



# Butterfly network: a convolutional neural network with a new architecture for multi-scale semantic segmentation of pedestrians

M. A. Alavianmehr<sup>1</sup> · M. S. Helfroush<sup>1</sup> · H. Danyali<sup>1</sup> · A. Tashk<sup>2</sup>

Received: 16 August 2022 / Accepted: 14 December 2022 / Published online: 2 February 2023  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

## Abstract

The detection of multi-scale pedestrians is one of the challenging tasks in pedestrian detection applications. Moreover, the task of small-scale pedestrian detection, i.e., accurate localization of pedestrians as low-scale target objects, can help solve the issue of occluded pedestrian detection as well. In this paper, we present a fully convolutional neural network with a new architecture and an innovative, fully detailed supervision for semantic segmentation of pedestrians. The proposed network has been named butterfly network (BF-Net) because of its architecture analogous to a butterfly. The proposed BF-Net preserves the ability of simplicity so that it can process static images with a real-time image processing rate. The sub-path blocks embedded in the architecture of the proposed BF-Net provides a higher accuracy for detecting multi-scale objective targets including the small ones. The other advantage of the proposed architecture is replacing common batch normalization with conditional one. In conclusion, the experimental results of the proposed method demonstrate that the proposed network outperform the other state-of-the-art networks such as U-Net + +, U-Net3 +, Mask-RCNN, and Deeplabv3 + for the semantic segmentation of the pedestrians.

**Keywords** Butterfly network (BF-Net) · Convolutional neural network · Pedestrian detection · Semantic segmentation · State-of-the-art U-Nets

## 1 Introduction

Detecting pedestrians promptly and explicitly in a natural environment is a vital goal in artificial intelligence systems. Pedestrian detection is also an interesting subject in computer vision. Besides, it is a fundamental building block in different applications, such as intelligent transportation systems (ITS), traffic control monitoring, visual search, models of human behaviour, pedestrian tracking, pose estimation,

pedestrian detection on social networks, face detection, semantic segmentation and, recently, monitoring the social distance of pedestrians in the Covid-19 Pandemic [1–4].

Pedestrian detection is often achieved through three main methods: 1) handcrafted, feature-based methods [5], 2) deep learning methods, particularly Convolution Neural Networks (CNNs) [6, 7], and 3) hybrid methods [8]. While in hybrid methods feature extraction is done by deep learning, classification and localization are implemented based on algorithms, such as Support Vector Machine (SVM) or AdaBoost. However, an alternative strategy is to deploy handcrafted methods to generate proposals and deep learning methods to classify and localize pedestrians. In the handcrafted, feature-based method, there are two major classes, including channel, feature-based methods [9] and deformable, part-based methods [10]. The main challenges in pedestrian detection can be divided into four categories, including occlusion, domain adaptation, scale variance, and real-time detection. Detecting small-scale and occluded pedestrians in a live-stream manner is the most essential issue in this field.

To distinguish pedestrians, there are three significant steps: 1) proposal generation, 2) proposal classification,

✉ M. A. Alavianmehr  
ma.alavianmehr@sutech.ac.ir

M. S. Helfroush  
ms\_helfroush@sutech.ac.ir

H. Danyali  
danyali@sutech.ac.ir

A. Tashk  
asta@mami.sdu.dk

<sup>1</sup> Department of Electrical Engineering, Shiraz University of Technology, Shiraz, Iran

<sup>2</sup> Maersk Mc-Kinney Møller Institute (MMMI), University of Southern Denmark (SDU), Odense, Denmark

and 3) post-processing. Pedestrian detection methods are defined based on these three steps. Firstly, proposal generation aims to recognize a set of bounding boxes where there are possibilities to detect pedestrians. Two leading approaches to achieve this aim are Region Proposal Network (RPN) and Sliding Window (SW) algorithms. Secondly, proposal classification aims to divide generated regions into two groups, including positive (pedestrian) and negative (background) classes based on feature extraction. It is noteworthy that in deep learning approaches, the first and the second steps unite and create a unified architecture. Additionally, in such approaches, localization and classification are gained simultaneously. In the third step, which is the post-processing step, the extra bounding boxes are excluded. Then, one or more pedestrians are likely surrounded by bounding boxes; therefore, the extra bounding boxes are neglected. The most popular approach in the post-processing step is non-maximum suppression (NMS) [11]. It is necessary that the researchers be able to produce distinctive feature maps to have an undemanding job in the classification step.

However, since supplementary information has been added to the pedestrian detection process, the segmentation method has been applied lately. Before researchers made use of CNN in segmentation, Random Forest (RF) and Conditional Random Field (CRF) were employed in the learning process. Fundamentally, image segmentation in pedestrian detection is classified to two categories, namely semantic and instant segmentation [12]. These segmentation methods are commonly known as multi-task learning, owing to the use of a separate network to segment semantically along with pedestrian detection. For example, in [3], an instance segmentation has been implemented by adopting the Faster-RCNN feature map. Given that semantic information of the background image is employed to detect pedestrians, these methods are more precise. Meanwhile, the use of semantic information must not lead to false positives (FP). Semantic segmentation methods include a considerable amount of computation inasmuch as they contain complex detection and segmentation networks.

Generally, the modern segmentation methods can be considered either proposal-based methods [13] or mask-based methods [14]. Proposal-based methods comprise a two-phase detection, and each region produces a proposal which is later segmented as a mask. In this method, pedestrian localization and classification is more accurate. Moreover, it is noteworthy that in this method, each proposal may contain several pedestrians. Segmentation must therefore be conducted precisely, which is not easy to achieve. That is why the relation between an occluded or not occluded pedestrian might not be distinguished. However, in mask-based methods, this problem does not exist, and they are

commonly employed to detect small-scale pedestrians or detect pedestrians in a crowded background [12].

In semantic segmentation, classification is conducted by means of a super pixels approach. Different semantic segmentation methods have been proposed, however, most of them have issues in the same topics as downsampling and spatial invariance [15]. To solve the former issue, the Atrous convolution algorithm has been proposed, while CRF has been used to extract semantic and more precise information, which leads on to resolving the latter issue [16]. The major problem in proposing a CNN-based semantic segmentation is the necessity for providing pixel-wise ground truth images to be deployed in the learning process, as the supervised learning algorithms require pixel-level labelled images. Among such methods, there are some algorithms that provide a weakly-supervised semantic segmentation [17]. These methods do not entirely depend on labelled information. Labelling can be carried out on image level, bounding box level, scribble level or point level. Another problem is that CNN-based semantic segmentation methods are real-time applications. Due to their complex architecture, they consequently involve a great deal of computation restricting some practical applications such as, ADAS and robot sensing.

In this paper, a novel semantic segmentation approach based on convolutional neural networks is presented. The proposed structure overcomes the limitations of open datasets as well as the deficiencies of the conventional semantic segmentation networks with the class of U-Nets [18]. The proposed network provides more flexibility for deploying the down-sampled batches of images and convolutional kernels, by introducing an innovative combination of conditional batch normalization and sampling blocks followed by a new supervision strategy based on a list of new skip connections. Generally, the major contributions of this paper can be summarized as follows:

- The proposed network can determine low scale pedestrians as target objects.
- Its special supervision prevents the loss of information and therefore mis-training of the network.
- The semantically segmented pedestrians can be given to a part detection network for possible occlusion detection.
- The implementation speed of the proposed architecture is high enough for possible real-time processing of images data, especially for surveillance and live supervision purposes.

This paper is organized as follows: in Sect. 2, a literature review of the previously proposed strategies for semantic segmentation of target objects, including pedestrians, is presented. Next, Sect. 3 contains a detailed description of the proposed method for semantic segmentation of pedestrians from popular datasets. The implementation and comparative

results are presented in Sect. 4. Finally, this paper is concluded under Sect. 5.

## 2 Literature review

Detection of meaningful objectives with the aid of neural networks has been developed notable advances for different objectives like pedestrians. In this section, an overview of the methods and algorithms chronologically deployed for pedestrian detection is represented.

### 2.1 Generic pedestrian detection

Deep neural networks, although powerful in detection, can be computationally expensive. Deep learning-based pedestrian detection algorithms have originated from the region proposal CNN (R-CNN) detectors. Pedestrian detection can be categorized into two-stage detectors and single-stage detectors. In the first phase, two-stage detectors estimate proposals. Then, each of the proposals is sent to classification and bounding box regression as the next phase of detectors.

Two-stage-based pedestrian detection algorithms have originated from the R-CNN detectors. In Fast-RCNN [6], to increase the network speed, the whole image enters the convolutional neural network at once. On the other hand, a pooling layer is also used. However, the network speed is still low due to the deployment of the selective search algorithm. Faster-RCNN [7] was designed to reduce the number of hyperparameters. Tesema et al. exploited region proposal networks (RPNs) as fundamental detectors of pedestrians [8]. Moreover, they deployed a naive classifier to refine the pedestrian detection outcome. Zhang et al. presented an anchor region proposal network to detect human different parts of the body as well as heads and endeavor to integrate them to attain higher accuracy. They also utilized the post-processing NMS to improve the detection results [19].

A significant problem with two-stage generators is their slow pace, which inspired researchers to speculate on single-stage detectors. These detectors worked based on the object bounding boxes and object classes, but they do not need intermediate object proposals. YOLO [20] and SSD [21], as examples of single-stage detectors, have high operation speed, though their accuracy is their weak point. Lately, RetinaNet [22] employed a novel object detection loss function named Focal Loss to deal with the data imbalance between the background (no object) and the other classes. Despite being a single-stage detector, RetinaNet is more accurate than Faster R-CNN and is analogous in terms of speed to other single-stage detectors. Wei-Yen Hsu et al. proposed the ratio-and-scale-aware YOLO

method, which is based on YOLOv3; however, it provides a lot of improvement. They proposed a revolutionary feature map that transformed each positive instance into a feature vector to encrypt both density and diversity information simultaneously [20]. Besides, the occlusion-sensitive hard example mining method and occlusion-sensitive loss were designed by Jin Xie et al. [23]. Their methods explore hard instances depending on the occlusion level and allocate higher weights to the detection errors taking place at considerably occluded pedestrians. Additionally, Yi Tang et al. designed the first architecture that enhanced pedestrian detection performance with a state-of-the-art framework that not only increased pedestrian information automatically but also investigated the loss function policy [24]. Recently, Glenn Jocher et al. proposed the most efficient version of YOLO algorithm with the name of YOLOv5 [25]. This version of YOLO can detect target objects faster and with higher accuracy compared to the previously proposed versions. The current algorithm deploys genetic algorithm for finding the best anchors and uses mosaic augmentation for improving the accuracy of training procedure.

### 2.2 Pedestrian detection based on semantic segmentation

Some of the semantic segmentation methods are U-Net [18], EncNet [26], Gated shaped CNN [27], Deeplab [16], etc. In semantic segmentation methods, due to multi-scale detection and receptive field increase, some methods have been proposed, such as Deeplab V2 and V3 [28]. In another method such as PSPNet [29] general content information is utilized to improve the segmentation process. Also, some methods such as ParseNet [30] have employed large-scale kernels for convolution and designed a network including boundary refinement.

In research by Alavianmehr et al. [31], a new combinational region and semantic segmentation CNN approach for pedestrian accurate detection and localization from static images is designed. The primary process of CNN-based methods includes two steps, proposal extraction, and CNN classification. The proposed framework is a mixture of modern CNNs such as, YOLO and semantic segmentation networks like Fully Convolutional Networks (FCNs) [15], particularly those with a structure similar to those of U-Nets. Huazhen Chu et al. introduced an effective segmentation method named Part Mask R-CNN. According to this method, they applied Part Mask R-CNN to every body part of the pedestrian to model different body parts and produce parts annotations utilizing database annotations and their processing [32]. Qiming Li et al. designed a new efficient anchor-free network based on Conditional Random

Fields (CRFs) for multi-scale pedestrian detection [33]. To set about the incomplete occlusion and scale problems in pedestrian detection, Peiyu Yang et al. developed an effective Fully Convolutional Network (FCN) [34].

Successful training of the proposed FCNs requires thousands of annotated training images, and just augmentation cannot provide reliable training for FCNs, especially in the case of special imaging modalities. According to these issues, a U-shape architecture named U-Net was proposed. U-Net supplies an asymmetric structure for semantic segmentation. This structure has its deficiencies. For instance, it is not able to detect multi-scale target objects very accurately, especially in the case of low image contrast. The other deficiency of U-Net is that the greater contracting depth it has, the higher complexities it will return. To overcome these deficiencies and add flexibility and scalability to the U-Net structure, some inspirational networks such as U-Net++ [35], and U-Net3+ [36], have been proposed. These two networks keep the original framework of U-Net and add some novelties to the U-Net structure. The proposed novelties compensate for the existing shortcoming in conventional U-Net architecture. For example, adding full- and multi-scale supervisions, which is essential for multi-scale semantic segmentation, embedding skip connections that provide integration for connected components of the detected objects, and so on.

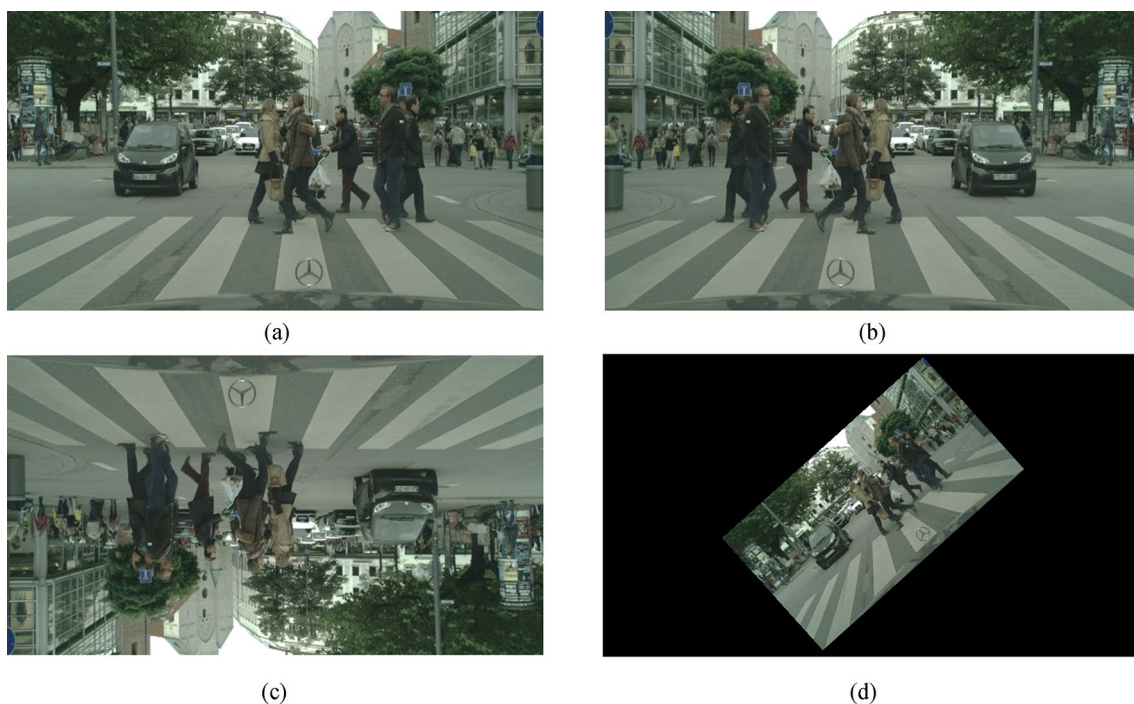
In addition, all the advantages and privileges provided by such state-of-the-art networks like U-Net3+ should be compatible with the application of pedestrian detection. The proposed method in this paper represents a new structure that provides multi-scale pedestrian detection based on the combination of multi- and full-scale supervision. Moreover, the proposed structure can detect low-scale pedestrians that could provide a horizon in front of occlusion detection.

### 3 Proposed method

In this section, the proposed framework for detecting pedestrians from popular datasets is explained. The proposed network has a novel structure so that it can be adapted to both online and real-time applications. The proposed method is folded into two parts to render a better explanation. Next, we examine the proposed pre-processing stage for feature extraction for fine-grained classification. Distinct parts of the proposed semantic segmentation architecture are explained in detail subsequently.

#### 3.1 Pre-processing: image augmentation

As we proposed a new structure for semantic segmentation of pedestrians, we should provide sufficient training and validation data for achieving an acceptable trained network



**Fig. 1** Applying sample augmentation to a sample image: **a** Original image, **b** Mirroring=Vertical (image x-axis), **c** Rotation=180°, and **d** Rotation=45



with a high volume of trainable parameters, even in the case of pre-trained backbones. For this case, one of the strategies for proving enough training images is augmentation. Image augmentation overcomes the under-fitting issue due to the low volume of input data. By deploying this procedure, all the trainable parameters would converge to their ultimate training stage. For this purpose, we deployed a list of common augmentation filters such as rotation, mirroring, and reflection. Figure 1 illustrates the deployed augmentation for a sample traffic image.

### 3.2 The proposed network

In this section, we introduced the proposed CNN network with its novel and innovative structure, which contains the feature extraction and selection procedures alongside semantic segmentation. Semantic segmentation is the process of classifying each pixel associated to a particular label. It does not differentiate across separate instances of the same object. On the other hand, Instance segmentation differs from semantic segmentation since it labels every instance of a particular object in the image dissimilarly.

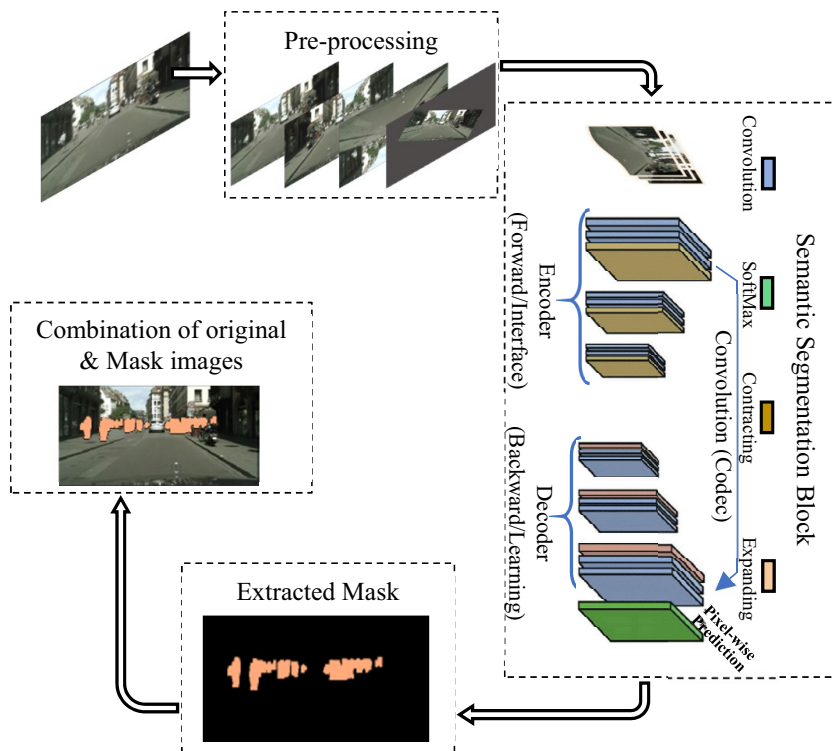
Figure 2 depicts the general block diagram of pedestrian detection system based on the semantic segmentation approach. The application of machine learning (ML) and deep learning technique in the segmentation of images has grown throughout the years.

In the context of this paper, the semantic segmentation approach is deployed for the detection of pedestrians. The pedestrian detection aims at the automatic driving aided system (ADAS), and it has a prominent role in the traffic surveillance strategies. The common structure for semantic segmentation purposes is fully convolutional networks (FCNs). Recently, there are also new proposed networks with U-shapes deployed for semantic segmentation purposes and therefore they are named as U-Nets. These networks are mainly deployed for semantic segmentation of biomedical objects because the focus of their application is more on the extraction of the detailed feature maps of the target objects. Such U-Net architectures are designed for extracting multi-scale target objects. Moreover, the architecture of networks like U-Net ++ and U-Net3 + has some deficiencies specially for detecting real world objects like pedestrians, so we have proposed a new architecture that is able to detect pedestrians from the relevant dataset images with more accuracy and less complexity.

Our proposed network with a new architecture can extract the fine features associated with pedestrian as target objects in a semantic segmentation manner. The node structures of different U-Nets and the one related to our proposed network is illustrated in Fig. 3.

As shown in Fig. 3, the proposed BF-Net does not have the skip connection complexities like UNet ++ and UNet3 +, because the obvious that the most important privilege of the proposed network over the competitive ones is that it has more flexibility to segment multi-scale

Fig. 2 General block diagram of a sample automatic object detection system



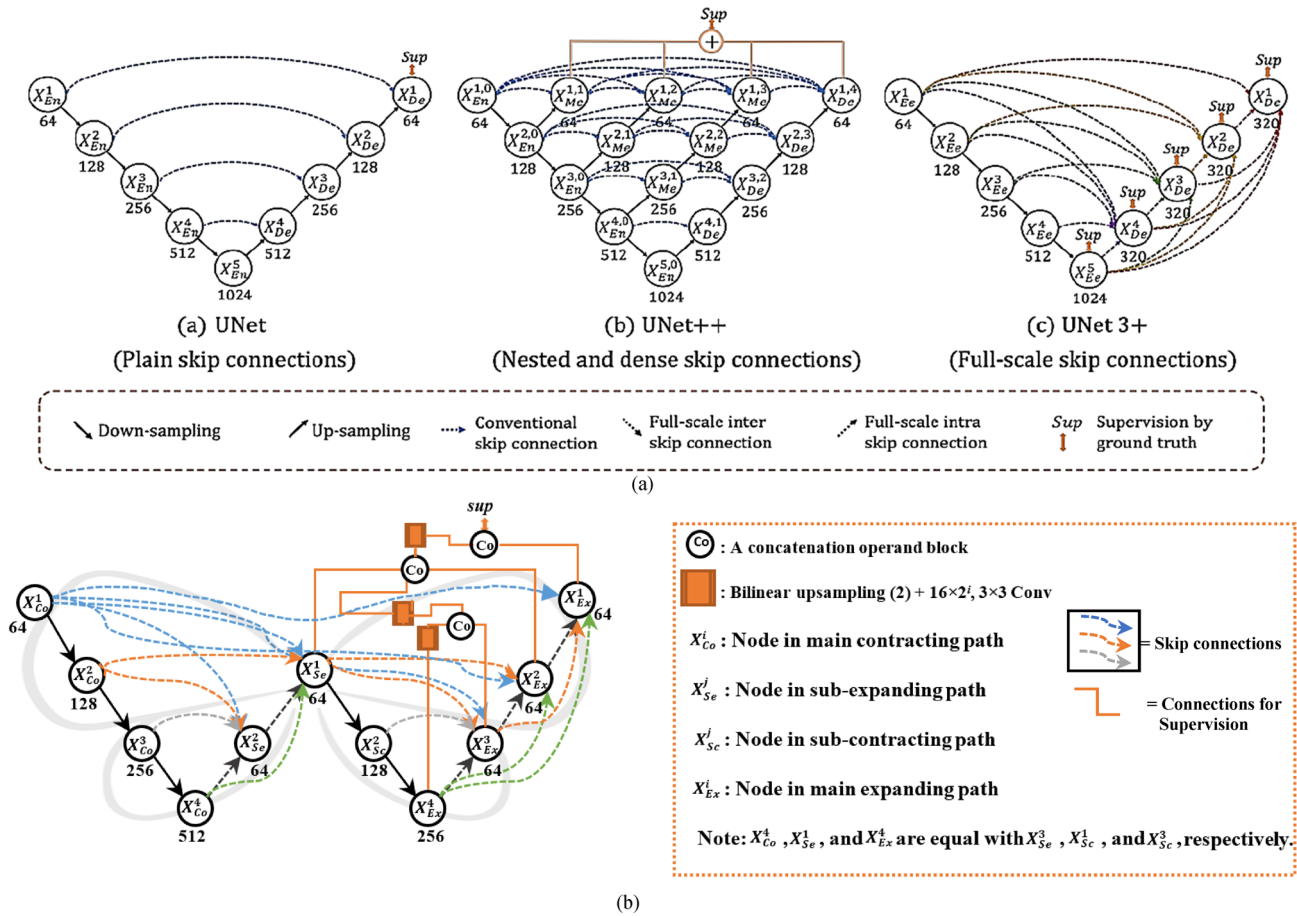


Fig. 3 a Comparing illustrations of the sample structures for the cutting-edge U-Nets [35], and b illustration of the node structure for the proposed network named as Butterfly Network (BF-Net) because of its resemblance to a butterfly

target objects due to its specific supervision. The full details of the proposed supervision strategy and its advantages are described in the next subsections.

### 3.3 Specific node structure of the proposed BF-Net

As the node structure of the proposed network is like the appearance of a butterfly, therefore we called it as

Butterfly-Network, and its abbreviation name is BF-Net. The proposed BF-Net comprises four types of nodes. Two of these nodes, named as  $X_{Co}^i$  and  $X_{Ex}^i$  are like the Contracting (Encoding) and Expanding (Decoding) nodes in Conventional U-Nets. The new nodes are  $X_{Se}^i$  and  $X_{Sc}^i$ , which are sub-expanding and sub-contracting nodes.

The new nodes are  $X_{Se}^i$  and  $X_{Sc}^i$ , which are sub-expanding and sub-contracting nodes.

$X_{Se}^i$  s that i demonstrates the ith downsampling or contracting layer along the coding direction, N stands for the number of the main contracting/Eanding nodes, and l is equal to the number of sub-contracting/sub-expanding nodes, defined with the following formula:

$$X_{Se}^i = \left\{ K \left[ \underbrace{C \left( D \left( X_{Co}^j \right) \right)_{j=1}^{N-l+i-1}}_{scales: 1^{st} \sim (N-l+i)^{th}}, C \left( X_{Co}^{N-l+i} \right), \underbrace{C \left( U \left( X_{Se}^j \right) \right)_{j=i+1}^l}_{scales: (N-l+i+1)^{st} \sim (l)^{th}} \right] \right\}, \quad \begin{matrix} i = 1 \\ i = 1, \dots, l - 1 \end{matrix} \tag{1}$$

where the function  $C(\cdot)$  stands for the convolution operation, the function  $K(\cdot)$  represents a convolution layer followed by a conditional batch normalization (CBN) and a ReLU activation. Moreover, the function  $D(\cdot)$  stands for a max-pooling layer with a pooling size  $2^{(N-l+j-1)}$  and  $U(\cdot)$  represents a bilinear up-sampling layer with a rate of  $2^{(l-j-1)}$ . Moreover, operand  $[\cdot]$  represents the channel dimension splicing and fusion. In addition to the definition of the sub-expanding nodes as  $\chi_{Se}^i$ , the definition for  $\chi_{Sc}^i$  nodes are as follows:

$$\chi_{Sc}^i = \begin{cases} \chi_{Se}^1, & i = 1 \\ H(\chi_{Sc}^{i-1}), & i = 2, \dots, l \end{cases} \quad (2)$$

where  $H(\cdot)$  applies the similar contracting procedure like the ones on the main contracting path. Finally, the feature map aggregation for  $\chi_{Ex}^i$  nodes is done based on the following equation:

$$\chi_{Ex}^i = \begin{cases} \chi_{Sc}^l \\ K \left( \underbrace{C(D(\chi_{Co}^j))_{j=1}^{N-l}, C(D(\chi_{Sc}^j))_{j=1}^{i+l-N-1}, C(\chi_{Sc}^{i+l-N})}_{scales: 1^{st} \sim (i+l-N)^{th}}, \underbrace{C(U(\chi_{Ex}^j))_{j=i}^{i-l}}_{scales: (i+l-N+1)^{th} \sim N^{th}} \right), & i = 1, \dots, N-1 \end{cases} \quad i = N \quad (3)$$

Due to the definitions represented for the different nodes existing in the proposed BF-Net, it is possible to illustrate the schematic structure of the proposed BF-Net as shown in Fig. 4 for an initial convolutional kernel size of 16, as mentioned in the node structure.

### 3.4 Proposed Multi-scale supervision

Deep supervision was introduced in U-Net ++, and full-scale deep supervision was proposed in U-Net3 + that adds some improvements to the final supervision results. However, the full-scale deep supervision faces a serious issue.

As the supervision for the last downsampling layer in the structure of U-Net3 + should be done with the down-sampled ground truth image, so the small-size or low-scale target

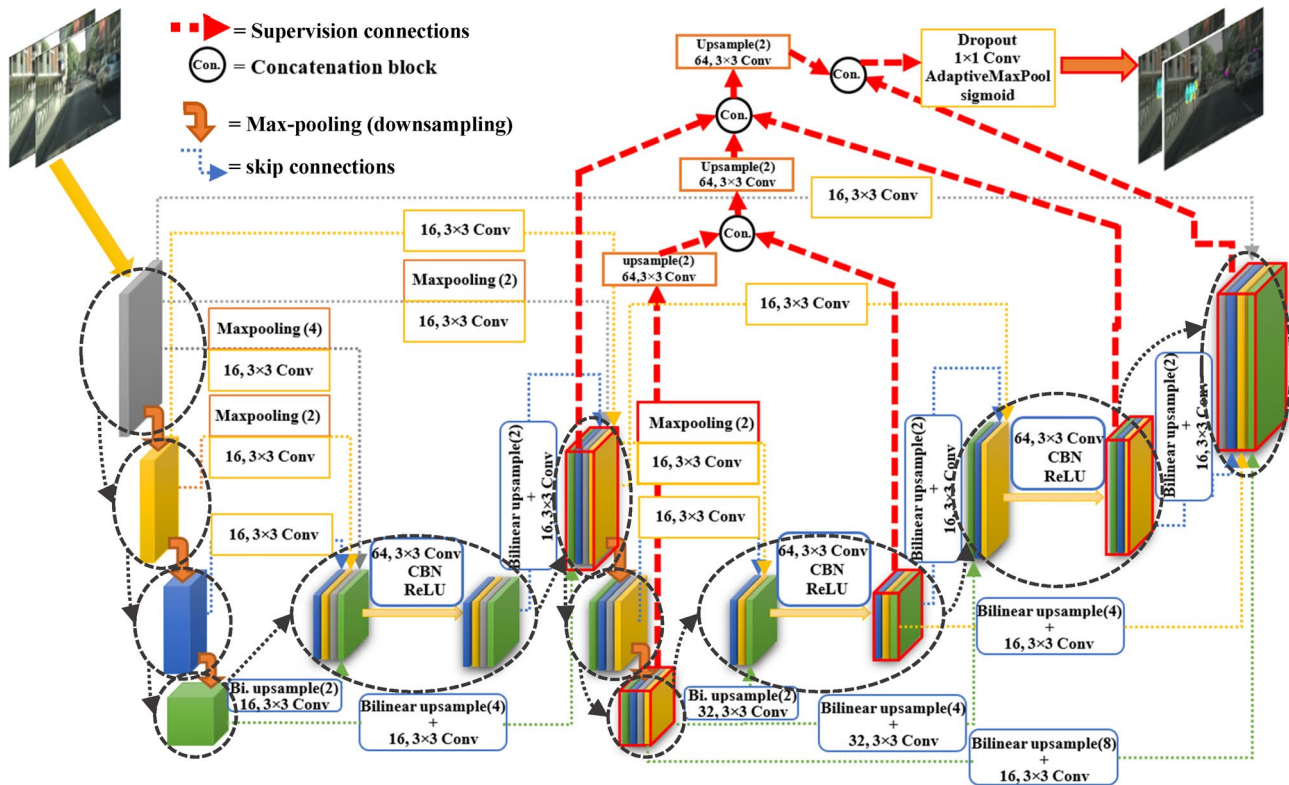


Fig. 4 Structure of the proposed BF-Net with  $N=4$  direct down/up sampling layer and 2 sub-up/sub-down sampling ones for the semantic segmentation application. The striped ellipsoids  $(\odot)$  and arrows

$(\odot)$  resemble the node structure of the proposed BF-Net shown in Fig. 3b. All the skip connections

objects may completely disappear due to downsampling. In that case, the supervision may lead to undesirable training results. This issue would be worse in the situation that there are multi-scale target objects in a very close contact with each other; so, the full-scale supervision may not take place efficiently. In this paper, a new approach toward conducting modified full-scale deep supervision is proposed.

For this purpose, we apply a bilinear upsampling followed by a 2D-convolution with an appropriate kernel to the  $\chi_{Sc}^{max}$ , so that it can be concatenated with matched  $\chi_{Ex}^i$  in the main expanding path. Subsequently, the concatenation of the blocks after resizing is applied to the output of all  $\chi_{Ex}^i$  s plus  $\chi_{Sc}^N$  s and  $\chi_{Se}^l$  s. Simultaneously, the outputs from each main expanding path are passed through appropriate upsampling and convolutional kernels, so that they can be concatenated with each other as well as the up-sampled sub-expanding  $\chi_{Sc}^{max}$ . In Eq. (4), the detailed formulas for the calculation of proposed supervisions (*Sup.*) for two different translations of BF-Net are shown:

$$Sup. = \begin{cases} K([C_3(U_3([C_2(U_2([C_1(U_1(\chi_{Ex}^4)), \chi_{Ex}^3])), \chi_{Ex}^2, \chi_{Sc}^1])), \chi_{Ex}^1]), & \text{for BF - Net}_{(4-2)} \\ K([C_4(U_4([C_3(U_3([C_2(U_2([C_1(U_1(\chi_{Ex}^5)), \chi_{Ex}^4), \chi_{Sc}^1])), \chi_{Ex}^3])), \chi_{Ex}^2]), \chi_{Ex}^1]), & \text{for BF - Net}_{(5-1)} \end{cases}$$

The important note about the sample supervisions mentioned in Eq. (4), is that the computational complexity of the proposed supervision is perpendicular with the number of main paths in BF-Nets. On the other words, the higher the number of main paths is, the more complicated the computation of the proposed supervision would be. According to Eq. (4), whenever the number of contracting layers increases from 3 to 4 phases, and so does the number of expanding layers, the computational complexity will increase as well. The number of main expanding or main contracting might change depending on data set and the use of real-time application. Consequently, the proposed BF-Net is simple, flexible, expandable, and powerful.

### 3.5 Hybrid loss function

After concatenating all the mentioned outputs, a code block containing a dropout followed by  $1 \times 1$  convolution with the number of kernels equal to the number of classes observed in the given dataset, followed by an adaptive max-pooling, and a proper activation function such as sigmoid is applied to the

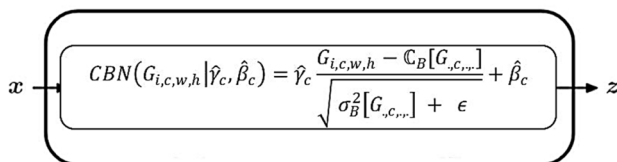


Fig. 5 General overview of a conditional batch normalization (CBN) block for embedding in BF-Net

concatenated output. In this case, a hybrid loss calculation method is deployed. The mathematical format of the hybrid loss is defined as:

$$L(G, P) = -\frac{1}{N} \sum_{c=1}^C \sum_{n=1}^N \left( g_{n,c} \log p_{n,c} + \frac{2g_{n,c}p_{n,c}}{g_{n,c}^2 + p_{n,c}^2} \right) \quad (5)$$

where  $g_{n,c} \in G$  and  $p_{n,c} \in P$  stand for the ground truth labels and predicted values for class  $c$  and  $n$ th pixel belonging to each batch of images, respectively. Accordingly,  $N$  shows the number of pixels inside each batch. This equation is just defined for the images of a batch, so the total loss function for the proposed BF-Net would be defined as:

$$L_{total} = \sum_{j=1}^e \omega_j \times L(G, P^j) \quad (6)$$

where  $e$  stands for the number of main expanding path blocks plus the first sub-contracting block (for instance, according to Fig. 5,  $e$  is equal to 5).  $\omega_j$  s index weights that

give each loss function calculated in Eq. (5).

For the accomplishment of the full supervision, it is necessary that the concatenated output of the resized outputs of the mentioned nodes be passed through a block, including a dropout,  $1 \times 1$  Convolution, an adaptive maxpooling, and finally, a sigmoid function for calculating the loss function.

### 3.6 Deployment of conditional batch normalization (CBN)

Another innovative contribution of the proposed BF-Net can be mentioned as the deployment of conditional batch normalization instead of conventional BNs. The basic idea for embedding batch normalization layers after convolutional layers is to provide faster convergence and keep heterogeneity of the processed data in each convolutional layer to prevent the optimization failures.

The conventional BN with equation  $BN(G_{i,c,w,h} | \gamma_c, \beta_c)$  has two  $\gamma_c$  and  $\beta_c$  that should be predicted from an embedding during the training procedure [37]. In other words, a BN reduces the internal covariant shift by normalizing feature maps belonging to each input mini-batch. However, the initialization of a network with less sensitivity to the initialization of the two parameters, i.e.,  $\gamma_c$  and  $\beta_c$  is very difficult. Accordingly, researchers suggested a conditional batch normalization (CBN) to estimate two change measuring parameters  $\delta\gamma_c$  and  $\delta\beta_c$  on the fixed primary numerical values, so that the target neural network will be initialized to produce outputs with a mean equal to zero



and a very small variance [38]. In that case the definition for a conditional batch normalization would be as follows:

$$CBN(G_{i,p,x,y}|\hat{\gamma}_c, \hat{\beta}_c) = \hat{\gamma}_c \frac{G_{i,p,x,y} - C_B[G_{..c,...}]}{\sqrt{\sigma_B^2[G_{..c,...}] + \epsilon}} + \hat{\beta}_c \quad (7)$$

where  $\{G_{i,...}\}_{i=0}^N$  stands for N samples in the form of a mini batch, and  $G_{i,p,x,y}$  is related to the  $p^{th}$  vector of feature maps of  $i^{th}$  sample at location  $(w, h)$ .  $\epsilon$  stands for a fixed value as stabilizing and regulating coefficient.  $\hat{\gamma}_c$  and  $\hat{\beta}_c$  are also defined as follows:

$$\hat{\gamma}_c = \gamma_c + \Delta\gamma_c \quad (8)$$

$$\hat{\beta}_c = \beta_c + \Delta\beta_c \quad (9)$$

where  $\Delta(\cdot)$ s stand for latent multi perceptron layers. Accordingly, the CBN is able to tune the independent feature maps based on different inputs, therefore assists in boosting the generalization ability of the network on inharmonious data [39].

The overall view of CBN block inside the structure of a sample BF-Net is illustrated in Fig. 5.

CBNs are more likely to be embedded inside a residual building block of each contracting/expanding path of a BF-Net. After describing the full details of the proposed network, it is time for evaluating the performance of the proposed method in comparison with the other cutting-edge U-Nets.

### 3.7 BF-net training

To implement the proposed BF-Net, we should train it based on all the points mentioned in the previous subsections. For instance, we should follow the node structures for down and

up sampling paths as introduced on Sect. 3.3. Moreover, we should consider the proposed multi-scale supervision alongside the hybrid loss function for the training procedure. The specific structure of the CBNs also plays an important role for the appropriate training and validation of the network. Selected optimization algorithm for training BF-Nets is Adam, since it is a combination of the best properties of the AdaGrad and RMSProp algorithms. We also adjusted the values for the number of epochs, the initial learning rate, and the patience term for the early stopping of the training procedure as 100,  $10^{-5}$ , and 20, respectively.

## 4 Implementation and comparative results



In this section, the performance evaluation of the proposed network is presented. In addition, the efficiency of the proposed network for the purpose of pedestrian detection is compared with the other state-of-the-art U-Networks. Accordingly, this section comprises of two subsections. Under the first sub-section, the validation datasets for the purpose of semantic segmentation of pedestrian images are introduced. The next sub-section is assigned to the qualitative and quantitative comparative results.

### 4.1 Pedestrian datasets

In this paper, two dataset containing images of the pedestrians and their pixel-wise annotations as ground truths are used for assessing the performance of pedestrian detection by the proposed network and comparing it with the performance of the other semantic segmentation networks. One of the deployed datasets is cityscapes [40, 41].

This dataset comprises of a large and diverse set of stereo video sequences recorded in street scenes from 50

**Table 1** Datasets for evaluating and comparing the performance of the pedestrian detection implemented by the proposed method and the other state-of-the-art networks

Dataset name	Dataset Description (e.g., Image sizes, Annotation Type, etc.)	Annotated pedestrians merged with original images
Cityscapes [40-41]	<ul style="list-style-type: none"> <li>5000 stereo images of size 1024x2048</li> <li>high quality pixel-level annotation</li> <li>Just 3475 images have fine annotation GTs</li> </ul>	
PennFudanPed [42]	<ul style="list-style-type: none"> <li>170 images with 345 labeled pedestrians, with the heights of labeled size fall into 180x390 pixels.</li> <li>All labeled pedestrians are straight up.</li> </ul>	

different cities, with high quality pixel-level annotations of 5000 frames in addition to a larger set of weakly annotated frames. The other dataset including pixel-wised annotation of pedestrians is PennFudanPad [42]. The brief introduction to all the deployed datasets for reporting the implementation results of this paper are shown in Table 1.

### 4.2 Qualitative and quantitative comparison results

Under this section, the qualitative and quantitative comparisons between the pedestrian detection outcomes of the proposed BF-Net and the state-of-the-art networks such as ResNet-50 & -101-based U-Nets are presented.

To achieve a good presentation of the proposed network in comparison with the other state-of-the-art networks such as U-Net3+, we should apply both BF-Net and U-Net3+ to a similar image belonging to the introduced datasets.

Figure 6 exhibits both qualitative and quantitative results for the semantic segmentation of pedestrians conducted by ResU-Net-50 based U-Net3+ (left) and the proposed ResU-Net-50 based BF-Net (right) in a sample cityscapes image.

It is worth mentioning that the true positive (TP), false negative (FN), false positive (FP) and true negative (TN) parameters are calculated based on a threshold 0.5 applied to the output results of the SoftMax layer.

As it can be inferred from the results shown in Fig. 6, the proposed BF-Net can reduce the number of false positives

(FPs) to increase the precision of the pedestrian detection. This advantage alleviates the precision and subsequently the dice score of pedestrian detection.

Some other qualitative results for semantic segmentation of pedestrians based on the proposed BF-Net and the other competitive U-Nets are depicted in Fig. 7. The results shown in this figure demonstrate the great ability of the proposed network to detect very low scale pedestrians comparing to the other state-of-the-art U-Nets. This ability is due to the proposed supervision strategy that describes in the previous section. It is noteworthy that substituting the conventional BN blocks with the CBN ones also help improving the segmentation results specially for the distracting objectives like riders and manikins.

The other quantitative comparison results are illustrated in Fig. 8. In the shown diagram of these two figures, the implementation results are reported from two aspects. One aspect is the illustration of the precision-recall curves, that are reported for the implementation of the proposed BF-Net compared with the state-of-the-art U-Nets based on the backbone of ResNet-50 and ResNet-101.

As the curves of the diagrams in Fig. 8 show, the BF-Net achieves a higher performance than conventional U-Net, U-Net++, and U-Net3+ for both introduced backbones.

After comparing the performance of the proposed BF-Net with the state-of-the-art UNet architectures, it is important to compare the performance of the proposed network with

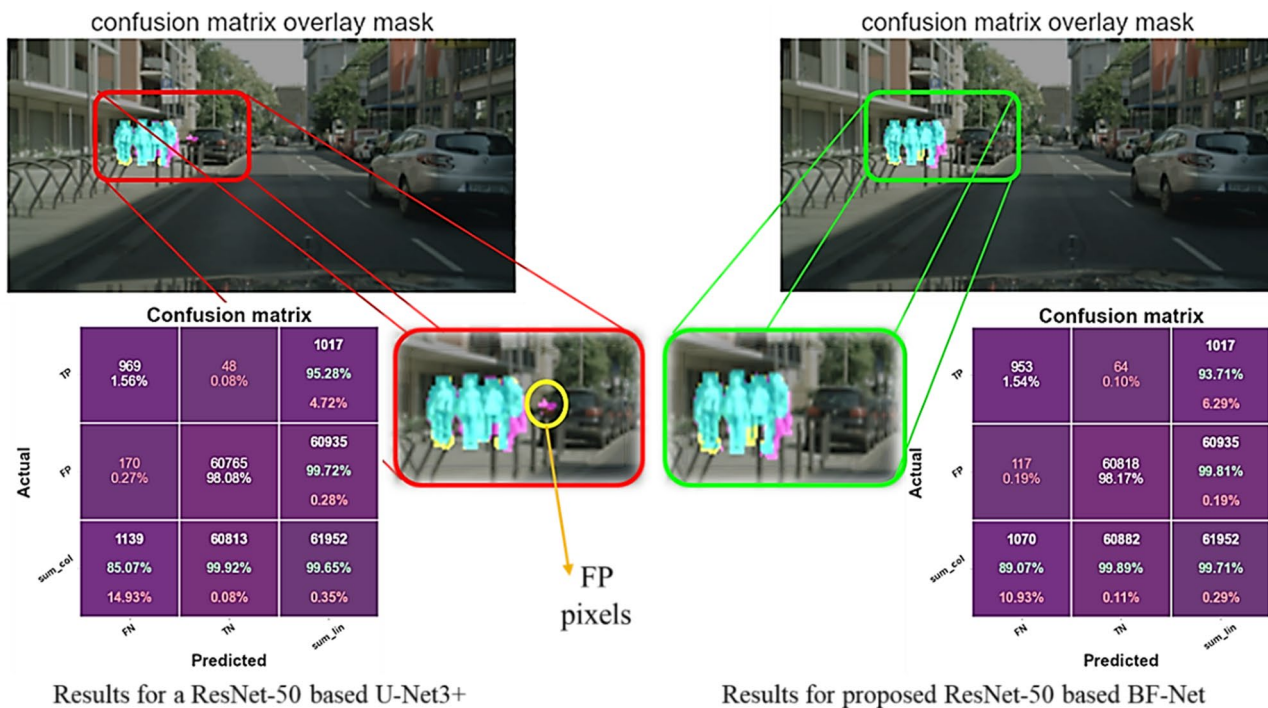
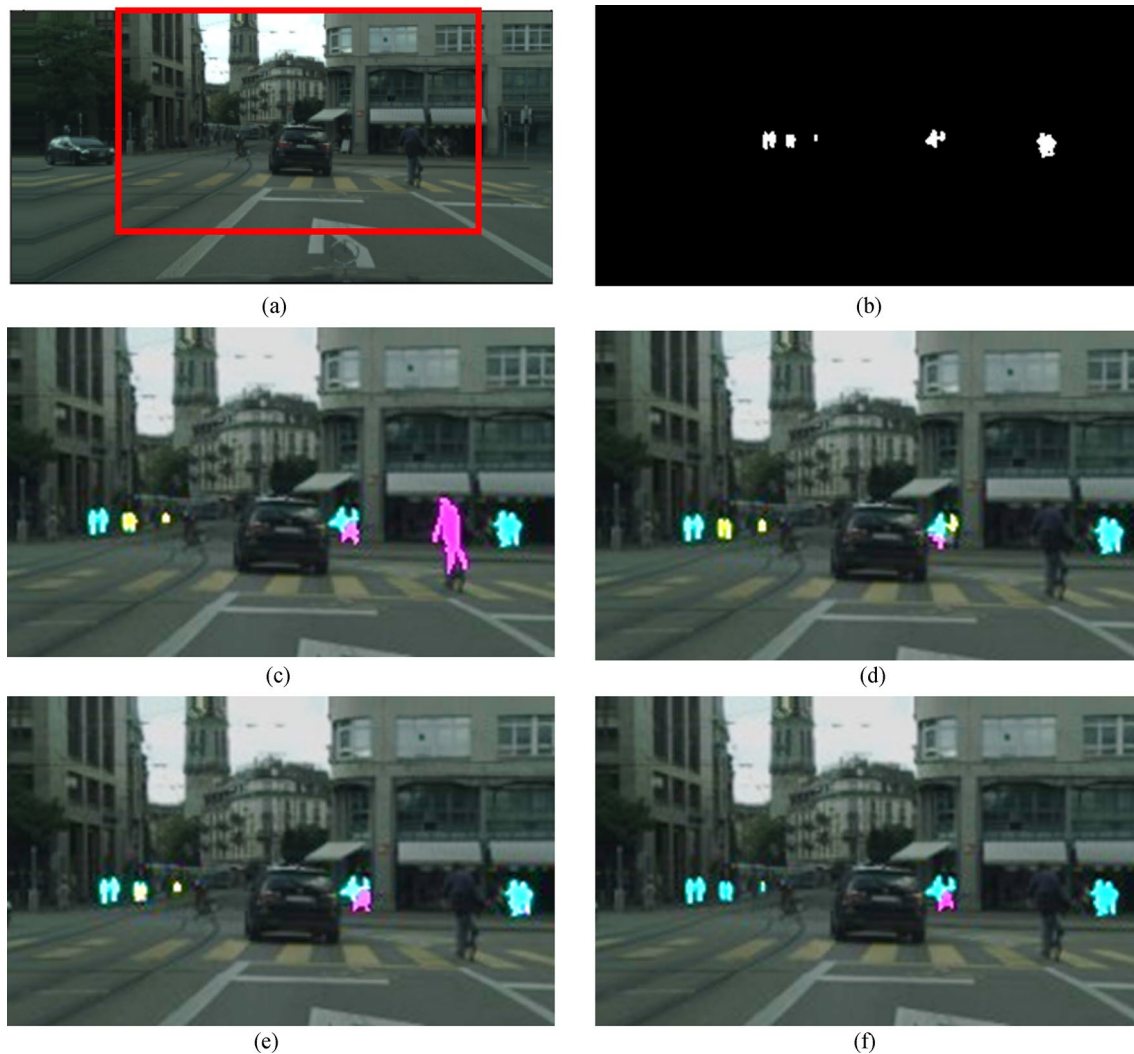


Fig. 6 Qualitative comparison of ResNet-50-based U-Net3+ (left), and proposed BF-Net (right) on a sample image belonging to the cityscapes dataset. Cyan areas: true positive (TP); Yellow areas: false negative (FN); Purple areas: false positive (FP)



**Fig. 7** Sample Qualitative comparison between the results of detecting very small-scale pedestrians based on different semantic segmentation networks: **a** original image from cityscapes dataset with original size, and **b** its associated binary ground truth, and pedestrians' detection results for: **c** U-Net, **d** U-Net++, **e** U-Net3+, and **f**

BF-Net. The images including segmentation results are cropped and resized for showing the segmentation results more clearly. Cyan areas: true positive (*TP*); Yellow areas: false negative (*FN*); Purple areas: false positive (*FP*)

two other state-of-the-art networks, i.e., Mask R-CNN and Deeplabv3+ with the same backbone ResNet50.

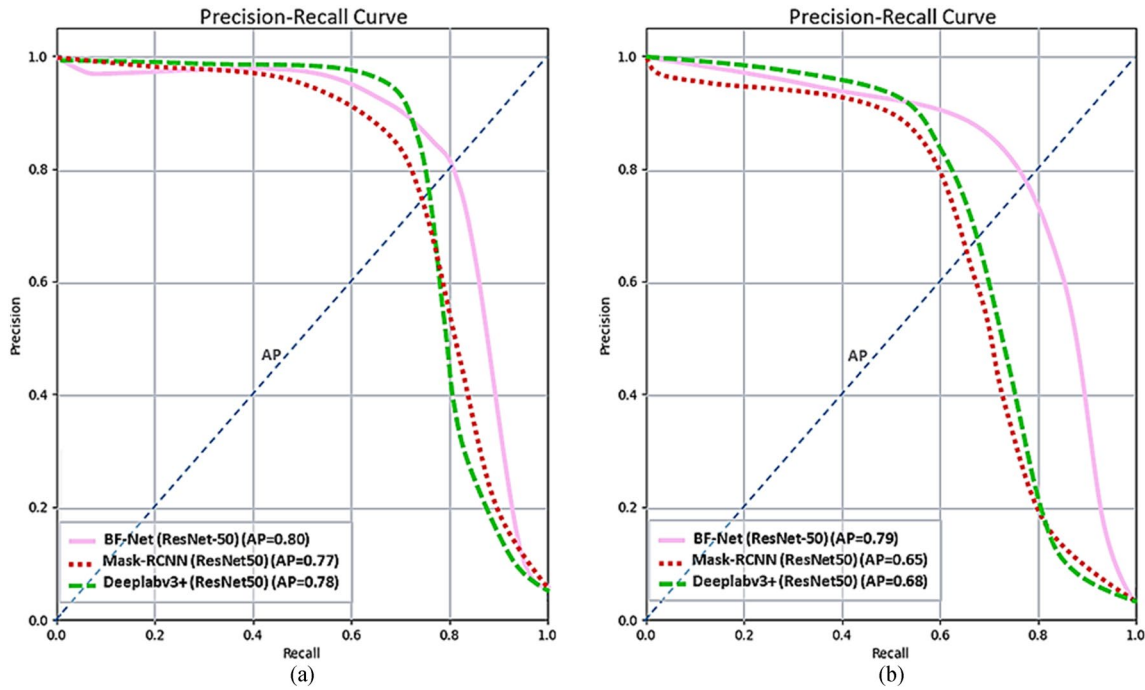
It is worth mentioning that we selected ResNets because in most of the other methods and references, these backbones lead to the best results with an acceptable level of complexity. Therefore, we built the foundation of all the implementation and simulations associated with our proposed method based on ResNet50 and ResNet101 backbones. Accordingly, a fair comparison between the efficiency of our proposed method and the other state-of-the-art ones can take place [43].

The precision-recall curves of implementing pedestrian detection by the means of BF-Net, Mask-RCNN and

Deeplabv3+ and applied to the images of both cityscapes and PennFudanPed datasets are illustrated in Fig. 9.

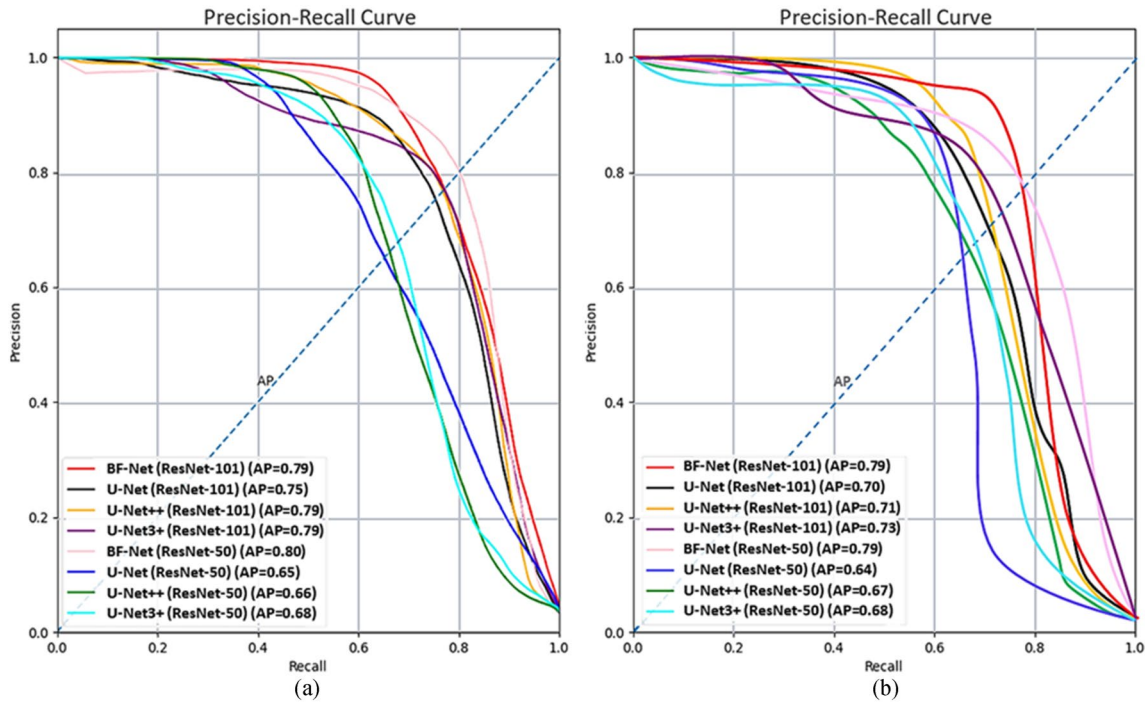
The other aspect of quantitative comparison is inference time plots shown in Fig. 10. In this plot, the inference time, the complexity, and  $F_1$ -scores (Dice score) are illustrated simultaneously for all the comparative networks. It can be observed from this plot that both BF-Nets have higher performance with a similar rate of complexity (the size of the depicted circles), and an implementation speed a bit slower than the other networks with similar backbones. Table 2 compares U-Net, U-Net++, U-Net3+ and the proposed BF-Net with two different ResNet-50 and ResNet-101 backbones in terms of segmentation results measured by Dice coefficient and IoU (mean  $\pm$  std) for the images in





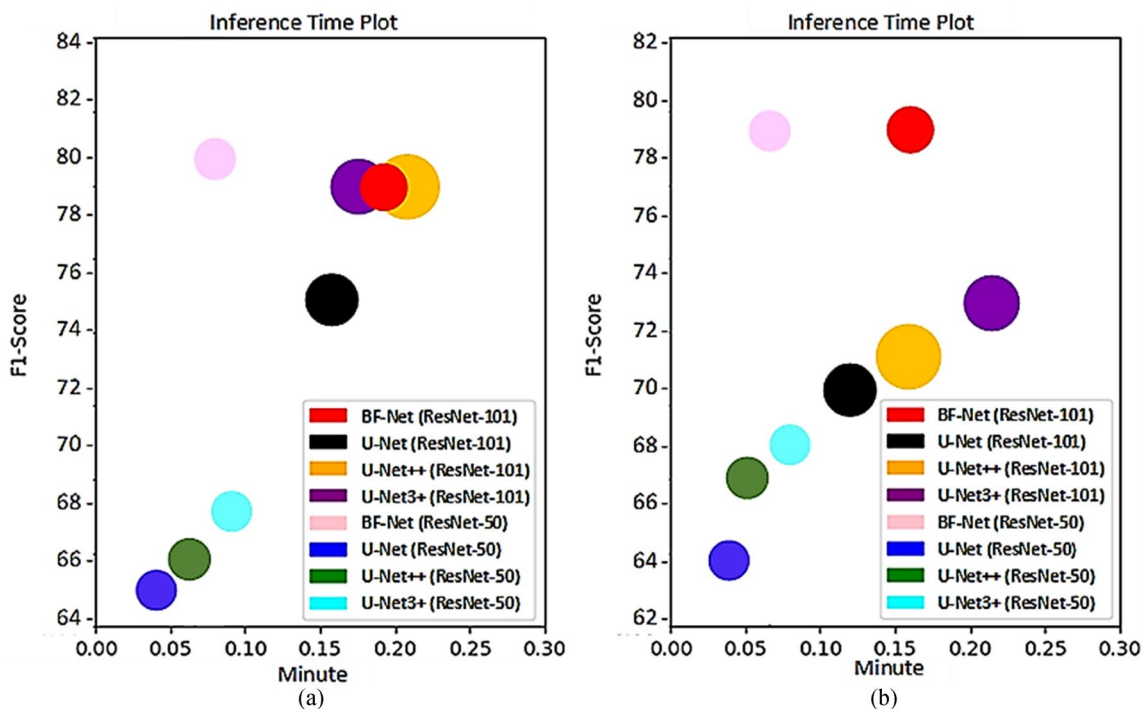
**Fig. 8** Comparison of BF-Net and the three other U-Nets with implementing two different ResNet-based networks: **a** Precision-Recall Curves with highlighted AP values for images of cityscapes dataset,

and **b** Precision-Recall Curves with highlighted AP values for images PennFudanPed dataset



**Fig. 9** Comparison between BF-Net, Mask-RCNN and Deeplabv3+ based on ResNet-50: **a** Precision-Recall Curves with highlighted AP values for the images of cityscapes dataset, and **b** Precision-Recall Curves with highlighted AP values for the images of PennFudanPed dataset





**Fig. 10** inference time, complexity (based on the size of parameters), and F1-Score of the proposed BF-Net and the comparative U-Nets for: **a** cityscapes dataset, and **b** PennFudanPed dataset. The inference

time is calculated by the time taken to process test images belonging to cityscapes dataset on a single NVIDIA GeForce RTX3080 GPU with 8 GB of dedicated memory

**Table 2** Comparison of pedestrian detection performance for test images in Cityscapes dataset between the competitive 3state-of-the-art U-Nets and:

(a) BF-Net with ResNet101 backbone	Architecture names/Criteria name			
	U-Net (ResNet101) (5) [17]	U-Net++ (ResNet101) (5) [34]	U-Net3+ (ResNet101) (5) [35]	BF-Net (ResNet101) (4-1) [Proposed]
Dice Coefficient ( $F_1$ -Score) ( $\uparrow$ )	75.21 $\pm$ 1.31	76.84 $\pm$ 0.73	77.39 $\pm$ 0.43	<b>78.95 <math>\pm</math> 0.15</b>
IoU (Jaccard) ( $\uparrow$ )	60.27 $\pm$ 1.90	62.39 $\pm$ 0.57	63.12 $\pm$ 0.27	<b>65.22 <math>\pm</math> 0.08</b>
No. trainable Param.s (Million) ( $\downarrow$ )	<b>16.8</b>	21.4	17.9	18.1
(b) BF-Net with ResNet50 backbone	Criteria name/Architecture names			
	U-Net (ResNet50) (5) [17]	U-Net++ (ResNet50) (5) [34]	U-Net3+ (ResNet50) (5) [35]	BF-Net (ResNet50) (4-1) [Proposed]
Dice coefficient ( $F_1$ -Score) ( $\uparrow$ )	64.92 $\pm$ 0.97	66.14 $\pm$ 1.02	68.07 $\pm$ 0.75	<b>80.24 <math>\pm</math> 0.34</b>
IoU (Jaccard) ( $\uparrow$ )	48.06 $\pm$ 0.54	49.41 $\pm$ 0.67	55.60 $\pm$ 0.45	<b>67.00 <math>\pm</math> 0.84</b>
No. trainable Param.s (Million) ( $\downarrow$ )	<b>12.6</b>	14.3	13.8	14.0

Bold values indicate the results of the proposed method which are superior to other methods

cityscapes dataset. For each evaluation, we investigate all possible confidence thresholds to report the best Jaccard Index ( $J.I.$ ) score. Larger Jaccard Index represents better

performance.  $J.I.$  mainly assess the degree of overlap between the predicted set  $P$  and the ground truth label set  $G$ , as showed in Eq. (10).

**Table 3** Comparison of pedestrian detection performance for the test images in PennFudanPed dataset between the competitive state-of-the-art U-Nets and:

	Architecture names/Criteria name	
(a) BF-Net with ResNet101 backbone	U-Net (ResNet101) (5) [17]	BF-Net (ResNet101) (4-1) [Proposed]
Dice Coefficient ( $F_1$ -Score) ( $\uparrow$ )	69.68 ± 1.64	<b>78.86 ± 0.94</b>
IoU (Jaccard) ( $\uparrow$ )	53.47 ± 0.83	<b>65.1 ± 0.47</b>
No. trainable Param.s (Million) ( $\downarrow$ )	<b>16.8</b>	18.1
	U-Net++ (ResNet101) (5) [34]	U-Net3+ (ResNet101) (5) [35]
Dice Coefficient ( $F_1$ -Score) ( $\uparrow$ )	71.13 ± 1.05	73.35 ± 0.73
IoU (Jaccard) ( $\uparrow$ )	55.2 ± 9.53	57.92 ± 0.37
No. trainable Param.s (Million) ( $\downarrow$ )	21.4	17.9
(b) BF-Net with ResNet50 backbone	U-Net (ResNet50) (5) [17]	BF-Net (ResNet50) (4-1) [Proposed]
Dice Coefficient ( $F_1$ -Score) ( $\uparrow$ )	64.21 ± 1.33	<b>78.91 ± 1.02</b>
IoU (Jaccard) ( $\uparrow$ )	47.29 ± 0.67	<b>65.17 ± 0.51</b>
No. trainable Param.s (Million) ( $\downarrow$ )	<b>12.6</b>	14.0
	U-Net++ (ResNet50) (5) [34]	U-Net3+ (ResNet50) (5) [35]
Dice Coefficient ( $F_1$ -Score) ( $\uparrow$ )	66.86 ± 0.73	68.15 ± 0.54
IoU (Jaccard) ( $\uparrow$ )	50.22 ± 0.37	51.69 ± 0.27
No. trainable Param.s (Million) ( $\downarrow$ )	14.3	13.8

Bold values indicate the results of the proposed method which are superior to other methods

$$J.I. = \frac{P \cap G}{P \cup G} \tag{10}$$

And the  $F_1$ -measure, is the same as the Dice coefficient:

$$\text{Dice\_Coefficient} = \frac{2 \times PG}{P + G} \tag{11}$$

Table 3 contains the comparative results between the performance of the proposed BF-Net and the other competitive U-Nets in the terms of Dice coefficient and IoU (mean ± std) for the images in PennFudanPed dataset.

Finally, Table 4 includes the comparative results of pedestrian semantic segmentation implemented by ResNet50 backbone BF-Net, Mask-RCNN, and DeepLabv3+ networks for test images belonging to both city-scapes and PennFudanPed datasets.

## 5 Discussion and conclusion

The semantic segmentation of pedestrians is a crucial preliminary step in various domains related to intelligent traffic systems, especially safe and secure automatic driving systems, and traffic surveillance ones. The proposed BF-Net provides a reliable and efficient architecture for pedestrian semantic segmentation with an acceptable level of computational complexity so that it could be deployed in real-time processing. In this section, the strengths and limitations of the proposed BF-Net are discussed under a separate subsection. Moreover, the subsection contains the impacts of this architecture on the community of the semantic segmentation of the pedestrians. The main contribution of the proposed architecture compared with the other state-of-the-art networks are concluded under conclusion and future work subsection.

### 5.1 Discussion

The proposed BF-Net architecture provides semantic segmentation of pedestrians with higher performance and lower complexity than the other state-of-the-art architectures, especially the ones with architectures like traditional U-Net.

As the results shown in Tables 2-4, the efficiency of the proposed BF-Net based on Dice coefficient ( $F_1$ -score) and Jaccard index (IoU) increases, while the computational complexity based on the number of the trainable parameters decreases. In other words, the efficiency and the computational complexity are inversely proportional, i.e., the higher the number of trainable parameters is, the less efficient the BF-Net performs. This relationship can be formulated as follows:

**Table 4** Comparison of pedestrian detection performance for the test images in PennFudanPed dataset between the competitive state-of-the-art U-Nets and:

Dataset Name	(a) cityscapes			(b) PennFudanPed		
	BF-Net (ResNet50) (4-1) [Proposed]	Mask-RCNN (ResNet50) [3]	Deeplabv3 + (ResNet50) [26]	BF-Net (ResNet50) (4-1) [Proposed]	Mask-RCNN (ResNet50) [3]	Deeplabv3 + (ResNet50) [26]
F <sub>1</sub> -Score (↑)	<b>80.24 ± 0.34</b>	77.25 ± 1.76	78.43 ± 0.95	<b>79.08 ± 0.53</b>	65.36 ± 0.87	68.28 ± 0.34
IoU (Jaccard) (↑)	<b>67.00 ± 0.86</b>	62.93 ± 0.89	65.51 ± 0.48	<b>65.4 ± 0.27</b>	48.54 ± 0.44	51.84 ± 0.17

(a) BF-Net with ResNet101 backbone,

(b) BF-Net with ResNet50 backbone

$$\text{No. of Trainable Param.s} \propto \frac{1}{\text{Efficiency}} \quad (12)$$

Besides the mentioned advantages, the proposed BF-net has some limitations. First, based on the precision-recall curves, the performance of the proposed BF-Net does outperform the other U-Nets for a specific range of thresholds. The reason could be related to the specific architecture of the nodes and the ways that the weights and biases trained. Second, the BF-Net is not faster than conventional U-Net and U-Net + +. In general, for a specified applications like real-time and online surveillance, the pedestrian segmentation scheme should be modified in such a way that the deployed strategy can create a trade-off between complexity and processing speed.

Most of the deep learning architectures for semantic segmentation mentioned in [43], suffer from deficiencies such as incompatibility of their convolutional blocks for pre-trained backbones. This problem occurs mostly because the semantic segmentation networks have no fully connected layers. The encoder-decoder models such as U-Nets have the same issue. The proposed BF-Net tries to solve this issue by deploying Multi-scale Supervision like U-Net3 + for increasing the accuracy of the segmentation for low scale objects. On the other side, by concatenating the supervisions and producing a unique one, like U-Net + +, the BF-Net succeeds to overcome the complexity of U-Net3 +. The proposed BF-net is also able to be compatible with different pre-trained backbones that may have a lot of convolutional and residual blocks, by using flexible direct and sub-paths. In addition, the proposed BF-Net can create a compromise between the convolutional blocks from the direct and sub paths by deploying appropriate skip connections.

## 5.2 Conclusion

In this paper, we proposed a new deep neural network architecture named as BF-Net for the purpose of pedestrian semantic segmentation. The novelty of the proposed method is that

it could be deployed for segmenting the pedestrians especially the ones that have several scales and different appearances in a series of sequential frames. The proposed network extracts all the available feature maps for the pedestrians as target objects, so that the occlusion issue can be observed and detected in the studied datasets. The implementation results for detecting pedestrians from the images of cityscapes and PennFudanPed databases demonstrate that the proposed method has a high ability to detect and even predict the existence of pedestrian in both stationary images and live video streams. As it is possible to embed the proposed skip connection in the structure of BF-Net into the feature pyramid network (FPN) existing in Mask R-CNN, then the next step as the future work would be replacing the plain skip connections of FPN with the redesigned skip connections of BF-Net.

**Availability of data and materials** The data that support the findings of this study are openly available in Cityscapes dataset at [<https://www.cityscapes-dataset.com/>], [40, 41] and PennFudanPad dataset at [<https://www.kaggle.com/datasets/psvishnu/pennfudan-database-for-pedestrian-detection-zip>], [42]. For more details please kindly refer to implementation and comparative results, part 4.1.

## Declarations

**Conflict of interest** All authors declare that they have no conflicts of interest.

## References

- Zheng, D., Xiao, J., Huang, K., Zhao, Y.: Segmentation mask guided end-to-end person search. *Sig. Process. Image Commun.* **86**(1), 115896 (2020). <https://doi.org/10.1016/j.image.2020.115876>
- Chen, L., Lin, L., Lu, X., Cao, D., Wu, H., Guo, C., Liu, C., Wang, F.: Deep neural network based vehicle and pedestrian detection for autonomous driving: a survey. *IEEE Trans. on Intell. Transp. Syst.* **22**(6), 3234–3246 (2021). <https://doi.org/10.1109/TITS.2020.2993926>
- He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(2), 386–397 (2020). <https://doi.org/10.1109/TPAMI.2018.2844175>

4. Bao, Q., Liu, W., Cheng, Y., Zhou, B., Mei, T.: Pose-guided tracking-by-detection: robust multi-person pose tracking. *IEEE Trans. Multimed.* **23**(20278824), 161–175 (2021). <https://doi.org/10.1109/TMM.2020.2980194>
5. Lai, C., Teoh, S.: A review on pedestrian detection techniques based on Histogram of Oriented gradient feature. *IEEE Stud. Conf. Res. Dev.* **9**(1), 47–64 (2014). <https://doi.org/10.1109/SCORED.2014.7072948>
6. Girshick, R.: Fast R-CNN. *IEEE Int. Conf. Comput. Vision* **15801732**(1), 2380–2504 (2015). <https://doi.org/10.1109/ICCV.2015.169>
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. *Proceed. Int. Conf. Neural Inform. Process. Syst.*, 1(2):91–99 (2015). <https://arxiv.org/abs/1506.01497>
8. Tesema, F., Wu, H., Chen, M., Lin, J., Zhu, W., Huang, K.: Hybrid channel-based pedestrian detection. *Neurocomputing* **338**(1), 1–8 (2020). <https://doi.org/10.1016/j.neucom.2019.12.110>
9. Liu, X., Toh, K., Allebach, J.: Pedestrian detection using pixel difference matrix projection. *IEEE Trans. Intell. Transp. Syst.* **21**(4), 1441–1454 (2020). <https://doi.org/10.1109/TITS.2019.2910093>
10. Wang, L., Xu, L., Yang, M.: Pedestrian detection in crowded scenes via scale and occlusion analysis. *IEEE International Conference on Image Processing, 2016-1210–1214* (2016). <https://doi.org/10.1109/ICIP.2016.7532550>
11. Yang, C., Li, L., Guo, X., Wang, Y., Ma, J., Jiao, L., Liu, F., Liu, X.: Region NMS-based deep network for gigapixel level pedestrian detection with two-step cropping. *Neurocomputing* **468**(1), 482–491 (2022). <https://doi.org/10.1016/j.neucom.2021.10.006>
12. Jiang, H., Liao, S., Li, J., Prinet, V., Xiang, S.: Urban scene based semantical modulation for pedestrian detection. *Neurocomputing* **474**(1), 1–12 (2022). <https://doi.org/10.1016/j.neucom.2021.11.091>
13. Lin, C., Lu, J., Zhou, J.: Multi-grained deep feature learning for robust pedestrian detection. *IEEE Trans. Circuits Syst. Video Technol.* **29**(12), 3608–3621 (2019). <https://doi.org/10.1109/TCSVT.2018.2883558>
14. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *European Conference on Computer Vision*, 11211(1), 1–10 (2018). [https://doi.org/10.1007/978-3-030-01234-2\\_49](https://doi.org/10.1007/978-3-030-01234-2_49)
15. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2017). <https://doi.org/10.1109/TPAMI.2016.2572683>
16. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018). <https://doi.org/10.1109/TPAMI.2017.2699184>
17. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. *IEEE/CVF Conf. Comput. Vis. Pattern Recogn.* **4**(1), 18–23 (2018). <https://doi.org/10.1109/CVPR.2018.00733>
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 9351(1), 234–241 (2015). [https://link.springer.com/chapter/https://doi.org/10.1007/978-3-319-24574-4\\_28](https://link.springer.com/chapter/https://doi.org/10.1007/978-3-319-24574-4_28)
19. Zhang, K., Xiong, F., Sun, P., Hu, L., Li, B., Yu, G.: Double anchor R-CNN for to human detection in a crowd. *J. Mobile Inform. Syst.* **2022**(2), 1–12 (2019). <https://doi.org/10.1155/2022/4012252>
20. Hsu, W., Lin, W.: Ratio-and-scale-aware YOLO for pedestrian detection. *IEEE Trans. Image Process.* **30**(2), 934–947 (2021). <https://doi.org/10.1109/TIP.2020.3039574>
21. Li, Y., Pang, Y., Cao, J., Shen, J., Shao, L.: Improving single shot object detection with feature scale unmixing. *IEEE Trans. Image Process.* **30**(2), 2708–2721 (2021). <https://doi.org/10.1109/TIP.2020.3048630>
22. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(2), 318–327 (2020). <https://doi.org/10.1109/TPAMI.2018.2858826>
23. Xie, J., Pang, Y., Khan, M., Anwer, R., Khan, F., Shao, L.: Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection. *IEEE Trans. Image Process.* **30**(1), 3872–3884 (2021). <https://doi.org/10.1109/TIP.2020.3040854>
24. Tang, Y., Li, B., Liu, M., Chen, B., Wang, Y., Ouyang, W.: AutoPedestrian: an automatic data augmentation and loss function search scheme for pedestrian detection. *IEEE Trans. Image Process.* **30**(1), 8483–8496 (2021). <https://doi.org/10.1109/TIP.2021.3115672>
25. Jocher, G., Chaurasia, A., Stoken, A., Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Mai Thanh Minh. *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference (v6.1)*. Zenodo (2022). <https://doi.org/10.5281/zenodo.6222936>
26. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context Encoding for Semantic Segmentation. *Proceed IEEE/CVF Conf. Comput. Vis. Patt. Recogn.* **30**(1), 7151–7160 (2018). <https://doi.org/10.1109/CVPR.2018.00747>
27. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-SCNN: gated shape CNNs for semantic segmentation. *Proceed. IEEE/CVF Int. Conf. Comput. Vis.* **30**(1), 1–5 (2019). <https://doi.org/10.1109/ICCV.2019.00533>
28. Yurtkulu, S., Şahin, Y., Unal, G.: Semantic segmentation with extended Deeplabv3 architecture. *Sign Process Commun Appl Conf* **1**(1), 1–5 (2019). <https://doi.org/10.1109/SIU.2019.8806244>
29. Zhao, H., Shi, J., Qi, X., Wang, X.: Pyramid scene parsing network. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **1**(1), 2881–2890 (2017). <https://doi.org/10.1109/CVPR.2017.660>
30. Liu, W., Rabinovich, A., Berg, A.: ParseNet: Looking Wider to See Better. *Proceedings of the IEEE Comput. Vis. Patt. Recogn.* **1**(1), 2881–2890 (2015). <https://arxiv.org/abs/1506.04579#:~:text=We%20present%20a%20technique%20for,the%20features%20at%20each%20location>
31. Alavianmeh, M. A., Helfroush, M. S., Danyali, H., Tashk, A.: A New Approach toward Pedestrian Detection based on A Mixture of Region Proposal and Semantic Segmentation Deep Convolutional Neural Networks. *The 11th Iranian and the 1st Int. Conf. Mach. Vis. Image Process.*, 1(2), 1–8 (2020). <https://mvip2020.ut.ac.ir/paper?manu=39055>
32. Chu, H., Ma, H., Li, X.: Pedestrian instance segmentation with the prior structure of semantic parts. *Pattern Recogn. Lett.* **149**(1), 9–16 (2021). <https://doi.org/10.1016/j.patrec.2021.05.012>
33. Li, Q., Qiang, H., Li, J.: Conditional random fields as message passing mechanism in the anchor-free network for multi-scale pedestrian detection. *Inform Sci* **550**(2), 1–12 (2021). <https://doi.org/10.1016/j.ins.2020.10.049>
34. Yang, P., Zhang, G., Wang, L., Xu, L., Deng, Q., Yang, M.: A Part-aware multi-scale fully convolutional network for pedestrian detection. *IEEE Trans. Intell. Transport. Syst.* **22**(2), 1125–1137 (2021). <https://doi.org/10.1109/TITS.2019.2963700>
35. Zhou, Z., Siddiquee, M., Tajbakhsh, N., Liang, J.: U-Net++: redesigning skip connections to exploit multi-scale features in image segmentation. *IEEE Trans. Med. Imag.* **39**(6), 1856–1867 (2019). <https://doi.org/10.1109/TMI.2019.2959609>



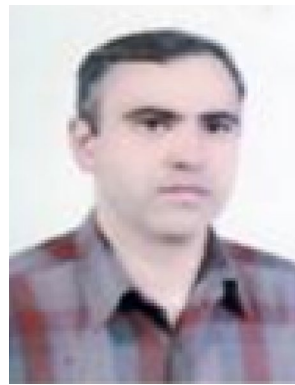
36. Huimin, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y., Wu, J. U-Net 3+: A full-scale connected U-net for medical image segmentation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 22(2), 1–10 (2020). <https://doi.org/10.1109/ICASSP40776.2020.9053405>
37. Ioffe, S., Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*, 37(1), 448–456 (2015). <https://proceedings.mlr.press/v37/ioffe15.html>
38. Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A. Modulating early visual processing by language. *Part of Advances in Neural Inform Process Syst.* 30(1), 1–10 (2017). <https://arxiv.org/abs/1707.00683>.
39. Perez, E., Vries, H., Strub, F., Dumoulin, V., Courville, A. Learning visual reasoning without strong priors. *Int. Conf. Comput. Vis. Patt. Recogn.* 1(1), 1–10 (2017). <https://arxiv.org/abs/1707.03017>.
40. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B. The cityscapes dataset for semantic urban scene understanding. *Proceed. IEEE Conf. Comput. Vis. Patt. Recogn.*, 2(1) 1–10 (2016). <https://arxiv.org/abs/1604.01685>.
41. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B. The Cityscapes Dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2(1), 1–10 (2015). <https://www.cityscapes-dataset.com/citation/>
42. Wang, L., Shi, J., Song, G., Shen, I.: Object detection combining recognition and segmentation. *Asian Conf. Comput. Vis.* 1(1), 189–199 (2007). [https://doi.org/10.1007/978-3-540-76386-4\\_17](https://doi.org/10.1007/978-3-540-76386-4_17)
43. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(7), 3523–3542 (2022). <https://doi.org/10.1109/TPAMI.2021.3059968>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



**Mohammad Ali Alavianmehr** received the B.Sc. degree in electrical engineering from Busher University, Busher, Iran, in 2006, the M.S. degree in electrical engineering from the Sistan & Baluchestan University, Zahedan, Iran, in 2012, and he is currently a PhD student at Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz, Iran. His research interests include different aspects of image and video processing, data hiding, pattern recognition and machine learning.



**Mohammad Sadegh Helfroush** received the B.Sc. degree in electrical engineering from Shiraz University, Shiraz, Iran, in 1993, the M.S. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 1995, and the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, in 2006. He is currently a Professor with the Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz. His research interests

include different aspects of signal and image processing, pattern recognition, and machine learning.



**Habibollah Danyali** received the B.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 1991, the M.Sc. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 1993, and the Ph.D. degree in computer engineering from the University of Wollongong, Wollongong, NSW, Australia, in 2004. From 1994 to 2000, he was with the Department of Electrical Engineering, University of Kurdistan, Sanandaj, Iran, as a Lecturer. He is

currently a Professor with the Department of Telecommunication Engineering, Shiraz University of Technology, Shiraz, Iran. His research interests include medical image processing, scalable image and video coding, data hiding, remote sensing, and machine learning.



**Ashkan Tashk** received the B.Sc. degree in electrical engineering from the Shiraz University, Shiraz, Iran, in 2006, the M.Sc. degree in electrical engineering from Shiraz University of Technology, Shiraz, Iran, in 2009, and the Ph.D. degree in Telecommunications Systems from Shiraz University of Technology, Shiraz, Iran, in 2015. He is currently a Postdoc student at Unit of Applied AI and Data Science, Mærsk Mc-Kinney Møller Institute (mami), University of Southern Denmark (SDU),

Odense, Denmark. His research interests include medical image processing, pattern recognition, image classification and object detection.