EAE-Net: effective and efficient X-ray joint detection[*]

WU Zhichao[1†], WAN Mingxuan[1], BAI Haohao[2†], MA Jianxiong[2], and MA Xinlong[2]**

1. School of Control Science and Engineering, Tiangong University, Tianjin 300387, China

2. Orthopaedics Institute, Tianjin Hospital, Tianjin University, Tianjin 300211, China

(Received 12 July 2023; Revised 3 April 2024)

The detection and localization of bone joint regions in medical X-ray images are essential for contemporary medical diagnostics. Traditional methods rely on subjective interpretation by physicians, leading to variability and potential errors. Automated bone joint detection techniques have become feasible with advancements in general-purpose object detection. However, applying these algorithms to X-ray images faces challenges due to the domain gap. To overcome these challenges, a novel framework called effective and efficient network (EAE-Net) is proposed. It incorporates a context augment module (CAM) to leverage global structural information and a ghost bottleneck module (GBM) to reduce redundant features. The EAE-Net model achieves exceptional detection performance, striking a balance between accuracy and speed. This advancement improves efficiency, enabling clinicians to focus on critical aspects of diagnosis and treatment.

Document code: A **Article ID:** 1673-1905(2024)10-0629-7

DOI https://doi.org/10.1007/s11801-024-3129-y

The detection and localization of bone joint regions in medical X-ray images have emerged as a pivotal task in contemporary medical diagnostics[1]. This critical undertaking plays an indispensable role in the assessment and treatment of various bone and joint disorders[2], encompassing fractures, arthritis, and deformities. Accurate identification and localization of bone joint regions enable healthcare professionals to make informed decisions regarding patient care, treatment planning, and surgical interventions.

Traditional approaches to bone joint detection heavily rely on physician expertise and domain knowledge, which are typically subjective, reliant on clinical experience, and vulnerable to human errors. These conventional methods often involve manual interpretation of medical images, where the accuracy and consistency of detection can vary among different practitioners. The subjective nature of these approaches introduces a level of uncertainty and inconsistency in the detection and localization of bone joint regions. Moreover, human errors, such as fatigue or cognitive biases, can further impact the reliability and accuracy of the results. Therefore, there is a growing recognition of the necessity to develop automated and objective methods that can overcome these limitations and provide robust and reliable bone joint detection.

With the rapid advancements in general-purpose object detection in computer vision, automated bone joint detection techniques have become more feasible. The aim of general-purpose object detection is to identify and localize objects of interest in an image, simultaneously determining their positions and categories. It can be categorized into one-stage[3] and two-stage algorithms. Two-stage algorithms extract features using candidate boxes and utilize convolutional neural networks (CNN)[4] for classification and regression, offering the advantage of high precision but at the cost of slower detection speed, as exemplified by Faster-RCNN[5]. Conversely, one-stage algorithms employ CNN to directly classify and locate objects, providing the primary benefit of faster detection speed, albeit with the drawback of lower detection accuracy, as seen in the you only look once (YOLO) series[6].

However, the application of general-purpose object detection algorithms for bone detection encounters significant challenges due to the substantial domain gap between natural images and X-ray[7] images. Firstly, X-ray images often inherently contain inevitable noise resulting from the imaging process, which tends to cause local edge contours in X-ray images to appear more blurred compared to natural images. This situation can lead to inaccurate feature extraction from X-ray images and generate false

[*] This work has been supported by the National Key R&D Program of China (No.2018YFB1307802), and the Tianjin Science and Technology Plan Project (No.18PTLCSY00070).

[†] These authors contributed equally to this work and should be considered co-first authors.

[**] MA Xinlong is a professor at the Institute of Medical Engineering and Translational Medicine, Tianjin University. He received his master degree from Tianjin University in 2004. His research interests are mainly in orthopedics, digital orthopedics, orthopedic biomechanics and sports medicine. E-mail: maxinlong8686@sina.com

detections when applying traditional general-purpose object detection models. Additionally, X-ray images often have a significant amount of black background due to the requirement of filling the image's surroundings with a solid black color to accommodate device specifications. However, when utilizing traditional feature extractors, often results in the generation of numerous redundant feature maps that resemble the background. The generation of these redundant feature maps can significantly slow down the entire process.

To address the aforementioned challenges, we propose a novel framework called effective and efficient network (EAE-Net), specifically designed to enhance the effectiveness and efficiency of object detection in X-ray images. To leverage the global structural information inherent in X-ray images, we introduce a context augment module (CAM)[8]. This module employs global pixel attention to capture crucial context information across the entire image. By incorporating global context, our model effectively reduces false detections and improves overall detection accuracy. Moreover, to mitigate the impact of redundant features, we integrate a novel convolution layer known as ghost bottleneck module (GBM)[9] into the backbone of our model. This layer enables the generation of optimized feature maps through a more efficient linear feature transformation, ensuring fast inference without compromising accuracy. By combining these two innovative design elements, our EAE-Net model achieves exceptional detection performance, striking a balance between accuracy and speed. This advancement not only saves valuable time and effort for clinicians, but also empowers them to focus more on critical aspects of diagnosis and treatment.

Next, we begin with analyzing the unique characteristics of X-ray images. Subsequently, we present the framework of our proposed approach, named EAE-Net. Finally, we provide experimental results to validate our arguments.

As shown in Fig.1, we used a general-purpose detection model (YOLOV4)[10] to detect X-ray images and visualize the feature maps and heatmaps of the final layer of the model. We found two distinct characteristics in the feature maps of X-ray images: local region blur and background feature redundancy. Firstly, X-ray images often inherently contain inevitable noise resulting from the imaging process, which tends to cause local edge contours in X-ray images to appear more blurred compared to natural images. For example, in the blue region, we can find that the visual representation of the right ankle appears blurred, which leads to inaccurate feature extraction from X-ray images and generates false detections when applying traditional general-purpose object detection models. However, despite the blurring of local edge contours in X-ray images, the global structural information remains clear. This global structural information, which includes the human body skeleton prior, serves as valuable cues for effectively recognizing bone regions. Therefore, in this paper, we propose a novel global attention mechanism with CAM that leverages this information to extract better visual features.

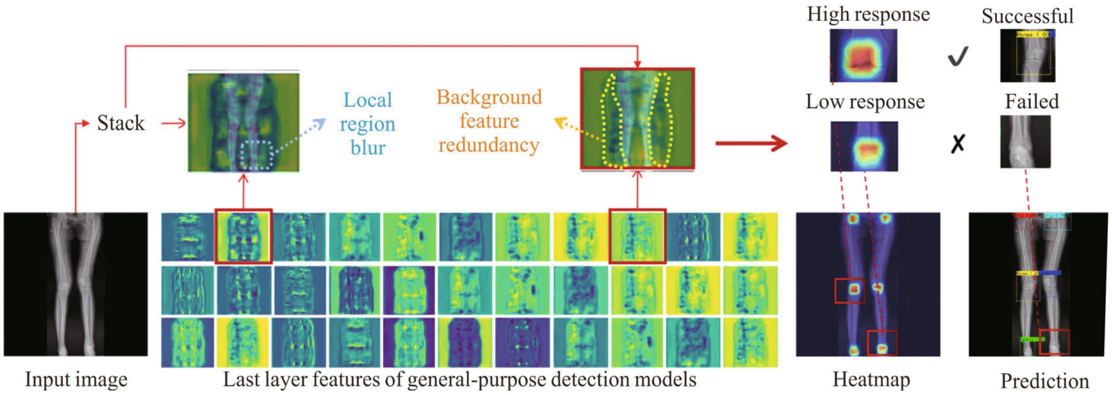

Fig.1 Distinct characteristics of X-ray image

Secondly, X-ray images often have a significant amount of black background due to the requirement of filling the image's surroundings with a solid black color to accommodate device specifications (1 024×1 024). For instance, in the yellow region, there is a high response from a significant number of non-body parts (background). To address this issue, we propose an efficient feature extraction approach GBM that employs a linear operation to accelerate the overall processing by altering the generation method of redundant features.

Based on the above two motivations, we design our proposed EAE-Net. As depicted in Fig.2, EAE-Net consists of three main components: efficient feature extraction module, effective feature augment module, and detection head. The efficient feature extraction module is responsible for extracting multi-scale features[11] from the image, which plays a crucial role in subsequent localization and recognition tasks. The effective feature augment module enhances the extracted features by fusing information from different scales and phases. The detection head module then adapts the outputs, converting the features into target box coordinates, confidence scores, and classification probabilities for each detected object in the frame.

Compared to the general-purpose object detection models, our proposed model incorporates additional special designs. Firstly, we introduce a lightweight module called GBM, which reduces redundant calculations and accelerates the inference speed. In the neck part, we propose a CAM to achieve more comprehensive features and improve detection accuracy. Next, we will introduce these two designs in more detail.

GBM is a lightweight module compared with previous designs. As shown in Fig.1, we observe that X-ray images have a significant amount of black background due to the requirement of filling the image's surroundings with a solid black color to accommodate device specifications. However, when utilizing traditional feature extractors, often results in the generation of numerous redundant feature maps that resemble the background. To address this issue, we propose an efficient feature extraction module, GBM, which employs a linear operation to accelerate the overall processing by altering the generation method of redundant features.

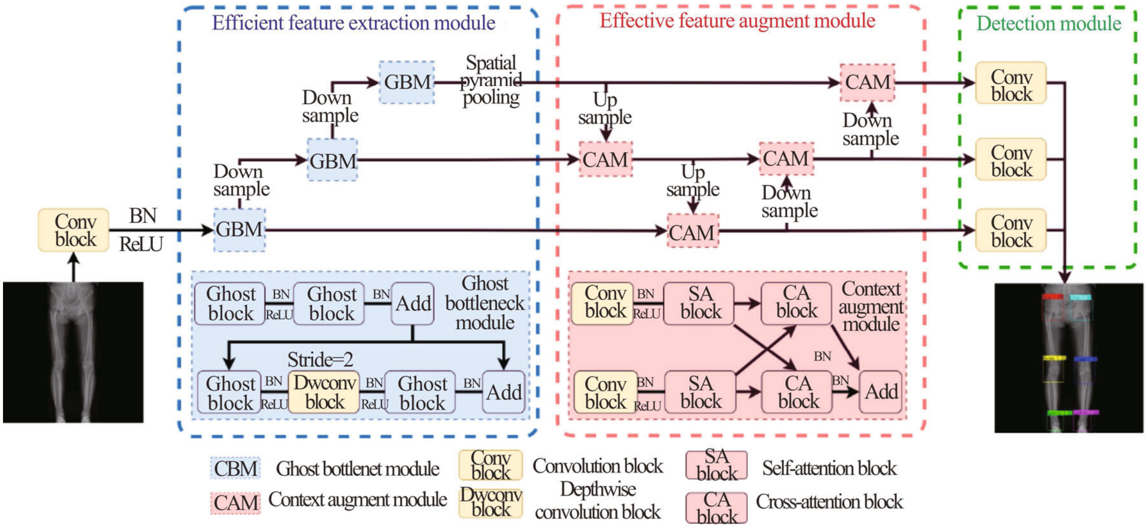

Fig.2 Overall framework of EAE-Net

As shown in Fig.3, the GBM exhibits a structure similar to the ResNet[12] bottleneck, with the ghost module serving as its core component. To show the effectiveness of the ghost module, we compare three different convolutions in Fig.3. Fig.3(a) is an ordinary convolution operation, and each box in the figure is the feature of a channel. The ordinary convolution operation performs convolution on all channels, which is slower and consumes more memory in the case of large channel dimensions or stacking multiple layers. Fig.3(b) shows the cross stage partial (CSP) network[12], which only performs complex convolution on partial channels, and then fuses the features of the original other channels and the encoded channels by transition operation to obtain the output. By reducing the number of connections and complexity, CSP-based networks can get better features significantly faster. Different from the above two types of convolutions, the ghost module uses a set of intrinsic feature maps, which are obtained by standard convolution, and then applies some transformations with cheap cost, such as depth-wise convolution and pointwise linear projection, to produce more feature maps that could fully reveal the information of the intrinsic features.

For a standard convolution operation, given input data $X \in R^{h \times w \times c}$, where h is the height of the input feature map, w is the width of the input feature map, and c is the number of input channels, the operation of any convolution layer generating n feature maps is shown as

$$Y = X * f + b, \tag{1}$$

where b is the deviation term, $f \in R^{c \times k \times k \times n}$ is the convolution kernel of the feature layer, and $Y \in R^{h' \times w' \times n}$ is the output feature graph with n output channels. In this convolution process, the amount of calculation *FLOP* is huge, and its operation is shown as

$$FLOP = c \times k \times k \times n \times h' \times w', \tag{2}$$

where h' and w' are the height and width of the output feature map respectively, and k is the size of the convolution kernel f.

The ghost module is a convolution operation on part of the feature graphs, that is, a standard convolution is used to complete m original output feature graphs $Y' \in R^{h' \times w' \times m}$, where $m \leq n$, the operation is shown as

$$Y' = X * f', \tag{3}$$

where $f' \in R^{c \times k \times k \times m}$ is the convolution kernel used by the feature layer and does not contain the deviation term. In order to further obtain the required n feature maps, a series of simple linear changes are carried out on the obtained m-dimensional feature maps to generate s similar feature maps:

$$y_{ij} = \phi_{i,j}(y'_i) \quad \forall i = 1, \ldots, m, j = 1, \ldots, s, \tag{4}$$

where y_i' is the ith original feature graph in Y', and $\phi_{i,j}$ is the jth linear calculation used to generate the jth similar feature graph y_{ij}. The calculation of *FLOP* is shown as

$$FLOP = c\times k\times k\times \frac{n}{s}\times h'\times w' + \frac{n}{s}\times d\times d\times (s-1)\times h'\times w', \qquad (5)$$

where d is the average kernel size per linear operation. From Eq.(5), it can be seen that the ghost module divides the calculation into two parts, one part is the ordinary convolution operation, the other part is the linear transformation operation, combined with Eq.(2), the compression ratio of the model is about s, which greatly reduces the number of parameters.

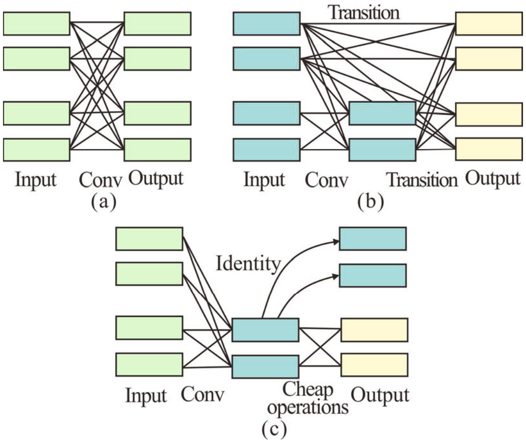

Fig.3 Three types of convolution: (a) Standard; (b) CSP; (c) Ghost

The CAM serves as a feature enhancement module to improve performance. It is important to note that there exists a domain gap between natural images and X-ray images, with X-ray images often exhibiting blurred local edge contours while retaining clear global structural information. Therefore, the objective of this paper is to utilize the global structural information as contextual cues to obtain more comprehensive features. To effectively leverage the correlation of global pixels, CAM employs a dual-attention mechanism to process features from different scales. This involves calculating the self-attention of individual features and subsequently calculating the cross-attention of correlated features. By adopting this approach, a more thorough exploration of global information can be achieved. In the following sections, we provide a detailed description of the structure of the self/cross attention block.

As shown in Fig.4, when given an input feature map X, we can calculate the global attention as the context to augment X to Y. In this way, global context is added to the features, which could avoid blurred local edge contours with clear global structural information.

$$Attention = \mathrm{softmax}(Q_x K_x^{\mathrm{T}}),$$
$$Y = X + Attention\cdot V_x. \qquad (6)$$

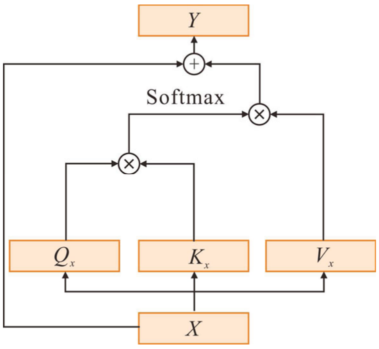

Fig.4 Self/cross attention block

After that, the head part in EAE-Net will utilize three scales of feature maps for bone detection, and get the final bounding box prediction through some post-processing operations.

To optimize the network, we consider the complete intersection over union (CIoU) loss function. Different from distance intersection over union (DIOU)[13] which considers directly minimizing the normalized distance between the predicted image and the actual image, the overlap area of two frames, distance from the center point and length and width are three important factors that should be considered in target regression loss. Therefore, CIOU loss is adopted as the loss function in this paper:

$$\mathrm{Loss}_{\mathrm{CIOU}} = 1-\mathrm{IOU}+\frac{\rho^2(b,b^{gt})}{C^2}+\alpha v, \qquad (7)$$

where α is the weight and v is used to measure the aspect ratio. An influence factor av operation is added to the calculation of CIOU loss, and the aspect ratio is considered, which makes the convergence speed and precision of CIOU loss higher.

The dataset for this paper was obtained from the Tianjin Institute of Orthopaedics, and the source data consists of a total of 515 full-length radiographic images of the human lower limb, each with a resolution of 1 024×1 024 and containing three channels of RGB. In this paper, the data set is augmented to 7 210 images using 11 data enhancement methods, including random pixel addition, vertical flip, maximum pooling, HSV transform and adaptive histogram equalization, which can effectively suppress overfitting during model training. After expanding the dataset using the above data enhancement methods, the dataset is divided into a training set and a test set in the ratio of 9: 1, with 5 840 images in the training set, 649 images in the validation set and 721 images in the test set. Our goal is to detect ankles, hips, knees from a single radiograph of a human lower limb. Fig.5 shows the unprocessed X-ray image and Fig.6 shows the expected detection results. Each image typically contains three pairs. We set these six areas that need to be detected (left ankle, right ankle, left knee, right knee, left hip, right hip) to the six categories. The experiments were conducted using PyTorch 1.8 deep learning framework and the code was run on Windows 10 operating system via Python and accelerated by CUDA 11.0. The hardware used for running the experiment is

Intel i7-10700KF for CPU and NVIDIA RTX3090 for GPU.

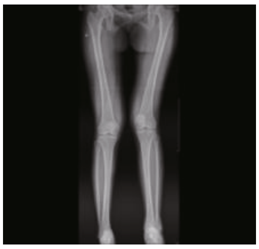

Fig.5 Original human lower limb X-ray image

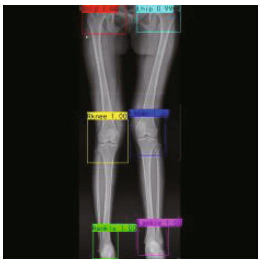

Fig.6 Expected X-ray image detection results

Our EAE-Net model has an input image size of 416×416, supervised by the CIOU loss, and the final results are evaluated using mean average precision (*mAP*). Tab.1, Fig.7 and Fig.8 show the comparison of our proposed model with the existing general purpose object detection baselines. It shows that our model has comparable or better performance in all metrics, and there is a significant improvement in processing speed. The *mAP* of our proposed EAE-Net model achieves an average improvement of 0.5 compared to the best results reported in previous approaches, which shows the effectiveness of our method. Besides, the model introduced in this paper improves the frames per second (*FPS*) by 55%, which enables real-time detection. The EAE-Net model showcases its improved performance while maintaining lightweight characteristics, making it suitable for the efficient automatic joint detection task. Moreover, according to the various metrics of each class, it can be found that the left knee and hip are more difficult to detect, and we analyzed that it may be because the structure of the place is more complex, nevertheless, ours still achieved satisfactory results.

Tab.1 Comparison between our improved model and the baseline

Metric	Model	Target areas					
		Left knee	Right knee	Left hip	Right hip	Left ankle	Right ankle
mAP	Faster-RCNN	90.83	98.64	99.86	90.41	99.10	98.97
	CenterNet[14]	93.20	96.91	99.69	93.39	98.12	98.79
	YOLOV3[5]	95.00	97.16	98.90	94.15	97.65	96.51
	YOLOX[15]	92.47	97.71	98.79	93.91	99.45	98.31
	YOLOV4	95.41	98.41	99.41	95.69	98.59	99.54
	EAE-Net (Ours)	96.53	99.15	100	96.32	99.65	99.67
FPS	Faster-RCNN	12	13	14	12	12	14
	CenterNet	15	16	15	17	15	16
	YOLOV3	18	17	18	16	17	16
	YOLOX	21	20	22	20	21	21
	YOLOV4	22	22	23	21	22	23
	EAE-Net (Ours)	33	33	34	35	35	33

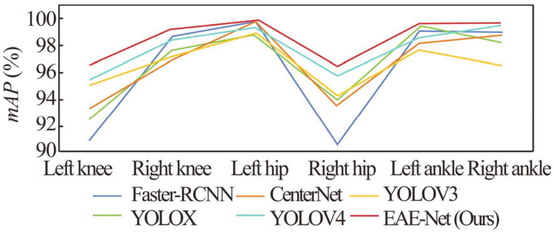

Fig.7 *mAP* of EAE-Net

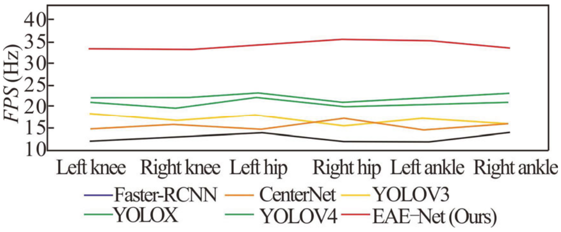

Fig.8 *FPS* of EAE-Net

We also conducted an ablation study on two key components of our framework, namely GBM and CAM, as shown in Tab.2. When GBM is excluded, the *FPS* is reduced by nearly 50%, while the *mAP* remains almost unchanged. This demonstrates the effectiveness of GBM in accelerating the inference process without compromising accuracy. Additionally, when CAM is not utilized, the *mAP* is decreased by approximately 0.7. This showcases the effectiveness of the proposed CAM in capturing global context and reducing false detections in X-ray images. By incorporating CAM into our model, we are able to leverage the global structural information present in the images, leading to improved accuracy in object detection tasks.

Qualitative analysis experiments the best previous approach (YOLOV4) and our proposed solution, EAE-Net is shown in Fig.9. We visualize the features of the final layer of both models. We can observe that compared to the blurry feature maps of YOLOV4, the features of EAE-Net are more focused on the body parts (as indicated by the yellow region), which demonstrates that the design of EAE-Net effectively eliminates the influence of redundant features. Additionally, in the feature maps of EAE-Net, there is a clear response in the local joint areas of the body (as shown by the blue region), while YOLOV4 only exhibits a few feature maps with this phenomenon. This also indicates that EAE-Net can utilize the CAM module to explore prior knowledge of the human body, thereby helping to clarify initially blurry regions in the local areas. Additionally, in Fig.10, we compare the heatmaps of different approaches. It can be observed that in the right ankle region, EAE-Net obtains a higher response compared to YOLOV4, accurately predicting the position of the joint. This further demonstrates the effectiveness of our proposed solution.

Tab.2 Ablation results of EAE-Net

CAM	GBM	*mAP* (%)	*FPS* (Hz)
√		98.15	22.17
	√	97.27	31.83
√	√	98.56	33.83

Furthermore, in order to demonstrate the practical effectiveness of the introduced model, we visualized the model's predictions between EAE-Net and more previous approaches in Fig.11. We performed the inference on an average performance PC and expanded the image size to 1 024×1 024 for more accurate detection. It was found that our model was able to accurately localize and classify all joints with higher confidence.

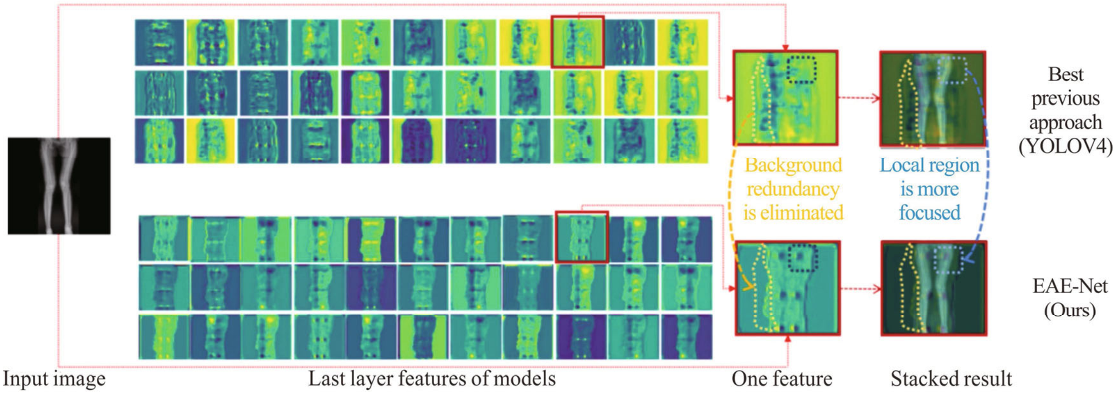

Fig.9 Comparison of features maps

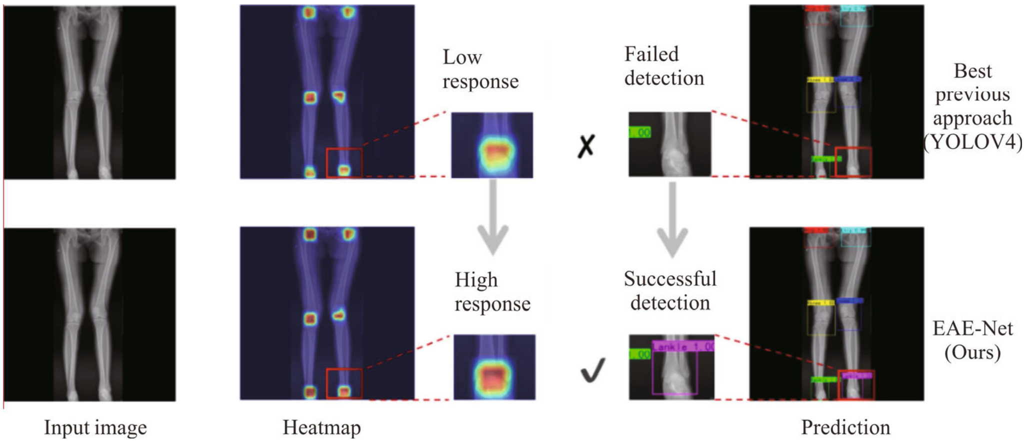

Fig.10 Comparison of heatmaps

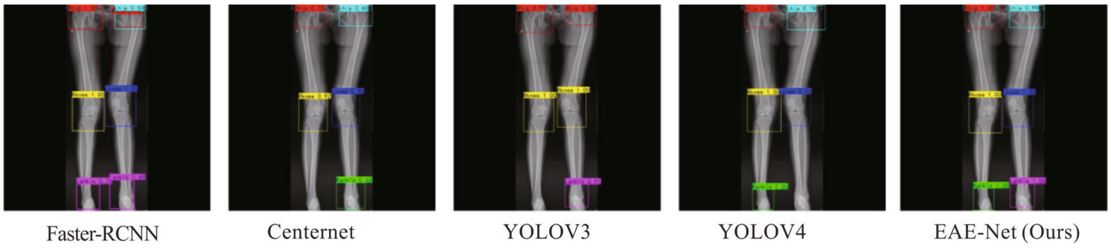

Fig.11 Comparison of detection results of different methods

In summary, this paper explores the detection and localization of bone joint regions in medical X-ray images. Traditional methods rely on subjective interpretation, leading to variability and errors. Automated bone joint detection techniques have become more feasible with advancements in general-purpose object detection. However, applying these algorithms to X-ray images faces challenges due to the distinct characteristics of X-ray images and the domain gap with natural images. To address these challenges, a novel framework called EAE-Net is proposed. It incorporates a CAM to capture global structural information and a GBM to reduce redundant features. The EAE-Net model achieves exceptional detection performance, balancing accuracy and speed. This advancement saves time for clinicians and allows them to focus on critical aspects of diagnosis and treatment.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] LITJENS G J S, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis[J]. Medical image analysis, 2017, 42: 60-88.

[2] MIOTTO R, WANG F, WANG S, et al. Deep learning for healthcare: review, opportunities and challenges[J]. Briefings in bioinformatics, 2018, 19(6): 1236-1246.

[3] WANG C Y, LIAO H Y M, YEH I H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//Conference on Computer Vision and Pattern Recognition (CVPR), June 14-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 1571-1580.

[4] TAGHANAKI S A, ABHISHEK K, COHEN J P, et al. Deep semantic segmentation of natural and medical images: a review[J]. Artificial intelligence review, 2021, 54(1): 137-178.

[5] REN S, HE K, GIRSHICK R B, et al. Faster RCNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems 28, December 7-12, 2015, Montreal, Quebec, Canada. Ottawa: NIPS, 2015, 28: 91-99.

[6] REDMON J, FARHADI A. YOLOV3: an incremental improvement[EB/OL]. (2018-04-08) [2023-05-23]. https://arxiv.org/abs/1804.02767.

[7] HAN Y, CHEN C, TEWFIK A H, et al. Pneumonia detection on chest X-ray using radiomic features and contrastive learning[C]//International Symposium on Biomedical Imaging, April 13-16, 2021, Nice, France. 247-251.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//IEEE International Conference on Learning Representations, May 3-7, 2021, Austria. New York: IEEE, 2021.

[9] HAN K, WANG Y, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 1577-1586.

[10] LI J, XU Z, XU L. Vehicle and pedestrian detection method based on improved YOLOV4-tiny[J]. Optoelectronics letters, 2023, 19(10): 623-628.

[11] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[12] HE K M, ZHANG X Y. Deep residual learning for image recognition[C]//Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[13] JIANG P, ERGU D, LIU F, et al. A review of YOLO algorithm developments[C]//Proceedings of the 8th International Conference on Information Technology and Quantitative Management, July 9-11, 2021, Chengdu, China. Amsterdam: Elsevier, 2021, 199: 1066-1073.

[14] ZHOU X Y, WANG D Q, KRHENBÜHL P. Objects as points[EB/OL]. (2019-04-16) [2023-05-23]. https://arxiv. org/abs/1904.07850v1.

[15] ZHANG H, LU C, CHEN E. Obstacle detection: improved YOLOX-S based on swin transformer-tiny[J]. Optoelectronics letters, 2023, 19(11): 698-704.