architecture with top-down pathway and lateral connections, and it outputs multi-layer in pyramidal hierarchy, as shown in Figure 3(b). With the ASPP layout in parallel and FPN in serial, multi-level and multi-scale, feature maps can be generated, which is beneficial for small proposals or pedestrian detection involving occlusion.

Figure 2(e) shows how the backbone to extract features is specifically designed for pedestrian detection. It complies the same design as ResNet50 up to *stage 4*, while keeping the feature map size after *stage 4* (i.e. *DetNet-body1* and *DetNet-body2*). Based on the individual performance of the full body and head-shoulder parts in detection, we designed two multi-level prediction branches. The Head-Shoulder Prediction Branch (in Figure 2(d)) combines FPN with multi-level feature maps, i.e., *stage 3, stage 4, DetNet-body1, DetNet-body2*,

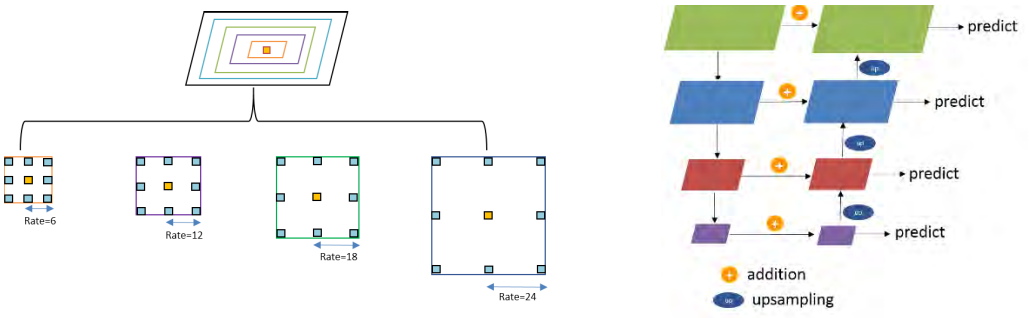


FIGURE 3. The architectures of atrous spatial pyramid pooling (ASPP) and feature pyramid network (FPN), which are combined in our strong pyramid feature modules to make the receptive field as large as possible without sacrificing feature map spatial size. (a) Atrous Spatial Pyramid Pooling (ASPP). (b) Feature pyramid network (FPN).

while the Full Body Prediction Branch directly uses these 5 feature maps. The two branches are thus concatenated to next adaptive fusion layers to adjust their weights for the final output.

B. TWO-BRANCH PREDICTION FOR HEAD-SHOULDER AND FULL BODY

Within the PedJointNet backbone introduced in Section 3.1, an auxiliary structure embedded in the network produces proposal boxes. Each added feature layer can produce bounding boxes and scores with fixed number for pedestrians instances through a set of convolutional filters inspired by conventional SSD. For each box out of the total number of filters, k, 2 class scores for head-shoulder and full body, and 4 offsets coordinates relative to the default box are calculated. This process yields $(2+4) \times kmn$ outputs for $m \times n$ feature map.

For common two-branch networks, the general solution is to first generate region proposals for the two classes through the sharing backbone network, and the two sub-networks with same architecture are learned to predict each individual task. However, the varying learning challenges and convergence rates of two branches may fail to make them serve as complementary features for accurate pedestrian detection. In accordance with the features of the head-shoulder and full body, the two branches are designed separately to learn and predict to ensure efficiency and accuracy. Figure 4 shows the difference between the structures of the traditional two-branch network architecture used for pedestrian detection [1], [8], [9] and our two-branch network architecture. Based on modified DetNet backbone introduced in previous part, our two branches have different proposals according to the inner difference between the full body and head-shoulder predictions. The way that scores are fused in the two structures also differs, as will be explained the next section.

C. THE FUSION LAYER OF ADAPTIVE WEIGHTED LOSS

Traditional two-branch fusion methods usually use regular softmax operation to increase the complementarity and robustness of two branches [1], [8]. However, for two-branch prediction we cannot have a pre-specified ratio of importance

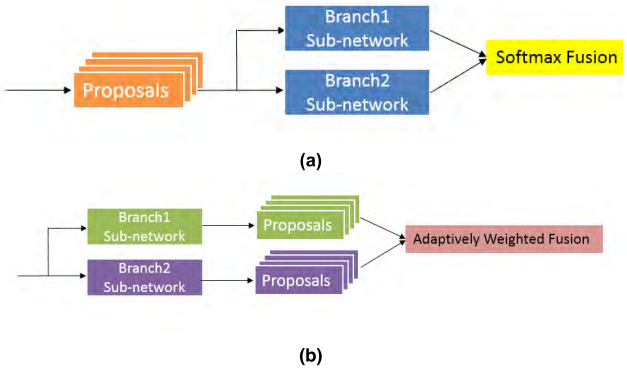


FIGURE 4. Comparison of our two-branch network architecture with traditional design. (a) The common proposals are fed into two separate branches (as shown in the two blue boxes). (b) We design two different sub-networks (as shown in the green and purple boxes) for the two branches according to the inner features of head-shoulder and full body.

for each task. The weight of each branch is expected to be learned by the network automatically.

Our fusion layer of adaptive weighted loss aims to coordinate the relationships in learning the head-shoulder and full body parts of pedestrians. Specifically, the weight of each attribute can be dynamically adjusted during training.

Intuitively, the head-shoulder and full body regions should be well correlated. The adaptive weighting function for the two branches is learned as follows. The loss function in general pedestrian detection usually consists of class confidence over all classes and bounding box regression derived from Fast R-CNN. In this paper, we follow the same bounding box regression formulation as in Fast R-CNN. The difference between the two-branch fusion in traditional methods and our work lies in the softmax computation, "SC" for short, which is a component of class confidence computation.

SC_1 in traditional two-branch fusion methods [1], [8] adopts regular softmax computation in (1) without adaptive weighting, where s_1^1, s_2^1 represents positive detection scores, and s_1^0, s_2^0 represents negative detection scores for the head-shoulder and full body branch, respectively.

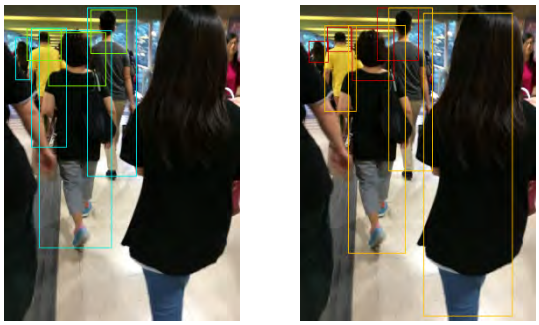
While in our proposed adaptive weighted fusion SC_2 in (2), we assign scalar values λ_1, λ_2 to weight the importance

of the head-shoulder and full body, respectively, which are initialized as 1. Then, $softmax_1$ and $softmax_2$ are caculated using SC_1. During training, λ_1 and λ_2 are updated through the mean loss differences between two intervals every period k, the details of which can be seen in [33]. This is an efficient yet effective approach to tracking the loss trend in order to dynamically adjust the importance of each branch.

$$SC\_1 = \frac{\exp\left(s_1^1 + s_2^1\right)}{exp\left(s_1^1 + s_2^1\right) + \exp\left(s_1^0 + s_2^0\right)} \quad (1)$$

$$SC\_2 = \lambda_1\, softmax_1 + \lambda_2\, softmax_2 \quad (2)$$

Figure 5 features an illustration of why our adaptive weighted fusion performs better than traditional fusion does. Assuming the detection threshold is 0.3, the full body score and head-shoulder score of the woman closest to the camera will be 0.31 and 0.25, respectively.



(a) (b)

FIGURE 5. Comparison of two-branch fusion between traditional method and our adaptive weighted fusion. (a) Traditional method. (b) Our work.

The final score SC_1 of the traditional method will be 0.29 according to (1). While in our adaptive weighted fusion in (2), the λ_1, λ_2 would be 0.7 and 0.65 respectively, resulting in a final score SC_2 of 0.38. Therefore, the woman will not be missed using the adaptive weighted fusion approach as the threshold set to be 0.3.

IV. EXPERIMENTS AND ANALYSIS

We evaluate the effectiveness of our proposed PedJointNet on CityPersons [4], CUHK-SYSU Person Search Dataset [22] and TownCentre [34], which are popular pedestrian detection datasets.

A. EXPERIMENT SETTINGS

1) DATASETS

Robust pedestrian detection should be designed to address cases involving partial occlusions and large numbers of people per image, i.e., closer to daily surveillance applications. It is thus important to train our model on datasets that consider volume, diversity, and occlusion.

CityPersons is a new collection of person annotations derived from the Cityscapes dataset [35]. In a total dataset of 5,000 images, there are 35,000 people, 13,000 of which lack region annotations. Both the popular Caltech [15] and KITTI [36] pedestrian datasets, despite having a large number of frames, have lower density groups of people. While our work focuses on occlusion circumstances, CityPersons includes two times more occlusion cases than Caltech dataset, and the occlusion patterns in CityPersons enhance its diversity, make the dataset a more suitable test base. Specifically, in the reasonable subset (<= 0.35 occlusion), CityPersons has more occlusion cases while Caltech is dominated by full visible pedestrians. As validated in [4], CityPersons performs well on small scale and occluded pedestrians and achieves higher localization ability.

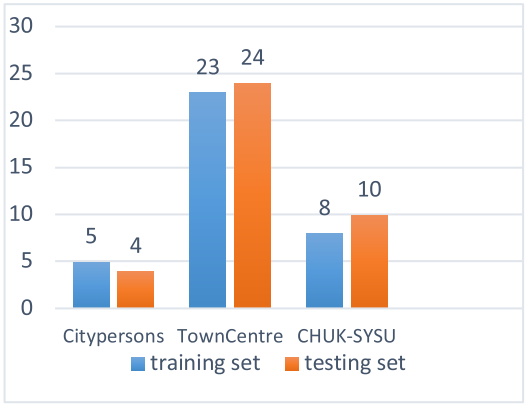


FIGURE 6. Comparison of the average number of persons presented in an image for the training and testing sets.

TownCentre is a pedestrian tracking dataset in a busy town center street. We use the first 3600 frames of the video for training and validation, and the last 900 frames for testing. In addition, we generate XML annotation files for the head-shoulder parts according to the given full body annotations.

CUHK-SYSU Person Search Dataset contains 18,184 images collected from public areas and movies with high density pedestrians. We use the subset version provided by [21], which includes 16,907 annotated images with head-shoulder and full body parts. We divide it into a training set and test set that contain 13,399 and 3,508 images, respectively. The reason we chose this dataset is that it was the only public dataset we could find that contained accurate head-shoulder and full body bounding box annotations. In addition, images in this dataset have significant variations in perspective, illumination, and background, which are very helpful for creating a pedestrian detection system for real applications. Figure 6 shows the average number of persons presented in an image for the training and testing sets used in the experiments.

2) TRAINING DETAILS

Our method is implemented using a Tensorflow backend, with 1 GTX 1080Ti GPUs for training.

For CUHK-SYSU and TownCentre datasets, our backbone network adopts pretrained weights on ImageNet, and all added layers were randomly initialized with he_normal [37]. The network was completely trained in 338,000 iterations, using the step decay schedule, with an initial learning rate

TABLE 1. comparison of different backbone networks used in the CUHK-SYSU dataset.

Backbone	Number of stages	MR^{-2}	
		IoU=0.5	IoU=0.75
VGG16	Stage1-4	16.01	48.94
ResNet50	Stage1-4	16.23	25.49
Our modified DetNet	Stage4,5,6	13.27	27.12
	Stage4,5,6,out	**9.40**	**18.28**
	Stage4,5,6,7,out	11.26	18.67

of 1e-4, and decreased by a factor of 0.75 and a step size of 2. For CityPersons, we also included experiments with the model initialized from CUHK-SYSU and trained it across a total of 240,000 iterations with a learning rate of 1e-5.

3) OPTIMIZER

Original DetNet was optimized via Stochastic Gradient Descent (SGD) with a weight decay of 0.0001 and momentum of 0.9. To design an effective convergence pedestrian detection network, we investigate several successful stochastic optimizers in training deep networks. SGD with momentum and Adam are using gradient updates scaled by square roots of exponential moving averages of squared past gradients. However, these algorithms tend to fail to converge to an optimal solution in many applications, e.g. learning with large output spaces [38]. Therefore, we investigate two recently proposed methods based on Adam: Amsgrad [38] and AdamW [39].

Amsgrad is a new variant of Adam. By endowing Adam with ''long-term memory'' of past gradients, Amsgrad is expected not only to solve the convergence issues, but also lead to improved empirical performance. AdamW used in our PedJointNet, decouples the optimal choice of weight decay factor from the learning rate for both regular SGD and Adam.

Figure 7 shows the comparison of convergence rate for Adaw, AdamW, Amsgrad, and SGD with momentum, respectively. The experimental results show that AdamW converges faster than the other three optimizers, while also achieving the smallest loss in training. After exploring, we believe that AdamW's success is due to its normalized weight decay and the use of cosine annealing and warm restarts for Adam, resulting in a more robust hyper-parameter selection, better final performance, and a better anytime performance.

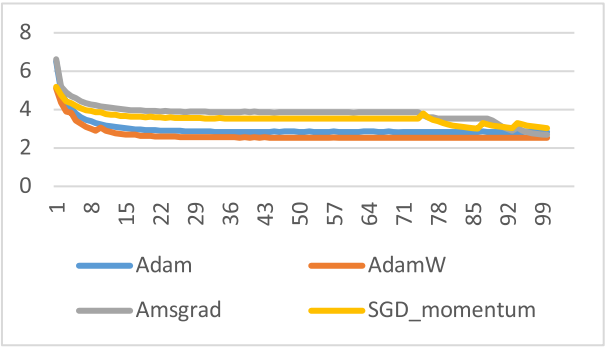


FIGURE 7. Convergence rate comparison for Adaw, AdamW, Amsgrad, and SGD with momentum.

B. ABLATION EXPERIMENTS

In this part, we experiment ablation studies on the CUHK-SYSU, CityPersons, and TownCentre pedestrian benchmarks to illustrate the effectiveness of our proposed PedJointNet.

1) DIFFERENT BACKBONE NETWORK

Table 1 shows the comparison of different backbone networks used in the CUHK-SYSU dataset. It turns out that the adopted DetNet with *stages 4, 5, 6*, *out*achieves the lowest miss rate MR^{-2} (lower is better).

We compared the performance of the current mainstream backbone network VGG16 and ResNet50 with our modified DetNet on the CUHK-SYSU dataset by selecting two different IoU thresholds, 0.5 and 0.75, as shown in Table 1.

The miss rates of VGG16 and ResNet50 (the feature map is reduced by 32 times) are higher than DetNet. It is worth noting that we also did a blind study for different stages of DetNet, comparing the impact of *stage 4, 5, 6*, *stage 4, 5, 6*, *out* and *stage 4, 5, 6, 7, out* stages. The results prove that our DetNet with *stage 4, 5, 6, out* can achieve the lowest miss rate, which is 9.40, 18.28 at IoU = 0.5 and IoU = 0.75 respectively. This result indicates that the fixed spatial resolution after *stage 4* (the feature map is reduced by a factor of 16), and after which each stage beginning with an expansion bottleneck and a 1x1 convolution projection can perform better on pedestrian detection.

2) TWO-BRANCH MODULES AND ADAPTIVE WEIGHTED LOSS LAYER

To further validate the effectiveness of our multi-level strategy, i.e., FPN and ASPP, Table 2 shows comparison when different two-branch modules are used, or if the adaptive weighted loss layer is applied.

From the results in Table2, two-branch modules with adaptive weighted layer generally boost the detection mAP. And two branches combined with the adaptive weighted layer obtain mean average precision (mAP) with 0.76, 0.75 in head-shoulder branch and full-body branch respectively.

And it appears that the FPN and ASPP modules with the adaptive weighted layer will obtain the best performance for

FIGURE 8. Comparison of pedestrian detection results with other state-of-the-art methods. The first column shows the input images. The remaining columns show the detection results of our method (green rectangles and blue rectangles represent head-shoulder and full body results, respectively), YOLOv3 [25], ALFNet [12] and bi-box regression [8] (full body annotated with yellow rectangles and red rectangles indicates head-shoulder), respectively. It can be seen that our head-shoulder and full body can be used as complementary attributes to successfully detect most pedestrians under different occlusion scenarios.

the head-shoulder branch. Notably, for the full body branch modules, involving FPN and ASPP cannot provide accuracy gain in CUHK-SYSU dataset, as the pre-trained weights can be helpful for full body detection. These results indicate that the dilated bottleneck is important for pedestrian detection, especially for small head-shoulder detection. Since the relatively small head-shoulder needs more contextual cues from multi-level feature maps, and multi-scale information must be correctly encoded. With the ASPP layout in parallel and FPN in serial, multi-level and multi-scale, which is beneficial for small head-shoulder detection involving occlusion.

TABLE 2. The effectiveness of different two-branch modules with and without the adaptive weighted loss layer.

Head-Shoulder Branch								
FPN	√		√		×		×	
ASPP	√		×		√		×	
Adaptive weighted loss layer	√	×	√	×	√	×	√	×
CUHK-SYSU mAP	**0.76**	0.75	0.73	0.72	0.72	0.73	0.69	0.64
Full Body Branch								
FPN	√		√		×		×	
ASPP	√		×		√		×	
Adaptive weighted loss layer	√	×	√	×	√	×	√	×
CUHK-SYSU mAP	0.71	0.68	0.74	0.73	0.74	0.74	**0.75**	0.73

TABLE 3. A comprehensive comparison of our method with other state-of-the-art methods showing the citypersons MR^{-2}, CUHK-SYSU and towncentre mean average precision score for each method.

Method	CityPersons	CUHK-SYSU	TownCentre
YOLOv3 [25]	R: 15.36 H: 54.89	0.69	0.79
ALFNet [12]	R: 14.13 H: 54.32	0.71	0.87
Bi-box Regression[8]	R: 13.85 H: 53.59	0.73	0.87
Ours	**R: 13.45** **H: 52.17**	**0.78**	**0.88**

C. COMPARISON TO STATE-OF-THE-ART METHODS

We compared our proposed PedJointNet with YOLOv3 [25], ALFNet [12] and bi-box regression [8] which are state-of-the-art pedestrian detection methods, as shown in Table 3.

1) CITYPERSONS

R signifies a reasonable set, while H refers to a heavy set from the CityPersons dataset. For the Reasonable setup, pedestrians are taller than 50 pixels and occlusion is less than 35%. For the Heavy setup, the height and visibility ranges of pedestrians are $[50, \infty]$ and $[0.2, 0.65]$ respectively. In CityPersons, our PedJointNet achieves a miss rate MR^{-2} of 13.45%, 52.17% in the Reasonable and Heavy subsets, which is lower than what ALFNet and bi-box regression achieve.

2) CUHK-SYSU AND TOWNCENTRE

As CUHK-SYSU contains various occlusions and multiple variants, e.g., different illumination and angles, although YOLOv3 and ALFNet specifically aim to simplify pedestrian detection into a single stage and detect small objects specifically, they only achieve mAPs of 0.69 and 0.71 respectively, as they cannot handle occlusions and small instances well. Our PedJointNet has the best performance with 0.78 mAP. The increased detection accuracy results from the complementary outputs produced by our two-branch mechanism.

Note that bi-box regression is designed for occlusion handling, it lacks two-branch adaptive weighted operation, while our adaptive weighted loss scheme can refine the weights of the head-shoulder and full body parts, thus achieving highest accuracy under various occlusion situations (0.78, 0.88 in CUHK-SYSU and TownCentre, respectively). Some testing samples under challenging variants can be seen in Figure 8.

V. CONCLUSION

In this paper, we propose a new two-branch architecture for pedestrian detection, called PedJointNet. The backbone is fully adapted for object detection and leverages two kinds of feature pyramid modules to build two branches specifically for head-shoulder and full body part prediction. Then, we build a novel sharing mechanism that enables weight adjustment dynamically and adaptively according to different learning ability of each branch during training. This novel design achieves an accuracy rate that is competitive with that of current state-of-the-art pedestrian detectors, especially in occasions that involve occlusion.

In the future, we would like to introduce more useful guidance features for our pedestrian detection approach, including other parts of the pedestrian and the semantic features of the pedestrian. The proposed PedJointNet can therefore further be applied to people with arbitrary occlusions.

REFERENCES

[1] G. Brazil, Y. Xi, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 4950–4959.
[2] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7774–7783.
[3] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 443–457.
[4] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3213–3221.
[5] 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.*, Washington, DC, USA, Jun. 2014, pp. 580–587.
[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Washington, DC, USA, Dec. 2015, pp. 1440–1448.
[7] 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.*, Montreal, Quebec, Canada, 2015, pp. 91–99.
[8] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 135–151.
[9] 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.
[10] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 21–37.
[11] 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.*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.
[12] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning Efficient Single-stage Pedestrian Detectors by Asymptotic Localization Fitting," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 618–634.
[13] G. Gan and J. Cheng, "Pedestrian detection based on HOG-LBP feature," in *Proc. 7th Int. Conf. Comput. Intell. Secur.*, Sanya, Hainan, China, Dec. 2012, pp. 1184–1187.
[14] Y. Xin, X. Shan, and S. Li, "A combined pedestrian detection method based on haar-like features and hog features," in *Proc. 3rd Int. Workshop Intell. Syst. Appl.*, Chaves, Portugal, 2011, pp. 1–4.
[15] P. Dollar, 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.
[16] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Washington, DC, USA, Dec. 2015, pp. 1904–1912.
[17] S. Wang, J. Cheng, H. Liu, and M. Tang. (Apr. 12, 2018). "PCN: Part and context information for pedestrian detection with CNNs." [Online]. Available: https://arxiv.org/abs/1804.04483
[18] K. Simonyan and A. Zisserman. (Sep. 4, 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[20] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: Design backbone for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 334–350.
[21] G. Chen, X. Cai, H. Han, S. Shan, and X. Chen, "HeadNet: Pedestrian head detection utilizing body in context," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, May 2018, pp. 556–563.
[22] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3415–3424.
[23] 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.*, Boston, MA, USA, Jun. 2015, pp. 5079–5087.
[24] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 7263–7271.
[25] J. Redmon and A. Farhadi. (Apr. 8, 2018). "YOLOv3: An incremental improvement." [Online]. Available: https://arxiv.org/abs/1804.02767
[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 2980–2988.
[27] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. (Jan. 23, 2017). "DSSD: Deconvolutional single shot detector." [Online]. Available: https://arxiv.org/abs/1701.06659
[28] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Oct. 2017, pp. 3486–3495.
[29] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1505–1512.
[30] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2056–2063.
[31] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.
[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[33] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Adaptively weighted multi-task deep network for person attribute classification," in *Proc. ACM Multimedia Conf.*, California, CA, USA, 2017, pp. 1636–1644.
[34] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 3457–3464.
[35] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223.
[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Rhode Island, RI, USA, Jun. 2012, pp. 3354–3361.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Washington, DC, USA, Dec. 2015, pp. 1026–1034.
[38] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018, pp. 1–23.
[39] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018, pp. 24–37.

CHIH-YANG LIN (M'11) received the Ph.D. degree in computer science and information engineering from National Chung-Cheng University, Chiayi, Taiwan, in 2006. From 2007 to 2009, he was with the Advanced Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan. In 2009, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, as a Postdoctoral Fellow. In 2010, he joined Asia University, Taichung, Taiwan, where he became an Associate Professor, in 2013. From 2014 to 2017, he was the Chair of the Department of Bioinformatics and Medical Engineering, Asia University, where he is currently an Assistant Professor. He is also an Associate Professor with the Department of Electrical Engineering, Yuan-Ze University, Taoyuan, Taiwan. He has authored or coauthored over 100 papers and holds patents. His research interests include computer vision, machine learning, image processing, and the design of surveillance systems. He received best paper awards from the Pacific-Rim Conference on Multimedia (PCM), in 2008, and best paper awards and the Excellent Paper Award from Computer Vision, Graphics and Image Processing Conference, in 2009 and 2013. He has served as the Session Chair, the Publication Chair, or a Workshop Organizer for many international conferences, including AHFE, ICCE, ACCV, IEEE Multimedia Big Data, ACM IH&MMSec, APSIPA, and CVGIP.

HONG-XIA XIE received the B.S. degree in Internet of Things from the Zhengzhou University of Aeronautics, China, in 2016. She is currently pursuing the master's degree in communication and information systems with Fujian Normal University.

She has participated in projects including head–shoulder segmentation and pedestrian detection. Her research interests include object detection and segmentation based on deep learning.

HUA ZHENG received the Ph.D. degree in optical engineering from Zhejiang University, Hangzhou, China, in 2007. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. Since 2010, he has been an Associate Professor with the College of Photonic and Electronic Engineering, Fujian Normal University. He has authored or coauthored over 30 papers and holds several patents. His research interests include image processing, weak signal detection, signal processing, wireless communication, embedded systems, and the design of optical intelligence instruments.