Coupled Network for Robust Pedestrian Detection With Gated Multi-Layer Feature Extraction and Deformable Occlusion Handling

Tianrui Liu, *Member, IEEE*, Wenhan Luo, *Member, IEEE*, Lin Ma, *Member, IEEE*, Jun-Jie Huang, *Member, IEEE*, Tania Stathaki, *Member, IEEE*, and Tianhong Dai

***Abstract*— Pedestrian detection methods have been significantly improved with the development of deep convolutional neural networks. Nevertheless, detecting ismall-scaled pedestrians and occluded pedestrians remains a challenging problem. In this paper, we propose a pedestrian detection method with a couple-network to simultaneously address these two issues. One of the sub-networks, the gated multi-layer feature extraction sub-network, aims to adaptively generate discriminative features for pedestrian candidates in order to robustly detect pedestrians with large variations on scale. The second sub-network targets on handling the occlusion problem of pedestrian detection by using deformable regional region of interest (RoI)-pooling. We investigate two different gate units for the gated sub-network, namely, the channel-wise gate unit and the spatio-wise gate unit, which can enhance the representation ability of the regional convolutional features among the channel dimensions or across the spatial domain, repetitively. Ablation studies have validated the effectiveness of both the proposed gated multi-layer feature extraction sub-network and the deformable occlusion handling sub-network. With the coupled framework, our proposed pedestrian detector achieves promising results on both two pedestrian datasets, especially on detecting small or occluded pedestrians. On the CityPersons dataset, the proposed detector achieves the lowest missing rates (i.e. 40.78% and 34.60%) on detecting small and occluded pedestrians, surpassing the second best comparison method by 6.0% and 5.87%, respectively.**

***Index Terms*— Pedestrian detection, coupled network, gated feature extraction, squeeze network, multi-layer feature, occlusion handling, deformable RoI-pooling.**

I. Introduction

PEDESTRIAN detection has long been an attractive topic in computer vision and has significant impact on both research and industry. Pedestrian detection is essential to the understanding of a scene, and has a wide range of applications such as video surveillance [1], [2], robotics automation [3], [4] and intelligent driving assistance systems [5]. Pedestrians are usually within a complex environment, like for example images captured by an intelligence vehicle driving in urban scenarios mostly having changing and complex backgrounds. Apart from the complex backgrounds, the pedestrian detection task is also challenged by a large variety of poses and appearances in real life scenarios. Depending on how close the pedestrian is to the camera, the size of pedestrians in the captured images can vary within a large range. Partial occlusions between pedestrians and their surrounding instances under crowded environments further raise additional challenges to this task.

Recent advances of Deep Neural Networks (DNNs) have made significant improvements on the performance of the pedestrian detection task. The detection difficulties due to complex environment and high variances on appearances can be largely resolved by powerful feature representations generated by DNNs. Nevertheless, as investigated in [6], it still remains challenging to detect pedestrians of small size or being heavily occluded. For small pedestrians, the image resolution is relatively low, therefore there is less visual information which can be unitized for feature representation and subsequent classification and location estimation. For detecting pedestrians with occlusions, if a rigid detection model is used for the holistic body structure, normally a very low confidence score will be obtained for the detecting window which may lead to missing detections.

Fusing global and local information together has been proven to be effective in many visual tasks. The Deformable Part Models (DPM) detector [7] is a successful example of incorporating a global model and local part models. Similarly, in [8] a pyramid pooling module is designed to effectively extract a hierarchical global contextual prior. The pyramid module is then concatenated with the local convolutional network features to improve the performance on scene parsing tasks. In [9], two sibling networks are coupled together in which one sibling network targets on extracting global information and the other one targets on extracting local information. The global sibling network applies region of interest (RoI) pooling [10] to predict a global score of this RoI; the local sibling network uses position-sensitive region of interest (PSRoI) pooling [11] to generate score maps which are

Manuscript received December 14, 2019; revised August 10, 2020 and October 8, 2020; accepted November 2, 2020. Date of publication November 25, 2020; date of current version December 4, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding author: Jun-Jie Huang.)*

Tianrui Liu was with the Department of Electrical and Electronic Engineering, Imperial College London (ICL), London SW7 2AZ, U.K. She is now with the Department of Computing, ICL, London SW7 2AZ, U.K. (e-mail: t.liu15@imperial.ac.uk).

Wenhan Luo is with the Tencent AI Lab, Shenzhen 518000, China (e-mail: whluo.china@gmail.com).

Lin Ma is with Meituan, Beijing 100102, China (e-mail: forest.linma@gmail.com).

Jun-Jie Huang and Tania Stathaki are with the Department of Electrical and Electronic Engineering, ICL, London SW7 2AZ, U.K. (e-mail: j.huang15@imperial.ac.uk).

Tianhong Dai is with the Department of Biological Engineering, ICL, London SW7 2AZ, U.K.

Digital Object Identifier 10.1109/TIP.2020.3038371

sensitive to local parts of the object. By coupling the global confidence with the local part confidence together, one can obtain a more reliable prediction.

Targeting on developing a robust pedestrian detection method which can simultaneously address the problems of detecting pedestrians of small size or with partial occlusions, we propose a pedestrian detection framework with two coupled networks. Each of the two sub-networks is focusing on solving one problem for pedestrian detection. The gated multi-layer feature extraction sub-network aims to adaptively generate discriminative features for different pedestrian candidates in order to robustly detect pedestrians with large variations on scales; the deformable occlusion handling sub-network targets on handling the occlusion problem of pedestrian detection by using deformable regional RoI-pooling.

The proposed gated multi-layer feature extraction sub-network consists of squeeze networks and gate networks. A squeeze network is applied to reduce the dimension of RoI feature maps pooled from each convolutional layer. It is an essential component in the gated multi-layer feature extraction network which helped to achieve a good balance between performance and model complexity. What follows is a gate network applied to the feature maps to decide whether features from this layer are essential for representing this RoI. We investigate two gate units to manipulate and select features from multiple layers of a DNN, i.e., a spatio-wise gate unit and a channel-wise gate unit. The expectation is that features manipulated by the proposed two gate units should be able to possess stronger inter-dependencies among channels and among spatial locations, respectively.

In the deformable occlusion handling sub-network, we propose to use deformable regional RoI-pooling which can better fit the non-rigid parts of human body than using the traditional RoI-pooling [10]. The deformable occlusion handling network is fully convolutional and with deformable RoI pooling. Deformable RoI-pooling uses shiftable pooling grids which can adapt to the local parts of deformable pedestrians. The offsets of the shifting pooling grids are learnt through additional convolutional layers in the deformable occlusion handling sub-network.

The contributions of this work are as follows. First, we propose a coupled framework for pedestrian detection which can address problem of detecting small pedestrians and detecting pedestrians with occlusions simultaneously. Second, the proposed gated feature extraction sub-network can adaptively extract multi-layer convolutional features for pedestrian candidates. The two different types of gate units we proposed can manipulate the RoI feature maps in the channel-wise and spatio-wise manner, respectively. Thirdly, the deformable occlusion handling sub-network enables more robust detection through a deformable RoI-pooling which can adaptively adjust the relative positions of the pooling grids for body parts which can better fit the deformable/occluded pedestrians. With the coupled detection framework, the two sub-networks use two complimentary ways of detection to reinforce the robust detection results, leading to state-of-the-art performance on the challenging CityPersons [12] and Caltech [13] pedestrian datasets.

II. Background

A. Traditional Pedestrian Detection Methods

Traditional features that have been exploited for pedestrian detection include Haar-like features [14], Scale-Invariant Feature Transform (SIFT) [15], Local Binary Pattern (LBP) [16], edgelets [17], and Histogram of Oriented Gradient (HOG) [18], etc. Among them, HOG and its variations, such as Aggregated Channel Features (ACF) [19], Local Decorrelated Channel Features (LDCF) [20] and Checkerboards [6], are considered as arguably the most successful hand-engineered features for pedestrian detection. These features are used in conjunction with a classifier, for instance boosted forests, to perform pedestrian detection via classification.

Traditional detection approaches [18]–[20] which use a holistic model for the entire pedestrian body structure normally do not work well when the pedestrian is partially occluded. Deformable Part Models (DPM) [7] is one of the most successful early attempts on occlusion handling for pedestrian detection. It applies a part-based model which contains a root-filter (analogous to the HOG filter) and a set of part-filters associated with deformation costs measuring the deviation of each part from its ideal location.

In [21], the DPM framework has been extended for crowded pedestrian detection by using a two-pedestrian detector to reinforce the detection scores of the single-pedestrian detector. Similarly, a set of occlusion patterns of pedestrians have been explored in [22] to deal with inter-pedestrian occlusions.

The above methods were the dominant approaches for pedestrian detection before the emerging of deep neural networks based pedestrian detection methods. The traditional methods utilise hand-engineered features which are difficult to have sufficient feature representation ability for the challenging pedestrian detection scenarios.

B. Deep Neural Network Based Pedestrian Detection Methods

Deep Neural Networks (DNNs) based pedestrian detection methods have demonstrated superior detection performance. One of the main reasons is that DNNs are able to generate features with a stronger discrimination capability through end-to-end training when compared to the hand-engineered feature representations.

Region Convolutional Neural Networks (R-CNN) method [23] and its variations [10], [24] have achieved an excellent detection accuracy in general object detection tasks. These methods make use of a two-stage detection strategy. That is, a small number of highly potential candidate regions are first proposed using a candidate proposal method and then classification is performed based on the extracted CNN features from these candidate regions. R-CNN [23] and Fast R-CNN [24] apply selective search [25] for region proposal, while Faster R-CNN [10] replaces selective search with a built-in Region Proposal Network (RPN) that can effectively generate proposals.

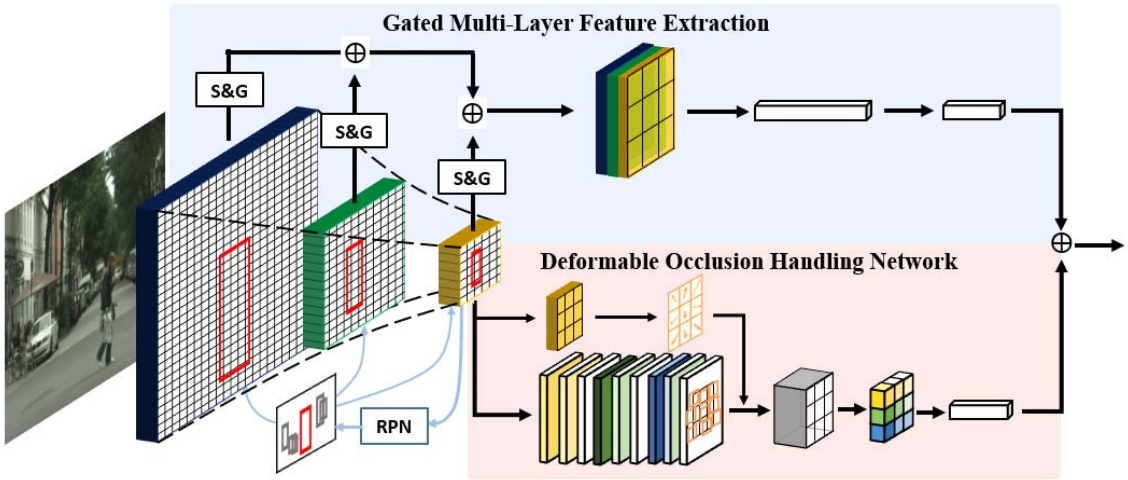


Fig. 1. Overview of the proposed pedestrian detection framework. The backbone network is the VGG16 network which has five convolutional blocks. For illustration purposes, only 3 convolutional blocks are shown in this figure. The Region Proposal Network (RPN) is used to generate a number of pedestrian proposals which will be carefully examined. The proposed pedestrian detection framework contains two coupled sub-networks: the upper sub-network is a gated multi-layer feature extraction network which can provide highly discriminative CNN features especially for small pedestrians under complex backgrounds; the lower sub-network is a deformable occlusion handling stream which applies deformable RoI-pooling and is used to generate robust features for occluded pedestrians. The "S&G" in the upper sub-network denotes the squeezing network and the gated network.

As investigated in [26], directly applying Faster-RCNN for pedestrian detection does not lead to satisfying results since small pedestrians which dominate common pedestrian datasets are usually not well detected. In [12], [26], [27], the detection anchors of the Faster-RCNN detector are tailored in order to accommodate for the pedestrian detection task. However, the low feature resolution at a deep layer limits pedestrian detectors from accurately detecting pedestrians with small size.

To take advantage of the CNN features from multiple layers, [26] proposes to concatenate features from multiple layers of a CNN and replace the downstream classifier of Faster-RCNN with boosted forests to improve performance on detecting hard samples. In [26] and [28], a fixed feature combination is used for all pedestrians. Scale-aware multi-resolution RCNN [29] method proposes to choose the most suitable feature combination to detect pedestrians according to their scales. The limitation is that the multi-layer feature combinations is not automatically learned and only considers a limited scale range. In our recent work [30], we study the automatic selection of combination of multi-layer features for detecting pedestrians of various sizes.

There are also DNN-based methods in the literature targeting on the problem of detecting pedestrians with occlusions. DeepPyramid DPM method [31] utilizes features extracted from a learned CNN feature extractor and has shown improved performance compared to the traditional DPM-based methods [7], [21], [22]. A multi-label learning method [32] is proposed to improve the performance of part detectors and reduce the computational cost of integrating multiple part scores. However, the body parts used in these methods are manually designed, which may limit their performance. Some recent pedestrian detection methods [33], [34] introduce novel loss functions to solve the occlusion problem when detecting pedestrians in crowded scenarios. In occlusion-aware RCNN [34], an occlusion-aware region-of-interest (RoI) pooling unit is proposed to integrate the visible prediction into the network for occlusion handling. Five part anchors at hand-selected positions are used for pooling the regional features of five body parts. The problem is that pedestrians are not rigid object and can deform under different poses and when facing different directions.

III. Method

A. Overview of the Coupled-Network

Our motivation is to build a pedestrian detection framework which is robust for detecting pedestrians of small scale and with possible occlusions. In order to jointly achieve these two goals, the proposed detection network couples two sibling sub-networks, that is, a gated multi-layer feature extraction network and a deformable occlusion handling network.

Fig. 1 shows the block diagram of the proposed pedestrian detection framework. The 13 convolutional layers of the VGG16 network [35] are employed as the backbone network whose the second, forth, seventh, tenth and thirteenth convolutional layers are followed by a max-pooling layer. The convolutional layers before each of these pooling layers can be regarded as convolutional blocks. The mulit-layer CNN features are extracted from the last convolutional layer of each block, i.e., $Conv1\_2$, $Conv2\_2$, …, $Conv5\_3$. For simplicity, we refer them as $Conv1$, $Conv2$, …, $Conv5$. A Region Proposal Network (RPN) is used to generate a number of pedestrian proposals which will be carefully examined. There are two coupled sub-networks which are built on top of backbone network and used for jointly learning and inference. The upper sub-network is the gated multi-layer feature fusion network which can adaptively provide discriminative features for each pedestrian candidate, and the

lower sub-network is a deformable occlusion handling network which uses deformable RoI-pooling to produce robust features for occluded pedestrians. In order to balance the contribution of the two sub-networks in training as well as inference stages, we perform normalization using an additional convolution layer at each sub-network and use element-wise summation for network coupling. The entire detection framework can be end-to-end trained.

In the following two sub-sections, we will describe the implementation details of the gated multi-layer feature extraction sub-network and the regional occlusion handling sub-network, respectively.

B. Gated Multi-Layer Feature Extraction Sub-Network

The large and small size pedestrians in a captured image differ in two main aspects, that is, spatial resolution and visual appearance. Feature maps from different layers of a Convolutional Neural Network (CNN) have different reception field and spatial resolution and represent different levels of abstraction. Therefore, for pedestrians of different sizes, features that can best balance the representation ability and the feature resolution should come from different CNN layers. For this reason, it is a good intuition to use different feature representations for detecting objects of different sizes. In [29], the authors presented a Scale-Aware Multi-resolution (SAM) method which achieves a good feature representation by choosing the most suitable feature combinations for pedestrians of different scales. However, the multi-layer feature combinations are not automatically learned and there can only be a limited number of simple combinations of features to be tested.

In this paper, we aim at investigating a more advanced approach which can learn to select the best multi-layer feature combination for detecting pedestrians of various sizes. The proposed gated multi-layer feature extraction sub-network takes the features from all the five convolutional blocks as the input and will thereafter select the most discriminative feature components for different pedestrian candidates based on features from different layers. The gated multi-layer feature extraction sub-network realizes an automatic re-weighting of the multi-layer features from different layers of the backbone network using gate units. Nevertheless, the gated network requires additional convolutional layers which induce a deeper RoI-wise sub-network at the cost of a higher complexity and a higher memory occupation. To remedy this issue, our gated sub-network includes squeeze units which are used to reduce the dimension of feature maps.

As illustrated in Fig. 2, features maps from each convolutional block of the backbone network are first compressed by a *squeeze network*, then the RoI features pooled from the squeezed feature maps are passed through *gate networks* for feature selection, and are integrated at the concatenation layer. The output of the RoI pooling layers at all the convolutional blocks are feature tensors of the same spatial size but with different channel dimensions. The concatenation is then performed along the channel dimension.

A squeeze unit is used to transform input feature maps to a lightweighted representation. Let us denote the input feature maps as $\boldsymbol{F} = \left[\boldsymbol{f}_1, \cdots, \boldsymbol{f}_{C_{in}}\right] \in \mathbb{R}^{H \times W \times C_{in}}$ which has spatial size $H \times W$ and is of C_{in} channels. The squeezed feature maps are denoted as $\widehat{\boldsymbol{F}} = \left[\widehat{\boldsymbol{f}}_1, \cdots, \widehat{\boldsymbol{f}}_{C_{out}}\right] \in \mathbb{R}^{H \times W \times C_{out}}$ and has spatial size $H \times W$ of C_{out} channels with $C_{out} < C_{in}$. $\boldsymbol{F}$ is squeezed to $\widehat{\boldsymbol{F}}$ through 1×1 convolution:

$$\widehat{\boldsymbol{f}}_i = \boldsymbol{v}_i * \boldsymbol{F}, \tag{1}$$

where $\widehat{\boldsymbol{f}}_i$ is the i-th feature element of the output feature map, $\boldsymbol{v}_i$ is the i-th learned filter in the squeeze network for $i = 1, \cdots, C_{out}$, and '$*$' denotes convolution.

The reduction ratio of the squeeze network is defined as $r = \frac{C_{in}}{C_{out}}$. We find experimentally that by properly selecting the reduction ratio r the squeeze network can reduce the RoI-wise sub-network parameters without noticeable performance downgrading. Taking VGG16 $Conv4$ feature maps as an example, $Conv4$ feature maps which are of $C_{in} = 512$ channels can be squeezed by a ratio of $r = 2$ so as to be reduced to the lightweighted feature maps with $C_{out} = 256$ channels.

The RoI pooling will be performed on the squeezed lightweighted feature maps. The RoI pooled features will then be passed through a gate unit for feature selection.

A gate network will be used to manipulate the RoI pooled features to highlight the most suitable feature channels or feature components for a particular RoI, while suppressing the redundant or unimportant ones. As shown in the block diagram in Fig. 2, the output of a gate unit is used to perform an element-wise product with the RoI pooled features, deciding how the input feature would be manipulated.

In general, a gate unit consists of a convolutional layer, two fully connected (fc) layers and a Sigmoid function at the end for output normalization. The convolutional layer and fc layers are associated with ReLU activation functions for non-linear mapping.

Given RoI pooled feature maps $\boldsymbol{R}$, the output of a gate unit $\boldsymbol{G}$ can be expressed as:

$$\boldsymbol{G} = \sigma\left(\boldsymbol{W}_2 \delta\left(\boldsymbol{W}_1 \delta\left(\boldsymbol{C}_1 * \boldsymbol{R}\right)\right)\right), \tag{2}$$

where $\{\boldsymbol{C}_1, \boldsymbol{W}_1, \boldsymbol{W}_2\}$ are the learnable parameters of the gate unit; $\sigma(\cdot)$ denotes the Sigmoid function and $\delta(\cdot)$ denotes the ReLU activation function [36]. $\boldsymbol{G} \in \mathbb{R}^{h_g \times w_g \times c_g}$ is of spatial size $h_g \times w_g$ and has c_g channels.

The output of a gate unit $\boldsymbol{G}$ will be used to manipulate the regional feature maps $\boldsymbol{R}$ through an element-wise product:

$$\widehat{\boldsymbol{R}} = \boldsymbol{G} \odot \boldsymbol{R}, \tag{3}$$

where $\odot$ denotes the element-wise product.

The manipulated features $\widehat{\boldsymbol{R}}$ have the same size as its input RoI pooled feature maps $\boldsymbol{R}$, and will have enhanced information that is helpful for identifying the pedestrian within this RoI.

We have designed two types of gate units based on how the RoI pooled feature maps will be manipulated, namely, a spatio-wise selection gate unit and a channel-wise selection gate unit. The expectation is that the features manipulated

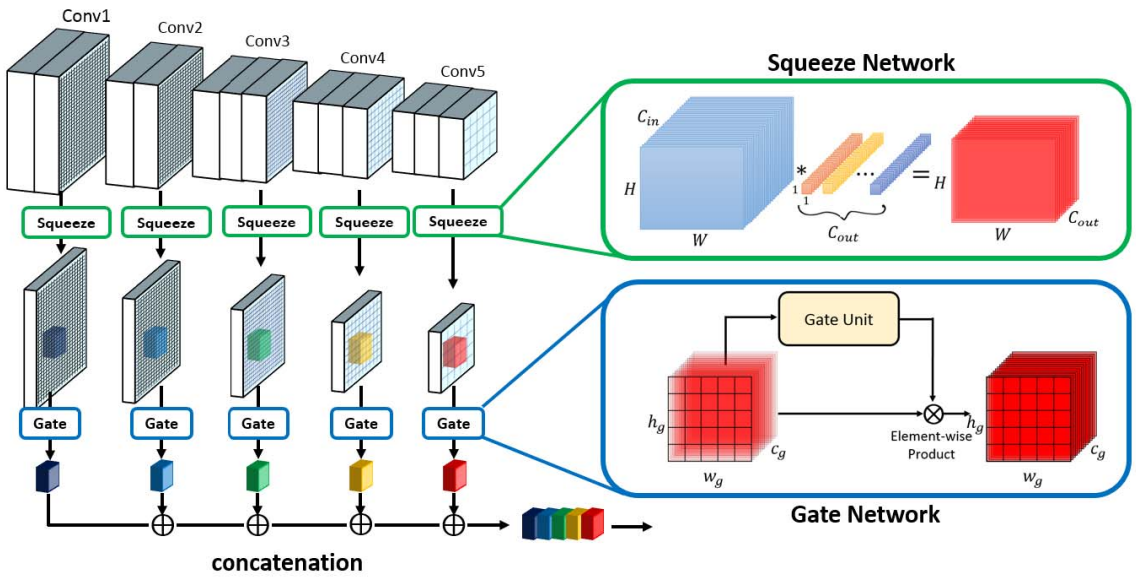


Fig. 2. Overview of the proposed gated multi-layer CNN feature extraction network. Feature maps from each convolutional block of the backbone network are first compressed by the **squeeze networks** for dimension reduction. The squeezed lightweighted feature maps are then passed through **gate networks** for feature selection, and are integrated at the concatenation layer. The gate units will manipulate the CNN features to highlight the most suitable feature channels or feature components for a particular RoI, while suppressing the redundant or unimportant ones.

by the output of the spatio-wise selection gate unit and the channel-wise selection gate unit will be able to have increased inter-dependencies among different channels and among different spatial locations, respectively. Therefore, the manipulated features will be more adaptive to the content within this RoI in terms of spatial variances and visual appearance variances, respectively. By applying the two different gate models, the proposed pedestrian detector performs differently.

1) Spatio-Wise Selection Gate Unit: The output of a spatio-wise selection gate unit will be used to enhance the features at spatial locations where the features are essential for detecting pedestrians within this RoI. By reinforcing features at important spatial locations, the learned features are expected to be more robust for detecting pedestrians with partial occlusions.

The spatio-wise selection gate unit outputs a 2-dimensional (2D) map $\boldsymbol{G}$ of size $(h_g, w_g, c_g) = (h, w, 1)$. It will be used to perform an element-wise product with the RoI pooled feature maps $\boldsymbol{R}$ which is of size $h \times w \times c$ through a 1×1 convolution.

As shown in Fig. 3, through 1×1 convolution, the resulting 2D map has the same spatial resolution as the input feature. The 2D map is then passed through two fully connected (fc) layers and a Sigmoid function for normalization. The obtained 2D spatial mask $\boldsymbol{G}$ will be used to modulate the feature representation for every spatial location of the input feature. The feature values from all C feature channels at spatial location (i, j) will be modulated by the coefficient $\boldsymbol{G}(i, j, 1)$.

2) Channel-Wise Selection Gate Unit: The output of a channel-wise selection gate unit will be used to enhance the feature channels that are important for detecting pedestrians within this RoI. By reinforcing essential feature channels, the learned features are expected to be more adaptive to the feature resolution and visual appearance of the object.

The channel-wise selection gate unit generates a vector of size $(h_g, w_g, c_g) = (1, 1, C)$ through n depth-wise separable convolution [37]. As shown in Figure 4, this vector is further passed through two fully connected layers and a Sigmoid function. The obtained $\boldsymbol{G}$ thereafter is used to perform a modulation with the convolutional features along the channel dimension. All the feature values within the k-th ($k \in [1, C]$) channel will be modulated by the k-th coefficient of $\boldsymbol{G}(1, 1, k)$.

An illustration of depth-wise separable convolution is given in Fig. 5. There are c_g numbers of kernels applied separately on the input feature map of size $h_g \times w_g \times c_g$. Each filter, of size $h_g \times w_g \times 1$, is convolved with a single channel of the input feature, resulting in a map of size $h_g \times w_g \times 1$. Then the separate output maps are stacked together to generate a vector of $h_g \times w_g \times 1$.

C. Deformable Occlusion Handling Sub-Network

To better solve the occlusion problem, we propose a sibling deformable occlusion handling sub-network with deformable regional RoI-pooling [38] which will be able to generate robust features for pedestrians with partial occlusions. The deformable regional RoI-pooling performs regional RoI pooling on shiftable local grids which can better adapt to the position variations of the pedestrian parts, therefore will be capable of better handling occlusion problems in pedestrian detection.

1) Baseline of the Occlusion Handling Sub-Network: The baseline network for this pipeline is the Region-based Fully Convolutional Network (RFCN) [11]. The idea of the RFCN detector is based on the intuition that one can localize an object even only with partial information of the object.

As given within the green box of Fig. 6, different parts of an object is considered separately by generating $C_{cls} \times k \times k$

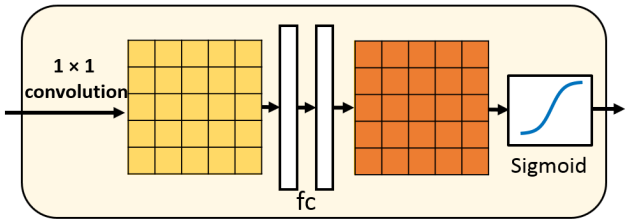


Fig. 3. The gate unit for spatio-wise selection. In this gate unit, the (squeezed) RoI-pooled features will be transformed to a 2D map via 1×1 convolution. The 2D map is followed by two fully connected layers to generate another 2D map to be fed into the Sigmoid function. The output of this gate unit will be used to perform element-wise production in spatial domain with its input RoI features as given in Fig. 2.

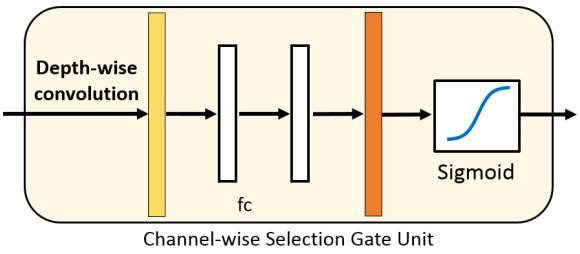


Fig. 4. The gate unit for channel-wise selection. In this gate unit, the (squeezed) RoI-pooled features will be transformed to a 1D vector via depth-wise separable convolution. This vector is followed by two fully connected layers to generate another vector to be fed into the Sigmoid function. The output of this gate unit will be used to perform element-wise production in the channel direction with its input RoI features.

regional position sensitive feature maps, each considering feature representation corresponding to an object class and at a grid location on a $k \times k$ grid. For illustration, we use $k = 3$ hence there are $3 \times 3 = 9$ local grids in an RoI. The relative positions of each local grid with respect to the RoI are top-left, top-center, top -right, …, bottom-right.

Let us denote with $\boldsymbol{S}^c_{(i,j)}$ the feature maps corresponding to class c and grid location (i, j). The detection score is determined by combing the votes from these regional position sensitive feature maps. This operation is named Position-Sensitive RoI (PSRoI) pooling. PSRoI pooling enables the extraction of a fixed length feature representation for objects of arbitrary size from the $C_{cls} \times k^2$ regional position sensitive feature maps. For class c and grid (i, j), the extracted feature is obtained as the average of the feature maps $\boldsymbol{S}^c_{(i,j)}$ within grid (i, j), i.e.,

$$\boldsymbol{s}(c, i, j) = \frac{1}{n_{c,i,j}} \sum_{\boldsymbol{p} \in grid(i,j)} \boldsymbol{S}^c_{(i,j)}(\boldsymbol{p}_0 + \boldsymbol{p}), \tag{4}$$

where $\boldsymbol{p}_0$ is set to the position of the upper-left grid, $\boldsymbol{p}$ enumerates spatial positions within the grid location (i, j), and $n_{c,i,j}$ is the number of pixels in grid (i, j) and is used as a normalization term.

2) Occlusion Handling Sub-Network Using Deformable Regional RoI-Pooling: The proposed deformable occlusion handling sub-network applies deformable PSRoI pooling within regional-based fully convolutional network. Possible deformation and occlusion of the candidate pedestrian would be more adaptively handled through learned shiftable pooling grids for RoI pooling. The schematic representation for our proposed occlusion handling sub-network is shown in Fig. 6.

Given the convolutional feature maps of the input image that have been extracted from the backbone network, as well as the candidate pedestrian proposal RoI output from the RPN network, the feature maps are first passed through a convolutional layer to generate a bank of regional position sensitive feature maps. For pedestrian detection, there are only $C_{cls} = 2$ classes, i.e., pedestrian class and background class. Therefore, the deformable occlusion handling sub-network will generate $2k^2$ regional position sensitive feature maps, with k^2 feature maps responsible for the pedestrian class and the other k^2 feature maps for the background class.

We apply deformable PSRoI pooling [38] to obtain a fixed length feature representation for each RoI. Deformable PSRoI pooling is more adaptive than the regular PSRoI pooling [24] and has adaptive and shiftable pooling grids for each candidate pedestrian proposal. The deformable PSRoI pooling uses shiftable pooling grids which can adapt to the local parts deformation of pedestrians. The extracted feature $\boldsymbol{s}'(c, i, j)$ is the average of the feature maps $\boldsymbol{S}^c_{(i,j)}$ within a shifted grid:

$$\boldsymbol{s}'(c, i, j) = \frac{1}{n_{c,i,j}} \sum_{\boldsymbol{p} \in grid(i,j)} \boldsymbol{S}^c_{(i,j)}(\boldsymbol{p}_0 + \boldsymbol{p} + \Delta p_{ij}), \tag{5}$$

where offsets Δp_{ij} for grid location (i, j) of the RoI are learnt to shift the pooling grids.

As shown in Fig. 6, the offsets Δp_{ij} are learnt through additional convolutional layer (see the orange pathway). The feature maps pooled from each shifted grid are used to vote for the detection score.

With the deformable PSRoI pooling, the regional features can be flexibly pooled from a set of RoI bins in order to better cover foreground instance (i.e. the candidate pedestrian) region. The localization capability can therefore be enhanced, especially for occluded and deformable pedestrians.

D. Network Training

The gated multi-layer feature extraction sub-network and the deformable occlusion handling sub-network are coupled to achieve a simultaneous robust detection on small scaled pedestrians and partially occluded pedestrians. To have a balanced energy on the feature generated by two sub-networks, the feature representation from each sub-network will go through a 1×1 convolution before they are combined through element-wise summation. The combined feature representation will then be used for pedestrian classification and bounding box regression.

1) Label Assignment: The training classification labels are assigned based on the overlapping ratio between a proposal and the ground-truth bounding box. We assign a binary class label to each proposal for network training as in Faster-RCNN [10]. A positive label will be assigned if it overlaps with any ground-truth proposal with Intersection-over-Union (IoU) higher than 0.7, or has the highest IoU with a ground-truth proposal. Proposals will be regarded as negative if the maximum IoU with all ground-truth proposals is lower than 0.3.

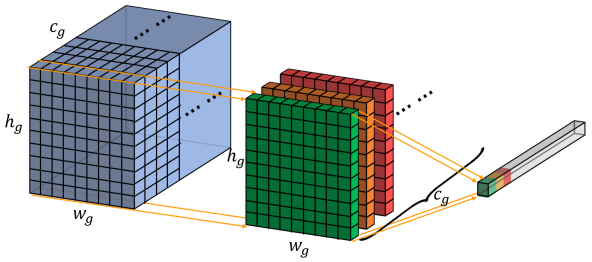


Fig. 5. Illustration of depth-wise separable convolution. There are c_g numbers of kernels applied separately onto the input feature map of size $h_g \times w_g \times c_g$. Each filter, of size $h_g \times w_g \times 1$, is convolved with a single channel of the input feature, resulting in a map of size $h_g \times w_g \times 1$. Then the separate output maps are stacked together to generate a vector of $h_g \times w_g \times 1$.

2) Loss Function: Pedestrian detection aims to determine whether the content within a candidate region is a pedestrian or not and estimate an accurate bounding box coordinate for pedestrian candidates. The loss function of the couple network contains a classification loss term $\mathcal{L}_{cls}$ and a regression loss term $\mathcal{L}_{reg}$ which are summed over all the N proposal regions within a training batch data:

$$\mathcal{L}_{couple} = \sum_{i=1}^{N} \mathcal{L}_{cls}(C_i^*, C_i) + \alpha \mathcal{L}_{reg}(B_i^*, B_i), \tag{6}$$

where C_i is the predicted probability of the i-th candidate bounding box region being a pedestrian and B_i is the predicted bounding box coordinates; C_i^* and B_i^* are the ground-truth label of candidate region and the ground-truth bounding box positions, respectively; α is a scalar and is used to adjust the contributions of the two terms, and is empirically set as 1; $\mathcal{L}_{cls}(\cdot, \cdot)$ denotes the classification loss term which is cross entropy loss over pedestrian class and non-pedestrian class, and $\mathcal{L}_{reg}(\cdot, \cdot)$ denotes regression loss term which is a smoothed L_1 loss, i.e.,

$$\mathcal{L}_{\text{reg}}\left(B^*, B\right) = \sum_{j \in \{x, y, w, h\}} \text{smooth}_{L_1}\left(B_j^* - B_j\right), \tag{7}$$

where x, y, w, and h denote the box's center coordinates and its width and height. smooth_{L_1} is the is a smoothed L_1 loss of t, i.e.,

$$\text{smooth}_{L_1}(t) = \begin{cases} 0.5t^2 & \text{if } |t| < 1, \\ |t| - 0.5 & \text{otherwise.} \end{cases} \tag{8}$$

3) RPN: In terms of the RPN for candidate proposal, we use 9 anchors of different scales with single aspect ratio of $\gamma = 0.41$.[1] The RPN slides over the convolutional feature maps output from the feature extraction network to perform box regression and classification simultaneously. The cost function for training the RPN network also contains a classification loss term and a regression loss term. In our work, we train the RPN and coupled network jointly in an end-to-end manner. That is, the loss terms of the coupled network and the RPN are summed up together for back-propagation.

[1]The ratio $\gamma = 0.41$ is the average width-height ratio of pedestrian bounding box in pedestrian benchmarks (i.e., the Caltech [13] dataset and the Citypersons dataset [12]).

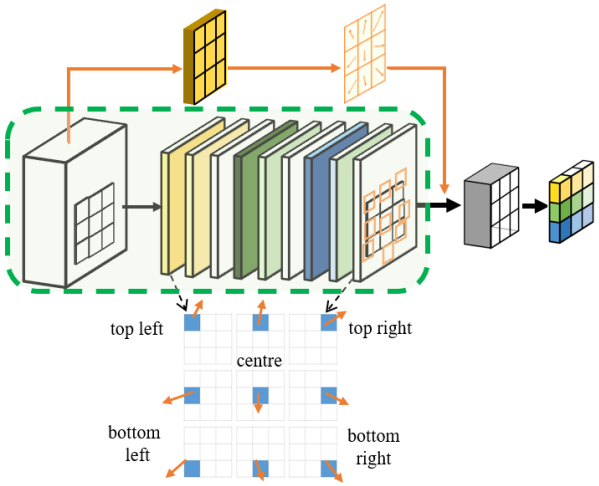


Fig. 6. The proposed deformable occlusion handling sub-network. The green box illustrates the position sensitive feature maps. Taking the CNN feature maps of an entire image as the starting point, a bank of $2k^2$ regional feature maps are generated by convolution. Here we use $k = 3$ for illustration purpose. The $3 \times 3 = 9$ spatial grids describing the relative position of each region with respect to the RoI, *i.e.*, top-left, top-center, top-right, …, bottom-right. The orange pathway learns offsets through additional convolutional layer so that the feature maps pooled from each shifted grid can be used to vote for the occlusion score map.

4) Optimization: The network weights of the backbone network (i.e., from $Conv1$ to $Conv5$ convolutional blocks) are initialized from the network pre-trained using the ImageNet dataset [39], while the network weights of other convolutional layers are initialized as a Gaussian distribution with mean 0 and standard deviation 0.01. Stochastic Gradient Descent (SGD) with momentum is used to optimize the network weights of the propose pedestrian detection network. The learning rate of the algorithm is initialized at 1×10^{-3} and was reduced by a factor of 10 for two times during the training. The momentum λ is set to 0.9 and weight decay is set to 5×10^{-4}. During training, a single image is processed in each mini-batch, and for each image there are 256 randomly sampled proposals used to compute the loss for this mini-batch. The whole network is fully convolutional and benefits from end-to-end approximate joint training and multi-task learning.

IV. Experiments

A. Datasets and Evaluation Metrics

CityPersons: CityPersons [12] is a recent pedestrian detection dataset built on top of the CityScapes dataset [40] which is for semantic segmentation. The dataset includes 5, 000 images captured in several cities of Germany. There are about 35, 000 persons with additional ∼13,000 ignored regions in total. Both bounding box annotation of all persons and annotation of visible parts are provided. We conduct our experiments on CityPersons using the reasonable training/validation sets for training and testing, respectively.

We evaluated the CityPersons dataset under four subsets of different ranges of pedestrian size and difference occlusion levels as shown in Table I. Evaluations are measured using the

TABLE I

THE DEFINITION OF REASONABLE, SMALL, OCCLUSION, AND ALL SUBSETS IN CITYPERSONS DATASET

	Reasonable	Small	Occlusion	All
Height	$[50, \inf)$	$[50, 75]$	$[50, \inf)$	$[20, \inf)$
Visibility	$[0.65, 1]$	$[0.65, 1]$	$[0.2, 0.65)$	$[0.2, 1]$

log average miss rate (*MR*) of false positive per image (FFPI) ranging from 10^{-2} to 10^{0} (denoted as MR_{-2}).

Caltech: The Caltech-USA dataset [13] and the improved annotations [6] are used for training and evaluation. The training set contains 4,250 images which are obtained by extracting every 30^{th} frame from the Caltech videos. The new high quality annotations provided by [6] which correct some inaccurate annotations are used in our experiments. The testing set contains 4024 images of size 480×640. Following the evaluation of Caltech benchmark [13], only bounding boxes restricted in the range of $x \in [5, 635]$, $y \in [5, 475]$ are evaluated. We evaluate the detection performance on the following subsets: (1) Reasonable: height [50, inf), visibility [0.65, 1]; (2) Medium: height $\in$ [30, 80], visibility [0.2, 1]; (3) Heavy: height [50, inf), visibility [0.2, 0.65).

B. Experimental Setup

For the Citypersons dataset, we follow the same hyper-parameters in the Deform-ConvNet [38] source code for fair comparison: the number of epochs is 7, the learning rate starts with 1×10^{-3} and the scheduling step is at 5.333, the warm-up step is used with a smaller learning rate 1×10^{-4} for 4,000 min-batches. Online Hard Example Mining (OHEM) [41] is used for training. Among N proposals, only the top n ($n = 300$) RoIs which have the highest loss are used for back-propagation. For the Caltech dataset, we set the learning rate to 5×10^{-3}.

In order to reduce over-fitting we use data augmentation, which flips the images horizontally. All experiments are performed on a single TITAN X Pascal GPU.

C. Effectiveness of Squeeze Ratio

The squeeze ratio r affects the network in terms of feature capacity and computational cost. To investigate the effects of squeeze ratio, we conduct experiments using features from multiple convolutional layers that have been squeezed by $r = 1, 2, 4, 8, 16, 32$. The performances are compared in Table II.

We find that squeeze network can reduce the RoI-wise sub-network parameters without noticeable performance deduction. We use the reduction ratio $r = 2$ as it is a good trade-off between performance and computational complexity.

D. Effectiveness of the Gated Multi-Layer Feature Extraction Network

We compare our gated multi-layer feature extraction network with a modified version of Faster-RCNN detector [10] (we denote it as the *Baseline1* detector thereafter in this paper). The Faster-RCNN detector was modified in terms of the region proposal network (RPN). For general object detection [10], a three-scale three-ratio anchor is used to generate 9 proposals at each sliding position. For pedestrian candidate proposal, we use anchors of a single ratio of $\gamma = 0.41$ with 9 scales.

TABLE II

MISSING RATE (MR%) ON CITYPERSONS VALIDATION SET USING DIFFERENT SQUEEZE RATIOS r

squeeze ratio	Reasonable	Small	Occlusion	All
$r = 1$	14.49	39.65	56.97	43.70
$r = 2$	14.35	42.02	55.60	43.02
$r = 4$	14.63	44.33	56.34	42.93
$r = 8$	14.85	39.98	58.06	44.52
$r = 16$	14.72	40.18	53.97	43.07
$r = 32$	15.12	41.68	57.82	44.31

TABLE III

COMPARISON OF PEDESTRIAN DETECTION PERFORMANCE (IN TERMS OF MR%) USING DIFFERENT GATE UNITS ON THE CITYPERSONS DATASET

Model	Reasonable	Small	Occlusion	All
FasterRCNN[*Baseline1*]	16.44	40.46	56.19	44.6
Spatial-wise Gate	13.64	41.17	**52.37**	**40.65**
Channel-wise Gate	**13.49**	**37.62**	53.53	41.76

The *Baseline1* detector [10] only adopts the $Conv5\_3$ feature maps for feature representation and for the following classification and regression prediction. The limited feature resolution of $Conv5\_3$ restraints the capability for detecting small pedestrians. For our "spatio-wise gate" model and the "channel-wise gate" model in Table III, we use features extracted from the proposed gated multi-layer feature extraction network applying the two gate models respectively. We dilate the $Conv5$ features by a factor of two for both the *Baseline1* detector and our proposed gated sub-networks. The dilated convolution is proposed in [42] for better performance in semantic segmentation tasks. A dilated convolution "inflates" a filter by inserting spaces between the kernel elements. The technique enlarges the receptive field without increasing the filter size and hence, induce little additional costs. Using dilate convolution is crucial for detecting small instances.

As can be seen from Table III, both the spatio-wise gate model and the channel-wise gate model make improvements upon the *Baseline1* detector. These results demonstrate the effectiveness of our proposed gated multi-layer feature extraction method. More specifically, the spatio-wise gate model makes more improvements on the "Occlusion" subset, while the channel-wise gate model makes more improvements on the "Small" subset.

For our experiments on the coupled network in Section IV-G, we select the channel-wise gate model to be used in the gated multi-layer feature extraction sub-network. The reason is that the channel-wise gate model performs better on detecting small pedestrians, which is the main responsibility of the gated multi-layer feature extraction sub-network. The problem of detecting occluded pedestrians can be further addressed by the occlusion handling sub-network.

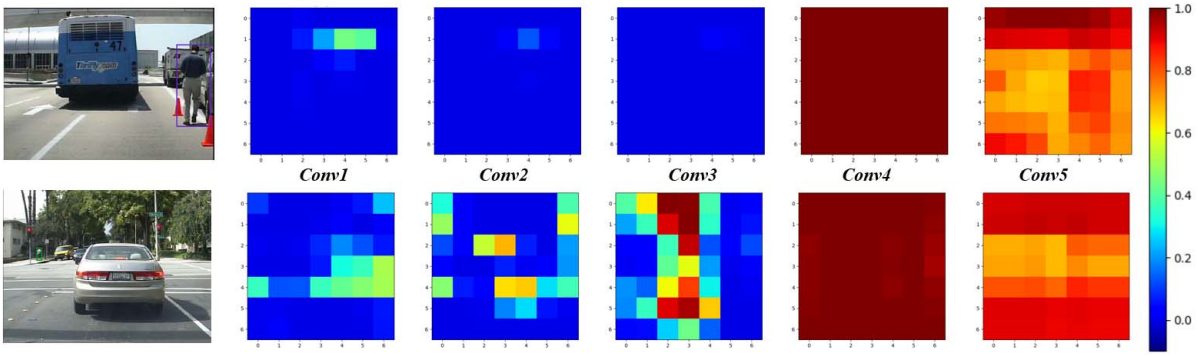


Fig. 7. Examples of visualization of the gated multi-layer feature selection using spatial-wise model. The learned 2D gate masks for a small and a large pedestrian are different. For the small pedestrian, the gate model select a portion of features at the $Conv1$ layer, which is barely use by the large pedestrian. On the other hand, the large pedestrian selects to use more of features at the $Conv5$ layer than the small pedestrian.

E. Visualization of Gated Multi-Layer Feature Selection

An example is given in Fig. 7 to visualize the spatio-wise gate selection. As described in Section III-B, the output of the gate unit for spatio-wise selection is 2D masks. As can be seen from Fig. 7, the learned gate masks for a small and a large pedestrian are different. For a small pedestrian, the spatio-wise gate model selects a portion of features at the $Conv1$, $Conv2$ and $Conv3$ layer, which are barely used for the large pedestrian. On the other hand, the large pedestrian selects to use more of features at the $Conv4$ and $Conv5$ layer. The results, as has been expected, indicate that it is beneficial for small pedestrians to make use of feature representations from shallow layers which have higher resolution, while large pedestrians can benefit from features from deeper layers which have higher levels of abstraction.

F. Effectiveness of Deformable RoI-Pooling for Occlusion Handling

We evaluate the effectiveness of deformable RoI-pooling for occlusion handling by comparing its performance on the Citypersons dataset to our *Baseline2* detector.

The *Baseline2* detector is a modified version of the RFCN [11]. We re-implemented the RFCN detector [11] method using the VGG16 backbone network. Although the original RFCN detector in [11] utilizes the ResNet [43] of 101 layers which achieves better results for general object detection than using the VGG16 network. We show in our experiments (see Table IV) that the performance for pedestrian detection using the ResNet is not as good as using VGG16. The reason is that the down-sampling rate for convolutional layers in the ResNets is too large for the network to provide feature maps with sufficient resolution to detect small pedestrians. Moreover, we modified the RPN network for pedestrian candidate proposal in the same manner as we have done to the *Baseline1* detector.

The results of the RFCN [11] detector (in terms of missing rate) on the validation set with/without using deformable RoI-pooling layer are compared in Table V. "RFCN + Deformable" denotes the improved version of RFCN which has been incorporated with deformable RoI-pooling for occlusion handling. As we can see, by applying deformable RoI-pooling, the improved RFCN detector has better performance in detecting pedestrian of the "Small" and "Occlusion" subsets. The overall performance, which can be evaluated using the "Reasonable" and "All" subsets, is therefore improved.

We give examples of visualization of deformable RoI-pooling method using in the occlusion handling sub-network for pedestrian detection in Figure 8. The upper example is a fully visible pedestrian, while the lower case is an occluded pedestrian.

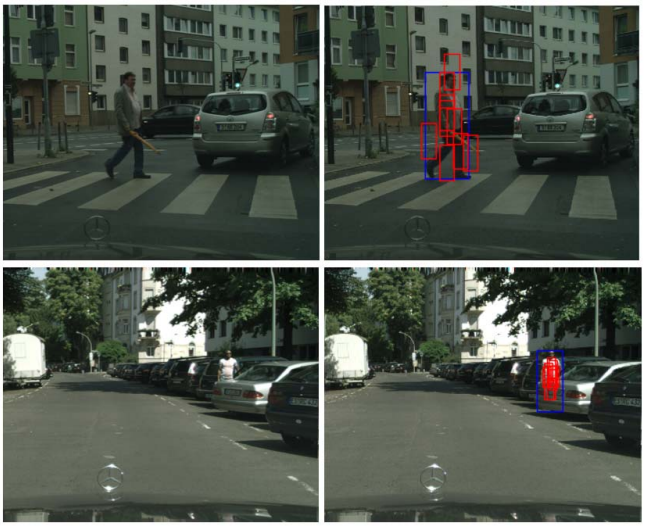
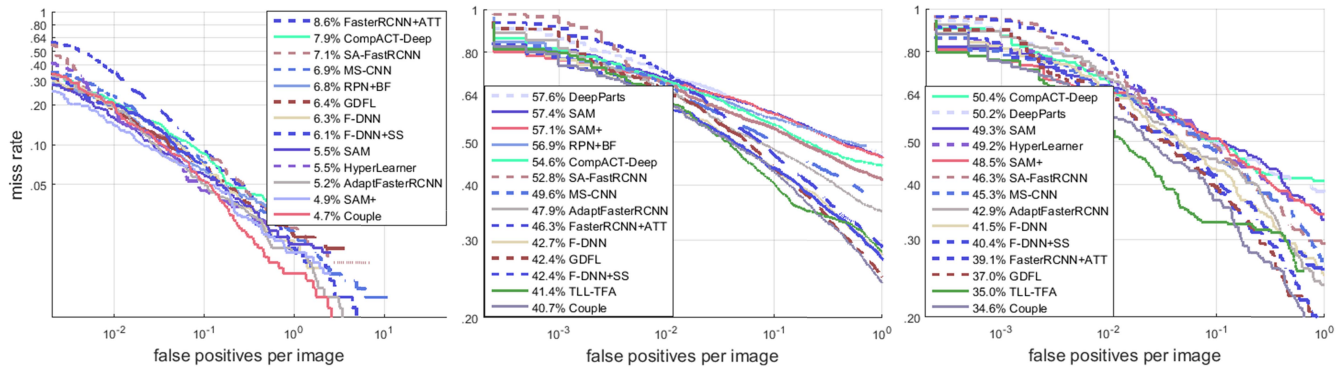


Fig. 8. Examples of visualization of deformable RoI-pooling the occlusion handling sub-network for pedestrian detection. The upper example is a fully visible pedestrian, while the lower case is an occluded the pedestrian.

TABLE IV

COMPARISONS OF THE PERFORMANCE (IN TERMS OF MISSING RATE MR%, THE LOWER THE BETTER) OF THE RFCN DETECTOR FOR PEDESTRIAN DETECTION USING DIFFERENT BACKBONE NETWORKS, I.E, RESNET101 AND VGG16, RESPECTIVELY

Model	Reasonable	Small	Occlusion	All
RFCN-VGG16	16.19	42.95	54.70	45.19
RFCN-ResNet101	18.80	48.16	60.53	47.84



(a) Reasonable (b) Medium (c) Heavy

Fig. 9. On the Caltech dataset under (a) Reasonable, (b) Medium and (c) Heavy evaluation protocols, our proposed method is compared with state-of-the-art pedestrian detection methods.

Fig. 10. Qualitative results from our proposed detector on the CityPersons dataset. Green bounding boxes denote the ground-truth and red ones denote predicted bounding boxes. The first and row give some detection examples of small-size pedestrians; the second and row give some detection examples of occluded pedestrians, the third row shows some more challenging cases where pedestrians are crowed. The proposed pedestrian detection method performs well even on these challenging cases.

TABLE V

COMPARISONS OF THE PERFORMANCE OF THE RFCN DETECTOR FOR PEDESTRIAN DETECTION WITH/WITHOUT USING DEFORMABLE ROI POOLING. "RFCN+DEFORMABLE" DENOTES THE IMPROVED VERSION OF RFCN WHICH HAS BEEN INCORPORATED WITH DEFORMABLE ROI POOLING FOR OCCLUSION HANDLING

Methods	Reasonable	Small	Occlusion	All
RFCN-VGG16 [Baseline2]	16.19	42.95	54.70	45.19
RFCN + Deformable	15.17	39.34	54.15	43.59

TABLE VI

COMPARING THE PROPOSED DETECTOR TO STATE-OF-THE-ART METHODS (IN TERMS OF MISSING RATE MR%, THE LOWER THE BETTER) ON CITYPERSONS VALIDATION SET

Methods	Reasonable	Small	Occlusion	All
FRCNN+ATT+part [44]	15.96	-	56.66	-
CityPersons [12]	15.40	-	-	-
TTL [45]	14.40	-	52.00	-
CoupleNet [Baseline3]	14.36	38.56	52.95	42.25
Repulsion Loss [33]	13.22	42.63	56.85	44.45
OR-CNN [34]	12.81	42.31	55.68	42.32
CoupleNet+Gate+Occlusion **[Proposed]**	**12.37**	**38.31**	**49.81**	**40.39**

G. Comparison to State-of-the-Art Methods

1) CityPersons Dataset: We compare the our proposed pedestrian detector with several state-of-the-art pedestrian detectors, including [44], FRCNN+ATT+part [44], Adapted Faster RCNN [12], TTL [45], Repulsion Loss [33], OR-RCNN [34]. The TTL [45] method, is proposed for small-scale

Fig. 11. Qualitative results from our proposed detector on the Caltech dataset. Green and red bounding boxes denote the ground-truths and our predictions, respectively. We show some challenging cases where pedestrians are small and occluded.

pedestrian detection using temporal feature aggregation. Repulsion Loss [33] and OR-RCNN [34] are proposed to address the occlusion problems for pedestrian detection.For fare comparison, all the performances are evaluated on the original image size (1024×2048) of the CityPersons validation dataset.

We also implement a *Baseline3* detector which is the coupled network of our *Baseline1* detector and *Baseline2* detector. That is, one of the sub-networks adopts the PSRoI pooling [11] in Baseline2, while the other sub-network employs the RoI pooling as in the *Baseline1* detector.

From Table VI, we can observe that the proposed method surpasses the other approaches on all the subsets. The most notable improvements are on the "Occlusion" and the "Small" subsets. Our methods outperforms the state-of-the-art method OR-RCNN [33] with an advantage of 6% on the "Small" subset, which highlights the effectiveness of our method for small-size pedestrian detection. In the case of the "Occlusion" subset including some severely occluded pedestrians, our proposed method achieves the best MR^{-2} performance of 49.81%), surpassing the second best pedestrian detector by a large margin of 5.87%. These results clearly demonstrate the effectiveness of our proposed method for occlusion handling.

2) Caltech Dataset: In Figure 9, our pedestrian detectors are compared with the state-of-the-art pedestrian detection methods, namely, MRFC [46], CompACTDeep [47], SA-FastRCNN [48], MS-CNN [49], RPN+BF [26], HyperLearner [50], OA-RCNN [34], F-DNN/F-DNN+SS [51], FasterRCNN+ATT [44], GDFL [52] and TLL-TFA [45]. On the "Reasonable" subset which is widely used to evaluate pedestrian detectors, the proposed method outperforms the second best method by 0.2 in terms of MR_{-2}. When it comes to the "Small" and "Occlusion" subsets, our method has achieved the best performance (i.e. 40.78% and 34.60%), surpassing the previous state-of-the-art method GDFL [52] and PDOE [53] which are recent competitive methods on detecting small occluded pedestrians. These improvements on the later two subsets indicate that our proposed coupled network is effective in detecting small-scale and occluded pedestrians.

H. Qualitative Results

1) CityPersons Dataset: Fig. 10 shows some exemplar detection results by our proposed pedestrian detector on the CityPersons dataset. In the first and second row, we show some examples of detected small pedestrians; in the third row we show some occlusion cases which have been successfully detected by our proposed method. From the qualitative results, we observe that our method is mostly successful on detecting even some small and heavily-occluded pedestrians. This demonstrates the effectiveness of our proposed coupled network on simultaneously addressing the problems of detecting small and occluded pedestrians in complex environment.

2) Caltech Dataset: We also show some qualitative results of our pedestrian detector on the Caltech dataset in Figure 11. Some challenging detection examples in this dataset are given. As can be seen, the proposed detection method can successfully detect small pedestrians even under crowed circumstances. For some extremely hard cases where the pedestrians are heavily occluded by cars and leaving merely the head region visualized, our pedestrian detector still performances well.

V. Conclusion and Discussion

In this paper, we proposed a robust pedestrian detection method with coupled network. The proposed coupled network consists of a gated feature extraction sub-network which exploits different combinations of multi-resolution CNN features for pedestrian candidates of different scales and a occlusion handling sub-network which applies a deformable position sensitive pooling model. The two sub-networks provide two complimentary ways of detection to reinforce the robust detection results. Owing to the coupled framework, the proposed detection method can address the problem of detecting small and occluded pedestrians simultaneously. Extensive experiments on two widely used pedestrian detection datasets (i.e. CityPersons and Caltech) demonstrate the effectiveness of our proposed method for detecting pedestrians in urban scenarios. Our proposed method outperforms previous state-of-the-art methods by a large margin, where the largest improvement is registered for small-scale and heavily-occluded pedestrians.

In the future, we can investigate the usage of temporal information since the pedestrian datasets are usually in video/ image sequences. The temporal information of consecutive sequences in the videos can provide extra contextual clues for better feature discrimination, especially for the occlusion cases where image-level information is relatively weak and ambiguous. Moreover, the proposed detection approach can be extended for general object detection. The rational of proposing the coupled network for pedestrian detection is to address two challenges of the pedestrian detection task simultaneously. This mechanism can also be beneficial to the detection of other small size and potentially occluded or deformable objects.

References

[1] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 361–374, Feb. 2014.

[2] M. Bilal, A. Khan, M. U. K. Khan, and C.-M. Kyung, "A low-complexity pedestrian detection framework for smart video surveillance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2260–2273, Oct. 2017.

[3] R. Kümmerle, M. Ruhnke, B. Steder, C. Stachniss, and W. Burgard, "Autonomous robot navigation in highly populated pedestrian zones," *J. Field Robot.*, vol. 32, no. 4, pp. 565–589, Jun. 2015.

[4] H. Wang, B. Wang, B. Liu, X. Meng, and G. Yang, "Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle," *Robot. Auto. Syst.*, vol. 88, pp. 71–78, Feb. 2017.

[5] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2020.

[6] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1259–1267.

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

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[9] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu, "CoupleNet: Coupling global structure with local parts for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4126–4134.

[10] 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. (NIPS)*, 2015, pp. 91–99.

[11] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 379–387.

[12] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3213–3221.

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

[14] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 97, Jun. 1997, pp. 193–199.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[16] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[17] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, Aug. 2007.

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

[19] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[20] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1751–1760.

[21] W. Ouyang, X. Zeng, and X. Wang, "Single-pedestrian detection aided by two-pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1875–1889, Sep. 2015.

[22] S. Tang, M. Andriluka, and B. Schiele, "Detection and tracking of occluded people," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 58–69, Oct. 2014.

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

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

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

[26] L. Zhang, L. Lin, X. Liang, and K. He, "Is Faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 443–457.

[27] T. Liu and T. Stathaki, "Faster R-CNN for robust pedestrian detection using semantic segmentation network," *Frontiers Neurorobotics*, vol. 12, p. 64, Oct. 2018.

[28] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3210–3220, Jul. 2017.

[29] T. Liu, M. Elmikaty, and T. Stathaki, "SAM-RCNN: Scale-aware multi-resolution multi-channel pedestrian detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2018, pp. 1–13.

[30] T. Liu, J.-J. Huang, T. Dai, G. Ren, and T. Stathaki, "Gated multi-layer convolutional feature extraction network for robust pedestrian detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3867–3871.

[31] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 437–446.

[32] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3486–3495.

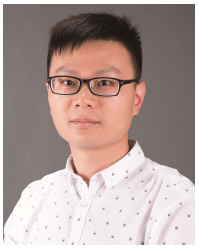
[33] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.

[34] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 637–653.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[36] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[37] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[38] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[40] M. Cordts *et al.*, "The cityscapes dataset," in *Proc. Workshop Future Datasets Vis. (CVPR)*, 2015, pp. 1–4.
[41] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.
[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: http://arxiv.org/abs/1412.7062
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[44] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.
[45] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 536–551.
[46] A. D. Costea and S. Nedevschi, "Semantic channels for fast pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2360–2368.
[47] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3361–3369.
[48] 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.
[49] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 354–370.
[50] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3127–3136.
[51] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.
[52] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 732–747.
[53] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–151.

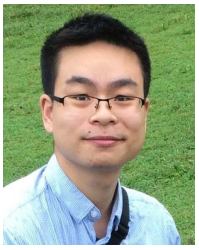
Tianrui Liu (Member, IEEE) received the B.Eng. degree (Hons.) from the Department of Electronic Information Engineering, The Hong Kong Polytechnic University, in 2013, the M.Phil. degree from The University of Hong Kong, in 2016, and the Ph.D. degree from the Electrical and Electronic Engineering Department, Imperial College London (ICL), U.K., in 2019. She is currently a Postdoctoral Researcher with BioMedIA Group, Department of Computing, ICL. Her research interests include image processing, computer vision and medical image analysis, more specifically, object detection, and ultrasound image/video processing.

Wenhan Luo (Member, IEEE) received the B.E. degree from the Huazhong University of Science and Technology, China, in 2009, the M.E. degree from the Institute of Automation, Chinese Academy of Sciences, China, in 2012, and the Ph.D. degree from Imperial College London, U.K., 2016. His research interests include several topics in computer vision and machine learning, such as motion analysis (especially object tracking), image/video quality restoration, object detection and recognition, and reinforcement learning.



Lin Ma (Member, IEEE) received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013.

He was a Researcher with Huawei Noah's Ark Laboratory, Hong Kong, from 2013 to 2016. He was a Principal Researcher with Tencent AI Laboratory, Shenzhen, China, from 2016 to 2020. He is a currently a Researcher with Meituan, Beijing, China. His current research interests include in the areas of computer vision, multimodal deep learning, specifically for image and language, image/video understanding, and quality assessment. He received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was a recipient of the Microsoft Research Asia Fellowship in 2011. He was a finalist in the HKIS Young Scientist Award in engineering science in 2012.



Jun-Jie Huang (Member, IEEE) received the B.Eng. degree (Hons.) in electronic engineering and the M.Phil. degree in electronic and information engineering from The Hong Kong Polytechnic University in 2013 and 2015, respectively, and the Ph.D. degree from the CSP Group, Imperial College London (ICL), London, U.K., in 2019. He is currently a Postdoctoral Researcher with the Communications and Signal Processing (CSP) Group, Electrical and Electronic Engineering Department, ICL. His research interests include image restoration, inverse problems, continuous-domain signal processing, deep dictionary learning, and deep learning.

Tania Stathaki (Member, IEEE) received the master's degree in electronics and computer engineering from the Department of Electrical and Computer Engineering, National Technical University of Athens, and the Ph.D. degree in signal processing from the Imperial College London, U.K. She was a Lecturer with the Department of Information Systems and Computing, Brunel University, London, and an Assistant Professor with the Department of Technology Education and Digital Systems, University of Piraeus, Greece. She is currently a Reader (an Associate Professor) with the Department of Electrical and Electronic Engineering, Imperial College London. She has intensive research experience in image processing, computer vision and machine learning and, more specifically, image fusion, image registration, change detection, object detection and recognition, and object tracking.

Tianhong Dai received the B.Eng. degree (Hons.) from the University of Liverpool in 2015 and the M.Sc. degree from Imperial College London in 2016. He is currently pursuing the Ph.D. degree from the Biologically Inspired Computation and Inference (BICI) Group, Imperial College London. His research interests include deep reinforcement learning, medical imaging, and computer vision.