Computers and Electrical Engineering 99 (2022) 107781

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

An infrared pedestrian detection method based on segmentation and domain adaptation learning☆

Jianlong Zhang [a], Chishuai Liu [a], Bin Wang [a,*], Chen Chen [b], Jianhui He [a], Yang Zhou [c], Ji Li [d]

[a] School of Electronic Engineering, Xidian University, Xi'an, 710000, Shaanxi, China
[b] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, 710000, Shaanxi, China
[c] The Ministry of water resources of China, Beijing, 101400, China
[d] The Goldenwater Information Technology Co. Ltd., Beijing, 101400, China

ARTICLE INFO

Keywords:
Infrared pedestrian detection
Multi-task learning
U-Net
Domain adaptation
Swin Transformer
Semantic segmentation
Gradient inversion

ABSTRACT

Benefiting from the capability of night viewing, infrared images has been widely applied to surveillance systems as an effective complement to visible-light images. However, the development of infrared pedestrian detection is still impeded by weak features and limited diversity of infrared images. Aiming at these two problems, we designed a multi-task learning framework for pedestrian detection by incorporating a semantic segmentation branch and a domain adaptation branch. Composed of UNet network with Swin Transformer, the semantic segmentation could apply spatial constraints to pedestrian detection. The domain adaptation branch aligns the features between infrared and visible-light images to improve the scene diversity. In addition, three tasks shared a basic feature extraction network to reduce computation cost. The experiment results show that the average precision (AP) of our method is superior to the EfficientDet network by 2.0% on the XDU-NIR2020 dataset and 2.2% on the CVC-09 dataset respectively.

1. Introduction

In recent years, The development of artificial intelligence technology has contributed to the progress of the Internet of Things [1,2], content distribution [3] and other fields. Among these fields, computer vision plays a vital role, for example, the lightweight and efficient segmentation network empowers the city surveillance camera with the ability to find potential dangers in time [4]. In addition, computer vision is also the fundamental technology of automatic driving and intelligent monitoring. Pedestrian detection, one of the most important functionalities, could locate pedestrians in images or videos to realize obstacle avoidance and keep pedestrians and vehicles in safe status. The performance of current pedestrian detection methods for visible images has made great progress, but in the low-illumination environment, the accuracy decreases sharply. Infrared pedestrian detection is more suitable for the low-illumination case of automatic driving, therefore, some commercial vehicles were equipped with infrared imaging devices to assist driving and give pre-warning about the emergence of pedestrians. In addition, benefiting from the low hardware cost and high imaging quality of infrared imaging, infrared imaging is also applied to current mainstream monitoring scenes, e.g., residential entrance or factory park, to realize 24-hours effective monitoring. Herein, we proposed an efficient infrared detection method based on a multi-task learning network, which has a lower false alarm rate and good generalization ability.

☆ This paper is for regular issues of CAEE. Reviews were processed and recommended for publication by Co-Editor in Chief Prof Huimin Lu.

* Corresponding author.
E-mail address: bwang@xidian.edu.cn (B. Wang).

https://doi.org/10.1016/j.compeleceng.2022.107781
Received 9 July 2021; Received in revised form 31 January 2022; Accepted 1 February 2022
Available online 19 February 2022

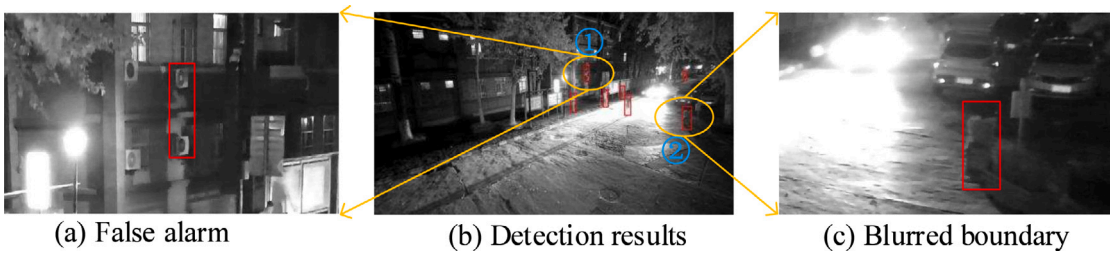

Fig. 1. XDU-NIR2020 dataset.

The rest of the paper is arranged as follows. Section 2 briefly introduces the related work of infrared pedestrian detection. Section 3 introduces the principles and implementation of the proposed method in detail. Section 4 gives the experimental results and analysis. Section 5 summarizes our work.

2. Related work

Early pedestrian detection algorithms mainly used traditional image features for recognition and localization. The most representative feature is the histograms of oriented gradients (HOG) [5] proposed by Navneet Dalal. Dollar proposed the integrated channel feature (ICF) [6] and achieved high detection performance on the INRIA dataset [5] by combining ICF and the other features, e.g., color and gradient. However, the hand-craft features above are extremely susceptible to background interference and cannot maintain robustness in complex real environments.

In recent years, convolutional neural networks have become the mainstream technique for pedestrian detection [7]. A representative method, Faster RCNN [8], is a typical two-stage detection network. In its first stage, the proposal boxes are obtained through the region proposal network, and the refined detection results are obtained through the regression classification network in the second stage. EfficientDet [9] uses EfficientNet [10] as a feature extraction network, utilizes weighted bi-directional feature pyramid network (BiFPN) [9] to fuse multi-scale features, and surpasses a mainstream single-stage detection network in accuracy and speed. Heo proposed an adaptive boolean-map-based saliency (ABMS) method [11] to improve detection accuracy for infrared images. It combines YOLO-V2 [12] features with ABMS to make the target more salient and obtains competitive results on the far-infrared image datasets. Yingfeng Cai et al. proposed an infrared pedestrian detection method based on visual saliency [13] and obtained a detection model with high accuracy and speed. Wang D et al. proposed a novel infrared pedestrian detection network [14] in a per-pixel prediction fashion to solve the problems of feature shortage and diversity in infrared images and achieved better performance on the South China University of Technology (SCUT) dataset [15]. Yunfan Chen et al. proposed an infrared camera system [16] to overcome shortcomings of low resolution and suppress noise in infrared images, which used an attention-guided coding–decoding convolutional neural network to effectively highlight pedestrian features while eliminating background interference. However, these infrared pedestrian detection methods still have the potential to promote unless the following problems are well addressed.

(1) Infrared images are featured by low contrast, low signal-to-noise ratio and the blurred boundaries between the target and the background. This makes the features learned from infrared images are not enough to avoid interference from the background. Fig. 1(b) shows the detection results of the detection network EfficientDet on the XDU-NIR2020 dataset[1] with a confidence of 0.25; Fig. 1(c) is an enlarged view of area ② in Fig. 1(b), and the pedestrian contours are blurry; Fig. 1(a) is an enlarged view of area ① in Fig. 1(b), and false alarms appear in the non-pedestrian activity area. As recall rates rise, false alarms will rise further.

(2) The infrared pedestrian detection datasets are mainly collected by vehicle-mounted cameras or surveillance cameras, and the scenes are very similar, which makes the generalization ability of pedestrian detection is not satisfactory. Fig. 2 shows three images of the CVC-09 dataset with a numbered interval of 10. The difference between the backgrounds in the sampled frames is small. It means that the sample diversity is limited, which is a reason for the poor generalization ability of the pedestrian detection methods in different scenes.

To address the above problems and improve detection accuracy, we utilize multi-task learning framework to combine detection and segmentation that is to filter out the interference from background. Meanwhile, in our method, EfficientDet is taken as the core detection network, segmentation and domain adaptation as the additional branches to improve the detection performance in different scenes. In addition, for the sake of saving computation cost, detection, segmentation and domain adaptation share the same features.

The main advantages of our method are as follows. (1) The segmentation network based on U-Net [17] and Swin Transformer (Swin TR) [18] structures are both used to predict the pedestrian activity area, then the detection parts in non-pedestrian areas are filtered out to reduce false alarms while keeping a high recall rate. (2) In the training process, the domain adaptation method is introduced to fuse the features from both visible light images and infrared images in the local and global manners with the consistency constraints. Therefore, visible light images are combined with infrared images to improve the generalization performance of infrared pedestrian detection model.

[1] XDU-NIR2020 dataset is a self-collected and non-public dataset at present.

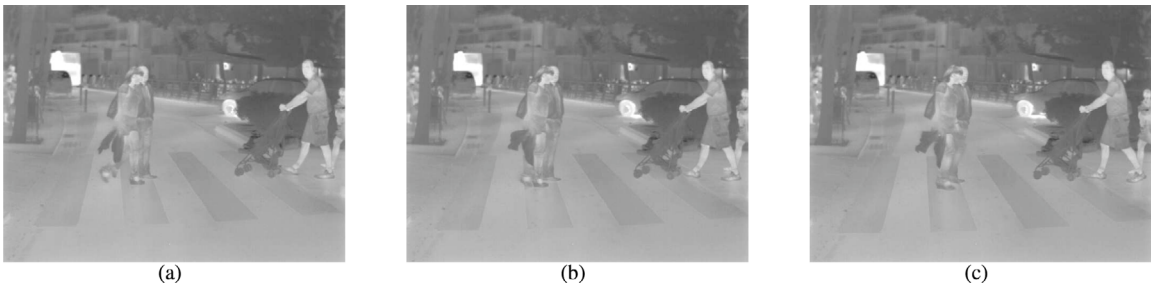

Fig. 2. Samples in CVC-09.

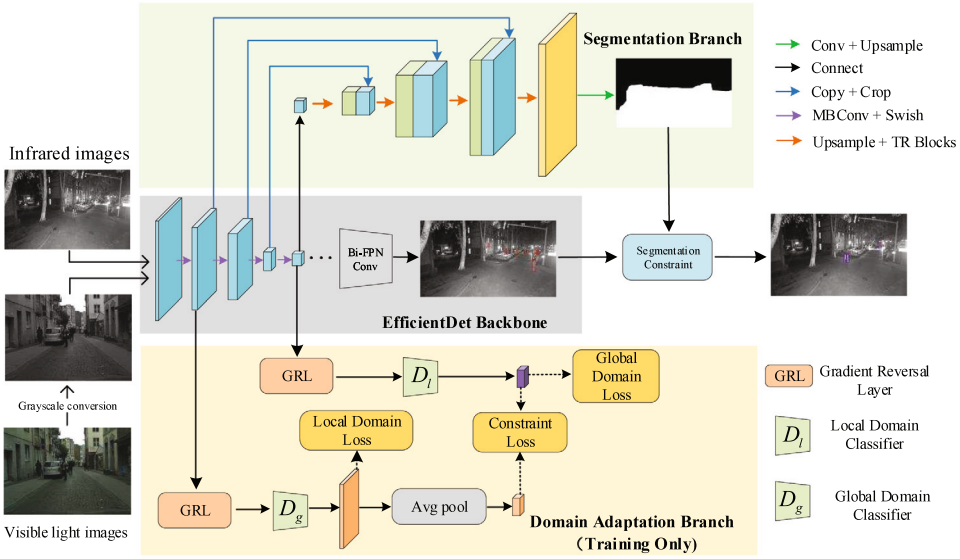

Fig. 3. Structure of multi-task learning based infrared pedestrian detection.

3. Algorithm principles and framework

The proposed infrared pedestrian detection method, shown in Fig. 3, is composed of the following parts, i.e., (1) the main network for object detection based on EfficientDet, (2) the segmentation branch with U-Net and Swin Transformer, and (3) a domain adaptation branch with gradient reversal layer (GRL) [19]. To be specific, the patch merging module [18] of Swin TR is replaced by bilinear interpolation module to improve the global relevance during the process of up-sampling for the segmentation task. For the domain adaptation task, the feature extractor of EfficientDet focuses on common features of different domains by domain classifier and GRL. The visible images are used for training like as same as infrared images. The domain adaptation branch mainly completes the expansion of the training sets, therefore it is only worked in the process of training.

3.1. Semantic segmentation task

The framework of the segmentation branch, shown in Fig. 4, uses the structure of encoder and decoder like U-Net network. Meanwhile, to decrease the network complexity, the encoder shares P2–P5 layers of the EfficientDet, and the decoder consists of fusion operations and up-sampling operations with Swin TR. In order to utilize the multi-scale features extracted by the decoder, the outputs of D4–D2 are cascaded to D1 with the same resolution.

Classical segmentation networks, e.g., U-Net, adopt convolution operation to extract image features. But the convolution focuses on the local correlation of features more than the global correlation, which is not conducive to the large scene segmentation in our task. This problem is tackled better by Transformer model [20]. Transformer is composed of attention module and full connection layer to calculate the correlation in all of the input vectors to obtain a stronger global relevance. However, the computation complexity of Transformer is square of the size of input image, which costs more memory as well as power and restricts promotion and application of transformer in high-resolution images.

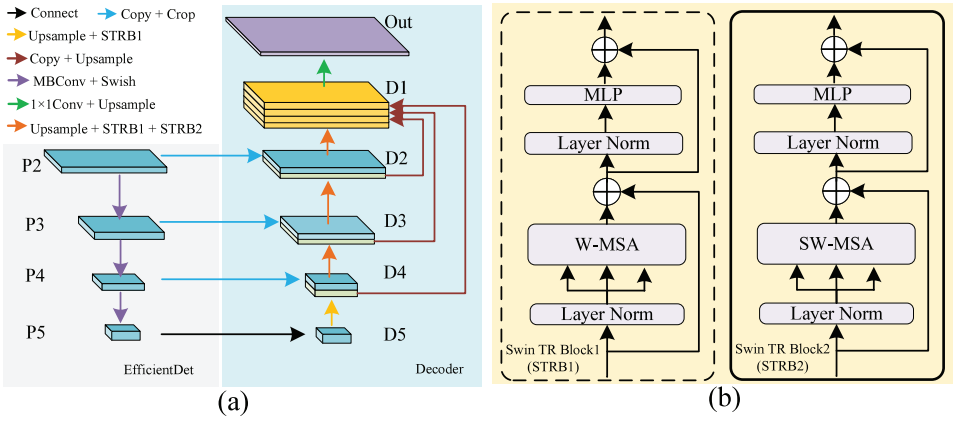

Fig. 4. Structure of the segmentation branch.

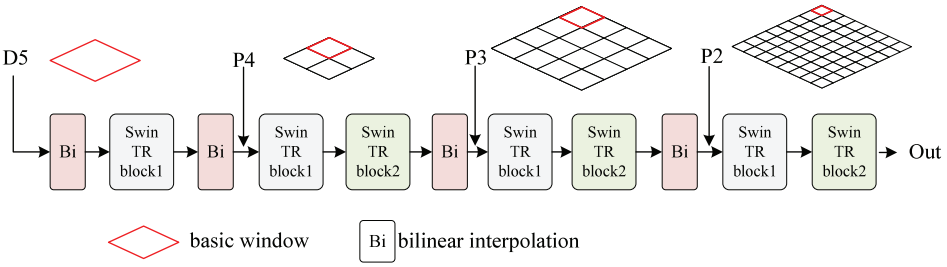

Fig. 5. Diagram of the decoding for the segmentation branch.

To realize efficient segmentation, we utilize Swin TR [18] with hierarchical attention, in which feature maps of each hierarchical layer are divided into many equal-size windows, and only self-attention within the window is calculated instead of all features, so the computational complexity can be effectively reduced. The structure of Swin TR, shown in Fig. 4(b), consists of two cascaded parts, i.e., Block1 that calculates correlation within the windows by W-MSA module, and Block2 that calculates the correlation between adjacent windows with SW-MSA module. When Swin TR, however, is applied to the decoder of U-Net network, there are still adaptations to make as follows. (1) Swin TR implements down-sampling by Patch Merging module in which multi-dimensional features in the same window are mapped into a vector to reduced dimensionality. It is obvious that this module cannot complete the up sampling of decoding layer features. Therefore, bilinear interpolation is introduced to replace the Patch Merging. This process of implementation is shown in Fig. 4(a). (2) In Swin TR, the default size of the attention window is 7×7, however, the encoder features have different size in each layer, which should guide Swin TR to generate attention window with corresponding size. In our method, the size of minimum feature layer D5 is defined as the basic window. D4–D2 are correspondingly divided into 4, 16 and 64 basic windows to match the feature sizes in decoder. The schematic diagram of the decoding is shown in Fig. 5. The red box is the basic window, which is the same size as D5.

The segmentation branch, herein, predicts the candidate area of pedestrian activity (M0 in Fig. 6) which usually contain noise and holes. Therefore, morphological operations of dilation and erosion [21] are used to remove noise and holes, and a complete pedestrian active area (M1) is obtained. In this way, the objects outside M1 area are filtered out to reduce the ratio of false alarms and achieve better effects under a high recall rate.

The general loss function of semantic segmentation model is the cross entropy L_{bce} which pays the attention to all points of the prediction results equally. However, in the task of pedestrian activity area prediction, the amount of foreground points is less than that of background points. Considering that the points belonging to objects play more important role in our task, Dice similarity coefficient L_{Dice} [22] is used. Our final loss function is shown as formula (1).

$$L_{\text{seg}} = \alpha L_{\text{bce}} + \beta L_{\text{Dice}} \quad \text{s.t. } \alpha + \beta = 1 \tag{1}$$

where α and β are loss weight for L_{bce} and L_{Dice}, respectively.

3.2. Domain adaptation task

The existing image enhancement algorithms, e.g., zoom, translation, rotation, are not effective in increasing the scene diversity of the dataset. Considering the shortage of suitable infrared pedestrian datasets, we introduced visible light data into our training dataset by domain adaptation method to improve its scenes generalization capability.

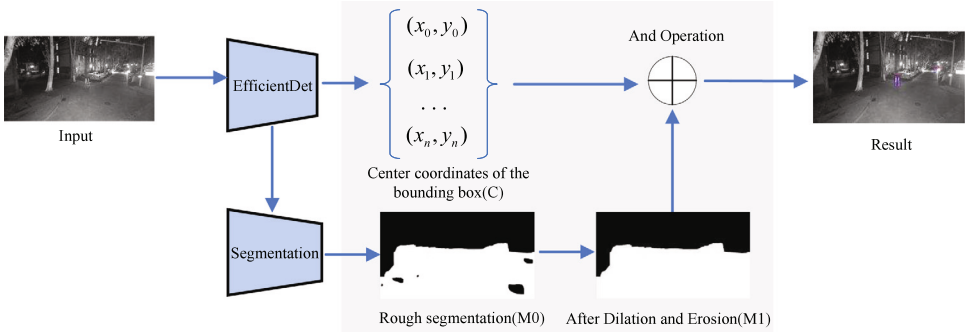

Fig. 6. The segmentation constraints.

The domain adaptation method is mainly applied in unsupervised method for object detection, and its core modules are domain discriminator and gradient inversion layer. The domain discriminator is used to calculate the domain probability of the image belonging to. By using gradient inversion layer (GRL), feature extraction of EfficientDet is updated along the direction of increasing domain classification loss function, so that the feature space of different domains can be aligned which contributes to network training on the data from different domains. DA-Faster R-CNN [23] is a classic domain adaptation method based on Faster RCNN, which constrains the domain output results on instance-level and image-level by consistent regularization. SW-Faster R-CNN [24] abbreviated as SW, proves that instant-level feature alignment has little effect for domain alignment, low-level feature alignment narrows the style characteristics (e.g., hue, contrast, illumination) of the two domains and contributes to domain feature alignment.

Borrowed the consistency regularization structure and SW, alignment is applied to local feature (P2) and global feature (P5), meanwhile, the consistency constraint is employed to decrease the alignment difference between different layers, as shown in Fig. 7. The local feature alignment and global alignment are realized by minimizing the following loss, i.e.,

$$L_{\text{local}} = \frac{W_{\text{local}}}{(n_s + n_t)\,HW} \sum_{i=1}^{n_s+n_t} \sum_{w=1}^{W} \sum_{h=1}^{H} \left(y_{\text{iwh}} - y'_{\text{iwh}}\right)^2 \tag{2}$$

$$L_{\text{global}} = -\frac{W_{\text{global}}}{(n_s + n_t)} \sum_{i=1}^{n_s+n_t} \left(y_i \log y'_i + (1 - y_i) \log\left(1 - y'_i\right)\right) \tag{3}$$

where n_s and n_t represent the number of visible light images and infrared images in each training batch. H and W represent the height and width of the prediction probability map, respectively. y_{iwh} denotes the output of domain classifier in each location, and y'_{iwh} denotes the corresponding domain label. y_i denotes the domain classification probability of the ith image, and y'_i denotes the corresponding domain label. W_{local} and W_{global} are used to adjust the proportion of local and global domain adaptation loss, and both are set as 0.5 in our work.

In order to avoid the ambiguity of domain classification, the consistency loss function L_{cst} is designed to calculate the loss between the average classification probability $P2$ and the classification probability $P5$. L_{cst} is shown as formula (4).

$$L_{cst} = \frac{W_{cst}}{n_s + n_t} \sum_{i=1}^{n_s+n_t} \left| \frac{1}{P_{Loc}} \sum_{w=1}^{W} \sum_{h=1}^{H} p^i_{w,h} - p^i_g \right| \tag{4}$$

where the definitions of n_s, n_t, H and W are consistent with formula (3). P_{Loc} is the number of feature points in the output probability map of the local domain classifier. $P^i_{w,h}$ and p^i_g denote the local domain classification probability in each location and global classification probability of the ith image respectively.

4. Experiments

4.1. Dataset

To verify the performance of our method, the near-infrared dataset XDU-NIR2020 and far-infrared dataset CVC-09 are taken as the target domain data, and visible light dataset WiderPerson [25] and CrowdHuman [26] are taken as the source domain data. The XDU-NIR2020 dataset contains 16 498 images with a resolution of 2560 × 1440. Its images were all captured by the surveillance camera where 13 113 images are used for training and 3385 images for testing. The images of CVC-09 dataset were collected by vehicle camera and divided into day and night groups. Since infrared pedestrian detection is mainly applied in low illumination environment, the night group of CVC-09 is selected as the target domain data to verify the performance of our method. It contains 5081 images of 640 × 480. WiderPerson is a dataset for pedestrian detection, and not limited to traffic scenes. It consists of a train set of 8000 images, a verification set of 1000 images and a test set of 4382 images. In order to match the number of infrared datasets, we randomly selected 4000 images from CrowdHuman's train set and added them to WiderPerson's train set. Many images of the above datasets are shown in Fig. 8.

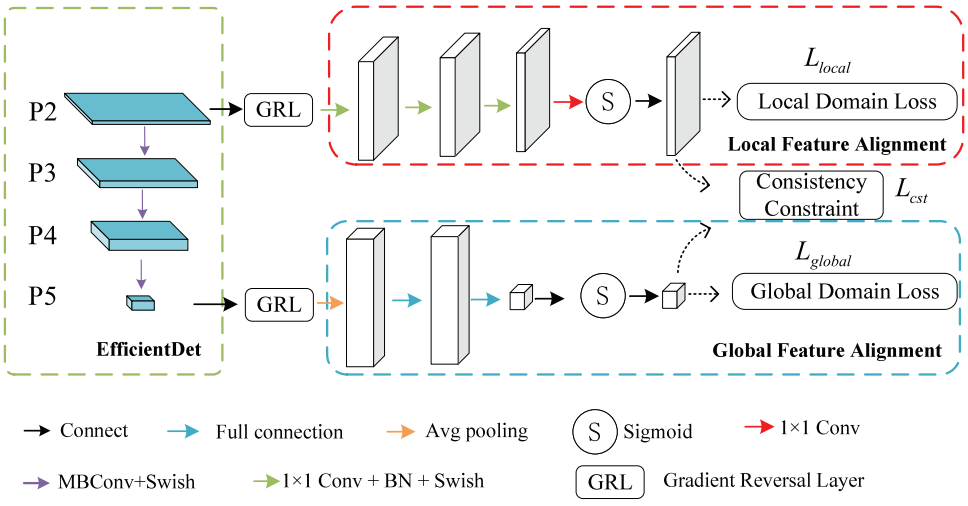

Fig. 7. Structure of the domain adaptation branch.

Fig. 8. Samples of the datasets.

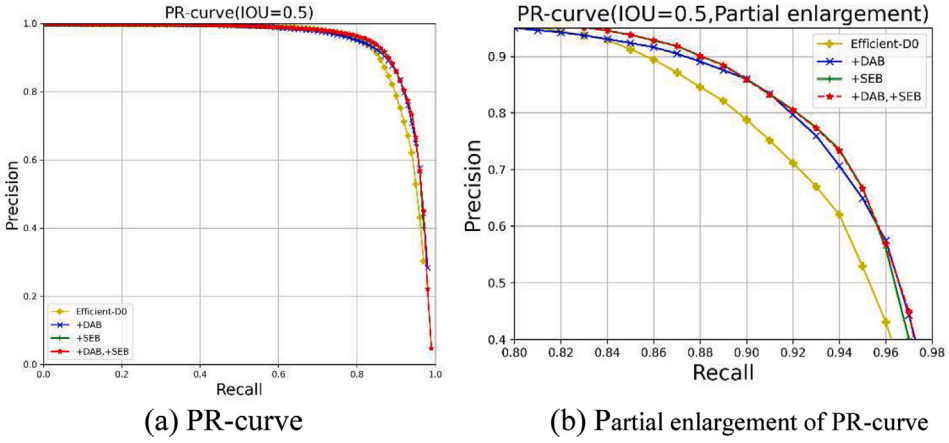

Fig. 9. PR curves of ablation experiments.

4.2. Implementation details

Since it is difficult to train all of three branches of networks and make them converged at the same time, the network training of our method has two stages. In the first stage, target domain and source domain data are utilized to train the detection branch and the domain adaptation branch. In the second stage, the weights of the detection and domain adaptation branch are frozen, and only the segmentation network is trained. It means that training of segmentation network only uses the data of target domain.

The object detection branch uses EfficientDet-D0 as the detection model. AdamW [27] is selected as the training optimizer with the learning rate as 0.0003, batch-size as 8; the segmentation branch learning rate is set as 0.000001; the domain adaptation branch learning rate is set as 0.0003; and the epochs of two stages are set as 40. Our training model is implemented in Python 3.7 using

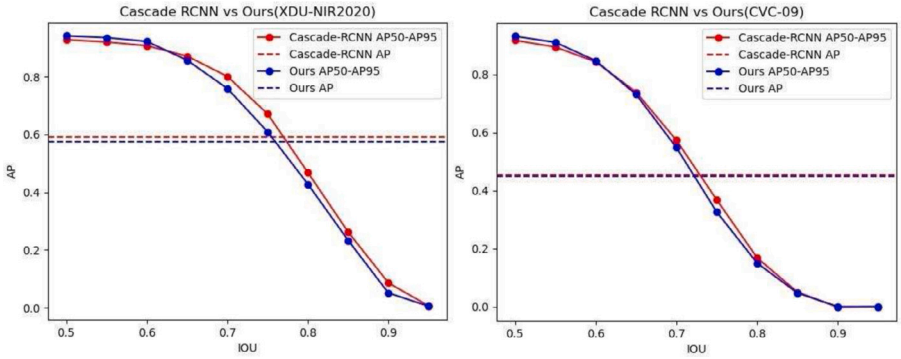

Fig. 10. The comparison between Cascade RCNN and our method. The horizontal axis indicates that the IOU samples from 0.5 to 0.95 every 0.05, and the vertical axis indicates the corresponding AP index.

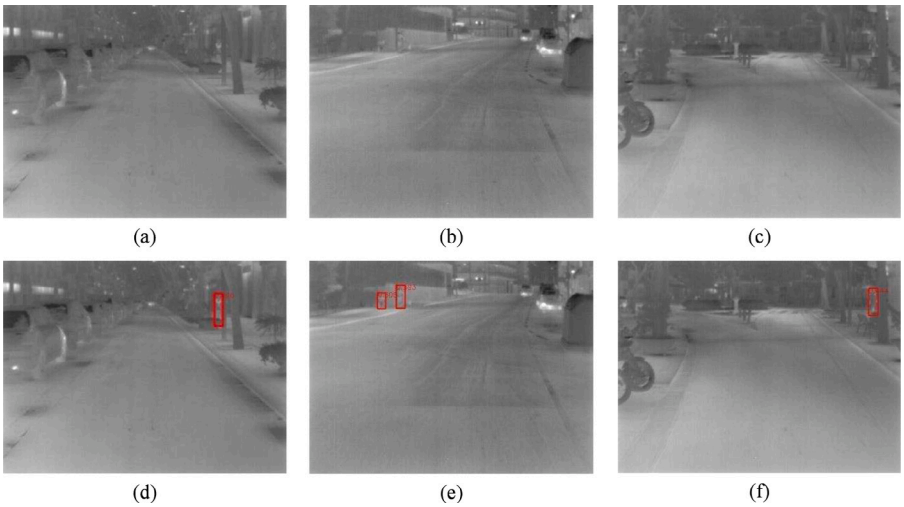

Fig. 11. The detection results of some images in CVC-09 dataset. The first line is the detection result of EfficientDet-D0, and the second line is the detection result of our method. (d) (e) (f) solve some missed alarms in (a) (b) (c) respectively.

Table 1
Pedestrian detection ablation experiment based on multi-task learning (XDU-NIR2020).

Method	SEB	DAB	AP	AP50
EfficientDet-D0	–	–	0.554	0.924
+DAB	–	✓	0.566	0.934
+SEB	✓	–	0.568	0.937
+DAB,+SEB	✓	✓	0.574	0.941

PyTorch 1.8 on a PC with Intel(R) Core (TM) i7-7820X CPU @ 3.60 GHz, Nvidia GTX 2080Ti×2. The performance is measured in terms of average precision AP and AP50.

4.3. Ablation experiment

We took EfficientDet-D0 as the backbone network, and combined domain adaptation branch (DAB) and segmentation branch (SEB). These two branches were verified by ablation analysis in XDU-NIR2020 dataset. The results are shown in Table 1. AP is increased by 1.2% when adding domain adaptation branches. AP is improved by 1.4% when adding segmentation branches. When adding these two branches both, AP is increased by 2.0%. It shows that both domain adaptation branches and segmentation branches could effectively improve the performance of infrared pedestrian detection. PR curve of ablation experiment is shown in Fig. 9. According to the local enlarged PR curve of Fig. 9(b), the accuracy of detection network is greatly improved by only adding segmentation branch or domain adaptation branch, especially for the case of high recall rate. Adding two branches both integrates the advantages of segmentation and domain adaptation, and makes the detection accuracy slightly improve.

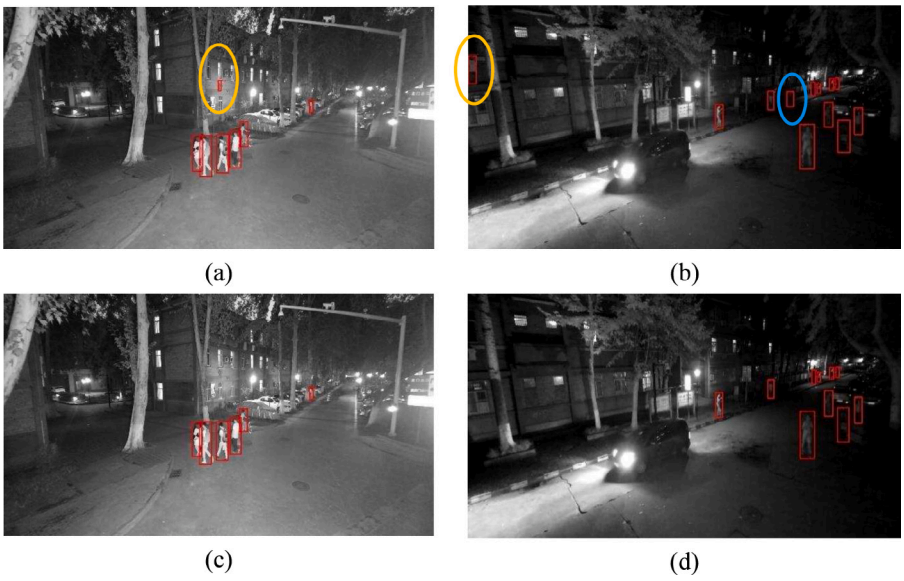

Fig. 12. (a) and (b) are the detection results of EfficientDet-D0. (c) and (d) are the detection results of our method. The segmentation branch filters out the false alarms of the non-pedestrian parts in the orange area. The domain alignment branch filters out the false alarm caused by the background blur in the blue area.

Table 2
Our method is compared with other detection algorithms (XDU-NIR2020).

Method	Image scale	Frame rata	AP	AP50
FCOS-R101	1280 × 720	12.8	0.559	0.919
CascadeRCNN-R101	1280 × 720	10.2	0.592	0.929
EfficientDet-D0	1280 × 720	42.5	0.554	0.924
our method	1280 × 720	27.1	0.574	0.941

Table 3
Our method is compared with other detection algorithms (CVC-09).

Method	Image scale	Frame rata	AP	AP50
FCOS-R101	1280 × 960	10.2	0.435	0.890
CascadeRCNN-R101	1280 × 960	8.5	0.456	0.917
EfficientDet-D0	1280 × 960	38.2	0.427	0.903
our method	1280 × 960	26.1	0.449	0.930

4.4. Comparison of mainstream methods

We selected three typical object detection methods, e.g., Cascade RCNN [28], FCOS [29] and EfficientDet [9] as the mainstream methods for verification. The comparison results are shown in Tables 2 and 3. Compared with single-stage network FCOS, our method has faster detection speed and higher AP and AP50 indexes. The AP of our method is lower than that of Cascade RCNN, but the detection speed is about three times faster than that of Cascade RCNN. Fig. 10 shows the average precision whose IOU threshold value with an interval of 0.05 from 0.5 to 0.95 on XDU-NIIR2020 dataset, which is called as AP50–AP95. The AP50–AP60 of our method are higher than that of Cascade RCNN, the AP65–AP95 are lower than that of Cascade RCNN, and these scores indicates that the accuracy of border regression in our method is worse than that of Cascade RCNN. When the demand for border regression is not strict, our proposed method has the advantages of accuracy and speed. Even under the constraints of high IOU, compared with excellent two-stage network Cascade RCNN, the average accuracy of the proposed method is not significantly inferior. Compared with EfficientDet-D0, our method improves AP and AP50 by 2.0% and 1.7% on XDU-NIR2020 dataset, and 2.2% and 2.7% on CVC-09 dataset.

Since our method combined the segmentation with detection, its speed is slightly reduced. Figs. 11 and 12 show the comparison results between our method and EfficientDet. We selected some detection results from the CVC-09 dataset, as shown in Fig. 11, in which the pedestrian activity area occupies almost the whole images. Obviously, for Fig. 11, the segmentation constraint is limited, but our method detects some missing alarm in Efficientdet, indicating that the domain adaptation branch plays an important role. Fig. 12 shows the comparison between EfficientDet and our method on XDU-NIR2020 dataset. For Fig. 12(a) and (b), the false alarms in orange area are obviously out of the pedestrian activity area, so they are filtered by the segmentation branch of our

method. In Fig. 12(b), the false alarm in blue area appeared due to the fuzzy background, and this false alarm was eliminated by the domain adaptation network of our method. The above comparative experiments show that segmentation task and the domain adaptation task are necessary and effective. Compared with the traditional single-stage network, the detection accuracy of our method is significantly improved. Compared with the traditional two-stage network, the detection speed of our method is greatly improved while the detection accuracy is close to the two-stage network.

5. Conclusion and future work

We proposed an infrared pedestrian detection method based on multi-task learning framework where three tasks, i.e., detection, segmentation and domain adaptation, work together. The EfficientDet completes main detection task. The semantic segmentation branch is introduced to predict the pedestrian activity area to reduce the false alarm of detection. The domain adaptation branch is employed to increase the diversity of scenes. The ablation and comparison experiments show that our method could effectively improve the performance of the infrared pedestrian detection, however, the method of domain adaptation still relies on the diversity of available domain public dataset. In the future, generative adversarial networks will be considered to produce infrared images for auxiliary training.

CRediT authorship contribution statement

Jianlong Zhang: Conceptualization, Writing – original draft, Methodology, Project administration. **Chishuai Liu:** Writing – original draft, Acquisition of data. **Bin Wang:** Conception and design, Writing – review & editing. **Chen Chen:** Analysis and/or interpretation of data, Validation, Investigation. **Jianhui He:** Writing – review & editing, Term, Conceptualization. **Yang Zhou:** Data curation. **Ji Li:** Software.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.compeleceng.2022.107781.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020YFB1807500), the Aeronautical Science Foundation of China (2018ZC81001), the National Natural Science Foundation of China (62072360, 61902292, 61971331, 62001357), the Key Research and Development Plan of Shaanxi province (2020JQ-844, 2021ZDLGY02-09, 2019ZDLGY13-07, 2019ZDLGY13-04), the Key Laboratory of Embedded System and Service Computing (Tongji University) (ESSCKF2019-05), Ministry of Education, the Xi'an Science and Technology Plan, China (20RGZN0005) and the Xi'an Key Laboratory of Mobile Edge Computing and Security (201805052-ZD3CG36).

References

[1] Qiu T, Wang X, Chen C, Atiquzzaman M, Liu L. TMED: A spider-web-like transmission mechanism for emergency data in vehicular ad hoc networks. IEEE Trans Veh Technol 2018;67(9):8682–94.
[2] Qiu T, Zhang S, Si W, Cao Q, Atiquzzaman M. A 3D topology evolution scheme with self-adaption for industrial internet of things. IEEE Internet Things J 2020.
[3] Fu S, Ma L, Atiquzzaman M, Lee Y-J. Architecture and performance of SIGMA: A seamless mobility architecture for data networks. In: IEEE international conference on communications, vol. 5. IEEE; 2005, p. 3249–53.
[4] Sun J, Li Y. Multi-feature fusion network for road scene semantic segmentation. Comput Electr Eng 2021;92:107155.
[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition, vol. 1. Ieee; 2005, p. 886–93.
[6] Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 2011;34(4):743–61.
[7] Brunetti A, Buongiorno D, Trotta GF, Bevilacqua V. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing 2018;300:17–33.
[8] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 2015;28:91–9.
[9] Tan M, Pang R, Le QV. Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 10781–90.
[10] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019, p. 6105–14.
[11] Heo D, Lee E, Ko BC. Pedestrian detection at night using deep neural networks and saliency maps. Electron Imaging 2018;2018(17):060401–3.
[12] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 7263–71.
[13] Cai Y, Liu Z, Wang H, Sun X. Saliency-based pedestrian detection in far infrared images. IEEE Access 2017;5:5013–9.
[14] Wang D, Lan J. PPDet: A novel infrared pedestrian detection network in a per-pixel prediction fashion. Infrared Phys Technol 2021;103965.
[15] Xu Z, Zhuang J, Liu Q, Zhou J, Peng S. Benchmarking a large-scale FIR dataset for on-road pedestrian detection. Infrared Phys Technol 2019;96:199–208.
[16] Chen Y, Shin H. Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network. Appl Sci 2020;10(3):809.
[17] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015, p. 234–41.

[18] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. 2021, arXiv preprint arXiv:2103.14030.
[19] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. PMLR; 2015, p. 1180–9.
[20] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Adv neural inf process syst. 2017, p. 5998–6008.
[21] Gil JY, Kimmel R. Efficient dilation, erosion, opening, and closing algorithms. IEEE Trans Pattern Anal Mach Intell 2002;24(12):1606–17.
[22] Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision. IEEE; 2016, p. 565–71.
[23] Chen Y, Li W, Sakaridis C, Dai D, Van Gool L. Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition. 2018, p. 3339–48.
[24] Saito K, Ushiku Y, Harada T, Saenko K. Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 6956–65.
[25] Zhang S, Xie Y, Wan J, Xia H, Li SZ, Guo G. Widerperson: A diverse dataset for dense pedestrian detection in the wild. IEEE Trans Multimed 2019;22(2):380–93.
[26] Shao S, Zhao Z, Li B, Xiao T, Yu G, Zhang X, et al. Crowdhuman: A benchmark for detecting human in a crowd. 2018, arXiv preprint arXiv:1805.00123.
[27] Loshchilov I, Hutter F. Fixing weight decay regularization in adam. 2018.
[28] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 6154–62.
[29] Tian Z, Shen C, Chen H, He T. Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9627–36.

Jianlong Zhang received the B.Eng. and M.Sc. degree in electronic engineering from Northwestern Polytechnical University, and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1998, 2001, and 2007 respectively. He worked currently in Xidian University. His research direction includes pattern recognition, machine learning and computer vision.

Chishuai Liu received the B.Eng. in Electronic and Information Engineering from Wuhan University of Science and Technology, Wuhan, China, in 2019. Since 2019, he has been working on Master's Degree in Xidian University, Xi'an, China. His research interests include object detection, machine learning and pattern recognition.

Bin Wang received the B.Sc. and M.Sc. degrees in electronics and information system from Northwest University and the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 1999, 2002 and 2010, respectively. He worked currently in Xidian University. His research interest is image segmentation and analysis.

Chen Chen (M'09-SM'18) received the B.Eng., M.Sc. and Ph.D. degrees in telecommunication from Xidian University, Xi'an, China, in 2000, 2006, and 2008, respectively. He worked currently in Xidian University. His research interest is 6G communication and edge computing.

Jianhui He received the B.Eng. in integrated circuit design Major of integrated system from Huaqiao University, Fuzhou, China, in 2018 and M.Sc. degree in electronic engineering from Xidian University. He has been study in computer vision, object detection and semantic segmentation.

Yang Zhou received the B.Eng. in communication engineering from Xidian University, Xi'an, China, in 2000. Since 2000, he has been with the Ministry of water resources of China. His research interests include wireless communication and computer engineering.

Ji Li received the B.Eng. in computer science and technology and the M.Sc. degree in software engineering from Beihang University, Beijing, China, in 2002 and 2007 respectively. He has been with the Ministry of Water Resources Information Center, China, and engaged in consulting, designing and researching of Goldenwater Information Technology Co. Ltd. for a long time.