Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CRPA: You only look once with Cross-directional Regional Pooling Attention in mandibular fracture CT

Zhou Tao
North Minzu University
Shiyi Yu
20247520@stu.nmu.edu.cn

North Minzu University
Kaixiong Chen
North Minzu University
Huiling Lu
Ningxia Medical University
Ke Song
North Minzu University

Article

Keywords: Object detection, YOLO, Pooling, weighting binary relation, Cross-location Region-pooling Attention

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11121157/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CRPA: You only look once with Cross-directional Regional Pooling Attention in mandibular fracture CT

Zhou Tao [a,c,1], Shiyi Yu [a,c,*,2], Kaixiong Chen [a,3], Huiling Lu [b,c,4] and Ke Song [a,c,5]

[a]School of Computer Science and Engineering, North Minzu University, Yinchuan, 750021, China.

[b]School of Medical Information Engineering, Ningxia Medical University, Yinchuan, 750004, China.

[c]Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan, 750021, China.

Abstract: Mandibular fracture detection is very important in clinical diagnosis. However, the CT images of mandibular fractures have the characteristics of complex fracture morphology, uneven location distribution and small fracture line, which makes it difficult for the YOLO model to fully model the structural characteristics and spatial dependence of the fracture area. In view of the above problems, this paper proposes an improved YOLO mandibular fracture detection model based on Pooling Attention Binary Relation (PABR). Firstly, a pooling attention binary relation representation framework is constructed, and the pooling response between different regions is uniformly represented as a weighted binary relation, and its properties are discussed. Secondly, based on this framework, a Cross-directional Regional Pooling Attention (CRPA) module is designed. By decoupling the horizontal and vertical regional pooling responses and coupling their cross-directional relationships, the proposed module enhances spatial dependencies while preserving positional information, thereby improving the representation capability of the model for complex fracture structures. Finally, CRPA is embedded in the YOLO network, and compared with the classical pooling attention module and the representative improvement methods in recent years. The experimental results show that the mAP50, mAP50-95, Recall, Precision and F1 values of the proposed method on the self-built mandibular fracture CT dataset reach 95.90 %, 45.80 %, 90.80 %, 94.80 % and 92.73 %, respectively. All indicators are better than the baseline YOLO model and comparison method. This study is of great significance for the auxiliary diagnosis of mandibular fractures.

Keywords: Object detection; YOLO; Pooling; weighting binary relation; Cross-location Region-pooling Attention;

1. Introduction

The mandible is one of the largest facial bones and one of the most frequently injured bones in the maxillofacial region [1]. Anatomically weak structures, such as the mandibular angle, symphysis, mental foramen region, and condyle, together with the prominent position of the mandible, its mobility, and relatively limited bony support, make it highly susceptible to fractures under external impact. Following a fracture, patients often present with limited mouth opening and malocclusion, which may further lead to abnormalities in facial anatomy and function. Therefore, CT and X-ray imaging are commonly employed to visualize fracture regions for clinical diagnosis. In recent years,

YOLO-based models have demonstrated promising performance in fracture detection tasks. Zhou et al. [2] propose Mandible-YOLO for mandibular fracture detection, which enhances the representation of fracture features through multi-scale convolutions, a spatial-channel fusion strategy, and global-local perception modules. Chen et al. [3] are the first to apply YOLOv9 to fracture detection and employed its Programmable Gradient Information (PGI) mechanism to alleviate information loss in fracture detection from X-ray images. Meza et al. [4] integrate a Spatial-Channel Synergistic Attention mechanism into YOLOv8 for hand bone fracture detection, and the experimental results demonstrated that spatial-channel attention is an effective strategy for enhancing fracture detection performance. Zhou et al. [5] propose the M3YOLOv5 model based on YOLOv5, which incorporates a global-local feature extraction module, a shallow-deep feature interaction module, and a multi-scale feature extraction module, thereby effectively improving the perception and representation of mandibular fracture features. Yari et al. [6] evaluate the performance of YOLOv5 for mandibular fracture detection in panoramic radiographs, and the experimental results demonstrated its promising capability in detecting most types of mandibular fractures. However, since fracture regions usually appear as elongated crack-like structures with distinct directional characteristics, existing methods, although effective in enhancing feature representation, still lack sufficient modeling of directional dependencies and spatial positional information. As a result, they are unable to adequately characterize the continuity and structural properties of fracture lines.

To enhance feature representation, pooling-based attention mechanisms have been widely incorporated into YOLO models. Methods such as SE [7], ECA [8], and CBAM [9] exploit global average pooling or max pooling to capture global contextual information and improve feature representation through channel recalibration. Subsequently, Coordinate Attention (CA) [10] decomposes two-dimensional global pooling into two one-dimensional directional pooling operations, thereby introducing positional information to a certain extent. However, these methods still fundamentally rely on single-directional or global aggregation operations, making it difficult to effectively characterize the structural relationships among different spatial regions. For mandibular fractures, which typically exhibit distinct directional characteristics and long-range spatial dependencies, relying solely on single-region pooling is insufficient to fully capture the continuity and spatial distribution patterns of fracture lines.

Although existing pooling attention mechanisms are capable of acquiring global or local contextual information through pooling operations, a unified and explicit theoretical formulation of the relationships among different spatial regions remains unavailable. Analysis reveals that, despite variations in their architectural designs, most pooling attention mechanisms follow a common "pooling–encoding–weighting" paradigm. In this paradigm, regional features are first extracted by pooling operations, inter-feature relationships are subsequently established through encoding modules, and feature representations are finally recalibrated by learned weights. Therefore, pooling attention mechanisms can essentially be interpreted as a form of regional relationship modeling. inary relations are important mathematical tools for describing mapping relationships between two sets and provide a unified way to characterize association patterns among objects. Inspired by this concept, a Pooling Attention Binary Relation (PABR) representation framework is proposed, in which the pooling, encoding, and weighting processes in pooling attention mechanisms are uniformly formulated as weighted binary relations, and their structural characteristics are analyzed from the perspective of relation properties. Based on this framework, a Cross-directional Regional Pooling Attention (CRPA) module is further developed. By jointly modeling the relationships among regions in the horizontal and vertical directions, the representation capability of the model for fracture boundaries, disrupted structures, and spatial positional information is enhanced.

The main contributions of this paper are as follows:

Pooling attention binary relation: From the perspective of binary relation, this paper generalizes the weighting operation to binary relation. On the basis of weighting binary relation, this paper proposes the pooling attention representation framework, pooling attention binary relation, and discusses some properties, such as reflexivity, symmetry and transitivity.

Cross-directional Regional Pooling Attention (CRPA): A Cross-directional Regional Pooling Attention (CRPA) module is proposed. The module establishes cross-directional regional correlations through horizontal-vertical pooling and employs horizontal and vertical pooling to obtain decoupled directional representations. The cross-directional features are then fed back to enhance the representations of both directions, achieving a unified modeling of directional feature decoupling and cross-directional relation coupling. By preserving positional information while strengthening spatial dependencies, the proposed module improves the representation capability of the model for complex fracture structures and subtle fracture lines.

The proposed method was implemented based on YOLOv8, and the effectiveness and generalization capability of CRPA were validated on both the mandibular fracture CT dataset and the LUNA16 dataset. Experimental results demonstrate that CRPA achieves the best overall performance among classical pooling-based attention mechanisms on the mandibular fracture dataset.

2. Related Work

2.1 Pooling

Pooling, also known as down-sampling, is used to reduce the dimension of feature maps during training, thereby decreasing the image size, reducing computation complexity, and extracting higher-level features. In this process, pooling cause information loss. Currently, many pooling operations work to reduce information loss. Bayraktar et al. [11] propose conditional pooling. It removes the noise according to the condition operator, and then performs the pooling operation. It is superior to traditional pooling in noise processing and fine-grained down-sampling. Mishr et al. [12] propose probabilistic pooling CNN(PP-CNN), which combines three different pooling architectures: MaxPooling, MinPooling, and MaxMinPooling. This innovative CNN classifier takes advantage of different pooling to capture a wider range of features and improves the accuracy of image classification by integrating positive and negative signals during activation. Pratik et al. [13], in order to reduce the information loss of traditional pooling, use two-dimensional discrete wavelet transform for downsampling. In addition, in order to retain the features of different pooling, MAC pooling is proposed, which is a three-branch pooling strategy. Zhao et al. [14] propose T-Max-Avg pooling in order to minimize the shortcomings of Max pooling and average pooling. In this method, a learnable threshold T is used to decide a specific pooling operation on the input data. Ozdemir et al. [15] propose a pooling method based on local binary pattern(LBP) in order to solve the problem of losing location information in traditional pooling methods. This method captures the correlation between pixels in a specific region around each pixel through LBP, so local spatial location information is preserved in the pooling. In this paper, Avg-TopK pooling is used in cross-location region pooling to obtain more discriminative features and avoid noisy features.

The down-sampling operation leads to information lose in pooling, so many studies reduce its use. However, pooling is used in the form of attention because it extracts key features. Pooling attention encodes the pooling results to obtain the weight value, and then uses this weight to weight

the feature map, so as to enhance the important features and suppress redundant features. This approach avoids information loss caused by the down-sampling operation. Zhang et al. [16] propose a new lightweight Channel Pooling Attention mechanism called Partial Channel Pooling Attention(PPA). PPA differs from existing convolutional attention blocks by allowing direct interaction between channels, which simplifies data representation, reduces redundancy, and lets channels share complementary information more efficiently. Ouyang et al. [17] improve on CA and propose Efficient Multi-Scale Attention Module. EMA module ensures effective spatial distribution of semantics by refactoring the channel dimension into the batch dimension and organizing it into multiple sub-features. Sun et al. [18] propose Successive Pooling Attention (SPAM). Through an innovative successive pooling mechanism combined with self-attention mechanism, SPAM effectively extracts deeper semantic features and fuses multi-scale features. In order to retain the location information in channel attention, Wang et al. [19] propose Vector Pooling Attention Module (VPA), which combined CA attention and ECA attention to realize the combination of spatial attention and channel attention. Li et al. [20] propose an attention-based pooling method (NFP). NFP uses deep convolution to encode the attention weights of neighboring features, which effectively guides local spatial pooling to learn discriminative features.

2.2 Binary Relation

Binary relation is a basic concept in mathematics to describe the relationship between two sets. Let A, B be two nonempty sets. A binary relation R from set A to set B is a subset of $A \times B$. For any ordered couple $< a, b >$ in the set $A \times B$, if $< a, b > \in R$, then a is related to b by R, denoted aRb.

Definition 1 Ordered Couple: a two-tuple of two elements a and b arranged in a certain order is called an ordered couple, denoted $< a, b >$, where a is its first element and b is its second element.

The ordered couple $< a, b >$ possesses the following properties.

(1) When $a \neq b$, $< a, b > \neq < b, a >$.

(2) Given $< a, b >$ and $< u, v >$,$(< a, b > = < u, v >) \Leftrightarrow (a = u \wedge b = v)$

Definition 2 Cartesian Product: Let A, B be two nonempty sets, with the element in A as the first element and the element in B as the second element to form an ordered couple. The set of all these ordered couples is called the Cartesian product of A and B, denoted $A \times B$. $A \times B = \{<$

$a, b > | a \in A \wedge b \in B\}$.

Definition 3 Binary Relation: Let A, B be two nonempty sets, and the binary relation R defined by any subset of $A \times B$ is called a binary relation from A to B. It is denoted as $R: A \rightarrow B$ or aRb. R possesses the following special relations:

(1) When $R = \emptyset$, R is called the empty relation from A to B.

(2) When $R = A \times B$, R is called the universal relation from A to B.

(3) When $R = I_A = \{< a, a > | a \in A\}$, R is called an identity relation on A

According to the definition of a binary relation, it has the following properties:

Suppose R be a relation on the nonempty set A,

(1) R is reflexive on set A if $\forall a(a \in A \rightarrow < a, a > \in R) = 1$.

(2) R is anti-reflexive on set A if $\forall a(a \in A \rightarrow < a, a > \notin R) = 1$.

(3) R is symmetric on set A if $\forall a \forall b\big((a \in A \wedge b \in A) \wedge < a, b > \in R \rightarrow < b, a > \in R\big) = 1$

(4) R is anti-symmetric on set A if $\forall a \forall b\big((a \in A \wedge b \in A) \wedge (< a, b > \in R \wedge < b, a > \in R) \rightarrow a = b\big) = 1$.

(5) R is transitive on the set A if $\forall a \forall b \forall c((a \in A \wedge b \in A \wedge b \in A) \wedge (< a, b > \in R \wedge < b, c > \in R) \rightarrow < a, c > \in R)$=1.

2.3 Attention mechanism for YOLO

YOLO are widely used in recent years due to its effective balance between computation complexity and performance. In addition, attention mechanisms are proven to effectively improve the performance of various visual tasks, especially in tasks such as medical image object detection. Therefore, it is an effective way to improve the performance of YOLO with attention.

Xiao et al. [21] propose FPM module in the feature fusion part of feature pyramid. This module uses channel and spatial attention to filter redundant features of feature fusion of different sizes, and then highlights small targets. Zhao et al. [22] propose an attention module based on Mamba. It uses Mamba to scan the image in different dimensions to obtain global context information, which improves the focus of the YOLO model on key regions and reduces the impact of redundant information. Zhou et al. [23] design a cross-modal and cross-scale global local attention YOLOv5 model. The model makes full use of the complementary advantages of Transformer and CNN to enhance the feature extraction ability of the lesion region. Pan et al. [24] design the Residual

Pyramid Separable Attention Module (RPSA) for yolov5, so as to balance the channel information loss caused by skip connections and further improve the performance of the network in detecting intestinal polyps. Li et al. [25] introduce a bottleneck attention module into the YOLOv5 detection head, which introduced CA attention into the bottleneck structure to effectively suppress the interference of complex background. Wang et al. [26] design a normalization-based Attention Module in YOLOv8, which uses batch Normalization to generate a scaling factor and weights the input feature map to highlight salient features and suppress the background. PRIADANA et al. [27] introduce Sequential Efficient Attention Module into C2F in YOLOv8, which adds a large kernel decomposition convolution block after ECA to extract spatial features. The salient features in the channel and space can be extracted by this method.

3. Methodology

3.1 Pooling Attention Binary Relationship

Pooling attention is mainly used to model the dependencies between Spaces or channels. Its construction process conforms to a unified paradigm, but currently, there is still a lack of a universal representation framework to systematically describe this paradigm. To solve this problem, the pooling attention binary relation is proposed. In this section, the pooling attention binary relation framework is defined and then it is used to represent Cross-directional Regional Pooling Attention (CRPA).

In the pooling attention operation, there are pooling operation and attention weighting operation. For the pooling operation, the pooling kernel set $K = \{k_1, k_2, \dots, k_m\}$ is defined, where $k_i (i = 1,2, \dots, m)$ represents the pooling kernel, and m is the number of pooling kernels. m is determined according to the specific pooling operation. The feature map set $F = \{f_1, f_2, \dots, f_n\}$ is then defined, where $f_j (j = 1,2, \dots, n)$ represents the feature maps to be pooled, and n is the number of feature maps.

Definition 4 Pooling Ordered Couple: A two-tuple formed by elements k in the set K of pooling kernels and f in the set F of feature maps arranged in a certain order, denoted as $< k, f >$. Where k is the first element and f is the second element.

The set of all ordered couples is called the Cartesian product about the pooling kernel set K and the feature map set F, the first element from the pooling kernel set K and the second element from the feature map set F in ordered couples, denoted as $K \times F = \{< k, f > | k \in K \wedge f \in F\}$

Definition 5 Pooling Binary Relation: Let K be the set of pooling kernel and F be the set of feature map, $\forall f_j \in F(j = 1,2,\dots,n)$, $\forall k_i \in K(i = 1,2,\dots,m)$ and $< k_i, f_j > (i = 1,2,3,\dots,m, j = 1,2,3,\dots,n)$. The subset R_P of $K \times F$ consisting of these pooling ordered couples is pooling binary relation.

Due to the particularity of the pooling binary relation, there are the following properties:

(1) In the pooling binary relation, the pooling kernel and the feature map are not the same, so the pooling binary relation is anti-reflexive.

(2) In the pooling binary relation, the binary relation from the set K to the set F is different from the binary relation from the set F to the set K, so the pooling binary relation is anti-symmetric.

(3) In the pooling binary relation, only the pooling set K acts on the set F of feature maps, the pooling binary relation is anti-transitive.

In pooling attention, the pooling results are processed through a transformation operation to obtain the weight matrix of the feature map. The weight matrices corresponding to all feature maps are denoted as the set $W = \{w_1, w_2, \dots, w_n\}$, n is the number of weight matrices.

Definition 6 Weighting Ordered Couple: A two-tuple is called a weighting ordered couple, which is formed by elements w in the set W of pooling kernels and f in the set F of feature maps arranged in a certain order, denoted as $< w, f >$. Where w is the first element and f is the second element.

Definition 7 Weighting binary relation: Let W be the set of weight matrices, F be the set of feature maps, $\forall f_j \in F(j = 1,2,\dots,n)$, $\exists w_i \in K(i = 1,2,\dots,n)$, $< w_i, f_j > (i = j)$. The subset R_{rw} of $W \times F$ consisting of these weighting ordered couples is called a weighting binary relation.

Here is a further discussion on the weighting binary relation R_{rw}:

(1) In the weighting binary relation R_{rw}, $\forall f_i(i = 1,2,\dots,n) \in F$, $\exists w_j(j = 1,2,\dots,n) \in W$, $< w_j, f_i > \in R_{rw}$.

(2) In weighting binary relation R_{rw} , $\forall f_i(i = 1,2..,n), < f_i, f_i > \notin R_{rw}$, $\forall w_j(j = 1,2,\dots,n), < w_j, w_j, > \notin R_{rw}$,, so R_{rw} is anti-reflexive.

(3) In weighting binary relation R_{rw} , $\forall f_i(i = 1,2,\dots,n) \in F$, $\exists w_j(j = 1,2,\dots,n) \in W, if < w_j, f_i > \in R_{rw}$, then $< f_i, w_j > \notin R_{rw}$, so R_{rw} is anti-symmetric.

(4) In the weighting binary relation R_{rw} , because $\exists w_j(j = 1,2,\dots,n), < w_j, f_i > \in R_{rw}, \nexists w_k(k = 1,2,\dots,n), < f_i, w_k > \in R_{rw}$, so $\nexists < w_j, w_k > \in R_{rw}$, R_{rw} is anti-transitivity.

The pooling attention binary relation is used to represent CRPA. The specific process of CRPA is shown in Figure 3. In the pooling, it divides the location region along the diagonal. Each pooling domain includes horizontal axis-vertical axis features. The pooling domain includes Horizontal axis-Vertical axis pooling, Horizontal axis pooling, and Vertical axis pooling. To simplify the expression, the feature map size is 4×4, and the pooling is taken as an example of Horizontal axis-Vertical axis pooling. Firstly, its pooling kernel set K and feature map set F are defined, as shown in Formula.(1) and Formula.(2) respectively

$$K = \{ \quad , \quad , \quad , \quad \} \tag{1}$$

$$F = \{ \quad , \quad , \ldots , \quad \} \tag{2}$$

According to Definition 2, the Cartesian product of the pooling kernel set K and the image set F are shown in Formula.(3).

$$K \times F = \{< \quad , \quad >, < \quad , \quad >, \ldots, < \quad , \quad >, < \quad , \quad >\} \tag{3}$$

Where$<$ $,$ $>$is the pooling ordered couple, the number of elements in the set $K \times F$ is $4 \times n$, and n is the number of feature maps. In the pooling operation, The first feature map is taken as an example, then the pooling binary relationship of the pooling operation is shown in Formula.(4) :

$$R_P = \{< \quad , \quad >, < \quad , \quad >, < \quad , \quad >, < \quad , \quad >\} \tag{4}$$

The relation matrix is another description of the relation, let $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, respectively, and let r be the relation from A to B.

$$R_{ij} = \begin{cases} 1 & if \quad a_i \; r \; b_i \\ 0 & otherwise \end{cases} \tag{5}$$

In order to clearly represent the pooling operation of the pooling kernel set K on the feature map set F, this paper presents the relation matrix of this pooling binary relation, as shown in Figure 1.

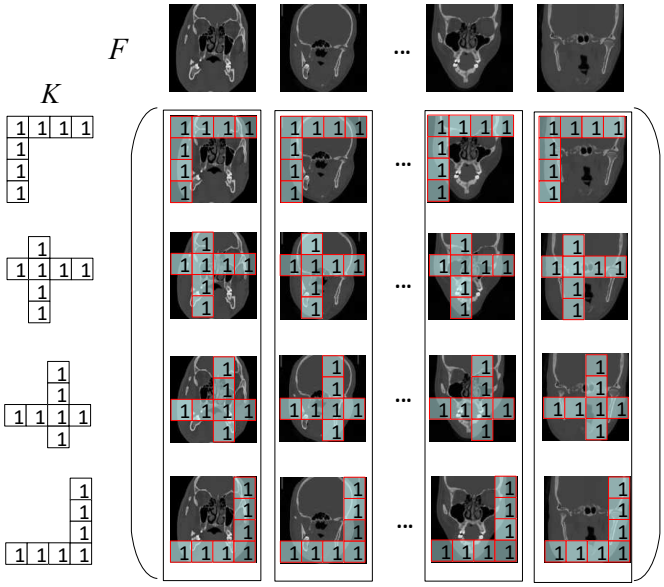


Figure 1. Cross-location Region-pooling relationship matrix

As shown in Figure 1, this relation matrix is an all-ones matrix , so this pooling binary relation is a universal relation. The pooling results of CRPA are the Horizontal axis-Vertical axis pooling plus the Horizontal axis pooling and the Vertical axis pooling, respectively, to obtain a couple of one-dimension pooling results. This result is encoded by two branches to obtain a couple of weight matrices, denoted as $[w^H, w^V]$. Then the weight matrix set W of CRPA is shown in Formula.(6):

$$W = \{[w_1^H, w_1^V], [w_2^H, w_2^V], \dots, [w_n^H, w_n^V]\} \tag{6}$$

According to the definition of weighting binary relation, the CRPA weighting binary relation is shown in Formula.(7) :

$$R_{rw} = \{< [w_1^H, w_1^V], \text{} >, < [w_2^H, w_2^V], \text{} >, \dots, < [w_n^H, w_n^V], \text{} >\} \tag{7}$$

In order to clearly represent the weighting binary relationship between the set of weight matrices W and the feature map F, this paper gives the weighted relationship matrix of CRPA, as shown in Figure 2. The weight matrix of W is $w_j^H \times w_j^V$ (j=1,2,... ,n).

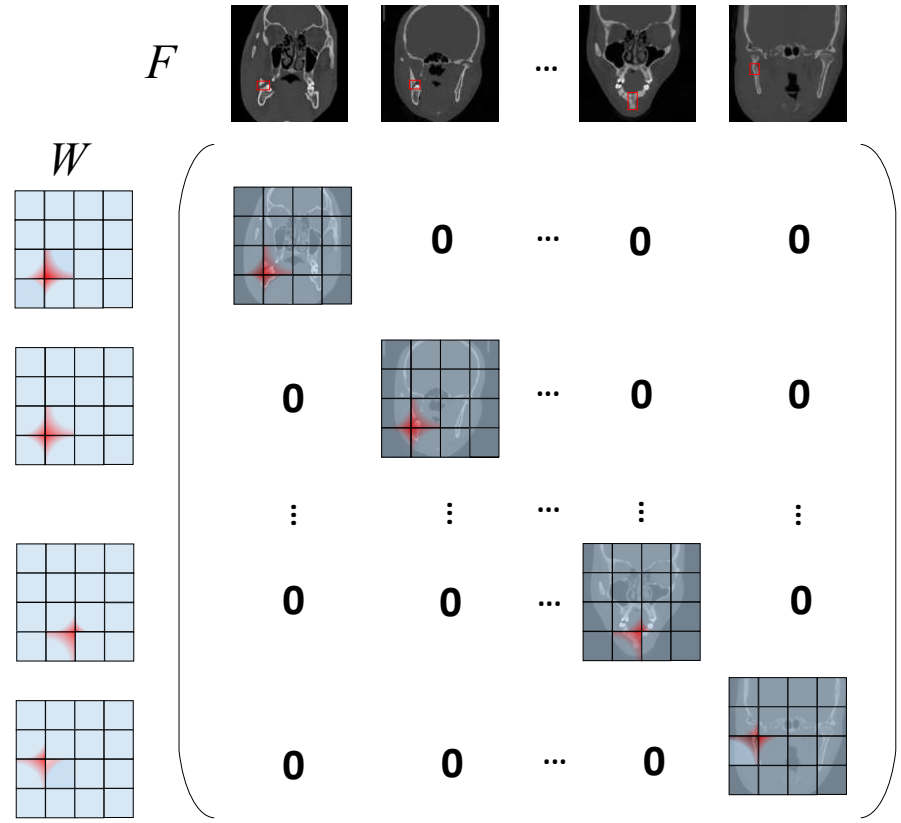


Figure 2. Cross-directional Regions Pooling Attention Weighting binary relation matrix. The red region is the region with larger weight value, which represents the most concerned region of the model.

3.2 Cross-directional Regional Pooling Attention

To address the challenges of subtle fracture lines, complex fracture morphologies, and uneven spatial distributions in mandibular fracture detection, a Cross-directional Regional Pooling Attention (CRPA) mechanism is proposed, as illustrated in Figure. 3. The proposed module is centered on cross-directional regional pooling, which enhances the spatial dependency modeling capability of feature maps by establishing correlations among regions from different directions. Specifically, the input feature map is first processed by horizontal regional pooling and vertical regional pooling to obtain decoupled directional representations while preserving positional information in each direction. Subsequently, a cross-directional regional pooling representation is constructed based on the two directional pooling features to establish correlations between the horizontal and vertical regions. The cross-directional features are then fed back to enhance the representations of both directions, enabling unified modeling of directional feature decoupling and cross-directional relation coupling. Finally, the encoded features are used to generate attention weights for recalibrating the input feature map, thereby enhancing the model's ability to perceive

critical regions, such as fracture boundaries, disrupted structures, and subtle fracture lines. The detailed implementation of the proposed module is described in the following sections.

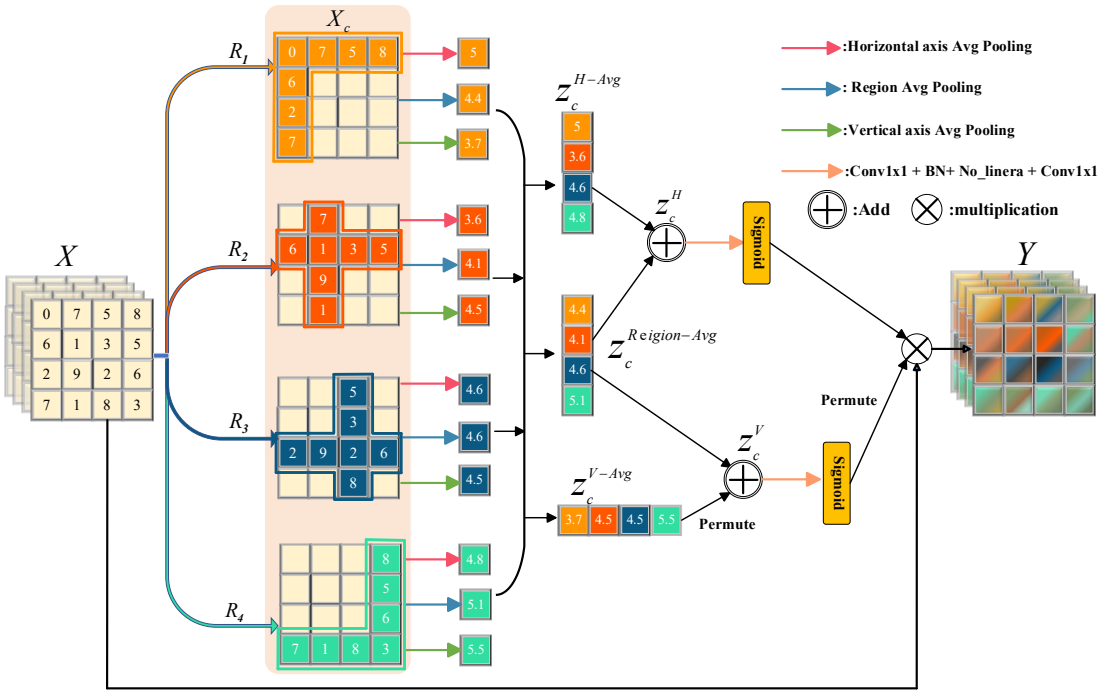


Figure 3. Cross-directional Regions pooling Attention modules. The size of the input feature map is taken 4×4. The figure depicts the calculation process of the c^{th} channel.

3.2.1 Cross-directional Regional Pooling

Channel attention mechanisms typically employ global pooling to capture global contextual information. However, global pooling tends to lose spatial positional information. Coordinate Attention (CA) alleviates this issue to some extent by encoding positional information through horizontal and vertical pooling, respectively. Nevertheless, the two directional pooling operations are performed independently, lacking information interaction between directions and thus failing to adequately model the relationships among regions from different directions. To address this limitation, we propose a cross-directional regional pooling strategy. Built upon horizontal and vertical pooling, the proposed strategy introduces horizontal-vertical cross-directional pooling to establish correlations among regions from different directions, thereby enabling the interactive modeling of directional information. The entire pooling process consists of three steps: (1) horizontal-vertical cross-directional pooling, (2) horizontal and vertical pooling, and (3) fusion of the three pooled features to generate the feature representations for subsequent attention weight

generation.

Step 1: Cross-directional regional pooling**.**

In this pooling operation, given an input feature map $\in R^{C\times H\times W}$, the horizontal-vertical pooling result of the k-th $(k = 1,2,3,4)$region in the c-th channel is defined as follows:

$$z_c^{Region-Avg}(R_k) = \frac{1}{H+W}\left[\sum_{0\le i<W} X_c(R_k, i) + \sum_{0\le j<H} X_c(j, R_k)\right] \tag{8}$$

Cross-directional regional pooling simultaneously aggregates regional feature responses along the horizontal and vertical directions. By explicitly establishing correlations between the two directions, it enables interactive modeling of cross-directional regional information and provides richer spatial structural information for subsequent directional feature enhancement and attention weight generation.

Step 2: Horizontal Pooling and Vertical Pooling.

The pooling operation, given an input feature map $X \in R^{C\times H\times W}$, the horizontal pooling result of the k-th $(k = 1,2,3,4)$region in the c-th channel can be formulated as Eq. (9), while the vertical pooling result can be formulated as Eq. (10):

$$z_c^{H-Avg}(R_k) = \frac{1}{W}\sum_{0\le i<W} X_c(R_k, i) \tag{9}$$

$$z_c^{V-Avg}(R_k) = \frac{1}{H}\sum_{0\le j<H} X_c(j, R_k) \tag{10}$$

Horizontal pooling and vertical pooling independently aggregate regional features along the two directions, enabling decoupled directional representations and allowing the model to preserve spatial positional information in each direction.

Step 3: Cross-directional Regional Pooling Feature Fusion.

To simultaneously preserve directional positional information and enhance the spatial correlations between different directions, the horizontal-vertical pooling result of each region is fused with its corresponding horizontal pooling and vertical pooling results, respectively, thereby generating the final cross-directional regional pooling representation for the c-th channel of X. This process is formulated in Eqs. (11) and (12):

$$z_c^H = z_c^{Region-Avg} + z_c^{H-Avg} \tag{11}$$

$$z_c^V = permute\left(z_c^{Region-Avg}\right) + z_c^{V-Avg} \tag{12}$$

where $z_c^H \in R^{1\times H\times 1}$, $z_c^V \in R^{1\times 1\times W}$. Extending this process to all channels of the feature map X, the outputs of cross-directional regional pooling can be obtained as $z^H \in R^{C\times H\times 1}$, $z^V \in R^{C\times 1\times W}$. By feeding the cross-directional regional relationships back into the horizontal and vertical feature representations, the proposed fusion strategy achieves unified modeling of directional feature decoupling and cross-directional relation coupling. Consequently, it strengthens the spatial dependencies between different directions while preserving precise positional information, thereby providing richer feature representations for subsequent attention weight generation.

3.2.2 Cross-directional Regional Pooling Attention Block

To fully exploit the directional features and regional correlation information aggregated by cross-directional regional pooling, we propose a Cross-directional Regional Pooling Attention Block (CRPAB). Built upon the outputs of cross-directional regional pooling, the proposed block employs a dual-branch architecture to separately encode the horizontal and vertical directional features. By integrating cross-directional regional correlations while preserving precise spatial positional information, CRPAB achieves unified modeling of directional feature decoupling and spatial dependency coupling. Subsequently, the encoded bidirectional features are utilized to generate the corresponding attention weights, which are applied to the input feature map for adaptive enhancement of critical regions, such as fracture boundaries, disrupted structures, and subtle fracture lines. As a result, the proposed module improves the feature representation capability of the model for complex fracture targets. The overall architecture of the proposed attention block is illustrated in Fig. 4.

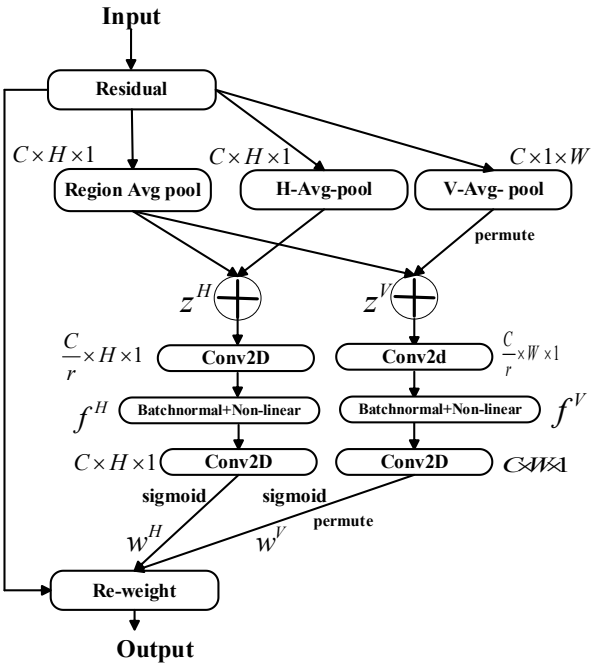


Figure 4. Cross-directional Regional Pooling Attention Block (CRPAB)

Specifically, for the outputs $z^H \in R^{C\times H\times 1}$ and $z^V \in R^{C\times 1\times W}$ of Cross-directional Region Pooling, these are input into two 1×1 convolution transformation functions, respectively. As expressed in Formula (13) and (15):

$$f^H = \delta(BN(Conv(z^H))) \tag{13}$$

$$f^V = \delta(BN(Conv(z^V))) \tag{14}$$

Where $f^H \in R^{\frac{c}{r}\times H\times 1}$ and $f^V \in R^{\frac{c}{r}\times 1\times W}$ are intermediate features. r is channel reduction rate, which is used to reduce the complexity of the model. δ is the nonlinear activation function. Then, the other two 1×1 convolutions are used to transform the two intermediate feature maps into tensors with the same number of channels as the input F, respectively, the process are expressed as Formula(15) and (16).

$$w^H = \sigma\big(Conv(f^H)\big) \tag{15}$$

$$w^V = \sigma\big(Conv(f^V)\big) \tag{16}$$

Where σ is the Sigmoid function. Finally, w^H and w^V are expanded to the same size as the input feature map F and used as attention weight matrices. Thus the output Y of CRPA is expressed as formula(17)

$$Y_c(i,j) = X_c(i,j)\times w_c^H(i,j)\times w_c^V(i,j) \tag{17}$$

3.3 Implementation

For the mandibular fracture detection task, the proposed Cross-directional Regional Pooling

Attention (CRPA) module is integrated into the YOLOv8 architecture. As illustrated in Fig. 5, CRPA is embedded into the Bottleneck block to enhance the extraction of fracture-related features.

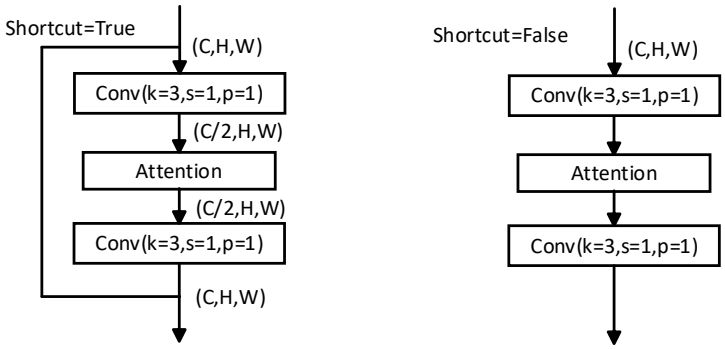


Figure 5. CRPA 模块嵌入 Bottleneck 结构示意图.

4. Experiments

4.1 Experiment Setup

Experimental environment: The experiments are conducted on a workstation running the 64-bit Windows 10 Pro operating system, equipped with 256 GB of RAM, an Intel Xeon Gold 6154 3.0 GHz CPU, and an NVIDIA Titan RTX GPU with 24 GB of memory. The deep learning environment is based on CUDA 12.2, Python 3.8, and PyTorch 1.12.0.

Parameter settings: During training, the model parameters are optimized using the Stochastic Gradient Descent (SGD) optimizer. The initial learning rate is set to 0.1, the minimum learning rate is set to 0.001, and the weight decay is set to 0.005.

4.2 Datasets

Two datasets are used in the experiments. The first dataset was a mandibular fracture CT dataset collected from the Department of Oral and Maxillofacial Surgery of a tertiary hospital in Ningxia, China. It consists of 2,384 mandibular fracture CT images and is categorized into three fracture types according to the anatomical location of the fracture: mandibular body fractures, mandibular angle fractures, and condylar fractures. Each CT slice contains one fracture region. The fracture regions were manually annotated by experienced radiologists using the LabelImg tool, and the corresponding annotations were stored in XML format. To improve the generalization ability of the model and increase the difficulty of the dataset, random flipping and Gaussian noise augmentation are applied to the images. The augmented images replace the original ones, ensuring that the data splitting procedure does not assign the same original image and its augmented variants to both the training and test sets simultaneously. The dataset is divided into training, validation, and

test sets with proportions of 70%, 15%, and 15%, respectively.

The other dataset is luna16, which is from the 2016 Lung Nodule Analysis competition and is a subset of the largest public lung nodule dataset LIDC-IDRI. LIDC-IDRI includes 1018 low-dose CT images of the lung. LIDC-IDRI deletes CT images with slice thickness greater than 3mm and lung nodules smaller than 3mm, and the LUNA16 dataset is left. This dataset contains the location and diameter annotations of 1186 lung nodules in the detection task. The dataset is split into training, validation, and test sets with a ratio of 70%, 20%, and 10%, respectively.

4.2.1 Ethics approval and informed consent

This study was reviewed and approved by the Biomedical Research Ethics Committee of North Minzu University (Approval No. 2025-16; approved on 11 March 2025). All methods were carried out in accordance with relevant institutional guidelines, applicable regulations, and ethical standards. This study did not directly recruit human participants. The medical imaging data used in this study were obtained from publicly available datasets and hospital-collected datasets. According to the approved ethics review documentation, the hospital-collected medical imaging data had undergone ethical review by the collaborating institution before being used in this study. All patient-related medical images were de-identified prior to analysis to protect patient privacy and confidentiality.

Under the protocol approved by the Biomedical Research Ethics Committee of North Minzu University, no additional informed consent was required for the secondary analysis of these de-identified medical imaging data.

4.3 Comparison with Other Methods

To validate the effectiveness of the proposed Cross-directional Regional Pooling Attention (CRPA) module for mandibular fracture detection, comparative experiments are conducted on the Mandibular fracture CT dataset using YOLOv8s as the baseline model. Under the same network architecture and training settings, CRPA is compared with several representative attention mechanisms, including SE, ECA, CBAM, CA, FCA, EMA, and Conv2Former. The batch size and the number of training epochs are set to 4 and 100, respectively. For a fair comparison, the reduction ratio r is uniformly set to 16 for all attention modules. The experimental results are presented in Table 1.

Table 1 Comparison results on the mandibular fracture CT dataset

Settings	Pram(M)	mAP50	mAP50-90	Recall	Precision	F1
Baseline(YOLOv8s)	11.13	93.80%	44.40%	86.40%	92.80%	89.52%
+SE	11.15	94.00%	43.80%	88.80%	91.60%	90.23%
+ECA	11.13	94.60%	46.30%	91.00%	92.90%	91.94%
+CBAM	11.55	94.44%	46.40%	90.10%	91.90%	91.00%
+CA	11.17	94.60%	45.55%	90.80%	91.50%	91.15%
+FCA[28]	11.15	94.50%	45.90%	90.80%	92.70%	91.74%
+EMA[29]	11.14	94.70%	45.10%	90.10%	91.50%	90.79%
Conv2former[30]	11.92	94.40%	43.90%	90.80%	92.10%	91.44%
+CRPA	11.17	95.90%	45.80%	90.80%	94.80%	92.73%

Table 1 presents the comparison results of CRPA and different attention mechanisms on the mandibular fracture CT dataset. It can be observed that the introduction of attention modules consistently improves the detection performance over the baseline model, indicating that pooling attention effectively enhances the representation of fracture regions. Among all the compared methods, the proposed CRPA achieves the best overall detection performance. With only 11.17M parameters, CRPA improves mAP50 from 94.60% to 95.90%, mAP50-95 from 45.55% to 45.80%, Precision from 91.50% to 94.80%, and F1-score from 91.15% to 92.73%, representing improvements of 1.30%, 0.25%, 3.30%, and 1.58%, respectively, compared with CA. Compared with the baseline YOLOv8s, CRPA improves mAP50, Recall, Precision, and F1-score by 2.10%, 4.40%, 2.00%, and 3.21%, respectively. These results demonstrate that, by introducing cross-directional regional associations and achieving unified modeling of directional feature decoupling and cross-directional relation coupling, CRPA is able to preserve positional information more effectively while enhancing spatial dependencies without increasing model complexity, thereby improving the detection performance for complex fracture structures and subtle fracture lines.

To further validate the feature modeling capability of CRPA for fracture regions, Grad-CAM is employed to visualize the feature responses of different models, as shown in Fig. 6. It can be observed that although the baseline model and other pooling attention methods are able to focus on the fracture regions, varying degrees of background activation and dispersed attention are still present. In contrast, the heatmaps generated by CRPA are more consistently concentrated on the fracture regions marked by the red boxes across all three samples.

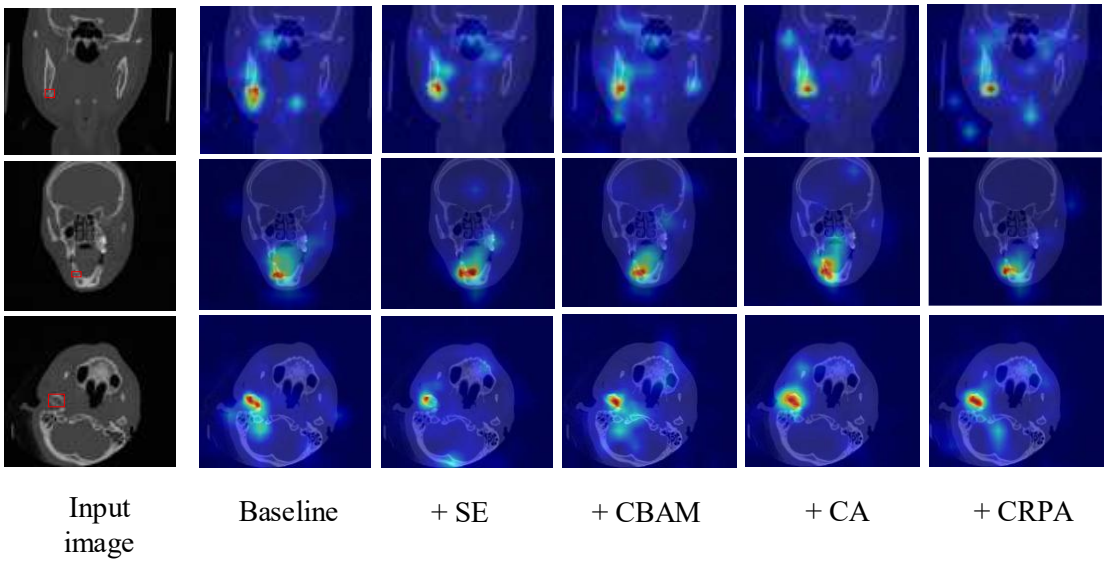


Figure 6. Heat map .the red box is the fracture region label. It is seen from the figure that the proposed attention pay more attention to the target region.

To further evaluate the generalization capability of the proposed CRPA module, experiments are also conducted on the LUNA16 dataset. Similarly, YOLOv8s is adopted as the baseline model, and CRPA is compared with several representative attention mechanisms, including SE, ECA, CBAM, CA, FCA, EMA, and Conv2Former, under the same network architecture and training settings. The batch size and the number of training epochs are set to 4 and 100, respectively. The experimental results are presented in Table 2.

Table 2 Comparison results on the LUNA16 dataset。

Settings	Pram(M)	mAP50	mAP50-90	Recall	Precision	F1
Baseline(YOLOv8s)	11.13	47.90%	21.40%	51.13%	56.50%	53.68%
+SE	11.15	60.10%	25.60%	60.50%	68.00%	64.03%
+ECA	11.13	62.70%	26.80%	55.50%	78.60%	65.06%
+CBAM	11.55	61.30%	28.50%	59.70%	67.20%	63.24%
+CA	11.17	63.90%	26.80%	58.80%	69.30%	63.62%
+FCA	11.15	61.00%	25.30%	58.80%	75.80%	66.23%
+EMA	11.13	59.90%	26.80%	58.80%	67.70%	63.01%
Conv2former	11.92	62.10%	26.20%	56.20%	76.10%	64.65%
+CRPA	11.17	65.30%	27.80%	63.90%	72.60%	67.95%

The experimental results are presented in Table 2. It can be observed that CRPA achieves the best overall detection performance, with mAP50, Recall, and F1-score reaching 65.30%, 63.90%, and 67.95%, respectively, outperforming all the compared methods. Compared with the most competitive method, CA, CRPA improves mAP50 and F1-score by 1.40% and 4.33%, respectively. In addition, compared with FCA, which achieves the second-highest F1-score, CRPA further

improves the F1-score by 1.72%. These results demonstrate that the unified modeling of directional feature decoupling and cross-directional relation coupling enables CRPA to preserve positional information while enhancing spatial dependencies across different directions, thereby improving detection performance on different datasets and exhibiting strong generalization capability.

4.4 Ablation Experiments

To investigate the effect of different reduction ratios $r$on the performance of CRPA, ablation experiments are conducted on the mandibular fracture CT dataset with r set to 8, 16, and 32, respectively. The experimental results are presented in Table 3.

Table 3 Ablation study on the reduction ratio r

Settings	Pram(M)	mAP50	mAP50-90	Recall	Precision	F1
Baseline(YOLOv8s)	11.13	93.80%	44.40%	86.40%	92.80%	89.52%
+CRPA_8	11.21	94.60%	44.70%	89.60%	91.19	90.39%
+CRPA_16	11.17	95.90%	45.80%	90.80%	94.80%	92.73%
+CRPA_32	11.16	95.30%	**46.40%**	**91.60%**	**94.20%**	**92.74%**

As shown in Table 3, all reduction ratios improve the detection performance of the model, demonstrating the robustness of the proposed CRPA module. Among them, when the reduction ratio $r$is set to 16, the model achieves the best overall performance, with mAP@50, Precision, and F1-score reaching 95.90%, 94.80%, and 92.73%, respectively. Although setting $r = 32$yields slight improvements in mAP@50:95, Recall, and F1-score, it leads to decreases in mAP@50 and Precision. Considering the overall detection performance and model complexity, the reduction ratio $r$is finally set to 16 in this study.

5. Conclusion and Future Work

To address the challenges of subtle fracture lines, complex fracture morphology, and uneven spatial distribution in mandibular fracture CT images, research is conducted on pooling attention mechanisms and YOLO-based object detection models. Unlike previous studies, binary relations are introduced into pooling attention, and a Pooling Attention Binary Relation (PABR) framework is proposed. This framework provides a new perspective for applying binary relations to deep learning and addresses the lack of a unified representation framework for pooling attention mechanisms. Furthermore, based on an analysis of existing pooling attention methods, a Cross-directional Regional Pooling Attention (CRPA) module is developed under the PABR framework. Rather than extending existing pooling attention mechanisms by introducing additional operations,

CRPA proposes a novel pooling strategy, providing new insights into the design of pooling mechanisms. Owing to its simple and lightweight architecture, CRPA can be seamlessly integrated into YOLO models. By incorporating CRPA into the YOLOv8 network, an improved model for mandibular fracture detection is constructed. The experimental results demonstrate that the proposed method achieves an mAP50 of 95.90%, an mAP50-95 of 45.80%, a Recall of 90.80%, a Precision of 94.80%, and an F1-score of 92.73% on the mandibular fracture CT dataset. Compared with the baseline YOLOv8s, improvements of 2.10%, 4.40%, and 3.21% are achieved in mAP50, Recall, and F1-score, respectively, demonstrating superior detection performance. Moreover, compared with representative attention mechanisms, including SE, ECA, CBAM, CA, FCA, EMA, and Conv2Former, CRPA achieves the best or competitive overall performance. To further validate the generalization capability of CRPA, additional experiments are conducted on the LUNA16 dataset. The results indicate that CRPA consistently outperforms existing pooling attention methods across different datasets, demonstrating that the proposed cross-directional regional pooling strategy possesses strong robustness and generalization capability, while also highlighting its potential for application to a wide range of medical object detection tasks.

Future research are further extended in the following aspects, In the pooling attention binary relation, the set F is the feature map, W is the weight matrix, and then the weighting binary relation from W to F is defined. With the emergence of new pooling attention, the set W and weighting relation R_{rw} are redefined to accommodate more diverse weighting ways. In addition, the binary relation not only represents pooling attention, but also extends to other operations in deep learning, such as convolution operations and various types of attention mechanisms, by flexibly defining two sets. More comprehensive regional relationship modeling strategies for CRPA should be further investigated. The current study focuses on cross-directional regional relationships between the horizontal and vertical directions. Future work may explore relationship modeling strategies based on diagonal directions, multi-scale regions, and dynamic region partitioning to further exploit the spatial structural information contained in feature maps and enhance the representation capability of CRPA.

Author contributions

T.Z. contributed to the conceptualization, supervision, and review of the manuscript. S.Y. designed the methodology, implemented the model, conducted the experiments, analyzed the results, prepared the figures, and wrote the original manuscript. K.C. contributed to data curation, validation, and manuscript revision. H.L. contributed to resources, clinical guidance, supervision, and manuscript revision. K.S. contributed to methodology, software implementation, experimental validation, visualization, and manuscript revision. All authors reviewed and approved the final manuscript.

Corresponding author：Shiyi Yu(Email: 20247520@stu.nmu.edu.cn)

Funding Declaration

This research was funded by the National Natural Science Foundation of China (Nos. 62561002 and 62576009), the Ningxia Natural Science Foundation Project (No. 2025AAC020006), and the Ningxia Leading Talent Cultivation Project for Science and Technology (No. 2025GKLRLX26).

Data availability

The mandibular fracture CT dataset generated and/or analyzed during the current study is not publicly available because it contains clinical imaging data and is subject to patient privacy, ethical, and institutional data-sharing restrictions. Researchers interested in accessing the dataset may submit a reasonable request to the corresponding author, Shiyi Yu (20247520@stu.nmu.edu.cn). Requests will be reviewed in accordance with applicable ethical and institutional requirements. Access may be provided only for legitimate academic and scientific research purposes. Approved users must protect patient confidentiality, must not attempt to identify or re-identify individuals, and must not redistribute the data to third parties without prior authorization.

The LUNA16 dataset analyzed during the current study is publicly available from the LUNA16 Grand Challenge repository at https://luna16.grand-challenge.org/Data/.

References

[1] Zhou, Z., Luo, S., Chen, J., et al. (2025). Finite element analysis of bioabsorbable plates versus miniature titanium plates in mandibular fracture fixation in different bone qualities. Chinese Journal of Tissue Engineering Research, 29(4), 818.

[2] Zhou, T., Wang, H., Chen, K. X., Zhang, Y., & Li, M. (2025). Mandible-YOLO: The fracture region is detected only once. Biomedical Signal Processing and Control, 106, 107724.

[3] Chien, C. T., Ju, R. Y., Chou, K. Y., Hsu, Y. H., & Lin, C. C. (2024). YOLOv9 for fracture detection in pediatric wrist trauma X-ray images. Electronics Letters, 60(11), e13248.

[4] Meza, G., Ganta, D., & Gonzalez Torres, S. (2024). Deep learning approach for arm fracture detection based on an improved YOLOv8 algorithm. Algorithms, 17(11), 471.

[5] Zhou, T., Wang, H., Du, Y., Liu, F., Guo, Y., & Lu, H. (2024). M3YOLOv5: Feature enhanced YOLOv5 model for mandibular fracture detection. Computers in Biology and Medicine, 173, 108291.

[6] Yari, A., Fasih, P., Hosseini Hooshiar, M., et al. (2024). Detection and classification of mandibular fractures in panoramic radiography using artificial intelligence. Dentomaxillofacial Radiology, 53(6), 363–371.

[7] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7132-7141). IEE.

[8] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11534-11542). IEEE.

[9] Woo, S., Park, J., Lee, Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19). Springer.

[10] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13708-13717). IEEE.

[11] Bayraktar, E., & Yigit, C. B. (2024). Conditional-pooling for improved data transmission. Pattern Recognition, 145, 109978.

[12] Mishra, N. K., Singh, P., Gupta, A., & Joshi, S. D. (2025). PP-CNN: Probabilistic pooling CNN for enhanced image classification. Neural Computing and Applications, 37(6), 4345-4361.

[13] Pratik, S., Sharma, P., Nayak, D. R., & Balabantaray, B. K. (2025). WMCF-Net: Wavelet pooling-based multiscale contextual fusion network for polyp classification. Biomedical Signal Processing and Control, 107, 107727.

[14] Zhao, L., & Zhang, Z. (2024). A improved pooling method for convolutional neural networks. Scientific Reports, 14(1), 1589.

[15] Özdemir, C., Doğan, Y., & Kaya, Y. (2024). A new local pooling approach for convolutional neural network: Local binary pattern. Multimedia Tools and Applications, 83(12), 34137-34151.

[16] Zhang, J., & Slamu, W. (2024). Partial channel pooling attention beats convolutional attention. Expert Systems with Applications, 237, 121436.

[17] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., et al. (2023). Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[18] Sun, L., Cheng, S., Zheng, Y., Wu, Z., & Zhang, J. (2022). SPANet: Successive pooling attention network for semantic segmentation of remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 4045-4057.

[19] Wang, X., Kang, M., Chen, Y., Jiang, W., Wang, M., Weise, T., et al. (2023). Adaptive local

cross-channel vector pooling attention module for semantic segmentation of remote sensing imagery. Remote Sensing, 15(8), 1980.

[20] Li, X., Wu, Y., Fu, Y., Tang, C., & Zhang, L. (2022). Neighbour feature attention-based pooling. Neurocomputing, 501, 285-293.

[21] Xiao, J., Guo, H., Zhou, J., Zhao, T., Yu, Q., Chen, Y., et al. (2023). Tiny object detection with context enhancement and feature purification. Expert Systems with Applications, 211, 118665.

[22] Zhao, Z., & He, P. (2024). YOLO-Mamba: Object detection method for infrared aerial images. Signal, Image and Video Processing, 18(12), 8793-8803.

[23] Zhou, T., Liu, F., Ye, X., Wang, H., & Lu, H. (2023). CCGL-YOLOV5: A cross-modal cross-scale global-local attention YOLOV5 lung tumor detection model. Computers in Biology and Medicine, 165, 107387.

[24] Pan, X., Mu, Y., Ma, C., & He, Q. (2024). TFCNet: A texture-aware and fine-grained feature compensated polyp detection network. Computers in Biology and Medicine, 171, 108144.

[25] Li, Y., Feng, Y., Zhou, M., Xiong, X., Wang, Y., & Qiang, B. (2024). DMA-YOLO: Multi-scale object detection method with attention mechanism for aerial images. The Visual Computer, 40(6), 4505–4518.

[26] Wang, H., Liu, J., Zhao, J., Zhang, J., & Zhao, D. (2025). Precision and speed: LSOD-YOLO for lightweight small object detection. Expert Systems with Applications, 269, 126440.

[27] Priadana, A., Nguyen, D. L., Vo, X. T., Choi, J., Ashraf, R., & Jo, K. (2025). HFD-YOLO: Improved YOLO network using efficient attention modules for real-time one-stage human fall detection. IEEE Access, 13, 41248–41258.

[28] Qin, Z., Zhang, P., Wu, F., & Li, X. (2021). Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 783-792).

[29] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z. (2023). Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1–5). IEEE.

[30] Hou, Q., Lu, C. Z., Cheng, M. M., & Feng, J. (2024). Conv2former: A simple transformer-style convnet for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 46(12), 8274-8283.