An Optimized YOLO Model with Local Attention for Arthritis Lesion Detection on X-ray Images

Yaqi Liu
College of Computer Science
Sichuan University
Chengdu, China
yaqiliu@stu.scu.edu.cn

Li Yang
North Sichuan Medical College
Nanchong, China

Tingting Wang
Department of Rheumatology and Immunology
Dazhou Central Hospital
Dazhou, China

Jianhong Wu
Department of Rheumatology and Immunology
Dazhou Central Hospital
Dazhou, China

Tao He*
College of Computer Science
Sichuan University
Chengdu, China
tao_he@scu.edu.cn

Zhang Yi
College of Computer Science
Sichuan University
Chengdu, China
zhangyi@scu.edu.cn

***Abstract*—In the field of medical image processing and analysis, automated lesion detection in arthritis is crucial for clinical diagnostic assistance and presents a challenging task. To enhance the performance of automated detection, we integrated a Criss-Cross attention module into the YOLOv8 model in object detection within X-ray images of hand joints, which was aimed at better capturing pathological features between hand joints. We analyzed the relationship between the two distinct methods of utilizing attention, specifically extracting global attention weights and extracting local attention weights, and the features exhibited by arthritis, comparing the impact of different attention mechanisms on arthritis lesion detection within the object detection task. Specifically, we validated our approach on large-scale medical datasets for Osteoarthritis (OA) and Rheumatoid Arthritis (RA), respectively. Experimental results demonstrate that utilizing attention weights to extract attention within specific local regions in the YOLOv8 model yields superior results in arthritis lesion detection compared to using no attention mechanism or extracting global attention weights.**

***Index Terms*—YOLO, arthritis, object detection, attention mechanism**

I. Introduction

Arthritis is a term used to describe inflammatory diseases that occur in the joints and surrounding tissues of the human body, caused by inflammation, infection, degeneration, trauma, or other factors. Clinical manifestations include redness, swelling, heat, pain, functional impairment, and joint deformities, with severe cases leading to joint disability and impacting the patient's quality of life. Common types of arthritis in the hands mainly include osteoarthritis (OA) and rheumatoid arthritis (RA).

RA can occur at any age, with the most common age range being between 40 and 70 years old [1]. RA usually has a hidden onset and affects all synovial joints, causing pain, stiffness, and swelling, particularly in the metacarpophalangeal joints (MCP), proximal interphalangeal joints (PIP), and wrists [2]. On the other hand, OA primarily affects the first carpometacarpal joint (first CMC), distal interphalangeal joints (DIP), and PIP, leading to issues such as joint osteophytes, subchondral sclerosis, and asymmetric joint space narrowing (JSN) [3]. In clinical practice, Kellgren-Lawrence Score (KL) and Sharp-van der Heijde Score (SHS) are commonly used as grading criteria for bone destruction in OA and RA imaging, respectively. KL and SHS are two widely employed indicators for assessing the severity of arthritis. KL scoring is mainly applied to assess joint X-ray images of OA patients to quantitatively measure the degradation of joint cartilage and bone structure, while SHS is primarily utilized to evaluate joint X-ray images of RA patients to quantify the extent of joint structure damage.

Over time, the disability rate of patients with hand arthritis significantly increases. Therefore, early diagnosis and intervention are crucial in reducing hand arthritis-related disabilities. However, traditional methods for detecting arthritis inflammation often rely on manual assessment, requiring doctors to analyze image data based on their experience and expertise. This manual assessment process is time-consuming and subject to subjective factors, making it challenging to ensure the accuracy of results. Currently, automatic detection of arthritis inflammation has become a hot topic in joint injury research. Utilizing computer vision and machine learning techniques, automatic detection methods can more accurately identify OA and RA from image data of OA and RA patients.

In the context of medical imaging analysis of joints, there exist several challenges, including the presence of high levels of image noise, poor clarity of joint structures, and the complex morphological and textural changes induced by joint lesions. Addressing these obstacles is crucial for accurate diagnosis and treatment planning. With the application of attention mechanisms in the field of image processing, it has also been introduced into medical image research, aiming to improve the performance and effectiveness of models by

*Corresponding author: Tao He{email: tao_he@scu.edu.cn}.

focusing on the crucial information in the images. In this paper, we aim to emphasize the information relevant to bone destruction in OA and RA datasets, reducing the impact of irrelevant information on the object detection results.

We applied common attention mechanisms, such as Non-local [4] and GCNet [5], to our object detection task but found that they did not yield remarkable improvements as they did in other image tasks. Through further analysis, we identified that obtaining global pixel-wise attention weights in our medical image object detection task led to the extraction of partially redundant features, causing overfitting and negatively affecting the model's performance. Based on this observation, our goal is to extract attention information only for the pixels important for the object detection task to achieve better results.

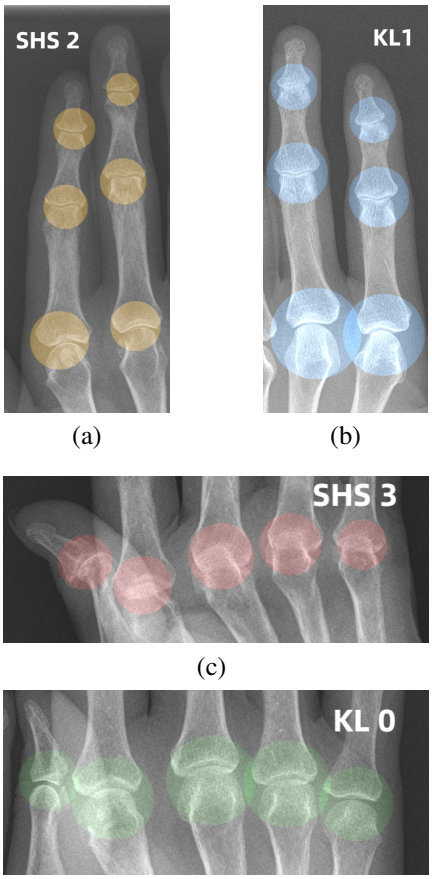


Fig. 1: Subfigures (a) and (b) present similar symptoms displayed in different joints of RA and OA patients on the same finger. Subfigures (c) and (d) exhibit similar symptoms observed in the corresponding joints of RA and OA patients on different fingers.

By analyzing the SHS and KL indicators, we have been able to uncover significant information within joint medical images. Specifically, for a particular joint, it exhibits a more pronounced connection with other joints on the same finger and joints located in the same position on different fingers. This correlation is depicted in Fig. 1, where subfigures (a) and (b) illustrate the relationship between SHS and KL scores on the same finger in single-hand X-ray images, and subfigures (c) and (d) demonstrate the correlation between SHS and KL scores on different fingers at the same position. To leverage this characteristic, we employ the Criss-Cross attention [6], which superimposes attention weights in a cross-shaped pattern around the current query position onto the original image. This enables the extraction of relevant information regarding the association between a specific joint and its counterparts on the same finger and different fingers, ultimately enhancing the model's performance.

The contributions of our work are as follows:

1) We examined attention mechanisms appropriate for classifying bone destruction on OA and RA datasets. By contrasting global pixel-wise mechanisms such as Non-local and GCNet with Criss-Cross Attention, we determined that the latter is more efficacious and subsequently enhanced its performance.
2) We have built upon the latest object detection model YOLOv8, integrating Criss-Cross Attention, which has resulted in substantial improvements in performance.
3) We evaluated the proposed model using private medical OA and RA datasets, which contributed to augmenting the dependability, and resilience of our model. It is noteworthy that the RA dataset constituted the largest-scale X-ray image dataset utilized in existing literature within the same field.

II. Related Work

A. Object Detection

Object detection is a crucial task in computer vision, aiming to precisely locate and recognize targets within images or videos. Traditional detection algorithms, based on manually extracted features, often yield subpar performance and struggle to meet the demands of modern applications. However, with the advent of deep learning and the prominence of Convolutional Neural Networks (CNNs), object detection has entered a new era of development.

CNN-based object detection methods can be broadly categorized into "anchor-based" and "anchor-free" approaches. The anchor-based methods can further be classified into "two-stage" and "one-stage" algorithms. Generally, two-stage algorithms, known for their high accuracy, involve two phases: generating region proposals from the image and then refining the final bounding boxes from these proposals. Notably, R. Girshick et al. (2014) [7] introduced the RCNN algorithm, which achieved significant results on the VOC-07 dataset. However, the redundant computation of overlapping box features led to slow detection speeds. To mitigate this issue, K. He et al. (2014) [8] proposed the SPPNet algorithm, significantly boosting inference speed, but with increased storage requirements. Subsequent to these advancements, various network structures emerged to address the challenges of two-stage algorithms. For instance, R. Girshick et al. (2015) [9]

presented the Fast RCNN, building on RCNN and SPPNet, and S. Ren et al. (2017) [10] introduced Faster RCNN, offering higher accuracy and faster processing.

On the other hand, anchor-based one-stage detection algorithms eliminate the need for a region proposal stage, achieving faster detection speeds by directly generating class probabilities and object position coordinates. One of the pioneering works in this category was the YOLOv1 model proposed by J. Redmon et al. (2015) [11], treating the entire image as network input and directly regressing boundary box positions and class predictions in the output layer, resulting in remarkable speed improvements. However, it had limitations in detecting small objects. Subsequently, J. Redmon et al. (2016) [12] introduced YOLOv2, which significantly improved accuracy, speed, and category coverage, becoming a better and more powerful solution. Over time, with the iteration and refinement of the YOLO model, several high-performance anchor-based single-stage detection models emerged. Nevertheless, these methods still face challenges due to the impact of anchor size, number, and aspect ratio on detection performance, as well as memory and computational time constraints.

B. Attention Mechanisms in Computer Vision

In order to mimic the human attention process of focusing on relevant and interesting information when dealing with data, the attention mechanism has been widely applied in the field of artificial intelligence. In computer vision tasks, the attention mechanism helps models better process image or video data by concentrating on regions relevant to the task at hand.

In the field of channel attention, the channel attention module adaptively obtains the weights of each layer of channels. The SENet (2018) proposed by J. Hu et al. [13] pioneered channel attention, with its core being a squeeze-and-excitation (SE) block used to collect global information, capture channel relationships, and improve representation ability. The queue module utilizes global average pooling to obtain global spatial information, while the exception module captures the relationships between different channels by using fully connected layers and non-linear layers, and outputs the obtained attention vectors. Afterwards, a series of improved models for SENet emerged, such as Z. Qin et al. (2020) [14] proposed FcaNet to improve the squeeze module, Q. Wang et al. (2020) [15] proposed the use of ECANet to improve the excitation module, while H. Lee et al. (2019) [16] proposed an SRM that simultaneously improves both the squeeze module and the excitation module.

Considering the different importance of the spatial information in the image, the spatial attention module is generated, which obtains the spatial weight matrix through training to represent the information of the key degree of the region in the image. The non-local mechanism based on self-attention was first introduced by X. Wang et al. in their seminal paper "Non-local Neural Networks" (2018) [4]. It addresses the limitation of traditional CNNs that primarily focus on local spatial contexts and allows models to capture long-range dependencies and global context.

Criss-Cross Attention was introduced by Z. Huang et al. (2018) [6]. They noticed that in traditional self-attention mechanisms, such as the non-local mechanism mentioned earlier, attention is calculated by comparing query features with all spatial positions in the image. However, this method may be computationally expensive, especially for high-resolution images. Criss-Cross Attention addresses this limitation by adopting a more effective attention solution that utilizes the criss-cross attention module to capture the horizontal and vertical relationships between spatial positions in the image, and indirectly obtains global information by using it twice, enabling the model to more effectively utilize global contextual information.

C. Automated Diagnosis of Arthritis Lesion Detection

In recent studies, H. J. Wang et al. (2022) [17] introduced a novel approach to address two vital tasks: joint space narrowing detection and classification. The outcomes on 135 hand images of RA patients obtained through their method exhibited remarkable performance, highlighting the potential of automated techniques in enhancing clinical diagnosis. K. Miyama et al. (2022) [18] proposed a two-stage automated assessment system for bone destruction, involving detection and classification. Both the staged models and binary classification methods displayed promising performance on 226 hand X-ray images. N. Chaturvedi et al. (2021) [19] incorporated an attention mechanism into the diagnosis of RA. They employed a CNN with an attention mechanism to predict scores for bone erosion. The extraction and overlay of global attention weights onto joint images contributed to improved model performance.

However, in the domain of automated arthritis detection, addressing the challenge of reducing the inefficiencies and additional computational costs arising from two separate models, while simultaneously maintaining the accuracy of object detection, has remained an unresolved issue. To overcome these challenges and establish dependable automated detection methodologies that advance clinical diagnosis, further research is essential.

III. METHODS

A. Criss-Cross Attention

Huang et al. (2018) [6] introduced the Criss-Cross Network (CCNet), which builds upon the Non-local approach by replacing the global attention mechanism with a criss-cross attention mechanism. This novel modification allows for a more efficient and effective way of capturing essential information.

As shown in Fig. 2, the Criss-Cross Attention Block computes the correlations with positions that are in the same row or column as the current position to assign varying weights to different parts of the input. Subsequently, the weighted information is added to the original image for output.

The main structure of the Crisis-Cross Attention module is shown in the Fig. 3. The initial input image undergoes three 1×1 convolutional layers to generate features Q,

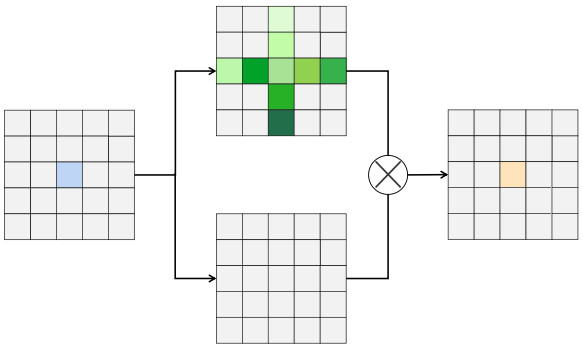


Fig. 2: Criss-Cross attention block.

K, and V, respectively. The dimensions of Q and K are $(C^{'} \times H \times W)$, reducing the original channel number C to $C^{'}$ and reducing computational complexity; The size of V remains $(C \times H \times W)$. After obtaining the feature map, Q and K proceed with Affinity operation to further generate the attention map: Affinity operations multiply the feature vector at a certain position in the feature map Q by the feature vector at a certain position u in the same row or column to obtain a new feature map D with a size of $(H + W - 1) \times H \times W$, thereby obtaining the relationship between each pixel and its corresponding row and column pixels.

Next, perform Softmax operations on the new feature map D to normalize the positional relationship exhibited by D. Finally, the correlation between positions is weighted through the aggregation operation to generate the final feature map.

At each position u in the spatial dimension of feature map V_u $\left(V_u \in R^C\right)$, a vector V and a set Φ_u can be obtained. Φ_u is the set of feature vectors in V, which are in the same row or column as position u. Context information is collected by the Aggregation operation where $H_u^{'}$ is the feature vector of the output feature map $H^{'}$ at position u:

$$H_u^{'} = \sum_{i \in |\Phi_u|} A_{i,u} \Phi_{i,u} + H_u \tag{1}$$

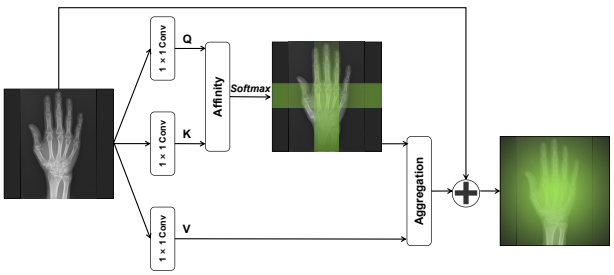
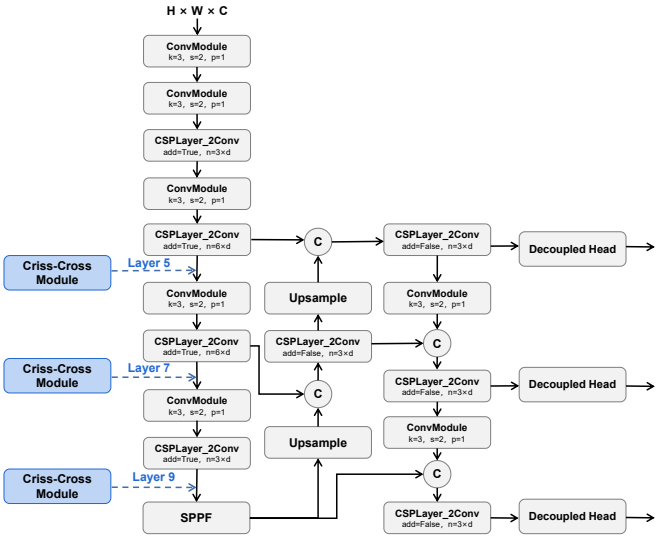


Fig. 3: Criss-Cross attention module.

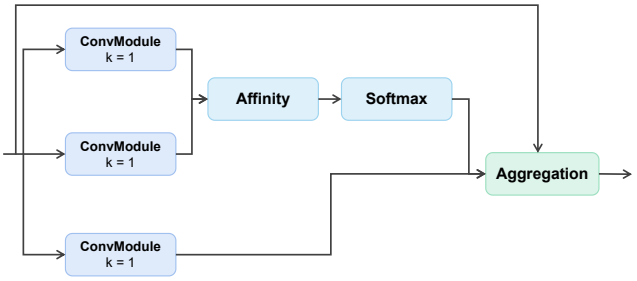
B. Improved YOLOv8 Model with Criss-Cross Attention

The YOLOv8 model draws upon the relevant designs of algorithms such as YOLOX [20], YOLOv6 [21], YOLOv7 [22], and PP-YOLOE [23]. It has transitioned from the original coupled heads to a decoupled head design, and shifted from YOLOv5 [24]'s "anchor-based" approach to an "anchor-free" approach, resulting in substantial improvements in model performance. Based on this state-of-the-art object detection model, we have enhanced its performance by incorporating an improved Criss-Cross Attention mechanism. This addition aims to boost the model's capabilities in object detection. Our main attempts and improvements are as follows.

In order to better explore the ability of attention mechanism to extract crucial information and adapt it to our automatic detection task, we conducted various experiments with the Criss-Cross Attention. Specifically, we focused on the step where we use convolutional operations in the Criss-Cross Attention Module to reduce the dimensionality and obtain feature maps with a new number of channels. During this step, we investigated the impact of the reduced dimensionality on the results. Specifically, we experimented with reducing the number of channels to one-half, one-fourth, and one-eighth of the original number of channels. By comparing the performance with different reduced channel numbers, we aimed to determine the most effective setting for our task.



(a) Model Structure



(b) Criss-Cross Module

Fig. 4: Subfigure (a) depicts the model's overall structure, and subfigure (b) elaborates on the Criss-Cross Module within it.

At the same time, in order to explore the effect of the atten-

tion mechanism on feature extraction at different levels of the input images and to further enhance crucial information while reducing noise or irrelevant details, we experimented with incorporating the improved Criss-Cross Attention layers at various levels of YOLOv8. Through these experiments, we aimed to find the most suitable combination of attention mechanisms for the automatic detection of arthritis inflammation, ultimately improving the model's robustness and generalization capabilities. Our model architecture is illustrated in Fig. 4, with subfigure A representing the holistic structure of the enhanced YOLOv8 model, and subfigure B showcasing the Criss-Cross module configuration. Specifically, we have incorporated the Criss-Criss Attention Module, which has shown promising results after dimensionality reduction, into the backbone of the YOLOv8 model. We conducted experiments to assess the impact of placing this module at the fifth, seventh, and ninth layers of the YOLOv8 model on its performance.

Finally, we conducted experimental evaluations on the datasets from patients diagnosed with Osteoarthritis (OA) and Rheumatoid Arthritis (RA) respectively. Based on metrics such as Precision, Recall, mAP50, and F1 score, we conducted a comprehensive performance comparison of the enhanced model on both datasets.

IV. Experiments

A. Image Preprocessing

1) Datasets: The dataset is divided into two parts, each containing hand joint images of patients diagnosed with Osteoarthritis (OA) and Rheumatoid Arthritis (RA) provided by Dazhou Central Hospital in Sichuan, China. Ethical approval was obtained by the ethics committee of the Dazhou Central Hospital. The personal privacy of subjects is protected because only the X-ray images of subjects are supported. Non-representative training images, such as those with finger misalignment, overlapping joints, and severe joint distortion, were excluded to reduce outliers' interference during model training.

There are two datasets available: the OA dataset and the RA dataset. The OA dataset comprises 113 patients aged over 20 years, encompassing a total of 216 hand X-ray images in PNG format. Each of these images' 21 joints has been annotated by three rheumatologists following the KL diagnostic criteria, used for bone destruction evaluation, resulting in KL scores. On the other hand, the RA dataset includes 528 patients aged 20 years and above, with a collection of 960 hand X-ray images in PNG format. Similarly, each of these hand images' 21 joints has been annotated by three rheumatologists using the SHS scoring system to assess bone destruction.

In both the OA and RA datasets, each image's 21 joints have been individually annotated by two rheumatologists. The annotations were then compared, and in cases of discordant results, a rheumatology specialist provided a revised annotation. Ultimately, the final annotations were determined collaboratively by the input of all three medical experts.

We randomly divided these two datasets into training, validation, and test sets in an 8:1:1 ratio, and the distribution of each dataset is presented in Table I.

Based on the severity scores of bone destruction in RA and OA images, we further categorized the corresponding joint regions with SHS and KL scores of 0, 1-2, and 3-4 into three classes: Healthy (s0), Mild (s2), and Severe (s3), following the classification approach inspired by the work of H. J. Wang et al.(2022) [17], which proves to be convenient for clear delineation. The statistical distribution of the joint categories is presented in Table II.

Table I: Distribution of each dataset.

	Train	Valid	Test	Total
OA	172	22	22	216
RA	768	96	96	960

Table II: Distribution of each joint category in datasets.

	Healthy(s0)	Mild(s2)	Severe(s3)	Total
OA	2174	1929	433	4536
RA	11906	5437	2817	20160

2) Data Augmentation: Data augmentation is a technique that involves applying random transformations or expansions to the original data during the training process, thereby enhancing the model's generalization ability. We applied data augmentation to the training data of both datasets. Specifically, we used horizontal flip and Mixup methods [25]. For the horizontal flip part, we achieved increased data diversity by horizontally mirroring the training images.

The mixup data augmentation algorithm is a method that enhances the training data by randomly selecting two samples' vectors and their corresponding labels. By utilizing linear interpolation, this technique generates new vectors and corresponding labels. This process effectively improves the model's generalization ability. The underlying formula can be expressed as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda)\, x_j \tag{2}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)\, y_j \tag{3}$$

Here, x_i, x_j are raw input vectors and y_i, y_j are one-hot label encodings. We set λ value of 0.5 as the parameter of the mixup algorithm and achieved good results.

B. Performance Evaluation Metrics

In evaluating the performance of the model, we mainly used metrics such as precision, recall, map50, accuracy, and F1 score. Their calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

Here, TP represents a positive sample predicted by the model as a positive class; TN represents negative samples predicted by the model as negative classes; FP represents negative samples predicted by the model as positive and FN represents positive samples predicted by the model as negative classes.

For each class, AP is calculated by computing the area under the Precision-Recall curve, with precision as the y-axis and recall as the x-axis. A higher AP indicates better model performance for that specific class. For all classes, mAP is obtained by computing the average of AP values across all classes, providing a comprehensive evaluation of the model's object detection performance. The formula for calculating mAP is as follows:

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{7}$$

Here, k denotes the number of categories, and AP_i represents the AP value for the specific object category i, where i is the index of the currently detected category among the total of k categories.

C. Results

On the OA and RA datasets, we compared the object detection results obtained from the YOLOv8 model with those from the improved YOLOv8 model, which incorporated Non-local, GCNet, and Criss-Cross Attention (taking the ninth layer as an example). On the RA dataset, the loss reduction curve during YOLOv8 model training is illustrated in Fig. 5. The curve demonstrates a smooth descent without any unusual fluctuations.

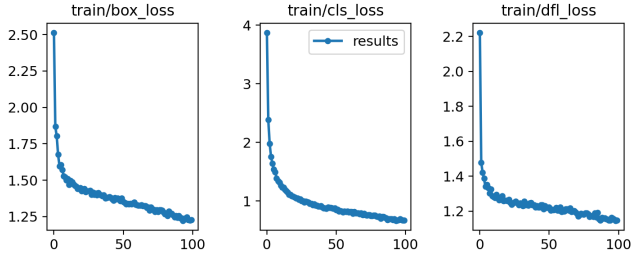


Fig. 5: The convergence process of the loss during training the YOLOv8 model on the RA dataset.

The comparison results are shown in the following table, with Table III presenting the outcomes on the OA dataset and Table IV showcasing the results on the RA dataset.

Through experiments, we found that the results align with our expectations. The addition of attention mechanisms to the YOLOv8 model improves the object detection performance for both RA and OA datasets. The GCNet, which optimizes global context modeling capability, performs better than Non-local, exhibiting faster detection speed and improved performance. However, compared to the Criss-Cross Attention that captures attention weights for local pixels, the approach of obtaining attention weights for global pixels lags slightly in overall performance.

Table III: Improved YOLOv8 model on OA test dataset.

	Precision	Recall	map50	F1 score
None	0.807	0.813	0.878	0.810
With Non-local	0.819	0.788	0.887	0.803
With GCNet	0.869	0.783	0.889	0.824
With CCNet	0.893	0.819	0.897	0.854

Table IV: Improved YOLOv8 model on RA test dataset.

	Precision	Recall	map50	F1 score
None	0.720	0.806	0.804	0.761
With Non-local	0.732	0.804	0.809	0.766
With GCNet	0.728	0.798	0.812	0.761
With CCNet	0.747	0.798	0.816	0.772

Next, in order to ascertain the impact of the varying degree of feature extraction due to the different levels of convolutional dimension reduction when calculating the Q and K feature maps within the Criss-Cross module on model performance, we conducted performance evaluations of the YOLOv8 model with the Criss-Cross Attention module, incorporating different dimensionality reduction channel numbers, on OA and RA datasets. Specifically, we embedded the Criss-Cross Attention module into the YOLOv8 model as the ninth layer. The results are presented in the two tables below, with Table V presenting the outcomes on the OA dataset and Table VI showcasing the results on the RA dataset.

Table V: With different CCNets on OA test dataset.

	Precision	Recall	map50	F1 score
$C' = \frac{1}{8}C$	0.893	0.819	0.897	0.854
$C' = \frac{1}{4}C$	0.796	0.879	0.900	0.835
$C' = \frac{1}{2}C$	0.864	0.779	0.890	0.819

Table VI: With different CCNets on RA test dataset.

	Precision	Recall	map50	F1 score
$C' = \frac{1}{8}C$	0.747	0.798	0.816	0.772
$C' = \frac{1}{4}C$	0.765	0.794	0.819	0.779
$C' = \frac{1}{2}C$	0.756	0.794	0.815	0.775

Here, C represents the number of channels in input to the Criss-Cross module, while $C^{'}$ signifies the size of the channel dimension after the dimension reduction of feature maps Q and K.

The results from Table V and Table VI reveal that the best outcomes are achieved when the channel dimension after reduction is set to an intermediate value. We continue to use this parameter for the subsequent experiments.

To delve deeper into the impact of extracting down-sampled features of varying resolutions within the backbone of the YOLOv8 model on its performance, we integrated the Criss-Cross Attention with improved dimensionality parameters into the YOLOv8 model, placing it at the fifth, seventh, and ninth layers. We then evaluated the performance of the enhanced model separately on OA and RA datasets. The results of these evaluations are presented in the two tables below with Table VII presenting the outcomes on the OA dataset and Table VIII showcasing the results on the RA dataset, showcasing the improved performance.

Table VII: YOLOv8 model with CCNet on OA test dataset.

	Precision	Recall	map50	F1 score
Without CCNet	0.870	0.774	0.883	0.819
CCNet in layer 5	0.819	0.788	0.887	0.803
CCNet in layer 7	0.867	0.859	0.917	0.863
CCNet in layer 9	0.796	0.879	0.900	0.835

Table VIII: YOLOv8 model with CCNet on RA test dataset.

	Precision	Recall	map50	F1 score
Without CCNet	0.720	0.806	0.804	0.761
CCNet in layer 5	0.761	0.783	0.820	0.772
CCNet in layer 7	0.749	0.782	0.816	0.764
CCNet in layer 9	0.765	0.794	0.819	0.779

The results on the OA dataset yield PR curves as illustrated in Fig. 6. Impressive performance is observed across all categories.

We tested the obtained model on the test sets, and the results are illustrated in Fig. 7. Subfigure (a) shows expert-annotated images, while sufigure (b) displays the output images predicted by the model. It's evident that the model demonstrates high overall prediction accuracy, although further improvements are needed in distinguishing between the healthy and mild categories.

D. Conclusion

Detection of joint arthritis lesions is a medically significant and challenging task in the realm of medical image analysis.

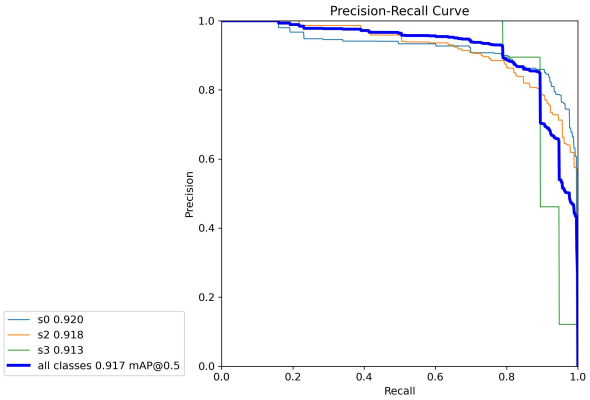
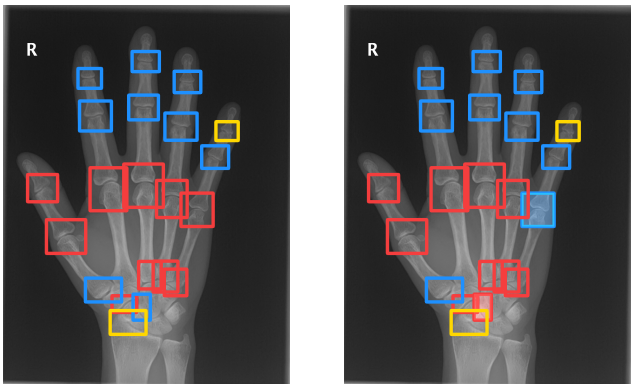


Fig. 6: The PR curve obtained by testing the OA dataset on the YOLOv8 model.



(a) (b)

Fig. 7: Subfigure (a) displays annotated images from the OA test set, and subfigure (b) displays the output images predicted by the model.

To address this challenge, our efforts are outlined below. Firstly, we performed joint localization and bone degradation assessment on two distinct datasets, OA and RA, using the latest object detection model, YOLOv8. We introduced attention mechanisms to enhance the model's performance, yielding improved results. Secondly, we compared the effects of various attention mechanisms and experimentally demonstrated that an attention mechanism focusing on the same position of different fingers and different positions of the same finger yields the most pronounced effects in our context. Subsequently, through further experimentation, we achieved the best performance using this attention mechanism. Lastly, we leveraged the largest dataset available in existing literature, enhancing the generalization ability of our model.

In the future, we aim to delve deeper into capturing essential information from hand joint inflammation by employing alternative attention mechanisms or relevant encoding approaches. By incorporating these insights into the target detection task, we anticipate achieving enhanced performance in joint lesion detection.

ACKNOWLEDGMENT

This work was supported by the National Major Science and Technology Projects of China under Grant 2018AAA0100201, the National Natural Science Foundation of China under Grant 62206189, the China Postdoctoral Science Foundation under Grant 2023M732427, the Basic research funds for central universities under Grant 2023SCU12091 and 2022 Dazhou City School Cooperation Special project under Grant 20226205.

REFERENCES

[1] Mate Gitanjali Subhash and AK Kureshi. An efficient cnn for hand x-ray classification of rheumatoid arthritis. *Microprocessors and Microsystems*, page 104822, 2023.

[2] Maria Kourilovitch, Claudio Galarza-Maldonado, and Esteban Ortiz-Prado. Diagnosis and classification of rheumatoid arthritis. *Journal of autoimmunity*, 48:26–30, 2014.

[3] KD Allen, LM Thoma, and YM Golightly. Epidemiology of osteoarthritis. *Osteoarthritis and cartilage*, 30(2):184–195, 2022.

[4] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[6] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[12] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021.

[15] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[16] HyunJae Lee, Hyo-Eun Kim, and Hyeonseob Nam. Srm: A style-based recalibration module for convolutional neural networks. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 1854–1862, 2019.

[17] Hao-Jan Wang, Chi-Ping Su, Chien-Chih Lai, Wun-Rong Chen, Chi Chen, Liang-Ying Ho, Woei-Chyn Chu, and Chung-Yueh Lien. Deep learning-based computer-aided diagnosis of rheumatoid arthritis with hand x-ray images conforming to modified total sharp/van der heijde score. *Biomedicines*, 10(6):1355, 2022.

[18] Kazuki Miyama, Ryoma Bise, Satoshi Ikemura, Kazuhiro Kai, Masaya Kanahori, Shinkichi Arisumi, Taisuke Uchida, Yasuharu Nakashima, and Seiichi Uchida. Deep learning-based automatic-bone-destruction-evaluation system using contextual information from other joints. *Arthritis Research & Therapy*, 24(1):227, 2022.

[19] Neelambuj Chaturvedi. Deepra: predicting joint damage from radiographs using cnn with attention. *arXiv preprint arXiv:2102.06982*, 2021.

[20] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[21] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

[22] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.

[23] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, et al. Pp-yoloe: An evolved version of yolo. *arXiv preprint arXiv:2203.16250*, 2022.

[24] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022.

[25] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.